pp 240-242 in Proc 8th Int Conf on User Modeling, Sonthofen, Germany, 2001

Using Document Structures for Personal Ontologies and User Modeling*

Sanghee Kim[1], Wendy Hall[1] and Andy Keane[2]

[1]Intelligence, Agents, Multimedia Group, Department of Electronics and Computer Science, University of Southampton, U.K., Tel: 44-(0)23-80-59-3256, Fax: 44-(0)23-80-59-2865
{sk98r,wh}@ecs.soton.ac.uk
[2]Computational Engineering and Design Center, School of Engineering Science, University of Southampton, U.K., Tel: 44-(0)23-59-2944, Fax:44-(0)23-80-59-3220
ajk@soton.ac.uk

Abstract. We present a new approach that makes use of the embedded structural information of the documents which a user frequently refers to for deriving a personalized concept hierarchy and for identifying user preferences concerning document searching and browsing.

Keywords. Personal ontology, Supported browsing, Structured document

1 Introduction

As it has come to be harder for a user to locate relevant documents quickly, the intelligent document organization system that proactively retrieves, categorizes, and ranks searched documents has become an essential component of personal desktop applications. Not only does this system is required to model the salient characteristics of the user, but it also has to integrate heterogeneous document sources, each of which has its own structural elements and vocabularies. Reviews of current approaches show that the embedded structural elements of the documents are often ignored in building such systems even though these present semantic clues about the kind of underlying document handling that would provide relevant descriptions of the document.

2 Personal Ontology Building

Personal ontologies provide a uniform representation of a user's personally archived documents by explicitly specifying structural elements and their meanings. Initial information about the user is bootstrapped by making use of existing user defined categories, such as file directories, folder hierarchies of e-mail messages or a user's homepage. Not only does this initialization reduce the user's reliance on system interaction, but it also provides good training examples for profiling the user. The development of the personal ontologies involves three steps. First, it requires the definition of a hierarchy of structural elements and the extraction of the values of the

* The work presented here has been supported by UTP (University Technology Partnership in Design) program sponsored by Rolls Royce Plc and BAE Systems.

specified tags from the documents. Given the structured documents, a set of top-level concepts can then be derived from the user-defined categories through a two-step category clustering process. Finally, the explicit specification of relations among the concepts can be defined by using the technique of formal concept analysis.

Formal concept analysis (FCA) is based on an applied lattice theory and defines the formal context consisting of a set of concepts, each of which is specified as: (C,A_1), where C is the extent that holds all objects belonging to the concept, and A_1 is the intent that comprises all attributes valid for those objects [1]. In the personal ontology, the object corresponds to a document and the associated attributes are extracted features. A concept lattice is constructed through an attribute exploration process that extensively investigates the combinations of related attributes to define co-relationships among the concepts. A super or a sub-concept is defined as: $(D,A_1)\subseteq (C,A_2)\Leftrightarrow D\subseteq C\ (\Leftrightarrow A_2\subseteq A_1)$ meaning that a concept C is a super-concept of D if and only if the attributes (A_2) of C are a subset of the attributes (A_1) of concept D.

3 Reinforcement Learning for Ranking Structured Documents

User preferences are modelled using two layers (i.e. global and local) in order to take into account semantic clues defined in the structural tags, so that terms can be differentiated which refer to different objects. The global profile represents overall user browsing preferences regarding the distribution of structural elements, each of which specifies its strength by a numeric weight. The local profile defines a set of features associated with importance weights for each element. Reinforcement learning is selected since its computation can be incrementally updated on-line and it directs a learner towards an optimal state in the future [3]. The feedback obtained from the user is through observation of whether or not the user clicks the specific document, and it is incorporated as the immediate reward which defines the effectiveness of ranking strategies. RLRSD (Reinforcement Learning for Ranking Structured Documents) ranks retrieved documents by their estimated relevance values to a user's query, so that the first placed document is presumably the most relevant. It also takes into account the feature differences of the ranked documents plus the previously learned user profiles in order to evaluate the value of next ranking strategy.

4 Evaluations

Two datasets were collected. One of the authors provided the first dataset which had a total of 1405 documents (1148 email messages, 166 bookmarked web pages, 88 texts in postscript format, and 3 web pages from a homepage) collected from 94 categories. 176 queries were simulated by assigning query terms from randomly selected document structures. The performance of RLRSD was compared to that of a flat vector model (FVM) which represents documents as an unstructured single layer. The approximately 37% higher precision of RLRSD shows the efficiency of using the structural information by assigning different weights to a feature that relates to different objects. We suggest that this is mostly due to a term confusion problem caused by the fact that no differences of structures are made in FVM. In other words, although it received feedback, it did not reflect the different contributions of a feature

which relates to different objects; instead it assigned a uniform weight to a feature irrespective of its linked objects.

For the second experiment, we downloaded the publicly available cystic fibrosis (CF) dataset. It consists of 1239 XML documents indexed with the term 'cystic fibrosis' in the National Library of Medicine's MEDLINE database [2]. It gathered subjective relevance judgments (i.e., highly relevant, marginally relevant, or irrelevant) made by four users against the selected 100 queries. We created an individual relevance score file per user to compare the ranking results by using different structural elements, to which the user referred in deciding the scores. A user5 was artificially created by combining the relevance scores produced by the four users.

User4 shows a unique result in that no specific preferences concerning the document structures that the user made use of in deciding scores are observed. In fact, while the other three users are domain experts, user4 is a bibliographer who presumably has better knowledge of citation and reference elements. User4's relatively higher precision rate in terms of the use of the citation, the reference, and the source tag compared to those of the other three users confirms our finding that user4 uses bibliographic-related tags heavily.

The performance difference (11%) between RLRSD and FVM on the second experiment was not so significant compared to that (37%) of the first experiment. We base our conclusion on the following observation. Dataset2 shared the same structural specification, while dataset1 originated from heterogeneous sources and thus the variations in the distribution of the structural elements are wider than those in dataset2. As RLRSD utilizes both the distribution of structural elements and features, it is clear that RLRSD on dataset2 could not take full advantage of its global profile compared to dataset1. Moreover, while there were only three kinds of user feedback on dataset2, dataset1 used a minimum of one and a maximum of thirty kinds of feedback. The performance of user5 (who received nine kinds of feedback) showed higher precision values than those of the other three users. This proves that the ranking algorithms can improve ranking performance when feedback reflects more detailed user preferences.

5 Future Work

We have not yet fully investigated the impact caused by varying the number of retrieved documents on the performance results as our work assigned a fixed maximum number to all local profilers regardless of their different strengths. In addition, our assumption that user clicked documents are mostly useful in identifying user preferences needs to be further validated through a real-user interaction with our system.

References

1. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis. Preprints, http://wwwbib.mathematik.tu-darmstadt.de/Math-Net/Preprints/Listen/pp97.html (1997)
2. Shaw, W., Wood, R., Tibbo, H.: The Cystic Fibrosis Database: Content and Research Opportunities. Int. J. Library and Information Science Research Vol. 13. (1991) 347-366
3. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. http://www-anw.cs.umass.edu/~rich/book/the-book.html (1998)