Using Windmill Expansion for Document Retrieval

Shao Fen Liang*, Paul Smart, Alistair Russell and Nigel Shadbolt
School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK

Abstract

SEMIOTIKS aims to utilise online information to support the crucial decision–making of those military and civilian agencies involved in the humanitarian removal of landmines in areas of conflict throughout the world. An analysis of the type of information required for such a task has given rise to four main areas of research: information retrieval, document annotation, summarisation and visualisation. The first stage of the research has focused on information retrieval, and a new algorithm, "Windmill Expansion" (WE) has been proposed to do this. The algorithm uses retrieval feedback techniques for automated query expansion in order to improve the effectiveness of information retrieval. WE is based on the extraction of human–generated written phases for automated query expansion. Top and Second Level expansion terms have been generated and their usefulness evaluated. The evaluation has concentrated on measuring the degree of overlap between the retrieved URLs. The less the overlap, the more useful the information provided. The Top Level expansion terms were found to provide 90% of useful URLs, and the Second Level 83% of useful URLs. Although there was a decline of useful URLs from the Top Level to the Second Level, the quantity of relevant information retrieved has increased. The originality of SEMIOTIKS lies in its use of the WE algorithm to help non–domain specific experts automatically explore domain words for relevant and precise information retrieval.

Keywords information retrieval, query expansion, retrieval feedback, humanitarian demining

1. Introduction

Our project, SEMantically-enhanced InformatiOn extracTion for Improved Knowledge Superiority (SEMIOTIKS), is a DIF DTC[1] project. It aims to meet the challenges faced by military and civilian agencies by maximising the potential of large–scale information repositories to support operationally effective decision–making.

The United States Department of Defence[2] suggests that there are approximately 55 million landmines in nearly 60 countries that cause over 10,000 casualties each year. Landmines have been in use for over a century, remaining long after a period of conflict has ended. The mines kill and maim civilians, and make large areas of land unusable for agriculture and development. In order to help civilians recover their countries and lives, humanitarian agencies are tasked with the clearance of landmined areas.

Although there are many opportunities for software-assisted support for humanitarian demining operations, the main concern of our project is to utilise open source information that supports crucial decision–making. Demining operations consist of a number of knowledge–intensive tasks, (e.g. decisions

* Corresponding author. Tel.: +44 238 059 3269
E-mail addresses: sfl@ecs.soton.ac.uk (S.F. Liang), ps02v@ecs.soton.ac.uk (P. Smart), ar5@ecs.soton.ac.uk (A. Russell), nrs@ecs.soton.ac.uk (N. Shadbolt)

[1] http://www.difdtc.com/

[2] http://www.humanitarian-de-mining.org/de-mining/threats/proliferation.asp

about the relative priority of mined areas), and an ability rapidly to exploit open source information. The latter is a critical element of success in these tasks.

Agencies involved in humanitarian demining typically have a number of information requirements:

1. Background: information about the history, geography, political situation, economic structure, infrastructure and culture of the country provides background knowledge about the area of operations. Baseline data is also necessary for aid organisations to be able to compare an emergency situation to previous conditions.

2. Situational Awareness: aid organisations need to know the latest situations and conditions on the ground, and the needs and locations of affected populations.

3. Operational/Programmatic: necessary information to plan and implement humanitarian assistance programs. This includes information about access routes, the activities and locations of other organizations, and the availability of humanitarian resources.

4. Analysis: humanitarian information needs to be interpreted in context and related to other semantic information. Analyses can include evaluations of issues and responses, projections about the future, and recommendations for policies and actions.

In order to achieve the objectives of SEMIOTIKS, the focus of the research needs to adhere to the four information categories. The first of these is information retrieval to gather relevant information in humanitarian demining domain. The second is semantic annotation to annotate the retrieved documents into their semantic meanings in a task–oriented form. The third area of focus is on automatic summarisation to summarise the annotated documents to support humanitarian demining decision–making; and the fourth is to focus on information browsing and visualisation in order to provide the end user with information in an attractive and relevant visual form; and within an easy–browsing environment.

This paper presents the results of using a retrieval feedback technique to improve the effectiveness of information retrieval. The paper is organised as follows: Section 2 presents related research on automatic query expansion techniques for information retrieval; it also describes the motivation behind our proposed query expansion algorithm. Our new algorithm is called Windmill Expansion (WE), which is based on extracting human–generated written phrases for automatic query expansion. The algorithm is presented in Section 3, along with details of its implementation. Section 4 evaluates our approach by measuring the usefulness of retrieved documents in relation to demining operations. Sections 5 and 6 present our conclusions and provide an overview of future work.

2. Related research

A key research challenge for humanitarian demining relates to information retrieval. Since the knowledge sources for demining are not easy to obtain, we need to develop an improved information retrieval mechanism. We immediately have a problem with the vocabulary, which has been discussed in Furnas et al.[1]. The vocabulary problem relates to the deleterious effect of incorrect word usage and short queries on the effectiveness of human computer interaction. This is a fundamental problem in current information retrieval research, but the problem can be remitted by using Query Expansion techniques [2].

A common approach to query expansion is to assess a collection of relevant information and then select a set of possible expansion terms which might help to retrieve more relevant documents. The process of selecting expansion terms can be performed by the user as Interactive Query Expansion (IQE), or by the machine as Automatic Query Expansion (AQE). It has been suggested that IQE allows greater control in terms of selecting the criterion of relevance; while AQE has the ability to access more relevant information to generate higher statistical relative terms [3]. In terms of our research, AQE is preferable because it does not require human expert knowledge in the humanitarian demining domain.

Using a thesaurus is one of the most common techniques for realising AQE. Thesauri can provide synonym sets and word relationships for query expansions as exemplified in the work of Cooper, Byrd [4]

and Carmel et al. [5]. However, in Voorhees's [6] study, expanding queries with their lexical-semantic relations provided little benefit in improving retrieval effectiveness than just used the original query if the original query was described well of the sought information. Analysing the relationships between contents and documents in a collection is another technique for AQE. Kim et al. [7] analysed the document concept to derive a set of similar terms for query expansion. Deerwester et al. [8] used factors' analysis to select statistical factors from documents to enhance retrieval effectiveness. However, the computational inefficiency of context analysis techniques is a key limitation to their use in information retrieval scenarios [9].

Another popular approach is so-called retrieval feedback [10][11][12]. This approach utilises an initial query to derive a set of top–ranked documents. These documents are assumed to be the documents of most relevance to the initial query. A set of possible expansion terms is then analysed to add to the initial query. This approach attracted interest in developing ranking algorithms to derive better expansion terms. Gordon [13] examined the similarity of human generated descriptions to each article, and found that the documents in a same cluster had the same terms preserved in the human produced description. His work motivated us to propose the new query expansion algorithm based on extracting key phrases from human written documents.

3. Windmill Expansion algorithm

The new algorithm we have proposed for retrieving documents is called Windmill Expansion (WE). It is based on the notion of extracting human written phrases to be used as the expansion terms. The idea is that the documents posted on the web are mostly written by humans, therefore extracting the human phrases is a natural way of finding the most popular natural language phrases in the humanitarian demining domain.

The pseudo code of the WE algorithm is shown in FIGURE 1. This algorithm is performed by creating a meta–search engine and uses the initial query "demining" to retrieve documents from the meta–search engine. The search engine retrieves a number of pages in the second statement. HTML tags and stopwords are removed from each retrieved page. Subsequently, the words that appear after "demining" are extracted as the expansion candidates, and they are stemmed by Porter Stemmer [14][15]. The stemming process aims to deal with morphological variants of the same term (e.g. "activities" and "activity" are counted as the same word). Finally, the top 10 most frequent terms are selected as the expansion terms.

```
Initial query "demining";
Retrieve N pages from meta search engine;
For each URL
{ remove HTML tags;
  remove stopwords
    extract $phrase =   "demining+$WordAfterDemining";
    record $phrase into @phrase }
For each $phrase in @phrase
{stem $phrase;
count frequency of $StemedPhrase;}
Select top 10 $StemedPhrase;
```

FIGURE 1. The pseudo code of the WE algorithm.

3.1. Retrieval feedback optimisation

How many retrieved pages should be processed to find the best of the top 10 frequent terms? In order to answer this question we have investigated 10 different categories and each one has a different number of retrieved pages. Category 1 contained the top 100 retrieved pages, category 2 contained the top 200 retrieved pages, category 3 contained the top 300 retrieved pages and so on. The reason for using up to 10 categories is because many search engines can only return around 1000 pages for each query.

The top 10 terms generated in each category were slightly different across the 10 categories because the number of pages in each category were different. In order to compare the difference among the 10 categories, the frequencies of a total of 100 terms in the 10 categories were calculated, and the top 10 most frequent terms were selected. The results are presented in TABLE 1. The 10 most frequent terms are listed in alphabetical order in the far left column in TABLE 1. The digital numbers from 1 to 10 in the top row represent the 10 categories. Each missing term of the 10 most frequent terms in the 10 categories is marked with an "X" in the corresponding cell. For example, "demining assistance" is missing in category 1 and "demining program" is also missing in categories 1, 2, 3 and 4. The overall missing terms are presented in the last row, where 4 terms are missing in category 1; 2 in categories 2, 3 and 4; 1 in categories 5, 6 and 7, and 0 in categories 8, 9 and 10. The correlation between the number of missing terms and the number of retrieved pages from categories 1 to 10 is that the number of missing terms decreases from 4 to 0, while the number of retrieved pages increases from 100 to 1000.

TABLE 1. Top 10 frequent phrases across 10 categories.

	1	2	3	4	5	6	7	8	9	10
demining activity										
demining assistance	x									
demining effort										
demining equipment										
demining operation	x									
demining program	x	x	x	x						
demining project					x	x	x			
demining team										
demining technology	x									
demining training		x	x	x						
Number of Missing Terms	4	2	2	2	1	1	1	0	0	0

This result indicates that the more documents we retrieved to perform WE, the fewer missing terms we encountered in the top 10 frequent terms. This information suggests that we need to retrieve the top 1000 pages for generating query expansion terms in order to achieve query expansion optimisation. Thus the 10 terms in TABLE 1 were the 10 expansion terms generated from the initial query of "demining" and were called Top Level terms.

The Top Level terms were used to re–perform WE to expand the query length from one word to three words, and the new terms generated were called Second Level terms. In fact, WE is able to expand query terms from 1 to 10, 10^2, $10^{3,}$ and so on, as the base 10 exponential function. However, in this experiment, we only implemented to the Second Level terms. FIGURE 2 shows all terms generated in Top Level and Second Level.

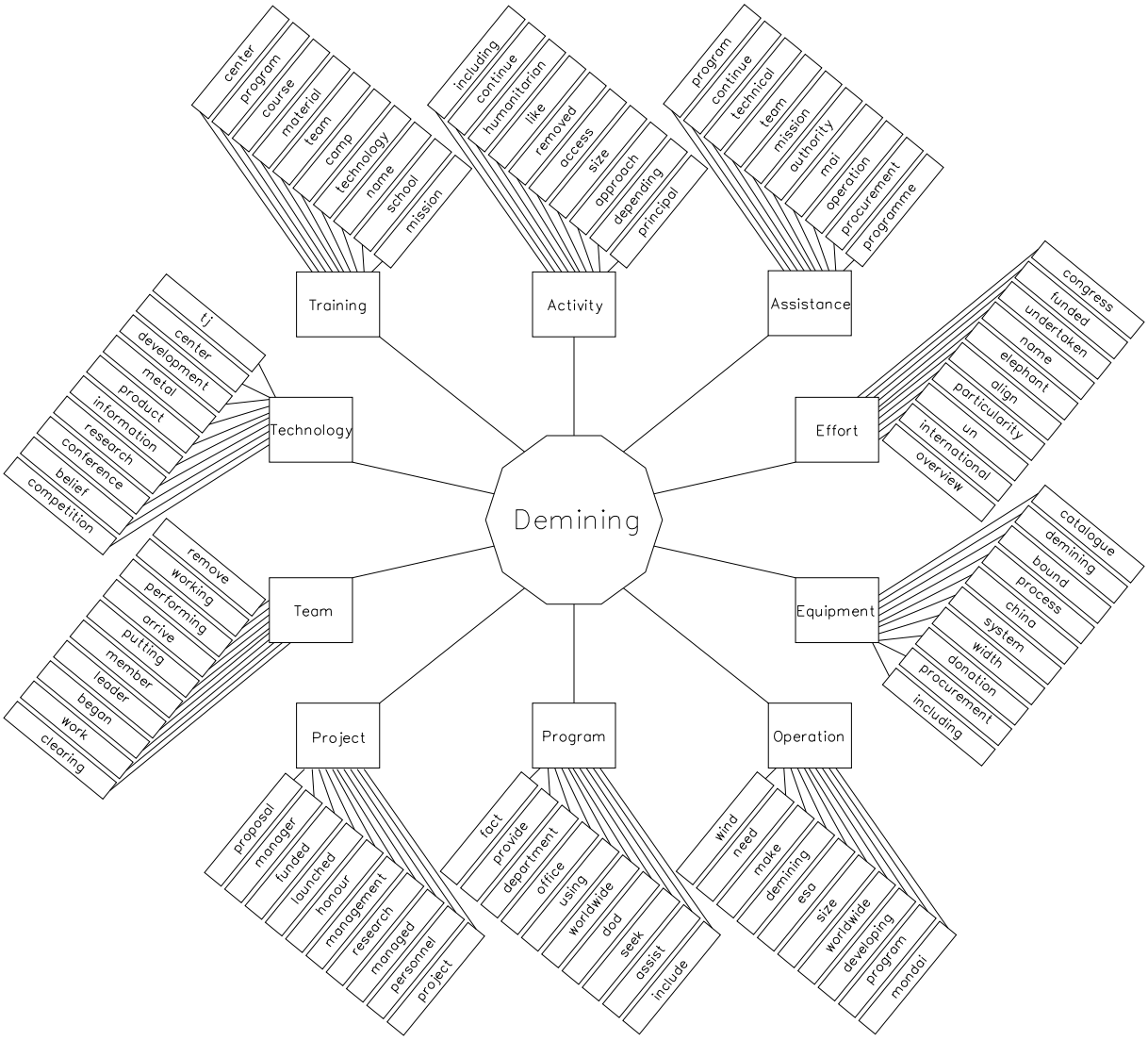

FIGURE 2 Top and Second Level terms of Demining.

The generated Second Level terms were manually examined and it was found that WE has the ability to extract not only human written phrases in this domain, but also the ability to extract terminology, stakeholder and named entity in the humanitarian demining domain. For example, the Second Level

term “mai” expanded from “assistance”, represents a type of blast mine such as MAI 2, MAI 68 and MAI 75. The Second Level term “un” from “effort”, represents the United Nations and “esa” in “operation”, represents the European Space Agency which is one of the stakeholders in humanitarian demining operations. The “tj” in “technology”, represents the authors T. J. Bloodworth who wrote articles about demining technology. The meaning of these terms was not obvious to us until these key words were used to retrieve documents; they were then found to be the specific query words of the humanitarian demining domain used to retrieve the most relevant documents.

In addition to the above experiments, we also ascertained whether the WE algorithm is a generic algorithm and able to perform other queries in a random domain. We randomly choose “Alzheimer” as the initial query from the Health domain and tried to generate two levels of expansion terms. TABLE 2 shows 10 Top Level expansions of “Alzheimer”. The most interesting term is the seventh term “dementia”, which gave us a clearer indication of this disease than the single query term “Alzheimer” itself.

TABLE 2 Top Level expansion terms of Alzheimer

alzheimer disease
alzheimer society
alzheimer research
alzheimer patients
alzheimer asssociation
alzheimer drug
alzheimer dementia
alzheimer scotland
alzheimer sufferers
alzheimer forum

FIGURE 3 show the Second Level expansion terms arising from the Top Level terms. These raise some interesting connections. For example, “sarasota” is the tenth expansion term arising from “Alzheimer sufferers”. Sarasota is where a team of research scientists work on new treatments for Alzheimer’s disease. “Inverclyde” is an expansion from “alzheimer scotland” and is a Scottish project providing support for Alzheimer’s sufferers and carers. These two examples show that the expansion terms enable the user to retrieve more relevant information about Alzheimer’s disease.

Although the retrieved documents were manually scanned to evaluate their relevance, a more objective method to prove their usefulness would provide better evidence. Therefore an advance evaluation was needed.

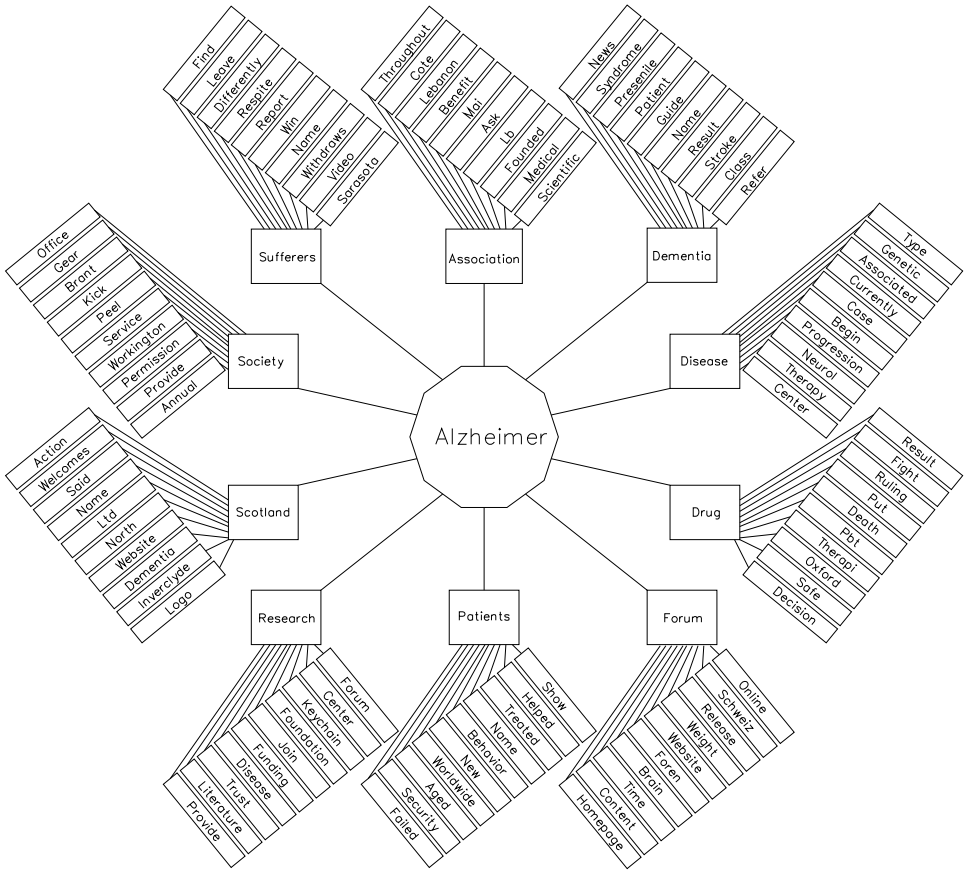

FIGURE 3 Top and Second Level terms of Alzheimer.

4. Evaluation of the WE algorithm

Precision and Recall [16] are commonly used for evaluating information retrieval techniques. However, these two metrics are not suitable in our case because of the lack of baseline data. Thus, according to the retrieval objectives of our project, measuring how much useful information could be provided by the WE algorithm was the focal point of our evaluation. Towards this objective, the evaluation was constructed under the assumption that the retrieved documents were relevant.

Our hypothesis was that using the WE algorithm for query expansion could improve retrieval effectiveness in terms of providing more useful information. In order to accept this hypothesis, the design of the evaluation focused on measuring the overlap between the retrieved documents. This was

because we did not want a high ratio of overlapping to reduce the number of useful documents. Thus, the fewer overlapped documents, the more useful information the algorithm provides.

To implement the evaluation, the Top Level terms were used to query our meta–search engine. Each Top Level query provides 10 URLs, thus a total of 100 URLs were retrieved. . In order to measure the degree of usefulness, the overlapping URLs which appeared in each query were counted and deducted from its total contributed pages. The results are presented in TABLE 3, where "demining effort", "demining operation" and "demining team" appear to have 1 page of overlap in each. Therefore they only contributed 9 pages to the final total of retrieved pages. The result shows that the Top Level query terms retrieved 97 out of 100 pages of useful information, so the percentage of usefulness is 97%.

TABLE 3 Contribution from Top Level terms - demining

Top Level expansion term	Number of overlapped pages	Total contributed pages
demining activity	0	10
demining assistance	0	10
demining effort	1	9
demining equipment	0	10
demining operation	1	9
demining program	0	10
demining project	0	10
demining team	1	9
demining technology	0	10
demining training	0	10
Total pages (out of 100)		97

TABLE 4 Contribution from Second Level terms - demining

Second Level category	1	2	3	4	5	6	7	8	9	10	
	Number of overlapped pages in Second Level category										Total contributed pages
demining activity	1	1	1	2	2	2	3	2	2	1	86
demining assistance	2	1	1	2	1	2	1	2	1	2	88
demining effort	2	1	2	2	1	1	2	2	2	1	88
demining equipment	2	1	2	1	2	2	2	2	3	1	84
demining operation	2	3	2	1	2	2	3	3	2	2	81
demining program	3	4	3	3	3	4	2	5	4	1	71
demining project	1	1	1	2	1	3	5	2	2	1	85
demining team	3	2	1	1	2	3	5	4	1	2	79
demining technology	4	2	3	4	4	2	2	2	3	3	75
demining training	2	2	2	2	3	1	2	3	1	2	83
Total pages (out of 1000)											817

The same process in TABLE 3 was applied to the Second Level terms. The 10 Top Level terms are placed in the first left column in TABLE 4, the 10 Second Level terms of each Top Level term are placed in numeric order from 1 to 10 and placed on the top row. The figure in each cell represents the number

of overlapping URLs retrieved by the corresponding Second Level term. For example the Second Level terms in "demining activity" have contributed to 86 out of 100 pages in this category. The overall quantity of retrieved useful URLs from the Second Level terms are 817 out of 1000 pages, which is nearly 82% and is a 15% drop in usefulness from the Top Level terms.

The evaluation of "demining" is reapplied to "Alzheimer". The contributions from the Top Level terms of Alzheimer is 94% (TABLE 5), and is approximately 84% from the Second Level terms (TABLE 6). We believe the drop is a trade–off in reduced usefulness but increased quantity. The results also imply that generating the third level expansion terms would cause a certain ratio of decline on the retrieval usefulness. However, it would also increase the number of retrieved documents from 1000 to 10000. This could be an opportunity to retrieve more information if necessary.

TABLE 5 Contribution from Top Level terms - Alzheimer

Top Level expansion term	Number of overlapped pages	Total contributed pages
alzheimer disease	1	9
alzheimer society	1	9
alzheimer research	1	9
alzheimer patients	0	10
alzheimer asssociation	0	10
alzheimer drug	0	10
alzheimer dementia	2	8
alzheimer scotland	0	10
alzheimer sufferers	1	9
alzheimer forum	0	10
Average		94

TABLE 6 Contribution from Second Level terms - Alzheimer

Second Level category number	1	2	3	4	5	6	7	8	9	10	
	Number of overlapped pages in Second Level categories										Total contributed pages
alzheimer disease		1						1	1		97
alzheimer society					1	3		2	1	2	91
alzheimer research	1				1	1	7	2		1	87
alzheimer patients	1		2		3	1			2		91
alzheimer asssociation		2				1		4	5		88
alzheimer drug		3	6	2		3		4	1	1	80
alzheimer dementia			1	1	1	3		8	3	2	81
alzheimer scotland		4	1	5	3	8	8	1	2	2	66
alzheimer sufferers	2	2	1	1	6	2		4			82
alzheimer forum	1	1	1	4	8	1	3	2	7		72
Total pages (out of 1000)											835

In additon, using either “demining” or “Alzheimer” resulted in above 90 % on the Top Level and above 80% on the Second Level of usefulness. Also the figures shown in both evaluations are quite similar. This proves the WE algorithm is a generic algorithm for query expansion.

5. Conclusion

In this paper, we have presented our initial attempts to develop efficient information retrieval techniques to support decision–making and situation awareness in the domain of humanitarian demining. The algorithm we have developed is called Windmill Expansion and supports query expansion in situations where access to domain–specific knowledge is limited. In our example, the initial query terms “demining” and “Alzheimer” were used to produce 10 Top Level and 100 Second Level expansion terms. In order to evaluate the usefulness of the WE algorithm, we measured the overlap across the retrieved URLs. The results show that the Top Level terms retrieved above 90% of useful URLs, and Second Level terms retrieved above 80%. Although there is a decrease of 10–15% in usefulness from Top Level to Second Level, the trade–off is an increase in the total quantity of contributed pages.

The WE algorithm has been shown to support the extraction of domain–specific terminology, named entities and stakeholder information in both Demining and Health domains. The evaluation results also prove that WE is a generic algorithm for query expansion in information retrieval.

6. Future Work

The WE algorithm provides a mechanism for processing and retrieving web documents. In the future we would like to enhance the WE ability from the current term weighting approach to contain semantic knowledge so as to obtain better sub category document retrieval. Dictionaries, thesauri or ontologies could be employed for concept recognition. We also like to test the applicability of an enhanced version of WE in fields such as law, economy, health and so on.

7. Acknowledgment

This work is supported by the Data and Information Fusion Defence Technology Centre (DIF DTC), a consortium of academic and industrial partners headed by General Dynamics UK Ltd. SEMIOTIKS is a joint academic/industrial project involving the University of Southampton and QinetiQ. We would like to thank QinetiQ research staff, particularly Christopher J.M. Booth, for their comments on an earlier draft of this paper. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the UK Ministry of Defence, or the UK Government.

References

[1] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, pp. 964-971, 1987.

[2] E. N. Efthimiadis, "Query Expansion," *Annual Review of Information Systems and Technology (ARIST)*, vol. 31, pp. 121–187, 1996.

[3] I. Ruthven, "Re-examining the potential effectiveness of interactive query expansion," in *the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, 2003, pp. 213-220

[4] J. W. Cooper and R. J. Byrd, "Lexical navigation: Visually prompted query expansion and refinement.," in *Proceedings of the second ACM international conference on Digital libraries*, Philadelphia, Pennsylvania, United States, 1997, pp. 237 - 246.

[5] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer, "Automatic query wefinement using lexical affinities with maximal information gain," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland 2002, pp. 283 - 290

[6] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland 1994, pp. 61 - 69.

[7] M. Kim, A. H. Alsaffar, J. S. Deogun, and V. V. Raghavan, "On Modeling of Concept Based Retrieval in Generalized Vector Spaces," in *the 12th International Symposium on Foundations of Intelligent Systems, Lecture Notes In Computer Science; Vol. 1932*, 2000, pp. 453-462.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1999.

[9] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Transactions on Information Systems (TOIS)*, vol. 18, pp. 79 - 112 2000.

[10] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic query expansion using SMART: TREC3.," in *the 3rd Conference on Text Retrieval (TREC-3)*, Gaithersburg, MD, 1995, p. .

[11] C. Carpineto, G. Romano, and V. Giannini, "Improving retrieval feedback with multiple term-ranking function combination," *ACM Transactions on Information Systems (TOIS)*, vol. 20, pp. 259-290, 2002.

[12] M. Song, I. Y. Song, R. B. Allen, and Z. Obradovic, "Keyphrase extraction-based query expansion in digital libraries," in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, Chapel Hill, NC, USA 2006, pp. 202 - 209

[13] M. D. Gordon, "User-Based Document Clustering by Redescribing Subject Descriptions with a Genetic Algorithm," *Journal of the American Society for Information Science*, vol. 42, pp. 311-322, 1991.

[14] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130-137, 1980.

[15] M. F. Porter, *An algorithm for suffix stripping*: Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997.

[16] G. Salton and M. E. Lesk, "Computer Evaluation of Indexing and Text Processing," *Journal of the ACM (JACM)* vol. 15, pp. 8-36, 1968.