# USING CONTEXT TO INTEGRATE HYPERMEDIA WITH INFORMATION RETRIEVAL SYSTEMS

by

Rui Miguel Neto Marinheiro

A thesis submitted for the degree of
Doctor of Philosophy

Department of Electronics and Computer Science,
University of Southampton,
United Kingdom

November 2000

UNIVERSITY OF SOUTHAMPTON

# ABSTRACT

FACULTY OF ENGINEERING

ELECTRONICS AND COMPUTER SCIENCE DEPARTMENT

Doctor of Philosophy

## USING CONTEXT TO INTEGRATE HYPERMEDIA WITH INFORMATION RETRIEVAL SYSTEMS

by Rui Miguel Neto Marinheiro

Hypermedia and information retrieval take different approaches in accessing information. For this reason it makes sense to integrate both systems in one where advantages can be enhanced and disadvantages eliminated.

This thesis presents the evolution and implementation of a new model of integration between these two information-seeking approaches both at document and abstract access levels bearing in mind the use of more structured information usually associated with hypermedia and more content information usually associated with information retrieval. In the development, an open and flexible information retrieval tool is considered as part of the model. Special care is taken with the modelling of the multimedia context indexing by considering hypermedia structure metadata from both document level and abstract level. Overall, this process enables a smoother transition between navigation and retrieval, whatever media is considered.

A method of evaluating has been proposed and performed in order to find metadata weight assignment strategies for context indexing and retrieval of multimedia. The fact that different hypermedia applications and their authors use different hypermedia development facilities and different authoring styles is taken into consideration by using five distinct test collections.

Future work and implications on the Web are also discussed.

# List of contents

# List of figures

# List of tables

# List of equations

# Acknowledgement

There are important things in life that we usually take for granted, and only too late do we come to know this. At the beginning of my stay in England it was quite hard. It is not the place that you miss but also the people you are used to interact with in everyday life. In Portuguese there is a quite unique word for defining this feeling of longing: *saudade*. In the same way I have arrived I will be leaving. Some parts of me will be left behind, but many more will be taken with me.

First of all I am taking with me an excellent example of supervision. Prof Wendy Hall was always able to have a smile for at the most difficult times and encourage me with kind words. Without her understanding, help and wisdom my research would not be possible.

Friends will always be friends, and they will stay forever with me. My personal growth during my stay in England was made with then. Making friendship with people from all around the world helps us to be more tolerant with others and also to understand our own culture.

Only away from one's family one understands how important they are to us. They are always there and always love us whatever happens. It is quite important to have this, and even away from them they could support me with love.

There are no words that can define how important Linda is for me. Without being incongruent I only want to say this and she knows why: "Linda, you are fantastic".

Finally I would like to thank the Portuguese Foundation for Science and Technology for their financial support during my PhD, under the program PRAXIS XI grant BD/5399/95.

# Chapter 1 - Introduction

## 1.1 Overview of research

Hypermedia and information retrieval have a different approach to the way they access information. They have advantages that can be further enhanced and disadvantages that can be eliminated by integrating these two approaches. This integration should help the user in navigation and in retrieval and increase the effectiveness and efficiency of each approach whichever the considered media.

Much work has been done on the integration of information retrieval systems with hypermedia systems using different approaches. These approaches have basically considered the hypermedia and information retrieval access in two ways:

At one end we see information retrieval improvements using hypermedia information presented at a document level. For these approaches the information retrieval functionality is then used for the benefit of hypermedia navigation. We see early examples of this in research made by Frisse (1988), Frei et al. (1992), Guinan et al. (1992) and Li et al. (1993). Navigational hypermedia at document level has also enabled multimedia information retrieval as seen in Dunlop et al. (1993).

At the other end, in works like Agosti et al. (1992), Arents et al. (1996) and Cunliffe et al. (1997), a kind of spatial or taxonomic hypermedia functionality at an abstract level was used to retrieve documents through navigation. The cognitive overhead required by the user to operate the information-retrieval mechanism is diminished by this kind of integration, since it is not necessary to implement a detailed retrieval strategy. In these approaches, it is the hypermedia functionality that is used for the benefit of information retrieval.

With this in mind we then developed an integration of hypermedia with information retrieval (Marinheiro et al. 1998) that used meta-information from both document and abstract levels to enable a better access to multimedia information. The multimedia access was made possible by using text descriptors. However their construction used a richer set of meta-information than that available in Dunlop et al. (1993), and multimedia retrieval was possible even in the absence of document level links. This integration has also been implemented in a transparent way and has used an existing text retrieval tool.

Information seeking in the Web is at present also generally divided into two different kinds of tools used to help the user in navigation. Altavista is perhaps the best known example of keyword/query based search engines, while Yahoo is perhaps the best known example of subject-classified lists.

In recent research (Marchiori 1997, Brin et al. 1998, Kaindl et al 1998, Mizuuchi et al. 1999) hypermedia at document level is still used for the improvement of information retrieval in keyword/query-based searches and (Amato et al. 1998, Mukherjea et al. 1999) abstract level navigation is used for ease of access to information in subject-classified lists.

The good preliminary results with our first integration and some conclusions drawn from its problems led us to develop a better model for the integration of information retrieval and hypermedia. For the same reason as with the first model, most of this better model has been implemented with the Microcosm system (Hall et al. 1996) since it makes it possible to have a richer set of metadata. This second model for open information retrieval integration allows more flexibility in information access, manipulation and retrieval. The merging of multiple indexes is possible.

This integration made at both document and abstract levels of navigation had in mind the use, for both access approaches, of more structure information usually associated with hypermedia and more content information usually associated with information retrieval.

We allowed both the content and context (used for multimedia access) indexing of information. Along with this, the context indexing of links was enabled, permitting a smoother transition between navigation and retrieval, whichever media was considered. Abstract level improvements are suggested and multimedia context indexing has also been refined and implemented with a better use of document and abstract level metadata.

Much work has also been done by other researchers in using neighbour documents, textual regions, labels or captions to index and retrieve multimedia (Frankel 1996, Smith 1997, Harmandas et al. 1997, Amato et al. 1998, Mukherjea et al. 1999, Srihari et al. 1999). All of these methods used different metadata retrieved from web pages, and in one way or another assigned a weight to different elements. However the weight assignment was unclear because of a lack of proper evaluation.

The goals of our evaluations where then to find strategies for metadata weight assignment to provide better context indexing of multimedia. Mukherjea et al. (1999) has done some work on this for the Web. However we go further by considering that different authors have different development styles and for this reason different weighting strategies have to be found.

Standards for document level navigation in the web have recently been proposed (XLink DeRose et al 2000, XPoint Daniel et al. 2000) and resource description standards have been also proposed (Dublin Core, Weibel et al. 1998; RDF, Lassila et al 1999) for abstract level navigation. With these implemented in XML (Bray et al. 1998) and with a new standard in progress (Chamberlin et al. 2000) for querying XML information, a richer Web is now possible, allowing our work to provide an even neater integration of hypermedia and information retrieval.

## 1.2 Outline of document

This thesis describes the evolution of a model for integrating hypermedia with information retrieval. An initial model was then extended to allow a better integration with our Microcosm based system. The bulk of evaluations were performed for the multimedia access facilities of this integrated version, which allowed context indexing of information.

Chapter 2 provides background information on hypermedia, notably the open hypermedia system Microcosm, and on information retrieval. Reasons for the integration of both systems are then given and this is followed by a description of past integration made at document and at abstract level, notably to allow the multimedia access of information. Some new standards for the Web are also contextualised with information retrieval querying facilities and open hypermedia navigation facilities

Chapter 3 describes our first model, which used an existing text information retrieval facility incorporated in the open hypermedia system Microcosm. The main goal of this model was to enable multimedia access to information using text descriptions for the multimedia, built using context information from the hypermedia network.

Chapter 4 describes the improvements made in our model of information retrieval and hypermedia integration, including the development of a new open information retrieval tool to allow a flexible integration with open hypermedia (the Microcosm system in this case). Structure and content, at both the abstract and the document level, is taken into account in a new system that allows a smooth transition between navigation and retrieval, enabling transparent multimedia information access.

Chapter 5 lays the foundation for the evaluation. The integrated system is presented with some user interaction examples. The text retrieval algorithms are evaluated with standard text collections, and a strategy for evaluating multimedia information retrieval is proposed.

Chapter 6 classifies different Microcosm applications according to the subject, size and availability of context information both at document level and at abstract level. This demonstrates that different applications have been developed according to different styles.

Chapter 7 presents evaluations on context link information at document level. Partial weighting strategies are shown for different metadata parameters. These weighting strategies will vary according to the style of development as shown in Chapter 6. The same pattern is followed in Chapter 8, which presents evaluations on abstract level metadata. Partial weighting strategies are also shown for different metadata at this level. Different conclusions are drawn from different application development styles.

Chapter 9 summarises all the evaluations by suggesting a global method of merging partial weighting strategies of metadata used in the context indexing of multimedia. It proves that multimedia access is optimised using different metadata weights.

With the experience obtained from our model, Chapter 10 discusses some implications for a Web-based integration of multimedia information retrieval. Directions for a new open and flexible information retrieval tool for the Web are given, adequate context indexing of multimedia on the Web is extrapolated from our model and compared with other researches, abstract level or meta-information improvement is suggested and new ways of context and content search integration in a Web integration are proposed.

Finally, Chapter 11 presents conclusions from our work, highlighting the novelties in our research. Some directions for future research are also given

Appendix A shows the link evaluation results on Microcosm applications further developed in Chapter 6, and Appendix B shows the text query selections used with each test collection.

# Chapter 2 - Background

## 2.1 Accessing information on hypermedia systems

The idea of hypermedia systems, although that name was not used at the time, goes back as early as 1945 to Vannevar Bush. He pointed out that the way information is searched in libraries by alphabetical search in subclasses of subclasses was not efficient since it did not work in the way humans are used to. He believed that human operations are done by association:

> *"With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain.*" (Bush, 1945)

He then proposed a machine called Memex that would work as a private supplement to an individual memory. The machine would have the storage capability and enough speed to allow easy consultation of all the material that the individual had previously indicated. The major concern of that machine would be the facility of associating any two items together in a way that, when one of those items was in view, we could instantly have access to the other item by easily selecting a button.

This explicit association would build defined trails along the stored items that could be organised non-sequentially and read in any way the individual would wish.

For many years the organisation of text remained sequential in spite of user needs for a better way of organising his/her information. However with the development and popularity of computer systems, better and faster ways of dealing with large amounts of information developed significantly. Thus a linear reading of documentation started to become obsolete in many applications as a poor method of searching for information. In the 1980's, many hypertext systems and models became popular, and some became commercially successful. In a good survey on hypertext/hypermedia, Conklin (1987) defines a more restricted concept of hypertext as being:

> *Windows on the screen are associated with objects in a database, and links are provided between these objects, both graphically (as label tokens) and in the databases (as pointers).* (Conklin, 1987)

Perhaps the most important aspect of hypertext/hypermedia is its ability to link information within and between documents. It is because of this ability that it is possible to have a non-linear organisation of information. However, there are other important features underlying hypermedia system.

Some of these are examined in more detail bellow:

- *Node/Document* - A basic unit or item in which any kind of information is kept, managed and presented using a window on a screen that is associated with that node. The node size may vary and its full content may be shown to the user, or just a part of it may be visible but with a scrolling facility allowing access to the entire content. A composite node may appear if there is an aggregation mechanism allowing one node to manage a collection of related nodes.

- *Links* - A logical connection between two related information items or between an information item and a node. The origin is the information item from which the connection emanates and the destination is where that connection ends. Many different types of links have been defined depending on the type of logical relation they represent.

- *Anchors* - This is the way a link is represented in a node. In a node each link is associated with an anchor that represents the information of the node used for link following. The content of this information and its location in the hypermedia system can be at different levels of functionality, within or outside the node (Davis 1995).

Traditionally, in hypermedia, links are used to connected nodes/documents in order to build a network of paths that can help the user/author on searching for information. Anchors are used to tie links with documents in some way, and they can be located with in documents, using mark-up in the data, or outside the documents, in a database. Links can also be stored with in nodes, our separately.

Unlike information-retrieval systems where the majority of research was directed to the representation of the semantic content of documents, the first hypertext systems were usually more concerned with issues related to the structure, management and presentation of documents (Agosti et al, 1996b). Later on, in some systems, concepts such as typed links and links descriptors, etc, were developed in order to give some semantics to the hypermedia network. With a semantic network of links it was then possible to help the user in navigation, solving problems associated with hypermedia systems, such as the difficulty of reaching relevant information in a large application.

However, user needs expanded further to the point where Malcon et al (1991) made a summary of the limitations of the current closed hypermedia systems. The development of open hypermedia systems was for them a key issue to be considered. In open hypermedia systems, it is possible to use any application within the hypermedia system via an appropriate protocol; no specialized representation is imposed upon the data; the data and the processes can be distributed; there is no distinction between the authors and the users and it is possible to add functionality in a easy way (Davis et al 1992, Hall et al. 1996, Grønbæk et al. 1996, Grønbæk et al. 1999).

The Dexter hypertext reference model defines a three-layer hypertext system: the *run-time* layer, the *storage* layer and the *within-component* layer (Halasz et al. 1994). Most open systems have tried to implement or at least can be compared to this reference model.

Intermedia (Yankelovich et al. 1988) was one of the first systems to achieve the separation of anchors (blocks) and links from documents. This allowed for an easy integration of hypermedia functionality onto the user's desktop environment.

Devise Hypermedia, DHM, (Grønbæk et al. 1996) is an open and extensible object-oriented hypermedia framework that explores link directionality with multiple sources or destinations, using *locspecs*, i.e., abstract specifications of location within documents. The hypermedia structures could be utilised in various ways by different users.

Hyper-G (Andrews et al. 1995) is a multi-user hypermedia system that also kept its links separately from the documents. It has a distributed architecture supported by a server that stores most of the information. An interesting feature is that it allows for keyword or whole text search of documents.

Another example of an open hypermedia system is Microcosm, first reported in Fountain et al. (1990). The following section describes this system.

### 2.1.1 The Microcosm open hypermedia system approach

Microcosm is an open hypermedia system developed at the University of Southampton. In the Microcosm model (Fountain et al. 1990, Davis et al. 1992, Hall 1994a, Hall et al. 1996) there are two ways of navigating through the information: navigation on the abstract level and hypermedia navigation on the document level, as can be seen in Figure 2-1. At the index/abstract level, the navigation is made through the classification of the documents until the relevant document is reached (Hall et al 1994b). The information about the classification of these nodes/documents is kept in a separate database that is managed by the Document Management System. This information covers thinks like logical index, document description, keywords, author name, etc. It therefore allows abstract navigation in a file manager style where directories are logical indexes and where the files are identified by a text description.

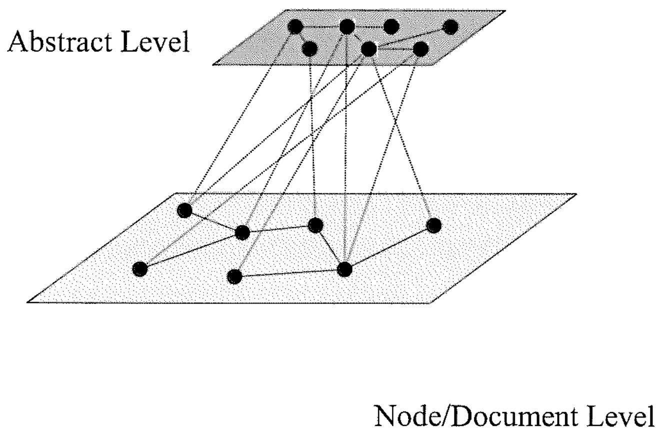


Figure 2-1- The two-layer model for navigation on Microcosm

In addition, the basic Microcosm model provides four different kinds of links at the document level that allow the author, or the user, to relate information. These are:

- *Specific links*, a tight connection between two fixed anchors. This link may be operated by a button highlighted by the system to indicate its presence. This type of link is important to create the back-bone link structure of the application at the document level that will allow the user to navigate between different items of information that may be obviously related or not
- *Local links*, a link with a fixed destination anchor, but with a source anchor, in a particular node/document, dependent on rules such as pattern matching, i.e. any occurrence of a particular object such as a text string. This kind of link is important for such things such as the definition of words or names, to give examples, or to show obviously related information, valid inside the node.

- *Generic links*, a link with a fixed destination anchor, but with the source anchor, in any document, dependent on rules such as pattern matching, i.e. any occurrence of a particular object such as a text string. This kind of link is important for the definition of words or names, to give examples, or to show obviously related information, valid for the complete application.
- *Dynamic computed links*, a link with a computed destination anchor, and a user typed or selected source. This last kind of link has been described in Li et al. (1992) and Li (1993) in order to integrate the Microcosm hypermedia system with some information-retrieval techniques and suggests starting points for navigation.

With the first three types of links, information is needed to define the link functionality type, the text link description showing the motivation for the link, the selection anchor of the source and destination of the link, etc. All this relevant information is used for the maintenance of the document links and to help the user in document level navigation.

Microcosm is an open hypermedia system since it does not use any mark-up link information in the document data files and because it can use data that comes from a variety of third-party applications, among other facilities. The lack of mark-up within the data is achieved by keeping all link information in separate link databases. This means it is possible to have different sets of links applied to the same information, allowing the co-existence of an author's link database (linkbase) and several user linkbases that each user can add to independently. The integration with third-party applications is achieved by using a model in which a number of autonomous processes communicate with each other by messages. The message passing system also allows the development of new tools that may easily be integrated into the system. Microcosm also separates the front end of the open hypermedia system from the back end. The front end consists of viewers and is responsible for user interaction. These can range from fully Microcosm aware viewers to unaware viewers (Davis et al 1994). The back end consists of a chain of filters (processes) and is responsible for many of the operations requested by the user. Connecting all this there is a "dipole" Document Control System - Filter Management System that is responsible for managing and integrating the whole system, as shown in Figure 2-2. Even with third-party applications without message passing capabilities it is possible to use the clipboard to integrate them into the hypermedia link service provided by Microcosm.

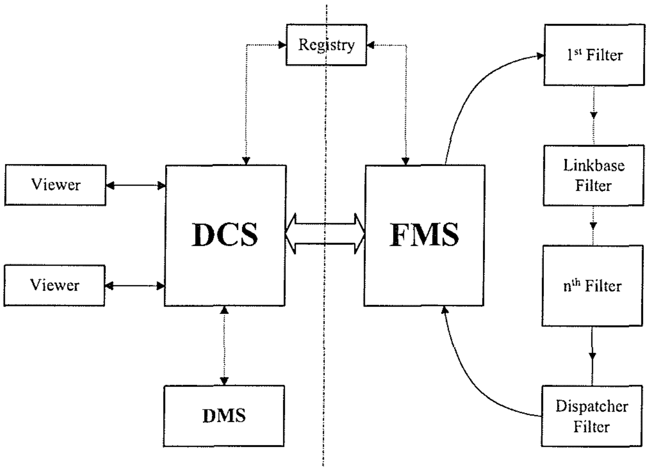


Figure 2-2 - The communicating processes that build up the Microcosm system

## 2.2 Accessing information in information retrieval systems

Information retrieval systems (Rijsbergen 1979, Salton 1989, Sparck Jones et al. 1997, Korfhage 1997, Baeza-Yates et al. 1999, Mitra et al. 2000) are mainly concerned with the representation, storage and retrieval of documents to support the user in finding relevant information. This definition is extremely wide, but the systems we are interested in are the ones where the systems will find documents stored in a computer, whose content is similar to a given query.

An indexing process is an important step in many information-retrieval systems, since it is impossible or impracticable to scan all documents using string matching and other direct techniques, in response to every query. The indexing process usually generates a set of index terms in such a way that its semantically combined meaning approximates to the content of the document or to the content of the user query. Then, by comparing the index terms of the queries and documents, a relation between them can be determined. According to the degree of that relation, relevant documents can then be selected and retrieved.

Different retrieval models can be used to represent information retrieval systems and procedures depending on the requirements to be imposed on the system:

- *Boolean retrieval model -*

In this model there is a comparison between a boolean query statement (set of terms connected by boolean operators *and*, *or* and *not*) with the term set used to identify the document content.

Although this model was widely accepted because of its low computing requirements, it is difficult to use. Boolean queries are difficult to formulate, and the number of retrieved documents is difficult to control. An imprecise or a broad request using the operator *or* can retrieve too many irrelevant documents, and a too precise query using the operator *and* may return no documents at all. Another disadvantage of this model is its inability to rank the results in a proper order in a way that could give a clue to the relevance of the retrieved document.

An example of such a retrieval model (Ortega et al., 1997) is present in the Yahoo Web search engine.

- *Vector space model -*

This model represents both queries and documents by vectors of weighted terms and computes global similarities between queries and documents. In this model there is the assumption that similar or related documents or similar documents and user queries are represented by a similar multidimensional term vector. The similarity is then just a function that determines a distance between the two representative term vectors.

This model is one of the most conceptually simple, but the justification for its parameters, such as the similarity function, are not derived from within the system but are instead chosen a priori by the system's author. Term relationships can be considered under this model but unfortunately that is not the standard, and the unrealistic assumption is often made that terms are independent of each other.

- *Clustered document environments -*

In this model (Crouch et al. 1989) there is a classification of all documents under clusters of related information structured in a hierarchical way that will be used to perform browsing or analytical retrieval.

This model provides easy access to all document information as it provides a browsing capability, although a search is much slower than in the vector space model, no matter how refined the available cluster classification is. However it is possible to access relevant documents, even if they are not related to queries, when the centroid (i.e., the representation of the cluster) that classifies that document is retrieved. With a browsing facility it is easy to access documents similar to specific items, since related documents are collected in a common group. Unfortunately, problems may arise in clustering very large files, or when dynamically maintaining the cluster classification.

- *Probabilistic retrieval model -*

This model (Rijsbergen 1979, Sparck Jones et al. 1997) is based on the computation of relevance probabilities for the documents of a collection. It can include term dependencies and relationships and important parameters such as the weighting of query terms, and the formulae for query-document similarities are determined by the model itself. In other models, such as the vectorial one, assumptions are often made for these parameters and formulae. In the probabilistic model the two main parameters are the probability of relevance and the probability of non-relevance of a document and upon these parameters much of the theoretical justification is made.

In spite of the theoretical advantages, this model did not lead to a significant improvement in retrieval effectiveness, mainly because of the difficulties in obtaining representative values for the required term-occurrence parameters. Since it is impossible to determine a priori all the relevant documents in a collection (otherwise there would be no reason for this model) a sample has to be frequently taken from user feedback. Using this sample the parameters can be calculated but do not have the required precision.

Among the facilities normally included in information retrieval systems are: vocabulary display, vocabulary expansion of query terms using a thesaurus, construction and storage of search protocols and operations with previously formulated queries, etc.

In the early days of information retrieval, there was greater concern with the retrieval of documents by means of their abstracts. However, when the full document content is taken into consideration, problems of heterogeneity are raised. It is then necessary to separate the content of the documents into different topics. Salton et al (1994) further developed the concept of a theme (semantically homogeneous text piece) and in Salton et al (1996) continued this work using segments. In a node, several information themes may be in focus. In a situation like this, if some information themes are not as developed as others, a bias on retrieval may result. So there is a need for analysis and retrieval of text passages, since full text statistical approaches may only be adequate enough with nodes that deal with single topic information where themes and segments are equally focused.

After splitting the text information into structural paragraphs (Salton et al 1994), statistical techniques were applied to find the similarity between all paragraphs in a certain node, making it possible to create links between related information. Because the information is split into a large number of small units, the number of links is very high. A study of the distribution pattern of the links inside a node could lead to the possibility of finding passages of themes and segments. A segment is defined as a continuous piece of text that is very well connected internally, but fairly disconnected from the adjacent text. Themes are generated by finding mutually similar text pieces that could be adjacent or not in the text. All sets of three similar paragraphs are joined in a centroid vector and all similar centroid vectors with three or more paragraphs merged to form a theme.

Having the nodes decomposed into themes and segments now makes it possible to implement passage retrieval. Based on this decomposition the authors implemented a different form for retrieving information by creating text transversals (retrieval of the best passages in a certain topic, which allows speed reading) and text summarisation (as the name implies, this summarises the information in a certain node by retrieving the most important paragraphs in a passage and connecting them with appropriate transition material). These kinds of tools seem to be of great use and interest to the user, and their implementation is easier after decomposition. However, authors faced a number of problems of coherency and understandability of the text transversal and text summarisation. However the use of a statistical approach is simpler and not so application dependent as linguistic analysis.

In this thesis we refer to the effectiveness of an information retrieval system. As in Rijsbergen (1979) or in Croft et al. (1989), retrieval effectiveness measures how successful a technique is in locating relevant information given a certain query, while retrieval efficiency is more concerned with space and time requirements.

The most widely accepted major measurements for partially determining the effectiveness of information-retrieval systems are "recall" and "precision". Recall indicates the performance of the system on retrieving all the relevant documents to a user query, and precision refers to the capability of the system in only retrieving the relevant documents. The combination of both measurements determines the effectiveness of the system that ideally should have high recall, by retrieving all the relevant documents, and high precision by retrieving only relevant documents. This combination can be calculated as a *Harmonic Mean* effectiveness measure (Baeza-Yates et al. 1999). Another possibility would be the E-measure (Rijsbergen 1979).

### 2.2.1 Searching in the Word Wide Web

Lately the majority of research in information retrieval has been developed for the World Wide Web. It is known that nowadays the large-scale size of the web, its dynamic change, resource format discrepancies and complexity makes it difficult to access relevant information. So there is a need for adequate indexing and retrieval of information, and here indexing means the extraction and representation of the semantic content and/or context of data. Different strategies have been developed and evolved for the Web (Poulter 1997, Schwartz 1998) to solve disorientation problems. There are basically two types: classified subject lists and keyword/query based search engines.

Many search engines like AltaVista (http://www.altavista.com) and Excite (http://www.excite.com) try to overcome this problem by automatically analysing and indexing the content of the entire web. The documents are retrieved by matching them with user-input text keywords. Unfortunately these search engines face many problems. The growing rate of the Web raises difficulties for the maintenance of indexes, keyword spamming or persuasion are used to bias retrieval results to undesired pages and the low precision on searches results in many unrelated pages being presented to users.

Another way to overcome this problem is the use of subject-based classified lists, which provide a useful browsable organization of information systematically arranged into categories, often using quite complex hierarchies (e.g. http://www.yahoo.com and http://www.sapo.pt). Systems that use this method try to describe the resources in the Web according to certain schema by providing information about information, i.e. metadata. This metadata can be for example, subject classification, author classification, web page ranking and web page interrelations. However, this method relies mainly on metadata inputted by humans to maintain the classifications.

This latter searching strategy using subject-based classified lists is, in a way, already an integration with hypermedia facilities. Some other examples will be given in section 2.4.2.

There are other specialized services aimed at helping users to select the right search tool from the large numbers of available options. These aids fall into two main types: directories, usually organizing the available search tools into categories, and meta-engines, usually sending user queries to multiple search engines and presenting the retrieved results in some way. An example of the first type is the All-in-One directory (http://www.allonesearch.com), and two examples of the later type are SavvySearch (http://www.savvysearch.com) and ONESEEK (http://www.oneseek.com). A model for describing the current query capabilities of the Web allowing their integration in a meta search engine has been suggested by Huang et al. (2000).

## 2.3 Reasons for the integration of information retrieval systems with hypermedia systems

### 2.3.1 Helping the user in navigation and retrieval

Hypermedia systems are different from information retrieval systems in the way that information is searched for. In hypermedia systems the search for information is achieved by navigation while in information retrieval a direct search is carried out. Each system carries advantages and disadvantages.

Hypermedia systems are usually user friendly, they can have useful interfaces and do not require any particular expertise of the user. This convenient facility for searching for information led to the success of several hypermedia systems, notably the World Wide Web. Hypermedia systems allow the user to choose their own path from many possible paths through the information. However, in large complex applications, this leads to a trial and error search strategy, which is random and time consuming, leading to disorientation problems. The user then feels "lost in hyperspace".

On the other hand, information-retrieval systems (mainly textual) have a more powerful content oriented way of searching for information, but unfortunately require more expertise from the user. The search for information is usually made by querying, and this leads to computer interpretation problems, requiring frequent reformulation of the query. The user needs a precise notion of the information he/she is seeking and unfortunately this is not always the case. Consequently, there is a high overhead to the user, since he/she has to analyse his/her information needs and organise the query search. Poor query formulation and inadequate user system interaction still occur even with skilled users.

One advantage of information-retrieval systems is that there are no disorientation problems, as the user always has to explicitly specify his/her information needs. Disorientation is however a problem that must be avoided in hypermedia systems. Conversely the effort required from the user in searching for information in information retrieval systems is a problem to be addressed but one that does not normally arise in hypermedia systems when browsing information. It makes sense then to integrate these two types of systems, in order to offset the disadvantages of one system with the advantages of the other.

A hypermedia system may allow the user an independent non-sequential reading of the information, that is, the freedom to browse through the information. But this freedom may not be beneficial if the user is not able to find the relevant information that he/she seeks. So, if a usable hypermedia system is required, it is necessary, in addition, to have a good way of providing clues to relevant information, and this usually leads us into the area of information retrieval. Traditionally, in these systems, a way of "linking" to the information is offered after the user makes an analytical query that he/she specifies in accordance with his/her search strategy. In this way, it is possible to fulfil the user's information needs by giving him/her more goal-oriented tools. The most interesting aspect is that this integration not only diminishes user disorientation problems, but provides new ways of viewing and changing applications and improving retrieval recall by considering all the information associated in the hypermedia network of the application.

Equally, an information-retrieval system may allow the user a comparably more precise and objective retrieval of information, but this facility may not be appreciated if in such a system the user is not able to specify a clear, meaningful query. The user may reach a point where it is impossible to retrieve any information at all or, conversely, retrieves information that is totally irrelevant. The user must then find a way to specify a query in a simpler manner. For this purpose, it might be important to know what kind or type of information is available in an application, and what its classification or structure is, before the query is formulated. With the integration of hypermedia facilities at the classification level it is possible to search for this relevant information in an interactive friendly way. Much of the overhead associated with information retrieval is then eliminated.

In this way it might also be feasible to incorporate feedback into the system whereby a query can be refined or reformulated to improve the recall and precision of the returned media.

### 2.3.2 Improving the effectiveness and the efficiency of the two systems

Apart from these major advantages of interaction with the user, arising from the integration of these two systems, there are other good points to take into account, notably the improvement of the effectiveness of both the information retrieval and hypermedia component.

Usually hypermedia systems deal with highly structured information, while classical information retrieval systems deal with fragmented units of information. Hypermedia systems are more concerned with the structure of information and not so much with the representation of their content, while in information retrieval systems the exploration of the content is usually more relevant than considerations of structure, as pointed out by Chiaramella et al. (1996). Thus, traditionally, content is for information retrieval as structure is for hypermedia. Therefore, with integration between the two systems, more use of the structure should be considered on the information retrieval side, and more use of the content on the hypermedia side.

The information retrieval side can be improved by considering information from the hypermedia network in the construction of the classification index, frequently used in these systems, or in the elaboration of better information retrieval algorithms. On the other hand the hypermedia side can similarly be improved by using content information in the construction of links.

Another advantage of the integration of these systems is the multimedia facility of hypermedia. In many hypermedia systems, different media are handled in a similar way, allowing easy access to both text and non-text documents. However in information retrieval systems that is more difficult. There are very well known effective and efficient algorithms for information retrieval where text is concerned (Rijsbergen 1979, Salton 89, Frakes et al. 1992) but for non-text files there is still a long way to go to attain the same levels of performance. This is because text information is already "encoded" (Narasimhalu et al., 1995), i.e., the information is already segmented into a discrete set of symbols such as words. This characteristic facilitates the indexing and classification of information and therefore facilitates efficiency of retrieval. Since with non-text files, such as pictures, the information is "unencoded", the apparently simple techniques of retrieval from text information cannot be applied. But in hypermedia systems the non-text information is in a way "encoded", since with the network of links in hypermedia it is possible to relate information from different media in a direct way. The multimedia non-text files are then to a certain extent "encoded" by their relation with other media. It may be possible to understand the multimedia content of non-text files by analysing their relation with the text files that are linked to them. Basically, there is then an indirect way of retrieving multimedia files using the hypermedia network information to "encode" in text the content of their information.

## 2.4 Previous integrations between information retrieval and hypermedia systems

Many previous systems have attempted to implement ideas on the integration of hypermedia systems with information retrieval systems as described in the previous section. However, the integration between the two systems usually depends on how the researcher or author viewed the retrieval of information and various aspects of hypermedia navigation, and at which level he/she believed that integration should be implemented.

In some information retrieval systems a two-layer model is adopted, separating the document level from the index/abstract level. However, for integration, there are usually different ways of viewing the hypermedia network in relation to the information retrieval tool: at the document level, or at the index abstract level or at both. If the hypermedia links were seen by the author as only explicit relations between nodes at the document level, then hypermedia information can be used to improve the definition of the index level and the search quality of the retrieval mechanism. If hypermedia was seen as a means of interconnecting indexes, that is, as a way of navigation at the index level to reach documents, and not exactly as a way of navigating direct connections between documents, then hypermedia was considered more as an interface tool by the author and as a way of semantically relating indexes for the information-retrieval tool.

Until now, most of the research on the integration of hypermedia systems with information retrieval systems has been incomplete. Usually the integration is about using hypermedia tools to interface with an information retrieval system or, alternatively, about the use of information retrieval tools to help the user in navigation.

### 2.4.1 Information retrieval tools to help the user in searching for information in hypermedia systems, by considering the hypermedia network at the document level.

As we have seen, a way of helping the user to find relevant information by browsing when he/she is lost is through the use of information retrieval techniques. Previously, many different approaches have been proposed and implemented to solve the problem of integrating information retrieval systems with hypermedia systems. The hypermedia structure was seen as a network of explicit links at the document level and the information-retrieval tool would work over the top of it at the classification level helping the user in navigation throughout the hypermedia network.

Generally, an attempt was made to use information in nodes or in typed or semantic links in an improved information retrieval tool to help the user in navigation through the hypermedia/text network by providing a means of orientation. Starting nodes (Frisse 1988, Frei et al. 1995, Marchiori et al. 1997, Brin et al. 1998, Mizuuchi et al. 1999), guided tours (Bernstein 1990, Guinan et al. 1992), best next match (Li 1993, Dean et al. 1999), and other ideas were tried. Hypermedia network information was also used to classify multimedia documents under a text information retrieval indexing space in order to retrieve multimedia information with text techniques.

One of the earliest to apply some of these concepts was Frisse (1988) with his hypertext medical handbook. In his work, the information contained in the handbook was divided into individual, small, fixed size cards using a hierarchical structure. In order to preserve the top down hierarchy structure of the handbook, links were made between each card and its parent. Labels were also associated with every card to make navigation through the hypertext easier. With a hierarchical hypertext structure, it can be difficult to move across the hierarchy and to find relevant information, because at times the structure is not apparent to the user. Therefore an information-retrieval process was designed in which a starting card, the starting point for browsing, was created following a user query. This was called the small-document approach for global navigation. After the creation of this starting card, the user was free to browse through the hypertext. This was called the graph-traversal approach for local navigation.

In Frisse (1988), the information retrieval was based on a statistical approach, where the results were dependent on the hypertext structure. This dependency was achieved by basing the score (importance of relevance) of each card on a statistical comparison between the query and the card itself and on the scoring of the immediately following cards (i.e., the child cards located on the hierarchy under the considered card). Further developments were achieved in Frisse et al (1989) by considering belief network algorithms with user feedback.

Unfortunately this card structure approach cannot be applied to large hypertext systems because the management would become impossible, as the system may return a large number of uninteresting cards (information), since in this approach recall can reach high values. In large applications the algorithm used to retrieve information also requires a reasonable amount of processing power from the computer. The hierarchical inter-dependence of this system is also a facet to be considered because it connects the information in a tight way. The retrieval of interdependent nodes was pioneering work, but as the author states, more work needs to be done on a network structure.

Even with a good starting point, navigation through a hypertext system can be a problem for a less informed user. Bernstein (1990) proposed a shallow system capable of automatic construction of a guided tour over a network of hypertext nodes, starting from a given node. However the type of algorithm used (best match first) for the construction of the tour did not account for the possible correct navigation order of the nodes and this cause the reader understandability problems.

Guinan et al. (1992) expanded the information-retrieval approach proposed by Frisse (1988) by adding the facility to create a planned tour through the hypertext system but with logical sequencing of nodes, in order to cover all the relevant information. To do this they used the concept of typed links, i.e., the classification of each link into one of a predefined set of categories. With this extra information added to the links they proposed that information retrieval could be improved. In this technique, the node content and its neighbours' content are used to retrieve the best nodes, and the typed links are used to structure the nodes in a logical order, which is called a planned guided tour.

Two points must be considered when typed links are used: the dependence of the set of types on the type of hypertext application used and authoring effort. Although in this approach the information retrieval is more fully integrated in the hypertext system, difficulties may arise in large hypertext systems. Also, in order to have a well-connected tour the hypertext application must have a relatively large number of links, and this may be a disadvantage for poorly connected hypertext systems.

Li et al. (1992) and Li (1993) also developed a method for integrating information retrieval and open hypertext systems.

Li used classical methods of statistical retrieval (Rijsbergen, 1979) by pre-indexing the content of all text nodes in order to achieve a quicker response time for retrieval using well-known algorithms. However, it must be pointed out that he improved retrieval by introducing concepts such as break words and developing further the use of phrase weighting. Unfortunately his approach has the major disadvantage of being static, i.e. it is only efficient in hypertext systems if there are no changes in the nodes, as often occurs in other information-retrieval systems. If a new file is inserted, removed or changed, into the hypermedia collection of documents, all the pre-indexed information must be recalculated for all nodes and, of course, this is a time consuming task.

One good advantage of Li's approach however is that the information-retrieval is considered as a different kind of link - a dynamic computed link - making the retrieval more transparent to the user. In this way the retrieval is no longer considered an independent tool. On the other hand, the information retrieval does not consider the structure of the hypermedia for retrieval improvement. The hypermedia link network is completely ignored.

Previous approaches only considered text information retrieval in hypertext/hypermedia systems. Although the work carried out in Frei et al. (1992) and Frei et al. (1995) was also based on a hypertext system, they used the semantic content of hypertext links to retrieve information, in the hope that this philosophy could be applied in future hypermedia systems.

They proposed a system, which improves information retrieval by considering for retrieval not only the content of the nodes but also the information stored in the semantics of links. The content of each semantic link was associated with a link description as well as a source and destination anchor and a set of structured link attributes such as creation time and author name. The description takes into account the neighbouring nodes (the destination and source node) of the link, but if the author (or the user during browsing) thinks that it is important, the description may be extended or completely changed.

To retrieve information, two different methods of searching are considered: the exhaustive search, which helps a user who does not know much about the hypertext collection by giving him/her a means of indicating some interesting starting points, and secondly the navigational search, which helps the user navigate through the hypertext collection by suggesting next best matches.

The suggested system incorporates good techniques such as the automatic construction of descriptors that can improve the quality of retrieval, helps the user with navigation, and, very importantly, allows the retrieval of non-text media. Because of the physical amount of information kept in the semantic links, problems in storage may arise in a large hypertext system, and these problems will increase if the system has a considerable number of links. On the other hand, if the system is sparsely connected, it is necessary to use greater maximum analysis distances, increasing the time for retrieval to an unacceptable level. However, this is one of the systems that takes the hypertext network more into account without too much extra effort on the part of the author.

A great problem with web search engines is that they rank text and not hypertext. The relationships of information obtained with links are not usually taken into consideration. Thus considering this hyper-information is fundamental

One of the best integrations achieved by using hypermedia to help information retrieval has been implemented by Brin et al. (1998). Their system Google (http://www.google.com) is a search engine based on the idea that popularity or visibility is a good criterion to rank web pages. In the web, links take us to where authors want to point us and those links should, but don't, work the other way around. However, the number of links pointing to a site is a good quality assessment of its importance. Google searches the web looking for these connections, building a database of link connections and anchor text. The system indexes documents considering not only their content, but also the content of the anchors of links pointing to them. The document relevancy ranking is based, among other factors, on the number of incoming links, i.e., it will give more importance to the most popular or visible pages. This kind of ranking, and also indexing, can give much better quality results than traditional search engines, but nevertheless some embarrassing results may arise. Popularity is not always a good criterion and for instance, if we ask to Google "What is more evil than Satan" the best ranking will be Microsoft's home page (O'Brien 2000)! In addition, if a web resource is not well known enough, due to its location, then it is going to have a low visibility even if it is of very good quality. Authors don't know the entire web and this system has limitations because of this assumption. In a nutshell, popularity is something completely different to quality and thus using it for higher scores can be a poor choice. According to Chakrabarti et al. (1999), this type of ranking based on popularity is called authority ranking. They suggest another method of improvement, namely hub ranking. Hubs are pages that convey authority, i.e., they point to many authority pages and links from them are therefore expected to point to relevant documents. An example of a hub is a directory list of search engines, or a web page with links to staff home pages in a research group.

Kaindl et al (1998) introduced a form of structure search based on content search. Its functionality was tested as a meta-search working on top of commercial search engines. The algorithm considers that neighbour documents, predecessors and successors, of all the retrieved documents, for a central query, set a constraint on the final set of valid documents to present to the user. They give examples of possible applications for this algorithm such as finding authority and hub documents, and constraining searches to a specific context. They lack a proper test of their algorithm, however their approach works well on the present architecture of the Web.

Marchiori (1997) proposed a system that could work on top of current search engines. This submits queries to commercial search engines and uses link information to improve precision results. He used a technique previously implemented in hypertext systems (Frisse 1988, Frei et al. 1995) that consisted of indexing documents not only using their content but also using portions of neighbour documents. They solved some problems expected on the web like backward linking, duplication of links, local links, frames, search engine persuasion, etc. This meta-search technique has proved to give better user satisfaction, however they didn't used standard precision/recall testing.

Mizuuchi et al. (1999) also consider links between documents, aiming to expand indexing to reflect a more complete description of each page. They based their work on the assumption that a page is inserted on a path starting from an entry point, and for that reason some of the description context of the page might have been left out. They proposed some methods to find those entrance paths and extract keywords from anchor selections, titles and heading information from documents on the path to increase the number of keywords for the target. This method can contribute to the increase of recall, but the problem on the web is often precision and they lacked a proper evaluation of the precision of this system.

### 2.4.2 Using hypermedia browsing functionality to aid searching in information retrieval systems, considering the hypermedia at the classification level

In the kinds of integration described here, the hypermedia is usually seen at the classification level. The hypermedia is a way of linking concepts to help the user in the retrieval of information. Navigation is achieved in several ways, such as between similar, broader or narrower concepts, but usually not through explicit links between the documents themselves. The background functionality of the system is supported under the information retrieval paradigm, and over the top of it there is the hypermedia interface helping the user to search for information.

Network structures have been used for some time in information retrieval research and there are retrieval methods that use automatic or manual links between documents and classification concepts and/or links between classification concepts. Cluster construction and retrieval of information is one of the research fields concerned with this (Rijsbergen, 1979; Salton, 1989). Clustering retrieval systems offer a way of grouping related documents under a classification tree, making them suitable for hypermedia integration. A browsing facility associated with a clustering retrieval tree will allow the easy location of information, by focusing the search on groups/clusters where documents are more closely related with the query, and by easily locating other similar groups/clusters of related documents.

Crouch et al (1989) developed the idea of cluster hierarchies in hypermedia information retrieval. They describe a graphical-traversal browsing interface that supports the analysis of the hierarchical clustering structure of a document collection. The cluster links between documents can be seen in two different viewers. A local viewer contains a fraction of the cluster tree where the user is actually navigating and where there is more detailed information about a specific subtree, and a global viewer that allows the user to have a more comprehensive way of viewing the search in relation to the full tree with a significantly large number of nodes. The user will conduct a search by specifying a query that can be refined during the search by removing or inserting concepts or by indicating which documents have been found to be relevant. To help the user in cluster navigation, besides visual orientation tools, information is provided for all nodes about the number of its relevant concepts in common with the query, and their associated correlational weight.

They tested this interface and concluded that its performance was a significant improvement compared with the automatic cluster retrieval systems. The authors were able to achieve this without using sophisticated user interfaces or expert systems to solve problems associated with the handling of information-retrieval searches.

Another model to overcome the limitations of hypermedia systems in relation to information retrieval operations was introduced by Agosti et al (1992). Their major concern was the explicit presentation to the user of the network of index terms and concepts that were used to represent the document collection. In their hypertext information-retrieval model, they implemented two associative information-retrieval tools to help the user in navigation. The associative reading tool was used to guide the user on browsing both the network of concepts and the hypertext at the document level. The semantic association tool was used to reveal the interpretation that the systems hold about a concept that the user uses to express his/her information needs.

Associative information retrieval tries to overcome the problems associated with exact-match retrieval. Many previous researcher have tried to use these techniques, and some mathematical background was given as early as Edmundson et al (1961) where word frequency was related with word significance. Later a probabilistic model (Rijsbergen 1977) was presented where a previous classification of the documents to be indexed was not required. The Agosti et al (1992) approach required a preliminary identification of the concepts in the documents that could be either created automatically or manually by experts. However they only tested the system using a manually created semantically connected thesaurus at the concept level. In the way they implement the system, the semantic association tool was able to return terms conceptually related to the terms specified in the user search, by analysing the relations between the terms that the thesaurus holds or by analysing the relation between the thesaurus terms and the document terms.

More clear tests should be performed to check its validity under an automatically created index space, where often the statistical relations between some connected terms do not make much sense for the user.

A similar type of integration in Taylor et al (1995) later developed in Cunliffe et al (1997) presented a notion of hypermedia navigation on a manually created index space, where semantic relationships existed between index terms. They decided on an index space with three ‘dimensions’, adequate to manually classify their collection of cultural heritage documents: time, space and subject. For the subject classification they used an accepted classification standard in social history. Such subject classification consists of controlled vocabulary, organised in a semantic hierarchy of broader and narrower terms. Under the space and time classifications, a hierarchical classification of broader/narrower space area divisions or time intervals was again adopted.

They defined semantic closeness as a distance between terms in the conceptual schema measured by the minimum number of semantic relationships that must be traversed in order to connect the terms. With this definition they were then able to build two hybrid navigation/query tools for the automatic traversal of relationships of the index space, with the same functional goal as the semantic association and associative reading tool proposed in Agosti et al (1992) but with a different implementation. In Cunliffe et al (1997), with the query generalisation tool, the system is able to retrieve a rank of terms semantically close to the set of terms in the current position, calculated using different algorithms in each index dimension. This allows the user to select alternatives to generalise the search until he/she finds the relevant information. What the authors called navigation via similarity tool was able to retrieve items located 'nearby' in the index space, in relation to a set of ten previous selected nodes/documents of interest.

The query by generalisation tool previously described helps the user in navigation of the index space, and the navigation via similarity helps the user in navigation at the document level. However, both tools have their implementation algorithms only at the index level, in spite of also using links at the document level as they said might be possible. Nevertheless their approach contains an important integration of information retrieval and hypermedia techniques, by providing the user with a smooth transition between browsing and querying. Finally, in the social history domain, as in some others domains, it is possible to find a sufficiently general semantically connected network of terms, but for the majority of domains that will not happen. The index space will therefore have to be created manually, by skilled indexers, and that will lead to time intensive developments of specific application dependent solutions.

Another specific solution developed by Arentes et al (1993) adopted a classification space consisting of three thesauri of index terms that were derived from standard accepted taxonomies on corrosion, material and environment concepts. Their two main concerns were the interface presentation of the index structure using a three-dimensional cube navigation tool, and the refinement of the hyperindices, i.e., index information organised in the form of a hypertext as defined in Bruza (1990). Arentes et al (1993) developed a semantic-aware version of Bruza's hyperindices by accounting for the meaning of the connectors between index terms and by identifying which node types were appropriate for a given combination of index terms. This allowed then to retrieve only the set of all index classification expressions that make semantic sense by considering semantically coupled thesauri for the domain of concepts when constructing the index expressions for each unit of information. The semantically valid index expressions can then be given to the user for navigation.

Their approach, as in previous solutions proposed by other researchers, relied heavily on the use of thesauri to support browsing searchs, but the restriction imposed by them on the connectivity of the hypermedia index structure helps user navigation. The good visual feedback given to the user by using a cube-navigation tool seems useful but is unfortunately restricted to the use of only three semantically coupled thesauri of index terms. Their menu solution to this problem suffers, on the other hand, from cognitive overhead and time search problems in using a complex thesaurus, and from the complexity of navigation to broader or to related terms.

Another approach was implemented by Boy (1991) where a knowledge-base component helps the user in the retrieval of contextual information by "experimental browsing" or "intentional search" of a hierarchy of multimedia descriptors classifying the documentation. The knowledge was represented as a set of ranked links to documents that are shown under a trigger condition (the selection of a descriptor) and a contextual condition (description of the type of information need, e.g., expert, novice, technical, etc). When the user selects a descriptor of interest the knowledge-base component is responsible for the ranking of the document references, in accordance with the context given by the user. The system is then tuned under that context by the user providing feedback on the most relevant documents.

Their system can then acquire the context under which certain documents were appropriate. With this component it was possible to adapt the system to the type of user and to his/her type of information needs, using his/her feedback. However they did not show how to formalise the contextual conditions. Another limitation of the system was the size of the contextual conditions, in order to avoid excessive calculation. This can be problematic in multi-user systems.

The main advances made in this area have been done by considering navigational facilities on top of information retrieval. In resume this advances have been:

- Semantic association of documents by the use of links;
- Visual interfaces for index navigation;
- Less overhead on query formulation;
- Querying refinement as browsing;
- Smother transition between navigation and retrieval;
- User adaptation by the use o navigation patterns.

However the majority of the solutions available in the are have some disadvantages. For instance the domain specific solution adopted in some systems restrict their use on a broader context. Also the use o manually created thesaurus and classification can have authoring problems.

### 2.4.3 Automatic construction of hypertext

To overcome the difficulty of creating a large number of manual links many researchers have tried to implement tools for the automatic construction of the hypermedia structures. Document structure has been used for the segmentation of text into hypertext nodes and to create a hierarchical skeleton that linked the nodes (e.g., Frisse, 1988) but their construction process was a weak manual one.

As mentioned before, Bernstein (1990) suggested a shallow approach capable of automatic identification of similar documents. Proper tools would then be available for the author to use for rapid linking. His approach was very light and, for the complexity of the algorithms used, the results were reasonable, but the reliance of the system on the human author to supervise it limited the amount of information that could be processed.

Golovchinsky (1997) presented a way to get over these problems of the automatic creation of links by using user feedback during browsing of the documentation. To select anchors, the system uses a heuristic approach by choosing capitalised words as candidates together with a statistical full text search engine to pick up query terms that distinguish well between documents. When the user selects an anchor to follow, the system will extract its context, consisting of the terms in the sentence that contain the anchor, the terms in previous selection "queries", user feedback, etc. This context is then used by the full text search engine to retrieve a ranked set of similar documents that will be organised in a newspaper style screen, i.e. the system shows the retrieved files in distributed order and taking up a screen space proportional to the ranking of the document.

This kind of approach to building a hypertext interface onto an information retrieval system seems quite user intuitive, but the creation of links is ephemeral. The authors did not show a way of creating permanent links, which are important for structuring a hypermedia application.

Allan (1996) suggested a much-improved method of hypertext construction when a statistical free text information retrieval tool is used. He was inspired by the theme generation addressed in Salton et al (1994) in which links were created between text segments. He started by constructing links between all the most similar text segments in the whole text collection. He was then able to give answers to problems such as link typing and number of links when using classical information-retrieval tools, by using merging techniques between the most similar of all text segments. The link types were identified when merging links by analysing their relative position in the documents.

As the author said, there is still some more work to do on heterogeneous documents or documents that are written in a non-regular style. However he has set out an important approach to solve problems that did not have a satisfactory answer previously, even if a better evaluation is required. Still, the computational requirements of his solution will not allow its application on interactive systems, and it is not feasible for most personal computers.

Agosti et al (1996a) implemented a tool that could construct a more structured hypertext system for information retrieval. They considered a hypertext structure with three different levels: document level, index term level and conceptual level. The links between documents were found using document-document similarity and links between terms and documents were inherent to the use of a statistical indexing procedure. Links between terms were found using term-term similarity and ranked by intra-document weight for finding the source and the destination of a link. Index terms were linked to concepts according to the association algorithm proposed in Agosti et al (1992) and their intra-document weight determined their ranking. It was possible to limit the number of links seen by the user by selecting a proper threshold level for document-document and term-term similarity. For the concept level they used a manually created semantically connected thesaurus to maintain the link structure at that level.

In an application domain where a manually created thesaurus is available, this approach seems to be of good value for the author, but when a thesaurus is not available then a manually created thesaurus must be implemented and therefore the process of automatic hypertext construction will not be fully automatic. They suggested that an automatic tool for the construction of a thesaurus could be used but they did not demonstrate whether such a thesaurus could be effective for automatic hypertext construction

### 2.4.4 Multimedia access of information

Currently there is no efficient system that can retrieve non-text information as well as text information. Several efficient (and some of them effective) retrieval techniques have already been developed for text information that uses natural language sentences as a query to retrieve the relevant text documents. However, as far as other media are concerned, it is easy to conclude that there is yet a long way to go to find a feasible method. Having said that, some work has already been developed in this field, and here we give examples such as Lewis et al. (1996a) and O'Docherty (1990).

In the multimedia retrieval approach taken by O'Docherty et al. (1990), they proposed the application of the object-oriented paradigm, suggesting that all these systems should use it as a way of dealing with complex heterogeneous data types and to make the design more efficient and less complex. To help the user with retrieval, they attempted to store semantic data in the system by interpreting the raw data automatically using knowledge bases.

Their technique of representing and dealing with information (using the artificial intelligence and object oriented paradigm) seems to be the closest to human interpretation. However, even today, systems using AI paradigms have not proved to be efficient, i.e., producing results with non-text information in a reasonable time and with low processing capabilities requirements. Even if the queries are pre-processed the quality of the results may not satisfy user requirements.

Lewis et al. (1996a) proposed an approach - MAVIS - that allows content-based navigation and content-based retrieval using a hypermedia link service for documents of any media type. With the introduction of the idea of signatures, rather than anchors as in hypermedia systems, it is possible to expand the model of navigation and retrieval. These signatures are features extracted from different media, and can represent for instance: a certain kind of shape or colour for images, a certain pitch in a sound file, etc. To navigate and retrieve information, signatures were compared, for example text, colour, shape, etc, rather than the content directly. This pre-processed information improves the efficiency of the system. The structure of the system is generic enough to allow the expansion of models and permit improvements, although direct analysis of non-text media still has a long way to go to attain the efficiency of the text medium. The number of models is still limited and it is difficult to represent non-text media properly. A similar early work on *media-base navigation* has also been developed by Hirata et al. (1993) at NEC.

Lewis et al. (1996a) pointed out that prior knowledge of the media may be important to improve the efficiency of navigation and retrieval, but such an implementation will be application dependent (e.g. Taylor et al., 1995). MAVIS capabilities were then expanded to form MAVIS 2 (Dobie et al 1999, Tansley 2000). MAVIS 2 used a multimedia thesaurus, which allowed associations of the different representations of the same object in different media, and allowed conceptual relations between different objects. To achieve this goal the system adopted a four-layer data model: the *raw* layer, the *selection* layer, the *selection expression* layer and the *conceptual* layer. This last layer, implemented with a multimedia thesaurus, allowed for different user search possibilities like concept navigation. Similar work has also been done by Hirata et al. (1996) and Hirata et al. (1997), where the concept of *object-based navigation* behaves like a multimedia thesaurus by associating the different media representation of the data to one concept, and by also allowing navigation through concepts. All these systems have however a reduced text capability, and no relations apart from the classification of the media in the thesaurus, can be used. Any links not at conceptual/classification level are ignored.

Nevertheless, an expansion of the idea of generic links, already implemented in Microcosm (Davis et al., 1992) for text, allows new capabilities in non-text media on hypermedia systems. It is possible to have links based on the content and not on the structure of the hypermedia. In a way, this is an integration of information retrieval within the linking functionality of the hypermedia system.

Dunlop (1991), Dunlop et al. (1991a), and Dunlop et al (1993) made important advances using probabilistic text retrieval techniques (Rijsbergen, 1979, Croft et al, 1979) to retrieve multimedia information in a hypermedia system. The early work of Frisse (1988), in which retrieval was dependent on structure, and previous work that used text descriptors for multimedia retrieval, presented the case that multimedia retrieval could be possible by analysing the linking network. Since a non-text node can have several links to and from it, starting at text nodes and finishing at text nodes, then it is possible to build a text descriptor node, based on the content of the neighbour nodes. This allows indexing of the text information in the descriptor node, and therefore the implementation of relatively non-complex computational retrieval mechanisms. This method is based on the assumption that the neighbour nodes would be related to the described node with common information topics, since links usually connect related information.

To test this theory, the retrieval efficiency of a hypertext application was compared using normal indexing techniques based on the content of all nodes with the same hypertext application where a few text nodes were replaced by descriptor text files. In this way they had a means for comparison of retrieval using text descriptors with retrieval using the content of the nodes. The results show that the retrieval efficiency using text descriptor files is not significantly less than the efficiency with the original nodes. However that could not happen if the system were not rich in links. The quality would degrade when only one or two links existed from, or to, the described node.

The Web search engines are useful for finding relevant information, but the most popular ones are essentially textual. Since most Web pages possess multimedia content there is a need for effective tools to find multimedia from the web. However, images do not appear in isolation, and they are usually accompanied with collateral text, and some work has been done on the exploration of the interaction of textual and non-textual information in the Web (Frankel et al. 1996, Harmandas et al. 1997, Mukherjea et al. 1999, Srihari et al. 1999). For all those integrations we know that text methods are powerful in matching context (Salton 1989), but don't give access to image content. Inversely, image methods determine similarity between images but use little semantics.

There has been an interest in using textual regions, labels or captions to index and retrieve images. All these methods use different metadata retrieved from web pages, and in one way or another assign a weight to different elements. However the weight assignment is unclear because they lack a proper evaluation of the importance of the different metadata elements according to different developing styles.

Webseer (Frankel et al. 1996, http://webseer.cs.uchicago.edu/) also utilizes both image and text content for picture searching. However it does not have text understanding apart from processing HTML tags, no attempt is made to extract descriptions of pictures, and searches are only implemented using text. All queries are in text and there is no similarity computation of image content. Image characteristics are inferred and these are used along with textual context information.

Harmandas et al. (1997) have extended some of Dunlop et al.'s (1993) work to allow multimedia retrieval in the Web. They exploit the linked nature of the Web to build the text description. This is achieved by using the information from the image caption around each image, the image caption of neighbouring images, the full text where the image is inserted and the full text of other links to that page. They show that considering the full text where the image is inserted is the option that gives the worst recall/precision on the retrieval, and that the information that provides the best recall/precision is the image text caption. However they restricted their tests to collections of museum paintings where many different text descriptions are usually available.

WebSeek (Smith et al 1997) uses text and visual information to store information, and search results are returned as thumbnail images with filenames and a link to the Web site where the image was found. Here text is used for the multimedia subject taxonomy tree classification search. Image algorithms are used for similarity matching.

A model of multimedia representation proposed by (Amato et al. 1998) shows a way forward in considering different features for image indexing, but unfortunately does not include text. Text context was only used to aggregate documents into interrelated concepts for navigation and retrieval and no support is given for ranking documents according to a possible text feature. They also lack a complete implementation of the system and an evaluation.

Mukherjea et al. 1997 and Mukherjea et al. 1999 have been developing AMORE, a system that allows searches based on keywords, images, or both and that presents results in many different visual ways. A possible matching technique would be to use text processing first to satisfy the general context of user interest, and then image processing to refine the results by similarity matching. The goal of the system was initially the retrieval of all type of documents but then became oriented to image retrieval. They extensively studied heuristics to adequately associate text captions and regions on web pages to images. They considered image URL, image name, title, alt text, anchor text, heading, surrounding text and short text, and created two evaluation criteria, usefulness and accuracy, to evaluate information to find the sources for high relevant keywords. However, they haven't made an overall evaluation of the system using the standard information retrieval metrics of precision and recall.

Webpic (Srihari et al. 1999) made possible the combination of text processing and image processing in both indexing and retrieval phases. Webpic presents possible methods for different similarity matching techniques, which are used according to user input, and the results sorted with the accorded criteria. Different metadata is extracted from the accompanying text using different techniques, like statistical text indexing, light parsing, etc, for the construction of a picture description template (PDT) which represents images characteristics. Image processing is done using various similarity matching techniques like colour histograms and face recognition. They then use different methods of combining content image and context text search giving different results, but they lack a proper evaluation of the data used.

Other examples of multimedia access in the web are: AltaVista's Multimedia Search (http://www.altavista.com), Yahoo's Image Surfer (http://isurf.yahoo.com) and Corbis (http://www.corbis.com/). AltaVista's Multimedia Search (http://www.altavista.com) allows keyword and visually similar search, but semantically similar search is not possible. Yahoo's Image Surfer (http://isurf.yahoo.com) stores a collection of images, cataloguing then into categories, allowing users to find images that match the category keywords. But HTML document search is not integrated, and image similarity is simplistic.

The main advances made in this area have been done mainly by considering the context of multimedia in relation to text media. In resume these advances have been done on:

- Good techniques to match the context of images;
- The use of thesaurus to associate, classify and retrieve multimedia along with text;
- The use of links to construct text descriptions for multimedia, by considering the all content of neighbour text documents;
- The use of textual regions, labels and caption to index and retrieve images;
- Some sort of integration of content and context searching of images.

However, there is still a long way to go when content information retrieval is considered. Context information retrieval (information retrieval that consider multimedia described using text extracted from the hypermedia) solves some problems for now, but there are still many aspects requiring improvements:

- Better consideration of abstract/classification information on context indexing;
- Better consideration of link information on context indexing;
- Distinction between different kind of authors and applications on context indexing;
- Better evaluation on the available context information;
- Better techniques to match content and not only context;
- The use of open retrieval models.

## 2.5 New World Wide Web standards

New standards are available or are being developed for the Word Wide Web. These standards may provide some ways forward for new methods of information retrieval and/or hypermedia navigation. A brief description of some of these standards follows along with some description of a few searching or navigation facilities already available.

### XML and querying languages

Some query languages have already tried to make structured queries on the traditional HTML web. The Squeal programming language (Spertus et al. 2000), built on top of SQL, used one schema to describe web pages and an implementation that answers queries about the schema. Giving the illusion that the web is entirely in a relational database, the information is fetched on demand. They make a rather poor assumption by selecting only one schema for the whole Web, when it is known how heterogeneous the Web can be.

But a new standard has been proposed by the W3C, World Wide Web Contortium, to replace the present HTML document markup language. XML, eXtensible Markup Language, (Bray et al. 1998) tries to provide many of the benefits of SGML not possible with HTML, but in a language that is easier to learn and use than the complete SGML (Lie et al 1999). Some of it advantages are user-defined tags, nested elements, separation between the structure of documents and their visualization, and a non-compulsory validation of document structure with respect to a Document Type Descriptor (DTD). XML is designed to provide easier information interchange between different data sources on the web.

Because of the popularity of XML, many query languages have been developed for it, but usually they are only concerned with querying content and structure of a certain type. At the time of writing the W3C was working on a standard (Chamberlin et al. 2000) for a query data model, algebra and query language for processing XML (see examples of XML Queries in Table 2-1), but meanwhile there are many proposed languages such as XML-GL (Ceri et al. 1999) and XML-QL

(Deutsch et al. 1999). The former is basically a logical-based graphical query language for XML. They use visual formalisms to represent both the content of XML documents and the syntax and semantics of queries. The later is a simple declarative relational query language that allows searching of semi-structured data, giving more support to semantics than to syntax. Pattern matching, among other techniques, allows for this functionality.

Example of a free text query to be expressed in XML:

- List books published by Addison-Wesley after 1991, including their year and title.

Solution in XQuery:

```
<bib>
   FOR $b IN document("http://www.bn.com")/bib/book
   WHERE $b/publisher = "Addison-Wesley" AND $b/@year > 1991
   RETURN
      <book year = $b/@year>
         $b/title
      </book>
</bib>
```

Example of a free text query to be expressed in XML:

- Prepare a (flat) figure list for Book1, listing all the figures and their titles. Preserve the original attributes of each <figure> element, if any.

Solution in XQuery:

```
<figlist>
   FOR $f IN document("book1.xml")//figure
   RETURN
      <figure>
         $f/@*,
         $f/title
      </figure>
</figlist>
```

Table 2-1 – Examples of XML Queries

A key distinction between data in XML and data in traditional database models is that data in XML is not rigidly structured. It is possible to have duplications of the same elements or missing elements, elements can have atomic values in some data or structure values in others, and a set of elements can have a heterogeneous structure. In addition XML can be used to represent rigidly structured data, i.e. records, as well as unstructured data, i.e. free text. So we need to have tools that allow for searching on both types of data. Database style query languages are important for the structured parts of XML documents, and information retrieval techniques for the unstructured parts. But the integration of these languages makes sense, since there are advantages in using information retrieval on structured data, when the structure is not known, or when there are discrepancies between different XML schemas.

Keyword search was integrated by Florescu et al. (2000) into an existing XML query language. They extended inverted files to be stored in a relational database, in a way that could also support keyword search along with other operations. The XML-QL query language (Deutsch et al. 1999) provides a solution to querying XML information, but assumes that a reasonable amount of knowledge about the structure is known. If this is not true, the addition of the *contains* predicate into XML-QL allowed for a search on XML elements, combining keyword search and regular (structured) query processing. However the implementation of the inverted file implies the possibility of having to use tens of thousands of tables, and this can be a problem in some relational databases. However, their evaluation was only done for performance reasons.

### Resource Description

Within the present Web structure a major problem facing information resource discovery tools is the absence of a mechanism for resource description. Therefore the notion of metadata for accessing information in the Web has been receiving a lot of attention. However there are some problems for the generation of metadata including trust, interoperability, author laziness and the fact that it can be time consuming.

Reliance on authored metadata is of key importance for creating a meaningful, trustworthy and usable Web. In the library world there is a culture of metadata creation and reliable sources are well known. But the Web got off to a bad start in this respect with the misuse of the <meta> header tag in HTML files. This tag should be used to properly assign information about documents (meta-information or metadata) but unfortunately it is common to find keyword spamming in this tag as author attempt to bias retrieval results for inappropriate reasons.

The domain specific solutions adopted by many systems causes an equally significant problem. There is a need for a standardisation on a widely accepted resource description schema, where applications can, without user intervention, interoperate and share information. Some standards being introduced recently are preoccupied with the construction of a semantic web of interconnected documents. Librarians have been using for some time the Dewey Decimal Code (DDC 1998), conceived in 1873 by Melvil Dewey. But new standards include the Resource Description Framework (RDF) (Lassila et al 1999) implementation of the Dublin Core (Weibel et al. 1998) encoded with XML (Bray et al. 1988). RDF can help information discovery tools, like search engines, and other tools to exchange and share metadata.

RDF is a foundation for processing metadata, recommended by the W3C in an attempt to introduce a language for machine-understandable descriptions of resources in the Web, without making assumptions about any domain specific semantics. The main goal of RDF is to facilitate the automatic processing of Web resources, and it can be used in resource discovery to help search engines, in cataloguing of content and content relationships, in content rating, in rights management, in privacy preferences and in digital signatures.

RDF metadata is generally encoded in a series of statements consisting of a resource or subject (anything addressable with an URI), a property or predicate (a characteristic, attribute or relation of the resource, e.g. ‘author’) and a value or object (e.g. ‘John Smith’). Resources can be the property values of other statements. RDF doesn’t impose a metadata schema. In practice we can mix different metadata schemas in a statement by using XML (Bray et al. 1988) Name Spaces. XML is not compulsory for RDF coding but nonetheless it is preferred for interoperability reasons.

### Open hypermedia and the new standards

RDF and open hypermedia are two strategies for describing relationships between resources, but they act at different levels. Both allow authored metadata to be stored separately, and to provide a structure to the information, thus adding value to it. However there are differences, notably in resource identification. The URI addressing of resources in RDF is more limited than LocSpecs (Grønbæk et al. 1996) in open hypermedia. They actually tackle different needs. While RDF is more suited to describing statements about documents and relationships between documents, open hypermedia is more concerned with explicitly structured construction of relationships between statements within documents.

XLink (DeRose et al 2000) and XPointer are more closely related to open hypermedia. Roughly XPointer is interested in locating regions of interest in documents, and these can be structured with XLink, which describes navigational hypermedia expressions.

XPointer is an extension of XPath (Clark et al. 1999), the XML Path Language, designed as a language for addressing structures in an XML document. These extensions allow XPointer to: locate points, regions and whole documents; find information by string matching (see example in Table 2-2) and use URI references with addressing expressions for fragment identification.

The XML Pointer Language, XPointer (Daniel et al. 2000) allows the identification of regions or fragments in any XML resource or document. To improve reliability, different locators are allowed (see example in Table 2-2), such as element types, attribute values, character content and relative position. The structures located with XPointer can be used for many purposes, but notably as link targets.

| |
|---|
| The following expression returns a range that selects the 17th occurrence of the string "Thomas Pynchon" occurring in a title element of doc.xml:<br>`doc.xml#xpointer(string-range(//title,"Thomas Pynchon")[17])` |
| This is an example that diferent locators can improve reliability. A XPointer like the following, with two parts, will have its first part fail if no DTD is available, but will have its second part succeed if the desired attribute's *name* is `id`:<br>`xpointer(id("chap1"))xpointer(//*[@id="chap1"])` |

Table 2-2 – XPointer examples

XLink (DeRose et al 2000), the XML Linking Language, allows the creation and description of explicit links between resources. Besides traditional links, XLink allows bi-directional links with different types, and links may be stored within or outside the documents containing the information content (see examples in Table 2-3). The traversal of links can deploy different actions (embed, replace, new or behaviour specified elsewhere) and may be initiated by users or automatically (on-load, on-request or behaviour specified elsewhere).

A linking element defining an out-of-line bi-directional extended link involving two remote resources, with different attributes for locators and arcs:

```
<extended>
  <locator href="#Fred" xlink:title="Fred" role="student"/>
  <locator href="teachers.xml#Joe" role="teacher"/>
  <arc  xlink:title="Fred tutor"
        from="student" to="teacher" show="embed"/>
  <arc  xlink:title="Joe student"
        from="teacher" to="student" show="new"/>
</extended>
```

`locator` locates a remote resource

`arc` defines traversal rules

`show` defines link behaviour

An external link database (linkbase) can be loaded using *external linksets.* An **external linkset** is a link with role **`xlink:external-linkset`**, e.g.

```
  <xlink:extended role="xlink:external-linkset">
    <xlink:locator href="..."/>
  </xlink:extended>
```

When processing a document with an *external linkset*, all links in the linkbase being referenced to should be fetched and processed.

Table 2-3 – XLink examples

XLink allows for linking on non-XML documents but the integration with third party applications is unclear in this specification, which is unfortunate as this problem has been addressed in open hypermedia for some time (Hall et al. 1996, Grønbæk et al. 1999). In addition there are still many structuring hypermedia features that are difficult to implement with XLink, for example, composites, guided tours and spatial or taxonomic hypermedia.

Open hypermedia research (Hall et al. 1996, Davis et al. 1996, Grønbæk et al. 1996, OHSWG 1998) has been working on creating support for user-controlled structures that can be stored separately from documents, and has more recently tried to introduce these features in the Web (Anderson 1997, Carr et al 1998, Grønbæk et al. 1999).

Open hypermedia structures can be used as metadata information for the Web (Grønbæk et al. 2000). The Webvise (Grønbæk et al. 1999) open hypermedia system has been extended allowing users to create and manage metadata with XML for web distribution.

## 2.6 Conclusions

Traditional hypermedia and information retrieval systems help the user to find relevant information in different ways. However these systems alone face problems that can be attenuated if they are integrated. The research community has understood the need for integration and hypermedia has progressed by using information retrieval facilities and information retrieval has progressed by using hypermedia facilities. Unfortunately, these approaches rarely considered information retrieval and hypermedia in an equal way, thus they usually neglected one of the sides of the integration.

Many researchers have also considered context access to multimedia, but the majority have not provided adequate evaluation of the validity of their models.

In the way we see the integration between the two systems, there is still a need for a better and more coherent model of integration. In the next chapter, an initial model of integration is proposed as a first approach to enable multimedia access using hypermedia structure. This model evolves in Chapter 4 where a better consideration of hypermedia and information retrieval at both the document and the abstract level is achieved by integrating the content and structure associated with each side of the integration.

After describing the implementation of the model, evaluations concerned mainly with multimedia access are then presented.

# Chapter 3 - First model, a step towards context descriptors

## 3.1 Introduction

As we have discussed previously, there is a need to implement an undemanding information-retrieval system integrated with a hypermedia system that can allow the user to navigate through the information without being lost in 'hyperspace'. Nevertheless, this need cannot prejudice the efficiency of the system, for any type of media, when the user is using it as an everyday tool.

Multimedia access using text queries, focused on the retrieval of any document type was one of the main targets of this first model. It was developed to check the validity of our hypotheses for multimedia access. Good ad-hoc results would then encourage us to further extend this model and then test it more consistently.

The next section (3.2) will introduce a way of accessing multimedia documents using the hypermedia context surrounding these documents. Suggestions for the construction of hypermedia context text descriptors are given by using metadata from the two levels offered in the open hypermedia system Microcosm, namely: document level and abstract level. This section ends by proposing a way of weighting the different metadata in the descriptor.

Section 3.3 describes the way in which this first integration was implemented. We intended to reuse an existing text information retrieval tool and for this reason techniques were developed to allow multimedia retrieval within the open hypermedia system Microcosm. This was done in a way that the meant integration would be automatic and transparent to the user.

## 3.2 Text description of multimedia files using context information from the hypermedia network

### 3.2.1 Accessing multimedia through hypermedia context

In many hypermedia systems that use information-retrieval tools, such as Microcosm (Davis et al. 1992) with Li's (1993) Computed Links filter, to reach non-text information through retrieval we have to expect that when text documents are retrieved, it is possible to get to non-text documents (related to the text documents) through navigation from the retrieved text documents. This assumption is based on the fact that hypermedia links usually connect related information.

As we have seen, Frisse (1988) in his hypertext medical handbook developed the concept of retrieval to aid navigation, using the dependency of documents/cards in the hypertext structure to improve the retrieval. In further work, Frei et al. (1992) used not only the content of text files but also the content of semantic links, and, as has been seen before, the semantic links were built with keywords retrieved from the source document and destination document of the link. This system considers not only the documents but also considers the link network between the documents. These previous approaches support the conclusion that the hypertext link structure at the document level is something to be considered for supporting information retrieval.

If it is not possible to retrieve information from non-text media as easily as from the text medium, then a text description of non-text media could allow an indirect application of text techniques to non-text media. Text descriptions for non-text media could be built directly by a human indexer, but this is a time consuming job and is also a technique that is very much based on the indexer's personal understanding of the media.

As was proposed by Dunlop et al. (1993), the existence of links could also be used to produce a computed text descriptor of non-text documents that could permit them to be retrieved directly from text queries. As has been seen, all text documents that are connected to a single non-text document can be considered as a set, where the common characteristic is the related information about the non-text document (see Figure 3-1). In their approach, they applied the cluster (set) centroid algorithm to calculate a descriptor that would be the average meaning of all the text documents represented in the set, and as they claimed to prove, that descriptor represents the non-text document in a reasonable way as seen in Figure 3-1.

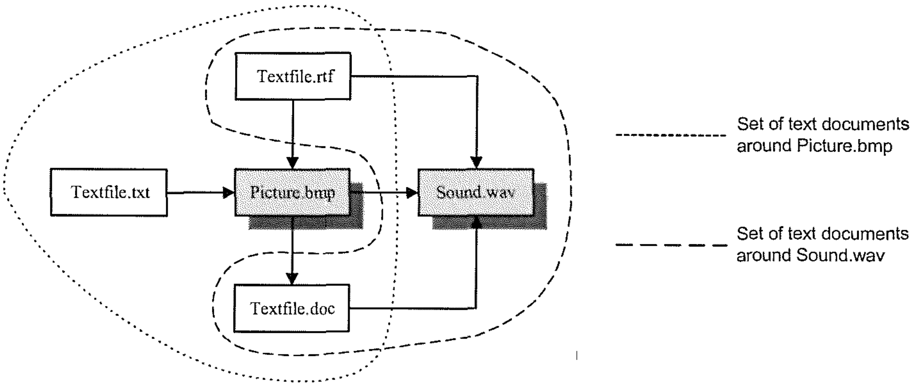

Textfile.rtf
Textfile.txt
Picture.bmp
Sound.wav
Textfile.doc
Set of text documents around Picture.bmp
Set of text documents around Sound.wav

Figure 3-1 - Links between text and non-text, build a set of related information around non-text

Our proposed model of text description to represent multimedia information is more extensive because it uses more than simply the content information from neighbouring documents (Marinheiro et al 1998). With the open hypermedia system Microcosm (Fountain et al 1991, Davis et al. 1992), where there is much more information available about the hypermedia network, a more precise text description of multimedia information is possible. An improved description can be achieved not only by extracting the relevant information from the hypermedia application at the document level extending the Dunlop et al. (1993) approach but also by extracting the relevant metadata information about the documents that is maintained at the abstract level for user browsing.

### 3.2.2 Metadata from the document level for multimedia text descriptors

As a first step we produce some improvements by refining the way the content information is extracted from the neighbouring documents of the multimedia files being described. In our approach the neighbouring documents of a multimedia document are the ones connected with at least one direct link to that multimedia document.

In their extraction operation, Dunlop et al. (1993) took into consideration the entire content kept in the neighbouring documents and not simply the contextual information in the document that could be related to the file being described. However, in Salton et al. (1994) and Salton et al. (1996) a different approach is suggested to retrieve information from text documents through decomposition of the text into 'text segments' and 'text themes'. It seems reasonable that a text segment, or theme, could be much more relevant to the construction of the non-text document descriptor than the whole text document, since it is possible that the same document/node contains a lot of information that may not be related to a particular destination document of a link that starts from a marginal sub-topic of the original document.

With this idea we can consider a different method of integration in the Microcosm open hypermedia system when extracting content information to build multimedia descriptors. So, instead of considering all the information kept in the documents, as Dunlop et al. (1993) did, it could be better to only consider the segment(s), or the theme(s), of the documents that have links to the non-text document to be described. In this way the quality of the descriptor should increase and as a result the information-retrieval efficiency of multimedia should increase as well. The Salton approach is, however, time consuming and in large applications this might be a significant disadvantage. We therefore considered the information (text) kept only in the link anchor selection to build reasonable descriptor more easily.

Further development could also be achieved by considering the phrase or the paragraph where the anchor is located. Nevertheless, for the initial model, it is possible to suggest that the anchor information will be more closely related to the described document than all the text information stored in the text file. This is justified by the work developed in Harmandas et al. (1997) on the World Wide Web, where they show that considering the full text where the image is inserted is the option that provides the worst recall/precision for multimedia retrieval, and that the information that provides the best recall/precision is the image text caption.

Another possible way of extending the model of text description was by considering other valid meta-information that is available in an open hypermedia system such as Microcosm. In Microcosm there is a small text description about each link, and this kind of link description can also be used to help in the building of the text file descriptor for each non-text file, in the same way that the link anchor can. Frei et al. (1992) have already used link description to improve recall/precision with good results, particularly with author description. However in their work the use of the link description for information retrieval required a time consuming computational model. But since we will only use the link description for the construction of the text file descriptors, which is not done at run time, then this problem will not occur.

Finally another way of extending the model of text description was achieved by distinguishing between different kinds of links. An advantage of Microcosm is that there are several different kinds of links for different functionality. Usually, generic links and local links are used for definitions, examples, and related information, while specific links are typically used for the back-bone navigational link structure of the application at the document level to allow the user to navigate between different items of information that may be obviously related or not. It is clear that, as the different links relate information in a different way, their use to improve information retrieval should be considered (Savoy, 1996). In particular, it should be possible to differentiate the use of the information associated with each link in the description text file.

Generic links usually relate information that is semantically closer to the content, so they are not particularly dependent on browsing information, whereas specific links may give more information about the hypermedia network functionality of an application, so they are less dependent on the content. It should be possible to choose a weight, dependent on the type of each link, which could allow the attribution of more relevance to generic links. This weight could represent the number of times the information referent to a link is repeated in the text file descriptor.

### 3.2.3 Metadata from the abstract level for multimedia text descriptors

Until now we have only considered the construction of the text file descriptor from the information taken from links that the system has at the document level. A lack of links at this level may degrade the retrieval efficiency of non-text media, as has been shown by Dunlop et al. (1993). In addition, sometimes the links to images are not relevant or relate information in a non-descriptive manner, as discussed in Harmandas et al. (1997).

There has been a significant amount of research on "hypermedia-like navigation" through a semantically connected set of concepts at what we call in this thesis an abstract level (Agosti et al., 1992, Arents et al., 1993, Cunliff et al., 1997). In their work, multimedia information retrieval was possible with the same efficiency irrespective of whether the media to be retrieved was text or non-text. The classification of documents against a semantically connected thesaurus could allow retrieval through the concepts of classification. In works like Arents et al. (1993), the creation of trails through the information using semantic coupled hyperindexes has also been suggested.

More recently this abstract level has been considered in the web as a subject or other kind of classification that allows for browsing searches. This is available in systems like Corbis (http://www.corbis.com/), Yahoo's Image Surfer (http://isurf.yahoo.com) and AltaVista's Multimedia Search (http://www.altavista.com), among others. Amato et al. (1998) and Mukherjea et al.

(1999) also give directions on the use of abstract levels for multimedia access in the web, but no context indexing of multimedia was implemented using this metadata.

In Microcosm, valid significant information is kept at the abstract level. This suggests its possible use for the correct building of the text file descriptor for multimedia. To implement navigation in Microcosm at the abstract level the author is usually required to create a small text description of each file, some keywords, and its classification under an abstract concept/index that is called the logical type. Thus, this important information should also be used in the same way as the information taken from the links to build the text file descriptors. The new system proposed here would then always have a small text description that could allow the retrieval of non-text media, independent of whether there are any relevant links.

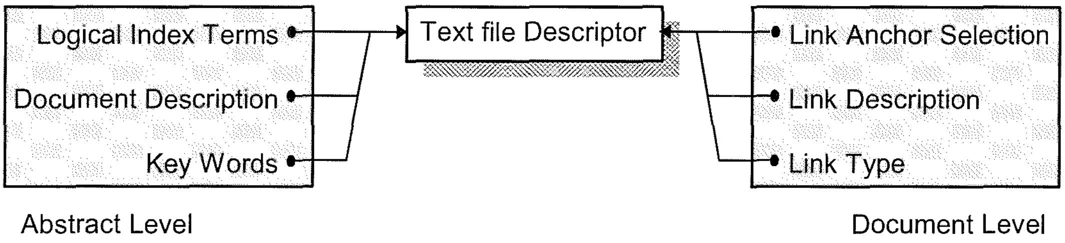


Figure 3-2 - Information that is used to build the text file descriptor for a non-text document

### 3.2.4 Calculation of the descriptor file

In summary, for the construction of the text file descriptors, our approach uses not only the information kept at the document level, as the majority of research has done until now, but also information maintained at the abstract level, improving the possibilities of better retrieval. Another novelty is the consideration of different link types that allow the distinction between different kinds of information taken from the document level and from the abstract level, as shown in Figure 3-2.

The proposed Equation 3-1 shown below, used to build the text descriptor files, makes it possible to change the weight of several metadata parameters to improve the retrieval quality of the information-retrieval tool to integrate with the Microcosm system:

$$l_g \sum_{L \subset Ge} \left( a \cdot L_{iS} + b \cdot L_{iD} \right) + \sum_{n \subset N} \left( \begin{array}{l} l_l \sum\limits_{L \subset Lc} \left( a \cdot L_{iS} + b \cdot L_{iD} + c \cdot n_{jD} + d \cdot n_{jA} \right) + \\ l_s \sum\limits_{L \subset Sp} \left( a \cdot L_{iS} + b \cdot L_{iD} + e \cdot n_{jD} + f \cdot n_{jA} \right) \end{array} \right)$$

Equation 3-1 - Weighting consideration for text file descriptor construction in the first model

Where:

L ___ link

Ge __ set of generic links connecting to the non-text node

Lc __ set of local links connecting to the non-text node

Sp __ set of specific links connecting to the non-text node

$L_{iS}$ __ link anchor selection on link i

$L_{iD}$ __ link anchor description on link i

$l_s$ ___ integer weight given to the specific link

$l_l$____ integer weight given to the local link

$l_g$ ___ integer weight given to the generic link

n____ node

N ___ set of text neighbour nodes around the non-text node to be described

$n_{jD}$ __ description of node j

$n_{jA}$ __ logical index of node j

a…f _ different user /author integer weights

Besides the different weighting depending on different link types, that has already been justified, there is also the possibility of changing the relative weighting between metadata coming from the document description, the document classification (logical type index), the link description and the link anchors. This allows the system to have different text description file construction strategies for different applications. Some hypermedia meta-information in Microcosm is author/user dependent and its importance in the construction of the descriptor must reflect author/user behaviour.

In Equation 3-1 the weights for link description, link anchor selection, document description and document classification are additive. By this we mean that they represent the number of times each piece of meta-information is repeated in the text file descriptor. Link type weight is however multiplicative, since there is no text associated with the information we get from knowing the link type. By this we mean that all other remaining metadata parameters weights containing text are actually multiplied by the weight for the link type.

In this first model, the document description and classification of neighbour documents for generic links is not considered. This is because for generic links, in theory, there is no need for source documents. We should also note that in Microcosm there are no generic links with source anchors on non-text files.

## 3.3 Integration of an open hypermedia system with text information retrieval to facilitate multimedia information retrieval

### 3.3.1 Goals for the first model of integration

In the implementation of this first model described in detail in Marinheiro et al (1998), there were two main goals to be achieved when considering the extension of an information-retrieval tool to incorporate the retrieval of multimedia information within the open hypermedia system Microcosm:

-The first goal was to design an integration model that allowed for the reuse of an existing text information-retrieval tool behaving like a “black box”. This allows for the future reuse of different text information-retrieval tools that can easily be inserted into the modular architecture of the Microcosm system;

-The second goal was to integrate the information-retrieval tool in a transparent way for the user/author, i.e. to do it in such a way that the user/author would be hardly aware of its existence, apart from times when it is necessary to change parameters. To implement this it is necessary that the system tracks down all the information about the hypermedia network that may change in the application and update the affected text file descriptors automatically.

### 3.3.2 Reusing of an existing text information-retrieval tool

In Microcosm, each file that exists in a specific application must be registered under the DMS (Document Management System). The registration information is kept in a database record that holds all the information necessary to control the file. The system permits the creation of a new field in that record and this allows for the creation of one extra field in each text file descriptor record containing the physical reference of the file that is being described. The reverse situation was implemented as well by creating a reference field in the described file record. This reference points to its descriptor, as shown in Figure 3-3. Consequently, there is a physical reference association between the descriptor and the described file, and it is therefore simple to find one if we have the other. This registration information allows the retrieval of a non-text file when its indexed text file descriptor is retrieved by a text retrieval tool, as will now be shown.

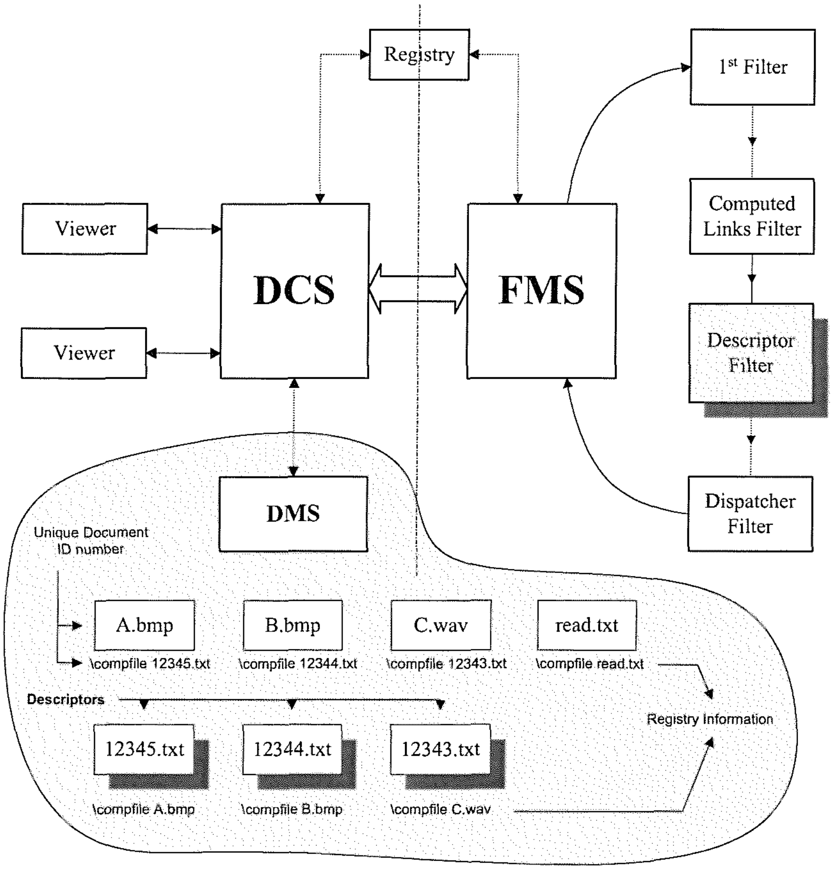


Figure 3-3 - Integration of the Descriptor Filter with the Microcosm System for the retrieval of multimedia information

It has been seen that in the Microcosm system the management of all information relevant to the link services is held in the back end and is processed independently from the front end, where the viewers interact with the user. These back end link services are managed by a sequence of filter modules that communicate with each other by messages. Each filter processes the messages that are relevant to the information that it is responsible for, and then passes on the processed message to the next filter, in a new message or not.

One such filter in the filter chain responsible for the text information-retrieval capabilities inside Microcosm is the Computed Links filter developed by Li et al. (1992) and Li (1993) as described in the previous chapter. The output of this filter is a set of ranked text files from the set of all text files in the application. This ranked set of text files contains suggestions for starting points for navigating at the document level of the hypermedia application. These starting points are documents that are similar to the text selection that the user has specified. For each suggested file a message is sent containing its reference. These will be processed by the Dispatcher Filter, which will show the brief abstract level document description of the suggested files, in the same order as the messages were received. The user then only has to select the desired file(s) in order to see them.

To satisfy our first objective of integration a new filter - the Descriptor filter - was developed and inserted into the Microcosm filter chain after the Computed links filter and just before the Dispatcher filter, as presented in Figure 3-4.

Since Microcosm messages are passed from each filter to the next in sequential order, the Descriptor filter, inserted in the filter chain after the Computed Links filter, has to keep track of the file registration information part of the messages that go through it. In this way it is sure to catch all non-text files represented by descriptor text files.

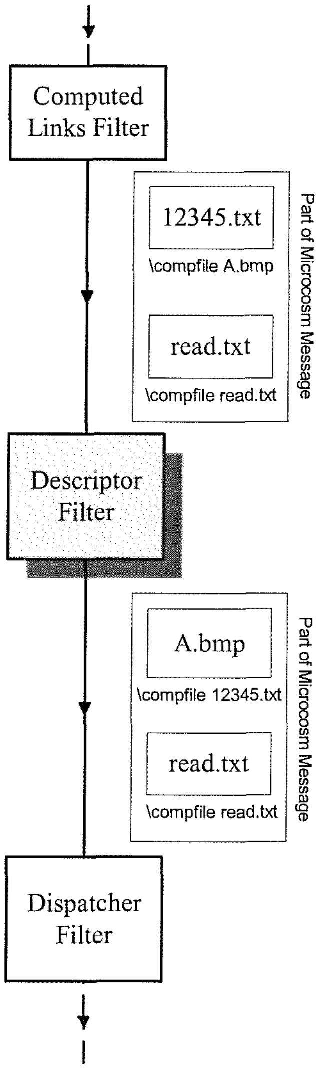


Figure 3-4 -The changing of messages to retrieve the right document

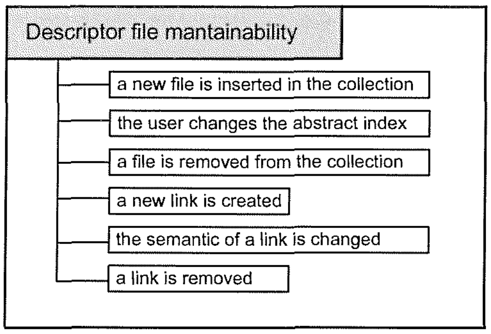
After the Computed Links filter passes on the text file descriptors and normal text files, the Descriptor filter has to check in the message for references to the described files. In the messages where this information is found, it is necessary to replace it so that the according described file is passed on as shown in Figure 3-3 and Figure 3-4.

In this approach, the Computed Links filter does not have to be aware that it is retrieving a descriptor file, nor does the Descriptor filter have to know how to retrieve information in response to a query. So, the independence of the two processes is assured. Finally, the Dispatcher Filter shows the user all the retrieved documents, which may know be references to every kind of media, for example: text, images, sound, videos, etc.

Using this method, it has been shown that it is possible to reuse an existing text retrieval tool and extend it to incorporate multimedia retrieval. In principle, the Computed Links filter could be replaced by any text retrieval tool that is able to receive and output messages. This approach benefits from the use of the file registration mechanism and the flexibility of the filter chain structure in Microcosm and also from the existence of one filter that retrieves text files.

### 3.3.3 Transparent integration with an existing text information retrieval tool

Now it is necessary to define how the Descriptor filter should behave in order to satisfy the second objective: transparency. One of the major problems with the Computed Link filter developed by Li et al. (1992) was its static behaviour, i.e. the actual content or structure of each application was not considered by the filter, unless the author specified it. In more detail, the Computed Links filter does not change the inverted file information necessary for retrieval when documents are inserted or removed from the collection. So every time the user/author inserts a new one or removes an old one, he/she has to recreate the inverted file. To avoid a similar problem in this new filter, it is necessary to carry out some other procedures to allow a transparent integration for the user/author, with the Microcosm system. Table 3-1shows the situations where new procedures are necessary.



Descriptor file mantainability

- a new file is inserted in the collection
- the user changes the abstract index
- a file is removed from the collection
- a new link is created
- the semantic of a link is changed
- a link is removed

Table 3-1- Situations to which the Descriptor filter reacts in order to maintain consistent descriptors

When a new document is created, a notification message is put into the filter chain. So, when the Descriptor filter detects that the new non-text document has been inserted, it selects an appropriate name and creates the text file descriptor with the information provided by the document registration attributes. When this Descriptor filter is inserted for the first time in the system, during the creation of all text file descriptor, it also asks the linkbases about all the links that are pointing to each described document. If any, it gathers all the text in the source anchors and descriptions and includes it in the text file descriptor, applying Equation 3-1 described previously. However, special care is taken in the storage of this information since all of it must be recognised by the Computed Links filter. Since only text information should be saved, separation between different parts of the text descriptor is achieved through the use of a different number of linefeeds. This might be important to allow the distinction between the different links in order to permit their maintainability when they change. When a file is removed from the collection it is only necessary to remove the descriptor text file.

When a new link is created, the Descriptor filter has only to catch the message that created the link in order to retrieve the text selection and link description from the message and store it properly, by applying Equation 3-1 for the text file descriptor update, together with information taken from the registry concerning the linked neighbour documents.

If a link is removed, the text concerning the source anchor and the link description stored in the text file descriptor could be taken out and when the document registration attributes or the user description is changed, the text file descriptor could also be updated. However for ease of implementation under these circumstances we opted for the construction of a new text descriptor. This new descriptor replaces the old one and uses the new meta-information along with the unchanged one.

The only time the user has to be aware of the information retrieval tool is when it is necessary to update the index for all the text files, and among them the context text descriptors for multimedia access.

## 3.4 Evaluations and conclusions

With this approach (as described in Marinheiro et al 1998), some ad-hoc qualitative experiments were undertaken to check on the validity of the hypotheses used in this first model.

Two different small applications were built with the same 20 documents but with different hypermedia structures, showing different strategies in the building of applications. The first application simulated the situation where the author did not construct the hypermedia application using the extra functionalities that Microcosm system provides. In this application the logical type index and text description at the abstract level were not valid or related to the content of the documents. At the document level the only valid information considered was the link anchor selection. The second application used the potential of the Microcosm system for the automatic construction of text descriptors. This was a better-planned hypermedia application with valid information both at the document and at the abstract level.

For the first application, the recall of multimedia files did not look very high for subjects related to the one addressed in the non-text document to be retrieved. In the second one however, since we had considered information from the abstract level and used the link description, the recall of multimedia files was improved even when non-text files had a small number of links connecting them.

With these first results, it was already possible to conclude that the integration worked well in applications developed for Microcosm, although further improvements were clearly still needed. In the ad-hoc experiments, it was noticed that the text description might be mainly dependent on two factors.

The first factor concerns the quality of information used at the abstract level. This information improves the retrieval of multimedia information through the use of text descriptors if the author has built the applications in a planned way. If care has not been taken in the insertion of information at the abstract level, that information cannot be beneficial for retrieval. However, for a planned hypermedia application, the good results obtained justified a better study of the abstract level for future models.

The second factor is the number of links connecting non-text documents with text documents. In a poorly connected hypermedia application the efficiency of retrieval seems to diminish. The consideration of segments (themes) as well as link anchors is therefore something that should be considered in the future. But the use of information taken from the abstract level allows the retrieval of multimedia files even in a situation where the hypermedia application is poorly connected.

This first model demonstrated that automatic text file descriptor construction using hypermedia metadata is important and relevant for multimedia access. It encouraged us to continue the research and expand the integration model, with a more flexible and powerful information-retrieval tool, which would allow better integration of information retrieval and hypermedia browsing for different media.

These first ad-hoc results also show a need for a more analytical study of the importance of the different hypermedia network metadata, such as link type and description and abstract level logical index and description, for the construction of test file descriptors.

The next chapter will present a second improved model for the integration of information retrieval with hypermedia browsing. Chapter 7, Chapter 8 and Chapter 9 give more analytical results regarding the relative importance of different metadata according to different authoring styles demonstrated in some applications presented in Chapter 6.

# Chapter 4 - Improving the model, a better integration of hypermedia and information retrieval

## 4.1 Introduction

As it has been observed (Agosti et al. 1996b), in hypermedia systems it is usual to use manual browsing through explicitly authored links, while in information retrieval systems there is usually the construction of analytical queries returned by a ranked set of documents. Agosti believes that although some recent research has attempted to integrate these two approaches, they still tend to appear as distinct phases of a user session.

Our goal for the improved model proposed in this thesis is then to find a better integration between information retrieval and hypermedia. Up to now, there have been different ways of viewing the integration of the two systems. The differences between the integration approaches can be classified under the two dichotomies: structure vs content and document level vs abstract level.

In a way, hypermedia systems and information retrieval systems can be a mirror image of each other. Usually in classical hypermedia systems there is more emphasis on the analysis of the structure of the information while in classical information retrieval systems the emphasis is placed mainly on the semantic content analysis of the information. Only later on in hypermedia systems did the semantics of links start to become important in helping the user in browsing. It then makes sense to say that structure is to hypertext as content is to information-retrieval systems. Further, we might say that links representing structure are to hypermedia systems what index terms representing content are to information-retrieval tools. Or, in other words, a network of links in a hypermedia system connecting documents, or parts of them, carry out the same task as the set of all indexing terms in an information-retrieval system. So to integrate the two systems, we must consider more structure and link information on the information-retrieval side and we must consider more content and index term information on the hypermedia side.

The abstract level and the document level dichotomy have already been present for a long time in many information retrieval systems for content retrieval of information as it was present later on in hypermedia systems for structural information. Thus, much of the effort in integrating these approaches has been made towards these two ways of viewing the abstract level, since in practice this level contains the same structure and content information, but gives different tools to access it. However, when the document level is considered, the integration between the two approaches has not always been regarded in the same way.

So, the integration between information retrieval systems and hypertext systems should be made both by considering structure and content at the abstract and at the document level in a new system where the user is not able to say if he/she is dealing with a hypermedia system or with an information-retrieval system.

The following section (4.2) presents a better model of integration with a new information retrieval tool that allows, among other facilities, multimedia access. This model tries to be open in the way it interacts with the exterior. Ideally this could be used with any system for multimedia access, including the open hypermedia system Microcosm. The flexible information retrieval tool that as been implemented to be used in this model has been described in section 4.3

Section 4.4 and 4.5 show the improved ways of integrating hypermedia and information retrieval when considering information respectively at the abstract level and at the document level.

Section 4.6 presents the way in which the different meta-information is considered using weights for the context indexing of multimedia.

## 4.2 An open integration with information retrieval

A difficulty that arose from the previous model was the limitations imposed by the information retrieval tool used in the integration - the Computed Links filter. It was a closed, inflexible, solution tied to the Microcosm processing structure. The information-retrieval algorithms and the communication protocol with the Microcosm filter chain were inserted in the same module. In addition there were some integration problems with this system, notably by not interpreting all available interaction information, which prevented the tool from having an up-to-date index tree that would reflect the real content of the system.

In the modelling and developing of a new information retrieval tool we had then in mind the independence and openness from any other hypermedia system (or even other systems, like databases) with which it could be integrated. To achieve this, a flexible three-layered model (as shown in Figure 4-1) has then been designed attributing to each layer a specific task. This three-tiered integration has a translation interface code layer, an information retrieval computation code layer, and a document access/translation layer. The model allows for the distribution of these different layers in different systems, which has actually been tested, outside the scope of this thesis, in a multi-agent environment (Moreau et al. 2000). However the adopted implementation in this thesis does not have a distributed approach. Instead, these layers communicate through a set of predefined APIs resembling the client-server paradigm. These layers have actually been implemented as Dynamic-Link Libraries (DLLs) in a MS Windows operating system environment.

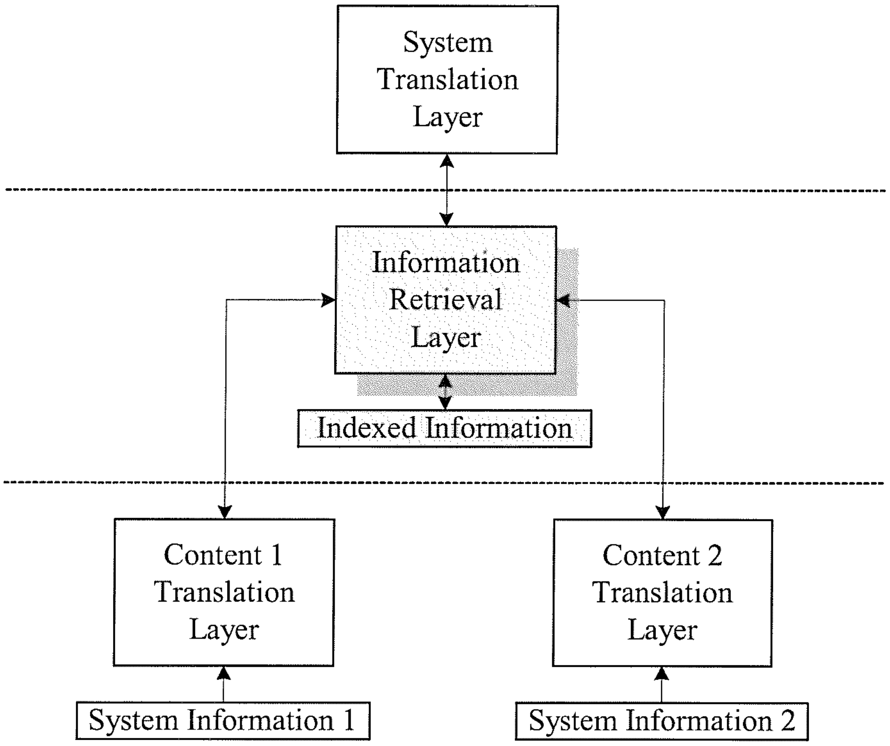


Figure 4-1 - A new independent information retrieval tool.

The content access/translation layer is responsible for the access to any information that is intended to be indexed. This information can be accessed by content and by context. When accessed by content different text formats can be used, and also the entire document or its parts can be considered. When accessed by context all meta-information available for this document is used to build a text descriptor and then any media format can be indexed in this manner. Context indexing is the method used for multimedia retrieval in this model. In theory this integration model can use any hypermedia system if a proper content translation layer DLL is developed. Actually more than one such content translation layer DLL can be used simultaneously. It is therefore possible to expand or change access facilities in this model if new or better methods are developed to access information for indexing. The DLLs are loaded dynamically when the required accessing methods are required and according to a configuration file. However it should not be forgotten that this layer must return an unformatted text buffer to the text information retrieval layer.

In this thesis we have used the Microcosm open hypermedia system and a proper DLL was developed for its access. The context indexing of information, by using text descriptors, is built according to mechanics that will be further explained in this chapter, and according to the weighting function explained in section 4.6.

To allow for an open independent integration a system interface layer must be used. This layer is responsible for the management of information received from the used system and for the translation of this information in such a way that the information retrieval tool and the system can interact. This must be done in a way that a proper integration is achieved. We can say that this layer together with the content translation layer is responsible for the interface between information retrieval facilities and other system facilities like hypermedia browsing.

The integration with Microcosm was achieved with the construction of an interface filter. Different messages are traversing through this filter, and the ones of interest for the integration must be analysed. These different messages can be about a link update, a link creation, a document inclusion, a request for indexing a document/item by content or context, a request for a removal from the indexes, a request for index merging, a verification of the type of indexing available for a certain document/item.

The model of integration allows, together with author-indexed information, the easy use of user-indexed information that can be changed according to a user's information needs. This is similar in a way to some hypermedia models such as Microcosm, where there are application/author linkbases (i.e. link databases) and user linkbases. With this facility, the user has the ability to specify his/her own topics of interest for future retrieval and to have the links created by him/her reflected in the information retrieval side of integration.

The different auxiliary index files are stored but only one may be used at run time for retrieval. For this reason, the management of this stored index information is achieved in Microcosm by using different interface layers, each one responsible for a different type of indexing, as shown in Figure 4-2. This could be author indexing, user indexing or any other specific subject indexing that we might wish to add.

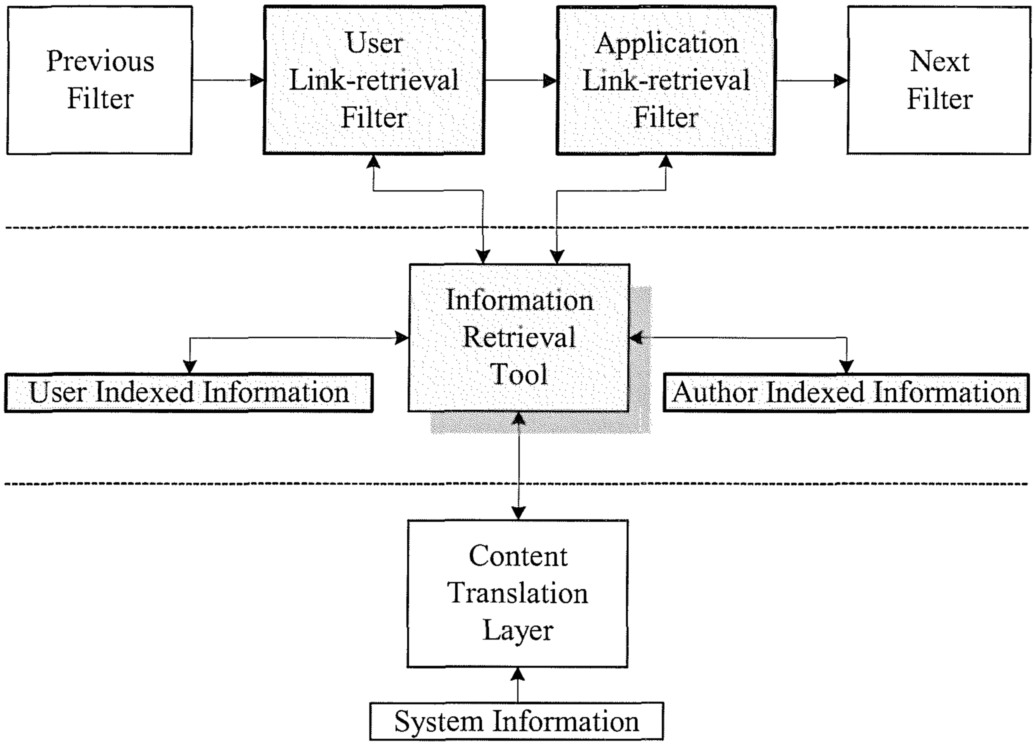

Previous Filter
User Link-retrieval Filter
Application Link-retrieval Filter
Next Filter
Information Retrieval Tool
User Indexed Information
Author Indexed Information
Content Translation Layer
System Information

Figure 4-2 - Integration of the new developed information retrieval filter with Microcosm

However, for a proper retrieval at run time, only one inverted index is used. This inversion is performed just after the merging of all non-inverted index information that is currently being used, as will be further explained in section 4.3.

In the next section the implemented central information retrieval layer responsible for the core text information retrieval algorithms will be further explained in detail. With this model it could be possible to change between different information retrieval tools as long as the interface APIs are kept the same. This layer provides a set of APIs to be used by used in theory by any system, as shown in Figure 4-3. These APIs allow the creation of auxiliary indexed information from scratch by using: *CreateStemIndex*. They also allow the addition, updating and deletion of items or even the merging of indexes, by using: *AddStemIndex* and *DeleteStemIndex*. The indexed information is inverted by using *InvertIndex*. This last method is actually used in our implementation, but it should not be part of an open model for information retrieval, since it makes the assumption that there is a need for an inverted index. This should be made transparent to the system in future implementations. *IsItemIndexed* and *WhichItemsAreIndexed* are used to check on the structure of indexing. *ItemMatchItem* and *StringMatchItem* are used for finding similarities to queries when they are respectively input by the user or based on existing indexed items.

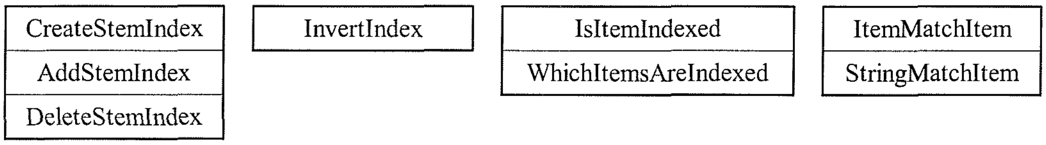


Figure 4-3 - Set of APIs provided by the information retrieval layer

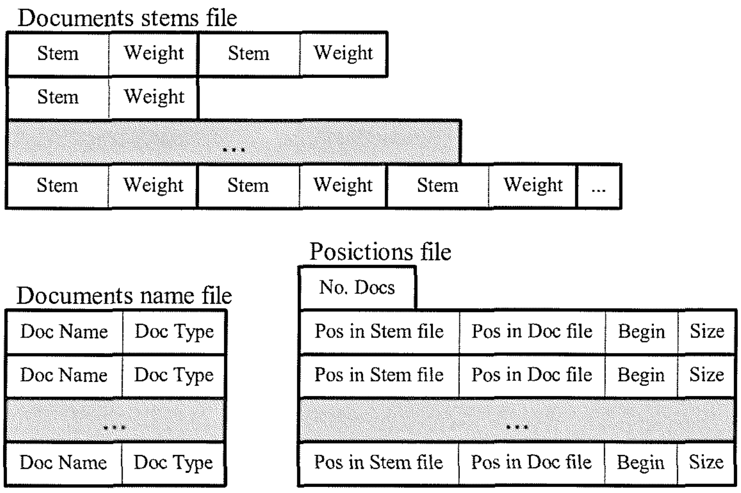
## 4.3 A flexible information retrieval tool

The algorithms that have been implemented in the information retrieval layer comply with the standard vector space model (Rijsbergen 1979, Salton 1989). This indexing and retrieval model has been proved to be at least as effective as the probabilistic model, with the advantage of being of easier implementation (Baeza-Yates et al. 1999, pp 34).

For performance reasons, in information retrieval, data is usually pre-processed in some way. Traditionally only one auxiliary data file (containing inverted stem references) is created in the vector model approach. However in our approach the pre-processing of auxiliary data has been divided in two distinct steps. In each step a set of auxiliary data files is created and this allows flexibility in the dynamic updating of these auxiliary data. Documents/Items can be added, removed or updated without having to recalculate all auxiliary data. It is even possible to merge different auxiliary data.

In the first step indexes for any text documents (that might be a context description of multimedia) are created. Each document is read and its content is processed through a set of algorithms (Frakes et al. 1992). Stop words (frequently used but having no real meaning for retrieval) are removed at first. Then all words are stemmed to allow for the indexing of variants of the same word to be indexed together. After this a frequency count of stems is performed for each document, and its inverse frequency calculated.

Three different intermediate auxiliary data files are created at this stage (see Figure 4-4). One file (documents stems file) contains for each document all the available stems and their weights. A second one (documents name file) contains documents or item names and types. A third file (positions file) contains for all documents the reference positions in the first two files and a beginning offset and size if a item that is a fraction of a document is being used. These three files hold pre-processed data that does not have to be recalculated when there is a change in the hypermedia document collection or when more than one index is considered at one time and then a merge of indexes is required.



Documents stems file

| | | | | | | |
|---|---|---|---|---|---|---|
| Stem | Weight | Stem | Weight | | | |
| Stem | Weight | | | | | |
| ... | | | | | | |
| Stem | Weight | Stem | Weight | Stem | Weight | ... |

Documents name file

| | |
|---|---|
| Doc Name | Doc Type |
| Doc Name | Doc Type |
| ... | |
| Doc Name | Doc Type |

Posictions file

| | | | |
|---|---|---|---|
| No. Docs | | | |
| Pos in Stem file | Pos in Doc file | Begin | Size |
| Pos in Stem file | Pos in Doc file | Begin | Size |
| ... | | | |
| Pos in Stem file | Pos in Doc file | Begin | Size |

Figure 4-4 - The three intermediate auxiliary data files structure, created in the first indexing stage

The document stem weights stored in the auxiliary data files are calculated in accordance to Equation 4-1, where $freq_{i,j}$ is the frequency of occurrence of stem $k_i$ in the document/item $d_j$ and '$\max freq_j$' is the total number of stems in document $d_j$. This equation gives more importance to words that occur more often, but their occurrence is normalized to the size of the document. The logarithm in the equation gives more importance to variations when they happened for low values of occurrence. We believe that a variation of occurrence between 1 and 2 might be more significant than a variation between 10000 and 10001.

$$w_{i,j} = \log\left(\frac{freq_{i,j}}{\max freq_j}\right)$$

Equation 4-1 – Stem weights for documents

In the second step a stem inverted index file (see Figure 4-5) is constructed using the three auxiliary data files created in the first step. These three files can actually be the merge of different index information when more than one index is used at one time. Therefore every time there is a change in these three files, the inverted index file has to be recalculated. This inverted index file is the one to be actually searched in the similarity matching algorithm.

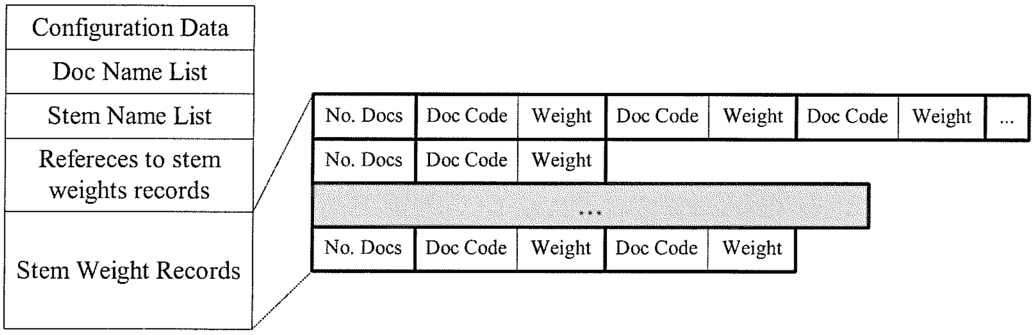


Figure 4-5 – The structure of the inverted stem index created in the second indexing stage

Two find similar documents to text queries we also have to pre-process the queries, as shown in Figure 4-6. Stop words are also removed and the remaining ones stemmed. After this the queries are expanded using synonyms. This allows for the retrieval of relevant information that can expressed in queries in different ways.

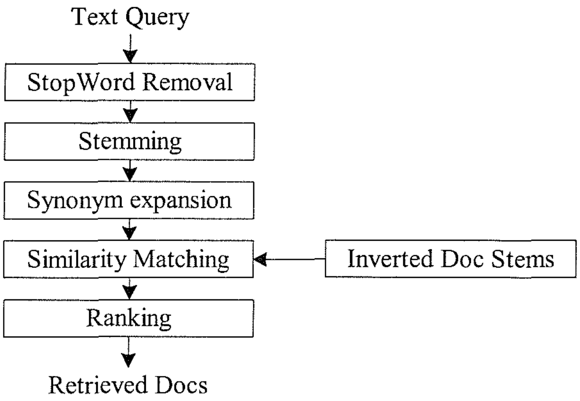


Figure 4-6 - Text query processing

The similarity-matching algorithm calculates the inner product between the stem weights of documents and the stem weight of queries (see Equation 4-3). The query stem weight $w_{i,q}$ for stem $k_i$ depends on the inverse document frequency (log(N/n$_i$), where N is the number of documents in the collection and n$_i$ the number of documents with that stem) and on the occurrence of each stem in the query $freq_{i,q}$ (see Equation 4-2).

$$w_{i,q} = freq_{i,q} \log\left(\frac{N}{n_i}\right)$$

Equation 4-2 - Query stem weight

$$sim_{q,j} = \sum w_{i,j} \cdot w_{i,q}$$

Equation 4-3 – Document similarity

The ranking of the best matches of retrieval is the output of this tool. The reference to the documents or items to be returned are ordered by their similarity to the initial query.

## 4.4 Analysis of a better integration at the document level

This section will describe the better implemented integration of information retrieval with hypermedia systems when considering structure and content information from the document level, that can allow a smooth transition between retrieval and navigation whatever media is considered.

At first (4.4.1) we show the use of more information supported by the structure of the hypermedia by using the link anchor context that will be used alongside user/author specific selections for indexing. This allows the retrieval of more specific information. The use of link anchor context is also more refined in this model of multimedia retrieval of documents using text descriptors. The model has also been further expanded to allow for the retrieval of text and non-text documents from within non-text documents.

The second aspect of document level integration is the extension of the model of the generic links by considering more content information on hypermedia. Alongside exact matching, to find generic links we will use similarity matching. By indexing the context of a generic link anchor, the move of the generic link matching algorithm to the information-retrieval side of the integration is possible and then, there will be a merge between the two ways of accessing information which implies a smoother transition between navigation and retrieval.

### 4.4.1 Considering more structure information on information retrieval

In the first part of the work (Chapter 3) we have seen the importance of links for the automatic creation of descriptions in order to incorporate the multimedia retrieval of information. However one aspect of the implementation considered for simplicity is the use of only the link selection anchor, instead of further context surrounding the link in the text file, for the automatic construction of the descriptor. Further context might give better results when valuable information is available. It could be possible to retrieve non-text documents from text queries in a better way.

In this model, in addition to the link anchor selection, further context where the link is inserted, such as in a sentence, or a paragraph was considered for this reason. Possibly, the use of theme segment could also be considered. This could be constructed by checking the similarity between the paragraph where the anchor is present and the surrounding paragraphs, in a way that is simpler than the one implemented by Salton et al (1996). In the three-tiered model we would only had to add a new content translation layer for this algorithm

A limitation in the first part of the work (Chapter 3) was the fact that it was only possible to retrieve non-text files given a text selection. Now the model has been improved to allow the user to find text or non-text documents, or parts of documents, from other text or non-text documents. With the availability of an indexing reference for a non-text file, it is possible to find matches the other way around. When in a non-text document, by using the indexed description text as the source for a query, it is possible to retrieve any multimedia documents, text or non-text, by matching that descriptor within the indexed tree of the information tool.

Another approach to give more structured information to the content retrieval is possible by allowing the indexing of the context of the link in the indexing tree of the information-retrieval tool. When a link is created it should be noticed that it is usually connecting two smaller units of information that belong to a larger information unit than the full document where they are inserted. So the consideration of the context of links in indexing for retrieval purposes allows access to more specific information.

- Possible Indexed Information
  - Content of text documents
  - Context of text documents
  - Context of non-text documents
  - Anchor context of links
  - Specific author or user atomic selections

Table 4-1 - Possible information to be used in the indexing

User and author specific selection for indexing is also considered in the integration. It might be interesting to have the facility of accessibility to smaller units of information inside text files that otherwise could became lost in the full text document, and to allow users to reflect their information retrieval needs in the structure of information retrieval. This resembles in a way the Microcosm model, where there are application/author linkbases and user linkbases.

Finally it is possible to have different information in the pre-indexed files (see Table 4-1). This information can be indexed altogether or not.

### 4.4.2 Considering more content information on hypermedia

One of the major advantages in Microcosm is the usability of the generic link as a powerful way of creating links and as a more content oriented way of following links. Its way of working is in a way related to information retrieval. It is stored in a database, but to find it a boolean match is processed, an algorithm that was one of the first to be used in information-retrieval systems for its simplicity. It was been proved to be an algorithm of high precision (Salton 1989), although its recall is low. For example, if a generic link is created that has as an anchor such as the word *car*, then for the whole application that link can be followed when that word is selected. However if the word *automobile* is selected instead, the link cannot be followed in the same way. To improve the concept of generic links, more information can then be used for link following. To overcome this problem, a manually created thesaurus for expansion of the anchor selection (Lewis et al. 1996b) has been suggested and actually implemented with further features (Tansley 2000). However a thesaurus implementation might be difficult, time consuming and application dependent. As was pointed out in the previous section, the text surrounding a link is also important to describe the context in which the link is present. If the generic link could also be found in any text document by similarity matching of its context, then even in a sentence where the word *car* is replaced by the word *automobile* the identification of the generic link could be possible. In sum, the context of the link (e.g. sentence, paragraph or theme/segment) will allow its identification even when there is no exact match. So along with the exact boolean matching already available in Microcosm the model allows for similarity matching with a threshold that is performed by the information-retrieval side of the integration. For this purpose the indexing of context information surrounding the link, such as link anchors and link description with a new system translation layer is possible. We would be constructing a new information retrieval index just for finding links.

However an improved way of showing the availability of the links is possible. In Microcosm at present, if a user finds a word of interest for further reading, he/she will select that word to check for a possible generic link. If no link is found then the user will try to select the sentence where the word is present in order to execute a computed link. In a common Microcosm information seeking we have two distinct phases of user interaction. It makes sense then to integrate the two steps.

We give the user the facility of having possible generic link(s) to be followed shown at the same time as a rank of relevant documents related to the selection. The indexing of link context along with documents and other atomic information has been done, the two searching phases are then integrated in one. Information retrieval and navigation will then tend to merge. The new user action is not "show me links about this" or "show me similar documents" but "show me information".

## 4.5 Analysis of a better integration at the abstract level

In this part of the work we will describe a better way of integrating information retrieval with hypertext systems when considering structure and content information from the abstract level, that can allow a smooth transition between retrieval and navigation whatever the media considered.

Here at first we will suggest different alternatives to the construction of an improved abstract level, and the automatic classification of the text documents. Then an algorithm is suggested for the automatic classification of non-text media.

Secondly we show an improved way of using abstract level information for the automatic construction of text descriptors.

### 4.5.1 From the document level to the abstract level

This part of the model has not been implemented but only designed. We left here the description of this side of the integration to allow a better understanding of the interaction of the abstract level with the document level.

In many Microcosm applications the abstract level does not reflect the structure of the hypermedia network or the content of the documents. Often, only a separation of the documents under clusters of the same media type is adopted. With this occurring in testing the first implemented model, we observed that often the abstract information was invalid for the automatic construction of text descriptors and, for the same reasons, the use of that abstract level information was at times a poor alternative to access relevant information.

The failure of the present model of Microcosm in providing a good management tool for the processing of information at the abstract level may discourage authors from creating and maintaining applications in which information at that level is relevant.

To help the author to build a better abstract level we should then provide him/her with an automatic or semi-automatic way of constructing and maintaining that level or perhaps an easy way of importing and maintaining a previously constructed abstract level. Different alternatives are then suggested.

To construct a multimedia classification level is not an easy task whatever the media considered (text or non-text) or technique adopted (manual or automatic). However for the text medium there are simple methods of automatic construction, or as an alternative in many domains, a manually created classification is available for use. Different tools for the construction of the abstract level can then be adopt:

- *Automatic Clustering*

With this tool there would be a fully automatic classification of all text documents under different clusters in a hierarchic tree using one of the algorithms suggested in Rasmussen (1992). The author would have of course the facility of changing the classification of the created hierarchy.

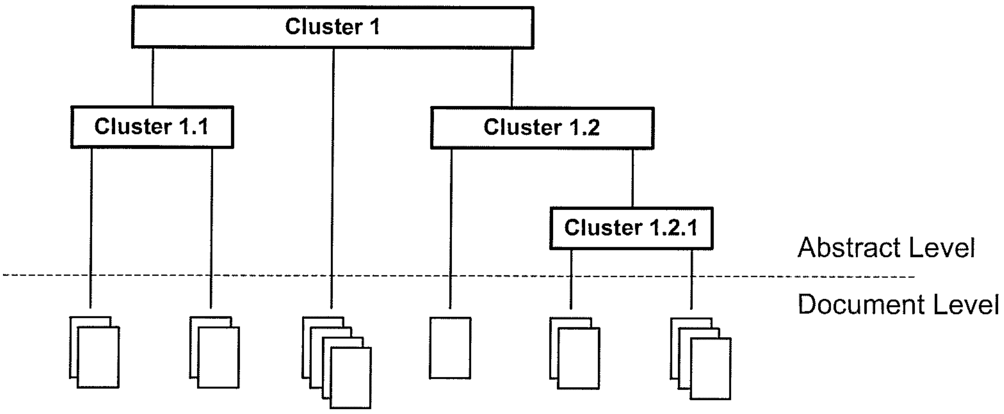


Figure 4-7 – An example of a possible hierarchical cluster based classifications at abstract level

- *Semi-automatic Clustering*

This tool is an extension of the previous one. Here the author could define a skeleton of a proposed classification hierarchy and then the tool would be responsible for the automatic classification of the text documents under that hierarchy that would be automatically expanded in a way that would allow a correct distinction between the incorporated documents.

If the tool would be evoked after the implementation of the link network structure at the document level then some algorithms could be used to find central documents of navigation such as the one suggested by Salton et al (1994), when he found that segments with a large number of output links would indicate important segments for summary construction. This has also been studied more recently by Chakrabarti et al. (1999). The direction of the link may also give a clue about the parent-child hierarchy relationships of documents.

• *Manually Created Thesaurus -*

In many specialised domains a semantically connected thesaurus of concepts is already available for use. The classification of documents could be done manually or an associative information-retrieval algorithm (Edmundson et al 1961, Agosti et al 1992) could be used to automatically index all the text documentation under the thesaurus.

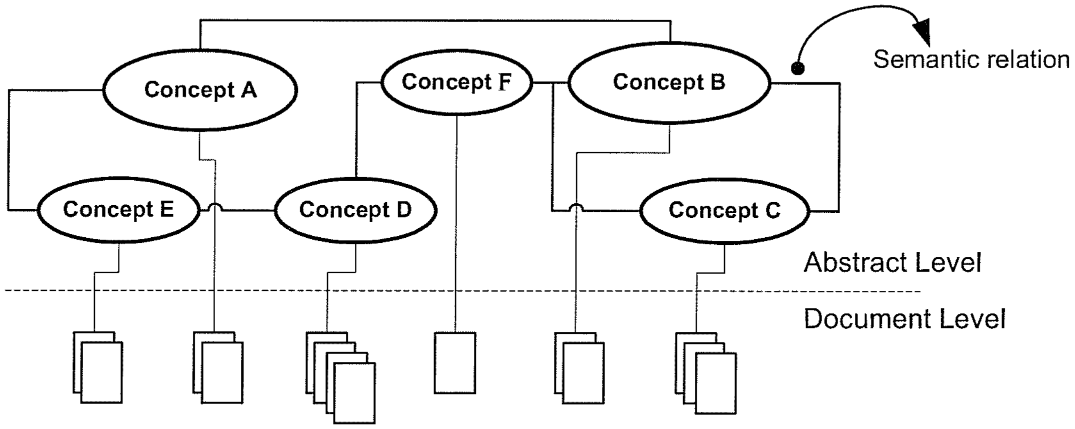


Figure 4-8 – An example of a possible semantically connected thesaurus at abstract level

• *Automatic Created Thesaurus*

With this tool, the semantic dependency of the words in the document collection could be used for the automatic construction of a semantically connected thesaurus (Foskett 1997, Srinivasan 1992, Edmundson et al 1961, Rijsbergen 1977). Afterwards the classification could be done automatically for all text documents using the same algorithms described in the previous section.

The automatic algorithms based on statistical procedures may have a problem of user/author understandability of the hierarchical network created for the abstract level, a problem that does not appear so deeply in a manually created classification space. However, manual methods require a large effort from the author as opposed to automatic algorithms.

The classification of text documents with a clustering algorithm is simpler than when comparing with a classification using a semantically coupled thesaurus.

However the distinction between different clusters is more blurred than the distinction between different concepts when a thesaurus is used.

To classify a non-text document under any of the previous classification models a manual procedure could be adopted or alternatively some automatic procedures could be implemented if the network of links at the document level were already available. The system would propose to the author the assignment to a certain concept depending on the number and type of links connecting that non-text document to other text documents already assigned under that concept, as outlined in

Figure 4-9, or depending on the minimisation of distances to the files they are linked to.

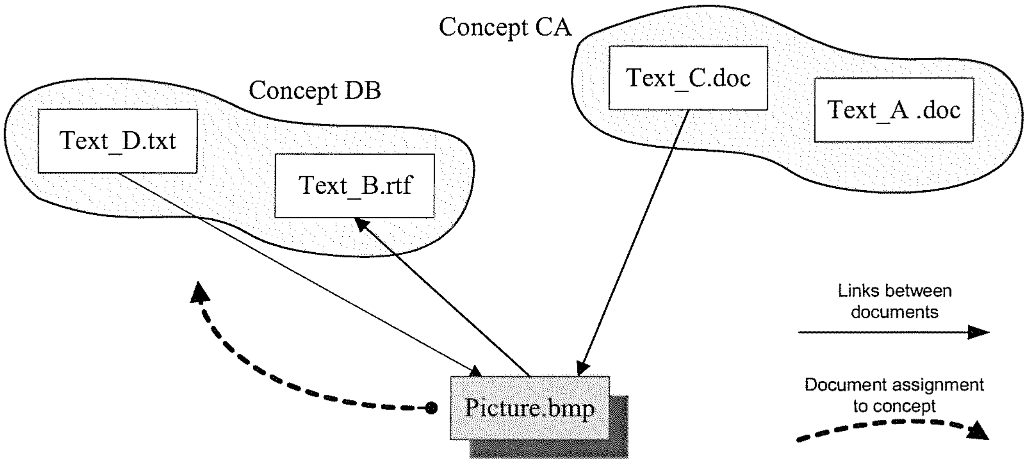


Figure 4-9 - Using link structure information to automatically classify non-text media

But with the available context indexes, a similarity matching could be performed in order to find the most suitable classification concept. However for the previous classification procedures and future alternatives some testing must be performed in order to find the best option for automatic classification of pictures.

### 4.5.2 From the abstract level to the document level

Besides the navigational features related to the abstract level, it is also possible to use its information for the retrieval of multimedia information. In the implemented integration, multimedia retrieval with automatically constructed text descriptors indexes are used that consider, among other things, the information from the abstract level. With its use it is possible to overcome problems that would arise in poorly connected hypermedia applications.

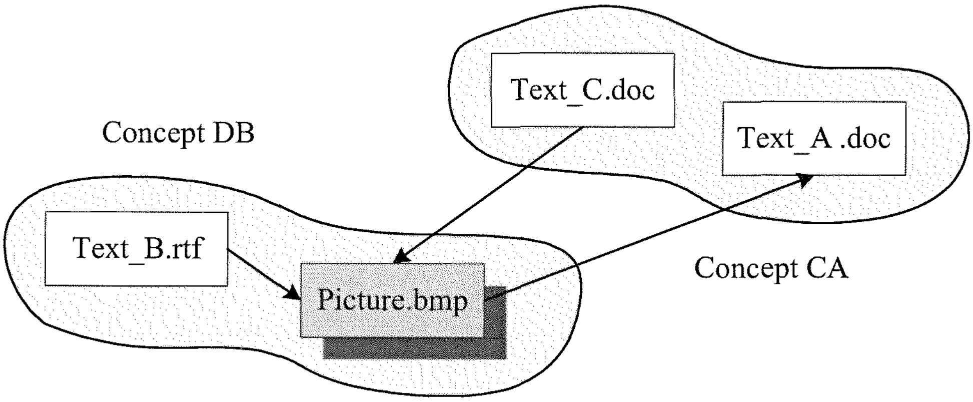


Figure 4-10 - Links connecting documents within the same classification, and in different classifications

Under the new model and improved way of using abstract level information for automatic descriptors index construction is used. The abstract level helps to distinguish between the relevance of different links when connecting non-text documents to text documents (see Figure 4-10). By assigning different weights to meta-information (related to the context of links) according to the location of their anchors we can ascertain whether links that have anchors in documents in the same classification concept / cluster are more relevant for the context descriptor indexing than when these documents are in distinct classifications. This idea is not Microcosm specific and certainly can be generalized to other systems that might include a classification scheme. If not a classification/clustering scheme would have to be appended.

Of course this is done in addition to the use of the classification terms in the context descriptor indexing of multimedia, already implemented in the first model of integration.

The method implemented here is somehow different from the method proposed in

Figure 4-9. Here the abstract level is used for a better context indexing of multimedia whereas in the other method the document level link information is used for a better multimedia classification at abstract level.

## 4.6 Calculation of the descriptor file

For the context indexing of documents, an adequate auxiliary text descriptor is constructed in advance. The approach adopted in this improved model is an expansion of the approach adopted in our first model described in Chapter 3. There we already used information kept at both document and abstract level. This improved the possibilities of better retrieval. Here however for the context indexing more context information is considered, a better use of link types is achieved, and a development in the consideration of the abstract level is used.

Supported by Equation 3-1, and with further research described in this chapter we propose an improved way (Equation 4-4) of calculating text descriptors for multimedia context indexing and retrieval. Equation 4-4 shown below, used to build the text context descriptor, makes it possible to change the weight of all considered metadata parameters. Here there is more independence between the weights for the different parameters. Ideally an improved retrieval quality could be achieved.

$$
\begin{array}{l}
a \cdot N_{Dcr} + b \cdot N_{Cls} + c \cdot N_{Key} + \\
\sum\limits_{l_i \subset L_{Ge}} \left( x_i \cdot \left( \begin{array}{l} d \cdot l_{iSrcAncSel} + e \cdot l_{iSrcAncSen} + f \cdot l_{iSrcAncPar} + \\ g \cdot l_{iDstAncSel} + h \cdot l_{iDstAncSen} + i \cdot l_{iDstAncPar} + \\ j \cdot l_{iDcr} + k \cdot n_{iDcr} + l \cdot n_{iCls} + m \cdot n_{iKey} \end{array} \right) \right) + \\
\sum\limits_{l_i \subset L_{Sp}} \left( z_i \cdot \left( \begin{array}{l} n \cdot l_{iSrcAncSel} + o \cdot l_{iSrcAncSen} + p \cdot l_{iSrcAncPar} + \\ q \cdot l_{iDstAncSel} + r \cdot l_{iDstAncSen} + s \cdot l_{iDstAncPar} + \\ t \cdot l_{iDcr} + u \cdot n_{iDcr} + v \cdot n_{iCls} + w \cdot n_{iKey} \end{array} \right) \right)
\end{array}
$$

Equation 4-4 - Weighting consideration for context indexing of nodes/documents in the second model

Where:

$N$ ____ Node to be indexed by context

$Dcr$ ___ Text description of the node ($N$ or $n_i$) or link ($l_i$)

$Cls$ ____ Classification terms of the node ($N$ or $n_i$)

$Key$ ___ Keywords of the node ($N$ or $n_i$)

$l_i$ _____ A particular link with an anchor in the context indexed node

$L_{Ge}$ ___ Set of generic links with an anchor in the context indexed node

$L_{Sp}$ ___ Set of specific links with an anchor in the context indexed node

$n_i$_____ Neighbour node connected through link $l_i$

$AncSel$ _ Text from the anchor selection of a particular link end ($l_{iSrc}$ or $l_{iDst}$)

$AncSen$ _ Text from the anchor sentence of a particular link end ($l_{iSrc}$ or $l_{iDst}$)

$AncPar$ _ Text from the anchor paragraph of a particular link end ($l_{iSrc}$ or $l_{iDst}$)

$l_{iSrc}$ __ The source anchor of link $l_i$

$l_{iDst}$ __ The destination anchor of link $l_i$

a...y___ Integer weights for different available meta-information

$x_i$_____ Integer weight dependent on the location of the generic link anchor

$z_i$_____ Integer weight dependent on the location of the specific link anchor

Much of the meta-information available from the hypermedia structure in a Microcosm application is author or user dependent and the development behaviour must be replicated in the context indexing. All these different weights allow for the adequate consideration of metadata in different applications by having different weighting strategies.

Besides the different weighting depending on different link types, that has already been justified, there is also the possibility of changing the relative weighting between metadata coming from the document description, the document classification (logical type index), the link description and the link anchors.

Almost all the weights in Equation 4-4 are additive. By this we mean that they represent the number of times each meta-information is repeated in the text file descriptor. However the weight for distinguishing links that connect documents in the same classification from ones that are not, is multiplicative. In practice, all the other weights for metadata parameters containing text, which are related to links, are actually multiplied by this weight for distinguishing links classified according to the location of the documents they connect.

In the first model, the document description and classification of neighbour documents for generic links was not considered. In the second model we use it because some preliminary tests have shown that sometimes in generic links there is a reference to the original source documents.

## 4.7 Conclusions

This chapter has presented an improved design and implementation of a better integration of hypermedia and information retrieval. In addition, in the implementation more structure has been considered for information retrieval and more content information for the hypermedia. A smoother transition between retrieval and navigation, whatever media is considered has been achieved. In the design of the model suggestions for a better abstract level have been made, although this has been the only part of the model not implemented. Nevertheless the abstract level has actually been considered and implemented in an improved way in the context indexing of documents.

A new open information retrieval tool has also been implemented. The way in which this integration has been developed makes it possible to append new information retrieval features or use new browsing metadata when they becomes available.

After the implementation, it is necessary to run some evaluation tests in order to check some of the hypotheses suggested in this thesis. For this reason proper Microcosm hypermedia applications must be chosen or built for the evaluation along with a set of queries and their related documents that previously had to be found manually. This is not an easy task and generally accepted document collections should be used with associated sets of queries. However, the majority of evaluations undertaken previously only considered information retrieval, few considered the integration of information retrieval and hypermedia.

There are some suggestions arising from previous evaluation work developed by Dunlop et al (1993). However, our approach is different and is described in the next. There we lay the foundations for the evaluation to be performed. Firstly, some evaluations with standard test collections are undertaken on the text information retrieval capabilities of the integrated tool. Following this some screen shots of the features available to users are given. Then a method for hypermedia context information evaluation is introduced. The goal of this evaluation is to achieve an accurate construction of context text descriptor indexes for multimedia access.

We expect that for different applications, different weights for the hypermedia meta-information must be used for the adequate context indexing of multimedia. Our evaluations give emphasis to the best weighting strategy to build such context descriptors to access non-text media in different applications. We also present an evaluation concerning the ability to retrieve text from multimedia using text descriptors.

# Chapter 5 - Laying the foundations for evaluation

## 5.1 Introduction

A new design and implementation of an integration of hypermedia with information retrieval is proposed in this thesis. In order to test our hypotheses we need to test the implementation of the integrated system,. This chapter introduces and discusses some first steps towards evaluation.

One of the novelties of our integrated system was the development and use of a new open and independent information retrieval tool. For this reason, some initial evaluations of standard test collections were made on the information retrieval capabilities of this tool. The results are shown in section 5.2

In section 5.3 of this chapter the implementation of the integrated system is introduced. All the features described in Chapter 4 that were actually implemented are presented in this section. The user interface is also outlined with examples, from an illustrative application.

The majority of the tests on the integrated system are concerned with access to multimedia information. Section 5.4 discusses the methodology of evaluation for multimedia access. We need to understand how hypermedia context information can be used to support the construction of context text descriptors for multimedia access.

Some previous work on multimedia access has already used link information (Dunlop et al. 1993, Frankel 1996, Smith 1997, Harmandas et al. 1997, Amato et al. 1998, Mukherjea et al. 1999, Srihari et al. 1999) but our approach goes beyond these by using a richer set of links in an improved way and by using abstract information. Our hypothesis is that for different applications the hypermedia information must be used in different ways for context descriptor building. Section 5.4 therefore focuses on the best strategy for building context descriptors to access non-text media in different applications.

## 5.2 Evaluating the information retrieval tool

First, the implementation of the information retrieval tool must be evaluated. To do this we have to choose one or more known test collections. This allows us to compare the performance of our information retrieval tool with the performance of other systems. There are some similarities between our information retrieval algorithms, and the ones used by Li (1993). So using the same test collections, the performance for plain text retrieval should be similar.

There are some widely used test collections (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/), which are available with Salton's SMART system. These collections are not large in comparison with the ones used nowadays in the Text Retrieval Conference (TREC), but they are more valid for small applications such as the ones created in Microcosm.

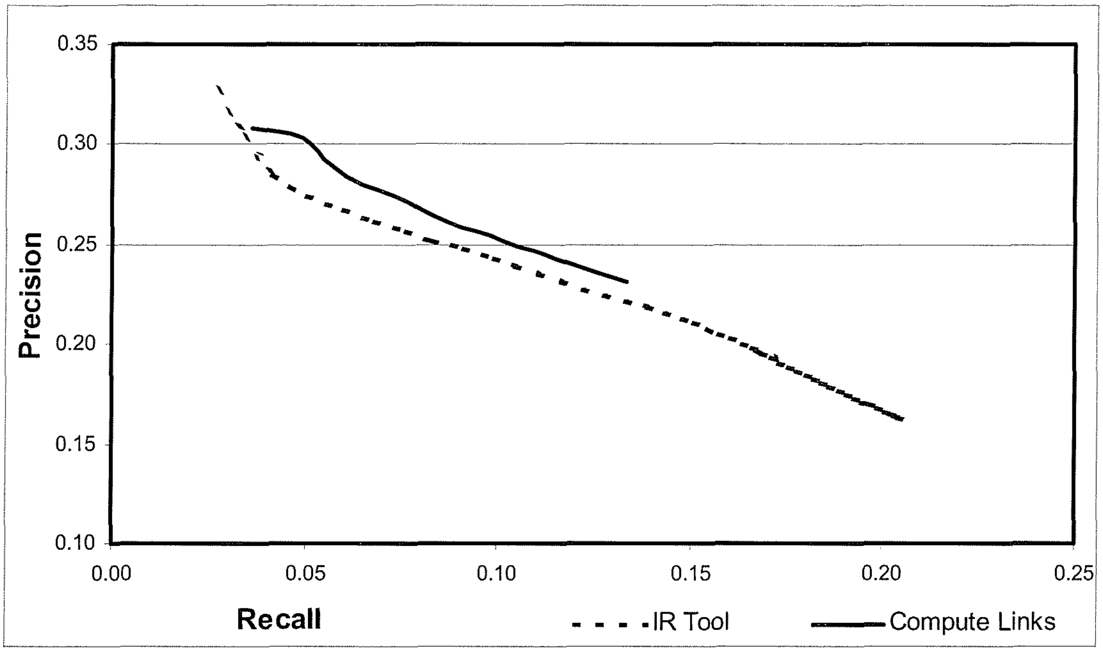


Figure 5-1 - Evaluation of the information retrieval tool with the CACM test collection

CACM is one of those collections. This collection has 3204 documents, 64 queries and 12.4 expected relevant documents per query on average. These queries are in natural language format. For this test, the title, body, journal name and author were indexed. Li (1993) tested his computed links filter with this collection, but the results we obtained with it are different from those stated in his thesis. However, they are similar to the ones achieved with our information retrieval tool (see Figure 5-1). Our hypothesis is that we were using a different version of the CACM test collection. However, the information requests in this test collection base are fairly specific and the average number of relevant documents is small, precision and recall figures tend to be low (Baeza-Yates 1998). Even so, the low figures obtained here that our version of CACM must have some errors, and for this we were unable to exactly reproduce Li's experiment.

The average (recall, precision) for expected relevant documents is (0.1007, 0.2117) with the information retrieval tool and (0.1021, 0.2580) with the computed links filter.

Due to the previous problem we decided to chose another test collection: Cranfield. With this we obtained satisfactory results on recall and precision. This collection has 1400 documents, 225 queries and 8.09 expected relevant documents per query on average. The queries are in natural language format and we indexed only the body of the documents.

With this collection both our information retrieval tool and the computed links filter obtained similar results, with the former performing slightly better. Precision is at a good level, and the recall of documents is higher than on CACM.

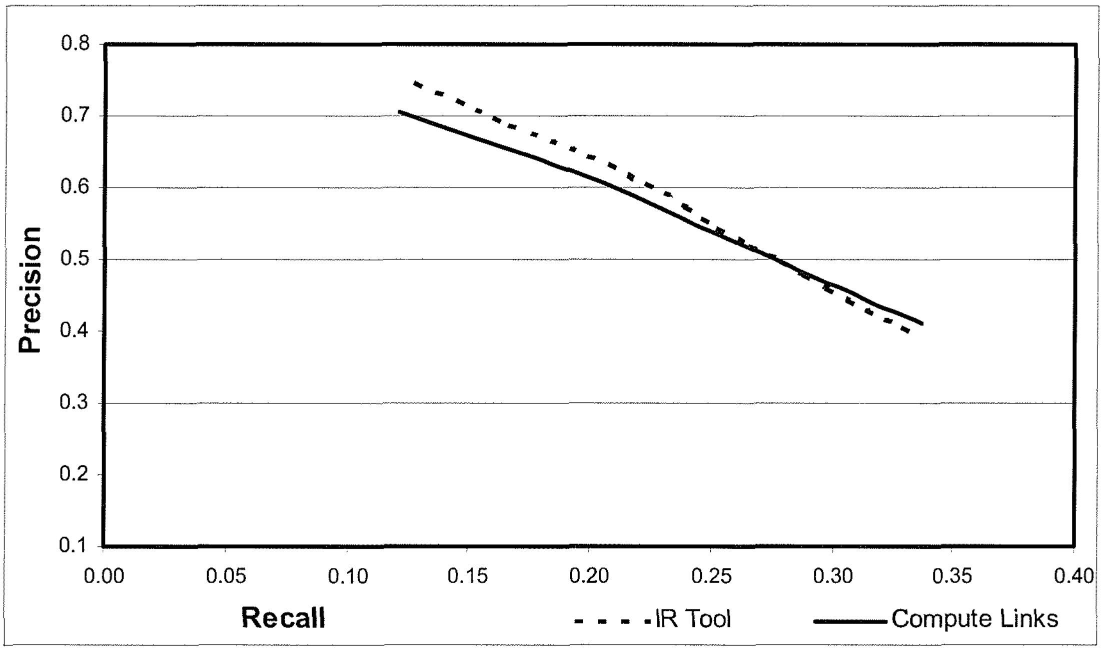


Figure 5-2 - Evaluation of the information retrieval tool with the Cranfield test collection

The average (Recall, Precision) for expected relevant documents is (0.2778, 0.5356) for the information retrieval tool and (0.2745, 0.5265) for the computed links filter. This values show that we have a good text information tool. Therefore we can use it on the multimedia context retrieval evaluations.

## 5.3 Presenting the integrated system

We have designed a new method of integrating information retrieval with hypermedia systems when taking into consideration structure and content information from the information level and the abstract level that can allow a smooth transition between retrieval and navigation, whatever media is considered.

As described in Chapter 4, a new information retrieval tool was developed, the implementation of which is open and independent from any particular system (i.e. the environment where the tool will be run) or file format.

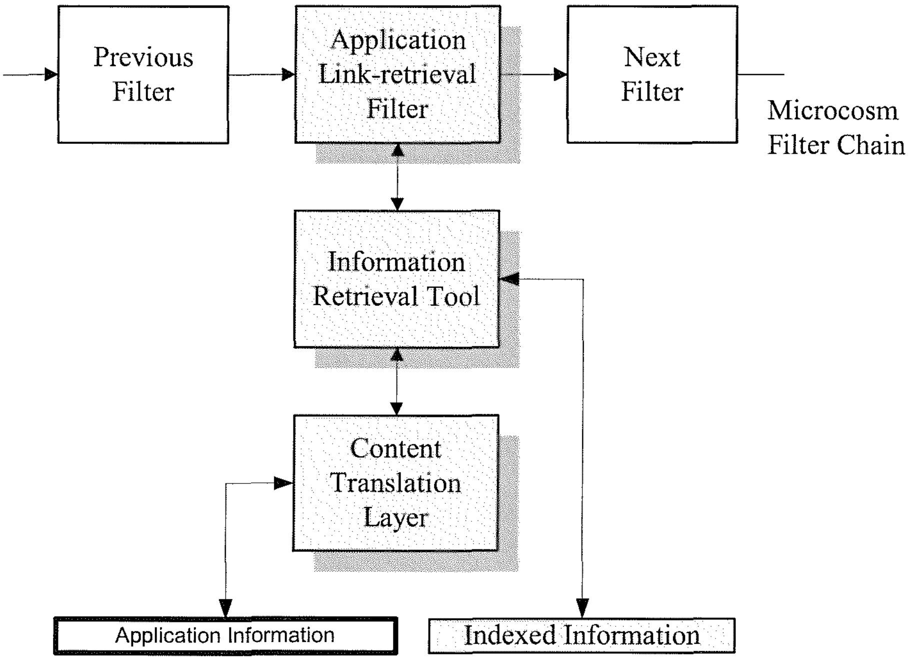


Figure 5-3 - Integration of the new information retrieval tool with Microcosm

To have a more flexible implementation we created three different modules, each one responsible for its specialised task. Figure 5-3 shows the integration of this tool with the Microcosm system. The information retrieval tool is responsible for core information retrieval management. It provides a set of APIs to be used by any system, and a system like Microcosm can use it within a filter. The resulting information retrieval filter is responsible for the management of information received from Microcosm and for using the received information in such a way that it can be used by the information retrieval tool and by Microcosm as shown in Figure 5-4. Finally, the content translation layer is more concerned with the information format of documents and links. It will retrieve raw data to be used by the information retrieval tool. It provides a set of APIs to read content from files, link context, etc. These APIs are loaded dynamically for all legacy formats, such as in Microcosm.

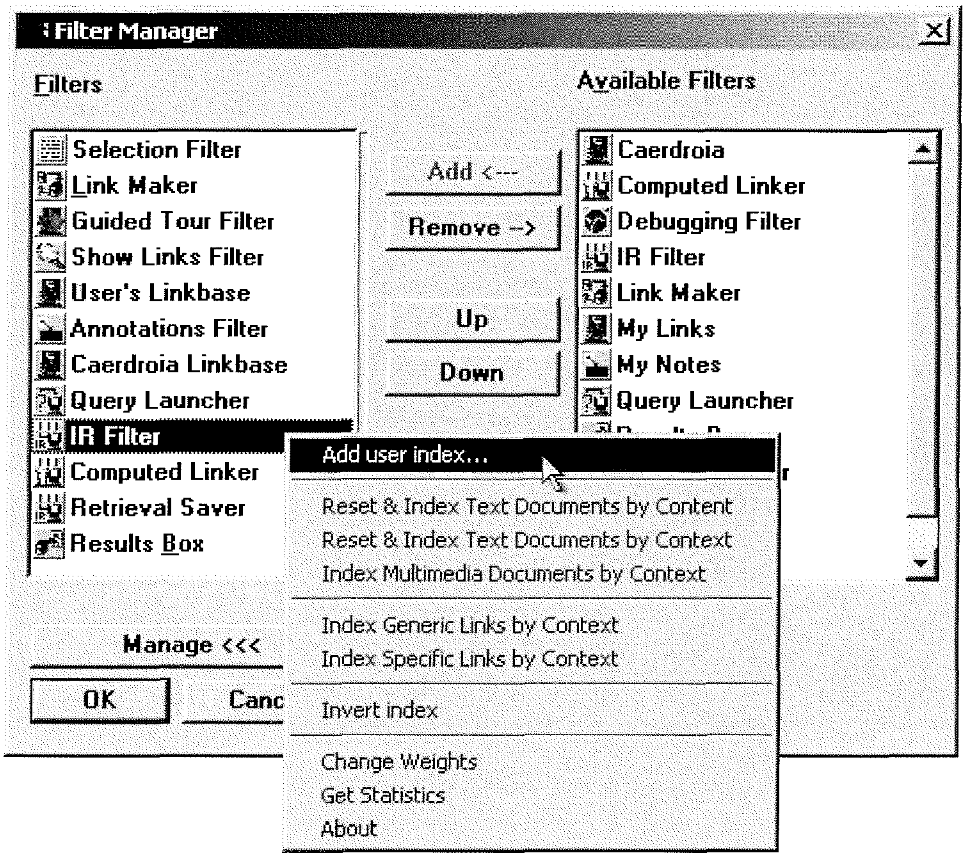
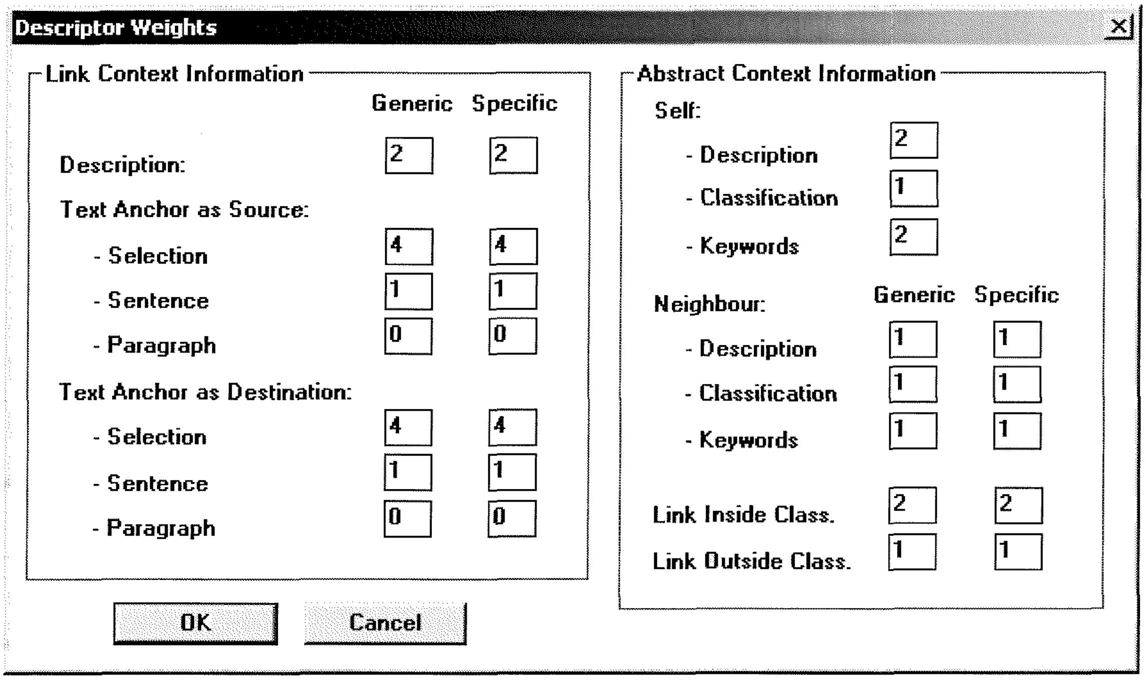


Figure 5-4 - Information retrieval tool integration and its interface in Microcosm

The model of integration allows, together with author-indexed information, the easy use of user-indexed information that can be changed according to a user's information needs. This is similar in a way to the Microcosm model, where there are application/author linkbases (i.e. link databases) and user linkbases. With this facility, the user has the ability to specify his/her own topics of interest for future retrieval and to have the links created by him/her reflected in the information retrieval component of integration. The two different indexed trees are stored but only one is used at run time. This indexed tree is created by merging the previous two.

Figure 5-4 shows Microcosm filter chain, with the list on the left showing the filters in use and the list on the right showing the available filters. Here some filters were inserted for testing and the Application Link-retrieval Filter is named here as 'IR Filter'. On this dialog box the user change the behaviour of the integration.



Descriptor Weights

Link Context Information

| | Generic | Specific |
|---|---|---|
| Description: | 2 | 2 |
| Text Anchor as Source: | | |
| - Selection | 4 | 4 |
| - Sentence | 1 | 1 |
| - Paragraph | 0 | 0 |
| Text Anchor as Destination: | | |
| - Selection | 4 | 4 |
| - Sentence | 1 | 1 |
| - Paragraph | 0 | 0 |

Abstract Context Information

| Self: | |
|---|---|
| - Description | 2 |
| - Classification | 1 |
| - Keywords | 2 |

| Neighbour: | Generic | Specific |
|---|---|---|
| - Description | 1 | 1 |
| - Classification | 1 | 1 |
| - Keywords | 1 | 1 |
| Link Inside Class. | 2 | 2 |
| Link Outside Class. | 1 | 1 |

Figure 5-5 - Dialog Box where we select the weights for the different elements

Along with other options available as part of the information retrieval integration, the user can index text documents in two different ways as shown in Figure 5-4. This can be done by content and/or by context. When we are interested in the information retrieval of any media type, text or non-text, we have to take account of the fact that there are differences in these two types of indexing. Qualitatively, content indexing is understandably better than context indexing, but if the same kind of retrieval quality on any file type is desired, we should also index text documents by context. However there is no such requirement if we are only concerned with the correct ranking of the retrieved documents, because of the different size of information available. All document indexes are normalized and for this reason text files can be indexed by content, and other types by context, without having any bias in the results.

Of course, an option for the indexing of non-text multimedia by context is also available as shown in Figure 5-4. This context indexing, for text and non-text media, is done in accordance with the designed model, where it is possible to attribute different weights to different metadata elements from the hypermedia network as shown in Figure 5-5.

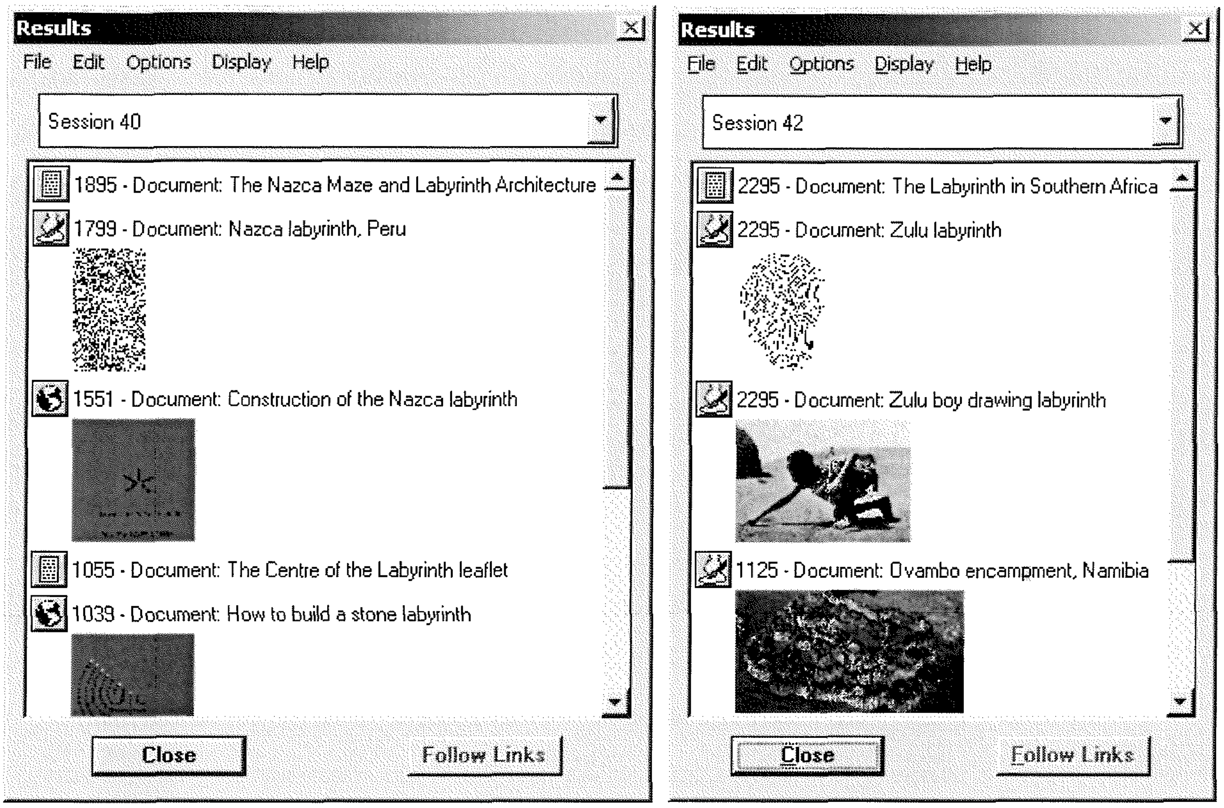
Zero is the default weight for all metadata and for this reason the user/author must input different values. Figure 5-5 is an example of different metadata weights.

The model of the integration presented here considers an increased number of metadata elements when compared to the first implementation. There is an extended consideration of link anchors, i.e. it is taken into account further context where the link anchor is inserted. Apart from anchor selections, sentence and paragraph are also used. In addition, there is a distinction between information coming from the different anchor ends of links. The use of subject classification at abstract level, removed from the tree-like classification of documents in Microcosm, has also been refined.

Link description, document description and keywords are used in many hypermedia systems. Therefore, we use this kind of information for information retrieval. In addition, besides the keywords and the description of the document itself the keywords and descriptions of its neighbours are also used, with a weight to be selected.

A limitation regarding insufficiency of links for context indexing is overcome in our system in a new way, by considering the classification of documents. It is then possible to determine a weight for the words contained under this metadata, and also a weight for the links connecting documents inside ('Link Inside Class.' field in Figure 5-5) and outside ('Link Outside Class.' field in Figure 5-5) the same classification. This last weight is multiplicative, i.e. it multiplies the influence of the weights for the other link metadata words.

Finally, for all information dependent on links, a distinction is made between generic and specific link types.


Results
File Edit Options Display Help
Close
Follow Links
Results
File Edit Options Display Help
Close
Follow Links

Session 40

1895 - Document: The Nazca Maze and Labyrinth Architecture

1799 - Document: Nazca labyrinth, Peru

1551 - Document: Construction of the Nazca labyrinth

1055 - Document: The Centre of the Labyrinth leaflet

1039 - Document: How to build a stone labyrinth

Session 42

2295 - Document: The Labyrinth in Southern Africa

2295 - Document: Zulu labyrinth

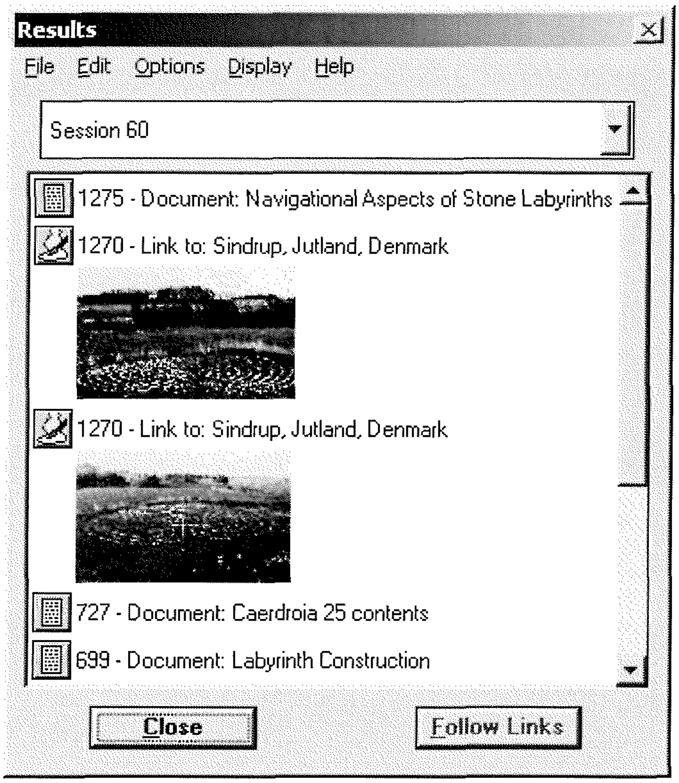
2295 - Document: Zulu boy drawing labyrinth

1125 - Document: Ovambo encampment, Namibia

Figure 5-6 - Retrieval of text and non-text documents after a string search for information

An illustration of this multimedia context indexing can be seen in Figure 5-6, where the selections "labyrinth architecture" and "Zulu" were used for searching. It is observed that we can have mixed retrieval of text and non-text documents. The description of all documents is shown, and for non-text, a small thumbnail is present.

A past limitation was the fact that it was only possible to retrieve non-text files given a text selection. The model was improved by allowing the user to find text or non-text documents, or parts of them, from text or non-text documents. In a non-text document, by using the text descriptor as the source for a query, it is possible to retrieve any multimedia documents, text or non-text, by matching that descriptor within the indexed tree of the information tool.

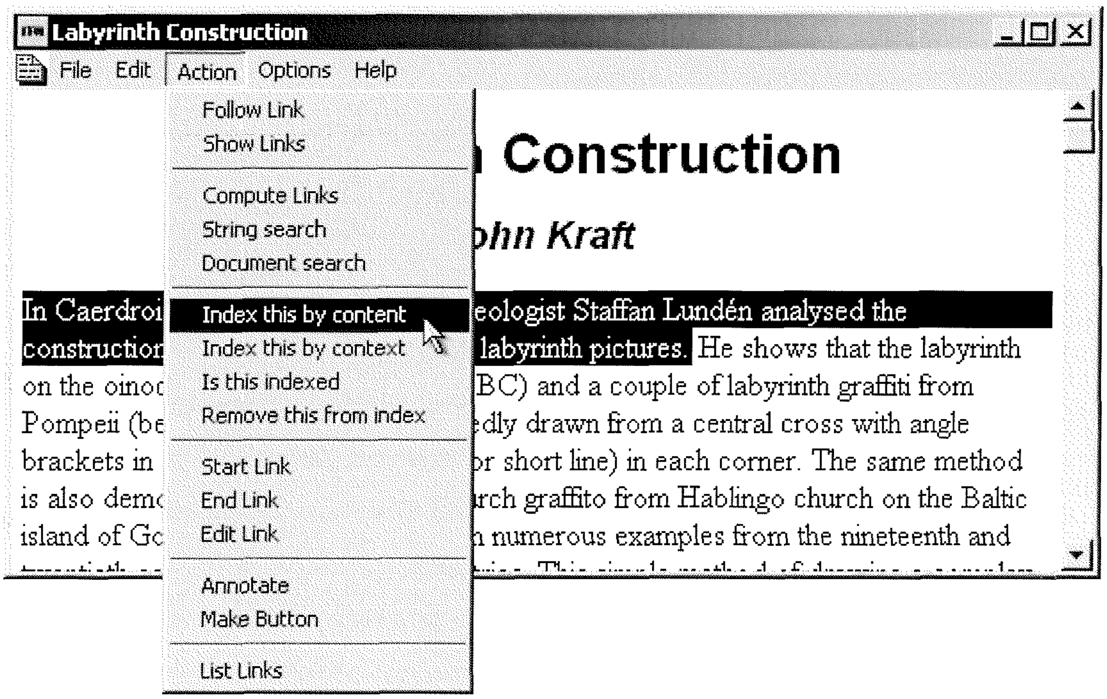


1275 - Document: Navigational Aspects of Stone Labyrinths
1270 - Link to: Sindrup, Jutland, Denmark
1270 - Link to: Sindrup, Jutland, Denmark
727 - Document: Caerdroia 25 contents
699 - Document: Labyrinth Construction

Figure 5-7 - Retrieval of documents and links after a string search for information

Another approach to give more structured information for content retrieval was to allow the indexing of the context of the links in the indexing tree of the information retrieval tool. This option is available in our integration separately for generic and for specific links as shown in Figure 5-4. The consideration of the context of a link in indexing for retrieval purposes allows us to have access to more specific information. In addition, this approach is similar to the merging of navigation and retrieval in one step. By merging the generic link anchors in the main index, the user has the availability of possible generic links shown along with the ranking of retrieved documents. The two searching phases are integrated into one, which can lead to a smoother transition between navigation and retrieval. An example of this strategy is shown in Figure 5-7, where the user was seeking for information on "stone labyrinths". The user is presented with a ranking of relevant information, which may be documents ('Document:' in the figure), of which a small document description is shown, or links ('Link:' in the figure), whose selection or link description is displayed instead, along with a small thumbnail if the destination is an image. This is an extension to the classic approach where only documents are retrieved, as presented in Figure 5-6.

A user/author-specific selection for indexing was also implemented. It makes sense to have the facility of accessibility to smaller units of information inside text files, which otherwise could become lost for the user in the full text document. The user, while traversing through information, can select pieces of text that he/she finds interesting, and ask for it to be indexed. Figure 5-8 shows how this is done inside a particular document.



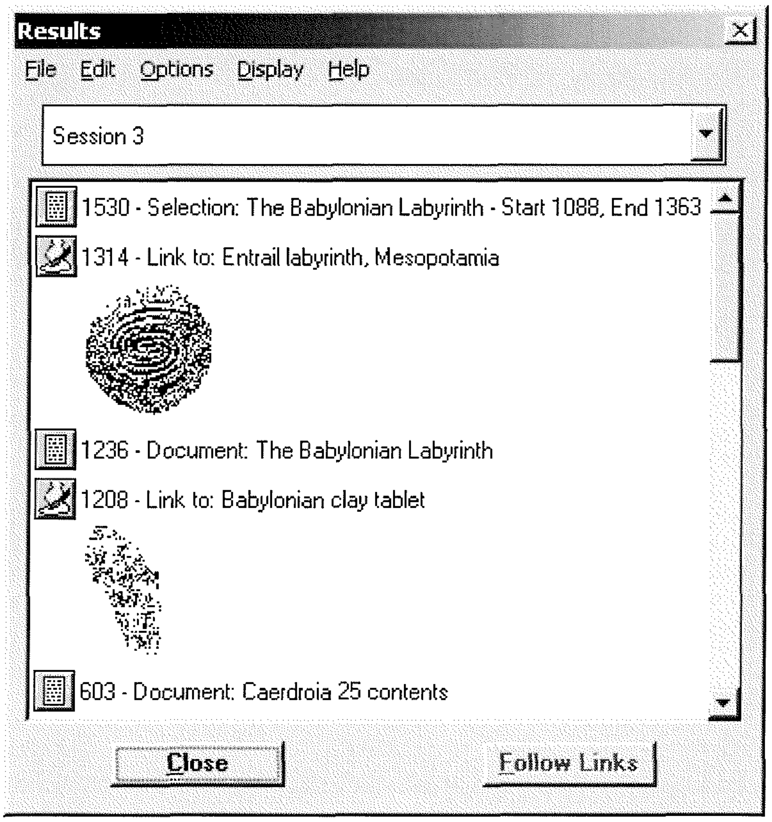
Construction

hn Kraft

In Caerdroi eologist Staffan Lundén analysed the
constructio labyrinth pictures. He shows that the labyrinth
on the oinoc BC) and a couple of labyrinth graffiti from
Pompeii (be dly drawn from a central cross with angle
brackets in r short line) in each corner. The same method
is also dem rch graffito from Hablingo church on the Baltic
island of Gc numerous examples from the nineteenth and

Figure 5-8 - Menu for accessing information retrieval and browsing facilities

The user, besides being able to index selections in documents, can also remove references of these selections from the index, or find whether a particular selection is already indexed. We should also take into account that this integration permits the easy management of single document indexing. If no selection is present, the functionality changes, and then the entire document can be indexed by context and/or content, removed from the index or even queried about their type of indexing. The information retrieval tool was implemented in such a way as to make re-indexing of all documents unnecessary when we just want to add or remove a reference. Instead the index of that particular document, or other information unit, is inserted or removed only.



1530 - Selection: The Babylonian Labyrinth - Start 1088, End 1363

1314 - Link to: Entrail labyrinth, Mesopotamia

1236 - Document: The Babylonian Labyrinth

1208 - Link to: Babylonian clay tablet

603 - Document: Caerdroia 25 contents

Figure 5-9 - Retrieval of documents, links and user selections after a string search for information

Figure 5-9 shows an example of this user-specific selection retrieval, while searching for "Babylonian labyrinth". The user selection is displayed in the ranking with a document description of where it is inserted, along with a start and an end offset. In the Caerdroia application some authored selections were indexed, and the one in particular retrieved on this example presented a high frequency for the word "Babylonian". We are then retrieving more specific information than when considering the whole document as one. In the example, the whole document is ranked in third position and is then less relevant than the user selection.

## 5.4 Evaluating the integration for multimedia access

To assess how the developed integration behaves for multimedia access we need to do some evaluation with proper test collections. Each of these test collections consists of a proper hypermedia application, a set of queries and their relevant associated documents.

Unfortunately, no widely accepted multimedia test collection is available which is as rich as Microcosm applications in hypermedia information, with assessment of retrieval and recall from given queries. Due to the absence of standard test collections, there is a need to find proper applications and build adequate test collections with these applications.

Some test collections were then partly created in Microcosm, which allow us to make some conclusions. Standard Microcosm applications were used together with a set of more or less random queries related to the subject of each of the applications.

These applications have a reasonable number of text files and an adequate number of links connecting those documents. Among other aspects, the study of the availability of link information will be discussed in the next chapter for the chosen applications, namely Archaeology, Cell Biology, French, the French Revolution and Tulip.

Text files for evaluation are of particular interest because they provide a means of comparing the content of documents with their context and of seeing to what extent they are related. To do that with non-text media would be rather difficult, since the hypermedia information surrounding the documents is rather textual. Therefore, by using text files we can have a better idea about the relations between the content of documents and the surrounding hypermedia information for multimedia access.

On the chosen applications, text files were indexed by content using the previously developed information retrieval tool. We then launched queries and saved the first fifteen retrieved documents for each of the applications. The queries were constructed by selecting text included in documents and asking for similar information, very much in a style that is common in Microcosm. The query text for each of the applications is included in Appendix B - Application testing selection queries.

Afterwards, these same text files were indexed by context, and then recall and precision were inferred by comparing the retrieval of these text documents indexed by context descriptors, with the expected ideal retrieval, and with a random retrieval. Ideally, the retrieval when using context descriptors, using the same queries, should be the same as the content retrieval.

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 3 | 2 |
| | Text Anchor as Source: | Selection | 3 | 2 |
| | | Sentence | 1 | 0 |
| | | Paragraph | 1 | 0 |
| | Text Anchor as Destination: | Selection | 3 | 2 |
| | | Sentence | 1 | 0 |
| | | Paragraph | 1 | 0 |
| Abstract information context | Self | Description | 3 | |
| | | Classification | 1 | |
| | | Keywords | 2 | |
| | Neighbour | Description | 2 | 1 |
| | | Classification | 1 | 1 |
| | | Keywords | 2 | 1 |
| | Link Inside Classification | | 1 | 1 |
| | Link Outside Classification | | 1 | 1 |

Table 5-1- Weight values for context information in the Caerdroia application example

The context descriptors can be built in many diverse ways since we considered different context information at different levels. For this reason a decision has to be made on the best way of considering all the information. The way to balance this information is to assign different weight values as described in Chapter 3 and Chapter 4.

As an example, and to test this approach, a simple test was made with a smaller illustrative application – Caerdroia. In this application there are 30 text documents, and we randomly placed 10 queries and saved 15 retrievals for each of those queries. We then assigned apparently reasonable weighting values (Table 5-1) for the different elements of context information on the hypermedia network.

The queries were placed, the retrieval results saved and the recall, precision and recall/precision graphics were plotted for the context indexed text files:

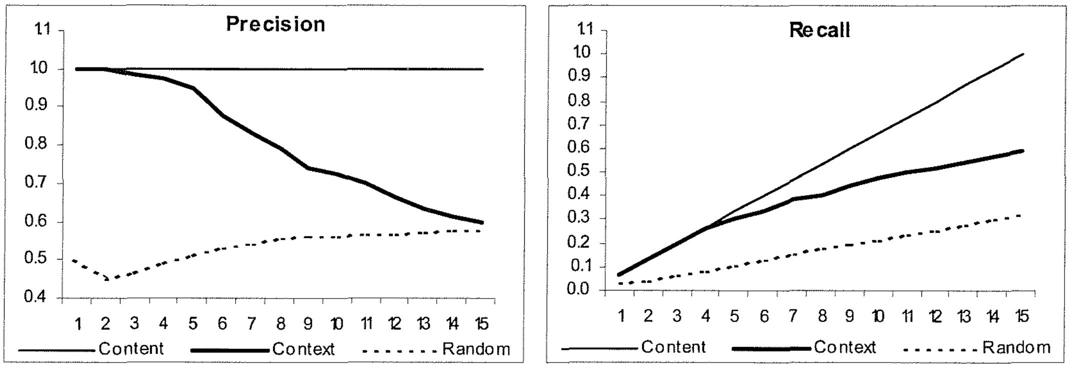


Figure 5-10 - Multimedia context retrieval precision and Recall for the Caerdroia application

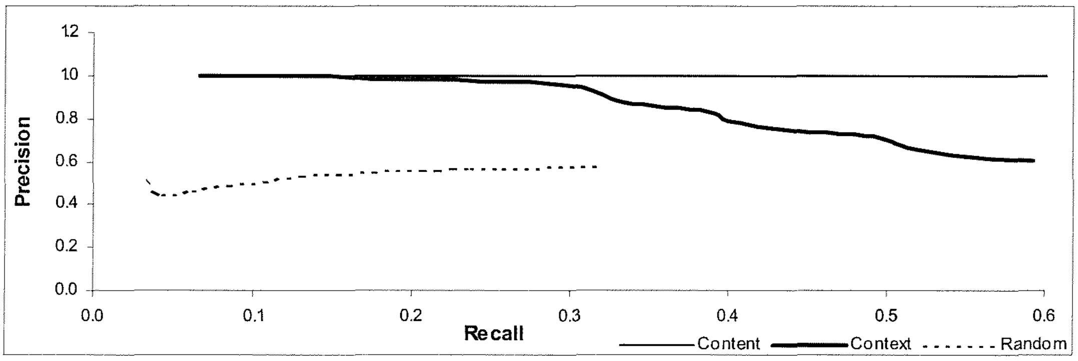


Figure 5-11 – Multimedia context retrieval Recall vs. Precision for the Caerdroia application example

These graphs show that context description is somewhere between the ideal content retrieval and random retrieval. The distance to the random retrieval is however not far enough. This is due to the small number of text files – 30 – in comparison with the number of relevant retrieved files for each query – 15. Such a small test collection does not give convincing conclusions, but gave us however the promise of obtaining good results if larger applications were used.

Some evaluation on the relative significance of the context information elements is required in such a way as to bring together context retrieval and content retrieval. This is rather important since we already know that authors, when designing hypermedia applications, use different styles with varying accuracy.

For each application, all the weights for all context elements in the hypermedia network are sequentially changed. Then the best weighting strategy for each particular application is determined.

At this point, we have to decide how much discrepancy there should be between the different weights. There are some limitations about how far we can go on the size of the weighting intervals. A binary variation is possible, which would roughly state the presence or absence of a particular context element. But to satisfy all the weighting combinations of the 27 available elements, 134217728 tests would have to be generated. Considering that each test requires indexing by context all the documents, placing the queries and then analysing the retrieval results, this could be quite time-consuming! All the tests are done automatically without any user intervention. Nevertheless, just as an example, the fastest tests can take about 2 minutes on a Pentium II – 350 MHz, although usually it takes much longer than that. So the experiment would take at least 610 years 263 days 12 hours and 16 minutes to finish. Even if the speed of the test were to increase 100-fold, so that each test took about 1 second, we would still have to wait for more than 2 years before we got the first results. At the same time, such a small variation in weights for each element does not distinguish much between then. So the interval of weight variation would have to be increased to better determine the relative significance of the context information elements for context retrieval. But time performance again becomes a serious issue if a bigger variation interval is considered. If a decimal interval were chosen for each element, instead of the 255 years we would have to wait for the binary option, we would now have to wait for about $8 \times 10^{18}$ centuries. Consequently, compromises have to be reached on our ideal option for testing of all parameters at the same time, with a more reasonable approach to the evaluation.

The combination of all possible weights is an exponential function of the number of elements considered in a certain test, and the amount of variation affected to the weighting. So the number of elements to be tested at the same time has to be limited.

Some of the weights were fixed and the remaining ones sequentially tested for all combinations. In this way a more realistic time scale solution is adopted. It is then necessary to choose a reasonable variation interval for those weights. A large interval selection creates time restrictions, and to overcome this problem only an insufficient number of elements could be tested at one time. Table 5-2 below gives an idea of the number of tests to run and the best-expected time to spend on those tests.

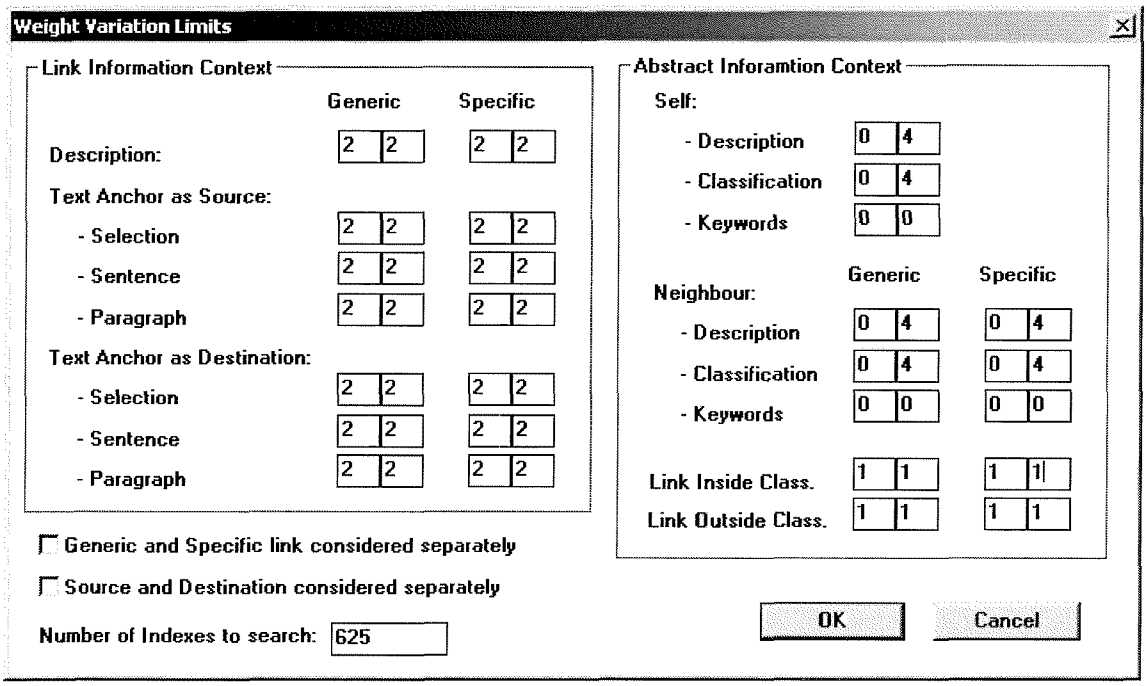
| | | | Weight Interval Size | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2 | 3 | 4 | 5 | 6 | 7 |
| Number of elements to test | 1 | No. tests | 2 | 3 | 4 | 5 | 6 | 7 |
| | | Time spent | 4m | 6m | 8m | 10m | 12m | 14m |
| | 2 | No. tests | 4 | 9 | 16 | 25 | 36 | 49 |
| | | Time spent | 8m | 18m | 32m | 50m | 1h12m | 1h38m |
| | 3 | No. tests | 8 | 27 | 64 | **125** | 216 | 343 |
| | | Time spent | 16m | 54m | 2h08m | **4h10** | 7h12m | 11h26m |
| | 4 | No. tests | 16 | 81 | 256 | **625** | 1296 | 2401 |
| | | Time spent | 32m | 2h42m | 8h32m | **20h50m** | 1d19h12m | 3d08h02m |
| | 5 | No. tests | 32 | 243 | 1024 | 3125 | 7776 | 16807 |
| | | Time spent | 1h04m | 8h06m | 1d10h08m | 4d08h10m | 10d19h12m | 23d08h14m |
| | 6 | No. tests | 64 | 729 | 4096 | 15625 | 46656 | 117649 |
| | | Time spent | 2h08m | 1d18m | 5d16h32m | 10d20h25m | 64d19h12m | 163d09h38m |

Table 5-2 - Number of tests to run and time scale

A balance between the number of elements to test simultaneously and the limits of the weighting interval had to be achieved. Since some weights were to be fixed at a certain value and the remaining ones changed around that value, an odd size interval had to be picked. Among the possible size options of three, five and seven the size interval of five was selected. In this way, tests could be run and the best weighting strategy estimated in around one day four of those context elements, and the remaining ones run for around four hours.

For almost all the tests, the weight of the unchanging elements was usually kept at two (the middle of the selected interval), and the remaining ones were changed between zero and four. To do this a small program was created that would change the weights, index the documents by context and place the queries. This program, Query Launcher, being a filter under the Microcosm system, was placed just before the Information retrieval filter, in the filter chain.

Figure 5-12 shows the dialog box that allowed us to select the variation intervals for the weights. We also created another program inserted just after the information retrieval filter, which would save all the retrieved documents for each test, making it possible to determine afterwards the recall and precision graphs for each of the weighting strategies.



Weight Variation Limits

Link Information Context

| | Generic | | Specific | |
|---|---|---|---|---|
| Description: | 2 | 2 | 2 | 2 |
| Text Anchor as Source: | | | | |
| - Selection | 2 | 2 | 2 | 2 |
| - Sentence | 2 | 2 | 2 | 2 |
| - Paragraph | 2 | 2 | 2 | 2 |
| Text Anchor as Destination: | | | | |
| - Selection | 2 | 2 | 2 | 2 |
| - Sentence | 2 | 2 | 2 | 2 |
| - Paragraph | 2 | 2 | 2 | 2 |

☐ Generic and Specific link considered separately

☐ Source and Destination considered separately

Number of Indexes to search: 625

Abstract Inforamtion Context

| Self: | | | | |
|---|---|---|---|---|
| - Description | 0 | 4 | | |
| - Classification | 0 | 4 | | |
| - Keywords | 0 | 0 | | |
| | Generic | | Specific | |
| Neighbour: | | | | |
| - Description | 0 | 4 | 0 | 4 |
| - Classification | 0 | 4 | 0 | 4 |
| - Keywords | 0 | 0 | 0 | 0 |
| Link Inside Class. | 1 | 1 | 1 | 1 |
| Link Outside Class. | 1 | 1 | 1 | 1 |

Figure 5-12 -Weight Variation Limits Dialog box on Query Launcher Filter

The next question was which weights to change at the same time, given that the hypermedia system being used, i.e. Microcosm, is basically organized at two levels. There is what may be called a document level with the explicit links interconnecting documents, on which the user can traverse through the information. On top of this we have what may be called an abstract level that joins documents with related characteristics. At this level the user has a top down approach for information searching.

For this reason, the tests on context descriptors were divided into two sets. In the first one, basically at document level, we evaluated metadata and context information associated with links between the documents. This evaluation can be found in Chapter 7. On the second one, we evaluated abstract information about the documents themselves and about the relations between links information and abstract information. This evaluation can be found in Chapter 8.

For these evaluations, the weight of the unchangeable elements was fixed to two, unless otherwise stated. Exception is made to the information taken from the knowledge of whether links connect documents in the same classification or not. The fixed weight for both options was set to one for reasons that will be explained later. The keyword weight was also kept to zero since the tested applications were not developed with that aspect in mind.

Between all the possible weight scenarios, we selected the best and the worst ones, and plotted their recall, precision and recall/precision graphs. We also present the obtained weights and the achieved best and worst values of recall, precision and effectiveness. Recall and precision measures have been used extensively to evaluate the performance of retrieval algorithms and we use here the definition presented by Rijsbergen (1979). Effectiveness is this thesis is a harmonic mean measure (Baeza-Yates 1999) of recall and precision. All these measures will take the ideal values of 1 when all relevant documents are retrieved and the value of 0 when no relevant document is retrieved.

The recall, precision and harmonic mean in the tests performed in this thesis are calculated in the following manner:

- **Recall** is defined as the proportion of relevant documents (the set R) that are retrieved, i.e.,

$$Recall = \frac{|Ra|}{|R|}$$

Equation 5-1 - Recall measure

- **Precision** is defined as the proportion of the retrieved documents (the set A) that are relevant i.e.,

$$Precision = \frac{|Ra|}{|A|}$$

Equation 5-2 - Precision measure

- **Harmonic Mean** which gives equal importance to precision and recall is defined here as:

$$HarmonicMean = \frac{2}{\frac{1}{P}+\frac{1}{R}}$$

Equation 5-3 – Harmonic Mean to measure effectiveness

This last measure gives an idea of the normalized size of the intersection between the set of relevant documents and the set of retrieved documents (see Figure 5-13). With this measure we have a way of knowing how effective a system is, and it is related in a way to the E-measure proposed by Rijsbergen (1979). Determination of its maximum value can be interpreted as an attempt to find the best possible compromise between recall and precision (Baeza-Yates 1999).

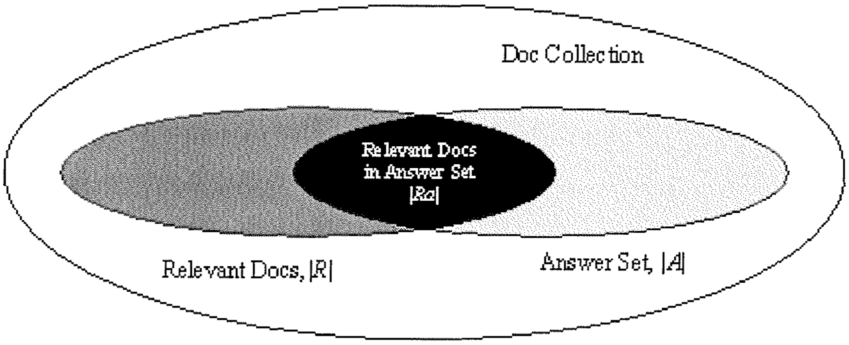


Figure 5-13 - Illustration of the relation between the relevant documents and the retrieved answer set

These precision and recall measures are calculated at each level of retrieval, and an average precision and recall is calculated for all queries. It is then possible to plot a graph of precision and recall versus the number of retrieved documents. The remaining graph is a plot of precision versus recall.

In the obtained graphs we can compare the context retrieval with the ideal retrieval, i.e. the retrieval obtained when the documents are indexed by content, as explained in the beginning of this section. For each of the evaluations we also plotted the curves for the random retrieval of documents.

At the end of each evaluation the obtained results are compared for best weight strategy with the style used by the author in developing the application. This weighting might be different for precision and recall, but effectiveness will give us the best compromise between these two measures.

Chapter 9 resumes an ad-hoc strategy to integrate the results from Chapter 7 and Chapter 8.

## 5.5 Conclusions

We have presented some tests made on the developed information retrieval tool. By comparison with Li's (1993) results, it is shown that a slightly better performance is achieved with implemented text algorithms used in the developed information retrieval tool. This assures us that we can effectively use the developed information retrieval tool to make some further tests on multimedia access capabilities, using context text descriptors.

The implemented integrated system was also introduced. The new way of using the open and independent information retrieval tool allows for several features. User and author indexing, content and context indexing, different granularity of indexed information, and the merging of hypermedia browsing and information retrieval search are permitted. Some examples of these features were discussed.

Due to computational restrictions on computers today, we have seen that there is a need to make multiple separate tests on the multimedia access capabilities. The tests will be performed using diverse metadata in such a way that we can find a weighting strategy for different type of applications. Tests on different metadata will then be expanded in Chapter 7, Chapter 8 and Chapter 9. But to better understand how to properly index multimedia by context we also need to know the style of the concerned applications. In the next chapter a classification of different Microcosm applications is discussed. These applications are then used to test our new information retrieval tool in the following chapters.

# Chapter 6 - Classification of different hypermedia applications developed in Microcosm

## 6.1 Introduction

The improved model was developed using the Microcosm system and for this reason we must use some of its applications to test our hypotheses.

We will be using different Microcosm hypermedia applications, authored in different styles. This will allow us to draw different conclusions from each of the applications. It would be difficult to identify an ideal, if there is one, fully implemented hypermedia application, and anyway, different authors, and different users, have different ideas about what is a good hypermedia application.

For constructing context descriptors, and evaluating their importance for multimedia information retrieval, we need to assess the type of application we are using. Since we are trying to find a relationship between the different metadata parameters, for better constructing context descriptors, then this is of key importance.

In the first place we will consider the size of different applications and briefly describe the subject of the selected ones.

Secondly we will study the distribution of the links within the hypermedia network: availability, type and meta-information associated with each link, are factors to be consider for each application.

It is also of key importance to know in which way the abstract information is used. Lastly, things such as the relevance of the document name description to the content, the number of organising classification clusters, and how they are assigned to those clusters are assessed.

## 6.2 The subject and size of the applications

Ten different Microcosm applications were collected, which were developed in different contexts. Some of these applications were developed using older versions of Microcosm but all of then have been adequately ported to the latest version.

Some of the applications were developed simply to introduce Microcosm to novices and to demonstrate its functionalities. Other applications had a more specific purpose and were developed as support material for different subjects or as self-assessed learning.

In order to judge how appropriate it would be to use these applications in information retrieval tests we first found out how many documents each application had. As we can see in Table 6-1 none of the applications are that large in size, Tulip being the largest one.

For reasons that will be explained later, we are particularly interested in applications with a reasonable number of text documents. From the ten available applications just five of them, Archaeology, Cell Biology, French, French Revolution and Tulip, have more than 100 text documents. Tulip is the largest application in this sense, with 374 text files. So, in our numerical tests we will be mainly using these five applications. The size of these applications is quite usual for Microcosm; we do not usually find them bigger. This must be considered in our conclusions, since for information retrieval we can generally make stronger statements when they are supported by a large test collection, but then we would be moving away from the kind of applications developed using this system. Microcosm applications are not generally composed of ad-hoc information. They are, on the contrary, applications typically developed with a specific goal, in a way that facilitates learning in a particular subject.

| | | Files | | |
|---|---|---|---|---|
| | | Total | Non-text | Text |
| Applications | Archaeology | 189 | 45 | 144 |
| | Caerdroia | 167 | 137 | 30 |
| | Cell Biology | 237 | 134 | 103 |
| | French | 226 | 80 | 146 |
| | French Revolution | 144 | 43 | 101 |
| | Pathology | 204 | 184 | 20 |
| | IDSI | 58 | 35 | 23 |
| | Romeo and Juliet | 68 | 27 | 41 |
| | Shell | 398 | 323 | 75 |
| | Tulip | 449 | 75 | 374 |

Table 6-1 - Applications size as indicated by the total number of files, non-text files and text files

A description of the subject of the five largest applications follows:

- The Archaeology Application is an adaptive teaching application and was designed for Archaeology students. It covers four dating techniques in archaeology (dendrochronology, archaeomagnetic dating, radiocarbon dating, thermoluminescence and obsidian hydration dating) and two sections on forgery and microscopy;

- The Caerdroia Application is the electronic implementation of Caerdroia issue 25, the journal of the Caerdroia Mazes and Labyrinths society. All the documents have been indexed and linked to create an application that is more than just the printed journal displayed on a screen, and additional material has been included to show all the possibilities offered by Microcosm.

- The Cell Biology Application discusses cell biology and mobility. The subject of cell motility is split into four broad categories: ciliary movement, flagellar movement, amoeboid movement and tissue cell movement. The application also presents information on a common subject relevant to all cell movement - the cytoskeleton;

- The French Application was developed as a package to offer a range of authentic resources in French, with related language-learning activities. This application is divided into four different subjects: underwater exploration, missiles and launchers, satellites, and earth exploration. These are presented in video, audio and text format and for each of these subjects there are some proposed grammar, lexical and other activities;

- The French Revolution Application: the People enter Politics is a tutorial publication of the UK Teaching and Learning Technology Programme (TLTP) History Courseware Consortium. This application is a collection of documents from different authors that focus on different aspects of French Revolution history. These subjects are chronologically organized in four sections: the Prologue of the revolution; the Estates General - establishing representative government; the storming of the Bastille - the people emerge; and bread and terror - the spectre of popular government.

| | | | Files | | | Links | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | on Non-text | | | on Text | | |
| | | | Total | Non-Text | Text | Total | Generic | Button | Total | Generic | Button | Total | Generic | Button |
| Archeology | Total | | 190 | 45 | 144 | 975 | 33 | 914 | 87 | 0 | 87 | 1705 | 66 | 1589 |
| | Average | per doc | | | | 9.4 | 0.3 | 8.8 | 1.9 | 0.0 | 1.9 | 11.8 | 0.5 | 11.0 |
| | | per doc with links | | | | 10.4 | 1.4 | 10.0 | 2.2 | | 2.2 | 12.7 | 1.4 | 12.3 |
| | Standard Deviation | per doc | | | | 10.5 | 0.8 | 10.0 | 1.2 | 0.0 | 1.2 | 11.0 | 0.9 | 10.5 |
| | | per doc with links | | | | 10.5 | 1.0 | 10.1 | 1.0 | | 1.0 | 10.9 | 1.0 | 10.4 |
| Caerdroia | Total | | 165 | 137 | 28 | 417 | 172 | 179 | 277 | 91 | 131 | 151 | 72 | 63 |
| | Average | per doc | | | | 2.6 | 1.0 | 1.2 | 2.0 | 0.7 | 1.0 | 5.4 | 2.6 | 2.3 |
| | | per doc with links | | | | 3.3 | 2.6 | 1.5 | 2.7 | 2.0 | 1.3 | 5.4 | 4.2 | 2.3 |
| | Standard Deviation | per doc | | | | 3.4 | 1.9 | 2.4 | 1.9 | 1.3 | 0.8 | 6.5 | 3.1 | 5.4 |
| | | per doc with links | | | | 3.3 | 2.6 | 1.5 | 2.7 | 2.0 | 1.3 | 5.4 | 4.2 | 2.3 |
| CellBiology | Total | | 238 | 134 | 103 | 2753 | 2429 | 265 | 139 | 12 | 125 | 2796 | 2429 | 265 |
| | Average | per doc | | | | 11.7 | 9.6 | 1.6 | 1.0 | 0.1 | 0.9 | 25.7 | 22.1 | 2.6 |
| | | per doc with links | | | | 33.0 | 33.4 | 2.9 | 2.1 | 0.3 | 1.9 | 46.6 | 48.0 | 3.6 |
| | Standard Deviation | per doc | | | | 19.5 | 33.7 | 4.1 | 3.3 | 1.2 | 3.1 | 26.2 | 39.3 | 4.9 |
| | | per doc with links | | | | 40.8 | 55.6 | 3.3 | 2.5 | 0.4 | 2.4 | 46.9 | 58.4 | 3.7 |
| French | Total | | 227 | 80 | 146 | 731 | 25 | 697 | 198 | 0 | 198 | 1029 | 29 | 984 |
| | Average | per doc | | | | 5.4 | 0.1 | 5.2 | 2.5 | 0.0 | 2.5 | 7.0 | 0.2 | 6.7 |
| | | per doc with links | | | | 5.9 | 1.6 | 5.8 | 3.1 | | 3.1 | 7.1 | 1.6 | 7.1 |
| | Standard Deviation | per doc | | | | 7.6 | 0.7 | 7.4 | 4.1 | 0.0 | 4.1 | 8.6 | 0.8 | 8.3 |
| | | per doc with links | | | | 7.8 | 1.9 | 7.6 | 4.3 | | 4.3 | 8.6 | 1.9 | 8.4 |
| French Rev | Total | | 146 | 43 | 101 | 372 | 79 | 292 | 140 | 0 | 140 | 453 | 149 | 304 |
| | Average | per doc | | | | 2.9 | 1.0 | 3.0 | 3.3 | 0.0 | 3.3 | 2.8 | 1.5 | 3.0 |
| | | per doc with links | | | | 3.3 | 74.5 | 3.4 | 3.3 | | 3.3 | 3.2 | 74.5 | 3.4 |
| | Standard Deviation | per doc | | | | 5.4 | 11.9 | 5.3 | 1.6 | 0.0 | 1.6 | 6.4 | 14.3 | 6.3 |
| | | per doc with links | | | | 5.6 | 69.5 | 5.5 | 1.5 | | 1.5 | 6.8 | 69.5 | 6.6 |
| Pathology | Total | | 203 | 184 | 18 | 349 | 160 | 188 | 153 | 0 | 153 | 326 | 320 | 6 |
| | Average | per doc | | | | 2.4 | 1.6 | 0.8 | 0.8 | 0.0 | 0.8 | 18.1 | 17.8 | 0.3 |
| | | per doc with links | | | | 3.1 | 320.0 | 1.0 | 1.0 | | 1.0 | 65.2 | 320.0 | 1.5 |
| | Standard Deviation | per doc | | | | 22.4 | 22.4 | 0.4 | 0.4 | 0.0 | 0.4 | 73.2 | 73.3 | 0.7 |
| | | per doc with links | | | | 25.4 | 0.0 | 0.1 | 0.1 | | 0.1 | 127.4 | 0.0 | 0.5 |
| Shell | Total | | 399 | 323 | 75 | 2234 | 19 | 2212 | 2122 | 6 | 2116 | 110 | 14 | 96 |
| | Average | per doc | | | | 5.6 | 0.1 | 5.5 | 6.6 | 0.0 | 6.6 | 1.5 | 0.2 | 1.3 |
| | | per doc with links | | | | 25.4 | 6.7 | 25.4 | 34.8 | 6.0 | 34.7 | 4.1 | 7.0 | 3.7 |
| | Standard Deviation | per doc | | | | 18.4 | 0.7 | 18.3 | 20.2 | 0.3 | 20.1 | 5.0 | 1.4 | 4.8 |
| | | per doc with links | | | | 32.1 | 4.1 | 32.2 | 34.2 | 0.0 | 34.1 | 7.7 | 5.0 | 7.6 |
| Tulip | Total | | 450 | 75 | 374 | 1205 | 527 | 677 | 49 | 0 | 49 | 1637 | 524 | 1111 |
| | Average | per doc | | | | 3.2 | 0.6 | 2.6 | 0.7 | 0.0 | 0.7 | 3.7 | 0.7 | 3.0 |
| | | per doc with links | | | | 3.5 | 4.9 | 2.9 | 1.5 | | 1.5 | 3.7 | 4.9 | 3.0 |
| | Standard Deviation | per doc | | | | 10.5 | 9.4 | 5.0 | 1.2 | 0.0 | 1.2 | 11.5 | 10.3 | 5.4 |
| | | per doc with links | | | | 3.5 | 4.9 | 2.9 | 1.5 | | 1.5 | 3.7 | 4.9 | 3.0 |

Table 6-2 - Evaluation at document level of links in the Microcosm applications

- The Tulip Application was developed by the Mental Health Group at the University of Southampton. The authors constructed Tulip as part of the Scholar Project, a campus wide scheme utilising new technologies as a platform for teaching students. It was designed for use by third year medical students, and covers a broad range of topics in psychiatry. The application has roughly three sections, which are interconnected:

Electronic textbook - including the psychiatry placement, third year handbook and clinical topics supported by a comprehensive glossary of psychiatric terms;

Case studies - patient's histories are presented in video, audio and text, and the student is asked to complete a mental state examination, reach a diagnosis and consider a management plan, with feedback about progress at each stage;

Quiz - a series of multiple choice questions on a broad range of clinical topics, which students may use to assess their progress. Linking to the clinical topics section provides feedback.

## 6.3 Availability of context information

### 6.3.1 Document level link information

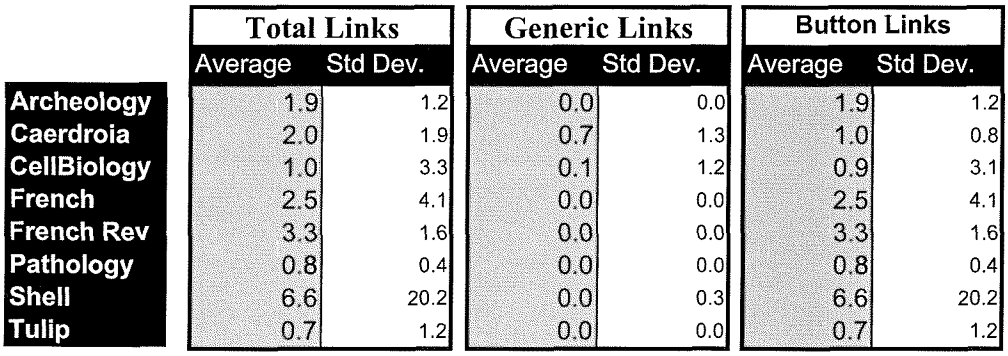
To study the distribution of links we ran a program to execute some tests with which it was possible to determine link availability and distribution. We basically searched for two kinds of interconnection: links from text documents to non-text documents, and links from text documents to text documents. In Appendix A the results of the tests and the distribution graphics are shown. A summary of the numerical information gathered is presented in Table 6-2. An example of the structure of the information about each link

### 6.3.1.1 Link availability

Considering link availability for constructing multimedia descriptors (Table 6-3), we can see that the Archaeology, Caerdroia, French, French Revolution and Shell applications have a good average of links connecting text files to non-text files. It is usually convenient to have more than two links per document as suggested by Dunlop (1993). However our system should also perform well under low link availability circumstances. The Shell and French applications have a high standard deviation on the average.

On these data a low standard deviation means that the majority of documents possess approximately the same number of links. A high standard deviation suggests that either there are a large number of documents with the number of links slightly away from the link average, or that there are a few documents with many links. This latter situation is particularly true in the Shell application where there is one document with over 100 specific links to/from text documents. This is due to the presence of a diagram document with many labels assigned to it.

Apart from Caerdroia, almost none of the applications use generic links to connect text to non-text documents. So we usually only have the specific links to relate non-text with text.



| | Total Links | | Generic Links | | Button Links | |
|---|---|---|---|---|---|---|
| | Average | Std Dev. | Average | Std Dev. | Average | Std Dev. |
| Archeology | 1.9 | 1.2 | 0.0 | 0.0 | 1.9 | 1.2 |
| Caerdroia | 2.0 | 1.9 | 0.7 | 1.3 | 1.0 | 0.8 |
| CellBiology | 1.0 | 3.3 | 0.1 | 1.2 | 0.9 | 3.1 |
| French | 2.5 | 4.1 | 0.0 | 0.0 | 2.5 | 4.1 |
| French Rev | 3.3 | 1.6 | 0.0 | 0.0 | 3.3 | 1.6 |
| Pathology | 0.8 | 0.4 | 0.0 | 0.0 | 0.8 | 0.4 |
| Shell | 6.6 | 20.2 | 0.0 | 0.3 | 6.6 | 20.2 |
| Tulip | 0.7 | 1.2 | 0.0 | 0.0 | 0.7 | 1.2 |

Table 6-3 - Link availability for constructing context multimedia descriptors

For evaluation proposes that we will explain later, we will have to consider links interconnecting text documents for building context descriptions (Table 6-4).

The Archaeology and French applications are the ones with more specific links per document, and also have a more uniform distribution of those links. However these applications don't possess many generic links.

| | Total Links | | Generic Links | | Button Links | |
|---|---|---|---|---|---|---|
| | Average | Std Dev. | Average | Std Dev. | Average | Std Dev. |
| **Archeology** | 11.8 | 11.0 | 0.5 | 0.9 | 11.0 | 10.5 |
| **Caerdroia** | 5.4 | 6.5 | 2.6 | 3.1 | 2.3 | 5.4 |
| **CellBiology** | 25.7 | 46.6 | 22.1 | 48.0 | 2.6 | 3.6 |
| **French** | 7.0 | 8.6 | 0.2 | 0.8 | 6.7 | 8.3 |
| **French Rev** | 2.8 | 6.4 | 1.5 | 14.3 | 3.0 | 6.3 |
| **Pathology** | 18.1 | 73.2 | 17.8 | 73.3 | 0.3 | 0.7 |
| **Shell** | 1.5 | 5.0 | 0.2 | 1.4 | 1.3 | 4.8 |
| **Tulip** | 3.7 | 11.5 | 0.7 | 10.3 | 3.0 | 5.4 |

Table 6-4 - Links for constructing context descriptors for text files

The Caerdroia application has a reasonable distribution of generic and button links, and a very good link average, i.e., average number of links per document. The problem is that it only has 28 text files. With such a small test collection it is difficult to make strong conclusions. Although we will use it in some examples, Caerdroia is definitely an application that we cannot consider for systematic evaluation with text documents.

The Cell Biology application on the contrary has a good number of specific links per document, and also possesses a better availability of generic links than any other application. On the other hand there is a high standard deviation on the distribution of those links. This suggests that there are few documents with a high number of generic links. We noticed the existence of 3 documents with over 200 generic links in them. This is clearly due to the presence of a glossary for the application, accessed through generic links.

In the French Revolution application there is a smaller number of generic links per document than in Cell Biology. However the standard deviation of this application is also quite high, and again this is due to the presence of a glossary file. The availability of specific links is also enough for test proposes.

The Shell application has a low average of links per document. This together with the total number of text documents lead us to not take this application into consideration in the evaluations.

As far as links connecting text files are concerned, the Tulip and French Revolution applications are quite similar. The link average is similar and Tulip also contains a glossary.

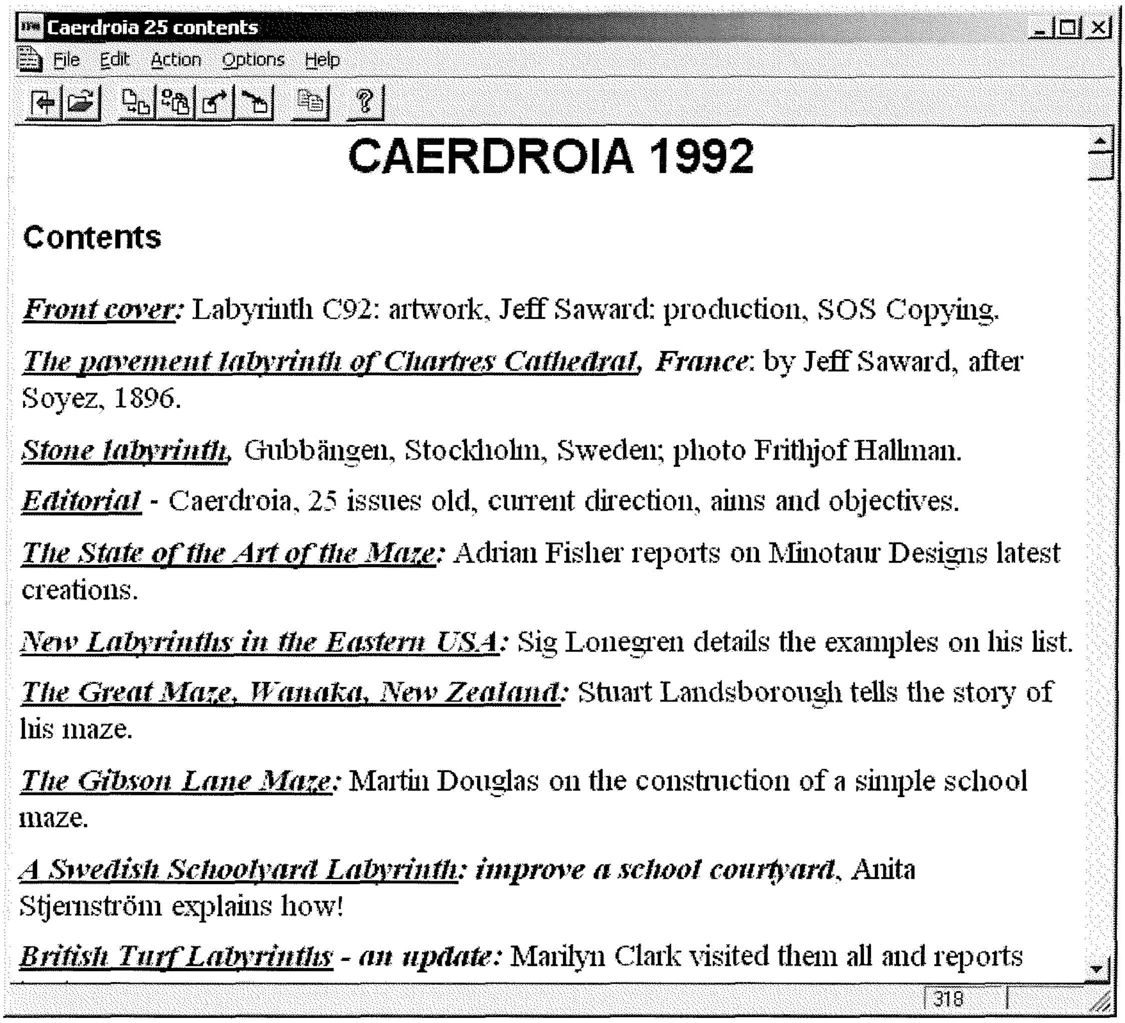
### 6.3.1.2 Qualitative assessment

It is also necessary to qualitatively inspect the way in which the links were created. We will be interested in things like the kind of link description being used, the type of text hold on anchors and in which way they are referenced in documents. The intent of the links should also be inspected. This information has been gathered by analysing the link database of each application. An example of the kind of information obtained is shown in Table 6-5. From the link database it was possible to retrieve the evaluation about anchor selection, anchor location and link description.

```
\DoctName              100.03.22.94.17.18.12.5480471
\SourceFile            100.03.22.94.17.18.12.5480471
\SourceDocType         TEXT
\DoctType              TEXT
\SourceLogType         \\MultiItem0/Caerdroia/Articles
                       \\MultiNumItems 1
\SourceSelection       Amalienborg hedge maze
\RealSourceSelection   Amalienborg hedge maze
\Selection             Amalienborg hedge maze
\SourceOffset          2069
\Offset                2069
\Description           Amalienborg, Copenhagen, Denmark
\LinkBaseRecordNum     47
\LinkBaseFile          C:\\MCM\\mazes\\maze.ddf
\ButtonAction          FOLLOW.LINK
\Action                UPDATE.LINK
\DestFile              100.03.21.94.14.30.47.12111292
\DestDocType           BITMAP
\DestSelection
\DestOffset            0
\UniqueDocID           2
```

Table 6-5 – Example of link information record held in a link data base for the Caerdroia application

To check on links intent, its location in documents has been inspected, and its importance to semantically connect concepts or to structure navigation was inferred. This as been done for all documents in every application. An example of navigations links is shown in Figure 6-1, where this kind of specific links is presented in bold, underlined and italic style in a document that describes the content of the application.



CAERDROIA 1992

Contents

***Front cover:*** Labyrinth C92: artwork, Jeff Saward: production, SOS Copying.

***The pavement labyrinth of Chartres Cathedral, France***: by Jeff Saward, after Soyez, 1896.

***Stone labyrinth,*** Gubbängen, Stockholm, Sweden; photo Frithjof Hallman.

***Editorial*** - Caerdroia, 25 issues old, current direction, aims and objectives.

***The State of the Art of the Maze:*** Adrian Fisher reports on Minotaur Designs latest creations.

***New Labyrinths in the Eastern USA:*** Sig Lonegren details the examples on his list.

***The Great Maze, Wanaka, New Zealand:*** Stuart Landsborough tells the story of his maze.

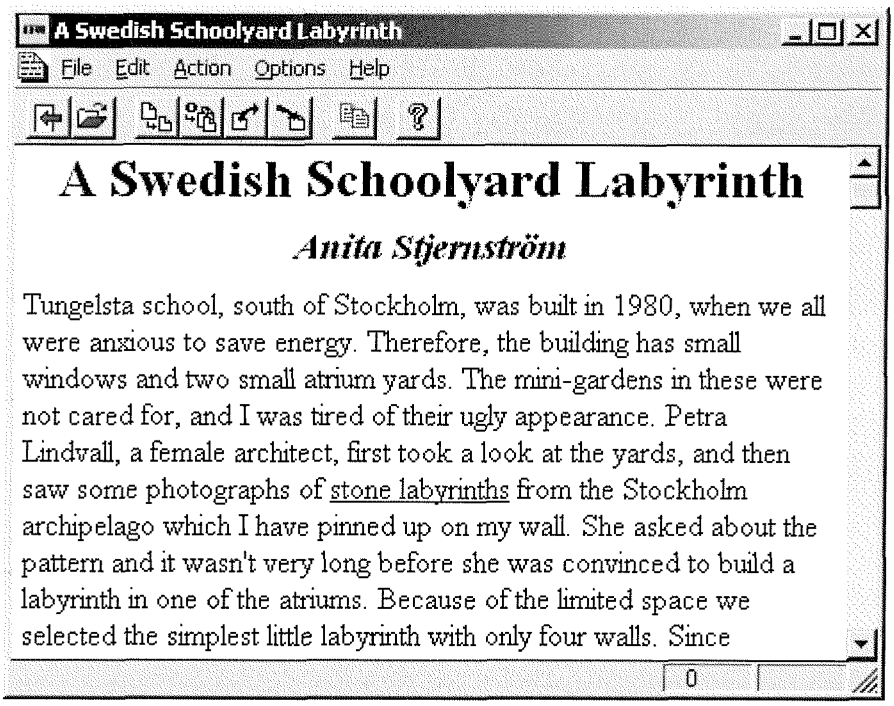
***The Gibson Lane Maze:*** Martin Douglas on the construction of a simple school maze.

***A Swedish Schoolyard Labyrinth: improve a school courtyard,*** Anita Stjernström explains how!

***British Turf Labyrinths - an update:*** Marilyn Clark visited them all and reports

Figure 6-1 – Example of navigational specific links in the Caerdroia application

On the other end concept specific links usually highlight only few words that appear in the middle of free text, as shown in Figure 6-2.



## A Swedish Schoolyard Labyrinth

*Anita Stjernström*

Tungelsta school, south of Stockholm, was built in 1980, when we all were anxious to save energy. Therefore, the building has small windows and two small atrium yards. The mini-gardens in these were not cared for, and I was tired of their ugly appearance. Petra Lindvall, a female architect, first took a look at the yards, and then saw some photographs of stone labyrinths from the Stockholm archipelago which I have pinned up on my wall. She asked about the pattern and it wasn't very long before she was convinced to build a labyrinth in one of the atriums. Because of the limited space we selected the simplest little labyrinth with only four walls. Since

Figure 6-2 – Example of conceptual specific links in the Caerdroia application

**Archaeology:**

In this application there are not that many generic links, however we don't have a glossary that aggregates all the destinations together. An advantage of generic links in this application is the information about the originating document and the reference to where the link was created. This allows us to expand the bound of the generic link anchor if necessary. The anchors of these links are words/concepts, and the destination is their definition/description. For generic links the description is always the same as the description of the destination document.

In this application there are also only six links with a destination offset, and the textual destination anchor selection is always absent.

We found the following information about specific links: the navigational links were usually at the end of the document, and the links describing concepts in the middle of sentences/paragraphs:

| | |
|---|---|
| Anchor selection | Meaningful textual anchors. |
| Anchor location | The majority of links are usually at the end of documents where anchor selection coincided with the sentence and the paragraph,<br>Also many single links inserted in the middle of a sentence and paragraph. |
| Link intent | Generally navigational links;<br>Many links connecting words/concepts to their definition/description. |
| Link description | Always the same as the description of the destination document. |

**Cell Biology:**

In this application there are a lot of generic links. However, from the 2429 available links only 50 of then do not link to a glossary. Glossary is split into different files, one for each letter of the alphabet.

The link description for the glossary generic links is similar to the source selection. In the non-glossary links the link description is generally equal to the description of the destination document.

These links almost never hold the reference to the original source document and the offset where the link was created. It is not usually possible to expand the bound of the generic link anchor. The anchors of these links are words/concepts, and the link traverses to their definition/description.

All glossary generic links have a destination offset, but usually no anchor selection. There is only one non-glossary link with a textual destination anchor selection. The same happens with specific links, but with these links there is no information about the destination offset.

The following information is also found about specific links:

| | |
|---|---|
| Anchor selection | Meaningful textual anchors. |
| Anchor location | The majority of links are inserted in the middle of sentences and paragraphs, some times more than one link at a time;<br>Very few links where anchor selection coincided with the sentence and the paragraph. |
| Link intent | The majority of links connect words/concepts to their definition/description;<br>Some navigational links. |
| Link description | Mostly the same as the description of the destination document. |

**French:**

The number of generic links in this application is small, and there is no glossary that aggregates all the destinations of those links together. Their link descriptions are generally the same as the description of the destination document. The anchors of these links are meaningful text, but sometimes abbreviations, pointing to resources and/or activities. They are actually working both as relating/expanding concepts and as navigation links for resource discovery. For half of these links the destination offset reference and a selection is present, which sometimes has the same characteristics as the source. In this half, the reference to the original source document and the offset where the link was created is available. It is then possible to expand the bound of the generic link source anchor.

For the specific links there is a destination anchor selection text for 124 links and destination offset for 116 links. Around 17% of links hold this metadata. The following metadata is also found about specific links:

| | |
|---|---|
| Anchor selection | Usually textual symbols, and the same words used many times as anchor;<br>Sometimes meaningful text. |
| Anchor location | The majority of links are usually at the end or the beginning of documents where anchor selection coincides with the sentence and some times the paragraph;<br>In few links the anchor selection is inserted in the middle of a sentence and subsequently in the middle a paragraph. |
| Link intent | The majority of links connect documents with proposed 'activities', mostly outside the scope of the original subject;<br>Some links advising further reading. |
| Link description | Predominantly the same as the description of the destination document. |

**French Revolution:**

In this application there is a reasonable number of generic links, all of which link to a glossary document. The link descriptions of these links are sometimes the same as the source selection but generally not. The anchors of these links are meaningful text about places, situation or people, which will be further explained or contextualized in the glossary. All the generic links have a destination offset, and a destination anchor selection. The reference to the original source document and the anchor offset is the same as in the destination, in all links apart from five where that information is present in a different document.

For the specific links there are no destination anchor selections and the destination offsets are present for seven links only. The following metadata about specific links is also found:

| | |
|---|---|
| Anchor selection | Usually textual symbols, and the same words used many times as anchors (ex. “Further information”);<br>Some links use meaningful text. |
| Anchor location | For a number of links the anchor selection coincides with the sentence and some times the paragraph;<br>In few links the anchor selection is inserted in the middle of a sentence and subsequently in the middle a paragraph. |
| Link intent | Usually suggesting further reading on related subjects. |
| Link description | Predominantly the same as the description of the destination document. |

**Tulip:**

In the Tulip application there are many generic links, of the 527 available links only 67 do not link to a glossary. For all generic links the description is the same as the description of the destination document. The anchors of these links are words/concepts, and the links traverse to their definition/description.

These links never contain the reference to the original source document or the offset where the link was created. In all the glossary generic links we have a destination offset, but no destination anchor selection. In the non-glossary generic links we find neither the destination offset nor the selection.

For the specific links there are no destination anchor selections and the destination offset is present for only 8 links. The following metadata is also found about specific links in Tulip:

| | |
|---|---|
| Anchor selection | Usually visual symbols, without any text;<br>Some meaningful text anchors. |
| Anchor location | Many of the links are inserted in a sentence coinciding with the paragraph;<br>In few links the anchor selection is inserted in the middle of a sentence and subsequently in the middle a paragraph. |
| Link intent | Many navigational links;<br>Many links to build up the functionality of a quiz;<br>Some of links connect words/concepts to their definition/description. |
| Link description | Predominantly the same as the description of the destination document. |

In all the applications used in our evaluations, the majority of links have no clear distinction between link description and document description, or distinction between link description and anchor selection. The Microcosm system provides this facility, but usually the authors are lazy about creating link descriptions. By default the system assigns the document description of the destination to the link description.

Authors have generally used generic links as glossaries, as seen in the Cell Biology, French Revolution and Tulip applications. The notable exceptions are Archaeology and French. However in the later application, the generic links do not relate semantically closed information as in the other applications. They are used as navigation links for resource discovery, with anchors on abbreviations. The Cell Biology and Tulip applications also possess some non-glossary generic links expanding concepts.

The use of specific links is generally broader, in these hypermedia applications. These links are generally used to support navigation, as seen to a greater extent in Archaeology and Tulip, or as concept relation/expansion as seen in Cell Biology and Archaeology, and to a lesser degree in Tulip. We found links for creating a quiz (Tulip application), resource discovering on related subjects (French Revolution application), and unrelated subjects (French application).

The anchor selections for specific links in the Archaeology and Cell Biology applications are generally meaningful. In the French and French Revolution applications the anchors are sometimes unrelated to the subject being discussed. These are the resource discovery links. In the Tulip application the anchor selection is almost always absent.

With regard to navigational links and resource discovery links, they are usually inserted in a sentence, and this sentence usually coincides with the whole paragraph. Links for relating/expanding concepts are usually inserted in the middle of the text, during the discussion of the subject.

### 6.3.2 Abstract level information

#### 6.3.2.1 Nature of abstract information

At this level, our applications use different kinds of abstract information. This is metadata that is not only considered by content but also by context.

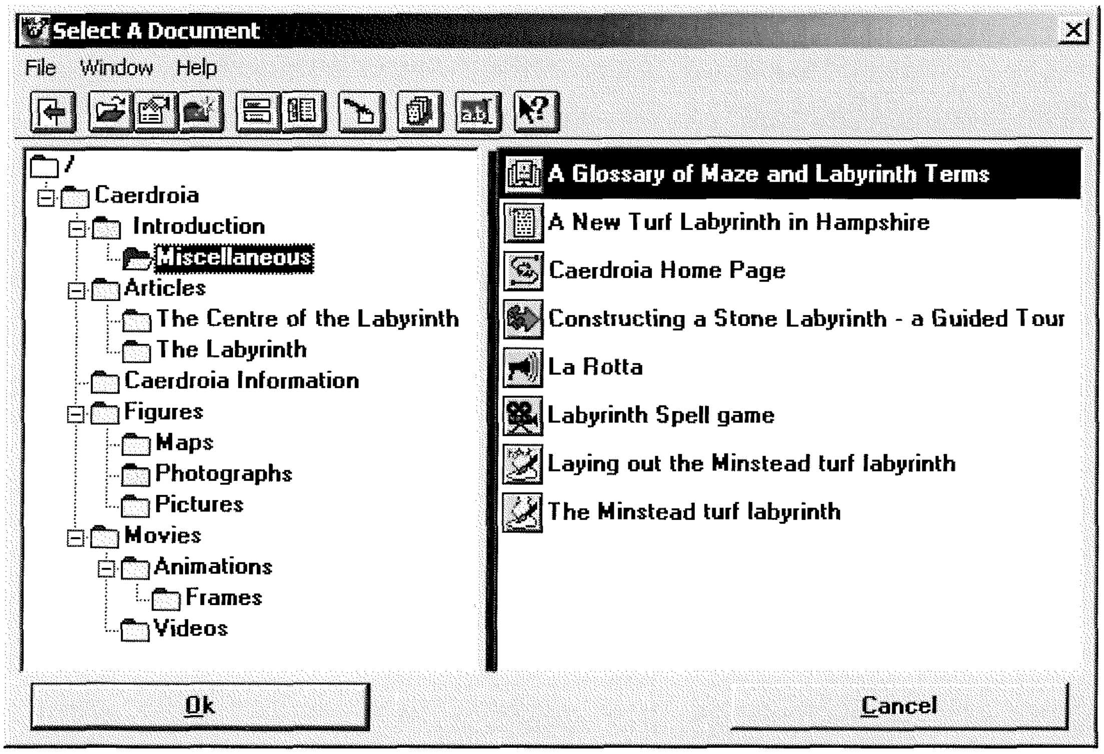
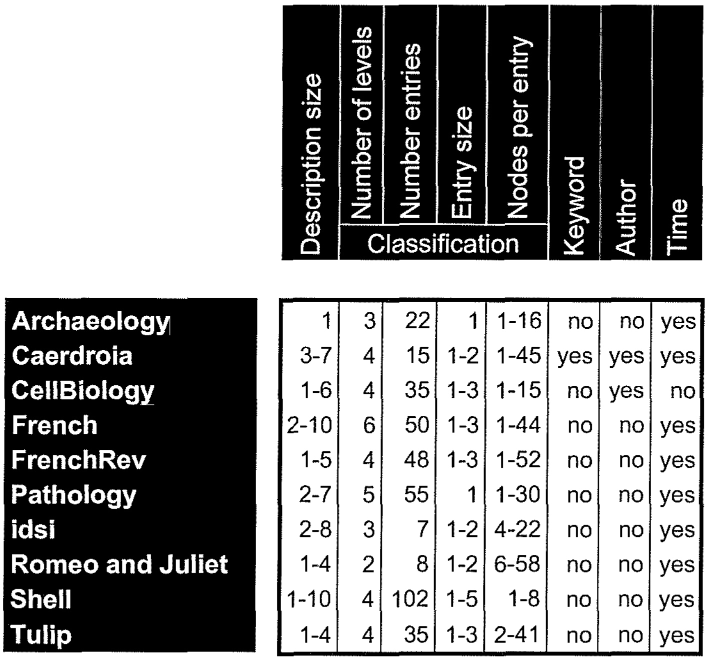


Figure 6-3 – Example of classifications and document description in the Caerdroia application

In the first place each document has a description assigned to it (see right pane example in Figure 6-3). The author introduces this data, along with some possible keywords. We should note that this description is distinct from the physical location, name and path, which is also available as an attribute. There is also the time when the file was placed in the application, and author. Finally, the document is included in the classification. This classification is organized as a tree of concepts in Microcosm, for ease of navigation (see left pane example in Figure 6-3). The same document can have multiple entries on the tree.

In fact, the Microcosm system allows for the use of many other attributes, but these were the ones that were included in our applications. Table 6-6 shows a summary of abstract information available for the different applications. In this table when two numbers are given they represent the two extremes of the found quantities.



| | Description size | Number of levels | Number entries | Entry size | Nodes per entry | Keyword | Author | Time |
|---|---|---|---|---|---|---|---|---|
| | | Classification | | | | | | |
| Archaeology | 1 | 3 | 22 | 1 | 1-16 | no | no | yes |
| Caerdroia | 3-7 | 4 | 15 | 1-2 | 1-45 | yes | yes | yes |
| CellBiology | 1-6 | 4 | 35 | 1-3 | 1-15 | no | yes | no |
| French | 2-10 | 6 | 50 | 1-3 | 1-44 | no | no | yes |
| FrenchRev | 1-5 | 4 | 48 | 1-3 | 1-52 | no | no | yes |
| Pathology | 2-7 | 5 | 55 | 1 | 1-30 | no | no | yes |
| idsi | 2-8 | 3 | 7 | 1-2 | 4-22 | no | no | yes |
| Romeo and Juliet | 1-4 | 2 | 8 | 1-2 | 6-58 | no | no | yes |
| Shell | 1-10 | 4 | 102 | 1-5 | 1-8 | no | no | yes |
| Tulip | 1-4 | 4 | 35 | 1-3 | 2-41 | no | no | yes |

Table 6-6 - Abstract context information availability

A more qualitative evaluation follows for the applications we used for our testing:

### Archaeology:

In this application, the document/node description contains only one word. The problem here is the comprehension of that associated word. It is usually a juxtaposition of several abbreviations related to the content of the document. However, there are some exceptions.

The words used in the document classification suffer from the same problem. At the first and second levels they are human-readable words, at the third they are not. The documents are then classified according to the major dating techniques in archaeology and the different levels of expertise required. Each document is present only once in the classification tree. There are not that many documents per entry. They are uniformly distributed amongst almost all the entries in the classification, with usually between 4 and 7 per entry.

There are only two different dates for all the documents.

### Caerdroia:

In Caerdroia we can usually find in the document description more than three words related to the content of the document. It works like a document title.

In the document classification a large number of words reflect the type of tree structuring. The documents are always catalogued according to their file format, and so many words represent this. Nevertheless documents have multiple entries in the tree, and they are also organized under a short book style index. The majority of entries possess between 3 and 10 documents. The exceptions are the entries representing certain file formats.

There is also present in the majority of documents between 2 and 5 human indexed keywords, the author of the document and a date, although the date is the same for all documents. This shows the introduction date of the document in the application and not its production date.

### Cell Biology:

The document descriptions in Cell Biology are used as a document title. There are some descriptions with one word, but usually we have more than two words related to the content of the document.

The documents are classified not only under different subjects of cell motility, but also under their file format. So we have a mixture of categorisation words in classification. There is a uniform distribution of around 8 documents per subject entry, but there are some documents classified under more than one subject. Under file format the distribution is not that uniform, with some entries without any documents, and others with a large number of then. There is also a separate entry for all glossary files.

The author attribute is uniformly the same on all documents. This indicates that the author refers to the application, and not the individual documents.

### French:

This is similar to other applications, as the document description also works like a document title. So there are some one-word descriptions but normally we have two or more. Frequently we also find in some document descriptions the repeated use of a few words.

In the French application the names of the entries are not very related to the subject of documents. Entries structure the application according to the usability of documents. The metadata reflect this resource classification. There is also a big disparity in the distribution of documents, with many empty entries, and some with a large proportion of total documents.

There is a diversity of dates attributed to documents. This reflects the date of document production.

### French Revolution:

In this application, document descriptions usually have more than two words related to the content of the document. They also work like a document title.

Documents are classified according to at least three different aspects: theme, author and media type. There is also a joint classification of document under an entry called "Title". The vocabulary used in the classification echo this, with names of people under author entries and the names of French revolution historical periods under themes. The distribution of documents per classification entry is not that uniform. Under the few theme entries we found between 16 and 52 documents. Under author entries documents are better distributed. We found on average around 4 documents per entry. The glossary is kept in a separate entry.

There is a diversity of different document dates.

**Tulip:**

Here we typically have one or two words in the document description, hinting at their content. There is however the repeated use of some words in various documents. These documents are used in implementing a quiz and their description reflects the enumeration of different questions.

Words are used in the classification for sorting text documents according to several clinical topics or case histories. Non-text documents are classified under their media type. There is also an entry for a glossary file. Under clinical topics we found a uniform distribution of around 14 documents per entry. Under case histories there are around 36.5 documents per entry. There are some empty entries.

The date is the same for almost all documents in the collection.

In summary, document descriptions usually consist of a small number of words related in some way to the content of the document they describe. The Archaeology application and some documents in the Tulip application used a mixture of symbols and short abbreviations for descriptions so it will be more difficult to build context descriptor for the files in these applications. In the French and Tulip applications there is the repeated use of a few words in some documents. In the remaining applications and in the remaining documents of the French and Tulip applications the description is similar to a human readable title.

The abstract information also consists of a classification of subject, file type, etc, of the different nodes/documents in a hierarchical tree. The words used in the classification can be of good value for context descriptors. The Caerdroia, Cell Biology and French Revolution documents have multiple entries in the classification tree. Here their authors also opted for a file format type (text, image, video, sound, etc) classification. But in the French Revolution application there is also an author classification apart from the common subject arrangement. It is possible that the French application might have the worst classification when we attempt a good context indexing of documents.

Microcosm has the facility for using keywords but only the Caerdroia applications applied this feature. The date is present in all applications apart from Cell Biology, but its use is irrelevant for context indexing, since it is generally the same for all documents, so it doesn't help to distinguish them. The Caerdroia and Cell Biology application exploit the author facility but only the former uses it for distinguishing between different authors.

#### 6.3.2.2 The location of source and destination documents of links with regard to document classifications

Links connect documents or pieces of information that are related in same way. But since these documents can also be located under a classification tree in Microcosm, it is possible to use their relative position to classify the links between them.

In one of the tests that we will be running, we will try to find out if links connecting documents with the same classification are more important for context indexing than links connecting documents in different classifications. Therefore, there is an interest in knowing where this information is located in the classification tree, with relation to links. A summary of the link distribution according to the classification location of documents is given in Table 6-7. This is different from the figures presented in Table 6-2, which show the number of links connecting documents

| Applications | Total | | | Text | | | Non-Text | | |
|---|---|---|---|---|---|---|---|---|---|
| | Outside Classification | Inside Classification | Inside Average | Outside Classification | Inside Classification | Inside Average | Outside Classification | Inside Classification | Inside Average |
| Archaeology | 187 | 747 | 1.00 | 124 | 731 | 1.00 | 63 | 16 | 1.00 |
| Cell Biology | 168 | 112 | 1.03 | 101 | 79 | 1.01 | 67 | 32 | 1.06 |
| French | 629 | 71 | 1.18 | 450 | 54 | 1.24 | 179 | 17 | 1.00 |
| French Revolution | 177 | 186 | 1.48 | 123 | 100 | 1.06 | 54 | 86 | 1.97 |
| Tulip | 77 | 535 | 1.00 | 27 | 529 | 1.00 | 48 | 0 | 0.00 |

Table 6-7 - Link distribution according to classification of the documents they connect

Inside and Outside Classifications fields represent the total number of links connecting, respectively, documents under the same classification and on different classifications. Inside Average represents the average number of common concepts that documents are sharing.

In this study we did not consider the classification of documents under their format type (video, sound, etc). This was because this information is not used during context indexing. In relation to context indexing, it makes more sense to consider only the entries where documents are aggregated in a more conceptual way.

The Cell Biology and French Revolution applications have a similar number of links connecting text documents in the same concept entry (Inside Classification column entry on Table 6-7), as opposed to distinct ones (Outside Classification column entry on Table 6-7). In the Archaeology and Tulip applications the majority of links connect documents in different concept entries, and the opposite is true in the French application.

After eliminating the format classification (the classifications referent to the documents type, like video, text, sound or image), the French application is the only one where text documents, connected with links, possess more than one common entry in the classification (Inside Average column entry on Table 6-7). This also takes place in part with the French Revolution application. Since in the other applications all documents generally have only one entry on the classification tree, the linked documents can also only have one common entry.

## 6.4 Conclusions

This chapter presents an evaluation of different Microcosm applications available for evaluation. Five of them were chosen, and described in more detail.

Apart from one application, their size is quite similar, as far as the number of text files is concern. Nevertheless the subject of their content is quite diverse.

The distribution of the links has been studied. It has been seen that for both type of links and type of link destinations there are a diverse number of links available per document. A qualitative assessment has been made for the links connecting text documents. It has been shown that this links in different application relate information in a different way, and available meta-information may also have different characteristics.

The abstract level information has also been inspected, and it has been seen once again authors adopted different styles during application construction. In some of the applications, document description was highly relevant to its content and in others it was not. We also found that the number of organising classification clusters, and how documents were assigned to those clusters, were different in different applications. Author and time metadata did not help to distinguish between different documents, and keyword metadata was only available for one application. The distribution of links according to the classification of the documents they connect, proved that did not have a consistent pattern. For two applications the majority of links connected documents in the same classification, for two other applications the distribution of links was uniform, while for the remaining application, the majority of links connected documents that did not share the same classification.

After analyzing all the metadata from both the document and abstract levels, it has been seen that the available applications are quite different. For this reason a proper evaluation of the different available metadata, and its comparison with the style adopted for the construction of the application is needed. In Chapter 7 and Chapter 8 we will describe this evaluation and comparison, in the interest of the best context indexing for multimedia access. Different weights will have to be adopted for different metadata, since different metadata is also of different quality for different applications.

# Chapter 7 - Considering context link information for multimedia retrieval

## 7.1 Introduction

In this chapter a set of tests will evaluate different metadata extracted from links at document level. This evaluation will be mainly concerned with the context indexing and searching of documents or parts of documents. The adopted methodology for this evaluation has been described in Chapter 5. We will be basically finding the relative significance of the evaluated metadata for constructing context descriptors.

For reasons explained in Chapter 5, it is necessary to separate the evaluation at this level so that we only test some metadata elements at one time. There we found out that a reasonable number of elements would be four or three. For this reason we aggregated related metadata in the same test trying to have some common elements among all the tests. Three different evaluations had then to be run at this level:

In the first evaluation (section 7.2) we attempt to determine what should be the bounds of the link anchors to consider for the construction of context descriptors. These include the anchor selection, sentence and paragraph without discriminating between different link types and the two different link anchor ends (source anchor and destination anchor).

The second evaluation (section 7.3) attempts to determine whether different weights for different links types are important for building context descriptors, and also evaluates how anchor selection and link description are related.

The third evaluation (section 7.4) on metadata extracted from links at document level, finds out whether distinctions between link directions are important, while relating link description and link anchor selection, i.e., whether there are contextual differences when the anchor is the source or the destination of a link.

A small summary is given at the end of each of the individual tests, comparing the obtained weight for the evaluated metadata, with the style adopted for the construction of the applications. This describes the best way to assign weights to possible similar applications.

The comparisons of the obtained weights for link description and link anchor selection on the different individual evaluations allows the three evaluations to be compared in Chapter 9.

## 7.2 Determining the bounds of the link anchor for context description

On the used application, the anchors are inserted in the documents in different ways. This evaluation examines how far we should go in considering context information taken from the link anchor - sentence, paragraph and selection - in order to build the context text descriptors.

Therefore, all the different weight combinations between 0 and 4 are considered for these 12 elements as shown in Table 7-1. However, the evaluation was approached in a different way: the weight variation of generic links and specific links were tied together, as well as the distinction between source and destination links. This is shown with the same grey filling. This gave 125 tests to run in this evaluation, which could be finished in a few hours.

The metadata being analysed are:

- Link Anchor Selection
- Link Anchor Sentence
- Link Anchor Paragraph

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 2 - 2 | 2 - 2 |
| | Text Anchor as Source: | Selection | 0 - 4 | 0 - 4 |
| | | Sentence | 0 - 4 | 0 - 4 |
| | | Paragraph | 0 - 4 | 0 - 4 |
| | Text Anchor as Destination: | Selection | 0 - 4 | 0 - 4 |
| | | Sentence | 0 - 4 | 0 - 4 |
| | | Paragraph | 0 - 4 | 0 - 4 |
| Abstract Information context | Self | Description | 2 - 2 | |
| | | Classification | 2 - 2 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 2 - 2 | 2 - 2 |
| | | Classification | 2 - 2 | 2 - 2 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 1 - 1 | 1 - 1 |
| | Link Outside Classification | | 1 - 1 | 1 - 1 |

Table 7-1 - Weight intervals for determining link anchor bounds

Below are some graphical and numerical results for each application, along with some comments on these results, followed by some brief conclusions on how to consider the bounds of anchors for context description.

## Archaeology:

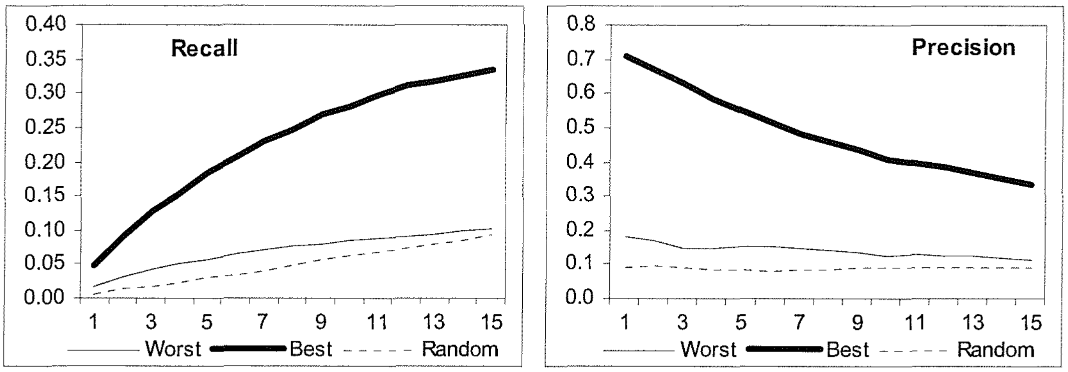


Figure 7-1 - Recall and precision graphs for Archaeology while evaluating link bounds

|  | Selection | Sentence | Paragraph | Average Recall |
|---|---|---|---|---|
| Best | 4 | 0 | 1 | 0.2279 |
| Worst | 0 | 0 | 0 | 0.0701 |

|  | Selection | Sentence | Paragraph | Average Precision |
|---|---|---|---|---|
| Best | 3 | 1 | 0 | 0.4868 |
| Worst | 0 | 0 | 0 | 0.1412 |

Table 7-2 - Recall, precision and weight values for Archaeology concerning link bounds

For this application, all anchor selections are expressive. This is one of the reasons why the weight for the selection is so high. We could expect that the anchors of navigational links would not be so highly relevant, but the poor abstract information might be the cause of this high value. The weighting of one for the paragraph also proves that there is a need for more descriptive information to improve recall.

In the best match for precision, the paragraph is not a factor to consider, but the sentence is. This is apparently predictable, since by considering more textual information from anchors one might improve recall, but if some of this text is also unrelated to the document at the other end of the link, this will definitely not improve precision. Moreover, the paragraph has a much larger concept expansion than a sentence.

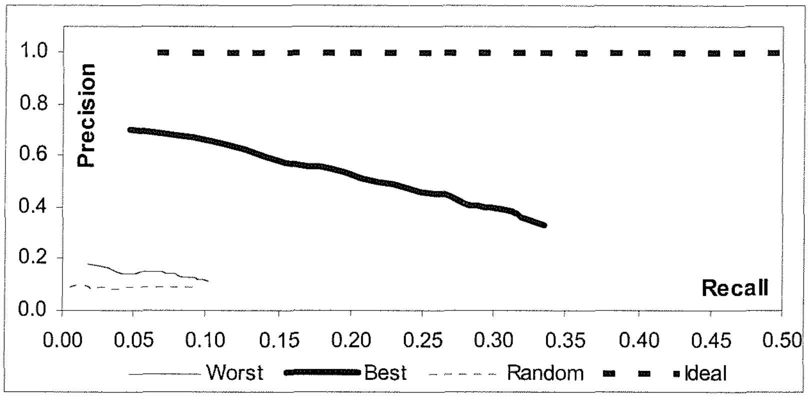


Figure 7-2 - Recall/precision graph for Archaeology while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| Best | 4 | 0 | 1 | 0.2279 | 0.4868 | 0.3104 |
| Worst | 0 | 0 | 0 | 0.0701 | 0.1412 | 0.0937 |

Table 7-3 - Best overall weight values for Archaeology concerning link bounds

The best overall match, if we consider equally precision and recall for the effectiveness, is 4, 0 and 1 for the selection, sentence and paragraph respectively. This match is the same as the best for recall, proving that there is a need to have high values of recall for the best effectiveness. Not using this information leads to the worst recall, precision and effectiveness. The proximity of this situation to the random one demonstrates the low importance of the abstract level. The difference between the worst and best solution proves how essential the anchor is for context description, under these circumstances.

## Cell Biology:

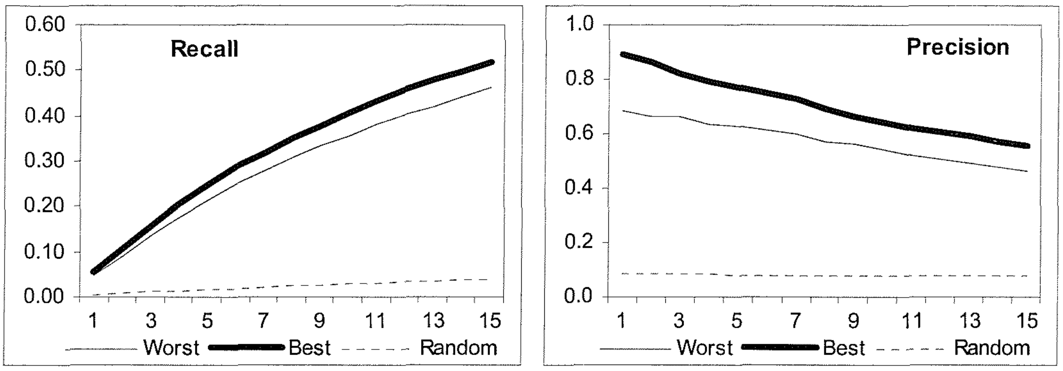


Figure 7-3 - Recall and precision graphs for Cell Biology while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall |
|---|---|---|---|---|
| Best | 4 | 1 | 0 | 0.3152 |
| Worst | 0 | 4 | 4 | 0.2876 |

| | Selection | Sentence | Paragraph | Average Precision |
|---|---|---|---|---|
| Best | 4 | 0 | 0 | 0.6914 |
| Worst | 0 | 4 | 4 | 0.5804 |

Table 7-4 - Recall, precision and weight values for Cell Biology concerning link bounds

Here, anchor selections are also expressive, leading once again to a high weight value for selection. A curious situation is that of the sentence. Considering it with a low weight together with a high weight for selection improves recall. But, if selection is not taken into account, its presence together with paragraph leads to the worst value of recall and also precision. This is easily explained, since in this application many times there is more than one link in the same paragraph, sometimes more than one link in the same sentence. This situation, where the same information is repeatedly indexed for different documents, does not help to create context descriptions that distinguish documents in a relevant way. This affects negatively precision and also recall.

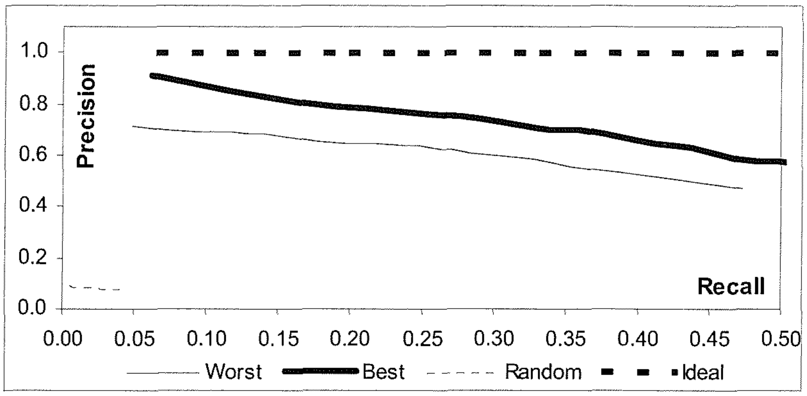


Figure 7-4 - Recall/precision graph for Cell Biology while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| Best | 4 | 0 | 0 | 0.3107 | 0.6914 | 0.4287 |
| Worst | 0 | 4 | 4 | 0.2876 | 0.5804 | 0.3846 |

Table 7-5 - Best overall weight values for Cell Biology concerning link bounds

The best option for effectiveness corresponds to the best option for precision. This proves that here there is not a big need to improve effectiveness with high values of recall by considering the sentence. Its absence will affect more positively effectiveness, by having higher values of precision. This is also due to the good abstraction level present in this application.

The distance of the worst situation to the random also proves the good quality of the abstract level, and the distance between the best and worst approach shows the quality of anchors in this application.

## French:

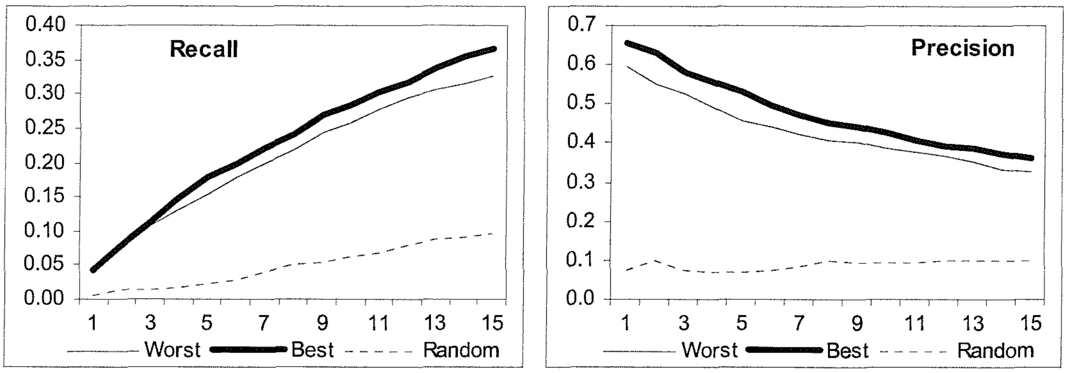


Figure 7-5 - Recall and precision graphs for French while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall |
|---|---|---|---|---|
| Best | 1 | 4 | 0 | 0.2304 |
| Worst | 0 | 0 | 0 | 0.2070 |

| | Selection | Sentence | Paragraph | Average Precision |
|---|---|---|---|---|
| Best | 1 | 3 | 0 | 0.4776 |
| Worst | 0 | 0 | 0 | 0.4426 |

Table 7-6 - Recall, precision and weight values for French concerning link bounds

The weight for the anchor selection in this application is not as relevant as in the previous ones. This situation is understandable, since here we often find as anchors, textual symbols and abbreviations. Also adding to the situation is the repeated use of some of these textual symbols in many anchors. However, in some specific links, and in generic links, the anchor selection is inserted, alone, in the middle of a sentence and a paragraph. Due to the circumstances, the recall is then improved, and to a lesser extent the precision, with a high weight for the sentence. We should also note the low value of precision.

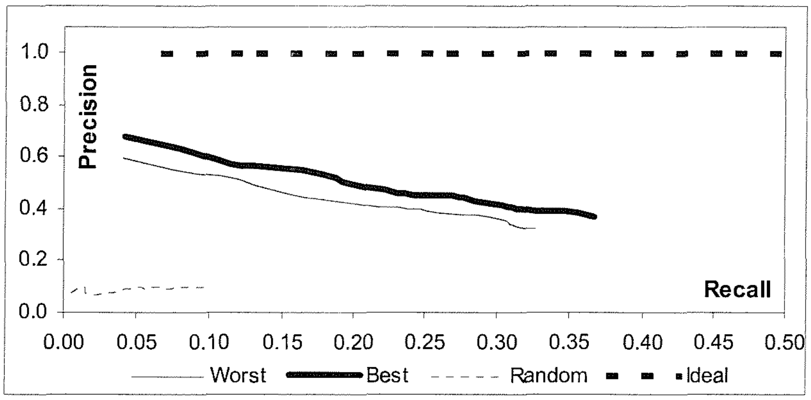


Figure 7-6 - Recall/precision graph for French while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| Best | 1 | 4 | 0 | 0.2304 | 0.4735 | 0.3099 |
| Worst | 0 | 0 | 0 | 0.2070 | 0.4426 | 0.2820 |

Table 7-7 - Best overall weight values for French concerning link bounds

The best overall match is the same as the best for recall. One, four and zero should be considered for the selection, sentence and paragraph respectively. Recall is then positively more affected by the presence of this information. Not using this information leads to the worst recall, precision and effectiveness.

## French Revolution:

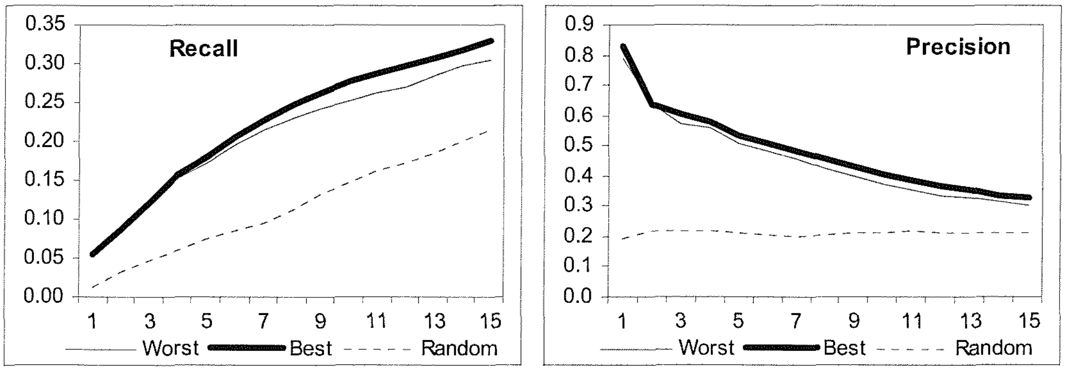


Figure 7-7 - Recall and precision graphs for French Revolution while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall |
|---|---|---|---|---|
| Best | 1 | 3 | 1 | 0.2261 |
| Worst | 0 | 0 | 0 | 0.2094 |

| | Selection | Sentence | Paragraph | Average Precision |
|---|---|---|---|---|
| Best | 1 | 2 | 0 | 0.4873 |
| Worst | 0 | 0 | 0 | 0.4611 |

Table 7-8 - Recall, precision and weight values for French Revolution concerning link bounds

The situation of the low weight for the anchor selection in this application is the same as for the French Revolution application. We often find as anchors, textual symbols and abbreviations, and also the repeated use of some selection text in many anchors. This proves that this type of anchor selection is not very relevant. But for some specific links the anchor selection is alone in the middle of a sentence and a paragraph. For this reason the recall is improved with a high weight for the sentence and low for the paragraph. However, the presence of the paragraph does not improve precision.

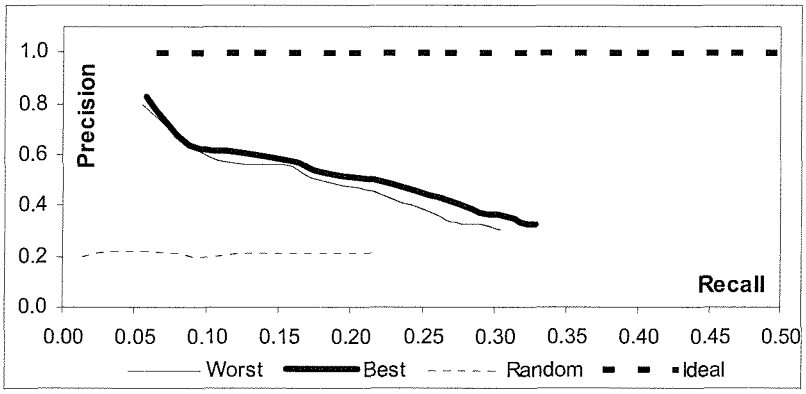


Figure 7-8 - Recall/precision graph for French Revolution while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| Best | 1 | 3 | 0 | 0.2256 | 0.4863 | 0.3082 |
| Worst | 0 | 0 | 0 | 0.2094 | 0.4611 | 0.2880 |

Table 7-9 - Best overall weight values for French Revolution concerning link bounds

The best overall weight is a balance between the best weight for recall and the best weight for precision. The paragraph is not present, but the sentence weight is the same as in the best for recall. In this application, the best and worst graphs are quite close. This points to the low quality of anchors for context description.

**Tulip:**

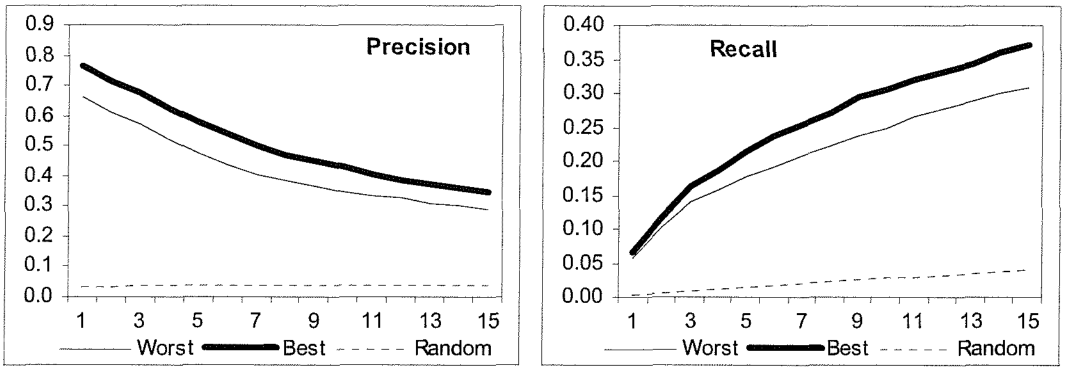


Figure 7-9 - Recall and precision graphs for Tulip while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall |
|---|---|---|---|---|
| Best | 1 | 3 | 1 | 0.2569 |
| Worst | 2 | 0 | 0 | 0.2252 |

| | Selection | Sentence | Paragraph | Average Precision |
|---|---|---|---|---|
| Best | 1 | 3 | 0 | 0.5415 |
| Worst | 2 | 0 | 0 | 0.4722 |

Table 7-10 - Recall, precision and weight values for Tulip concerning link bounds

For this application, many of the specific links are inserted in a sentence coinciding with the paragraph, making the weight for sentence quite high. Also for a few links, the anchor selection is inserted in the middle of a sentence, which in its turn is inserted in a paragraph. This also marks up the presence of the paragraph and a high value for the sentence on context description. This is all possible because there are some meaningful text anchors, but usually we only find visual symbols, without any text. The low importance of the selection is also emphasized by the weights on the worst scenario. Here selection is considered with a weight of two, when selection and paragraph are absent. One of the reasons for the system still considers that selection is due to the presence of generic links that possess expressive text.

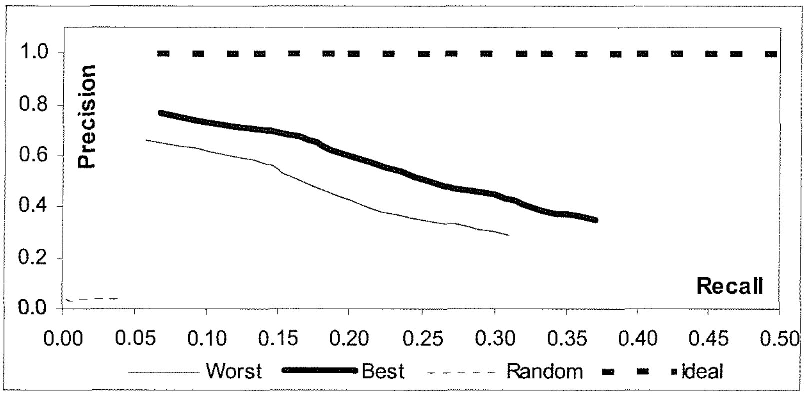


Figure 7-10 - Recall/precision graph for Tulip while evaluating link bounds

| | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| Best | 1 | 3 | 1 | 0.2569 | 0.5375 | 0.3476 |
| Worst | 2 | 0 | 0 | 0.2252 | 0.4722 | 0.3049 |

Table 7-11 - Best overall weight values for Tulip concerning link bounds

Due to the low availability of selection keywords in this application, the best match for effectiveness is the same as for recall. This happens because higher values of recall are here more important to improve this measurement.

### Summary:

We can already infer some conclusions about the bounds of anchors for context description. For all applications, looking into the distance between the curves for the worst and best scenarios, gives an idea about the quality of anchors for context description. Of course, their average values can also be compared. However, the offset of these curves gives a clue as to the quality of the remaining context metadata not analysed in this section.

| | | Best Recall | | | | Best Precision | | | | Best Effectiveness | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Selection | Sentence | Paragraph | Average Recall | Selection | Sentence | Paragraph | Average Precision | Selection | Sentence | Paragraph | Average Recall | Average Precision | Effectiveness |
| Archaeology | Best | 4 | 0 | 1 | 0.2279 | 3 | 1 | 0 | 0.4868 | 4 | 0 | 1 | 0.2279 | 0.4868 | 0.3104 |
| | Worst | 0 | 0 | 0 | 0.0701 | 0 | 0 | 0 | 0.1412 | 0 | 0 | 0 | 0.0701 | 0.1412 | 0.0937 |
| Cell Biology | Best | 4 | 1 | 0 | 0.3152 | 4 | 0 | 0 | 0.6914 | 4 | 0 | 0 | 0.3107 | 0.6914 | 0.4287 |
| | Worst | 0 | 4 | 4 | 0.2876 | 0 | 4 | 4 | 0.5804 | 0 | 4 | 4 | 0.2876 | 0.5804 | 0.3846 |
| French | Best | 1 | 4 | 0 | 0.2304 | 1 | 3 | 0 | 0.4776 | 1 | 4 | 0 | 0.2304 | 0.4735 | 0.3099 |
| | Worst | 0 | 0 | 0 | 0.2070 | 0 | 0 | 0 | 0.4426 | 0 | 0 | 0 | 0.2070 | 0.4426 | 0.2820 |
| French Revolution | Best | 1 | 3 | 1 | 0.2261 | 1 | 2 | 0 | 0.4873 | 1 | 3 | 0 | 0.2256 | 0.4863 | 0.3082 |
| | Worst | 0 | 0 | 0 | 0.2094 | 0 | 0 | 0 | 0.4611 | 0 | 0 | 0 | 0.2094 | 0.4611 | 0.2880 |
| Tulip | Best | 1 | 3 | 1 | 0.2569 | 1 | 3 | 0 | 0.5415 | 1 | 3 | 1 | 0.2569 | 0.5375 | 0.3476 |
| | Worst | 2 | 0 | 0 | 0.2252 | 2 | 0 | 0 | 0.4722 | 2 | 0 | 0 | 0.2252 | 0.4722 | 0.3049 |

Table 7-12 - Summary of the best and worst link bounds metadata weights for context indexing

Broadly speaking, there are two situations that lead to two different approaches to choosing weights:

In the first situation, there are good link anchors. By this, we mean anchors with expressive text that is related to the subject developed at the other end of the link. Under these circumstances, there are two applications: Archaeology and Cell Biology. Here we should always consider high values for the selection of anchors. It is true that the consideration of the sentence and the paragraph surrounding the selections might be also be important to improve recall, but paragraph never improves precision. This is apparently predictable, because by considering information further away from the selection we are also expanding the subject being reported. This improves recall but affects precision. So now, we have to see what is more important for effectiveness. Is it high values of recall or high values of precision? We have to check on other context data to make a choice. Archaeology and Cell biology applications roughly fall within the two limits for our options. The former has a poor abstract level, and the latter has a good one. So, with a poor abstract level, higher values of recall become more important to give high values of effectiveness. On the contrary, with a good abstract level, high precision implies high effectiveness. Of course, it might be possible to find an application that falls between these approaches.

The Archaeology application has many more specific links than the Cell Biology application. However many of the links in Archaeology were navigational. Nevertheless, the remaining links behaved in the same way as the majority of links in Cell Biology. The relative equal performance of these applications, when compared with the number of available specific links on them, demonstrates that links inserted in a sentence, which in turn is inserted in a paragraph, are the best for context description. The importance of generic links will be analysed in the next sub-section.

In the second situation, there is usually poor selection text in anchors. By this, we mean text whose relation to the subject developed at the other end of the link is not clear. It may even mean the absence of text in many anchors. This is definitely what happened with the Tulip application. Here the inclusion of the paragraph was important to improve effectiveness. But for all three applications, Tulip, French Revolution and French, a high weight for sentence is a key factor for higher values of effectiveness. We could say that the system balances out the lower quality of the selection with higher weights of sentence, and even some times with the inclusion of the paragraph.

The French and the French Revolution applications have a similar quality of text anchors. However, in the French application, the distance between the best and worst scenario is greater. This occurs because the French application has more links available per document. The French Revolution and the French applications also have similar values of maximum effectiveness, and this suggests a similar quality of abstract information.

## 7.3 Evaluating link description and link anchor selection according to link type

This evaluation determines how anchor selection and link description are used on two different link types – generic link and specific link. All the different weight combinations between zero and four were considered for these six elements as shown in Table 7-13.

We tied together the weight variation for distinction between the source and the destination anchor of the link. This is shown with the same grey filling. In this way, there were four independent parameters to inspect, giving 625 runs in each test. The majority of the tests took almost one day.

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 0 - 4 | 0 – 4 |
| | Text Anchor as Source: | Selection | 0 - 4 | 0 – 4 |
| | | Sentence | 0 – 0 | 0 - 0 |
| | | Paragraph | 0 – 0 | 0 – 0 |
| | Text Anchor as Destination: | Selection | 0 – 4 | 0 – 4 |
| | | Sentence | 0 – 0 | 0 – 0 |
| | | Paragraph | 0 – 0 | 0 – 0 |
| Abstract information context | Self | Description | 2 - 2 | |
| | | Classification | 2 - 2 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 2 - 2 | 2 - 2 |
| | | Classification | 2 - 2 | 2 - 2 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 1 - 1 | 1 - 1 |
| | Link Outside Classification | | 1 - 1 | 1 - 1 |

Table 7-13 - Weight intervals for selection and description when considering link types

The metadata being analysed are therefore:

- Link Anchor Selection
- Link Description
- Generic Link
- Specific Link

The weight for sentence and paragraph was assigned to zero in all tests. This makes sense given that usually with generic links there is no associated information about the source document or offset. In order to better compare the two link types we have to use similar available information from both links.

We then obtained the following results:

**Archaeology:**

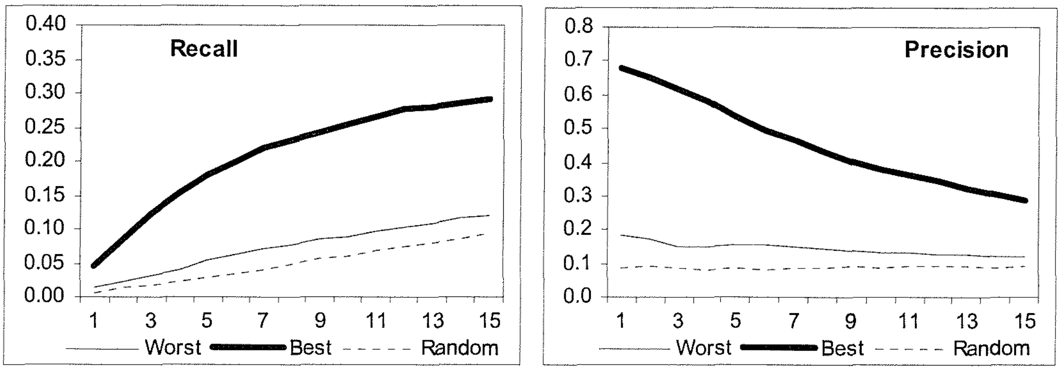


Figure 7-11 - Recall and precision graphs for Archaeology while evaluating link types

| | Generic Link Description | Specific Link Description | Generic Link Anchor Selection | Specific Link Anchor Selection | Average Recall |
|---|---|---|---|---|---|
| Best | 0 | 0 | 4 | 4 | 0.2088 |
| Worst | 0 | 0 | 0 | 0 | 0.0726 |

| | Generic Link Description | Specific Link Description | Generic Link Anchor Selection | Specific Link Anchor Selection | Average Precision |
|---|---|---|---|---|---|
| Best | 0 | 0 | 4 | 2 | 0.4609 |
| Worst | 0 | 0 | 0 | 0 | 0.1453 |

Table 7-14 - Recall, precision and weight values for Archaeology concerning link types

The graphs and values show a big distance between the worst and best situation. The achieved recall and precision is worse than the one found while testing the bounds of anchors in the Archaeology application. This happens here because paragraph and sentence information is not used to improve recall and precision respectively. However, a clearer distinction is achieved between the weight of the link description and the anchor selections. This is not surprising. In the Archaeology application, link description is equal to the destination document description, and with a poor document description, its data becomes less relevant. The consideration of the neighbour description metadata with a fixed weight value of two during this test justifies the assignment of the zero weight to link description. The worst scenario obtained here also performs better for the same reasons. With less link description (on the previous test the weight was fixed to two), there is a better recall and precision.

As regards precision, a distinction between generic links and specific links is also observed. A high weight for the former and a lower weight for the latter give the best precision. It should be noted that in the Archaeology application, there are many specific links per document supporting navigation, in a menu-like manner, and this is the reason why much of this information is repeated in many documents. It is clear that this does not help precision.

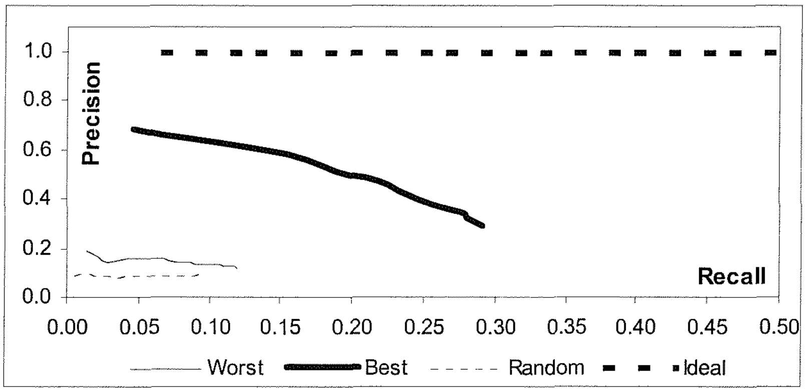


Figure 7-12 - Recall/precision graph for Archaeology while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | | | |
| Best | 0 | 0 | 4 | 4 | 0.2088 | 0.4579 | 0.2868 |
| Worst | 0 | 0 | 0 | 0 | 0.0726 | 0.1453 | 0.0968 |

Table 7-15 - Best overall weight values for Archaeology concerning link types

Again, the best match for effectiveness coincided with the best for recall. In this application, the best solution does not distinguish generic links from specific links. The low availability of generic links might be the reason why they do not improve recall in a different way than specific links. However, they provide higher values of precision. Nevertheless, since the system demands high recall values, due to its abstract level, precision becomes a minor factor.

### Cell Biology:

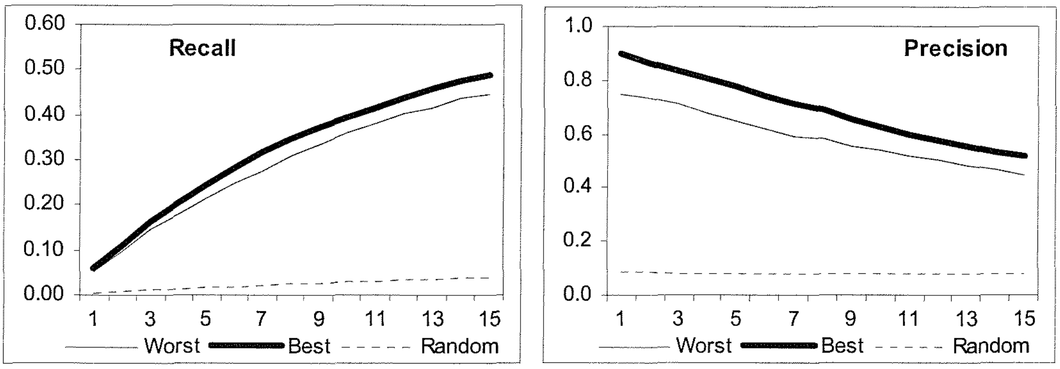


Figure 7-13 - Recall and precision graphs for Cell Biology while evaluating link types

| | Link Description | | Link Anchor Selection | | |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | Average Recall |
| Best | 2 | 2 | 4 | 4 | 0.3107 |
| Worst | 0 | 0 | 0 | 0 | 0.2878 |

| | Link Description | | Link Anchor Selection | | |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | Average Precision |
| Best | 2 | 2 | 2 | 4 | 0.7013 |
| Worst | 0 | 0 | 0 | 0 | 0.5894 |

Table 7-16 - Recall, precision and weight values for Cell Biology concerning link types

A close inspection of recall values shows, as in the previous application, that there is a lower maximum recall than while testing the bounds of link anchors. In addition, it can be observed that the worst scenario is a little better. Since the sentence should be considered in the best approach for recall, this maximum cannot be achieved because these factors are not taken into account during this test. The sentence, and also the paragraph, when considered gives the worst approach for recall and then this minimum is not achieved for the same reasons.

More revealing are the achieved results for precision. In the best precision scenario, there is a difference between the weighting of generic and specific links. However, generic links do not give as much precision as could be expected. This is justifiable if we consider the nature of the majority of generic links in the Cell Biology application. These links implement a glossary, and so are only available in a few documents. This biases the quality of context description towards these documents, which contributes to their ranking in higher positions. This does not help recall, but affects precision in a more negative way.

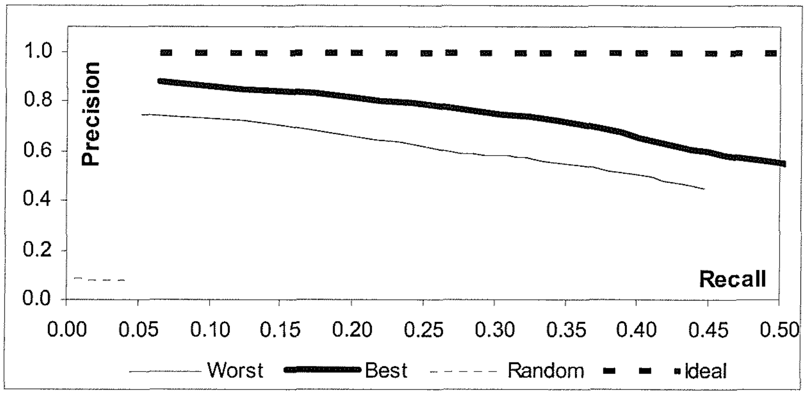


Figure 7-14 - Recall/precision graph for Cell Biology while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | | | |
| Best | 2 | 2 | 3 | 4 | 0.3105 | 0.6988 | 0.4300 |
| Worst | 0 | 0 | 0 | 0 | 0.2878 | 0.5894 | 0.3867 |

Table 7-17 - Best overall weight values for Cell Biology concerning link types

The best weight for effectiveness is between the best for precision and recall. Here we give a lower weight to generic links than when we were testing link anchor bounds. There is a balance between precision and recall to achieve the best performance.

## French:

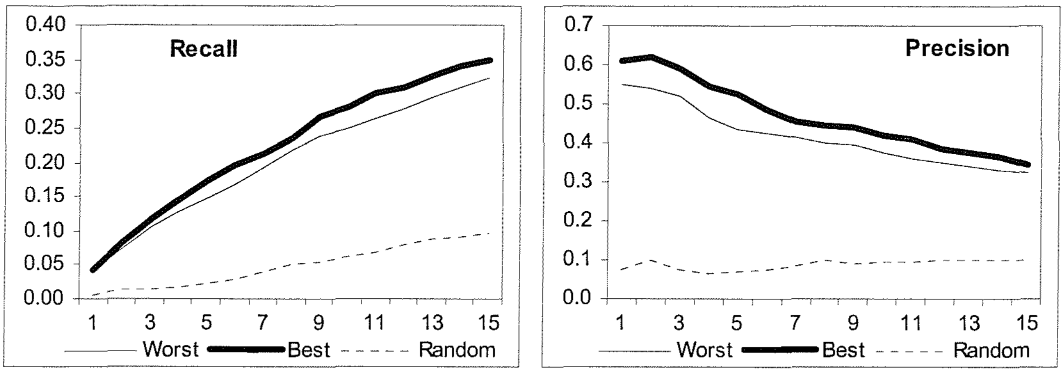


Figure 7-15 - Recall and precision graphs for French while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 2 | 2 | 2 | 1 | 0.2251 |
| Worst | 0 | 0 | 0 | 0 | 0.1996 |

| | Link Description | | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 1 | 1 | 2 | 1 | 0.4679 |
| Worst | 0 | 0 | 0 | 0 | 0.4327 |

Table 7-18 - Recall, precision and weight values for French concerning link types

This application does not have that many generic links, but even so there is a distinction between generic and specific links. For the French application, generic links do not support any glossary and, as in Archaeology, this might be the reason why they have a higher weight. Nevertheless, in this application, generic links are also used in a different way than the standard generic links. Another reason for having a higher weight for generic links might be the lower quality of selections on specific link anchors.

The best precision requires a lower weight for link description but, apart from that, the distinction between generic and specific links is the same as in recall.

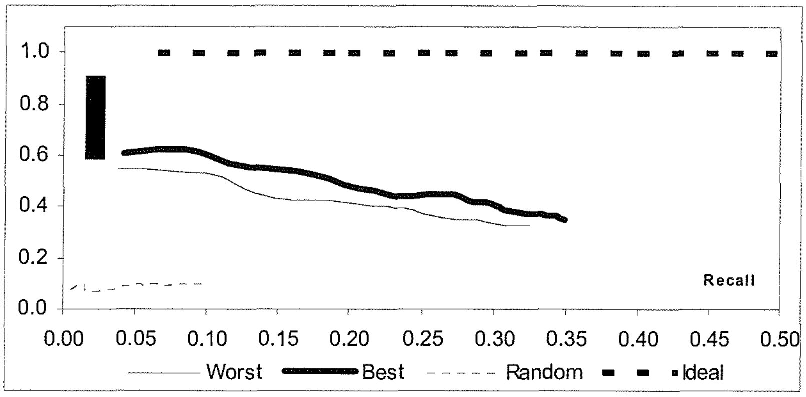


Figure 7-16 - Recall/precision graph for French while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | | | |
| Best | 2 | 2 | 2 | 1 | 0.2251 | 0.4677 | 0.3039 |
| Worst | 0 | 0 | 0 | 0 | 0.1996 | 0.4327 | 0.2732 |

Table 7-19 - Best overall weight values for French concerning link types

As we found while evaluating the bounds of link anchors, the system demands higher values of recall in this application. This leads the weights for effectiveness to be the same as for recall.

### French Revolution:

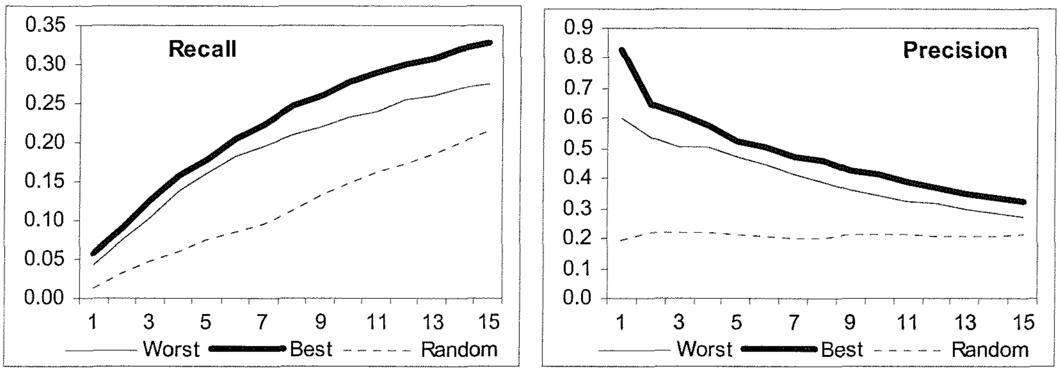


Figure 7-17 - Recall and precision graphs for French Revolution for evaluation link types

| | Link Description | | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 1 | 2 | 1 | 2 | 0.2242 |
| Worst | 0 | 0 | 0 | 0 | 0.1904 |

| | Link Description | | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 0 | 2 | 1 | 2 | 0.4880 |
| Worst | 0 | 0 | 0 | 0 | 0.4095 |

Table 7-20 - Recall, precision and weight values for French Revolution concerning link types

The French Revolution application has the worst performance for generic links. All these links simply implement a glossary file, and that is different from what occurs in the Cell Biology or Tulip applications. The fact that only one file is indexed with information from generic links prejudices recall and precision, since this file will have a larger set of keywords. This is the only application with a distinction between the weight for link description on generic and specific links. Here, generic link description is usually an expansion of anchor selection. Hence, the duplication of keywords in these two elements might also contribute to the decrease in the weight for each.

The worst recall and precision is also lower in this test, because link description is not used in the scenario.

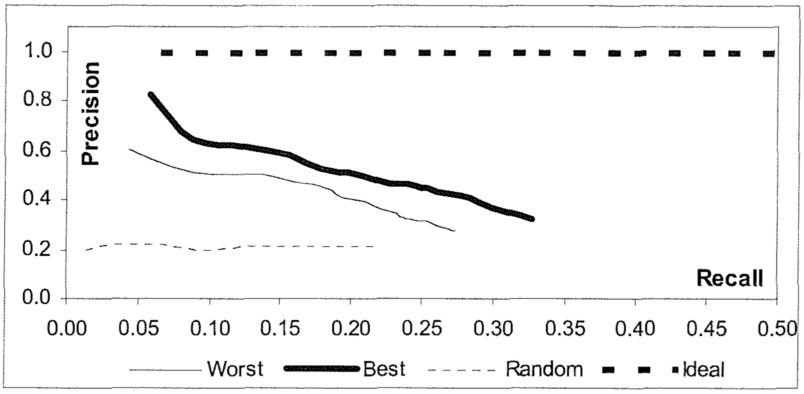


Figure 7-18 - Recall/precision graph for French Revolution while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | | | |
| Best | 1 | 2 | 1 | 2 | 0.2242 | 0.4870 | 0.3070 |
| Worst | 0 | 0 | 0 | 0 | 0.1904 | 0.4095 | 0.2599 |

Table 7-21 - Best overall weight values for French Revolution concerning link types

The weights, when considering effectiveness, are the same as when considering recall. However, effectiveness, along with recall and precision, is not significantly lower than when testing link bounds. There the system considered sentence information in the best scenario for context indexing, but the weighting for link description was fixed at two. Here sentence information is not considered, and link description has a lower weight. The absence of one element is balanced in a way by a more correct weight for the other.

## Tulip:

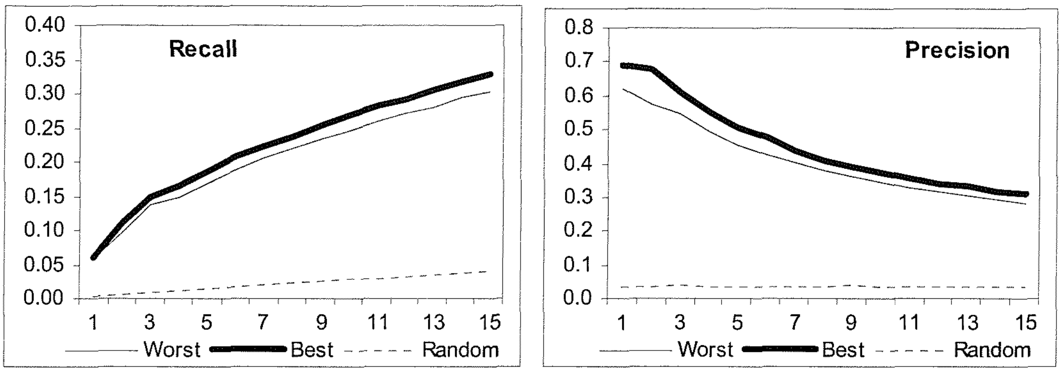


Figure 7-19 - Recall and precision graphs for Tulip for evaluation link types

| | Link Description | | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 2 | 2 | 4 | 1 | 0.2279 |
| Worst | 0 | 0 | 0 | 2 | 0.2199 |

| | Link Description | | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | |
| Best | 1 | 2 | 3 | 1 | 0.4803 |
| Worst | 0 | 0 | 0 | 2 | 0.4580 |

Table 7-22 - Recall, precision and weight values for Tulip concerning link types

The quality of anchor selections for specific links is quite low, and this can be observed again by the weight attributed to them in this application. Furthermore, the worst scenario is when specific link anchor selection is considered alone, with a weight of two.

The majority of generic links implement a glossary, but about 1/10 do not. This turns out to give a much higher weight to generic link selections, even when the remaining ones are used to implement that glossary. However, the link description weight for generic links should not have such a high value as for specific links, where we value precision. This is probably due to the repetition of the same word, e.g. 'Glossary', in all glossary generic links descriptions.

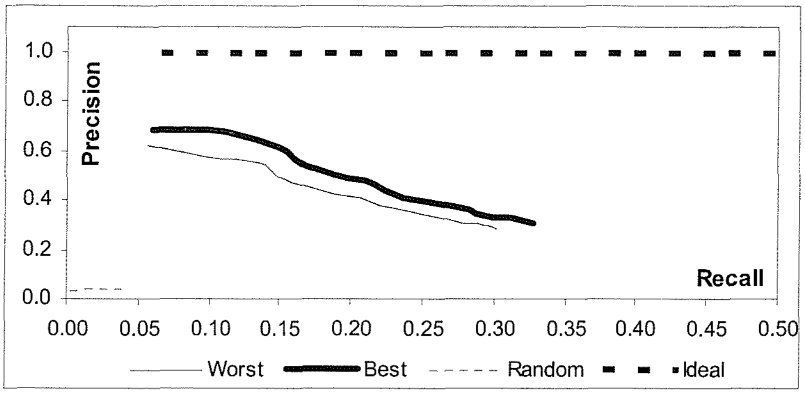


Figure 7-20 - Recall/precision graph for Tulip while evaluating link types

| | Link Description | | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| | Generic | Specific | Generic | Specific | | | |
| Best | 2 | 2 | 4 | 1 | 0.2279 | 0.4785 | 0.3087 |
| Worst | 0 | 0 | 0 | 2 | 0.2199 | 0.4580 | 0.2971 |

Table 7-23 - Best overall weight values for Tulip concerning link types

The absence of the sentence and paragraph elements entails a lower performance for this test. The best value for effectiveness, and also recall and precision, is much lower in this test. This confirms our statement about the quality of specific link selections in the Tulip application.

### Summary:

The differences between specific links and generic links can already be considered in studying their importance for context description building. Broadly speaking, in all applications generic links are usually used to evolve or expand a concept or idea. An exception is the French application, where they are also used to find resources and/or activities. Specific links have a more diverse application. Besides their use in relating concepts, they also structure applications for navigation, menu choosing and resource discovering.

| | | Best Recall | | | | | Best Precision | | | | | Best Effectiveness | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Link | Description | Link Anchor | Selection | Average Recall | Link | Description | Link Anchor | Selection | Average Precision | Link | Description | Link Anchor | Selection | Average Recall | Average Precision | Effectiveness |
| | | Generic | Specific | Generic | Specific | | Generic | Specific | Generic | Specific | | Generic | Specific | Generic | Specific | | | |
| Archaeology | Best | 0 | 0 | 4 | 4 | 0.2088 | 0 | 0 | 4 | 2 | 0.4609 | 0 | 0 | 4 | 4 | 0.2088 | 0.4579 | 0.2868 |
| | Worst | 0 | 0 | 0 | 0 | 0.0726 | 0 | 0 | 0 | 0 | 0.1453 | 0 | 0 | 0 | 0 | 0.0726 | 0.1453 | 0.0968 |
| Cell Biology | Best | 2 | 2 | 4 | 4 | 0.3107 | 2 | 2 | 2 | 4 | 0.7013 | 2 | 2 | 3 | 4 | 0.3105 | 0.6988 | 0.4300 |
| | Worst | 0 | 0 | 0 | 0 | 0.2878 | 0 | 0 | 0 | 0 | 0.5894 | 0 | 0 | 0 | 0 | 0.2878 | 0.5894 | 0.3867 |
| French | Best | 2 | 2 | 2 | 1 | 0.2251 | 1 | 1 | 2 | 1 | 0.4679 | 2 | 2 | 2 | 1 | 0.2251 | 0.4677 | 0.3039 |
| | Worst | 0 | 0 | 0 | 0 | 0.1996 | 0 | 0 | 0 | 0 | 0.4327 | 0 | 0 | 0 | 0 | 0.1996 | 0.4327 | 0.2732 |
| French Revolution | Best | 1 | 2 | 1 | 2 | 0.2242 | 0 | 2 | 1 | 2 | 0.4880 | 1 | 2 | 1 | 2 | 0.2242 | 0.4870 | 0.3070 |
| | Worst | 0 | 0 | 0 | 0 | 0.1904 | 0 | 0 | 0 | 0 | 0.4095 | 0 | 0 | 0 | 0 | 0.1904 | 0.4095 | 0.2599 |
| Tulip | Best | 2 | 2 | 4 | 1 | 0.2279 | 1 | 2 | 3 | 1 | 0.4803 | 2 | 2 | 4 | 1 | 0.2279 | 0.4785 | 0.3087 |
| | Worst | 0 | 0 | 0 | 2 | 0.2199 | 0 | 0 | 0 | 0 | 0.4580 | 0 | 0 | 0 | 2 | 0.2199 | 0.4580 | 0.2971 |

Table 7-24 - Summary of the best and worst link type metadata weights for context indexing

Looking at applications according to their generic links, we can distinguish them by determining how they aggregate those links. They can be used to link different distinguishable documents, to link to a unique glossary file, or both.

In the Archaeology and French applications, we have the first situation. They have a small number of generic links that do not support any glossary file. Under these circumstances, these links prove to be more precise than specific links. In these applications, specific links have a broader use apart from relating concepts, which can make generic links more suitable for descriptor construction. So, for achieving better precision, more weight should always be assigned to them. However, to balance recall and precision we might have to assign the same weight to both types, where the remaining context information is scarce. In the Archaeology application, due to the low quality of the abstract level, the best performance for effectiveness requires a similar weight for both specific and generic links. But in the French application, better performance is achieved when more weight is assigned to generic links.

The French Revolution application just has glossary generic links and in that way, they prove to be less efficient for constructing descriptors. This is understandable, since their context information will only be used to describe one file. This is bad for recall and worse for precision. In a situation like this, we should assign lower weight values to generic links. Here is also the only example where less weight should be assigned to link description on generic links for the best overall performance, due to the relation between link description and anchor selection.

The other applications, Cell Biology and Tulip, have a considerable amount of glossary generic links and a small amount of non-glossary ones. It is then trickier to infer the importance of generic links. In the Cell Biology application, where we have good specific links, it seems that generic link anchor selection is less relevant. However, in the Tulip application, where there are few text anchors on specific links, generic links have a more important role.

In summary, we can say that generic links on glossary files are not useful, always decrease precision and sometimes recall. This could be puzzling, because we tend to think that generic links relate information in a more content oriented way, but these links bias the retrieval of only one file, and that is the problem. Nevertheless, if they are used outside the glossary file, these links seem to contain better information for context descriptors. Unfortunately, not many of those links are available to enable us to make stronger statements. However, their nature makes us understand why it works in this way. It is not the mechanism used to find those links that makes the difference; what is important is the way they usually relate information, and generic links are usually used to relate concepts, or to expand ideas, while specific links are usually used for a broader range of purposes.

## 7.4 Evaluating link description and link anchor selection when the described document is either the source or the destination of the link

This evaluation determines how anchor selection is used for different link ends, i.e., when the described document is either the source or the destination of the link. Link description is also evaluated. Therefore, all the different weight combinations between zero and four were considered for these six elements as seen in Table 7-25.

The weight variations for the distinction between generic and specific link were tied together. This is shown in the table with the same grey filling. In this way, there were three independent parameters to inspect, and this gave 125 runs in each test. The majority of the tests took several hours.

The metadata being analysed are:

- Link description
- Link Source Anchor Selection
- Link Destination Anchor Selection

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 0 - 4 | 0 – 4 |
| | Text Anchor as Source: | Selection | 0 - 4 | 0 – 4 |
| | | Sentence | 0 - 0 | 0 – 0 |
| | | Paragraph | 0 – 0 | 0 – 0 |
| | Text Anchor as Destination: | Selection | 0 - 4 | 0 – 4 |
| | | Sentence | 0 – 0 | 0 – 0 |
| | | Paragraph | 0 – 0 | 0 – 0 |
| Abstract information context | Self | Description | 2 - 2 | |
| | | Classification | 2 - 2 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 2 - 2 | 2 - 2 |
| | | Classification | 2 - 2 | 2 - 2 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 1 - 1 | 1 - 1 |
| | Link Outside Classification | | 1 - 1 | 1 - 1 |

Table 7-25 - Weight intervals for selection and description when considering link anchor ends

For each of the applications, the retrieval performance graphs and values were obtained along with the best weighting policy for that situation:

### Archaeology:

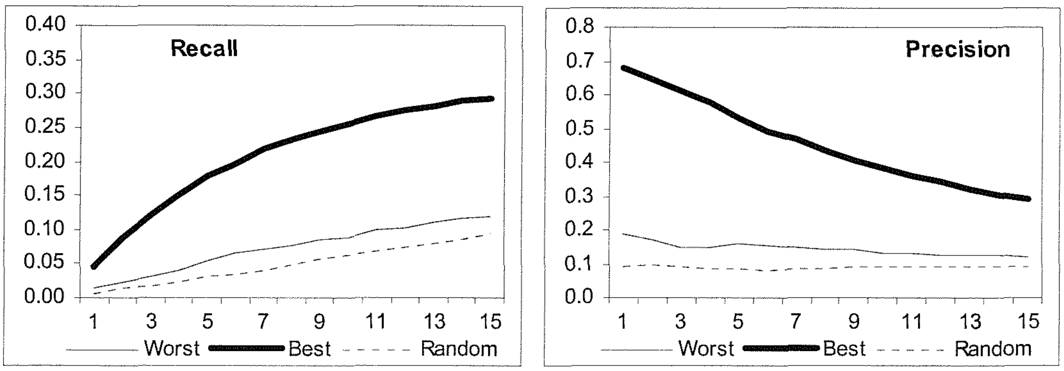


Figure 7-21 - Recall and precision graphs for Archaeology, evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 0 | 4 | 0 | 0.2088 |
| Worst | 0 | 0 | 0 | 0.0726 |

| | Link Description | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 0 | 3 | 0 | 0.4579 |
| Worst | 0 | 0 | 0 | 0.1453 |

Table 7-26 - Recall, precision and weight values for Archaeology, evaluating link anchor ends

The results obtained for this test are quite predictable. On a close inspection of the destination selections of links in Archaeology, no text information was found. Therefore, its consideration is irrelevant to context description. The achieved results for recall are then equal to the ones obtained while testing link types. For precision, there is however a decrease in performance, since link types are not differentiated.

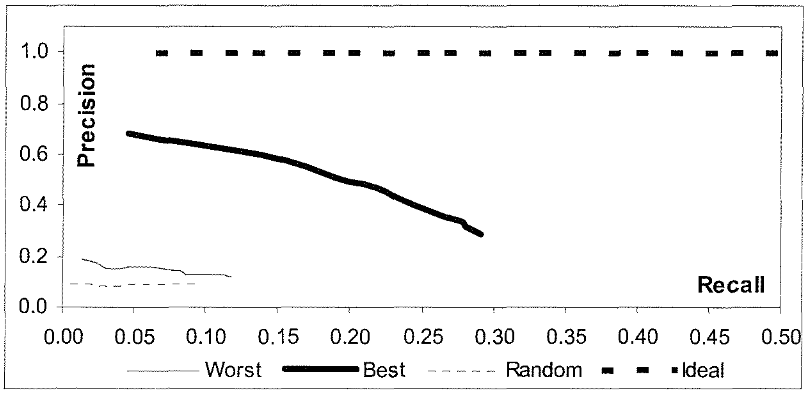


Figure 7-22 - Recall/precision graph for Archaeology while evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| | | Source | Destination | | | |
| Best | 0 | 4 | 0 | 0.2088 | 0.4575 | 0.2868 |
| Worst | 0 | 0 | 0 | 0.0726 | 0.1453 | 0.0968 |

Table 7-27 - Best overall weight values for Archaeology concerning link anchor ends

The results are the same as the ones obtained while testing link types. The only obvious conclusion to make with this application is that if the information is not there, we should assign a weight of zero to it. In this case, we do not have destination selection anchors, so they have a weight of zero.

### Cell Biology:

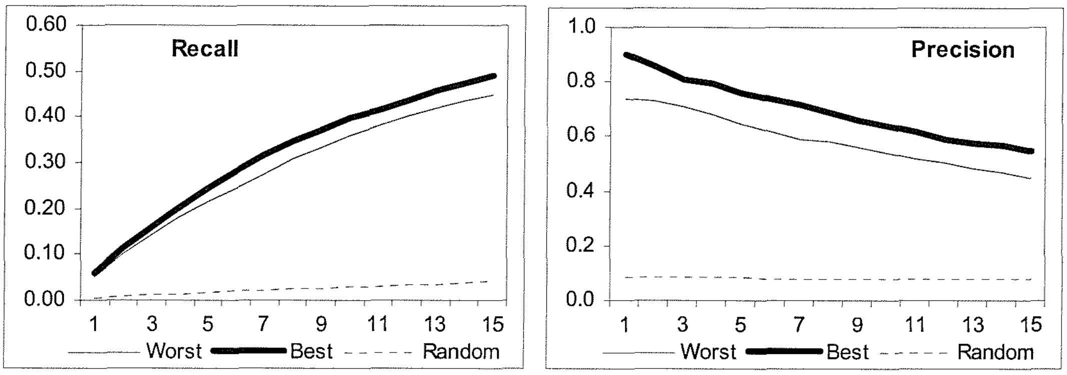


Figure 7-23 - Recall and precision graphs for Cell Biology, evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 4 | 1 | 0.3110 |
| Worst | 0 | 0 | 4 | 0.2875 |

| | Link Description | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 4 | 1 | 0.6919 |
| Worst | 0 | 0 | 4 | 0.5882 |

Table 7-28 - Recall, precision and weight values for Cell Biology, evaluating link anchor ends

There are only three links with destination selections in the Cell Biology application. This makes its weight quite low, and if used alone, it gives the worst performance. Its consideration however does not improve recall that much. While testing link types, similar results were obtained. Here, by comparison, there is also a similar, but lower precision, since we do not distinguish between link types.

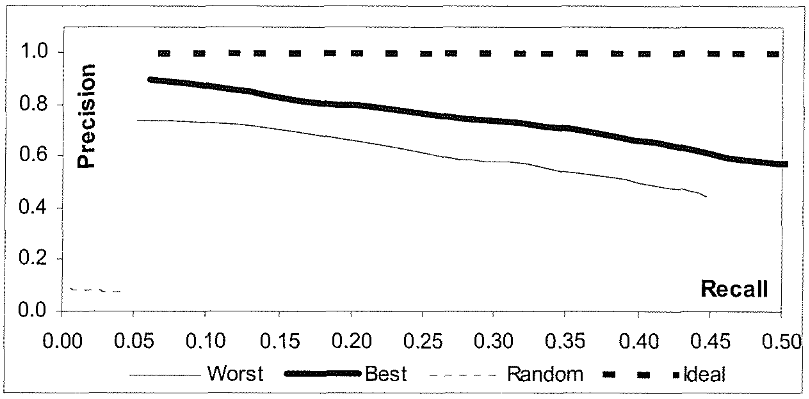


Figure 7-24 - Recall/precision graph for Cell Biology while evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| | | Source | Destination | | | |
| Best | 2 | 4 | 1 | 0.3110 | 0.6919 | 0.4291 |
| Worst | 0 | 0 | 4 | 0.2875 | 0.5882 | 0.3862 |

Table 7-29 - Best overall weight values for Cell Biology concerning link anchor ends

Again, insufficient information is available in this application to determine whether destination anchors are relevant or not for context indexing.

### French:

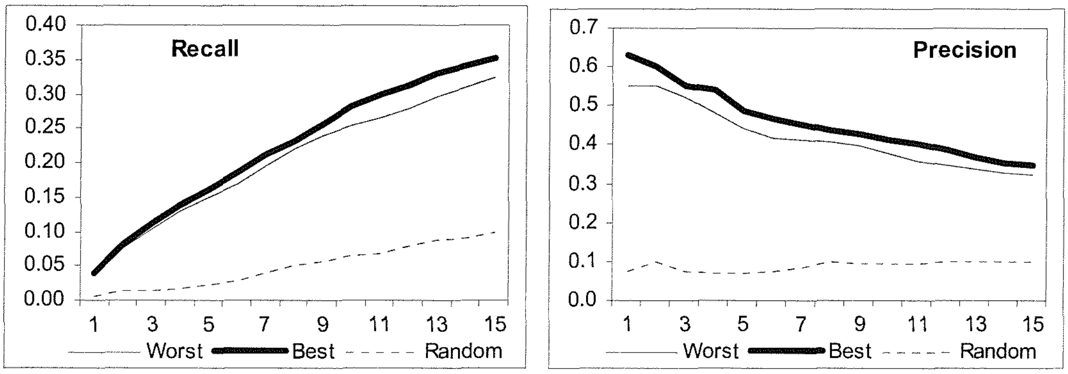


Figure 7-25 - Recall and precision graphs for French, evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 2 | 1 | 0.2256 |
| Worst | 0 | 0 | 0 | 0.1996 |

| | Link Description | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 2 | 1 | 0.4671 |
| Worst | 0 | 0 | 0 | 0.4327 |

Table 7-30 - Recall, precision and weight values for French, evaluating link anchor ends

In this application, there are a reasonable number of links where the destination anchor selection is present. Nevertheless, it seems that they are less relevant than the source anchors. This is partly due to their lower availability. Another reason is their similarity with the source anchor selection.

The worst performance is the same as when testing link types.

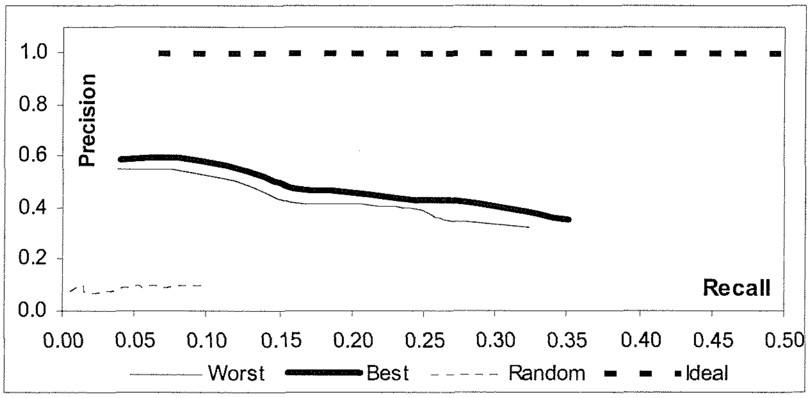


Figure 7-26 - Recall/precision graph for French while evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| | | Source | Destination | | | |
| Best | 2 | 2 | 1 | 0.2256 | 0.4671 | 0.3042 |
| Worst | 0 | 0 | 0 | 0.1996 | 0.4327 | 0.2732 |

Table 7-31 - Best overall weight values for French concerning link anchor ends

The retrieval values are quite similar to the ones obtained when testing link types. In this application, considering different weights for different link anchor ends has the same effect as considering different weights for different link types.

**French Revolution:**

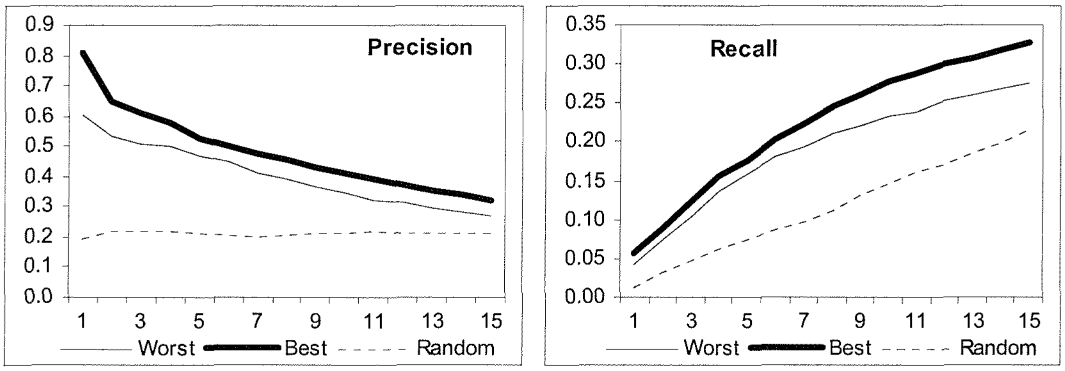


Figure 7-27 - Recall and precision graphs for French Revolution, evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 2 | 0 | 0.2240 |
| Worst | 0 | 0 | 0 | 0.1904 |

| | Link Description | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 1 | 2 | 0 | 0.4875 |
| Worst | 0 | 0 | 0 | 0.4095 |

Table 7-32 - Recall, precision and weight values for French Revolution, evaluating link anchor ends

Destination anchor selection is not present for specific links but only for generic links. However, since all these links have the same source document, this does not improve retrieval. That is the reason why this parameter should not be considered. Again, the non-distinction between generic and specific link in this test is balanced by the distinction of link ends. That is the reason why recall and precision values are close to the ones obtained while testing link types.

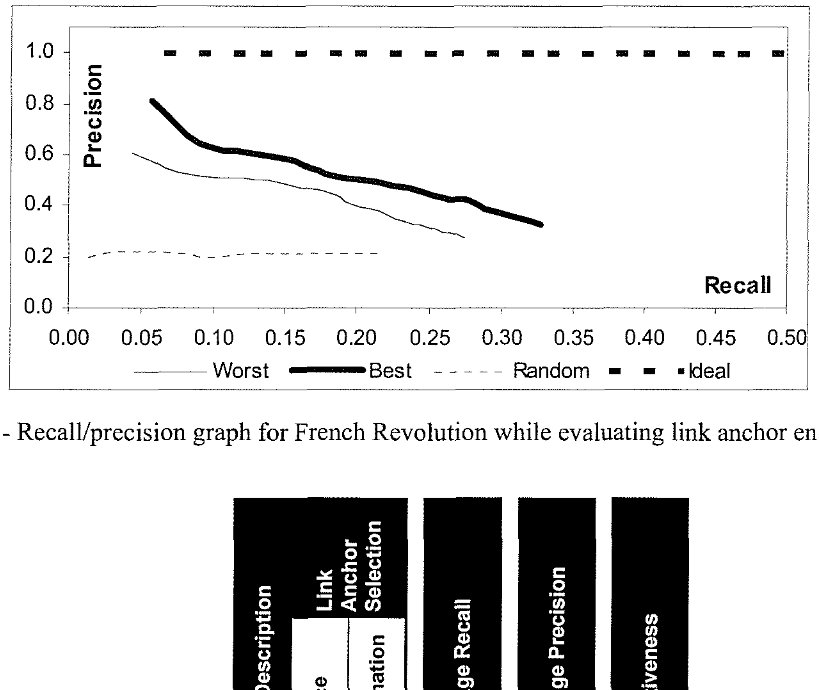
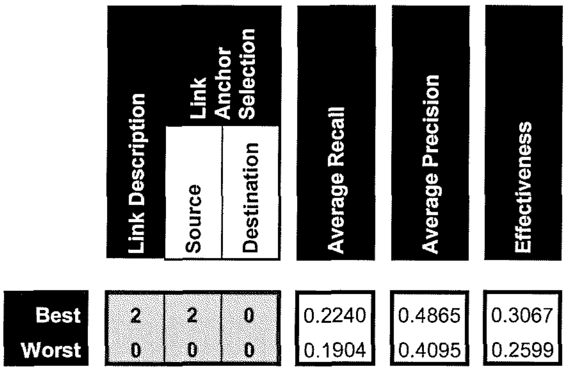


Figure 7-28 - Recall/precision graph for French Revolution while evaluating link anchor ends



| | Link Description | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| | | Source | Destination | | | |
| Best | 2 | 2 | 0 | 0.2240 | 0.4865 | 0.3067 |
| Worst | 0 | 0 | 0 | 0.1904 | 0.4095 | 0.2599 |

Table 7-33 - Best overall weight values for French Revolution concerning link anchor ends

Again, the best effectiveness is quite similar to the one obtained when testing link types. Here, excluding the destination anchor selection produces the same outcome as before, when considering lower weights for generic links.

The worst performance is the same as when testing link types.

## Tulip:

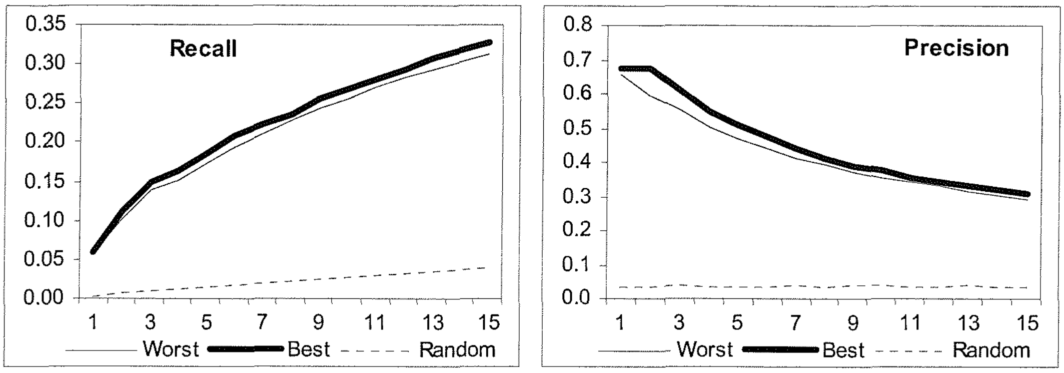


Figure 7-29 - Recall and precision graphs for Tulip, evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 2 | 2 | 0 | 0.2263 |
| Worst | 0 | 2 | 4 | 0.2195 |

| | Link Description | Link Anchor Selection | | Average Precision |
|---|---|---|---|---|
| | | Source | Destination | |
| Best | 1 | 2 | 0 | 0.4773 |
| Worst | 0 | 2 | 4 | 0.4573 |

Table 7-34 - Recall, precision and weight values for Tulip, evaluating link anchor ends

There are almost no destination anchor selections in this application. In addition, the available ones simply possess a number, or a letter. This gives a slight disadvantage to this end of the links, even given that specific links in this application do not have much relevant textual information in source anchors. This is proved by the worst achieved performance, quite similar to the one obtained when testing link types.

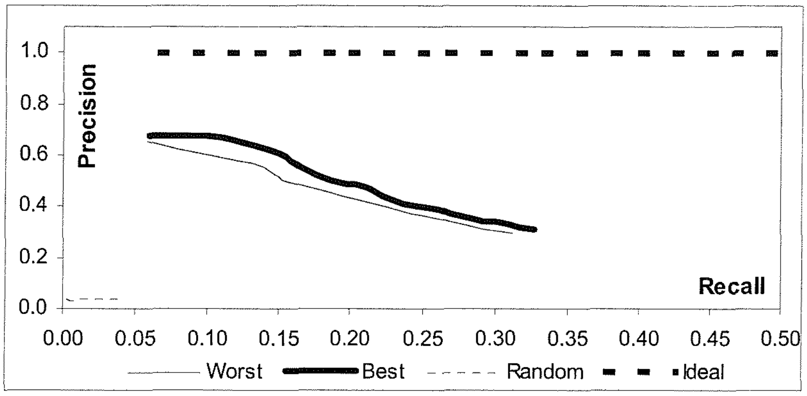


Figure 7-30 - Recall/precision graph for Tulip while evaluating link anchor ends

| | Link Description | Link Anchor Selection | | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|
| | | Source | Destination | | | |
| Best | 2 | 2 | 0 | 0.2263 | 0.4769 | 0.3069 |
| Worst | 0 | 2 | 4 | 0.2195 | 0.4573 | 0.2966 |

Table 7-35 - Best overall weight values for Tulip concerning link anchor ends

The best performance obtained here is lower than the one obtained while testing link types. Here there is no distinction between link types and this decreased the effectiveness. The discrimination of link ends was insufficient to balance the distinction of link types. This is due to the low availability and quality of destination anchor selections.

### Summary:

In summary, when testing link ends a lack of available information prevented strong conclusions. Generally, the weight for destination selection should be lower than for source selection. This is not due to the nature of the link end in itself. It just happens that authors did not choose to use this facility. They could, for instance, use this facility to highlight the sentence or paragraph where the relevant information in the destination document could be found. However, this did not happen.

| | | Best Recall | | | | Best Precision | | | | Best Effectiveness | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Link Anchor Selection | | | | Link Anchor Selection | | | | Link Anchor Selection | | | | |
| | | Link Description | Source | Destination | Average Recall | Link Description | Source | Destination | Average Precision | Link Description | Source | Destination | Average Recall | Average Precision | Effectiveness |
| Archaeology | Best | 0 | 4 | 0 | 0.2088 | 0 | 3 | 0 | 0.4579 | 0 | 4 | 0 | 0.2088 | 0.4575 | 0.2868 |
| | Worst | 0 | 0 | 0 | 0.0726 | 0 | 0 | 0 | 0.1453 | 0 | 0 | 0 | 0.0726 | 0.1453 | 0.0968 |
| Cell Biology | Best | 2 | 4 | 1 | 0.3110 | 2 | 4 | 1 | 0.6919 | 2 | 4 | 1 | 0.3110 | 0.6919 | 0.4291 |
| | Worst | 0 | 0 | 4 | 0.2875 | 0 | 0 | 4 | 0.5882 | 0 | 0 | 4 | 0.2875 | 0.5882 | 0.3862 |
| French | Best | 2 | 2 | 1 | 0.2256 | 2 | 2 | 1 | 0.4671 | 2 | 2 | 1 | 0.2256 | 0.4671 | 0.3042 |
| | Worst | 0 | 0 | 0 | 0.1996 | 0 | 0 | 0 | 0.4327 | 0 | 0 | 0 | 0.1996 | 0.4327 | 0.2732 |
| French Revolution | Best | 2 | 2 | 0 | 0.2240 | 1 | 2 | 0 | 0.4875 | 2 | 2 | 0 | 0.2240 | 0.4865 | 0.3067 |
| | Worst | 0 | 0 | 0 | 0.1904 | 0 | 0 | 0 | 0.4095 | 0 | 0 | 0 | 0.1904 | 0.4095 | 0.2599 |
| Tulip | Best | 2 | 2 | 0 | 0.2263 | 1 | 2 | 0 | 0.4773 | 2 | 2 | 0 | 0.2263 | 0.4769 | 0.3069 |
| | Worst | 0 | 2 | 4 | 0.2195 | 0 | 2 | 4 | 0.4573 | 0 | 2 | 4 | 0.2195 | 0.4573 | 0.2966 |

Table 7-36 - Summary of the best and worst metadata weights concerning link anchor ends

The only applications with a reasonable number of destination anchor selections are French and French Revolution. The difference between them is how they aggregate the links. In the former, the links have their sources in distinct documents. In a situation like this, we should consider destinations. Care should however be taken, because low availability and similarity with source selections might dictate a lower weight, as was the case in the French application. In the French Revolution application, all the links have the same source document. In a situation like this, we should not use destination anchor selection.

## 7.5 Conclusions

In this chapter three different evaluations have been made for each application, considering related metadata extracted from links at document level. This was made with the purpose of finding the best relative weighting of these related metadata, for the proper context indexing and searching of documents and/or their parts.

In the first evaluation (section 7.2) it was determined which were the bounds of link anchors that should be considered for context descriptor construction on different applications. For these applications different link editing styles were already found as described in Chapter 5, and this lead to different type of weighting consideration for the evaluated metadata. Broadly speaking we basically have two types of applications. Applications with links with expressive text anchors related to content of the other end of the link achieve better results with the selection as the bound of links. The other applications need instead the consideration of sentence and paragraph information. A more detailed analysis was developed in section 7.2.

The second evaluation (section 7.3) discovered the differences between the generic and specific type of links. Depending on whether generic links are aggregated together or not in a glossary, different weightings were obtained. When generic links are not aggregated in a glossary, they show a tendency to be more relevant for context indexing than average specific links. However if generic links are effectively aggregated in a glossary their significance for context indexing is quite reduced. A more detailed analysis was developed in section 7.3, presenting some strategies for weight assignment on similar applications.

The results obtained in the third evaluation (section 7.4) were less satisfactory. There we intended to evaluate the differences between information extracted from the source anchors and from the destination anchors. However the lack of destination anchors impeded the proper consideration of this information. The weight for destination should generally be low, however it was possible to find weighting differences by knowing if destination anchors are aggregated in one file or not. If a glossary aggregates together destination anchors they should not be considered for context indexing.

After these evaluations we need to further evaluate metadata from the abstract level. This will be described in next chapter (Chapter 8). There, we will also evaluate some link information along with other metadata, to allow us to relate the obtained weights in Chapter 9

# Chapter 8 - Considering abstract level information for multimedia retrieval

## 8.1 Introduction

The set of tests in this chapter will evaluate different metadata extracted from the abstract level. We will mainly be interested in proper context indexing construction. Chapter 5 laid the foundation for the evaluation methodology adopted here. It will be necessary to find the relative significance of the different metadata considered for the construction of context descriptors

For reasons explained in Chapter 5, it is once again necessary to split the evaluation at this level in such a way that only a few metadata elements (this could be between two and four) are considered at one time. As with the evaluations adopted for link context information at document level described in Chapter 7, here related metadata is also aggregated in each individual evaluation, but so that we have some common metadata elements share between all evaluations. Three evaluations have been undertaken at this level.

The first evaluation (section 8.2) determines the way in which different abstract metadata elements should be weighted for the best context description construction. The document description of it itself and its neighbours is evaluated along with the classification of the document itself and its neighbours.

The second evaluation (section 8.3) attempts to find the relation between parts of the abstract level information and the link context information. For this reason document description and classification of itself is evaluated along with anchor selection and description of links 'leaving' and 'arriving' at the document.

In the third evaluation (section 8.4) we judge whether it is relevant to differentiate between links that connect documents in the same classification from links connecting documents in distinct classifications. These are the only elements being considered here since there is no need to relatively compare the obtained weights with other metadata elements.

At the end of each individual evaluation we will summarise the obtained results, comparing the obtained weights, with the style adopted for the construction of the applications. The best strategy to assigned weights to possible similar applications is also attempted.

The comparisons of the obtained weights for document description and classification in the first two tests will allow the proper consideration of the weights for all the evaluated metadata to achieve the best weighting strategy, in Chapter 9.

## 8.2 Evaluating description and classification of documents and their neighbours, to build context descriptors.

This test analyses the relations between the different parameters available at the abstract level. We did not consider the use of the keyword metadata attribute because the authors of the test applications did not implement this feature.

All the different weight combinations between zero and four were therefore considered for six elements (see Table 8-1). The elements considered are Document Description, Document Classification, Neighbour Document Description and Neighbour Document Classification. However, the weight variations for the distinction between generic and specific links were tied together (the same grey filling in the table). In this way, there were four independent parameters to inspect, and this gave 625 runs for each test. The majority of the tests took over one day to run.

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 2 - 2 | 2 – 2 |
| | Text Anchor as Source: | Selection | 2 - 2 | 2 – 2 |
| | | Sentence | 2 - 2 | 2 – 2 |
| | | Paragraph | 2 - 2 | 2 – 2 |
| | Text Anchor as Destination: | Selection | 2 - 2 | 2 – 2 |
| | | Sentence | 2 - 2 | 2 – 2 |
| | | Paragraph | 2 - 2 | 2 – 2 |
| Abstract information context | Self | Description | 0 - 4 | |
| | | Classification | 0 - 4 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 0 - 4 | 0 - 4 |
| | | Classification | 0 - 4 | 0 - 4 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 1 - 1 | 1 - 1 |
| | Link Outside Classification | | 1 - 1 | 1 - 1 |

Table 8-1 - Weight intervals for description and classification of documents and their neighbours

In summary, the metadata being analysed are:

- Document description
- Document classification
- Neighbour document description
- Neighbour document classification

All the weights for link information were taken as equal to two, in order to have a means of comparison between the applications. For better classification information consideration, all entries in the tree relating to file formats were removed, as we foresaw that this information is not relevant and that it could influence the context indexing negatively.

For each of the applications, the retrieval performance graphs and values were obtained along with the best weighting policy, for that situation:

## Archaeology:

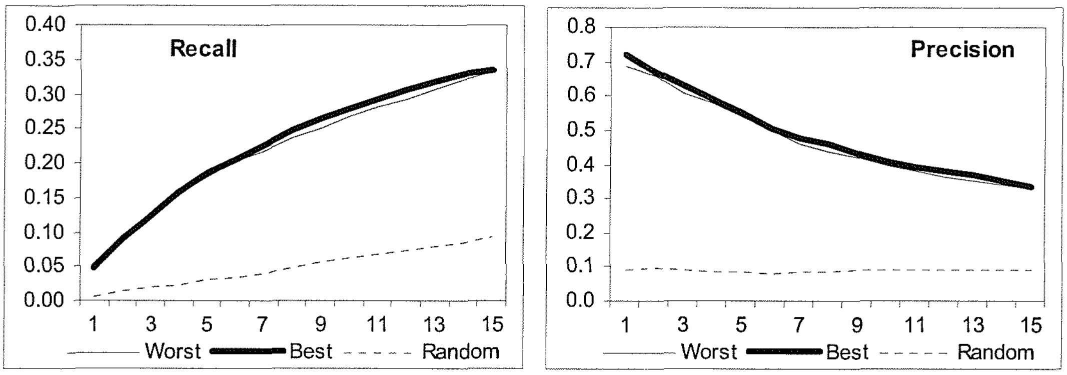


Figure 8-1 - Recall and precision graphs for Archaeology, evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 1 | 2 | 0 | 0 | 0.2266 |
| Worst | 0 | 0 | 4 | 3 | 0.2192 |

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 1 | 1 | 0 | 0 | 0.4864 |
| Worst | 1 | 0 | 4 | 4 | 0.4711 |

Table 8-2 - Recall, precision and weight values for Archaeology, evaluating abstract information

This application possesses the least valuable abstract level compared with other applications. This is reflected in the low weights given to those elements. The proximity between the worst and best recall and precision values also hint at this.

We should also note that considering a high weight for information from neighbours leads to the worst recall and precision.

However, with the best weights for precision and recall, only the classification and description from neighbours was not considered, and for recall, the Document classification has a weight of two. The classification has some relevant words related to the subject of the application and this improves recall. However, since these words are used in more than one document, when we index them, they become less pertinent to improved precision.

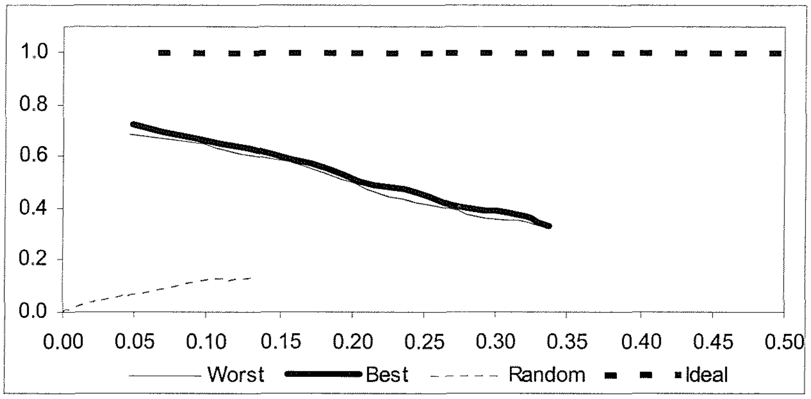


Figure 8-2 - Recall/precision graph for Archaeology while evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 1 | 1 | 0 | 0 | 0.2265 | 0.4864 | 0.3091 |
| Worst | 1 | 0 | 4 | 4 | 0.2193 | 0.4711 | 0.2993 |

Table 8-3 - Best overall weight values for Archaeology concerning abstract information

It appears that while evaluating abstract information, high precision is the answer to high values for effectiveness. This is contrary to what happened when evaluating link information, but since in this test we are already using a lot of information from the link level, there is no need for the highest recall, and so precision is more important for better effectiveness.

## Cell Biology:

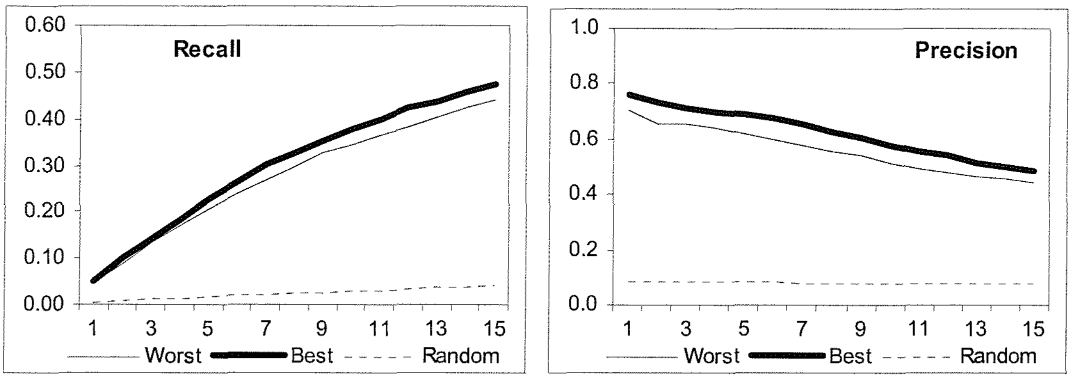


Figure 8-3 - Recall and precision graphs for Cell Biology, evaluating abstract information

|  | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 4 | 3 | 2 | 3 | 0.3022 |
| Worst | 0 | 0 | 0 | 0 | 0.2767 |

|  | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 4 | 3 | 2 | 3 | 0.6140 |
| Worst | 0 | 0 | 0 | 0 | 0.5606 |

Table 8-4 - Recall, precision and weight values for Cell Biology, evaluating abstract information

In the Cell Biology application there is a good document description working as a title. This is the reason why there is such a high weight assigned to it on both recall and precision. What could be surprising is that neighbour document description only receives medium weight. But we should notice that link description is quite similar to neighbour document description, and that in this test the metadata has a fixed weight of two. This might be one reason why neighbour document description has a lower weight. In addition we also expect neighbour document description to be further way from the content of the document than the description of itself. The classification weight is also quite high, and again the same for both recall and precision. This is possibly because we have a subject classification of documents, without many documents per entry.

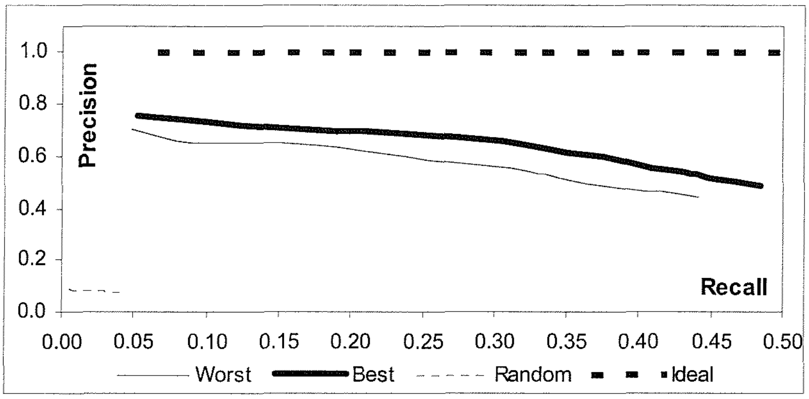


Figure 8-4 - Recall/precision graph for Cell Biology while evaluating abstract information

|  | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 4 | 3 | 2 | 3 | 0.3022 | 0.6140 | 0.4050 |
| Worst | 0 | 0 | 0 | 0 | 0.2767 | 0.5606 | 0.3705 |

Table 8-5 - Best overall weight values for Cell Biology concerning abstract information

Here we have the highest distance between the best and the worst values of effectiveness. This suggests that the quality of the abstract level in this application is quite good. Even so, if this information were not considered, it would achieve a better performance than any other system. This proves the good quality of link information.

## French:

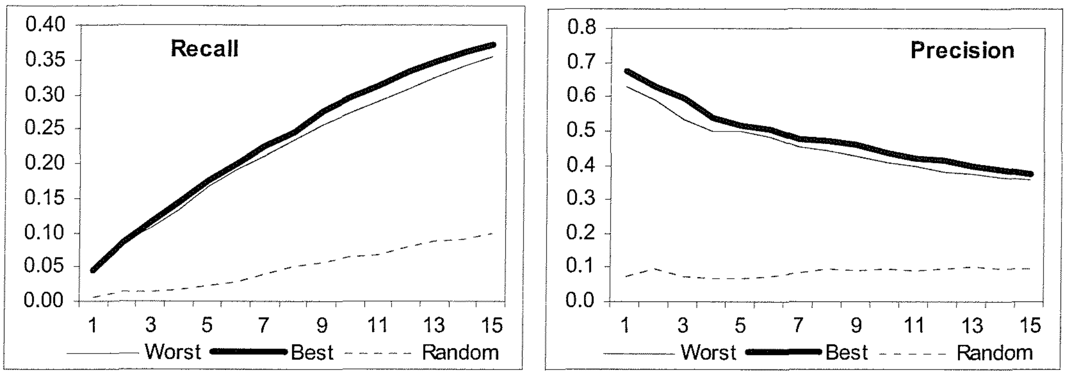


Figure 8-5 - Recall and precision graphs for French, evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 3 | 2 | 1 | 0 | 0.2357 |
| Worst | 0 | 0 | 0 | 4 | 0.2177 |

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 3 | 1 | 0 | 0 | 0.4886 |
| Worst | 0 | 0 | 0 | 4 | 0.4660 |

Table 8-6 - Recall, precision and weight values for French, evaluating abstract information

The French application has a low weight for classification information. It may also be noted that precision does not improve as much as recall when classification information is considered. The kind of words used in the tree entries and the number of documents per entry contribute to this.

The precision weight for neighbour document description is lower than the one on recall. Actually, it is equal to zero. This is due to the repetition of a few words in many documents. In addition, neighbours do not always have a subject relation, but sometimes have a resources relation for French learning. Considering more information from this metadata would decrease precision.

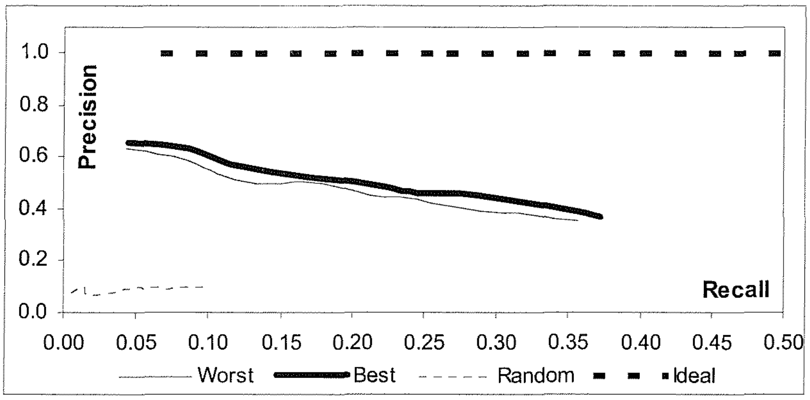


Figure 8-6 - Recall/precision graph for French while evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 3 | 2 | 1 | 0 | 0.2357 | 0.4850 | 0.3172 |
| Worst | 0 | 0 | 0 | 4 | 0.2177 | 0.4660 | 0.2968 |

Table 8-7 - Best overall weight values for French concerning abstract information

For best effectiveness, we have the same weights as for recall. We should note that considering neighbour classification does not help in context indexing. Actually, the opposite effect occurs. With this information present, the worst performance is achieved.

### French Revolution

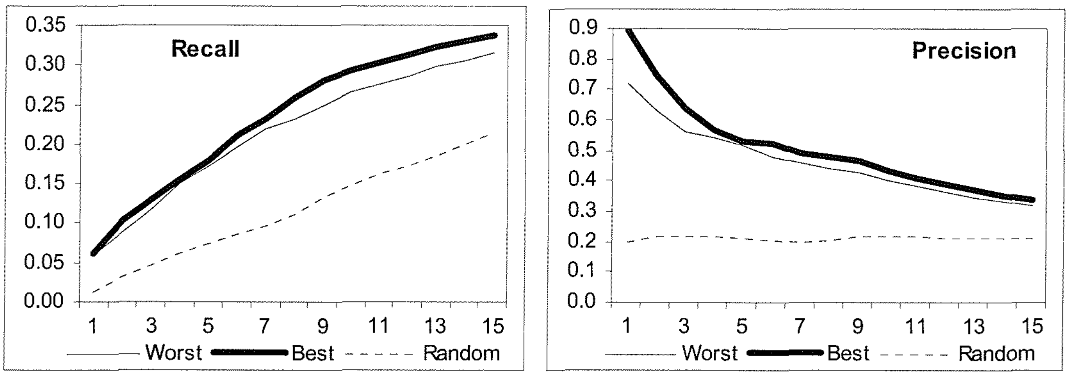


Figure 8-7 - Recall and precision graphs for French Revolution, evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 4 | 3 | 1 | 0 | 0.2351 |
| Worst | 0 | 0 | 0 | 1 | 0.2153 |

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 4 | 2 | 1 | 0 | 0.5139 |
| Worst | 0 | 0 | 0 | 2 | 0.4698 |

Table 8-8 - Recall, precision and weight values for French Revolution, evaluating abstract information

Document description functions here as a title with a considerable number of words. This type of metadata assures a high weight for it. However, the neighbour document does not have such a high weight for document description. A reason for this is that, in addition to its similarity with link description, the links in this application do not always connect semantically close documents, as happened, for instance, in the Cell Biology application. The classification weight is high for recall but slightly lower for precision. However, the neighbour classification should not be considered, and in addition its presence without any other elements leads to the worst values of recall and precision. This occurs because we have few subject classification entries in the tree, but a high number of documents per subject entry.

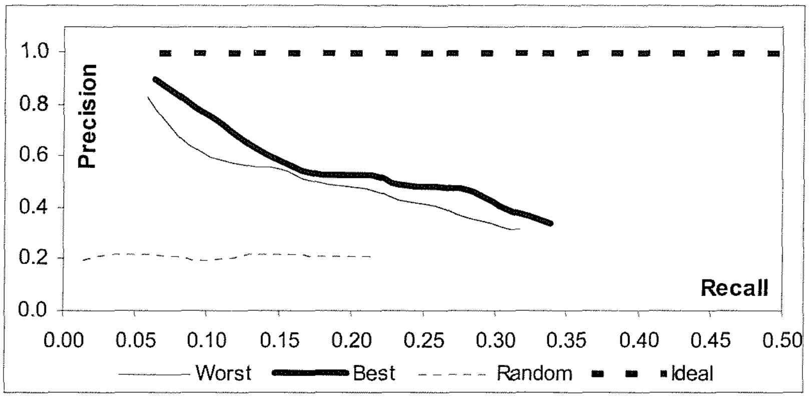


Figure 8-8 - Recall/precision graph for French Revolution while evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 4 | 2 | 1 | 0 | 0.2343 | 0.5139 | 0.3218 |
| Worst | 0 | 0 | 0 | 1 | 0.2153 | 0.4707 | 0.2954 |

Table 8-9 - Best overall weight values for French Revolution concerning abstract information

The distance between the worst and best effectiveness in this application is in the mid-range of the used applications for evaluation. It has good document description but deficient classification and the weights attributed to the different abstract metadata information show it.

## Tulip:

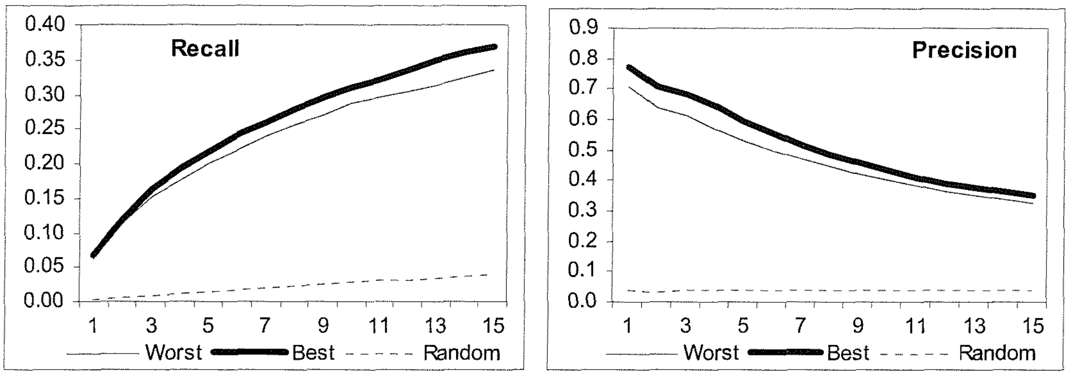


Figure 8-9 - Recall and precision graphs for Tulip, evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 3 | 3 | 1 | 1 | 0.2598 |
| Worst | 0 | 0 | 0 | 0 | 0.2379 |

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 3 | 3 | 0 | 1 | 0.5455 |
| Worst | 0 | 0 | 0 | 0 | 0.5010 |

Table 8-10 - Recall, precision and weight values for Tulip, evaluating abstract information

With this application, there are some documents where there is a repetition of a few words. This implies a not very high weight for document description. In addition, some documents implement a quiz, and the kind of description on documents is not always the most relevant.

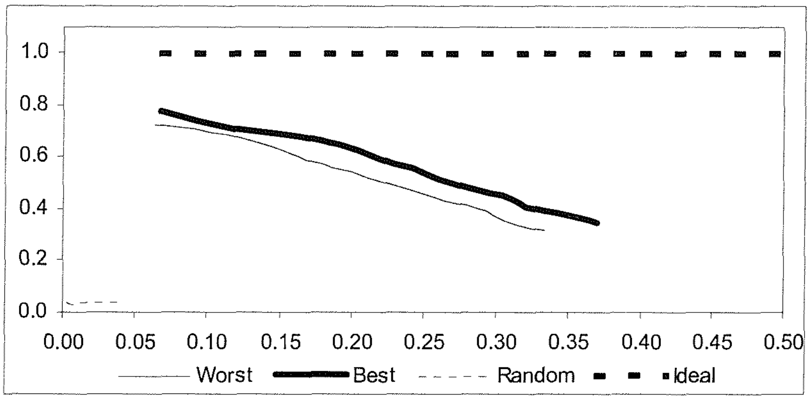


Figure 8-10 - Recall/precision graph for Tulip while evaluating abstract information

| | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 3 | 3 | 1 | 1 | 0.2598 | 0.5444 | 0.3518 |
| Worst | 0 | 0 | 0 | 0 | 0.2379 | 0.5010 | 0.3227 |

Table 8-11 - Best overall weight values for Tulip concerning abstract information

In our opinion, the Tulip application has the second best classification, and this is confirmed by a reasonable weighting of this metadata. The results were not as good as in the Cell Biology application because Tulip has more documents per subject tree entry. Nevertheless, the number of classification entries is similar. We should also be aware of the 'case histories' classification. They are not the most useful metadata for context indexing.

### Summary:

In summary, there is a mixed performance according to the different styles of abstract information. The two extremes can be found in the Archaeology and Cell Biology applications. The former has a poor abstract level, where document description and classification words are not usually human readable. On the contrary, the Cell biology application has good document description, links connect semantically related documents, and classification is made according to content subject without many documents per entry.

| | | Best Recall | | | | | Best Precision | | | | | Best Effectiveness | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Precision | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Average Recall | Average Precision | Effectiveness |
| Archaeology | Best | 1 | 2 | 0 | 0 | 0.2266 | 1 | 1 | 0 | 0 | 0.4864 | 1 | 1 | 0 | 0 | 0.2265 | 0.4864 | 0.3091 |
| | Worst | 0 | 0 | 4 | 3 | 0.2192 | 1 | 0 | 4 | 4 | 0.4711 | 1 | 0 | 4 | 4 | 0.2193 | 0.4711 | 0.2993 |
| Cell Biology | Best | 4 | 3 | 2 | 3 | 0.3022 | 4 | 3 | 2 | 3 | 0.6140 | 4 | 3 | 2 | 3 | 0.3022 | 0.6140 | 0.4050 |
| | Worst | 0 | 0 | 0 | 0 | 0.2767 | 0 | 0 | 0 | 0 | 0.5606 | 0 | 0 | 0 | 0 | 0.2767 | 0.5606 | 0.3705 |
| French | Best | 3 | 2 | 1 | 0 | 0.2357 | 3 | 1 | 0 | 0 | 0.4886 | 3 | 2 | 1 | 0 | 0.2357 | 0.4850 | 0.3172 |
| | Worst | 0 | 0 | 0 | 4 | 0.2177 | 0 | 0 | 0 | 4 | 0.4660 | 0 | 0 | 0 | 4 | 0.2177 | 0.4660 | 0.2968 |
| French Revolution | Best | 4 | 3 | 1 | 0 | 0.2351 | 4 | 2 | 1 | 0 | 0.5139 | 4 | 2 | 1 | 0 | 0.2343 | 0.5139 | 0.3218 |
| | Worst | 0 | 0 | 0 | 1 | 0.2153 | 0 | 0 | 0 | 2 | 0.4698 | 0 | 0 | 0 | 1 | 0.2153 | 0.4707 | 0.2954 |
| Tulip | Best | 3 | 3 | 1 | 1 | 0.2598 | 3 | 3 | 0 | 1 | 0.5455 | 1 | 1 | 0 | 0 | 0.2598 | 0.5444 | 0.3518 |
| | Worst | 0 | 0 | 0 | 0 | 0.2379 | 0 | 0 | 0 | 0 | 0.5010 | 0 | 0 | 0 | 0 | 0.2379 | 0.5010 | 0.3227 |

Table 8-12 - Summary of the best and worst metadata weights, concerning abstract information

If we now consider each of the elements separately, we can reach further conclusions on our policy of weight assignment.

A very high weight for document description requires a reasonable number of words, functioning as a title, related to the content subject of the document. That occurred in the French Revolution and Cell Biology applications. However, if the relation is not always accurate, or if there is a repetition of words, then it should be assigned a medium high weight, as in the French and Tulip applications. If the majority of words are not human readable, then document description should be assigned a low weight, as in the case for the Archaeology application.

Neighbour document description consideration is dependent not only on the same factors as the previous element, but also on the kind of relations implemented through links. However, it should be clear that the similarity between this information and link description in the majority of the applications dictates a lower performance for this parameter in this test, because here link description has a fixed weight of two. In applications with good document description words, neighbour description can be considered in different ways. A medium weight is assigned if there are links almost only relating subjects. The Cell Biology application is an example. However, if links also implement different kinds of relations then there should be a low weight for neighbour document description. The French revolution application is a good example of this. However, if document description is not always very good, neighbour description should always have a low weight. That wouldn't be true if we required the best precision. In this situation, for the French and Tulip applications, the weight should be zero. Finally, if document description is very poor, as in the Archaeology application, the neighbour description should not be considered.

The classification weighting is mainly dependent on three factors: how semantically close are the words in classification to the content of documents; how many subject entries are there in the classification tree; how many documents are there per entry. These last two factors are related such that, if there are fewer subject entries, then there will be a higher average of documents per entry. Semantically close classifications, with many entries and few documents per entry assure a medium-high weighting of this metadata, as in the Cell Biology and Tulip applications. If that is not the case, then a medium weighting applies, as in the French and French Revolution applications. There is also a distinction, with a lower weight, on precision. In the worst scenario, with few relevant words and many non human-readable ones, a low weight is assigned, as in the Archaeology application.

The element most affected by deficient classification is the neighbour classification. A classification not semantically close to the document or without many entries implies the non-consideration of the neighbour document classification words. Examples of this are the Archaeology, French and French Revolution applications. But, with a correct classification in many entries, there is a medium high weight (Cell Biology application) if there are few documents per entry, or a low weight (Tulip application) if there are many documents per entry.

## 8.3 Evaluating relations between abstract information and link context information

The elements considered in this test are link descriptions, link anchor selections, document description and document classification. All the different weight combinations between zero and four were considered for eight elements (see Table 8-13). However, the weight variation for the distinction between generic and specific link, and also between destination and source anchors, were tied together. This is shown in the table with the same grey filling. In this way, there were four independent parameters to inspect, and this gave 625 runs in each test. The majority of tests took at least half a day.

The analysed metadata are:

- Link description
- Link anchor selection
- Document description
- Document classification

| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 0 - 4 | 0 - 4 |
| | Text Anchor as Source: | Selection | 0 - 4 | 0 - 4 |
| | | Sentence | 0 - 0 | 0 - 0 |
| | | Paragraph | 0 - 0 | 0 - 0 |
| | Text Anchor as Destination: | Selection | 0 - 4 | 0 - 4 |
| | | Sentence | 0 - 0 | 0 - 0 |
| | | Paragraph | 0 - 0 | 0 - 0 |
| Abstract information context | Self | Description | 0 - 4 | |
| | | Classification | 0 - 4 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 2 - 2 | 2 - 2 |
| | | Classification | 2 - 2 | 2 - 2 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 1 - 1 | 1 - 1 |
| | Link Outside Classification | | 1 - 1 | 1 - 1 |

Table 8-13 - Weight intervals for evaluating relation the between abstract and link information

After running the tests, for each of the applications, the retrieval performance graphs and values along with the best weighting policy were obtained. Finally, some brief conclusions are given mainly on how to consider the distinction between link selection and document description.

### Archaeology:

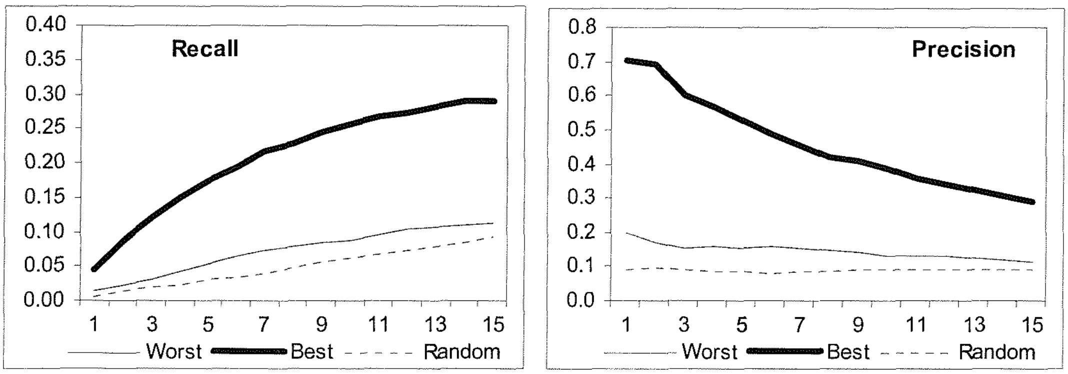
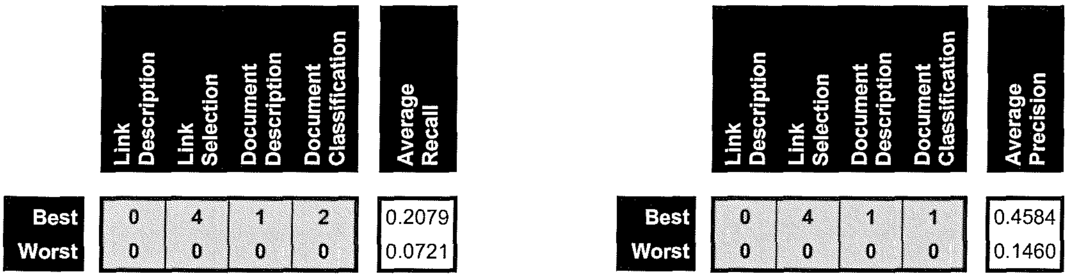


Figure 8-11 - Recall and precision graphs for Archaeology, evaluating abstracts and links



| | Link Description | Link Selection | Document Description | Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 0 | 4 | 1 | 2 | 0.2079 |
| Worst | 0 | 0 | 0 | 0 | 0.0721 |

| | Link Description | Link Selection | Document Description | Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 0 | 4 | 1 | 1 | 0.4584 |
| Worst | 0 | 0 | 0 | 0 | 0.1460 |

Table 8-14 - Recall, precision and weight values for Archaeology, evaluating abstracts and links

In this application link anchor information, or to be more precise link anchor selection is of a key importance. It requires a high weight value for the best performance. Again, the similarity between link description and neighbour document description, gives a weight of zero to the former because the latter has a fixed weight of two for this test. Document classification seems to be slightly better at improving recall than precision.

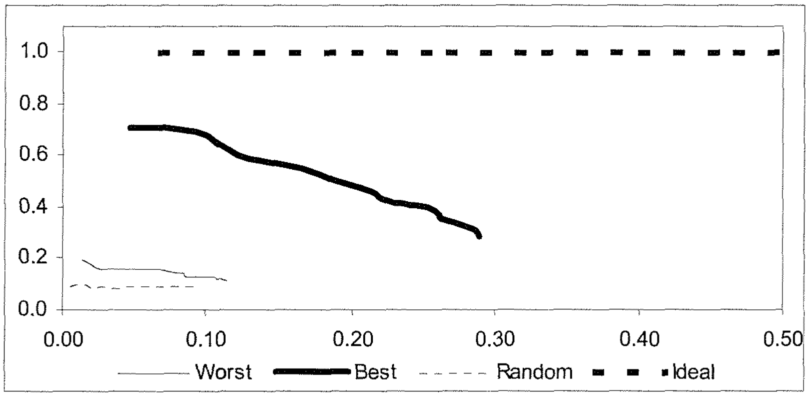
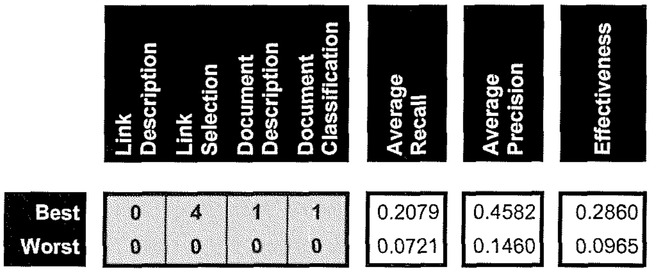


Figure 8-12 - Recall/precision graph for Archaeology while evaluating abstracts and links



| | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 0 | 4 | 1 | 1 | 0.2079 | 0.4582 | 0.2860 |
| Worst | 0 | 0 | 0 | 0 | 0.0721 | 0.1460 | 0.0965 |

Table 8-15 - Best overall weight values for Archaeology concerning abstracts and links

When there is the worst situation in these tests, the only information used is the document description of neighbours. Low recall and precision, very close to random retrieval, shows the very low quality of these metadata for context indexing in this application. The system will mainly rely on anchors for context indexing.

**Cell Biology:**

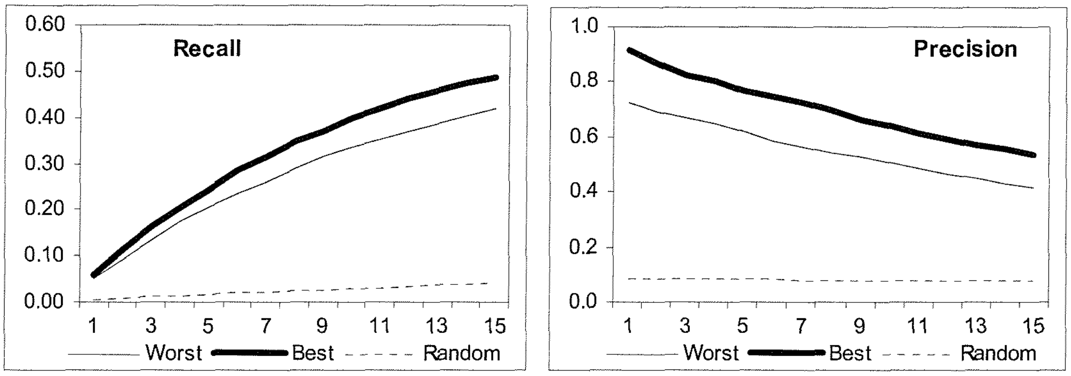
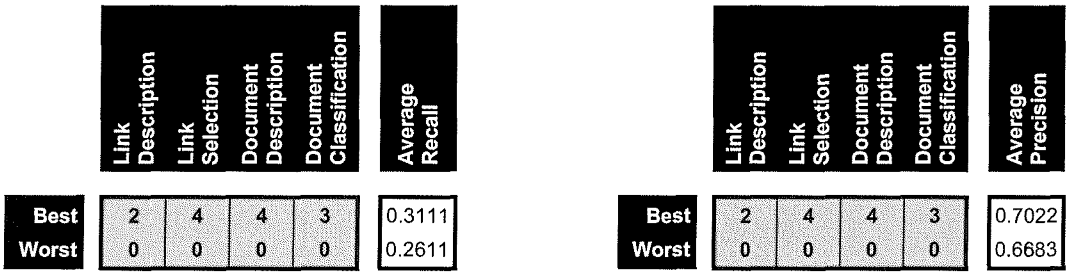


Figure 8-13 - Recall and precision graphs for Cell Biology, evaluating abstracts and links



| | Link Description | Link Selection | Document Description | Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 2 | 4 | 4 | 3 | 0.3111 |
| Worst | 0 | 0 | 0 | 0 | 0.2611 |

| | Link Description | Link Selection | Document Description | Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 2 | 4 | 4 | 3 | 0.7022 |
| Worst | 0 | 0 | 0 | 0 | 0.6683 |

Table 8-16 - Recall, precision and weight values for Cell Biology, evaluating abstracts and links

The Cell Biology application has high weights for anchor selections and document descriptions. The weights are slightly lower for document classification and medium for link description. Document classification gets a higher weight than the than the one fixed on two for the classification of neighbours. The medium weight of link description is again justified by the similarity with neighbour document description.

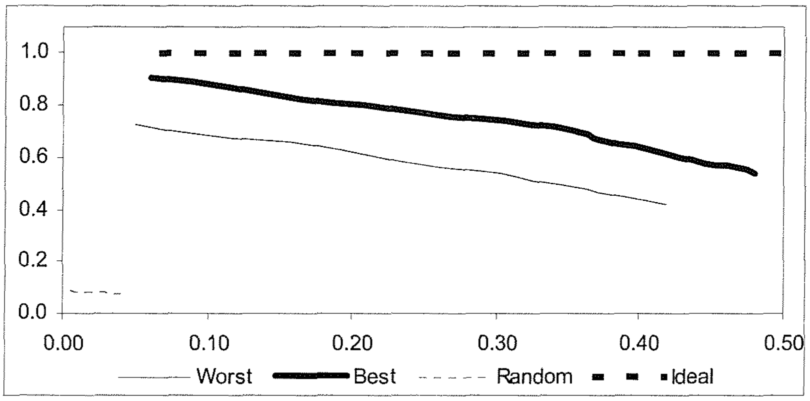


Figure 8-14 - Recall/precision graph for Cell Biology while evaluating abstracts and links

| | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 2 | 4 | 4 | 3 | 0.3111 | 0.7022 | 0.4312 |
| Worst | 0 | 0 | 0 | 0 | 0.2611 | 0.6083 | 0.3654 |

Table 8-17 - Best overall weight values for Cell Biology concerning abstracts and links

The analysis of effectiveness shows similar weights. Link anchor selection and document description are the most important metadata, followed by link description / neighbour document description, then by document classification and finally by classification of neighbours.

## French:

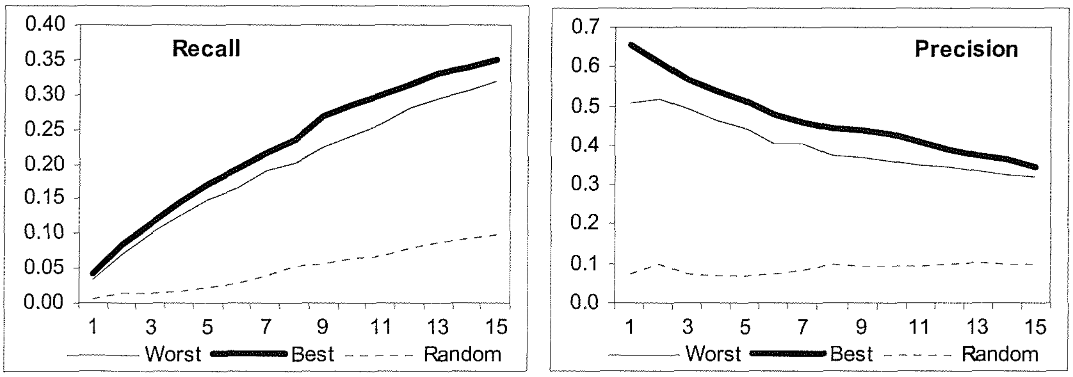


Figure 8-15 - Recall and precision graphs for French, evaluating abstracts and links

|  | Link Description | Link Selection | Document Description | Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 1 | 2 | 3 | 2 | 0.2252 |
| Worst | 0 | 0 | 0 | 0 | 0.1940 |

|  | Link Description | Link Selection | Document Description | Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 1 | 1 | 3 | 2 | 0.4678 |
| Worst | 0 | 0 | 0 | 0 | 0.4153 |

Table 8-18 - Recall, precision and weight values for French, evaluating abstracts and links

It may be seen in this application, there is a medium weight for document classification, and a medium-high weight for document description. This is in accordance with the weights obtained while testing abstract information, and here link description gets a low weight for the same reasons as neighbour document description got previously: the repetition of similar data. We are then more interested in the link selection weight, which proves to have the same importance as the classifications and is one step lower than document description, in terms of recall. However, a low weight is obtained when precision is considered.

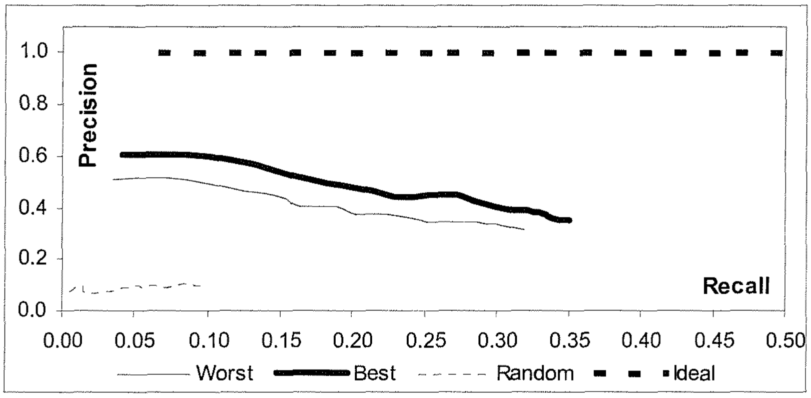


Figure 8-16 - Recall/precision graph for French while evaluating abstracts and links

|  | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 1 | 2 | 3 | 2 | 0.2252 | 0.4668 | 0.3039 |
| Worst | 0 | 0 | 0 | 0 | 0.1940 | 0.2234 | 0.2661 |

Table 8-19 - Best overall weight values for French concerning abstracts and links

The selection link anchor weight obtained for effectiveness is higher than what was expected, compared to the one obtained while testing anchor bounds. There, a weight of one was obtained, but there were also higher weights for anchor sentence and paragraph. Now in the absence of this metadata, a higher weight is obtained for selection. However, the relation between document description and anchor selection is maintained.

### French Revolution:

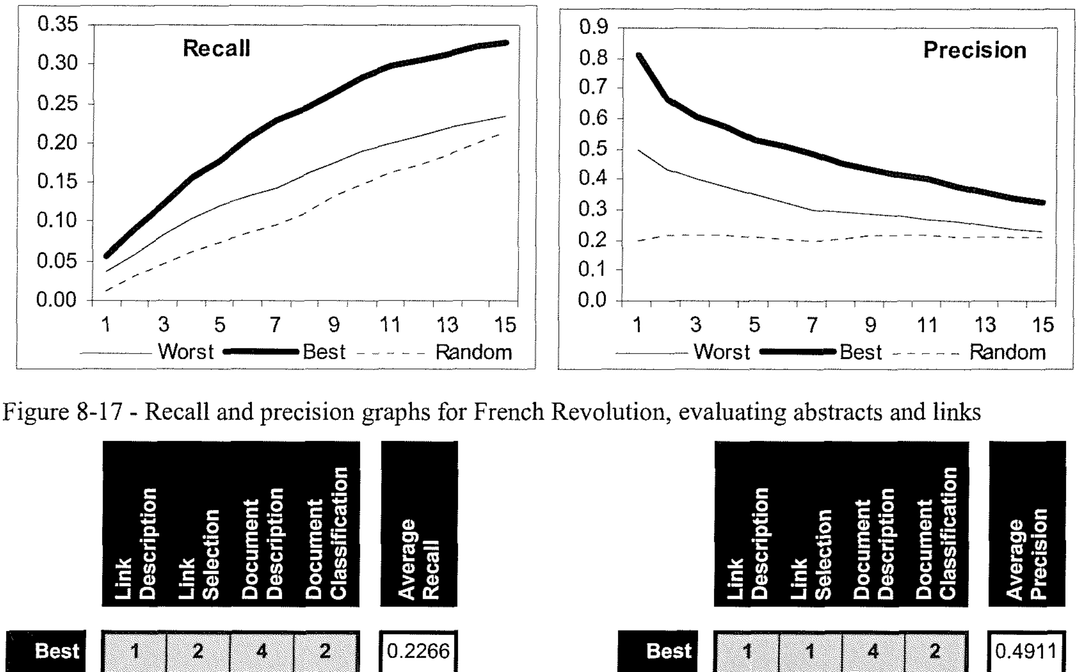


Figure 8-17 - Recall and precision graphs for French Revolution, evaluating abstracts and links

| | Link Description | Link Selection | Document Description | Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 1 | 2 | 4 | 2 | 0.2266 |
| Worst | 0 | 0 | 0 | 0 | 0.1529 |

| | Link Description | Link Selection | Document Description | Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 1 | 1 | 4 | 2 | 0.4911 |
| Worst | 0 | 0 | 0 | 0 | 0.3249 |

Table 8-20 - Recall, precision and weight values for French Revolution, evaluating abstracts and links

The strength of this application is the appropriate description of documents. French Revolution application also gets medium weight for document classification. For the same reasons as explained earlier, link description gets a low weight of one. In comparison with previous weights, link selection has a medium weight for recall and a low weight for precision. The repetition of some words in selections is a reason for the lower weight on precision.

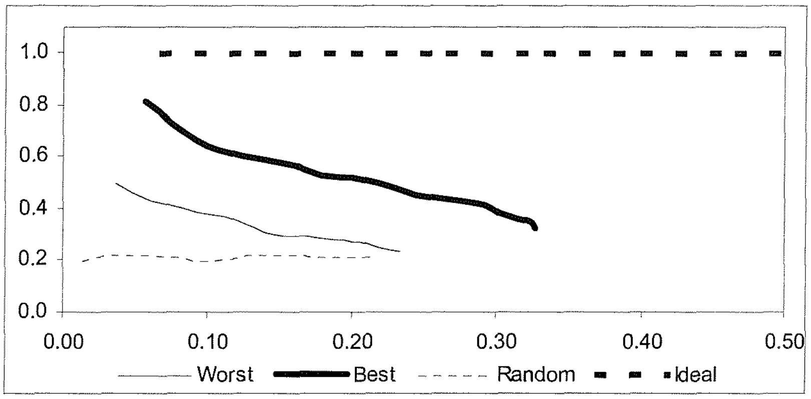


Figure 8-18 - Recall/precision graph for French Revolution while evaluating abstracts and links

| | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 1 | 2 | 4 | 2 | 0.2266 | 0.4908 | 0.3100 |
| Worst | 0 | 0 | 0 | 0 | 0.1529 | 0.3249 | 0.2049 |

Table 8-21 - Best overall weight values for French Revolution concerning abstracts and links

For best effectiveness, the selection weight is also higher than what was expected for the same reasons as for French application. However, the distinction between document description and anchor selection is also high.

### Tulip:

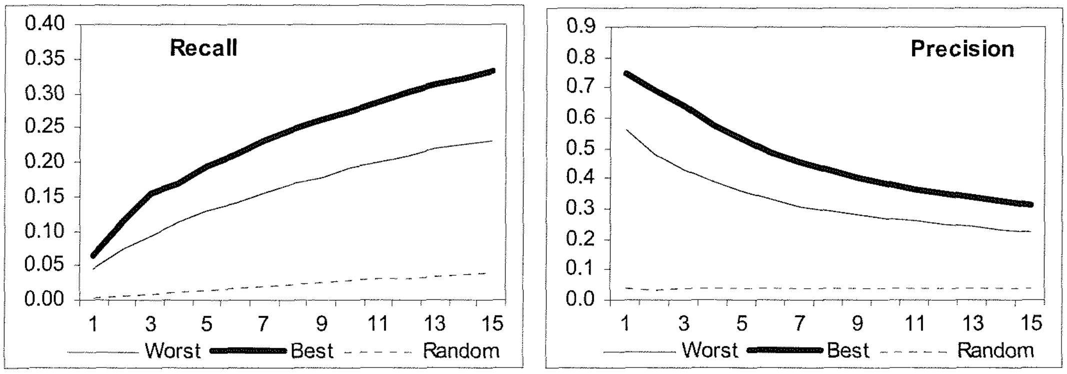


Figure 8-19 - Recall and precision graphs for Tulip, evaluating abstracts and links

| | Link Description | Link Selection | Document Description | Document Classification | Average Recall |
|---|---|---|---|---|---|
| Best | 1 | 1 | 3 | 3 | 0.2336 |
| Worst | 0 | 0 | 0 | 0 | 0.1605 |

| | Link Description | Link Selection | Document Description | Document Classification | Average Precision |
|---|---|---|---|---|---|
| Best | 1 | 1 | 3 | 3 | 0.4949 |
| Worst | 0 | 0 | 0 | 1 | 0.3467 |

Table 8-22 - Recall, precision and weight values for Tulip, evaluating abstracts and links

The Tulip application maintains similar relations between the analysed abstract metadata. Link selection gets a low weight in comparison with other applications.

Taking document classification alone, the worst value for precision is obtained.

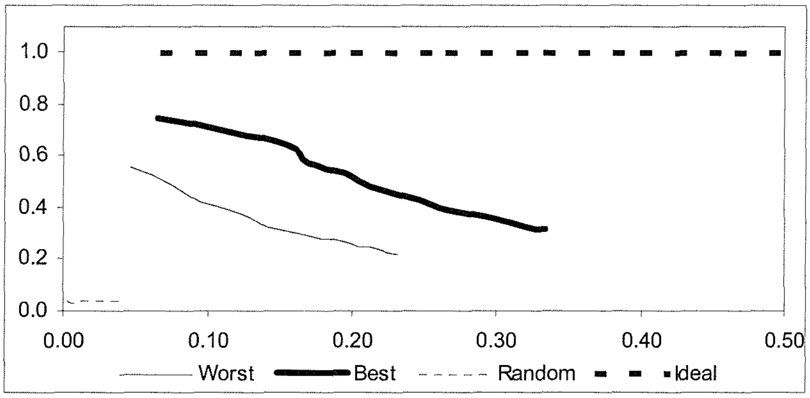


Figure 8-20 - Recall/precision graph for Tulip while evaluating abstracts and links

| | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|---|---|
| Best | 1 | 1 | 3 | 3 | 0.2336 | 0.4949 | 0.3174 |
| Worst | 0 | 0 | 0 | 0 | 0.1605 | 0.3471 | 0.2195 |

Table 8-23 - Best overall weight values for Tulip concerning abstracts and links

Tulip maintains the same relation between link selection and document description as the French Revolution application does, but both these weights are one step lower due to their quality. However, with better document classification a higher weight value for this metadata was achieved.

### Summary:

Finally, an idea may be obtained of the relation between some link information and abstract information. The analysed abstract information, document description and classification, achieved similar weights when compared with the previous test in section 8.2. So, the remaining parameters, link description and anchor selection, will be compared with this metadata.

| | | Best Recall | | | | | Best Precision | | | | | Best Effectiveness | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Link Description | Link Selection | Document Description | Document Classification | Average Precision | Link Description | Link Selection | Document Description | Document Classification | Average Recall | Average Precision | Effectiveness |
| Archaeology | Best | 0 | 4 | 1 | 2 | 0.2079 | 0 | 4 | 1 | 1 | 0.4584 | 0 | 4 | 1 | 1 | 0.2079 | 0.4582 | 0.2860 |
| | Worst | 0 | 0 | 0 | 0 | 0.0721 | 0 | 0 | 0 | 0 | 0.1460 | 0 | 0 | 0 | 0 | 0.0721 | 0.1460 | 0.0965 |
| Cell Biology | Best | 2 | 4 | 4 | 3 | 0.3111 | 2 | 4 | 4 | 3 | 0.7022 | 2 | 4 | 4 | 3 | 0.3111 | 0.7022 | 0.4312 |
| | Worst | 0 | 0 | 0 | 0 | 0.2611 | 0 | 0 | 0 | 0 | 0.6683 | 0 | 0 | 0 | 0 | 0.2611 | 0.6083 | 0.3654 |
| French | Best | 1 | 2 | 3 | 2 | 0.2252 | 1 | 1 | 3 | 2 | 0.4678 | 1 | 2 | 3 | 2 | 0.2252 | 0.4668 | 0.3039 |
| | Worst | 0 | 0 | 0 | 0 | 0.1940 | 0 | 0 | 0 | 0 | 0.4153 | 0 | 0 | 0 | 0 | 0.1940 | 0.2234 | 0.2661 |
| French Revolution | Best | 1 | 2 | 4 | 2 | 0.2266 | 1 | 1 | 4 | 2 | 0.4911 | 1 | 2 | 4 | 2 | 0.2266 | 0.4908 | 0.3100 |
| | Worst | 0 | 0 | 0 | 0 | 0.1529 | 0 | 0 | 0 | 0 | 0.3249 | 0 | 0 | 0 | 0 | 0.1529 | 0.3249 | 0.2049 |
| Tulip | Best | 1 | 1 | 3 | 3 | 0.2336 | 1 | 1 | 3 | 3 | 0.4949 | 1 | 1 | 3 | 3 | 0.2336 | 0.4949 | 0.3174 |
| | Worst | 0 | 0 | 0 | 0 | 0.1605 | 0 | 0 | 0 | 0 | 0.3467 | 0 | 0 | 0 | 0 | 0.1605 | 0.3471 | 0.2195 |

Table 8-24 - Summary of the best and worst metadata weights, evaluating abstract and link information

According to the obtained weights for link selection, there is a clear distinction between two kinds of applications:

In the first, if link selections have expressive text that is related to a subject developed at the other end of the link then we should assign the highest weight to it. The first two applications, Archaeology and Cell Biology, are examples of this and we already observed the relevance of this parameter when we tested link bounds.

However, the relation between link selection and other metadata can be somewhat different. With document description functioning as a title, related to its content, the same weight can be assigned to both document description and link selection, as in the Cell Biology application. However, if document description is quite poor and, in the extreme, if the majority of words are not human-readable then very distinct weights should be assigned for this parameter. The Archaeology application is an example of this where link selection and document description obtained the two weight extremes, four and one respectively.

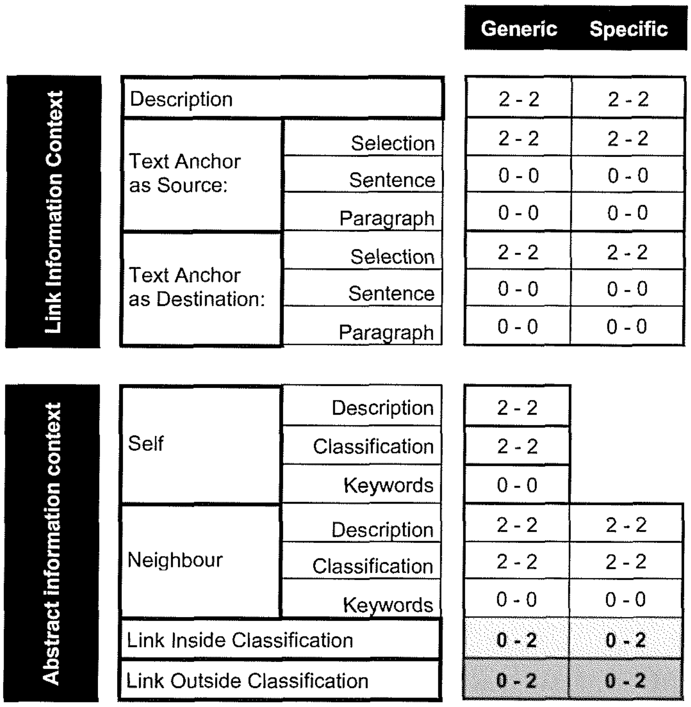
In the second, if links do not connect similar subjects to ones developed at the two ends of the links, or if there is repetition of some words in the anchor selection or even the absence of text in many anchors, then a low weight should be assigned to it. If we consider the anchor bounds tests, we would expect the selection weight to have the value of one. But here a slightly higher weight for two applications, French and French Revolution was obtained. For sure, their link anchor selections contains more text than in the Tulip application, and this might be the reason for this difference. In addition, the absence of anchor sentence and paragraph in this test necessitates a higher value for selection to improve recall. By way of confirmation, for precision their weight for selection is actually the same as in the Tulip application.

Still in this second kind of application, the comparison with abstract information is more difficult to grasp. Here, the weight for document description is always higher than for selection. Fortunately there are applications with good document descriptions, French Revolution application being the best example. This implies two units of difference between the weights for link selection and document description. Without such good document description, and depending on link selection, there would be one unit of difference, as in the French application, or two units of difference as in the Tulip application.

Finally, knowing the similarities between the link descriptions analysed here and neighbour descriptions in the previous test, similar weights could be expected for link descriptions in all applications. Those two parameters cannot be dissociated, as the results obtained prove. Nevertheless, it could be interesting to know the relation between the two parameters. However authors have chosen not to use different information for them.

## 8.4 Evaluating links taking account of the location of source and destination documents of links with regard to document classifications

This test determines whether it is relevant to distinguish between links that connect documents in the same classification and links connecting documents included in different classifications.



| | | | Generic | Specific |
|---|---|---|---|---|
| Link Information Context | Description | | 2 - 2 | 2 - 2 |
| | Text Anchor as Source: | Selection | 2 - 2 | 2 - 2 |
| | | Sentence | 0 - 0 | 0 - 0 |
| | | Paragraph | 0 - 0 | 0 - 0 |
| | Text Anchor as Destination: | Selection | 2 - 2 | 2 - 2 |
| | | Sentence | 0 - 0 | 0 - 0 |
| | | Paragraph | 0 - 0 | 0 - 0 |
| Abstract information context | Self | Description | 2 - 2 | |
| | | Classification | 2 - 2 | |
| | | Keywords | 0 - 0 | |
| | Neighbour | Description | 2 - 2 | 2 - 2 |
| | | Classification | 2 - 2 | 2 - 2 |
| | | Keywords | 0 - 0 | 0 - 0 |
| | Link Inside Classification | | 0 - 2 | 0 - 2 |
| | Link Outside Classification | | 0 - 2 | 0 - 2 |

Table 8-25 - Weight intervals for evaluating relation between abstract and link information

All the different weight combinations between zero and two were considered for four elements as seen in Table 8-25. The elements being considered are links to documents inside and outside classifications, and for this reason combinations between zero and four were not considered. These elements work in a different way to the others. Their weights actually multiply the weights of all other elements dependent on links, and so a factor of 4 for this weight would assign a weight of 16 to the other elements highly valued. This discrepancy would be too great in comparison with other elements not dependent on links, e.g. document description and document classification.

The weight variation for the distinction between generic and specific link were tied together. This is represented in the table with the same grey filling. In this way, there were two independent parameters to be inspected, and this gave nine runs in each test. The majority of the tests took just a few minutes.

The metadata analysed are:

- Links to document inside classification
- Links to document outside classification

From the first implementation of these tests, all classification information relating to file formats was removed from the entries in the tree. In this manner, we avoided consideration of documents in the same classification just by the fact that they possessed the same file format, and in this test that could happen with the text format.

For each of the applications, the retrieval performance graphs and values were obtained along with the best weighting policy, for that situation.

## Archaeology:

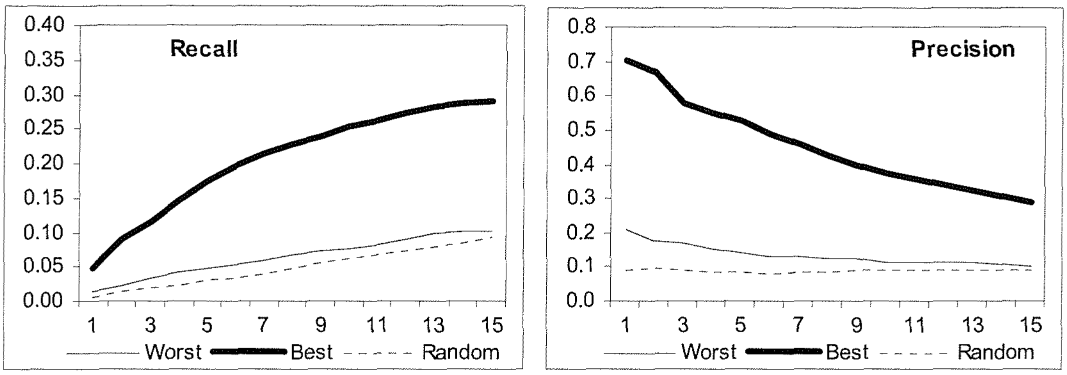
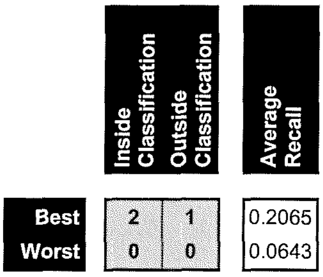
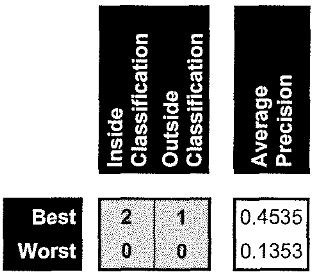


Figure 8-21 - Recall and precision graphs for Archaeology, evaluating classification inclusion



| | Inside Classification | Outside Classification | Average Recall |
|---|---|---|---|
| Best | 2 | 1 | 0.2065 |
| Worst | 0 | 0 | 0.0643 |



| | Inside Classification | Outside Classification | Average Precision |
|---|---|---|---|
| Best | 2 | 1 | 0.4535 |
| Worst | 0 | 0 | 0.1353 |

Table 8-26 - Recall, precision and weights for Archaeology, evaluating classification inclusion

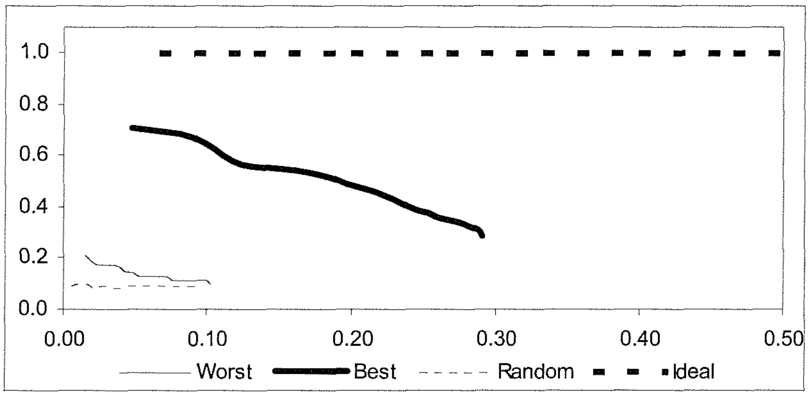
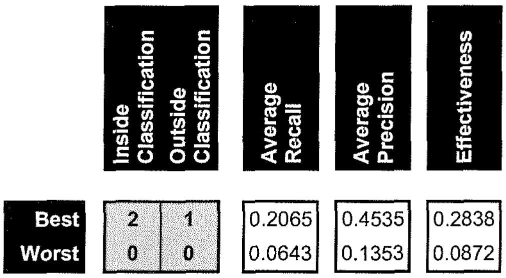


Figure 8-22 - Recall/precision graph for Archaeology, evaluating classification inclusion



| | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|
| Best | 2 | 1 | 0.2065 | 0.4535 | 0.2838 |
| Worst | 0 | 0 | 0.0643 | 0.1353 | 0.0872 |

Table 8-27 - Best overall weight values for Archaeology concerning classification inclusion

In the Archaeology application, the best performance is achieved for all retrieval metrics when more weight is given to links connecting documents inside the same classification.

Previously the conclusion was reached that classification words were not the best in this application; however the application broadly classifies documents in sets of common subjects. This last aspect might be a key factor for attributing a higher weight for links relating documents in the same classification. However, there is also a classification according to user expertise, and this could prejudice weight assignment but, to balance this, there are many more links connecting documents in the same classification than the contrary.

### Cell Biology:

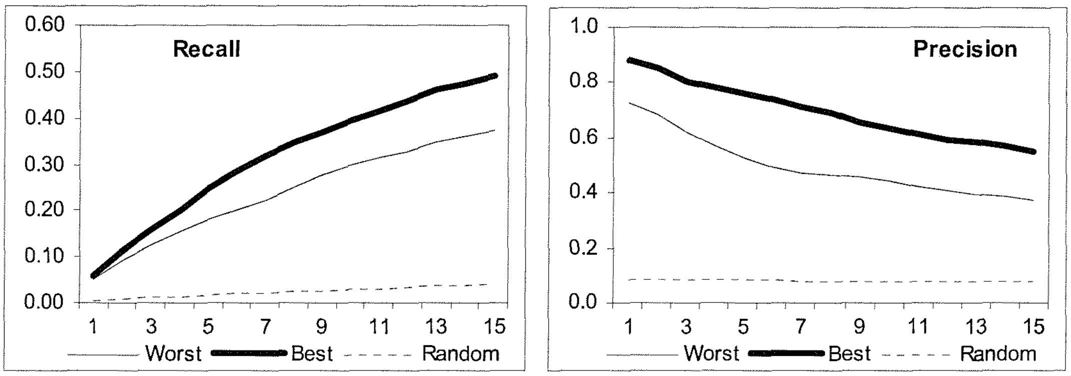


Figure 8-23 - Recall and precision graphs for Cell Biology, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall |
|---|---|---|---|
| Best | 2 | 1 | 0.3188 |
| Worst | 0 | 0 | 0.2380 |

| | Inside Classification | Outside Classification | Average Precision |
|---|---|---|---|
| Best | 2 | 1 | 0.6928 |
| Worst | 0 | 1 | 0.4986 |

Table 8-28 - Recall, precision and weights for Cell Biology, evaluating classification inclusion

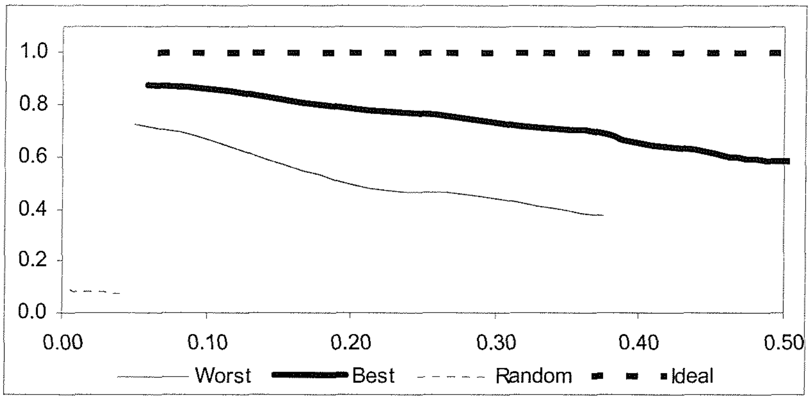
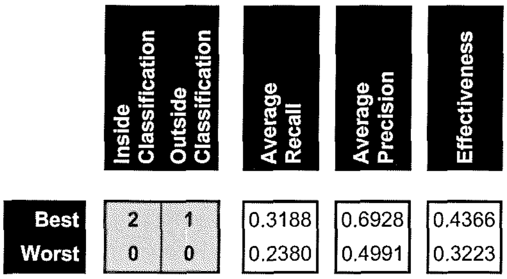


Figure 8-24 - Recall/precision graph for Cell Biology, evaluating classification inclusion



| | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|
| Best | 2 | 1 | 0.3188 | 0.6928 | 0.4366 |
| Worst | 0 | 0 | 0.2380 | 0.4991 | 0.3223 |

Table 8-29 - Best overall weight values for Cell Biology concerning classification inclusion

In the Cell Biology application, there is a very good classification of documents according to subject, with few documents assigned to each classification. This altogether can give a higher weight to information gathered from links that connect documents inside the same classification. This is in spite of having less links under this condition than links whose ends are in distinct classifications.

The worst scenario is the one where no information dependent on links is considered.

## French:

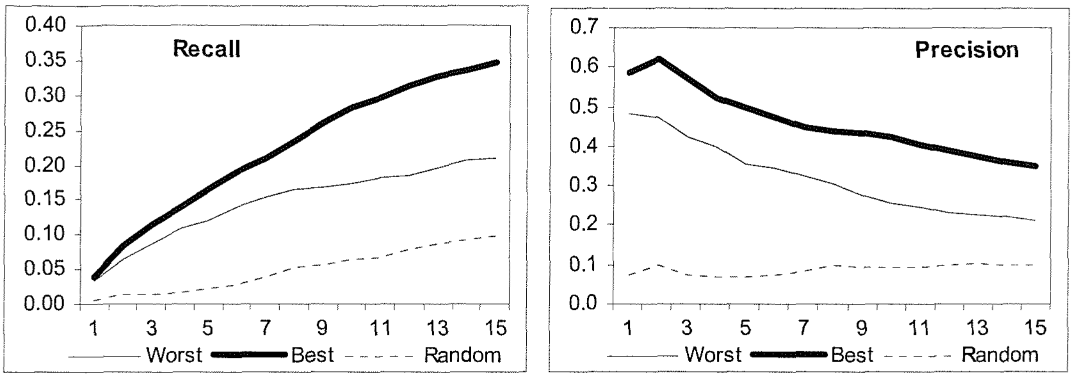


Figure 8-25 - Recall and precision graphs for French, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall |
|---|---|---|---|
| Best | 1 | 1 | 0.2227 |
| Worst | 0 | 0 | 0.1410 |

| | Inside Classification | Outside Classification | Average Precision |
|---|---|---|---|
| Best | 1 | 1 | 0.4600 |
| Worst | 0 | 0 | 0.3386 |

Table 8-30 - Recall, precision and weights for French, evaluating classification inclusion

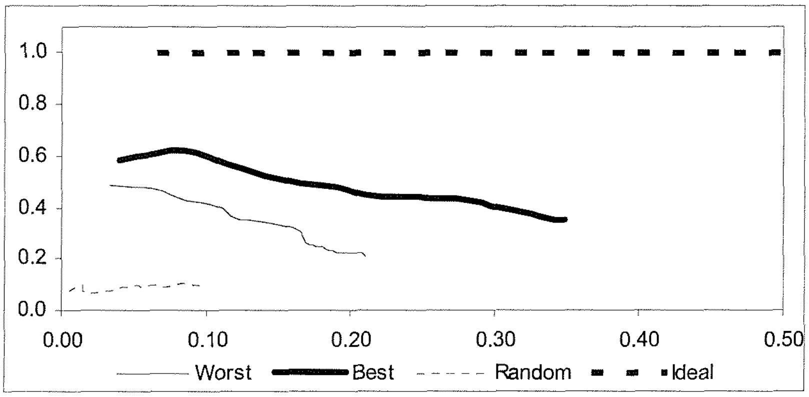


Figure 8-26 - Recall/precision graph for French, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|
| Best | 1 | 1 | 0.2227 | 0.4600 | 0.3001 |
| Worst | 0 | 0 | 0.1552 | 0.3386 | 0.2034 |

Table 8-31 - Best overall weight values for French concerning classification inclusion

In this application, the best performance is achieved for all retrieval metrics when we give the same weight to both kinds of link.

It has been observed that classifications are not so much related with the content subject of documents below them. This, together with an irregular distribution of documents and with some entries holding many documents, makes this weight assignment more reasonable. It should be noted that with a non-subject classification and with many documents per entry there are sets of documents that embrace broader concepts, and so it is possible to have more links connecting less unrelated documents, and that is troublesome for context indexing.

**French Revolution:**

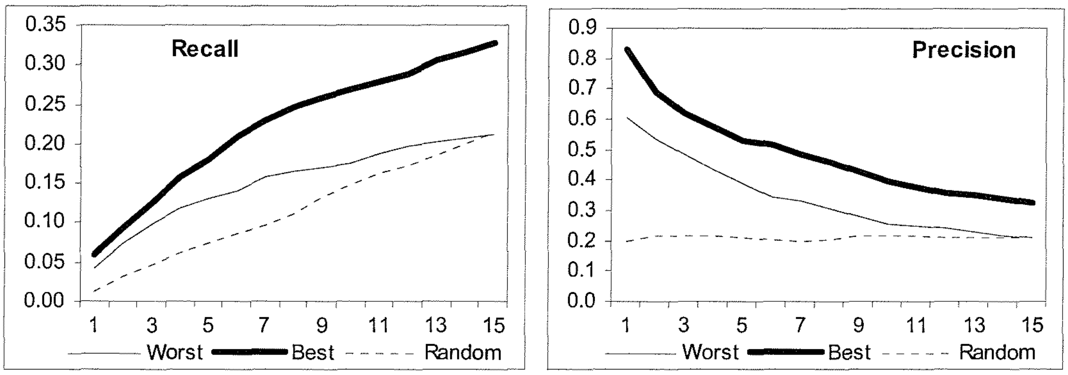


Figure 8-27 - Recall and precision graphs for French Revolution, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall |
|---|---|---|---|
| Best | 1 | 1 | 0.2229 |
| Worst | 0 | 0 | 0.1517 |

| | Inside Classification | Outside Classification | Average Precision |
|---|---|---|---|
| Best | 1 | 1 | 0.4889 |
| Worst | 0 | 0 | 0.3452 |

Table 8-32 - Recall, precision and weights for French Revolution, evaluating classification inclusion

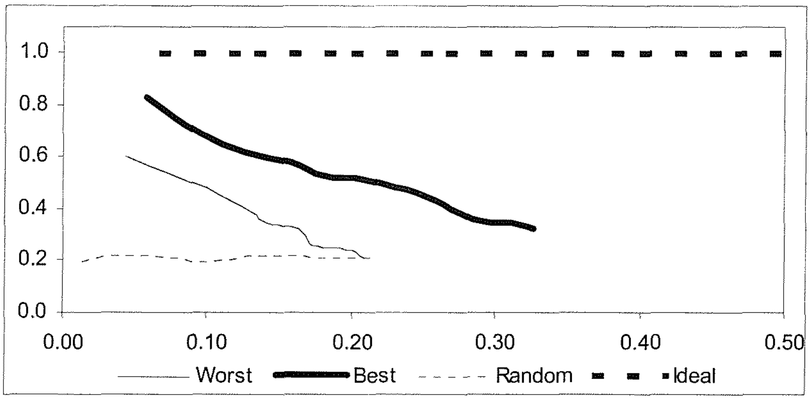


Figure 8-28 - Recall/precision graph for French Revolution, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|
| Best | 1 | 1 | 0.2229 | 0.4889 | 0.3064 |
| Worst | 0 | 0 | 0.1517 | 0.3452 | 0.2108 |

Table 8-33 - Best overall weight values for French Revolution concerning classification inclusion

Classifications in the French Revolution application are made according to subject themes. We could then expect a higher weight for links connecting documents in the same classification. However, this classification is implemented with very few themes, and the distribution of documents per entry is not the best. Under the few theme entries, there are many documents. Another problem is that a large majority of links connect documents in separate classifications.

In addition, glossary generic links in this application have the destination and original source anchors in the same document. Since we know from previous results that information from glossary generic links has lower weights, we can then expect that in turn there is not such high performance of links connecting documents in the same classification.

All the above might explain having a similar weight for both kinds of link, when we want to achieve the best performance for all retrieval metrics.

## Tulip:

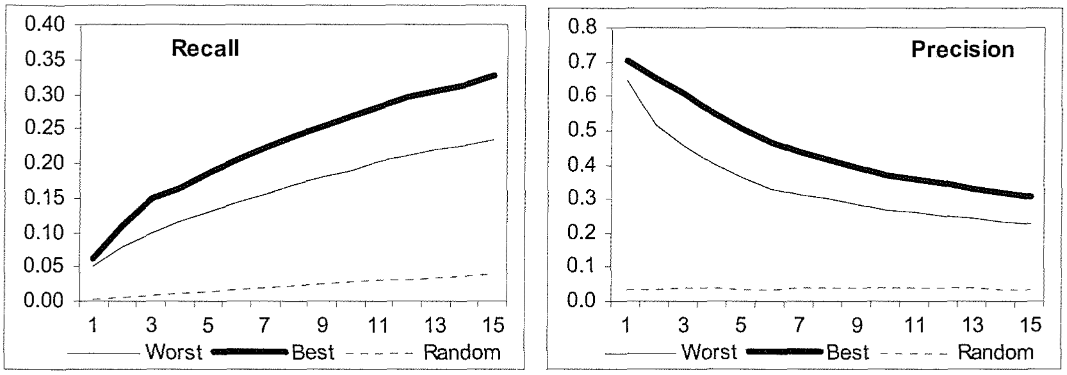


Figure 8-29 - Recall and precision graphs for Tulip, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall |
|---|---|---|---|
| Best | 2 | 1 | 0.2271 |
| Worst | 0 | 0 | 0.1631 |

| | Inside Classification | Outside Classification | Average Precision |
|---|---|---|---|
| Best | 2 | 1 | 0.4769 |
| Worst | 0 | 0 | 0.3602 |

Table 8-34 - Recall, precision and weights for Tulip, evaluating classification inclusion

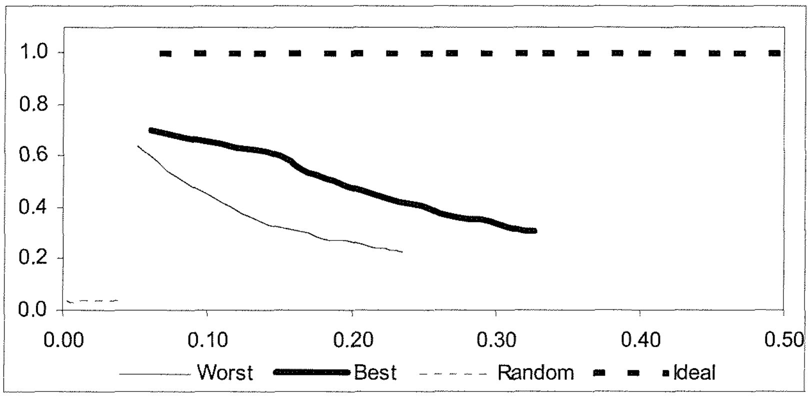


Figure 8-30 - Recall/precision graph for Tulip, evaluating classification inclusion

| | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
|---|---|---|---|---|---|
| Best | 2 | 1 | 0.2271 | 0.4769 | 0.3152 |
| Worst | 0 | 0 | 0.1631 | 0.3602 | 0.2246 |

Table 8-35 - Best overall weight values for Tulip concerning classification inclusion

In the Tulip application, the best performance is achieved for all retrieval metrics when more weight is given to links connecting documents inside the same classification than outside classification.

It has been seen that classifications in this application sort text documents according to many different subjects. This is the main reason for having this weight assignment. However, there is in contradiction a non-uniform distribution of documents. The majority of entries possess few documents, but there are some entries with many. Nevertheless, in agreement with this assignment, the bulk of links connects documents under the same subject classification.

The worst scenario is the one where no information dependent on links is considered.

**Summary:**

Finally, we can summarise how to assign multiplication weights to links connecting documents in the same and in different classifications.

| | | Best Recall | | | Best Precision | | | Best Effectiveness | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Inside Classification | Outside Classification | Average Recall | Inside Classification | Outside Classification | Average Precision | Inside Classification | Outside Classification | Average Recall | Average Precision | Effectiveness |
| Archaeology | Best | 2 | 1 | 0.2065 | 2 | 1 | 0.4535 | 2 | 1 | 0.2065 | 0.4535 | 0.2838 |
| | Worst | 0 | 0 | 0.0643 | 0 | 0 | 0.1353 | 0 | 0 | 0.0643 | 0.1353 | 0.0872 |
| Cell Biology | Best | 2 | 1 | 0.3188 | 2 | 1 | 0.6928 | 2 | 1 | 0.3188 | 0.6928 | 0.4366 |
| | Worst | 0 | 0 | 0.2380 | 0 | 0 | 0.4986 | 0 | 0 | 0.2380 | 0.4991 | 0.3223 |
| French | Best | 1 | 1 | 0.2227 | 1 | 1 | 0.4600 | 1 | 1 | 0.2227 | 0.4600 | 0.3001 |
| | Worst | 0 | 0 | 0.1410 | 0 | 0 | 0.3386 | 0 | 0 | 0.1552 | 0.3386 | 0.2034 |
| French Revolution | Best | 1 | 1 | 0.2229 | 1 | 1 | 0.4889 | 1 | 1 | 0.2229 | 0.4889 | 0.3064 |
| | Worst | 0 | 0 | 0.1517 | 0 | 0 | 0.3452 | 0 | 0 | 0.1517 | 0.3452 | 0.2108 |
| Tulip | Best | 2 | 1 | 0.2271 | 2 | 1 | 0.4769 | 2 | 1 | 0.2271 | 0.4769 | 0.3152 |
| | Worst | 0 | 0 | 0.1631 | 0 | 0 | 0.3602 | 0 | 0 | 0.1631 | 0.3602 | 0.2246 |

Table 8-36 - Summary of the best and worst metadata weights concerning classification inclusion

There are only two kinds of weight assignment, depending on the quality of the existing classification:

First of all, when a classification is made in accordance with many different subjects, with generally few documents per entry, we can assign a higher weight to links relating documents in the same classification. That is true even if some of the classifications are not the best or if some entries have a high number of documents.

Secondly, a classification can be made in accordance with different subjects, but with very few entries and too many documents per entry, or a classification that apparently does not aggregate semantically close documents. Under these conditions, we should select the same weight for both kinds of links.

It should be noted that on no condition should a weight of zero be given to any of the available kinds of links. This would not lead to optimal performance in any of the tested systems.

For all applications under these tests a lower performance for context indexing is achieved in the worst scenario, than in the tests implemented in the previous section. Here, in the worst scenario, the system only considers document description and classification. In the last section, there was only neighbour document description and classification. This suggests that descriptions and classifications of the document itself could be less relevant than the same information from neighbours. This is however misleading, since under these conditions, there is very little information for context indexing, and since in neighbours there are a higher number of words, even if sometimes unrelated to content, this can produce a better performance. With a sufficient number of words for context indexing from other metadata this situation did not occur, as has been seen from previous tests.

## 8.5 Conclusions

In this chapter three different evaluations have been made for each application, considering related metadata extracted from the abstract level. Some metadata extracted from links at document level was also considered in one of the tests. At the end of each test we found the best relative weighting of the analysed related metadata. All this was undertaken for the proper context indexing and searching of documents and/or their parts.

The first evaluation (section 8.2) analysed how document description and classification from itself and from its neighbours should be considered for context indexing. In essence it was found that classifications done according to the subject of the document's content prove to be more effective on the content indexing of documents. A uniform distribution of documents for many classifications also increased the relative significance of this metadata. Large document descriptions functioning as a title achieve a higher weight for this metadata. All neighbours metadata prove to be dependent not only on the content of the considered metadata but also on the relation they have through links to the indexed document. Links mostly relating subjects give higher weights to neighbours metadata. A more detailed analysis is presented in section 8.2.

In the second evaluation (section 8.3) we found some relations between some abstract level metadata and some link metadata from the document level. Link description and anchor selection weights were compared with document description and classification weights. Applications were differentiated by knowing whether text in link anchor selection had a proper subject relation to the document at the other end of the link or not. It was then possible to find a justification for the weight assignment to link anchor selection by analysing weight assignment to document description. In short, good text selection anchors achieve the same high weight as for document description when they work as a title. Otherwise, if document description has a different relation to the content, then it will get a lower weight in comparison with link anchor selection. Anchor selections without expressive text always possess a lower weight than the one for document description. The extent of this difference depends on the quality of document description.

In section 8.4 the third evaluation was performed. We could differentiate between links that connect documents in the same classification from links connecting documents in distinct classifications. Two types of applications appear in the results depending on whether documents are put into many different subject classifications or not. In summary, a higher multiplication weight is always given to links connecting document in the same subject classification. Links connecting documents in different classification never get a weight of zero but also never get a higher weight than the one obtained for the other type of links.

Following the evaluations made at document level in the last chapter and at abstract level in this chapter we can try to find an approximate best weight assignment for all metadata parameter. This will be discussed in next chapter. The common metadata evaluated in some of the tests in this and the previous chapter will help in this task.

# Chapter 9 - Summary of evaluations

## 9.1 Introduction

In the previous Chapter 7 and Chapter 8, several tests were run to determine the relative significance of weights for the different hypermedia metadata for context indexing of text documents. Due to computational limitations, it was necessary to divide the problem of finding the best weighting strategy into several smaller ones. In these smaller tests, similar metadata were analysed together to discover some constrained weighting strategies.

It was possible under each of these tests to justify the strategies by analysing different authoring aspects of the implemented applications. We then have a way of finding constrained weights after analysing an application.

The relative weights now have to be merged in one best overall match. In each separate test, care was taken to maintain some common elements to allow better integration of the achieved weights. With these common elements, there will be a way of comparing the weights for the constrained tests. In the following section (9.2) a strategy to find the best context indexing weighting for all metadata will be introduced. The precision, recall and effectiveness for all applications will be calculated when context indexed information (i.e., multimedia information) is access through text queries.

Section 9.3 discusses the precision, recall and effectiveness for all applications when text documents are accessed through multimedia, i.e., by using context indexed information queries.

## 9.2 Accessing non-text media from text

A summary of the individual tests on different metadata evaluated in the previous Chapter 7 and Chapter 8 is shown in Table 9-1. All three tests evaluating link context information had in common the link anchor selections and, in part, link descriptions. As for link anchor selection comparison, because destination anchors are a small minority of the anchors of all applications, we considered that the weights of source anchors in the third are an approximate average of the four anchor selection elements. This implies that the added influence of all the anchor selection weights for generic and specific links (test no.2) are an approximate average of the weight for source anchor selection (test no. 3).

In the abstract information analysis, the document description and classification test is repeated, along with link description and anchor selection in one of the tests to better determine the relations between all the weights at the two levels (see test no. 5 in Table 9-1). The weights of link description and neighbour document description have to be carefully determined. The dependency between the two in all tested applications implies that they should actually be considered as one. The nature of test no. 6 does not require a comparison with other elements.

| | | 1 | | | 2 | | | | 3 | | | 4 | | | | 5 | | | | 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Link Anchor | | | Link Description | | Link Anchor Selection | | Link Description | Link Anchor Selection | | Document Description | Document Classification | Neighbour Document Description | Neighbour Document Classification | Link Description | Link Selection | Document Description | Document Classification | Inside Classification | Outside Classification |
| | | Selection | Sentence | Paragraph | Generic | Specific | Generic | Specific | | Source | Destination | | | | | | | | | | |
| Archaeology | Weight | 4 | 0 | 1 | 0 | 0 | 4 | 4 | 0 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 4 | 1 | 1 | 2 | 1 |
| | Recall | 0.2279 | | | 0.2088 | | | | 0.2088 | | | 0.2265 | | | | 0.2079 | | | | 0.2065 | |
| | Precision | 0.4868 | | | 0.4579 | | | | 0.4575 | | | 0.4864 | | | | 0.4582 | | | | 0.4535 | |
| | Effectiveness | 0.3104 | | | 0.2868 | | | | 0.2868 | | | 0.3091 | | | | 0.2860 | | | | 0.2838 | |
| Cell Biology | Weight | 4 | 0 | 0 | 2 | 2 | 3 | 4 | 2 | 4 | 1 | 4 | 3 | 2 | 3 | 2 | 4 | 4 | 3 | 2 | 1 |
| | Recall | 0.3107 | | | 0.3105 | | | | 0.3110 | | | 0.3022 | | | | 0.3111 | | | | 0.3188 | |
| | Precision | 0.6914 | | | 0.6988 | | | | 0.6919 | | | 0.6140 | | | | 0.7022 | | | | 0.6928 | |
| | Effectiveness | 0.4287 | | | 0.4300 | | | | 0.4291 | | | 0.4050 | | | | 0.4312 | | | | 0.4366 | |
| French | Weight | 1 | 4 | 0 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 2 | 1 | 1 |
| | Recall | 0.2304 | | | 0.2251 | | | | 0.2256 | | | 0.2357 | | | | 0.2252 | | | | 0.2227 | |
| | Precision | 0.4735 | | | 0.4677 | | | | 0.4671 | | | 0.4850 | | | | 0.4668 | | | | 0.4600 | |
| | Effectiveness | 0.3099 | | | 0.3039 | | | | 0.3042 | | | 0.3172 | | | | 0.3039 | | | | 0.3001 | |
| French Revolution | Weight | 1 | 3 | 0 | 1 | 2 | 1 | 2 | 2 | 2 | 0 | 4 | 2 | 1 | 0 | 1 | 2 | 4 | 2 | 1 | 1 |
| | Recall | 0.2256 | | | 0.2242 | | | | 0.2240 | | | 0.2343 | | | | 0.2266 | | | | 0.2229 | |
| | Precision | 0.4863 | | | 0.4870 | | | | 0.4865 | | | 0.5139 | | | | 0.4908 | | | | 0.4889 | |
| | Effectiveness | 0.3082 | | | 0.3070 | | | | 0.3067 | | | 0.3218 | | | | 0.3100 | | | | 0.3064 | |
| Tulip | Weight | 1 | 3 | 1 | 2 | 2 | 4 | 1 | 2 | 2 | 0 | 3 | 3 | 1 | 1 | 1 | 1 | 3 | 3 | 2 | 1 |
| | Recall | 0.2569 | | | 0.2279 | | | | 0.2263 | | | 0.2598 | | | | 0.2336 | | | | 0.2271 | |
| | Precision | 0.5375 | | | 0.4785 | | | | 0.4769 | | | 0.5444 | | | | 0.4949 | | | | 0.4769 | |
| | Effectiveness | 0.3476 | | | 0.3087 | | | | 0.3069 | | | 0.3518 | | | | 0.3174 | | | | 0.3152 | |

Table 9-1 - Best weights summary, determined for individual groups of elements in all applications

From the individual tests, an overall weighting is obtained as can be seen in Table 9-2. These values are obtained using a strategy, as described in the following steps:

**1 –** The first step is to select an offset for anchor selections because this is a common element in the first three tests. The weights obtained for these selections in the second test should be assigned to source anchors, because the large majority of links are of this type.

**2 –** The relation between sources and destinations from test number 3 should be used to assign selection weights to destination anchors.

**3 –** From test number 1, other anchor context weights, sentence and paragraph, are obtained. We should use a multiplicative factor to the correctly bias the weights from this first test. This is taken from the relation between source anchor selections from test 3 and anchor selections from test number 1. In turn, the relation between source and destination in test 3 gives another multiplicative factor to use for the other weights, sentence and paragraph anchor, in destination anchors.

**4 –** With test number 2, link description weights can be obtained, by comparison with test 3. Care should however be taken because in the testing applications there is a strong correlation between link description and neighbour document description.

**5 –** From test number 5, document description and classification weights are obtained by comparing the anchor selection weight in this test with source selection weight in test 3.

**6 –** From test number 4, neighbour description and classification weight are obtained by comparing document description weights from tests 4 and 5.

**7 –** The weight obtained in the sixth test is used as it is. There is no need for comparison with other elements since this weight is multiplicative.

| | | | Archaeology | | Cell Biology | | French | | French Revolution | | Tulip | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Generic | Specific | Generic | Specific | Generic | Specific | Generic | Specific | Generic | Specific |
| Link Information Context | Description | | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| | Text Anchor as Source: | Selection | 4 | 4 | 3 | 4 | 2 | 1 | 1 | 2 | 4 | 1 |
| | | Sentence | 0 | 0 | 0 | 0 | 8 | 8 | 6 | 6 | 6 | 6 |
| | | Paragraph | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| | Text Anchor as Destination: | Selection | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| | | Sentence | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |
| | | Paragraph | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Abstract Information Context | Self | Description | 1 | | 4 | | 3 | | 4 | | 6 | |
| | | Classification | 1 | | 3 | | 2 | | 2 | | 6 | |
| | | Keywords | 0 | | 0 | | 0 | | 0 | | 0 | |
| | Neighbour | Description | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | | Classification | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 2 | 2 |
| | | Keywords | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Link Inside Classification | | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| | Link Outside Classification | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 9-2 - Best weights match for all elements in all applications

A summary of the chosen weights for each application and how they were arrived at is given below:

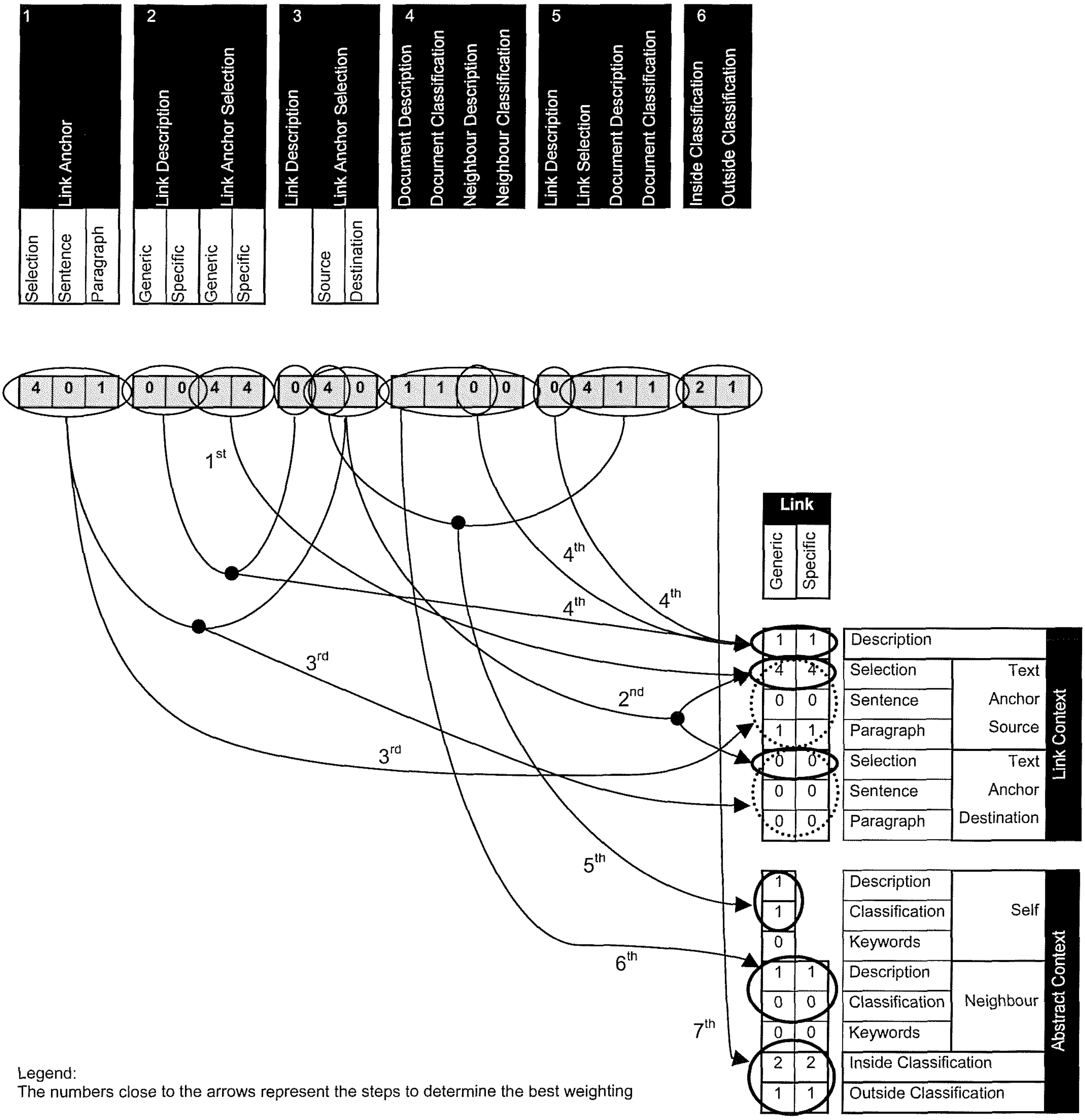
**Archaeology**

The weight for generic and specific link anchor selections is 4 in the link type test (test number 2) and 4 for the source link anchors on testing link ends (test number 3), but 0 for destinations. This implies the same weight of 4 for both types of links for selections of source anchors, but a weight of 0 for the destination anchors (see Table 9-3). For the same reason no other destination anchor information was considered. However, while testing link bounds, link selection also received a weight of 4. We then attributed a weight of one just for the paragraph of source link anchors for both link types.

During tests, a weight of zero was always obtained for link description and neighbour document description. But, while testing one of these elements in particular the weight of the other was fixed to two. Given that in the Archaeology application this information is the same for both these elements, then the weight of two will be distributed between them, giving one to each element.

When compared with the same selection weight of 4 obtained during test number 5, a weight of 1 can be assigned to both document description and classification. From test number 4, a weight of 0 for neighbour document classification is obtained.

The weights for test number 6 are unchanged since they do not depend on any other factors.



Legend:
The numbers close to the arrows represent the steps to determine the best weighting

Table 9-3 - Visualization of the steps for determining the best metadata weighting for the Archaeology application.

### Cell Biology

Link source selection gets a weight of 4 in test number 3, and since this is of the same magnitude as link selection weight in test number 2, and because there are almost no destinations anchors, the weights in test number 2 can be given for source anchors. As for destination anchors, we can assign a weight four times smaller (a weight of 1) since this is the same relation reported in test number 3. While testing link bounds, we found that selections were the only elements to consider for indexing. A weight of 0 is then assigned to all sentence and paragraph elements.

For this application, with the dependence between link description and neighbour document description, we obtained in all experiments a weight of 2 for both elements. This gives a weight of 4 to distribute between them, since in each test we had a weight offset of 2. We then attributed a weight of 2 to all link descriptions and neighbour document descriptions metadata.

In test number 5, anchor selection receives the same weight of 4, as obtained in other tests. This allows a direct relation with other abstract information: a weight of 4 and 3 respectively to document description and classification. From test number 4 a weight of 3 for neighbour document classification was obtained. Again, the exact weights obtained in test number 6 are used.

### French

The majority of link ends - link source selections - receive a weight of 2 in test number 3, and we then attribute the weights obtained while testing links types to this end of links. To the destination anchors, we proportionately assigned half of this weight, the same relation shown while testing link ends (see test number 3). As for other anchor information, we can see that during anchor bound testing a weight of one for selection was obtained. Compared with the weight obtained in test number 3 we have to give twice the weight to sentence for source anchors, 8, and the unchanged weight for destination anchors, 4.

In tests number 2 and 3 an added weight of 4 is obtained for link description and neighbour document description, but in tests number 4 and 5 we get an added weight of 3. Therefore, a weight of 2 was assigned to link description for both types of links, and just 1 to neighbour document description.

If with test number 5 we obtain a weight of 2 for anchor selection, we can then respectively assign weights of 3 and 2 to document description and classification. Comparing this with test number 4 a weight of 0 is given to neighbour document classification.

**French Revolution**

The weights obtained while testing anchors for different links types (see test number 2) were assigned to source anchors. Since the weight for destination anchors is 0, no context information relating to anchors is assigned to this end of link. This implies a weight of 0 for all this metadata. Since in test number 1 a selection weight of 1 was obtained, we will have to duplicate the weight to assign to source anchor selections.

Since the relation between source anchor and link description is one to one, the link description weights of test 2 are maintained for our best match.

In test number 5, anchor selection gets the same weight of 2 as obtained for source anchor in test 3. Then, we have a direct relation with other abstract information: a weight of 4 and 2 will be respectively attributed to document description and classification. Comparing this weight with test number 4, we choose a weight of 0 for neighbour document classification and 1 to its description. The weights of test 6 are used here unaffected.

**Tulip**

Again, the relation between source and destination anchors is 2 to 0. To all destination anchor metadata, a weight of 0 is assigned. The weights for source selections are obtained from test number 2, and for sentence and paragraph from test number 1. The relation between selection weights in tests number 1 and 3 determines a duplication of the weight for sentence, 6, and for paragraph, 2.

The weight for link description is maintained here as 2 for both link types.

In test number 5 source anchor gets a weight of 1. Therefore, we have to bias these values, multiplying them by 2. This implies a weight of 6 for both document description and classification and in turn, from test no. 4, a weight of 2 for both neighbour document description and classification metadata.

After determining the merged weights for all these applications, we ran the tests that compare content and context indexing, to determine whether the obtained performance is better than the individual tests.

| | Recall | Precision | Effectiveness |
|---|---|---|---|
| Archaeology | 0.2296 | 0.4925 | 0.3132 |
| Cell Biology | 0.3259 | 0.6826 | 0.4412 |
| French | 0.2368 | 0.4913 | 0.3196 |
| French Revolution | 0.2375 | 0.5135 | 0.3248 |
| Tulip | 0.2670 | 0.5327 | 0.3557 |

Table 9-4 - Retrieval metrics for overall best match

A comparison of Table 9-4 with Table 9-1, shows that we maximized the obtained effectiveness in this test. None of the individual test run previously achieved better results. However, if we consider precision, some of the obtained values are below the individual tests on some parameters. We should note that effectiveness is a harmonic mean of recall and precision, and for this reason it seems that precision is sometimes less significant for better effectiveness. This harmonic mean attempts to find the best possible compromise between recall and precision to maximize effectiveness, and so effectiveness takes priority, and not only precision.

The Cell Biology application was by far the best application in these tests. This had already been expected since the classifying applications, and was confirmed in the several tests run in **Chapter 8**. Good abstract level and very good link context information strengthened these results. At the other extreme are the Archaeology and French applications. The Archaeology application has a low quality abstract level but reasonable links, while the French application has the worst links but a slightly better abstract level. The French Revolution application, with similar link quality, performs better than French application does, since it has a better abstract level. The Tulip application is the second best application because of its abstract level, although its link information is of lower quality.

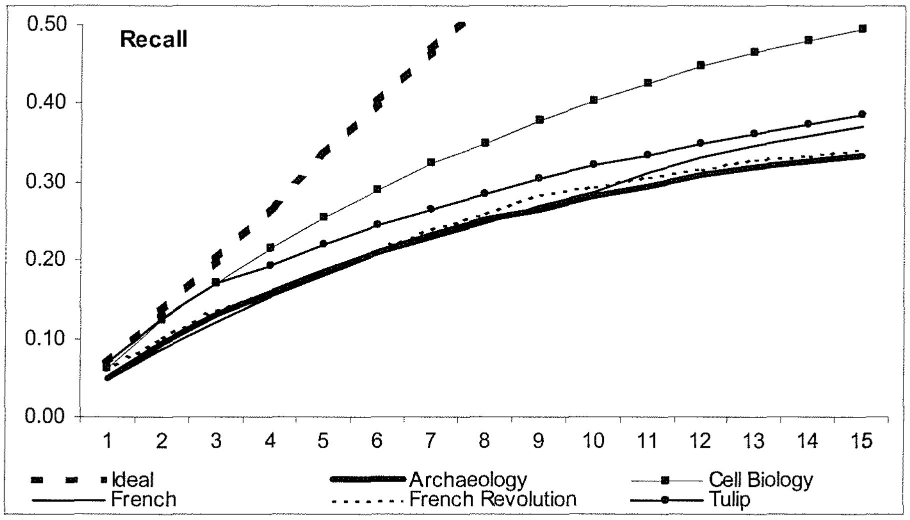


Figure 9-1 - Recall graphs of the best overall retrieval by context, for all applications

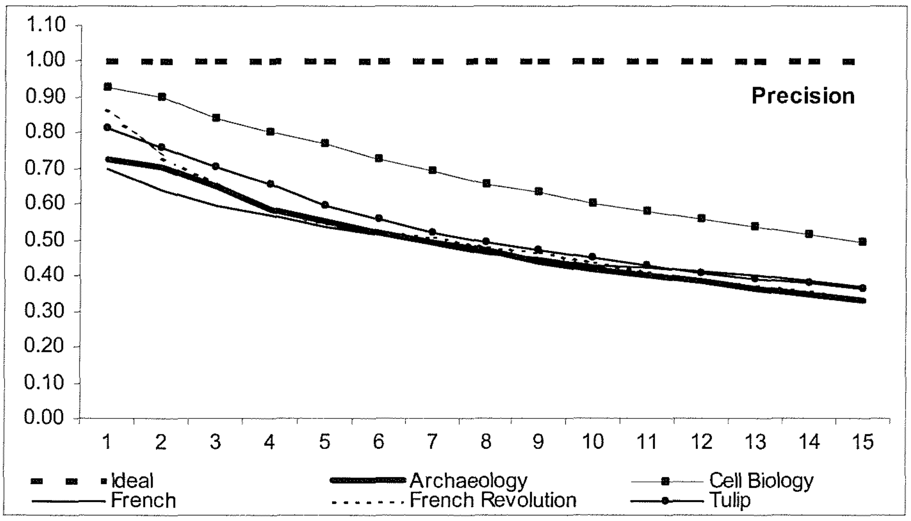


Figure 9-2 - Precision graphs of the best overall retrieval by context, for all applications

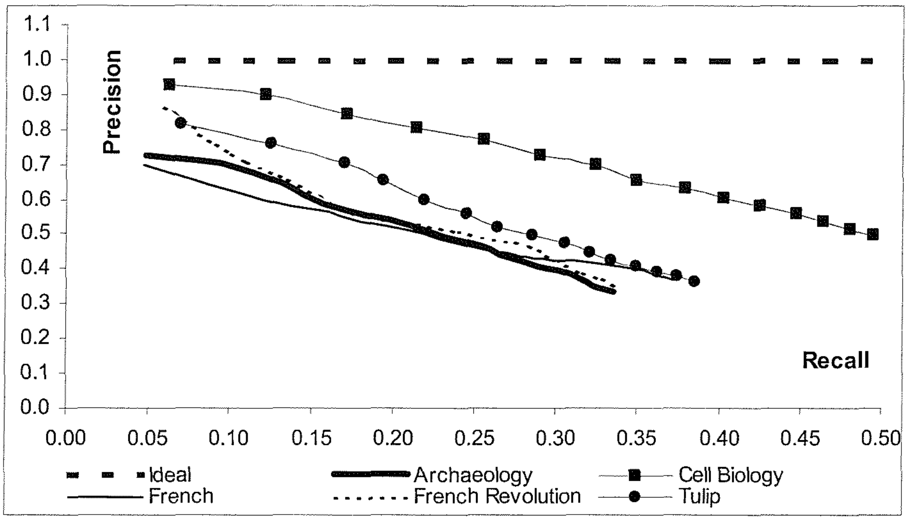


Figure 9-3 - Recall/Precision graphs of the best overall retrieval by context, for all applications

There is a clear need for careful analysis of all available hypermedia network metadata in applications. Different authors develop applications with different styles, giving varying significance to this metadata information. A way of discriminating between different information qualities is to assign weights to all metadata elements. In order to obtain the best weights, we had to consider small tests where we could vary fewer element weights, in order to reduce the complexity of the problem. The results in these tests show a correlation between the obtained weights and the quality of information used for indexing.

Afterwards, the different weights from different tests were integrated in order to find a possible best match. It would be impossible to find the best weights in just one test, and our approach is proved to obtain a better context indexing of documents. However, the ad-hoc method adopted to achieve the best overall match might not give the exact ideal best match. If the author of an application desires to use the integration proposed here, without having to create a test collection, then he/she has an approach that gets the best match in an ad-hoc way. However, for a more accurate result, he/she could vary the weights around the best match and change them in the direction of growing effectiveness, until all the weights stabilize on a maximum. Nevertheless, there is a reasonable guarantee that the addition of the weights obtained in separate tests leads to a maximum performance of retrieval.

We should also note that only text documents were context indexed and tested, and not other media, as justified in Chapter 5. It may be inferred that it is possible to achieve similar results in other media. However it is difficult to prove this, since there is no test collection where it is possible to do that, and also because metadata information is usually textual. Nevertheless, there are differences between the nature and usability of text files and non-text files in hypermedia applications. We cannot expect that non-text documents be used for the same purposes as text documents, and so the best weights for the metadata information might vary. Nevertheless, the obtained results clearly show that context indexing works for text files, by considering different metadata information with different weights.

## 9.3 Accessing text from non-text media

Evaluation of the facility to place queries within non-text multimedia files should not be difficult. The same previous strategy should be used, simulating multimedia files with text files. The text files will be indexed by content and, for all of them, the most relevant retrieval will be found and saved. Using these results, we can compare them with the ones obtained when the files are indexed by context, and therefore calculate different information retrieval metrics.

We indexed all the text documents by context, considering different weights for the metadata of the hypermedia network. We used the same weights for the best match for context retrieval obtained in section 9.1, and summarized in Table 9-2.

| | Recall | Precision | E-Mesure |
|---|---|---|---|
| Archaeology | 0.1602 | 0.3872 | 0.2267 |
| Cell Biology | 0.2090 | 0.4993 | 0.2946 |
| French | 0.1588 | 0.3586 | 0.2201 |
| French Revolution | 0.1974 | 0.4434 | 0.2731 |
| Tulip | 0.1608 | 0.3822 | 0.2264 |

Table 9-5 - Retrieval metrics for overall best match, considering the full documents as queries

From the values in Table 9-5 and graphs in Figure 9-4, Figure 9-5 and Figure 9-6, and by comparison with the ones in Table 9-4, for this specific selection of weight values, it can be concluded that using the full context indexing of documents for placing queries for information searching is not as efficient as placing selection queries as evaluated in section 9.1.

The Cell Biology application remains the best application for context indexing but the 'ranking' of the Tulip application decreases. The nature of queries makes the difference in these results. The Tulip application has many text documents with a similar content implementing a quiz. This might certainly bias either the retrieval of the context indexing or the retrieval of our relevant documents obtained from content indexing.

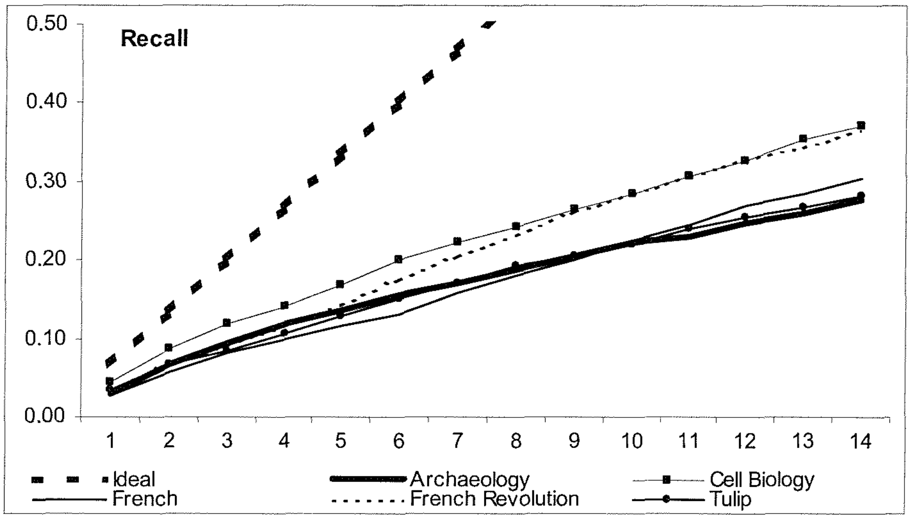


Figure 9-4 - Recall graphs for the retrieval by context, considering full documents

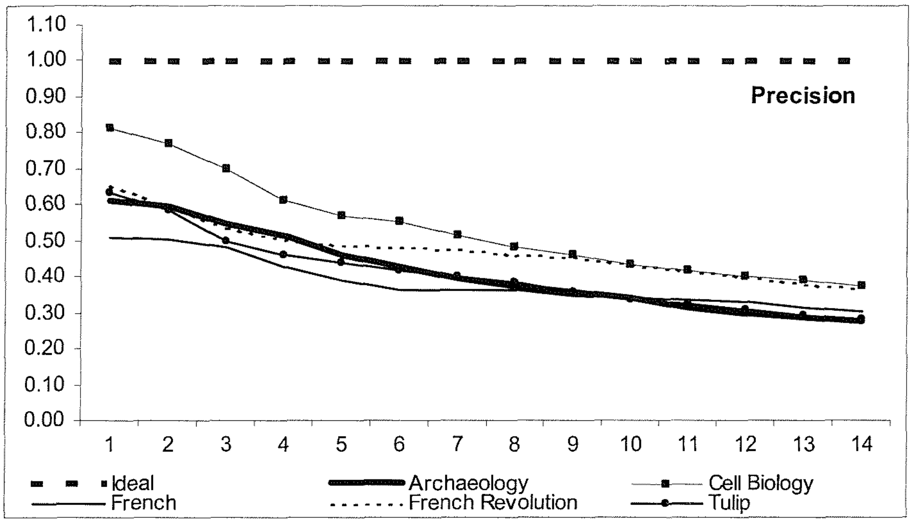


Figure 9-5 - Precision graphs for the retrieval by context, considering full documents

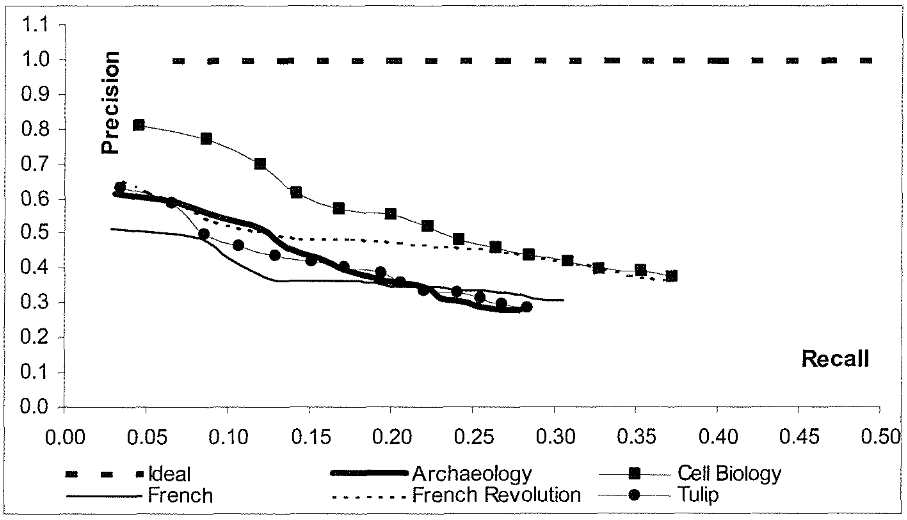


Figure 9-6 - Recall/Precision graphs for the retrieval by context, considering full documents

We don't however expect that with a different weight selection we would obtain a better performance. A justification for the lower values obtained here might again be the nature of the queries used. The queries in this test use all the words of the full indexed document, while for selection queries there are fewer words. A narrower subject is obtained with a selection query than with the full document and this implies better results.

## 9.4 Conclusions

In this chapter the relative weight assignment to the different metadata studied in the previous two chapters has been merged to achieve a best overall weighting for the best context indexing of documents.

In section 9.2 the common metadata evaluated in the separate tests, with the purpose of correlating the different tests, was studied to find a strategy of getting a better weight assignment for all metadata. The strategy, which can be applied in principle to any possible application, was explained, and for each evaluated application a small description was given in more detail. This strategy proved to achieve better weights since with it we got better results than the ones obtained in each of the individual evaluations. A closer inspection of the evaluated link description and neighbour document description metadata show a strong correlation. This was taken into consideration in the quest for a better weighting strategy.

Accessing context indexed documents through text queries (section 9.2) achieved better results than accessing text documents through context indexed queries (section 9.3). This can be due to the nature of the query itself and not to the way the best weighting was found.

Finally, we can conclude from the obtained results that if different applications are developed in different styles then extracted metadata has to be considered in different ways according to its own description qualities and to its relative importance when the description qualities of other metadata is considered.

Of course this strategy does not achieve the same results for all applications, but we are sure that it gives a better consideration to the weighting and it can achieve a better performance for each application. This best performance is then dependent on the overall quality of all the metadata. Between all the applications, the one that achieved best results was Cell Biology, due to the quality of all its metadata. The Archaeology and French application were the ones to achieve the worst results, due to the poor quality of one or other metadata element.

This chapter has demonstrated the importance of this multimedia context indexing evaluation. Other studies have already used link information, but very few have used any sort of abstract information and none (Dunlop et al. 1993, Frankel 1996, Smith 1997, Harmandas et al. 1997, Amato et al. 1998, Mukherjea et al. 1999, Srihari et al. 1999) has distinguished between different applications styles when extracting metadata.

The following chapter will show same implications for the web of the integration of information retrieval and hypermedia proposed in this thesis. The design of the integrated model and results obtained in this chapter leads to some improvements that could be achieved in browsing and retrieving approaches for the web.

Chapter 11 will further expand on the overall conclusions of this thesis.

# Chapter 10 – Implications for the Web-based multimedia information retrieval

## 10.1 Introduction

Some of the conclusions from the work we have undertaken may be extended and used to draw implications for web implementation.

The research conducted in this thesis is novel insofar as it develops a new method for integrating information retrieval with hypermedia, by considering content and context at both document and abstract level. The different aspects of this integration are interrelated but certain aspects merit particular attention:

- Our work used a new information retrieval tool that was independent of any system. The document and system interface were the only parts aware of the integration, and they could be changed if another hypermedia system was used;
- The capability of context indexing multimedia information was extended in order to enable access to non-text documents from text documents and vice-

versa. The majority of evaluations were carried out on the different aspects of this type of indexing;

- The indexing of anchor contexts was enabled, notably generic links, along with user specified selections. The purpose of this was to enable more focused retrieval of information, by making use of the hypermedia authoring context;
- The system also enables the integration of link and document searching. The goal was to make the transition between link and document searching smother;
- The model also suggests approaches for improving the construction and classification of documents, whatever media is considered.

## 10.2 The need for a new open, integrated information retrieval tool for the Web

With the current growth rate of the web, it will be difficult to keep up with the centralized indexing of all web pages. Searches on target localized user groups can be the solution for better searches. The information need of a known user group can be easier to predict, and the data may be better defined. However we will be restricted to a small view of the web. So, besides splitting the indexing problem, we have to find a way to allow for resource discovery outside this restricted search. Some support for distributed queries and indexes will have to be provided. This actually makes quite good sense, since web site owners are interested in having their web site properly indexed. Due to the diversity of information formats, hypermedia structures, security restrictions, etc, this indexing should be done locally, since authors understand the information they use better and the needs of their potential readers.

Some research in the open hypermedia community (OHSWG 1998) has been undertaken in order to create open hypermedia services (Grønbæk et al 1997, Anderson 1997, Carr et al 1998, Carr et al. 2000). However there has not been such an implementation of something similar to information retrieval services.

There is therefore a need for an open information retrieval tool that provides indexing services, which can be accessed within and outside of the web site context.

We have developed a tool, which in concept is open and facilitates functionalities that allow for the distribution of indexes and queries. It is not dependent on any file format or specific application solution, and it allows for the indexing of any type of data. This is done by separating the indexing functionally from the document interface and system interface. The provided functionalities focus on the maintenance of indices, i.e., the construction, addition, removing, etc of indices of any kind of information, such as all documents, or document regions, user selections, link context or other context, etc. If there is a way of interchanging indexes, we can merge indexes from different web sites. Moreover, we can merge a user specified index satisfying particular information needs, in the same way as we can already have tailored hypermedia structures stored in user link databases. The information retrieval tool we developed, besides allowing for traditional query processing into the content of the index, allows for querying on the structure of the index. At present these queries are about the availability of indexing and the type of indexing, which could be further extended to consider other type of index structure.

However in the web, the importance of having a centralized repository of some kind of information should not be rejected. It should be noted that with a totally distributed approach it would be easy to dismiss relevant information, and also allow for the easy spamming of retrieval. A centralized solution should be more concerned with the quality assessment of site indexes, rather than indexing the content of all web sites. Redirection of queries to site index services according to their importance in their subject areas, among other services, should be the goal of such resource discovery sites. To use the analogy of the librarian, he/she does not find for us the information we are looking for, but he/she can help us to discover how to look for that information, giving us valuable advice.

## 10.3 Context indexing multimedia on the Web

There has been interest in the exploration of the interaction of textual and non-textual information on the web.

Some work has been done in using neighbour documents, textual regions, labels or captions to index and retrieve images (Frankel 1996, Smith 1997, Harmandas et al. 1997, Amato et al. 1998, Mukherjea et al. 1999, Srihari et al. 1999). All of these methods used different metadata retrieved from web pages, and in one way or another assigned a weight to different elements. However the weight assignment was unclear because of a lack of proper evaluation.

For the correct context indexing of multimedia there is a need for correct weight assignment, and making that assignment according to the different developing styles of the authors, is of a key importance. Previous research lacks a proper evaluation concerning these aspects, and that has been rectified and demonstrated in our research.

Due to the diversity of styles in the web, for a global index approach, difficulties arise regarding the correct weight assignments. Nevertheless, it should be possible to derive from our research some rules to infer the style of developing in target sites. Ultimately different styles should be automatically taken into account throughout integration, mainly during indexing. But again, this problem is not so acute if we index information on target-localized sites for different user groups. This solution gives better searches and better metadata for context indexing since the information needs of a known user group can be easier to predict, and the data may be better defined.

An aspect to consider is the different nature of documents in the web and in the hypermedia system used for our evaluation. Our system uses metadata from many different sources relating to our documents, but in the web non-text multimedia is published, in relation to documents, in different ways. Two possible forms are: *inline* and *reference*. With the *inline* form different multimedia types are published in the same documents whereas with the *reference* form non-text multimedia is accessed through links The surrounding context can then change if one form or the other is present. Images are usually inserted inline with text documents, but this is not always the case for other media.

Work on context integration to improve retrieval in the web usually consists of two types, though not exclusively, namely: document structure analysis for multimedia searching and hypermedia structure analysis for precision enhancement.

The bulk of the work for multimedia access has been done on inner document structure analysis (Frankel 1996, Smith 1997, Mukherjea et al. 1999, Srihari et al. 1999). These researchers basically considered the structure of HTML documents and used different parts of it for context indexing. Image file names, image path, image captions, text on the vicinity of images, HTML titles, anchor text, and captions on other images are among the features which are used. Mukherjea's Amore system has the most extensive consideration of these different metadata and some evaluations on weights.

Hypermedia structure analysis, well described in Chakrabarti et al (1999), is not usually used for multimedia context indexing but rather for improving ranking and precision. In work developed as part of the Google system (Brin et al. 1998) a link anchor improves precision and as a side effect allows the retrieval of multimedia, although this has been though not properly explored. However in Harmandas et al.'s (1997) work, which is more focused on multimedia access, link information is explored more extensively. Improving indexing for text with entrance paths (i.e. set of document in the path that leads to the indexed document) in Mizuuchi et al. (1999) also suggests that is possible to use other structures for context indexing. In addition, the notion of hubs, as a complement to authority web pages, studied by Chakrabarti, has not yet been used for multimedia indexing. Some work could be done using this structure information.

Considering our work and all the previous approaches, work on the integration of information retrieval and hypermedia in the Web for multimedia access should be carried out by analysing both document inner structure and hypermedia link structure.

Compared with our approach, adaptations must be made to the metadata being considered, but the spirit remains the same. All the relevant metadata relating to multimedia, whether document or hypermedia link structure, should be evaluated for context indexing, considering the existence of different development styles both for documents and hypermedia.

At present, the language for the web is still dominantly HTML, which unfortunately presents many restrictions, and in the end we would not result in a proper integration, as happened with the other examples given here. But with the advent of new standards for web publishing, such as XML, XLink and XPointer, better and neater integration is possible.

XML (Bray et al. 1998), a restricted form of SGML, supports a better and more predictable definition of structured information and also a better merge of structured and non-structured information in documents. This allows consideration of more and better information from within documents. A more methodical organization of the data in documents eases the process of data mining for resource discovery.

XLink (DeRose et al 2000) and XPointer (Daniel et al. 2000) are related in a way to open hypermedia. Roughly, XPointer is concerned with locating regions of interest in documents, and this can be structured with XLink, which describes navigational hypermedia expressions.

With XPointer it is possible to better define the resource locations rather than relying on the simple absolute position available in tags inserted within HTML documents. If these located regions are related to multimedia, these relations can be used and weights inferred for context indexing. It is not compulsory, but usually the way we relate XPointer regions is by using links according to the XLink standard. XLink allows for the creation and description of different types of bi-directional links, stored within or outside documents. The traversal of links can deploy different actions and may be initiated by users or automatically. All these different aspects, that are further developed than in the Microcosm model, can also be considered for multimedia access, and should be evaluated in the same way as we did with the open hypermedia system Microcosm.

For instance, with standard HTML there is no link description or link destination selection, but with XLink and XPointer that is possible. These new standards are of key importance for considering more and diverse metadata, and that was not possible before with HTML alone, as was seen with previous solutions.

In addition, in the Web, image, and also general multimedia, access is possible through the use of classification browsing searches such as Corbis (http://www.corbis.com/), Yahoo's Image Surfer (http://isurf.yahoo.com), and AltaVista's Multimedia Search (http://www.altavista.com), among others. This subject classification is already used in our system for accessing multimedia by browsing, and many Web sites also already feature a structure synopsis of their content, or at least an organization into different directories. We have properly used and evaluated keywords from this level for the context indexing of images. This approach has not been fully implemented on the web and such integration could be quite important in the absence of other metadata. Mukherjea et al. (1999) considered context keywords for defining semantics but no concept relations were inferred. But a model for concept representation was proposed by Amato et al. (1998) and this shows a way forward to considering text context. However this context was only used to aggregate documents into interrelated concepts for navigation; little support is given in the model for relevance retrieval and no support is given for ranking documents according to a possible text feature. They also lack a complete implementation of the system and an evaluation.

Nevertheless the abstract information, or meta-information on top of documents, needs much improvement, and classifications proposed by Dublin Core (Weibel et al. 1998), use by Dewey Decimal Classification (DDC 1998), or others, implemented in RDF (Lassila et al 1999), could serve such purposes. With these new resource description standards, better and neater integration of hypermedia and information retrieval in the Web is possible.

Besides concept and other types of classification, RDF allows for other metadata consideration and control. This can be kept within documents or separately, can be resource descriptions about them or about their relations with others, and in addition, it provides interoperability between applications to share and exchange metadata. It will then be easier to integrate, merge, search, or mine metadata from different web sites. More contexts can then be considered not only for multimedia indexing but also for other information retrieval enhancements.

Moreover, our work uses classifications to discriminate between links and their importance for context indexing. This idea can be extended to the web, by allowing not only the discrimination of links according to the location of their ends in the subject classification, but also in other possible structures such as author classification, Web site location, path location, etc.

In summary, the way we see the integration of hypermedia, in this case the web, with information retrieval, is at three interconnected levels: at document inner structure level, at document outer structure or hypermedia link structure level, and at the abstract level, as illustrated in Figure 10-1.

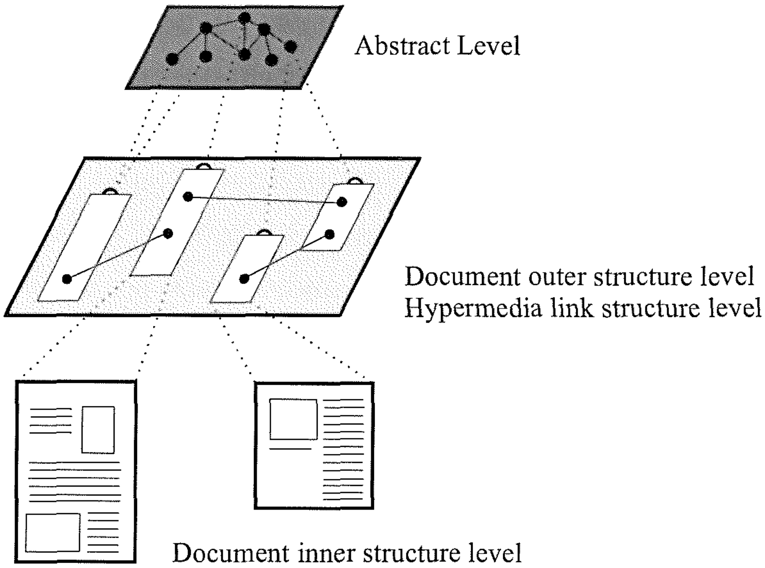

Abstract Level
Document outer structure level
Hypermedia link structure level
Document inner structure level

Figure 10-1 - Three structure levels available on the Web

This is somehow different from the evaluation in our implementation. There, the integration was studied at two levels: document outer structure or hypermedia link structure level and the abstract level, were the document outer structure level represents all the interconnections possible with links, i.e., the link structure, and the abstract level is a layer done with all classifications and descriptions about documents. The third level that we propose here comes from the fact that Web document content has a more complex structure, which can be considered for context indexing. This structure can be enhanced with XML, and that advantage can be used for the better integration of hypermedia and information retrieval.

## 10.4 Abstract level or meta-information improvement

With our system we have proved and evaluated the importance of having a properly defined abstract level for accessing information. But within the present Web structure a mechanism for resource description is absent. However there are many problems in the generation of that metadata: reliance, trust, interoperability, author laziness, time consumption, etc. But an equally significant problem is the domain specific solutions adopted by many systems, including the system we used.

RDF, Resource Description Framework, is a foundation for processing metadata, recommended by the World Wide Web Consortium in an attempt to introduce a language for machine-understandable descriptions of resources on the Web, without making assumptions about any domain specific semantics.

These metadata can be input in various ways: by authors, by users or groups of users (Grønbæk et al. 2000), or then automatically generated (Jenkins et al. 1999). Manual input is however time consuming and it is infeasible to classify the entire web in this way. Such techniques not being compulsory allow many authors still to choose not to include meta-information. This problem was already noted in Microcosm, and a solution for this, also on the web, might be the automatic or semi-automatic generation of metadata, such as term comparison of documents with classifications, identifying in the process other metadata (Jenkins et al. 1999), or metadata assignments of assessed documents to documents connected to them by links (Marchiori 1998).

But of course RDF is no solution on its own for creating a web of trust. The concentration in small communities of use might give more credibility to information in specific subject domains. Conceivably RDF could also be combined with addressing schemes such as XPointer to provide a finer granularity for meta-information.

## 10.5 Context and content search integration

One of the implemented developments in our integration of information retrieval and hypermedia was the allowance for the indexing of context information along with the content of documents. Different anchor context bounds can be indexed for different types and locations and these can be considered along with user specified selections.

The purpose was to allow for a more focused retrieval of information, by using the hypermedia authoring context in information retrieval. A side effect of this integration is a mechanism of link discovery with provided user selections. This double effect of document region searching and link searching, along with all-document searching, makes possible the merger of these two methods of resource discovery in one user interaction. The goal of a smooth transition between structure and content searching can then be achieved.

This context and content search integration can be better extended to the web with the new emerging standards of XPoint and XLink. With HTML, implementation could not be fully transferred to the Web, since the link structuring mechanism only allows for the use of specific / button links, which must be parsed with the document.

As explained earlier, XPointer specifies a language that allows for the support of addressing into the internal structures of XML documents. XPointer provides for specific reference to elements, character strings, and other parts of XML documents, and that can be used for specifying link anchors, user selections or other structures. All these context resource locations can be content indexed for resource discovery, in the same way we implemented in the Microcosm-based system, allowing access to atomic units of information determined by authors while editing those structures. There should also be a way of discriminating between different structures for content indexing, by classifying them according to the link characteristics defined with XLink, the location methods used in XPointer to find elements (e.g.: absolute location, relative location or string match), or other different sorts of metadata concerning XPointer elements. Conceivably RDF could be combined with XPointer to provide this metadata construction.

HTML provides a structuring mechanism for document content, but this is better explored in XML. This new structuring mechanism can better separate documents into their atomic information units, which have finer information granularity than traditional flat documents and then are more suitable for indexing, as far as information span is concerned. This document structure hasn't been considered during indexing in our implementation since our documents were flat, but in the web, and more so with the advent of XML, this information must be considered. But care must be taken if many of those information units are indexed separately from document content. Work on the hierarchical index construction (Lee et al. 1997) of documents and their querying suggests some solutions to overcome certain problems, such as index size and merging.

In addition, with extended indexing, the natural retrieval of links, and then structures, is possible. However user interaction for resource finding on the web is somewhat different, compared to our Microcosm-based system. In the later, the user makes a selection for showing links, notably generic links, and also for keyword information searching. In HTML there is no support per se for generic links. To overcome this problem, various solutions have been implemented. Methods for storing, managing and finding the location of this type of link are shown in work done by Hartman et al. (1997), Carr et al. (1998) and Kaindl et al. (1999). However the style of user interaction for link navigation, where a selection is required, similar to Microcosm, was only possible in the first version of the DLS system (Carr et al. 1995). Unfortunately at present, as far as we know, there is no support for finding generic links, or other types, through user selections. The user can only take a passive attitude towards linking, but the applicability of generic links might require a more active role (Hall 1994). At the limit, for hypermedia, every significant word could be the source of a link, but to have everything highlighted can be as bad as having nothing.

XPointer, with its string-range function, will allow for the location of content oriented anchors for links in the same style as source anchors of generic links. However the standard just defines exact string matching, and no fuzzy matching is allowed. Since this query-type facility is not available, generic link search will have to be extended in the same way as implemented in our system. We can context index those anchors in order to accommodate better retrieval of generic links, or we could extend our model by using a thesaurus. This would enhance the notion of generic links to a relation between a concept, made up of different definition keywords, and a resource. But the way the user interacts with the link must also be revised. It should be required from the user a more active action for generic link following. A query-type facility where selections serve both goals of link finding and content searching, similar to what has been implemented on our system, might be an ideal solution for resource discovery on the Web. This double effect of document region searching and link searching, along with full document searching would also favour in the Web the merger of these two ways of resource discovery in one user interaction. We would be closer to achieving the goal of smoother transition between structure and content searching.

Finally, XML structuring mechanisms allow the representation of structured and unstructured information in the same document. In addition, the structuring mechanism on XML is not as rigid as for databases. There has been a lot of work in finding a proper language for querying XML document structures, such as XML-GL (Ceri et al. 1999) and XML-QL (Deutsch et al. 1999). Many more were proposed at QL 1998, and for this reason a new standard is in progress (Chamberlin et al. 2000). However, we think that special care must be taken in the way structured and unstructured information is considered for querying. When the structure of the information is not known or when there are discrepancies between different XML DTDs in different implementations, problems might arise in properly specifying a query. So there must be a way that besides keyword search on the content of documents, in the same style as is possible today, also keyword searches the structure vocabulary and DTDs of documents. Florescu et al. (2000) suggest some ways forward in addressing this situation.

# Chapter 11 - Conclusions

From the research that we have been carrying out on the integration of information retrieval and hypermedia we can point out some achievements that have been attained

Supported by other research carried out on the integration of hypermedia and information retrieval we have developed a first model on the integration of both types of systems. We made possible the retrieval of any multimedia document through the use of text descriptors. This is not a new approach, however our implementation has improvements to this strategy. The consideration of more specific information items inside documents (like anchor selections) and link description was one of the first innovations. In addition abstract level information has also been considered, allowing the retrieval of multimedia even under low link availability circumstances.

Moreover, this integration had in mind the use of existing systems and wanted this integration to be transparent.

Supported by preliminary ad-hoc results obtained with the first implementation we further progressed our research for a better integration between hypermedia and information retrieval.

A new model for open information retrieval has been implemented. This allows a flexible and transparent integration with existing systems, as demonstrated with our Microcosm-based system.

One of our innovations in this improved model has also been the better integration between the information retrieval and hypermedia methods of working and accessing information, by considering more structure for retrieval, and more content for hypermedia. A more specific retrieval of information was enabled by the indexing of user selections, along with the link context. This link context indexing also allowed a better identification of generic links and all this in one phase of user interaction. This enabled the integration of link and document searching. Some suggestions were also given to improve abstract level information.

Special care has been taken in the multimedia context indexing with better use of context information from links at document level and classifications at abstract level. Access to non-text documents from text documents and vice-versa has also been enabled with the integration

A smooth transition between retrieval and navigation was achieved in the end whatever media is considered

One final innovation in the present work was the approach adopted to evaluations. From the results obtained we conclude that if different applications are developed in different styles then extracted metadata has to be considered in different ways according to its own description qualities and to its relative importance when the description qualities of other metadata is considered. Methods for metadata weight assignment were then proposed, which proved to achieve better context indexing results.

With new emerging standards for the Web, we suggested ways in which the work could be ported to a Web-based implementation, for a better integration of information retrieval and hypermedia in the Web.

We now name a few directions for future work:

- Elaboration of a model for open information retrieval services in the Web. Query representation, distributed processing and routing, possible index merging for different types of information and result merging should be taken into account. Ultimately each site should be responsible for its own indexing and then centralized solutions could make quality assessment of site indexes.
- Permitting, the context indexing of information in the Web by considering the three structure levels: document structure, link structure and abstract structure. This should be implemented with the new W3C standards for representing information in the Web. Proper evaluations on different type of Web application could also increase the performance of future and present systems.
- Abstract level improvements in the Web is possible, considering the new standards to share and exchange metadata among different sites. It is possible to find a whole new ways of representing information and accessing them across platforms. But much work has to be done at this level. Authors do not accept tedious development takes, and tools for semi-automatic classification constructions might have to be used.
- Finally we intend to find ways of merging structure and content seeking. Content information retrieval facilities should be better merged with the present navigation in browsers and also allow the use of structure information retrieval.

# Appendix A - Microcosm links statistics

In this appendix will follow a numerical characterization of different Microcosm applications according with link availability and distribution for constructing context descriptors for non-text media and text media. For these reason we basically present information for two kinds of interconnection: links from text documents to non-text documents, and links from text documents to text documents. To be more precise, what we are actually measuring is the number of text link ends. So we should notice that specific links, and some times generic links, connecting text documents will be counted twice, since a these links have two ends on a text files. This is the reason why adding the number links connecting text files with the number of links connecting text to non-text is higher that the total number of links present in the application. This is not an invalid assumption, since from the point of view of the document we can see a link; whatever the document is a source or a destination of it. So if we would like to find link average per document we should have this in count.

For every application we have two graphics showing the distributions of links from text files to text files, and links from text files to non-text files. Essentially they show how many documents we have with a certain number of links. But each of these graphics is effectively divided in three. From left to right we have the distribution for all the available links, for generic links and then for the specific links.

Archaeology

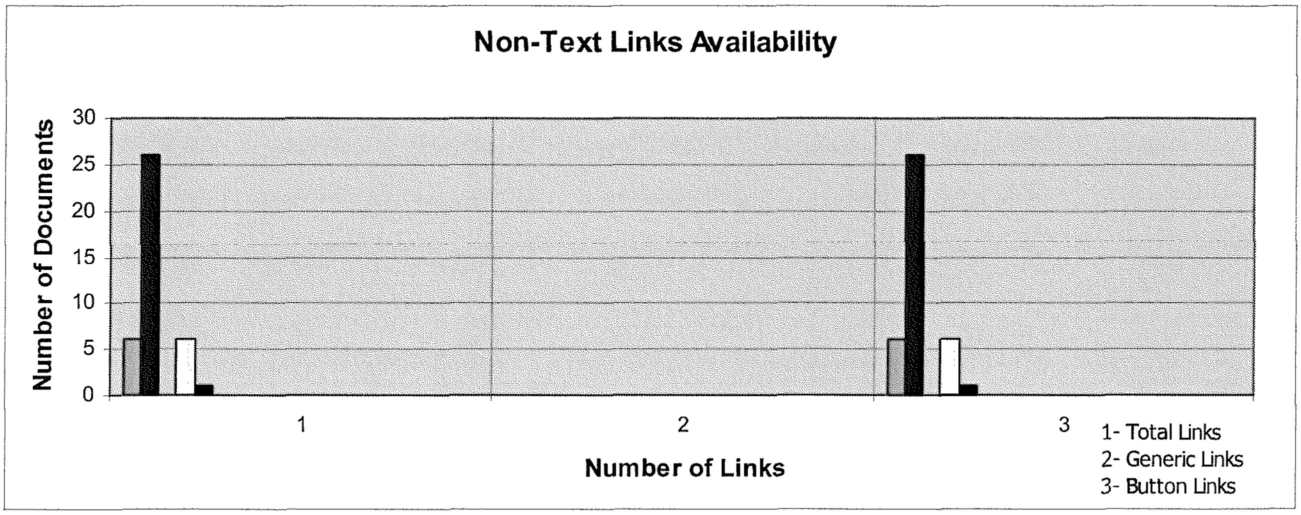
| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 9,4 | 10,5 | 10,4 | 10,5 | Links |
| 1,9 | 1,2 | 2,2 | 1,0 | Links on non-text |
| 11,8 | 11,0 | 12,7 | 10,9 | Links on text |
| 0,3 | 0,8 | 1,4 | 1,0 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 0,5 | 0,9 | 1,4 | 1,0 | Generic Links on text |
| 8,8 | 10,0 | 10,0 | 10,1 | Button Links |
| 1,9 | 1,2 | 2,2 | 1,0 | Button Links on non-text |
| 11,0 | 10,5 | 12,3 | 10,4 | Button Links on text |

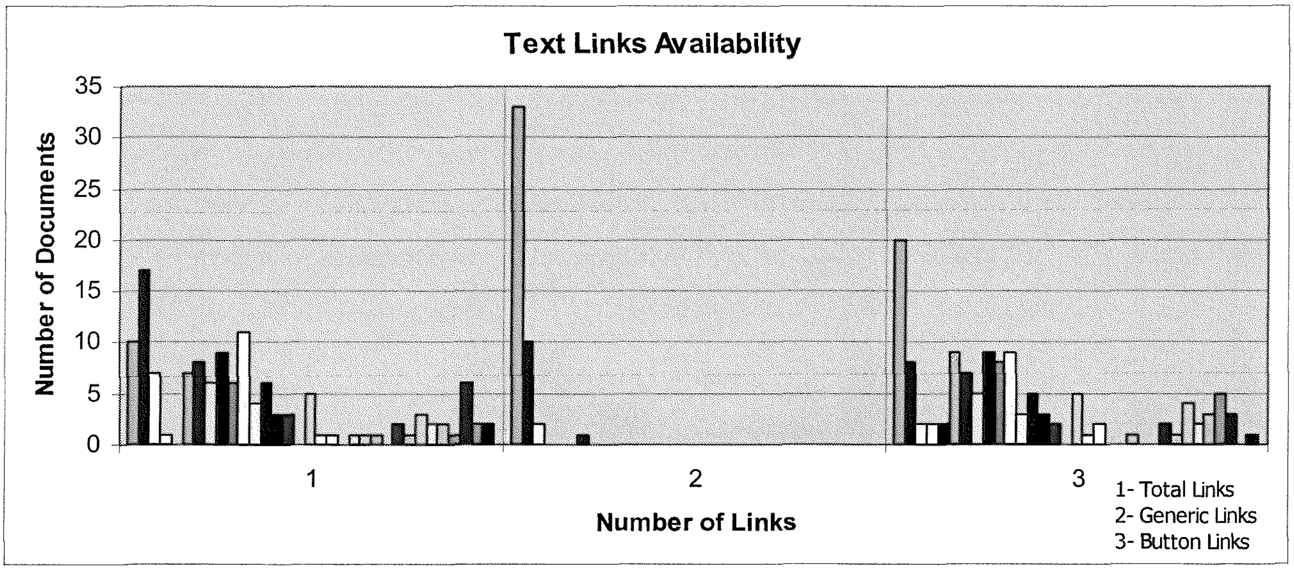
| | |
|---|---|
| 189 | files |
| 975 | links |
| 33 | generic links |
| 25 | local links |
| 914 | button links |

Valid links to build context descriptor:

| | |
|---|---|
| 45 | non-text files |
| 87 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 87 | button links on non-text files |

| | |
|---|---|
| 144 | text files |
| 1705 | links on text files |
| 66 | generic links on text files |
| 50 | local links on text files |
| 1589 | button links on text files |





Caerdroia

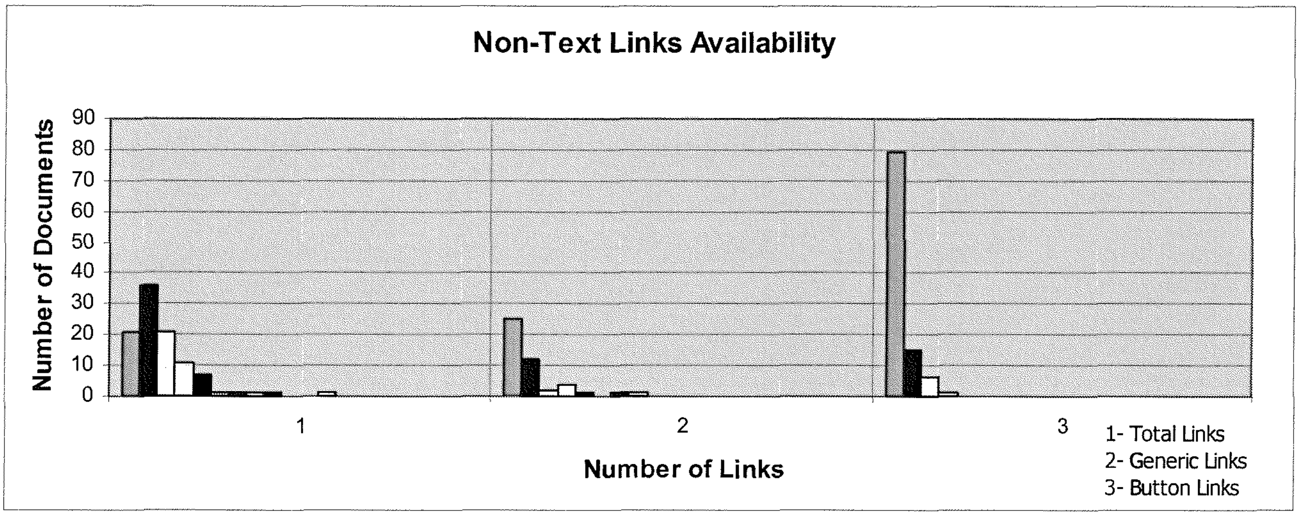
| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 2,6 | 3,4 | 3,3 | 3,6 | Links |
| 2,0 | 1,9 | 2,7 | 1,8 | Links on non-text |
| 5,4 | 6,5 | 5,4 | 6,5 | Links on text |
| | | | | |
| 1,0 | 1,9 | 2,6 | 2,3 | Generic Links |
| 0,7 | 1,3 | 2,0 | 1,6 | Generic Links on non-text |
| 2,6 | 3,1 | 4,2 | 3,0 | Generic Links on text |
| | | | | |
| 1,2 | 2,4 | 1,5 | 2,6 | Button Links |
| 1,0 | 0,8 | 1,3 | 0,6 | Button Links on non-text |
| 2,3 | 5,4 | 2,3 | 5,5 | Button Links on text |

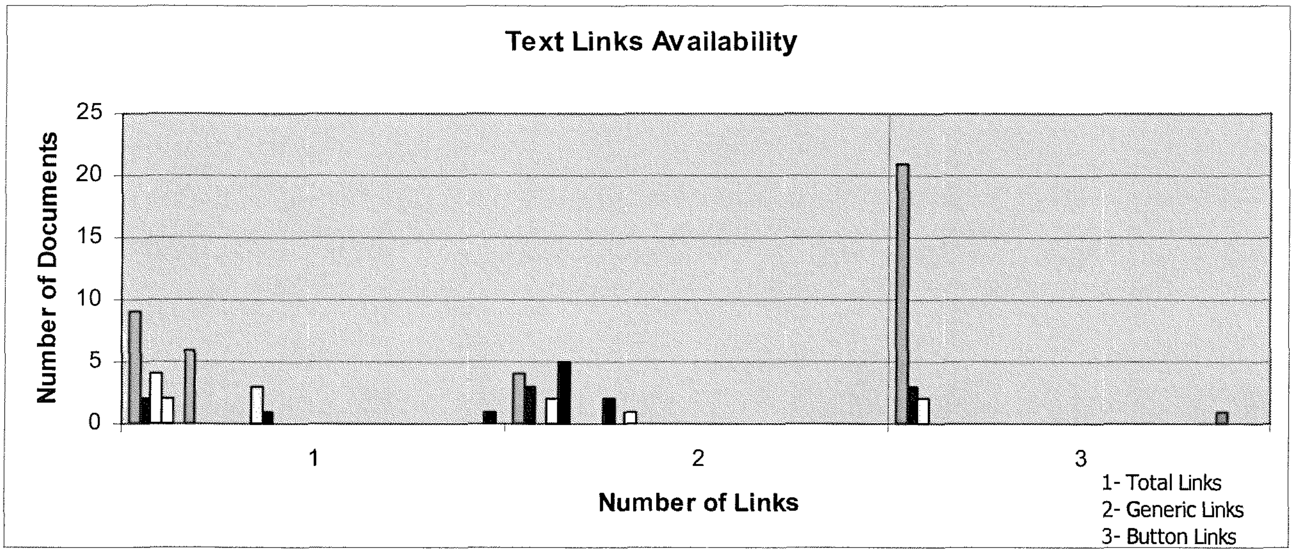
| | |
|---|---|
| 165 | files |
| 417 | links |
| 172 | generic links |
| 2 | local links |
| 179 | button links |

Valid links to build context descriptor:

| | |
|---|---|
| 137 | non-text files |
| 277 | links on non-text files |
| 91 | generic links on non-text files |
| 0 | local links on non-text files |
| 131 | button links on non-text files |

| | |
|---|---|
| 28 | text files |
| 151 | links on text files |
| 72 | generic links on text files |
| 4 | local links on text files |
| 63 | button links on text files |





CellBiology

| | |
|---|---|
| 237 | files |
| 2753 | links |
| 2429 | generic links |
| 0 | local links |
| 265 | button links |

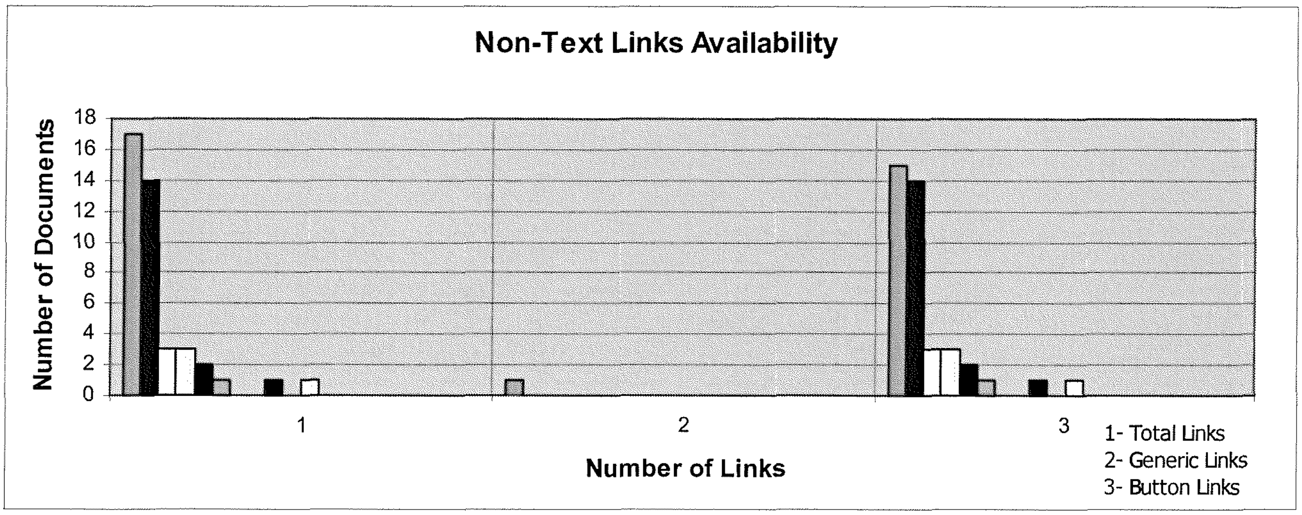
Valid links to build context descriptor:

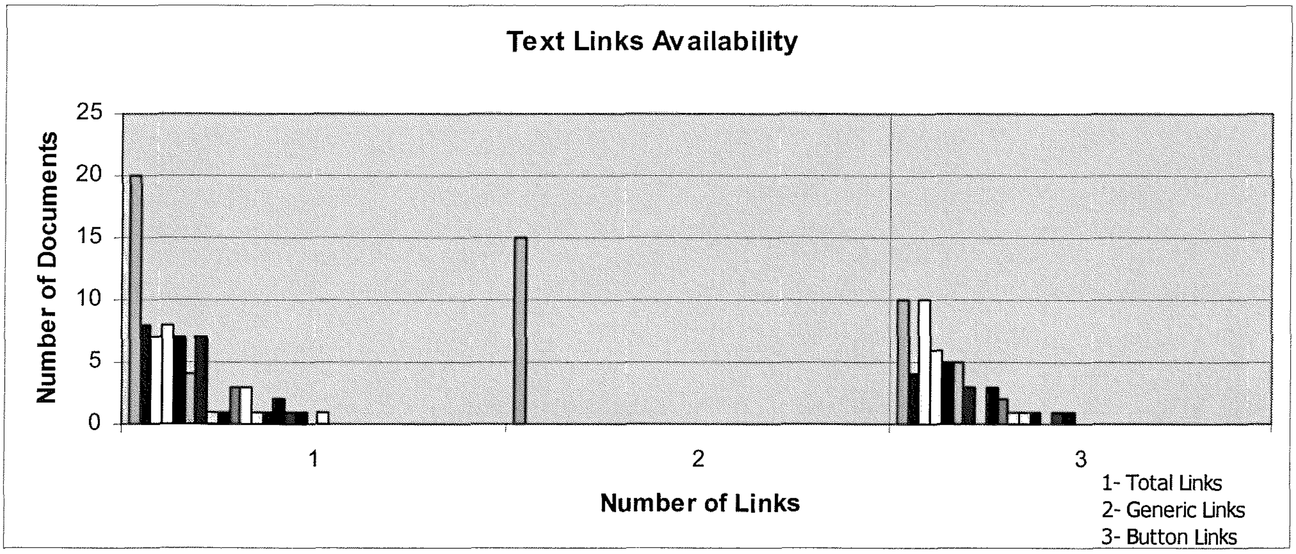
| | |
|---|---|
| 134 | non-text files |
| 139 | links on non-text files |
| 12 | generic links on non-text files |
| 0 | local links on non-text files |
| 125 | button links on non-text files |

| | |
|---|---|
| 103 | text files |
| 2796 | links on text files |
| 2429 | generic links on text files |
| 0 | local links on text files |
| 265 | button links on text files |

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | **Average** |
|---|---|---|---|---|
| 11.7 | 33.0 | 19.5 | 40.8 | Links |
| 1.0 | 2.1 | 3.3 | 2.5 | Links on non-text |
| 25.7 | 46.6 | 26.2 | 46.9 | Links on text |
| 9.6 | 33.4 | 33.7 | 55.6 | Generic Links |
| 0.1 | 0.3 | 1.2 | 0.4 | Generic Links on non-text |
| 22.1 | 48.0 | 39.3 | 58.4 | Generic Links on text |
| 1.6 | 2.9 | 4.1 | 3.3 | Button Links |
| 0.9 | 1.9 | 3.1 | 2.4 | Button Links on non-text |
| 2.6 | 3.6 | 4.9 | 3.7 | Button Links on text |

There are 3 text documents with more that 200 Generic Links





French

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 5,4 | 7,6 | 5,9 | 7,8 | Links |
| 2,5 | 4,1 | 3,1 | 4,3 | Links on non-text |
| 7,0 | 8,6 | 7,1 | 8,6 | Links on text |
| 0,1 | 0,7 | 1,6 | 1,9 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 0,2 | 0,8 | 1,6 | 1,9 | Generic Links on text |
| 5,2 | 7,4 | 5,8 | 7,6 | Button Links |
| 2,5 | 4,1 | 3,1 | 4,3 | Button Links on non-text |
| 6,7 | 8,3 | 7,1 | 8,4 | Button Links on text |

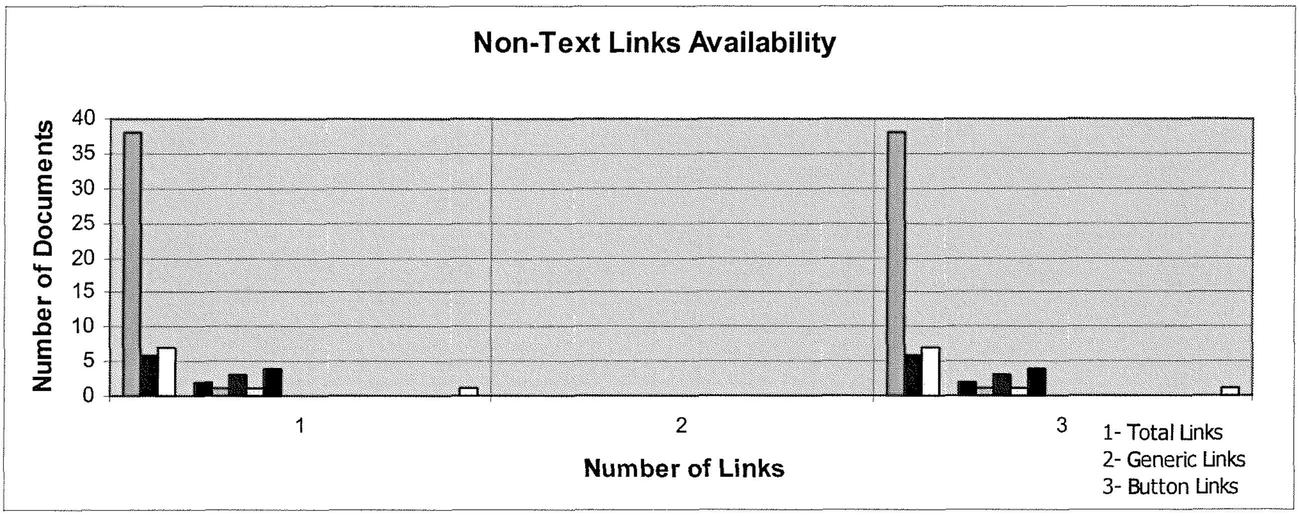
| | |
|---|---|
| 226 | files |
| 731 | links |
| 25 | generic links |
| 0 | local links |
| 697 | button links |

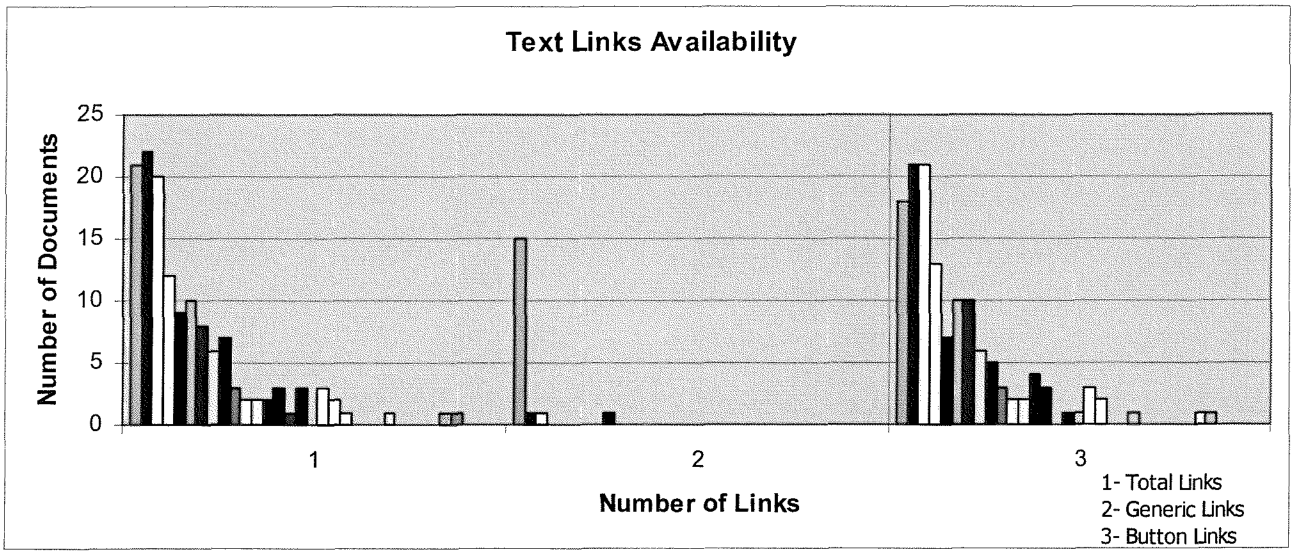
Valid links to build context descriptor:

| | |
|---|---|
| 80 | non-text files |
| 198 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 198 | button links on non-text files |

| | |
|---|---|
| 146 | text files |
| 1029 | links on text files |
| 29 | generic links on text files |
| 0 | local links on text files |
| 984 | button links on text files |

There is 1 text document with more that 50 Specific Links





FrenchRev

| | |
|---|---|
| 144 | files |
| 372 | links |
| 79 | generic links |
| 0 | local links |
| 292 | button links |

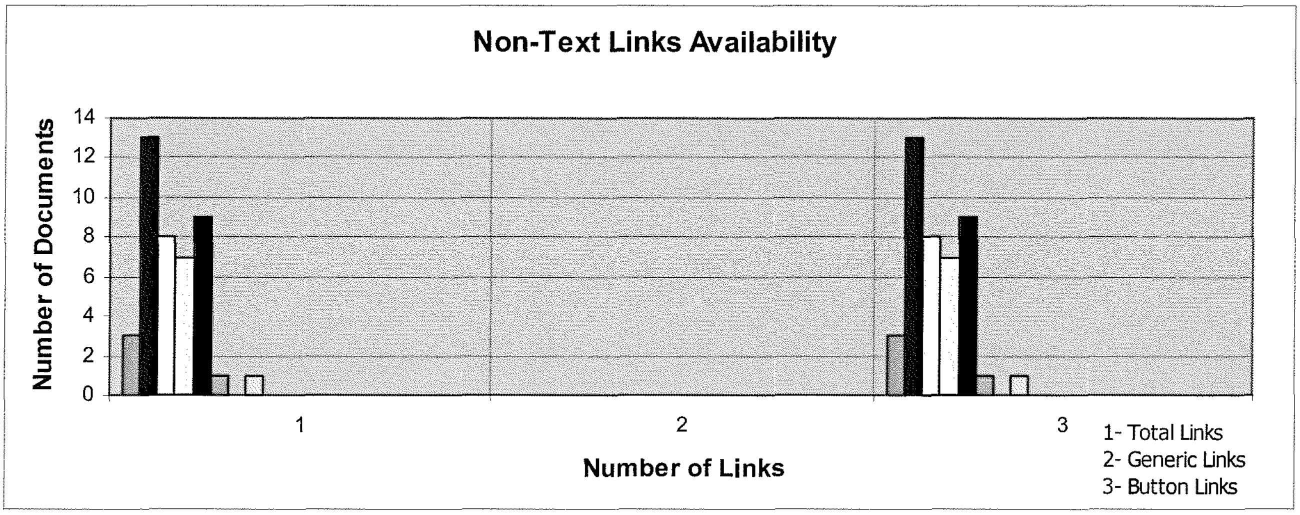
| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 2,9 | 5,4 | 3,3 | 5,6 | Links |
| 3,3 | 1,6 | 3,3 | 1,5 | Links on non-text |
| 2,8 | 6,4 | 3,2 | 6,8 | Links on text |
| 1,0 | 11,9 | 74,5 | 69,5 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 1,5 | 14,3 | 74,5 | 69,5 | Generic Links on text |
| 3,0 | 5,3 | 3,4 | 5,5 | Button Links |
| 3,3 | 1,6 | 3,3 | 1,5 | Button Links on non-text |
| 3,0 | 6,3 | 3,4 | 6,6 | Button Links on text |

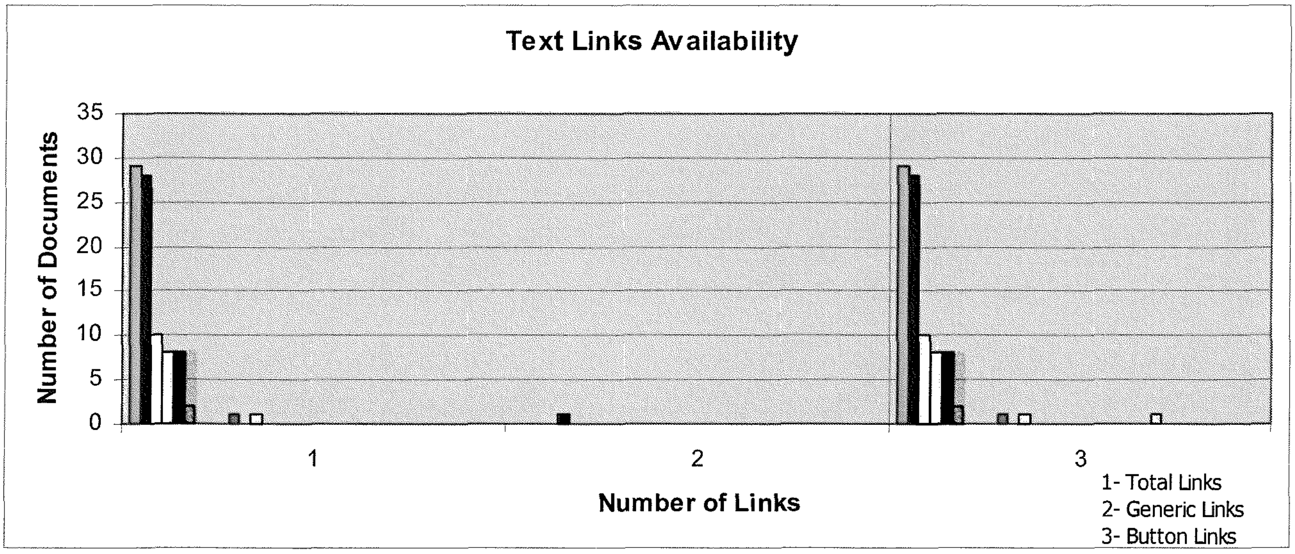
Valid links to build context descriptor:

| | |
|---|---|
| 43 | non-text files |
| 140 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 140 | button links on non-text files |

| | |
|---|---|
| 101 | text files |
| 453 | links on text files |
| 149 | generic links on text files |
| 0 | local links on text files |
| 304 | button links on text files |

There are 1 text document with more that 100 Generic Links





Pathology

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 2,4 | 22,4 | 3,1 | 25,4 | Links |
| 0,8 | 0,4 | 1,0 | 0,1 | Links on non-text |
| 18,1 | 73,2 | 65,2 | 127,4 | Links on text |
| 1,6 | 22,4 | 320,0 | 0,0 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 17,8 | 73,3 | 320,0 | 0,0 | Generic Links on text |
| 0,8 | 0,4 | 1,0 | 0,1 | Button Links |
| 0,8 | 0,4 | 1,0 | 0,1 | Button Links on non-text |
| 0,3 | 0,7 | 1,5 | 0,5 | Button Links on text |

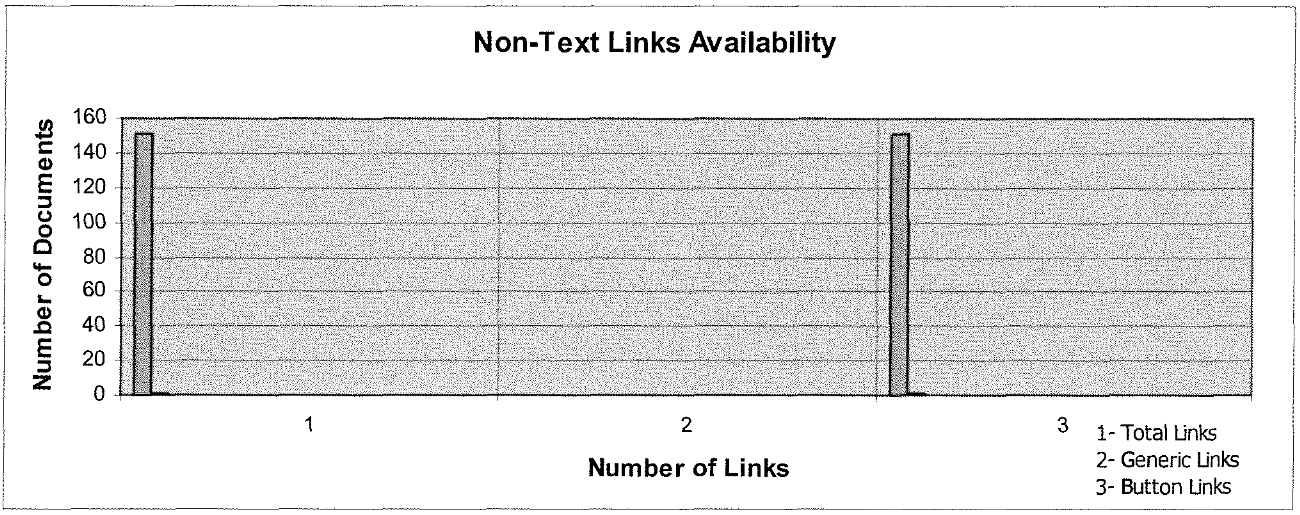
| | |
|---|---|
| 202 | files |
| 349 | links |
| 160 | generic links |
| 0 | local links |
| 188 | button links |

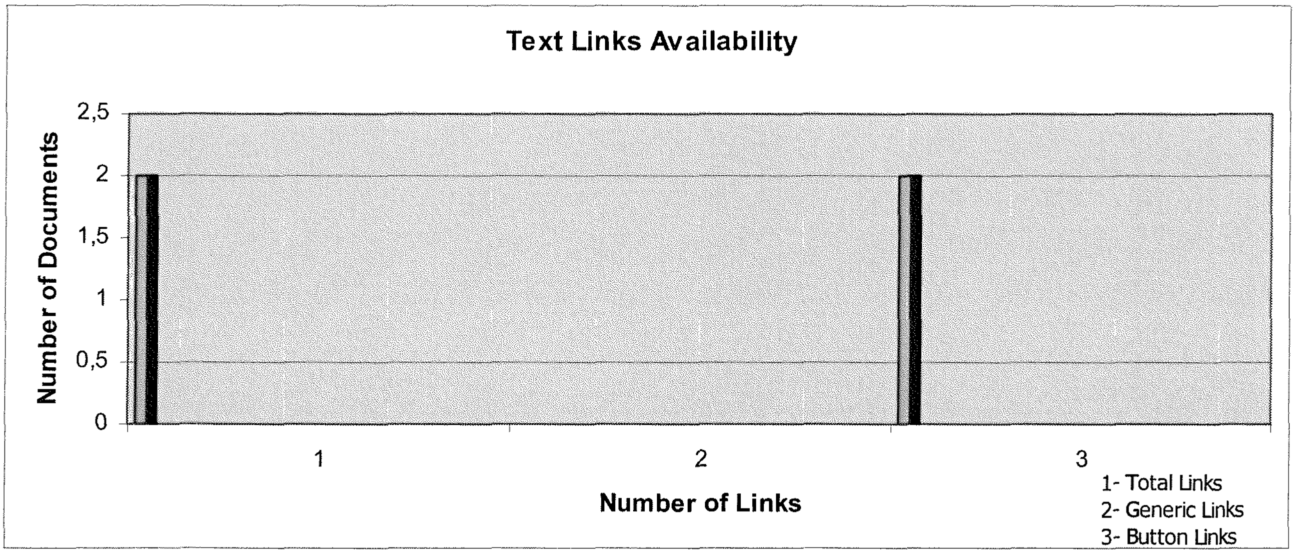
Valid links to build context descriptor:

| | |
|---|---|
| 184 | non-text files |
| 153 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 153 | button links on non-text files |

| | |
|---|---|
| 18 | text files |
| 326 | links on text files |
| 320 | generic links on text files |
| 0 | local links on text files |
| 6 | button links on text files |

There are 1 text document with more that 200 Generic Links





idsi

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 6,7 | 14,1 | 10,8 | 16,6 | Links |
| 4,6 | 13,5 | 11,7 | 19,6 | Links on non-text |
| 9,7 | 15,1 | 10,4 | 15,4 | Links on text |
| 4,8 | 11,7 | 9,7 | 15,1 | Generic Links |
| 2,0 | 5,4 | 7,5 | 8,4 | Generic Links on non-text |
| 8,1 | 15,1 | 10,2 | 16,3 | Generic Links on text |
| 1,8 | 5,9 | 3,1 | 7,5 | Button Links |
| 2,6 | 8,3 | 6,7 | 12,2 | Button Links on non-text |
| 1,6 | 3,9 | 1,9 | 4,1 | Button Links on text |

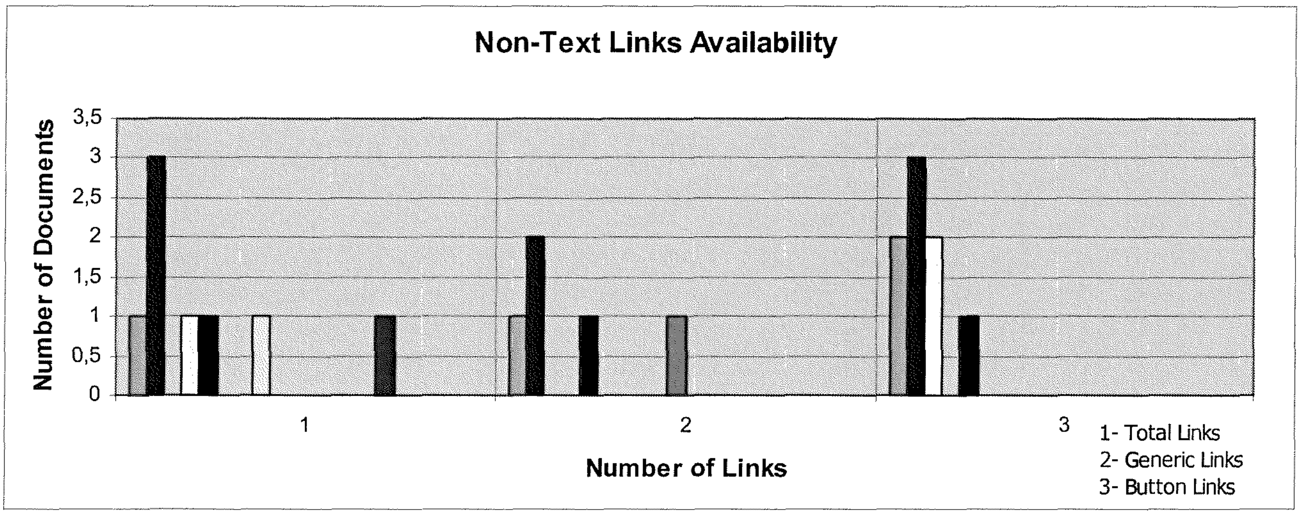
| | |
|---|---|
| 52 | files |
| 314 | links |
| 204 | generic links |
| 0 | local links |
| 108 | button links |

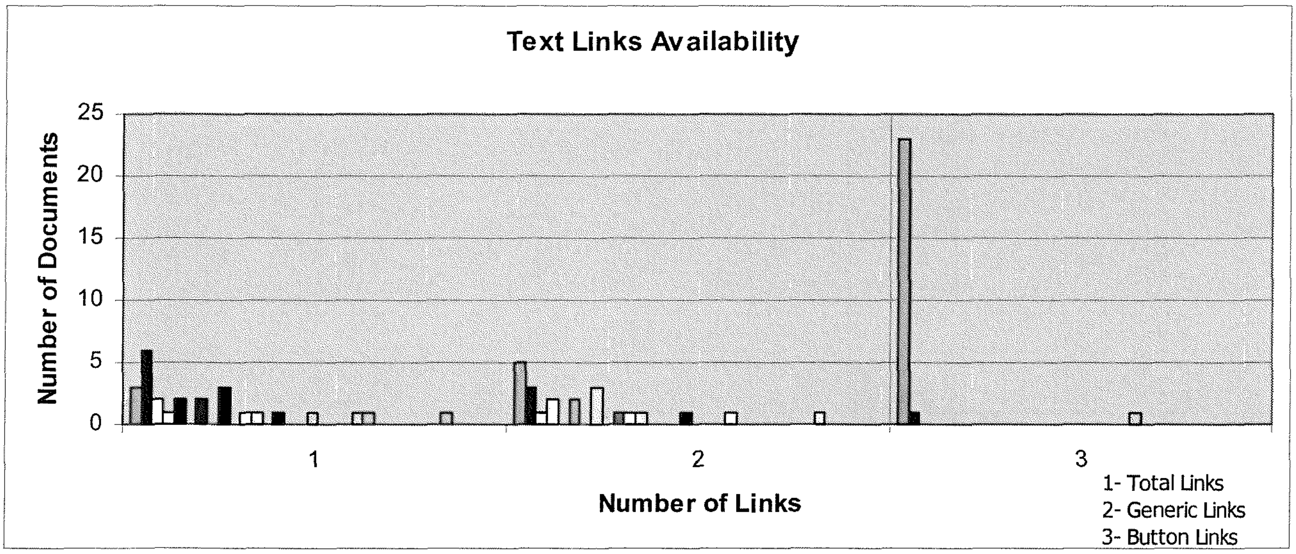
Valid links to build context descriptor:

| | |
|---|---|
| 23 | non-text files |
| 105 | links on non-text files |
| 45 | generic links on non-text files |
| 0 | local links on non-text files |
| 60 | button links on non-text files |

| | |
|---|---|
| 29 | text files |
| 282 | links on text files |
| 235 | generic links on text files |
| 0 | local links on text files |
| 47 | button links on text files |

There is 1 text document with more that 80 Generic Links





Posfix

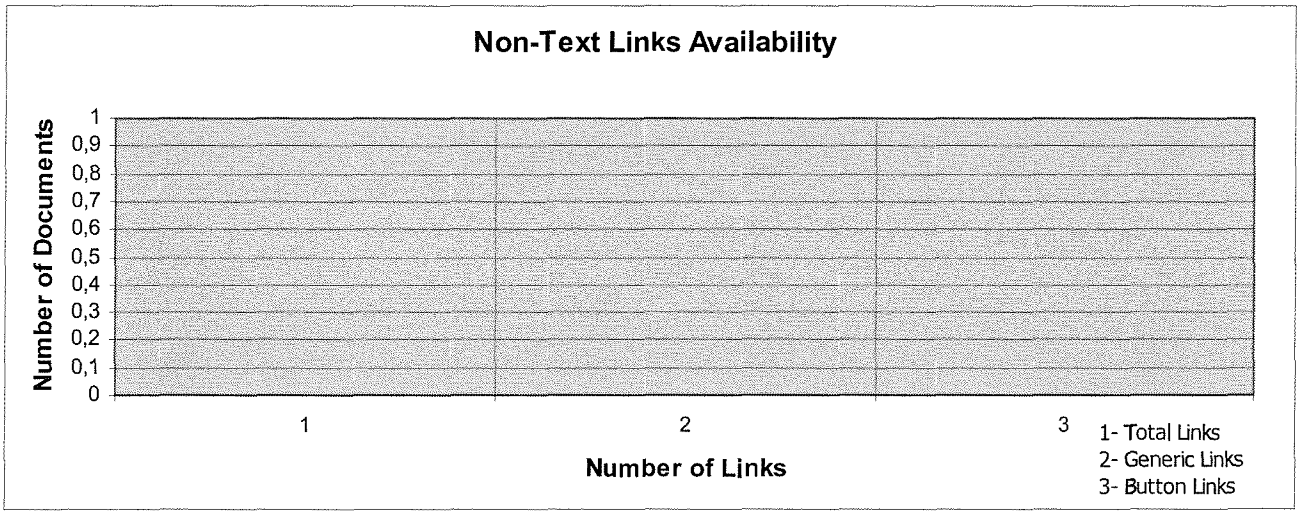
| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | Links |
| 0 | 0 | 0 | 0 | Links on non-text |
| 0 | 0 | 0 | 0 | Links on text |
| 0 | 0 | 0 | 0 | Generic Links |
| 0 | 0 | 0 | 0 | Generic Links on non-text |
| 0 | 0 | 0 | 0 | Generic Links on text |
| 0 | 0 | 0 | 0 | Button Links |
| 0 | 0 | 0 | 0 | Button Links on non-text |
| 0 | 0 | 0 | 0 | Button Links on text |

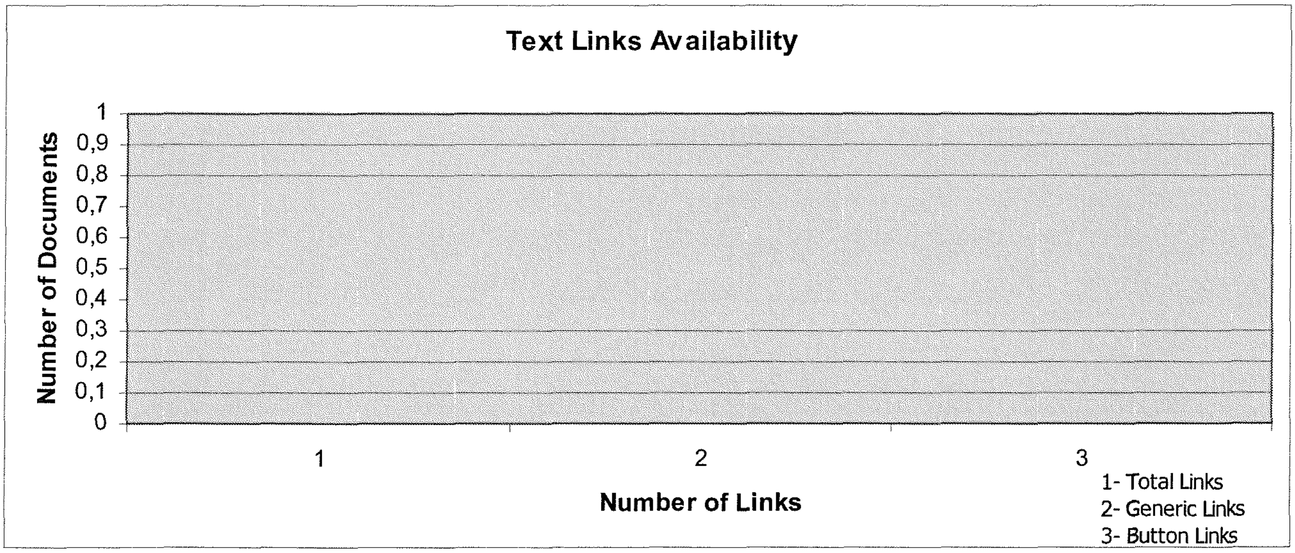
| | |
|---|---|
| 109 | files |
| 150 | links |
| 21 | generic links |
| 2 | local links |
| 121 | button links |

Valid links to build context descriptor:

| | |
|---|---|
| 109 | non-text files |
| 0 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 0 | button links on non-text files |

| | |
|---|---|
| 0 | text files |
| 0 | links on text files |
| 0 | generic links on text files |
| 0 | local links on text files |
| 0 | button links on text files |





RandJ

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 5,0 | 24,2 | 11,3 | 35,4 | Links |
| 0,1 | 0,4 | 2,0 | 0,0 | Links on non-text |
| 8,5 | 31,2 | 11,6 | 35,9 | Links on text |
| 2,8 | 23,6 | 199,0 | 0,0 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 4,9 | 30,7 | 199,0 | 0,0 | Generic Links on text |
| 2,2 | 6,4 | 5,1 | 8,9 | Button Links |
| 0,1 | 0,4 | 2,0 | 0,0 | Button Links on non-text |
| 3,7 | 8,0 | 5,2 | 9,1 | Button Links on text |

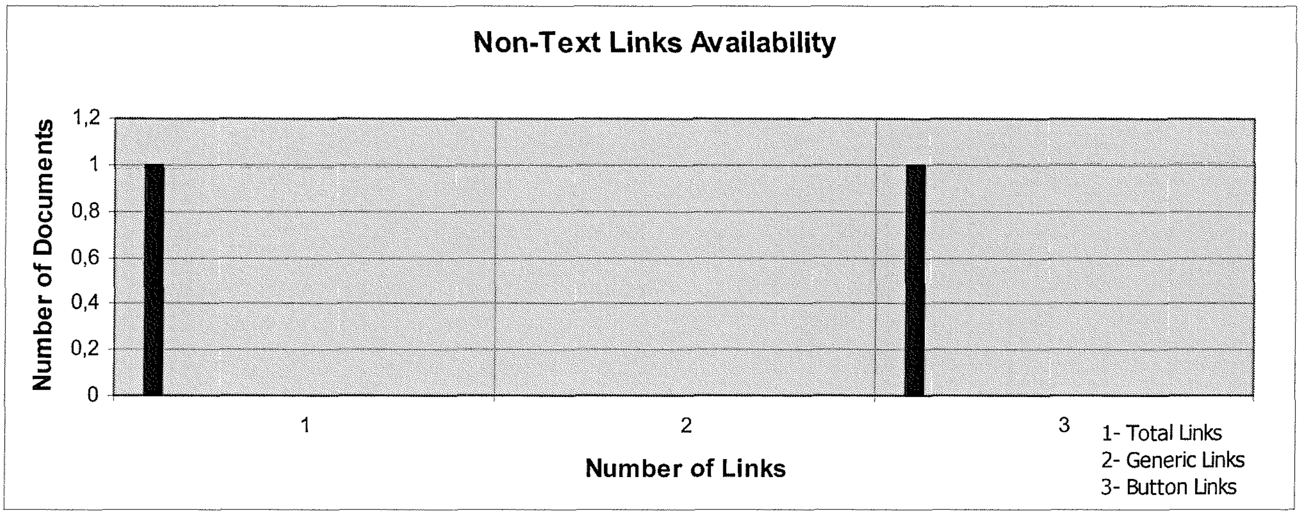
| | |
|---|---|
| 68 | files |
| 194 | links |
| 107 | generic links |
| 0 | local links |
| 87 | button links |

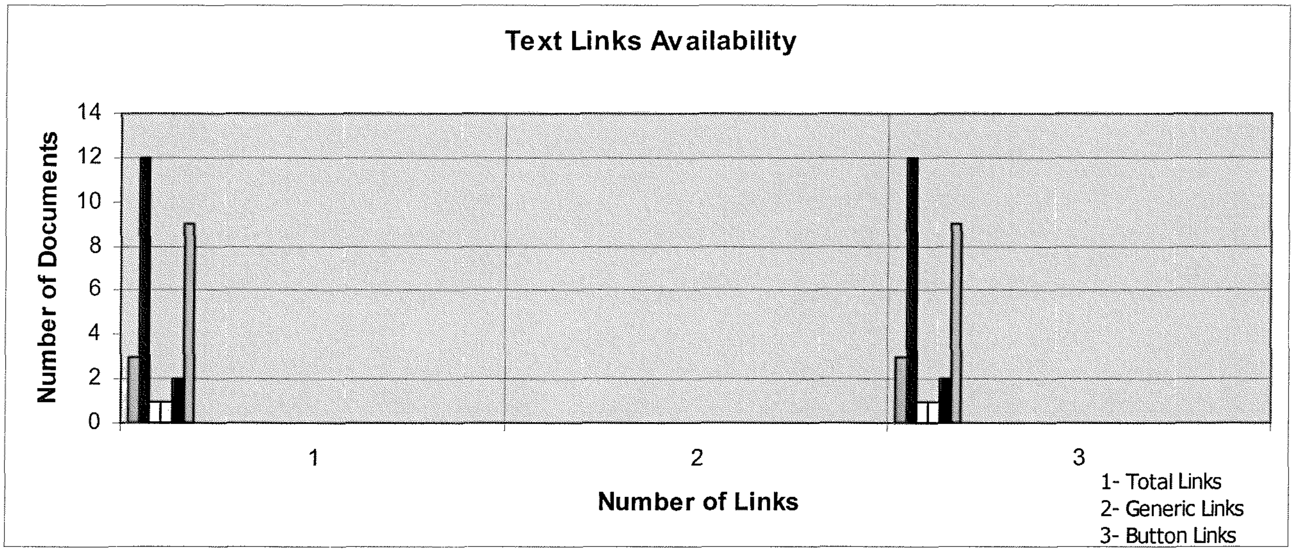
Valid links to build context descriptor:

| | |
|---|---|
| 27 | non-text files |
| 2 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 2 | button links on non-text files |

| | |
|---|---|
| 41 | text files |
| 364 | links on text files |
| 214 | generic links on text files |
| 0 | local links on text files |
| 150 | button links on text files |

There are 1 text document with more that 200 Generic Links





Shell

| | |
|---|---|
| 398 | files |
| 2234 | links |
| 19 | generic links |
| 0 | local links |
| 2212 | button links |

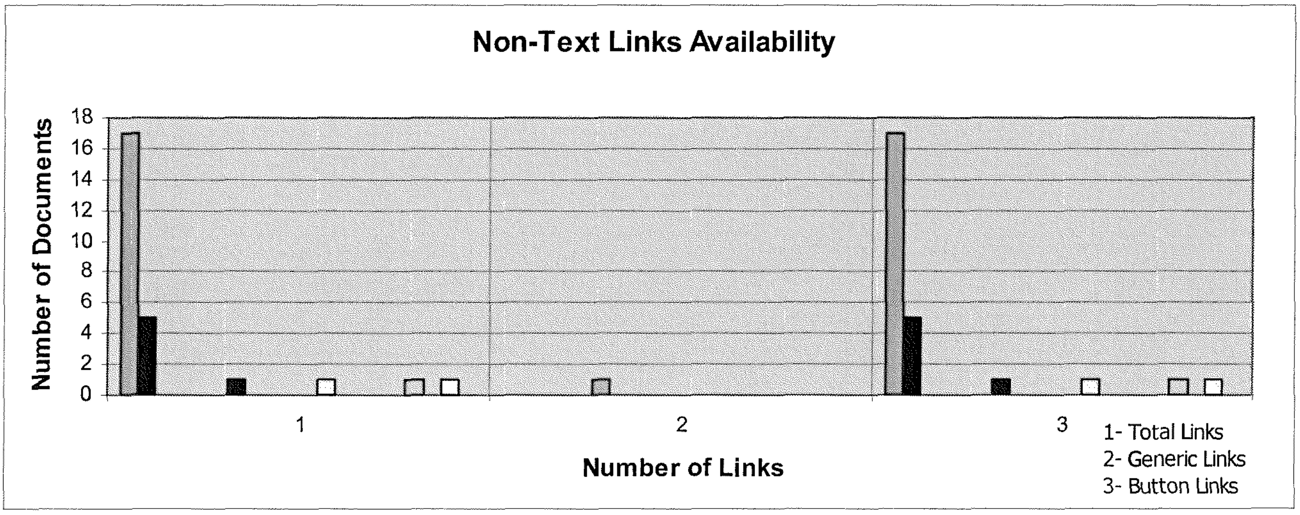
| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 5,6 | 18,4 | 25,4 | 32,1 | Links |
| 6,6 | 20,2 | 34,8 | 34,2 | Links on non-text |
| 1,5 | 5,0 | 4,1 | 7,7 | Links on text |
| 0,1 | 0,7 | 6,7 | 4,1 | Generic Links |
| 0,0 | 0,3 | 6,0 | 0,0 | Generic Links on non-text |
| 0,2 | 1,4 | 7,0 | 5,0 | Generic Links on text |
| 5,5 | 18,3 | 25,4 | 32,2 | Button Links |
| 6,6 | 20,1 | 34,7 | 34,1 | Button Links on non-text |
| 1,3 | 4,8 | 3,7 | 7,6 | Button Links on text |

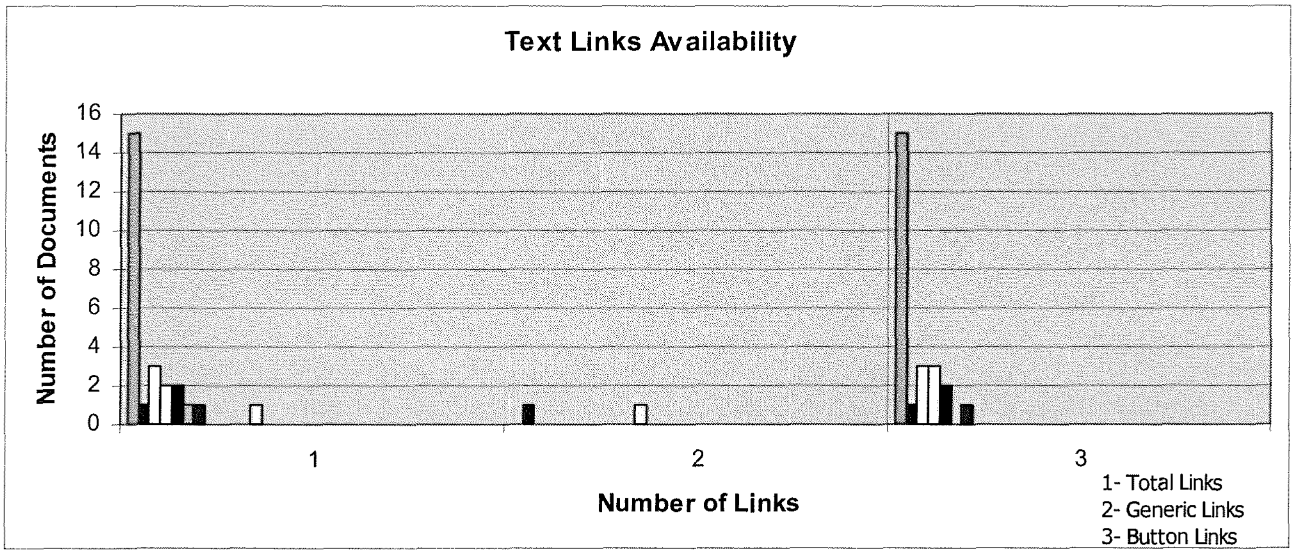
Valid links to build context descriptor:

| | |
|---|---|
| 323 | non-text files |
| 2122 | links on non-text files |
| 6 | generic links on non-text files |
| 0 | local links on non-text files |
| 2116 | button links on non-text files |

| | |
|---|---|
| 75 | text files |
| 110 | links on text files |
| 14 | generic links on text files |
| 0 | local links on text files |
| 96 | button links on text files |

There is 1 non-text document with more that 100 Specific Link





Tulip

| | |
|---|---|
| 449 | files |
| 1205 | links |
| 527 | generic links |
| 0 | local links |
| 677 | button links |

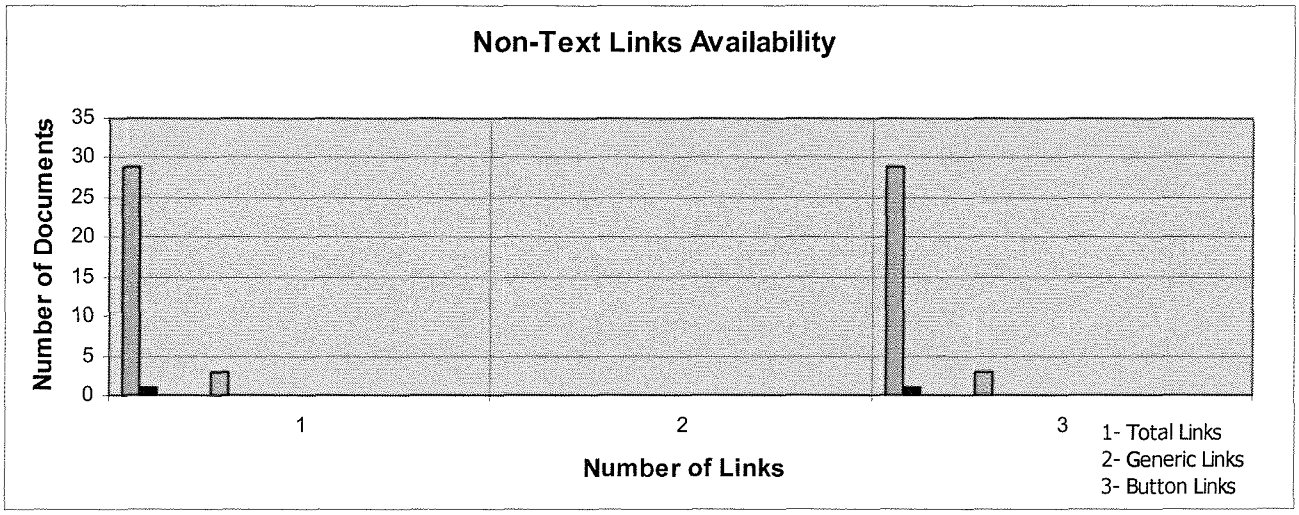
Valid links to build context descriptor:

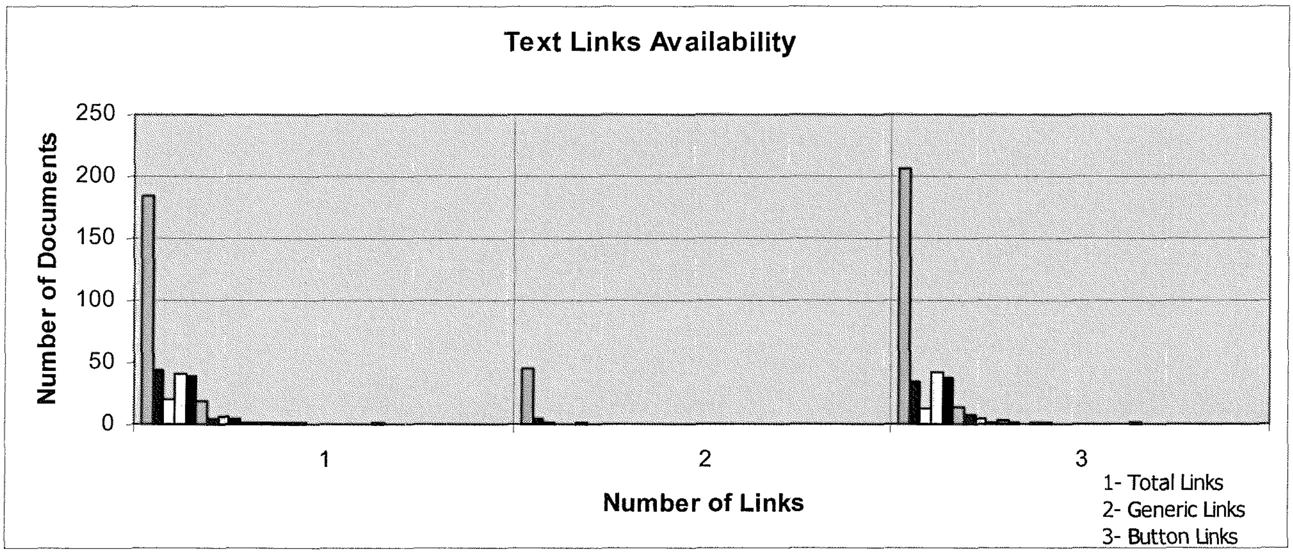
| | |
|---|---|
| 75 | non-text files |
| 49 | links on non-text files |
| 0 | generic links on non-text files |
| 0 | local links on non-text files |
| 49 | button links on non-text files |

| | |
|---|---|
| 374 | text files |
| 1637 | links on text files |
| 524 | generic links on text files |
| 0 | local links on text files |
| 1111 | button links on text files |

| Per Document | Standard Deviation per doc | Per document with links | Standard Deviation per doc with link | Average |
|---|---|---|---|---|
| 3,2 | 10,5 | 3,5 | 11,0 | Links |
| 0,7 | 1,2 | 1,5 | 1,4 | Links on non-text |
| 3,7 | 11,5 | 3,7 | 11,5 | Links on text |
| 0,6 | 9,4 | 4,9 | 26,7 | Generic Links |
| 0,0 | 0,0 | | | Generic Links on non-text |
| 0,7 | 10,3 | 4,9 | 26,7 | Generic Links on text |
| 2,6 | 5,0 | 2,9 | 5,2 | Button Links |
| 0,7 | 1,2 | 1,5 | 1,4 | Button Links on non-text |
| 3,0 | 5,4 | 3,0 | 5,4 | Button Links on text |

There are 1 text document with more that 200 Generic Links





# Appendix B - Application testing selection queries

## Archaeology

1. magnetic dating techniques
2. thermo-remanent magnetism
3. ancient buildings and on excavations
4. hydration layer which will increase in thickness over time
5. carbon-containing compound
6. thermoluminescence is used to date pottery and other clay objects
7. measures the energy trapped in the imperfections of mineral crystals which contain radioactive impurities
8. archaeomagnetic dating
9. obsidian hydration dating
10. magnetic dating techniques are particularly useful for the dating of stationary archaeological features
11. blocking temperature
12. temperature when the magnetic oxides lose their current magnetism
13. portable fired objects
14. dating in situ remain
15. archaeomagnetic properties of fired artefacts
16. requirements for magnetic dating
17. the benefits of magnetic dating
18. dating techniques in archaeology
19. earth's magnetic field
20. thermoluminescence can also be used to date baked clay
21. benefits of magnetic dating over radiocarbon dating
22. requirements for magnetic dating
23. blocking temperature
24. sediments from Petralona Cave, Greece
25. hoard of coins from Brittany
26. establishing a reference curve for dating
27. absolute and relative dating
28. radio carbon dating
29. the age of a tree can be calculated by counting its rings
30. dating by the presence of artefacts;
31. information about environmental conditions in the past
32. tree-ring calibration
33. not uniform in width, and vary with climatic fluctuations
34. eruption of Krakatoa near Java
35. absolute dating by tree-rings
36. combining historical information, tree-ring information, and trace element studies,
37. identifying forgeries among ceramic artifacts
38. measurements of remanent magnetization
39. etruscan pottery imitations
40. Barbetti et al., 1976
41. ancient pottery
42. etruscan poteery and recent imitations
43. scanning electron microscope

44. mineral and rock identification
45. the polarizing microscope
46. the microscope objective
47. pu l rpose of the polarizing microscope
48. geographical referencing
49. atomic absorption spectrometry
50. date of tool manufacture
51. hydration layer measurement
52. Mediterranean and the American cordillera areas
53. dating of obsidian geologically
54. strontium isotope analysis using a mass spectrometer
55. contamination of an archaeological stratum
56. obsidian hydration results
57. technique for measuring the hydration layer
58. Egyptian samples
59. sample collection
60. abrasion of the artefact
61. exposure of the obsidian to fire
62. samples of Lipari obsidian in the south of France
63. distribution of the material in Sicily and Malta
64. natural volcanic glass
65. forgeries
66. Cann and Renfrew
67. geochemical analyses
68. more sophisticated magnetic analyses
69. natural remanent magnetization
70. ceramics from Glozel are indeed forgeries
71. obsidian always contains some strontium,
72. radio carbon
73. stone age
74. initial concentration of radiocarbon
75. tree-ring dating
76. high-energy mass spectrometer systems
77. radiocarbon age determination
78. European oak calibration curve
79. mostly in the form of graphite
80. liquid scintillation counting techniques
81. radioactive isotopes
82. sediments from lakes and oceans
83. copper age
84. Monte Lessini
85. Greenland Eskinos
86. detection of ceramic forgeries
87. impurities in the clay
88. 'natural' thermoluminescence
89. date unburnt sediment
90. mineral separation

## Cell Biology

1. freshwater amoebae such as Amoeba proteus
2. larger amoebae, such as Physarum,
3. amoeboid movement
4. Dictyostelium discoidium
5. cortical microfilament meshwork
6. contraction of the gel-like ectoplasm
7. contraction of the cytoplasm by actomyosin
8. movement of endoplasm
9. the cytoplasm of a moving amoeboid cell
10. ectoplasm is found in the cortex of the cell
11. cell movement and cell behaviour
12. characteristic pseudopodia
13. duplex microtubules
14. flagella can be found in prokaryotes as well
15. the evolutionary origins of eukaryote cilia and flagella
16. cilia produce coordinated waves of movement which pass down the length of the organism
17. whereas flagella act in a kind of whip-like fashion
18. iliary locomotion is provided by the movement of adjacent cilia
19. each cilium performs an effector stroke and a recovery stroke
20. the cilia is brought about by the sliding of microtubules within the cilium
21. symplectic, in which the effector stroke and wave transmission occur in the same plane
22. metachronism is a physical phenomenon whose properties occur in space and time
23. antiplectic, in which the effector stroke and wave transmission occur in opposite directions
24. diaplectic metachrony may further be divided into dexioplectic
25. laeoplectic, in which it travels to the right
26. the gill of the mussel Mytilus is covered in millions of cilia

27. compare the movement of cilia in this sequence with that of Pleurobrachia and Opalina
28. Pleurobrachia is a ctenophore which moves by the action of specialized cilia
29. beat pattern here with those of the Mytilus gill and Opalina
30. mechanisms and control of microtubule
31. Cilia and Flagella
32. sliding interaction between adjacent microtubule doublets
33. Dynein, attached permanently to the A-tubule, binds to the B-tubule of the adjacent doublet
34. the dynein tilts and extends towards the disassembly end of the B-tubule
35. doublets to move relative to each other
36. Nexin links between the doublets prevent microtubules from moving
37. if the nexin links are removed using proteolytic enzymes the microtubules slide past each other
38. ATP induced sliding of tubules in Trypsin-treated Flagella of sea urchin sperm
39. he cross-section through a cilium shown here reveals nine microtubule doublets
40. each doublet consists of a complete A-tubule
41. also present are other proteins which bind adjacent microtubules together
42. intermittent contact with adjacent B-tubules
43. Opalina is a protozoon now classified with flagellates
44. Mytilus and Pleurobrachia
45. bacteria such as the spirochaete Spirillum are able to swim
46. bond which is broken by reduction and reformed by oxidation
47. the number of potential bond-forming residues on the S-ring is greater than that of the M-ring
48. Chromatium demonstrates the ability to change its direction of movement well
49. Wallin-Margulis theory of symbiotic evolution, stating that flagella probably arose “by an autogenous (i.e spontaneous) mechanism in an early non-flagellated eukaryote”
50. ancestors of the Hemiascomycetes
51. actin-based microfilaments and tubulin-based microtubules which arose in these ancestral cells
52. these protoflagella arose by spontaneous mutations
53. basal body cartwheel and other axonemal microtubules
54. similar to the gametes of the chytridiomycete fungi
55. helical bacteria which are believed to have developed a symbiotic relationship with their hosts which evntually became obligate to the point where the bacteria became part of the host's body
56. small spirochaetes in the termite hindgut are regularly attached to protists via specialised attachment sites
57. evolution of flagella from symbiotic spirochaetes
58. in some cases the surface of either the spirochaete or the protist became specialised for attachment
59. digestion-resistant spirochaetes were endocytosed by the hosts, and genetic information necessary for their protein synthesis was somehow transferred to the host nucleus
60. two main types of flagellar movement which can be observed: planar and helical
61. the protozoon Euglena shows a variation on this
62. the relative sliding of microtubules within the body of the flagellum
63. possible evolutionary significance of Spirochaetes
64. the cross-section through a cilium shown here reveals nine microtubule doublets
65. the dynein arms make intermittent contact with adjacent B-tubules
66. human sperm cells
67. the biology of developing systems
68. actin filaments in non-muscle cells are not arranged linearly as they are in muscle cells
69. three-dimensional meshwork with in the cytoplasm giving it a more gel-like consistency
70. thus small bipolar myosin aggregates can produce contraction of cytoplasmic gels
71. actin/myosin complex is acting in a similar way to the sarcomere of striated muscle
72. tissue cells orient themselves parallel to such physical stimuli as fibres, ridges and grooves
73. this phenomenon is a very efficient way of controlling the direction of cell

migration and is an almost universal property of cells

74. contact guidance has been explained as a form of capillarity in which parts of the cell margin are pulled outwards by the adhesions to fibres and grooves
75. bending of stress fibres is thought to interfere with their contractile functions
76. Abercrombie and Heaysman
77. contact inhibition is a significant phenomenon in biological terms
78. most often studied in neoplastic cells
79. loss of contact inhibition is responsible for the invasiveness of metastatic tumour cells
80. focal contacts are regions of ventral
81. shape of multiple small streaks and are found primarily near the front edge of the advancing cell
82. this adhesiveness is not as strong as that of the focal contacts themselves
83. cell migration and it's directional guidance
84. a crawling cell exerts a force on the substratum through a series of adhesions
85. the cell the actomyosin bundles-similar in many ways to myofibrils
86. posterior contacts detach and the posterior cytoplasm too is dragged forward
87. the lamellipodium is constantly advancing and withdrawing during the forward progression of the cell
88. a typical tissue cell, such as a chick embryo heart fibroblast
89. leading lamellipodium which exhibits a property known as ruffling
90. focal contacts which remain stationary when the cell moves
91. the crawling cell extends the lamellipodium
92. normal migration fibroblasts
93. contact inhibition wherein two cells stop moving if they come into contact and change direction before moving off
94. wide lamellipodia and many focal contacts
95. Py3T3 cells have fewer focal contacts
96. epithelial cell from Xenopus laevis, in vitro
97. neoplastic cells readily penetrate spaces between normal tissue cells and disrupt their organization
98. actively moving cells at the surface of a tumour penetrate and attach to normal cells
99. malignant cells can enter the bloodstream
100. one of the reasons for the invasive properties of neoplastic cells is their loss of contact inhibition
101. he studied the migratory patterns of normal and neoplastic mouse fibroblasts
102. formation of multilayers by the Py3T3 cells was by underlapping and not overlapping of cells
103. contact was made 'ruffle to side' rather than 'ruffle to ruffle' due to increased adhesions and narrowness of lamellipodia typical of neoplastic cells
104. neoplastic cells are cells which have undergone the process of transformation
105. normal culture conditions are not necessary and growth and division are no longer anchorage dependent
106. contact inhibition appears to be lost and growth occurs beyond confluency
107. the cytoskeleton becomes disorganised and the number of focal contacts is reduced
108. cyclosis is a phenomenon exhibited almost exclusively by large plant cells
109. large vacuoles and turgor
110. cytoplasmic streaming in order to mix and stir their contents and move materials around
111. continuous in those cells which have only a thin layer of cortical cytoplasm
112. giant algal cells such as Nitella
113. within the moving layer is a layer of thin fibrils containing actin filaments
114. this enables myosin molecules to move along them
115. this would provide a mechanism for the movement of those organelles
116. the stages of which can still be seen in bacteria and lower eukaryotes
117. a progression can be followed from the nuclear divisions of the bacteria
118. the evolution of the Mitotic Spindle
119. the chromatids are attached to the spindle
120. mitosis as seen in Xenopus
121. Dissodinium is a more advanced dinoflagellate
122. microtubules are organised in tunnels through the nucleus
123. epolymerization of microtubules is involved in the movement of chromatids towards the poles

124. polar fibres to move relative to each other
125. nuclear division, or mitosis, and cytoplasmic division, or cytokinesis
126. d kinetochore fibres extending from one pole to the kinetochore (or centromere) of a chromosome
127. migrate to opposite ends of the cell, from where microtubule assembly will begin, resulting in the completion of the spindle by the beginning of metaphase
128. chromosomes have become aligned at the equator of the spindle and during anaphase they divide to form chromatids
129. the disassembly end of a kinetochore fibre is at the pole
130. dynein-like ATPase and other microtubule as sociated proteins
131. mitoses observed in lower organisms form a gradual progression from simple bacteria right up to the higher vertebrates
132. interaction between actin and myosin which brings about contraction
133. the cycle can then be divided into four main stages
134. the myosin head undergoes a conformational change
135. the myosin head to assume its original configuration
136. the sequence you are watching is of cardiac muscle, the principles of contraction are the same for striated muscle as well
137. the basic contractile unit of skeletal muscle is the myofibril - a syncitial mass formed not by repeated nuclear divisions within a single cytoplasmic mass
138. myofibrils consist of many repeating units, known as sarcomeres, which contain actin and myosin in thin and thick filaments respectively
139. the sarcomere contracts it is the I-band which shortens
140. Myosin is a hexamer
141. interactions with actin in the production of contractile forces
142. obvious in the sarcomeres of muscle cells, but are also present in non-muscle cells
143. digestion with trypsin results in the production of light and heavy meromyosin
144. heavy chains of each myosin molecule have globular head regions
145. cleavage of the head region with the enzyme papain splits it into two subfragments S1 and S2
146. light chains of the myosin molecule vary greatly between tissues
147. phosphorylation by the enzyme myosin light chain kinase
148. they are known to have a role in regulating the actin-activation of the myosin
149. the molecular basis of muscular contraction
150. digesting with trypsin results in the production of light and heavy meromyosin
151. biomechanics of muscle contraction
152. mechanically integrating the contractile actions of actomyosin by linking the myofibrils at their Z- discs
153. together with actin and alpha-actinin, to be localized to the Z-line
154. alignment of sarcomeres during muscle contraction is maintained
155. desmin has a high affinity for actin
156. link actin to membrane sites such as intercalated discs in cardiac muscle and dense bodies in smooth muscle
157. intermediate filaments - looking for a function
158. association with neurofilaments in both astrocytes and neurons
159. less often in parallel arrays and lack the side branches of neurofilaments
160. although they have been found together with neurofilaments
161. intermediate filaments make up the third class of cytoskeletal elements
162. they differ from the other classes in not being of one type
163. keratin filaments-found in epithelial cells
164. found in all types of glial cells
165. intermediate filaments have a mainly structural role
166. they are involved in signalling and transport between the nucleus and the plasma membrane
167. keratins from different species are immunologically related
168. the keratins are formed from six or seven different polypeptides
169. several different triple chain-units may combine to form a single filament
170. avian feather keratins from different stages of development suggests that a family of keratin genes exists

171. intermediate filaments as integrators of cellular space
172. close association with microtubules and glial filaments in neuronal axons and dendrites
173. the arrangement of neurofilaments may be random or uniform
174. purified vimentin has been shown to bind to different fractions of erythrocyte membranes through two distinct domains
175. an amino-terminal domain binds specifically to plasma membran fractions
176. the central rod-like domain of the vimentin molecule has no affinity for either membrane fraction
177. the vimentin filaments form a network of nuclear-plasmalemmal connections
178. lamins, a group of proteins found exclusively on the inner surface of the nuclear envelope, bind to the carboxy-terminal of the vimentin molecule, suggesting that the interaction at the nuclear end is via the nuclear pores present in the envelope
179. aall intermediate filaments appear to be phosphorylated in vivo and, when isolated and purified, can be phosphorylated in vitro also
180. desmin and vimentin can be found in phosphorylated forms in muscle and non-muscle cells
181. the role of phosphorylation, however, is uncertain, but it is so ubiquitous that it may be that it serves to regulate the state of polymerization or aggregation of the various structures
182. Vimentin was first isolated from chick embryo fibroblasts
183. Vimentin-containing filaments form perinuclear aggregates in the form of a ring or a cap
184. serve in signalling and transport between the nucleus and the plasma membrane
185. Vimentin is also known to co-exist in cells with other intermediate filament types
186. some epithelial cells possess vimentin in addition to the characteristic keratin filaments
187. cross-linking proteins which form filament bundles or isotropic gels
188. filamin is a high molecular weight protein
189. found in the pseudopodia of amoeboid cells, in stress fibres and in membrane ruffles of chick embryo tissue cells
190. alpha-actinin is a rod-like protein associated with actin in muscle cells at the Z-disc
191. capping and severing proteins
192. cross-linking proteins
193. filament growth and to facilitate G-actin polymerisation
194. gelsolin is a non-muscle protein
195. Villin combines severing properties with the bundling abilities exhibited by alpha-actinin
196. Tropomyosin, on the other hand, acts to stabilize F-actin in non-muscle cells
197. microtubules self-assemble from tubulin dimers, microfilaments are produced by a similar process
198. an assembly end and a disassembly end, leading to treadmilling of a similar nature to that observed in microtubules
199. the assembly process is polarized
200. it is generally believed that F-actin formation occurs in two or three discrete steps
201. two 'nuclei' can join to form a longer segment, and once this has occurred the assembly/disassembly equilibrium is eventually reached
202. the finest of the fibres ,measuring 6nm in diameter and consist chiefly of the protein actin
203. globular actin (or G-actin) polymerizes in the presence of ATP to form long helices called filamentous actin (or F-actin)
204. actin microfilaments can be visualised using immunofluorescent labelling techniques
205. microfilaments are vital constituents in the actin/myosin systems which provide contractile forces for muscle and tissue cell, and amoeboid locomotion
206. the molecules of the cell matrix
207. actin binding proteins - regulators of cell architecture and motility
208. the distribution of Tau and HMW microtubule- associated proteins in different cell types
209. microtubules
210. microtubules contain proteins other than tubulin
211. assembly activity to the tau proteins
212. tubulin polymerization is aided not by tau proteins but instead by the HMW proteins

213. proteins stimulate microtubule assembly in vitro
214. they have been identified by immunocytochemical techniques
215. within the cytoplasm of the cell, microtubules are not arranged in a random fashio
216. the bast known MTOC is the centrosome, which usually lies close to the nucleus and consists of two centrioles plus pericentriolar material
217. MTOC's are the preferred site for microtubule assembly can be demonstrated experimentally
218. when colcemid was removed by simple washing, the centriole pairs within the centrosome separated and normal mitosis occurred
219. MTOC’s are made up of tubulin, Microtubule Associated Proteins (MAP’s) and RNA
220. anti-tubulin antibody staining has shown the presence of tubulin within the centrosome
221. microtubules differ from microfilaments in their composition
222. tubulin which is a globular protein with two sub-units
223. the axis of the protofilaments in the microtubule is parallel to that of the entire microtubule
224. microtubule consists of a number of axial segments joined end to end
225. is the ability to self-assemble (in a manner similar to that of microfilament assembly)
226. microtubule assembly in vivo is organized by various structures which provide a base from which the microtubule can grow
227. generally a microtubule has an assembly end and a disassembly end
228. however, radioactive labelling of dimers shows that individual molecules are being translocated from one end of the microtubule to the other in a continuous exchange process
229. microtubule assembly can be experimentally inhibited
230. the ability to move is a property of all cytoplasmic matter
231. ciliary movement.
232. flagellar movement.
233. amoeboid movement.
234. tissue cell movement.
235. cytoskeleton is intimately involved
236. main classes of proteins which make up the cytoskeleton
237. hese cytoskeletal elements are also involved in functions other than locomotion
238. microfilaments are involved in muscle contraction,
239. and intermediate filaments have been linked with signalling and transport to and from the nuclear membran

## French

1. la teneur en vapeur d’eau
2. maître d’oeuvre
3. des tassements
4. en maquette grandeur nature
5. Ariane retardé
6. Ariane retardé 3
7. exploration fusées et lanceurs
8. prises de vue sous-marines
9. la soucoupe plongeante du commandant Cousteau
10. à quoi ressemble le Nautile?
11. “creuser sous l’océan”
12. au sud-ouest des Açores, sa millième plongée
13. les travaux consistent à colmater les ouvertures
14. programmation annuelle des campagnes scientifiques
15. flotte de dix navires de recherche
16. trente-sept volcans
17. la canopée reste le dernier continent à découvrir
18. Nausicaa vient de faire sa première apparition publique
19. satellite d’observation du soleil, Soho
20. le territoire du MD-II
21. le ciel béarnais
22. l’Atalante a été la première à répertorier les sea-mounts,
23. satellite ISO
24. Elf Aquitaine cherche du pétrole
25. a DEMI-MOT: Dimensions, measures
26. la millième plongée du nautile

27. la fusée européenne Ariane 5
28. ACTIVITE 3d
29. un éboulis
30. exploration de la Terre
31. accéder au menu d'activités
32. exploration satellitaire
33. genre des adjectives
34. la maquette, grandeur nature, du lanceur Ariane 5
35. l'édification de l'Europe spatiale
36. ERS-2 suit une orbite polaire
37. Société Européenne de Propulsion
38. volcans sous-marins

## French Revolution

1. revolutionary tribunal
2. September massacres
3. bread and terror - the spectre of popular government
4. storming of the Bastille - the people emerge
5. "doubling the third". But when the Estates met, vote by head.
6. storming of the Bastille - the people emerge
7. June 14, 1791: Chapelier on organisations of workers. Translated by Tom Carter.
8. Calonne's plan of reform, approved by Louis XVI after several months of persuasion during the autumn of 1786, had three main elements. First came fiscal and administrative reforms designed to remedy once and for all the structural problems besetting the royal finances.
9. cahiers, cahiers de doléances. Statements of grievances presented to the Estates General by the deputies in 1789.
10. national popular elections are such a common occurrence today, that it is hard for us to imagine ourselves in the France of 1789. There, an institution in abeyance for one hundred and seventy five years was revived amid changed circumstances.
11. in England, as it became increasingly clear that the french revolution was not simply a matter of the french seeing the error of their ways and imitating and catching up with british institutions and liberties, a backlash began against those who continued to express
12. National Guard. by W. Doyle. The National Guard emerged from spontaneous
13. some possible essay questions. by W. Doyle. 1. Analyse the economic crisis of the years 1788-89, and estimate its political consequences.
14. third party rights holders for extracts used are also available.
15. François Furet, the Terror, and 1789. by David D. Bien In the last years François Furet has been writing almost faster than I can read.
16. perhaps the most famous of Gillray's anti-revolution caricatures, its enormous popularity led to the design being copied in several media, including pottery and medals.
17. National Guard
18. public safety
19. this cartoon indicates the bancruptcy of the regime.
20. commemoration of the return of the king to Paris.
21. though subsequently pilloried by english cartoonists the benefits of the 'Liberty Tree' were generally understood across european societies.
22. seen as capable of putting the finances to rights, while maintaining the king's prerogatives. But 'king of France and the nation' is not the same as 'absolute monarch'.
23. Oath in retrospect
24. protestants were granted civil rights
25. this powerful illustration draws on ancient metaphors to state the Rights of Man.
26. the Tennis Court Oath, 2
27. the women's march to Versailles marked a decisive moment in the transition from royal absolutism to a constitutional government.
28. French revolution,
29. third estate and the convention
30. council of five hundred
31. first republic
32. financial crisis

33. terror was not the only item on the agenda of that famous assembly.
34. the zenith of french glory
35. women in the French revolution
36. durable cultural creations
37. cereal crops
38. king was concentrating troops in the neighborhood of Paris
39. the crisis of the monarchy
40. the parlement of Paris made no trouble over registering the freedom of the grain trade
41. Sans-Culottes
42. commons or the tiers etat;
43. Estates-General
44. French liberty,
45. slavery
46. chronological curve of executions
47. constituent assembly
48. Cordeliers' Club
49. vestments of the revolution
50. the high point of the parlements' power
51. legislative assembly
52. national assembly
53. the Rights of Man - or - Tommy Paine, the little american Taylor, taking the Measure of the Crown, for a new pair of Revolution Breeches, 23 May 1791
54. the idea that clergy, nobility and third estate should work in harmony for common benefit was widely diffused in prints, china and glassware throughout the nation.
55. committee of public safety
56. republican calendar
57. a publication of the TLTP history Courseware Consortium
58. the Estates General - establishing representative government

## Tulip

1. acute organic brain syndrome
2. alcoholic hallucinosis
3. anxious personality disorder
4. borderline personality disorder
5. catatonic schizophrenia
6. depressive cognitions
7. hebephrenic schizophrenia
8. laurence-moon-biedl syndrome
9. munchausen's syndrome
10. post-traumatic stress disorder
11. schizoid personality disorder
12. trifluoperazine
13. wernicke's encephalopathy
14. alcohol dependence
15. recommended maximum alcohol consumption
16. symptoms of alcohol dependence
17. epileptic fits
18. epidemiology
19. assessment and diagnosis of alcohol dependence
20. self-help organisations
21. alcoholics anonymous
22. cognitive- behavioural models
23. drug treatments
24. benzodiazepines
25. antipsychotic
26. maintaining abstinence from alcohol
27. withdrawal symptoms
28. amphetamine
29. overdose
30. barbiturate
31. ecstasy
32. lsd and magic mushrooms
33. treatment of drug abuse and dependence
34. symptoms of drug induced psychosis
35. psychoactive drugs
36. solvent abuse
37. treatment of drug abuse
38. hypothalamic dysfunction
39. anorexia nervosa
40. eating disorders
41. bmi
42. bulimia nervosa
43. mental health act
44. patient does not consent to admission
45. enforce admission and treatment
46. programme of rewards for weight gain
47. guardianship order
48. depression postnatally
49. early parental loss
50. management of bipolar disorder
51. modern diagnostic classification systems
52. depressive episodes
53. hypomania is a milder form of mania
54. bipolar affective disorder
55. sleep disturbance
56. psychomotor retardation
57. mood disturbance in manic episodes
58. poor pre-morbid personality

59. depression bereavement
60. drug induced depression
61. management of depression
62. depression may present with physical symptoms
63. differential diagnosis
64. postnatal depression
65. antidepressant drugs
66. comparison of ssris and tricyclics
67. effectiveness of ect
68. acute episode of depressive disorder
69. social interventions
70. side-effects of lithium
71. toxic effects of lithium
72. relapse rates
73. counselling and social work treatment
74. psychosocial and pharmacological approaches
75. problem solving approaches
76. mental state examination
77. amitriptyline
78. family therapy
79. day hospital
80. using simple records of thoughts and feelings
81. rate of speech
82. abnormalities of thought
83. illusions are distorted perceptions
84. predisposing individuals to anxiety disorders
85. mechanisms of dissociation
86. agoraphobia symptoms
87. symptoms of anxiety
88. phobic anxiety disorder
89. obsessive-compulsive
90. stress and adjustment disorders
91. classical conditioning
92. planning treatment
93. antidepressants
94. overlap between depressive disorders and anxiety disorders.
95. panic attacks
96. dissociative amnesia
97. alzheimer's disease
98. relaxation is a useful skill to teach patients
99. post-traumatic stress
100. patient's social needs
101. phobia of a particular object
102. abnormal protein synthesis
103. electroconvulsive therapy
104. loss of concentration
105. relatives support group
106. depression in the elderly
107. treatment for dyssocial personality disorder
108. impulsive
109. psychotherapy
110. paranoid
111. illicit drugs
112. presenting complaint
113. medical history
114. psychiatric history
115. health services
116. employment agencies
117. emotional reactions
118. chronic social problems
119. head injury
120. parkinson's disease
121. involvement of the family
122. positive symptoms
123. auditory hallucinations
124. dopamine hypothesis
125. schizophrenia in monozygotic twins
126. hebephrenic
127. delusional perception
128. acute dystonia
129. problem solving therapy
130. supportive psychotherapy
131. short term treatment
132. symptomatic relief
133. depressive cognitions
134. cognitive therapy
135. child and adolescent psychiatry
136. psychological therapies
137. mental state examination
138. depressed mood
139. hypothyroidism
140. social workers
141. sigmund freud
142. reduced concentration

# References

Allan 1996

James Allan, "**Automatic Hypertext Link Typing**", The Seventh ACM Conference on Hypertext (Hypertext'96), pp 42-52, March 1996, ACM Press

Agosti et al. 1992

Maristella Agosti & Pier g. Marchetti, "**User Navigation in the IRS Conceptual Structure through a Semantic Association Function**", The Computer Journal, Vol. 35, No. 3, pp 194-199, 1992

Agosti et al. 1996a

Maristella Agosti, Fabio Crestani and Massimo Melucci, "**Design and Implementation of a Tool for the Automatic Construction of Hypertext for Information Retrieval**", Information Processing and Management, Vol. 34, No. 4, pp 459-476, 1996

Agosti et al. 1996b

Maristella Agosti, Alan F. Smeaton, "**Information Retrieval and Hypertext**", Kluwer Academic Publishers, 1996

Amato et al. 1998

Giuseppe Amato, Fausto Rabitti and Pasquale Savino: "**Supporting Image Search on the Web**", Computer Networks and ISDN Systems, Vol 30, no. 1, pp 604-616, 1998

Anderson 1997

Kenneth M. Anderson, "**Integrating Open Hypermedia Systems with the World Wide Web**", *Proceedings of Hypertext '97*, pp. 157-166, Southampton, UK. ACM Press. April 1997

Andrews et al. 1995

Keith Andrews, Frank Kappe and Hermann Maurer, "**The Hyper-G Network Information System**", Journal of Universal Computer Science, vol 1 (4), pp 206-220, 1995

Arents et al. 1993

Hans C. Arents & Walter F.L. Bogaerts, "**Concept-Based Retrieval of Hypermedia Information: from Term Indexing to Semantic Hyperindexing**" Information Processing and Management, Vol. 29, No. 3, pp 373-386, 1993

Baeza-Yates et al. 1999

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "**Modern Information Retrieval**", Addisson-Wesley, 1999

Bernstein 1990

Mark Bernstein, "**An Apprentice That Discovers Hypertext Links**", Hypertext: Concepts, Systems and Applications (Proceedings of ECHT'90), pp 109-122, 1990, A. Rizk, N. Streitz, and J. Andre Eds

Boy 1991

Gui Boy, "**Indexing Hypertext Documents in Context**", Proceeding of the ACM Conference on Hypertext, pp 51-61, 1991, ACM press.

Bray et al. 1998

Tim Bray, Jean Paoli and C. M. Sperberg-McQueen, "**Extensible Markup Language (XML) 1.0**", W3C Recommendation, 10 February 1998, http://www.w3.org/TR/1998/REC-xml-19980210

Brin et al. 1998

Sergey Brin and Lawrence Page. "**The Anatomy of a Large-Scale Hypertextual Web Search Engine**" Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7

Bruza 1990

Peter D. Bruza, "**Hyperindeces: A Novel Aid for Searching in Hypermedia**", Hypertext: Concepts, Systems and Applications (Proceedings of ECHT'90), pp 212-223, 1990, A. Rizk, N. Streitz, and J. Andre Eds

Bush 1945

Vannevar Bush, "**As We May Think**", Atlantic Monthly 176, pp 101-108, July 1945

Carr et al 1998

Les Carr, David DeRoure, Hugh Davis and Wendy Hall, "**Implementing an Open Link Service for the World Wide Web**", World Wide Web Journal, 1, 2. 1998.

Carr et al 2000

Les Carr, Wendy Hall and David DeRoure, "**The Evolution of Link Services**", ACM Computing Surveys, Symposium on Hypertext and Hypermedia

Ceri et al. 1999

Stefano Ceri, Sara Comai, Ernesto Damiani, Piero Fraternali, Stefano Paraboshi and Letizia Tanca, "**XML-GL: a graphical language for querying and restructuring XML documents**", *Eighth International World Wide Web Conference*, pp. 93-109, Toronto 1999

Chakrabarti et al. 1999

Soumen Chakrabarti, Byron E. Dom and David Gibson, "**Mining the Link Structure of the World Wide Web**", IEEE Computer, 32 (8), August 1999

Chamberlin et al. 2000

Don Chamberlin, Peter Fanhhauser, Massimo Marchiori and Jonathan Robie, "**XML Query Requirements**", W3C Working Draft, 15 August 2000, http://www.w3.org/TR/xmlquery-req

Chiaramella et al. 1996

Yves Chiaramella & Ammar Kheirbek, "**An Integrated Model for Hypermedia and Information Retrieval**", In Maristella Agosti & Alan F. Smeaton (eds.), Information Retrieval and Hypertext, pp 139-178, Kluwer Academic Publishers, 1996

Clark et al. 1999

James Clark and Steve DeRose, "**XML Path Language (XPath) Version 1.0**", W3C Recommendation, 16 November 1999, http://www.w3.org/TR/xpath

Conklin 1987

Jeff Conklin, "**Hypertext: An Introduction and Survey**", IEEE Transactions Computer, 1(9), pp 17-42, September 1987

Croft et al. 1979

W. Bruce Croft & D. J. Harper, "**Using Probabilistic Models of Document Retrieval Without Relevance Information**", Journal of Documentation, Vol. 35, No. 4, pp 285-295, 1979

Croft et al. 1989

W. Bruce Croft and Howard Turtle, "**A Retrieval Model for Incorporating Hypertext Links**", Proceedings of the ACM Conference on Hypertext, pp 213-224, November 1989, ACM Press

Crouch et al. 1989

Donald B. Crouch, Carolyn J. Crouch & Glenn Andreas "**The use of Cluster Hierarchies in Hypertext Information retrieval**", Proceedings of the ACM Conference on Hypertext, pp 225-237, November 1989, ACM Press

Cunliffe et al 1997

Daniel Cunliffe, Carl Taylor & Douglas Tudhope, "**Query-based Navigation in Semantically Indexed Hypermedia**", The Eighth ACM Conference on Hypertext, pp 87-95, April 1997, ACM Press

Daniel et al. 2000

Ron Daniel Jr., Steve DeRose and Eve Maler, "**XML Pointer Language (XPointer) Version 1.0**", W3C Candidate Recommendation, 7 June 2000, http://www.w3.org/TR/xptr

Davis et al 1992

H. Davis, Wendy Hall, Ian Heath, Gary Hill & R. Wilkins, "**Towards an Integrated Information Environment with Open Hypermedia Systems**", Proceeding of the ACM Conference on Hypertext, pp 181-190, 1992, ACM press.

Davis et al. 1994

H Davis, S Knight, W Hall, "**Light Hypermedia Link Services: A Study of Third Party Application Integration**", In Proceedings of the 1994 European Conference on Hypermedia Technology, Edinburgh, 18-23 September, 1994

Davis 1995

Hugh Davis, "**To Embed or Not To Embed…**", Communications of the ACM, August 1995

Davis et al. 1996

H. Davis, A. Lewis and A. Rizk, "**OHP: A draft proposal for a standard open hypermedia protocol**", 2nd Workshop on Open Hypermedia Systems, Washington, DC, University of California, Irvine, 1996, pp. 27-53

DDC 1998

OCLC Forest Press, "**Dewey Decimal System Home Page**", http://www.oclc.org/oclc/fp/index.htm, October 1998

Dean et al. 1999

Jeffrey Dean and Monika R. Henzinger, "**Finding related pages in the World Wide WebB**", *Eighth International World Wide Web Conference*, pp. 389-401, Toronto 1999

DeRose et al 2000

Steve DeRose, Eve Maler, David Orchard and Ben Trafford, "**XML Linking Language (XLink) Version 1.0**", W3C Candidate Recommendation, 3 July 2000, http://www.w3.org/TR/xlink/

Deutsch et al. 1999

Alin Deutdch, Mary Fernandez, Daniela Florescu, Alon Levy and Dan Suciu, "**A query language for XML**", *Eighth International World Wide Web Conference*, pp. 77-91, Toronto 1999

Dobie et al. 1999

Mark Dobie, Robert Tansley, Dan Joyce, Mark Weal, Paul Lewis and Wendy Hall, "**MAVIS 2: A New Approach to Content and Concept Based Navigation**", Proceedings of the IEE Colloquium on Multimedia Databases and MPEG-7, Vol 99, no. 56, pp. 9/1-9/5, 1999

Dunlop et al. 1991a

M. D. Dunlop & C. V. van Rijsbergen, "**Hypermedia & Probabilistic Retrieval**", Proceedings of the RIAO 91 Conference on Intelligent Text and Image Handling, pp 337-356, 1991

Dunlop 1991

Mark D. Dunlop, "**Multimedia Information Retrieval**", PhD Thesis, Glasgow University, 1991

Dunlop et al. 1993

Mark D. Dunlop & C.J. van Rijsbergen, "**Hypermedia and Free Text Retrieval**", Information Processing and Managment, Vol. 29, No 3, pp 287-298, May 1993.

Edmundson et al. 1961

H. P. Edmundson & R. E. Wyllys, "**Automatic Abstracting and Indexing - Survey and Recommendations**", Communications of the ACM, Vol 4., No. 5, pp. 226-234, 1961

Florescu et al. 2000

Daniela Florescu, Donald Kossmann and Ioana Manolescu, "**Integrating keyword search into XML query processing**", *Ninth International World Wide Web Conference*, pp. 119-135, Amsterdam 2000

Foskett 1997

D. J. Foskett "**Thesaurus**", In Karen Sparck Jones and Peter Willett, "Readings in Information Retrieval", Morgan Kaufmann Publishers, San Francisco, 1997

Fountain et al 1991

Andrew M. Fountain, Wendy Hall, Ian Heath & Hugh C. Davis, "**MICROCOSM: An Open Model for Hypermedia With Dynamic Linking**", Hypertext: Concepts, Systems and Applications (Proceedings of ECHT'90), pp 298-311, 1990, Cambridge University Press

Frakes et al. 1992

William B. Frakes and Ricardo Baeza-Yates, "**Information Retrieval: Data Structures & Algoritms**", Prentice Hall, 1992

Frankel et al. 1996

Charles Frankel, Michael Swain and Vassilis Athitsos, "**WebSeer: An Image Search Engine for the World Wide Web**", Technical Report 96-14, University of Chicago, Computer Science Department, August 1996

Frei et al. 1992

H.P. Frei & D. Stieger, "**Making Use of Hypertext Links when Retrieving Information**", Proceeding of the ACM Conference on Hypertext, pp 102-111, 1992, ACM press

Frei et al. 1995

H.P. Frei & D. Stieger, "**The use of semantic links in hypertext information retrieval**", Information Processing and Management, Vol. 31, No. 1, pp 1-13, 1995

Frisse 1988

Mark E. Frisse, "**Searching for information in a hypertext medical handbook**" Communications of the ACM, Vol. 31, No. 7, pp 880-886, July 1988

Frisse et al. 1989

Mark E. Frisse & Steve B. Cousins "**Information Retrieval from Hypertext: Update on the Dynamic Medical Handbook Project**", Proceedings of the ACM Conference on Hypertext, pp 199-212, November 1989, ACM Press

Golovchinsky 1997

Gene Golovchinsky, "**What the Query Told the Link: The Integration of Hypertext and Information Retrieval**", The Eighth ACM Conference on Hypertext, pp 67-74, April 1997, ACM Press

Grønbæk et al. 1996

Kaj Grønbæk and Randal H. Trigg, "**Toward a Dexter-based model for open hypermedia: Unifying embedded references and link objects**", *The Seventh ACM Conference on Hypertext (Hypertext'96)*, pp 149-160, March 1996, ACM Press

Grønbæk et al. 1997

Kaj Grønbæk, Niels Olof Bouvin and Lennert Sloth, "**Designing Dexter-based hypermedia services for the World Wide Web**", *The Eight ACM Conference on Hypertext*, pp 146-156, 1997, ACM Press

Grønbæk et al. 1999

Kaj Grønbæk, Lennert Sloth and Peter Ørbæk, "**Webvise: browser and proxy support for open hypermedia structuring mechanisms on the World Wide Web**", *Eighth International World Wide Web Conference*, pp. 253-267, Toronto 1999

Grønbæk et al. 2000

Kaj Grønbæk, Lennert Sloth and Niels Olof Bouvin, **"Open hypermedia as user controlled meta data for the Web"**, *Ninth International World Wide Web Conference*, pp. 553-566, Amsterdam 2000

Guinan et al. 1992

Catherine Guinan and Alan F. Smeaton, "**Information Retrieval from Hypertext using Dynamically Planned Guided Tours**", Proceeding of the ACM Conference on Hypertext, pp 122-130, 1992, ACM Press

Halasz et al. 1994

Frank Halasz and Mayer Schwartz, "**The Dexter Hypertext Reference Model**", Communications of the ACM, vol. 37 (2), 30-39

Hall 1994a

Wendy Hall, "**Ending the Tyranny of the Button**", Multimedia, pp 60-68, Spring 1994, IEEE

Hall et al. 1994b

Wendy Hall and Hugh Davis, "**Hypermedia link services and their application to multimedia information management**", Information and Software Technology, Vol. 36, No. 4, pp 197-202, 1994

Hall et al. 1996

Wendy Hall, Hugh Davis and G. Hutchings, "**Rethinking Hypermedia: The Microcosm Approach**", Kluwer, Boston 1996

Harmandas et al. 1997

V. Harmandas, M. Sanderson & Mark D. Dunlop, "**Information Retrieval by Hypertext Links**", Proceeding of SIGIR 97, 27-31 July 1997

Hartman et al. 1997

John H. Hartman, Todd A. Proebsting & Rajesh Sundaram, **"Index-based Hyperlinks"**, *Sixth International World Wide Web Conference*, TEC 127, pp. 41-47, 1997

Hirata et al. 1993

Kyoji Hirata, Yoshinori Hara, Naoki Shibata & Fusako Hirabayashi, "**Media-based Navigation for Hypermedia Systems**", Proceeding of the ACM Conference on Hypertext, pp 159-173, November 1993, ACM press.

Hirata et al. 1996

Kyoji Hirata, Yoshinori Hara, Hajime Takano & Shigehito Kawasaki, "**Content-oriented Integration in Hypermedia**", The Eighth ACM Conference on Hypertext, pp 75-86, April 1997, ACM Press

Hirata et al. 1997

Kyoji Hirata, Sougata Mukherjea, Yusaka Okamura, Wen-Syan Li & Yoshinori Hara, "**Object-based Navigation: An Intuitive Navigation Style for Content-oriented Integration Environment**", The Seventh ACM Conference on Hypertext (Hypertext'96), pp 11-21, March 1996, ACM Press

Huang et al. 2000

Lieming Huang, Matthias Hemmje and Erich Neuhold, "**ADMIRE: an adaptive data model for meta search engines**", *Ninth International World Wide Web Conference*, pp. 431-448, Amsterdam 2000

Jenkins et al. 1999

Charlotte Jenkins, Mike Jackson, Peter Burden and Jon Wallis, "**Automatic RDF metadata generation for resource discovery**", *Eighth International World Wide Web Conference*, pp. 227-242, Toronto 1999

Kaindl et al. 1998

Hermann Kaindll, Stefan Kramer and Luis Miguel Afonso, "**Combining Structure Search and Content Search for the World-Wide Web**", Proceedings of Hypertext 98, pp. 217-224, Pittsburgh PA USA, ACM 1998

Kaindl et al. 1999

Hermann Kaindll and Stefan Kramer, "**Semiautomatic Generation of Links: A Practical Solution**", Proceedings of Hypertext 99, pp. 3-12, Darmstadt Germany, ACM 1999

Korfhage 1997

Robrt R. Korfhage, "**Information Storage and Retrieval**", John Wiley & Sons, 1997

Lassila et al 1999

Ora Lassila and Ralph R. Swick, "**Resource Description Framework (RDF) Model and Syntax Specification**", W3C Recommendation, 12 February 1999, http://www.w3.org/TR/REC-rdf-syntax

Lee et al. 1997

Kyuchul Lee, Yong Kyu Lee and P Bruce Berra, "**Management of Multi-structured Hypermedia Documents: A Data Model, Query Language, and Indexing Scheme**", Multimedia Tools and Applications, 4, 199-223, Kluwer Academic 1997

Lewis et al 1996a

Paul Lewis, Hugh Davis, Steve Griffiths, Wendy Hall & Rob Wilkins, "**Media-based Navigation with Generic Links**", The Seventh ACM Conference on Hypertext (Hypertext'96), pp 215-223, March 1996, ACM Press

Lewis et al. 1996b

Paul H. Lewis, Hugh C. Davis, Mark R. Dobie & Wendy Hall, "**Towards Multimedia Thesaurus Support for Media-based Navigation**", First International Workshop on Image Databases and Multimedia Search, Amesterdan, 5th November 1996

Li et al. 1992

Z. Li, H. Davis & W. Hall, "**Hypermedia Links and Information Retrieval**", British Computer Society 14th Information Retrieval Colloquium, 13-14 April 1992, Lancaster University

Li 1993

Zhuoxun Li, "**Information retrieval for automatic link creation in hypertext systems**", PhD thesis, University of Southampton, Department of Electronics and Computer Science, 1993

Lie et al 1999

Håkon Wium Lie and Janne Saarela, "**Multipurpose Web Publishing using HTML, XML, and CSS**", Communications of the ACM, vol. 42, no. 10, pp. 95-101, October 1999

Malcolm et al 1991

Kathryn C. Malcolm, Steven E. Poltrock and Douglas Schuler, "**Industrial Strength Hypermedia: Requirements for a Large Engineering Enterprise**", Proceeding of the ACM Conference on Hypertext, pp 13-24, 1991, ACM press

Marchiori 1997

Massimo Marchiori, "**The Quest for Correct Information on the Web: Hyper Search Engines**", *Sixth International World Wide Web Conference*, TEC 123, pp. 265-276, 1997

Marchiori 1998

Massimo Marchiori, "**The limits of Web metadata and beyond**", Computer Networks and ISDN Systems, 30, 1-9, 1998

Marinheiro et al. 1998

Rui M N Marinheiro and Wendy Hall, "**Expanding a Hypertext Information Retrieval System to Incorporate Multimedia Information**", Proceedings of the IEEE 31st Hawaii International Conference on System Sciences, 6-9 January, 1998

Mitra et al. 2000

M. Mitra and B. B. Chaudhuri, "**Information Retrieval from Documents: A Survey**", Information Retrieval 2, pp 141-163, Kluwer Academic Publishers, 2000

Mizuuchi et al. 1999

Yoshiaki Mizuuchi and Keishi Tajima, "**Finding Context Paths for Web Pages**", Proceeding of the ACM Conference on Hypertext, pp. 13-22, 1999

Moreau et al. 2000

Luc Moreau et al., "**SoFAR with DIM Agents: An Agent Framework for Distributed Information Management**", The fifth International Conference and Exhibition on The Practical Application of Intelligent Agents and Multi-Agents, Manchester, UK, 2000

Mukherjea et al. 1997

Sougata Mukherjea, Kyoji Hirata and Yoshinori Hara, "**Towards a Multimedia Worlds Wide Web Information Retrieval Engine**", *Sixth International World Wide Web Conference*, TEC 130, pp. 177-188, 1997

Mukherjea et al. 1999

Sougata Mukherjea and Junghoo Cho, "**Automatically Determining Semantics for World Wide Web Multimedia Information Retrieval**", Journal of Visual Languages and Computing, 10, pp. 585-606, 1999

Narasimhalu et al 1995

Desai Narasimhalu & Mun-Kew Leong, "**Experiences with Content Based Retrieval of Multimedia Information**", Proceedings of the Final Workshop on Multimedia Information Retrieval (MIRO Glasgow '95), pp 6, September 1995, Ian Ruthven (Ed)

O'Brien 2000

Danny O'Brien, "**The fine art of Googling**", The Sunday Times, pp 47-50, 6th August 2000

O'Docherty et al. 1990

M. H. O'Docherty and C. N. Daskalakis, "**Multimedia Information Systems - The Management and Semantic Retrieval of all Electronic Data Types**", The Computer Journal, Vol 34, No. 3, pp 225-238, 1990

OHSWG 1998

OHSWG, Open Hypermedia Systems Working Group WWW site, 1998, http://www.ohswg.org/

Ortega et al. 1997

Michael Ortega, Yong Rui, Kaushik Chakrabarti, Kriengkrai Porkaew, Thomas S. Huang and Sharad Mehrotra, "**Supporting Ranked Boolean**

**Similarity Queries in MARS**", IEEE Transactions on Knowledge and Data Engineering, vol. 10 (6), pp. 905-925, 1998

Poulter 1997

Alan Poulter, "**The design of World Wide Web search engines: a critical review**", Program – Electronic Library & Information Systems, vol. 31, no. 2, pp. 131-145, April 1997

QL 1998

QL 1998, The Query Languages Workshop, Boston, December 5 1998

Rasmussen 1992

Edie Rasmussen, "**Clustering Algorithms**", In William B Frakes and Ricardo Baeza-Yates (eds.), Information Retrieval - Data Structure & Algorithms, pp 419-442, Prentice Hall, 1992

Rijsbergen 1977

C. J. van Rijsbergen, "**A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval**", Journal of Documentation, Vol. 33, No. 2, pp. 106-119, June 1977

Rijsbergen 1979

C.J. van Rijsbergen, "**Information Retrieval**", Butterworths, 2nd ed. London 1979

Salton 1989

Gerard Salton, "**Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**", Addison-Wesley, 1989

Salton et al 1994

Gerard Salton, James Allan, Chris Buckley & Amit Singhal, "**Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts**" Science, Vol. 264, pp 1421-1426, 3 November 1994

Salton et al 1996

Gerard Salton, Amit Singhal, Chris Buckley & Mandar Mitra, "**Automatic Text Decomposition Using Text Segments and Text Themes**", The Seventh ACM Conference on Hypertext (Hypertext'96), pp 53-65, March 1996, ACM Press

Savoy 1996

Jacques Savoy, "**Citation Schemes in Hypertext Information Retrieval**", In Maristella Agosti & Alan F. Smeaton (eds.), Information Retrieval and Hypertext, pp 99-120, Kluwer Academic Publishers, 1996

Schwartz 1998

Candy Schwartz, "**Web Search Engines**", Journal of the American Society for Information Science, vol. 49 (11), pp. 973-982, 1998

Smith et al 1997

John R. Smith and Shih-Fu Chang, "**Visually Searching the Web for Content**", IEEE Multimedia, pp 12-20, July-September 1997

Sparck Jones et al. 1997

Karen Sparck Jones and Peter Willett, "**Readings in Information Retrieval**", Morgan Kaufmann Publishers, San Francisco, 1997

Spertus et al. 2000

Ellen Spertus and Lynn Andrea Stein, "**Squeal: a structured query language for the Web**", *Ninth International World Wide Web Conference*, pp. 97-103, Amsterdam 2000

Srihari et al. 1999

Rohini K. Srihari and Zhongfei Zhang, "**Exploiting Multimodal Contenxt in Image Retrieval**", Library Trends, pp. 496-524, Fall 1999

Srinivasan 1992

Padmini Srinivasan, "**Thesaurus Construction**", In William B Frakes and Ricardo Baeza-Yates (eds.), Information Retrieval - Data Structure & Algorithms, pp 161-218, Prentice Hall, 1992

Tansley 2000

Robert Tansley, “**The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information**”, PhD Thesis, Southampton University, 2000

Taylor et al. 1995

Carl Taylor, Douglas Tudhope & Paul Beynon-Davies, “**A semantic modelling approach to hypermedia systems for museums**”, The Future for Europe’s Past - Conference Proceedings, 1.73 - 1.81, 1995

TREC

TREC, Text REtrieval Conference, http://trec.nist.gov

Yankelovich et al. 1988

N. Yankelovich, B. J. Haan, N. Meyrowitz and S. M. Drucker, “**Intermedia: The Concepts and the Construction of a Seamless Information Environment**”, IEEE Computer, vol. 21 (1), pp. 81-96, 1988

Weibel et al. 1998

S. Weibel and E. Miller, “**Dublin Core Metadata**”, http://purl.oclc.org/metadata/dublin_core/, http://purl.oclc.org/dc, November 1998