AN ALGORITHM FOR FINDING EFFECTIVE QUERY EXPANSIONS THROUGH FAILURE ANALYSIS OF WORD STATISTICAL INFORMATION RETRIEVAL

D. Wollersheim and W. Rahayu

Dept. of Computer Science and Computer Engineering
La Trobe University,
Address: Bundoora, Victoria 3086 Australia
Phone: +61 3 9479 1280 Fax: +61 3 9479 3060

ABSTRACT

Query expansion can improve word statistical based information retrieval, but problems occur due to the heterogeneous nature of underlying documents, and use of thesauri and ontologies not specifically designed for query expansion purposes. This paper performs a failure analysis on word statistical information retrieval, and uses this data to discover high value query expansions.

This process uses a medical thesaurus (UMLS) and a medical document test collection (OHSUMED). The UMLS based medical concepts within the OHSUMED documents and test queries are identified. Then, using the document-query relevance judgements, we explore the set of UMLS relationships that connect the concepts in the queries with the concepts in the relevant documents. From this, we increase the specificity of the query expansion process.

KEYWORDS

Query expansion, information retrieval, ontologies, UMLS, data mining

1. INTRODUCTION

Word-statistical information retrieval can be improved by using ontological connections to expand queries, but there is a problem in discovering which query expansions would be useful, because of the density of ontological interconnection, and the heterogeneous nature of the documents. It is necessary to prune the search space. This work outlines an algorithm to find the most useful query expansion directions, by using a document test collection to discover the ontological connections that exist between queries, and documents deemed as relevant to those queries.

Query expansion is the addition of related words to a query, with a purpose of improving information retrieval precision and recall. It 'widens the net' of a query. In ideal form, all relevant documents are retrieved, with a minimum loss of precision. The question is, which words to add, and therefore, which relationships to use.

In the example in figure 1, we have an input query *itch* and *hand*, and an ontology containing the relationship's *hand HAS-COMPONENT skin of hand*, *skin of hand IS-PART-OF skin*, and *itch IS-SYMPTOM-OF dermatitis*. This allows us to add the concepts "skin" and "dermatitis" to the original query, and retrieve documents containing this phrase.

The concepts and relationships that we use for query expansion (QE) come from the Unified Medical Language System (UMLS) [1, 2]. One component of UMLS is the Metathesaurus, a medical domain specific ontology. A key constituent of the Metathesaurus is a *concept*, which serves as nexus of terms across the different term sets. Concepts are interconnected by 8 categories of relationship, with varying degree of meaning. For example, the parent and child relationships imply subsumption, whereas the sibling (SIB) relationship only implies association.

There has been much interest in using UMLS to facilitate query expansion, with mixed results. Hersh et al found that query expansion degraded aggregate retrieval performance, but some specific instances of synonymy and hierarchy based QE improved individual query performance [3] . Other research has shown that UMLS based QE combined with retrieval feedback returned results significantly above baseline statistical retrieval [4, 5]. This interest is logical. UMLS is a rich resource. QE improves retrieval in other domains, using general purpose thesauri such as Wordnet ([6]).

First generation query expansion used a shotgun approach, expanding a query in every direction. This is not very successful because of the nature of the underlying data. Documents, written in natural language, are too heterogeneous for such a broad approach. The 'net' is cast too widely. In this vein, Srinivasan makes the case for using rough and fuzzy sets to decide query expansion direction, but the work is not evaluated [7, 8].

We need prune the search space. Instead of expanding the query in all directions, we will expand only in productive directions. We hypothesize that, given document set, a query, and a relevance relation joining the two, there will be a certain expansion of the query that will maximise its information retrieval effectiveness. This paper describes an algorithm for discovering the general characteristics of this maximally effective query expansion; by characterising the ontological path from concepts in the original query to those in the expanded query.

To do this, we need to know which documents are relevant to a query. Fortunately, this information exists in document test collections, which contain a set of documents, and queries, and expert derived relevance judgements connecting them. The specific collection we use is OHSUMED, developed by Hersh in 1994 [9]. While there are other medical test collections, we choose OHSUMED for several reasons. It is one of the largest available medical test collections, and because of the statistical nature of our algorithm, size is important. More importantly, because query expansion has been previously studied with OHSUMED [3], we will be able to judge the effectiveness of our algorithm.

In summary, this work will be done according to the following steps (see Figure 2 for more detail):

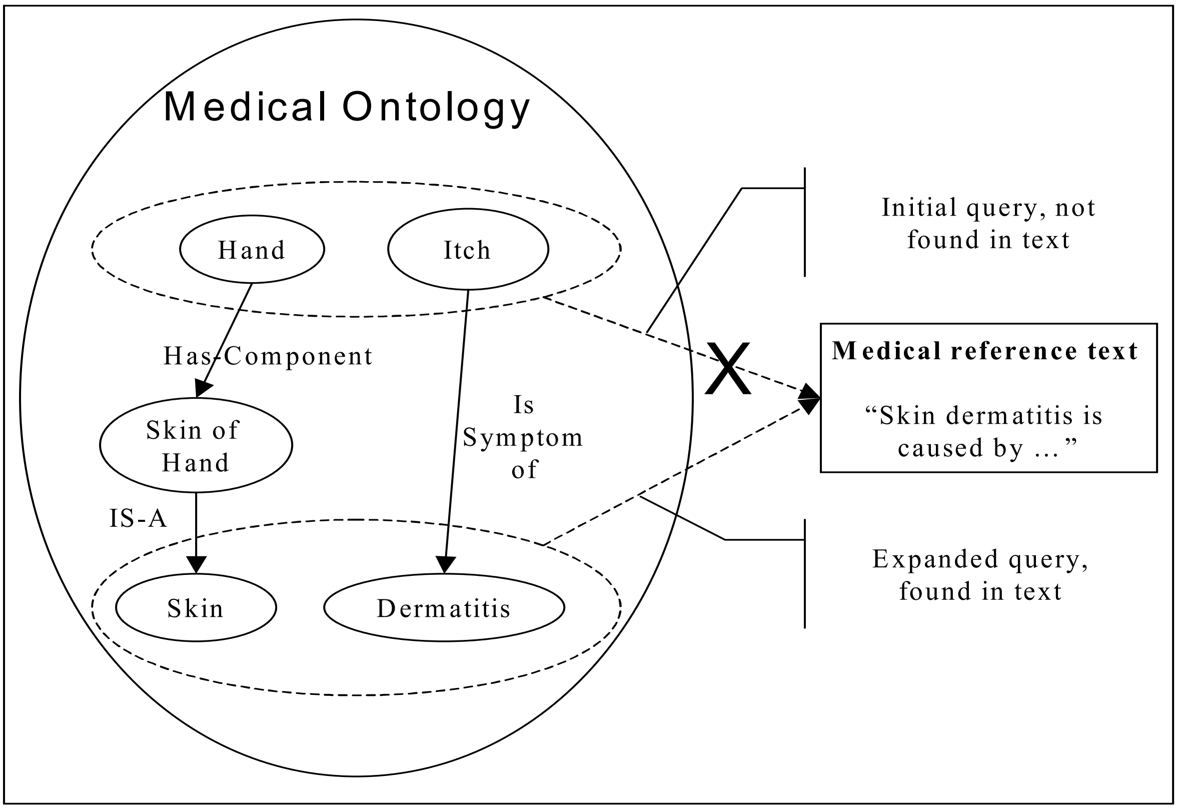

Figure 1: Example use of query expansion.

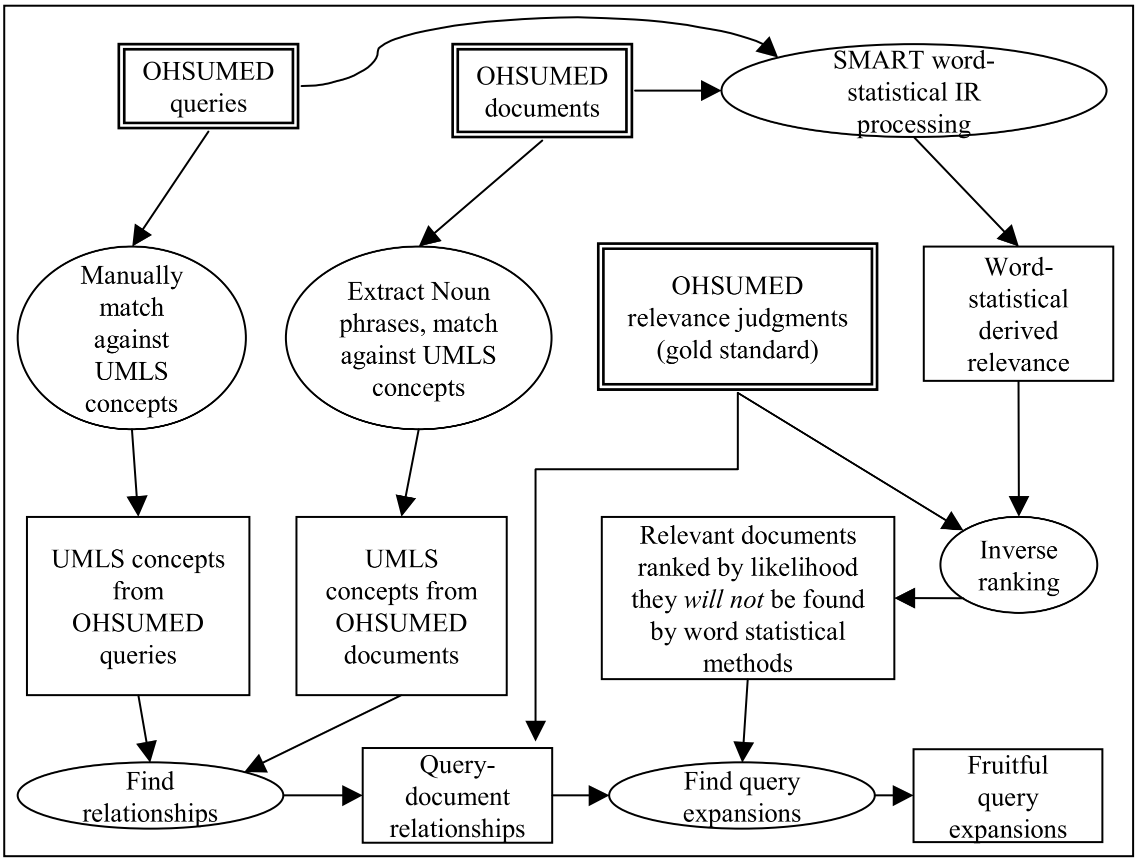
OHSUMED queries
OHSUMED documents
SMART word-statistical IR processing
Manually match against UMLS concepts
Extract Noun phrases, match against UMLS concepts
OHSUMED relevance judgments (gold standard)
Word-statistical derived relevance
UMLS concepts from OHSUMED queries
UMLS concepts from OHSUMED documents
Relevant documents ranked by likelihood they will not be found by word statistical methods
Inverse ranking
Find relationships
Query-document relationships
Find query expansions
Fruitful query expansions

Figure 2: Flowchart of document query matching methodology. Items in double lined boxes are the Ohsumed source documents, while those in single lined boxes are resources produced by the algorithm. Circles indicate processes.

1. using a concept dictionary, find the concepts represented in both the queries, and in the documents,
2. using an ontology, discover the closest relationships between the concepts from the queries and their relevant documents,
3. calculate the ranking of relevant documents according to word-statistical retrieval methods,
4. analyse relationships in 2, weighting the results inversely proportional to retrievability ranking found in 3, and
5. use information from 5 to suggest query expansion strategies.

We preferentially weight the documents not well retrieved by word-statistical methods because this will concentrate the relationship characteristics we are seeking. We want to improve the retrievability of documents not well retrieved by traditional methods. In effect, we perform a failure analysis of traditional IR.

Query expansion holds much promise. Information is becoming more explicitly interconnected. Ontological resources are increasing in number and complexity; for example, Wordnet, CYC, UMLS [10] [11]. Query expansion provides another use for these resources, and a way to test them against real world information needs. It is a simple, discrete, easily evaluated task.

This work in particular is opportune, and unique. OHSUMED is a well analysed medical IR test collection. Recent advances in automated indexing allow us to extract concepts at reasonable accuracy. The existence and maturity of UMLS, a resource of concepts and relationships, operationalise the semantic space around a query. Additionally, there are many test collections, and ontologies. This framework can be used to similarly analysed, compared, evaluated as to their usefulness for query expansion, and exploited.

Not only does this work advance the field of query expansion. Our input is a gold standard medical domain query collection, containing expert judgements about query-document relevance. By analysing this, we get insight into the questions, "How do humans join up queries and documents?", and "What is relevance?"

2. METHOD

This section describes in detail our methodology. First, we assign concepts to queries. This is done by hand, via a process similar to that of [3]. We repeatedly assign the most specific concept that is possible for the largest set of words in a query. For example, we assign the concept C0012739 to the phrase "disseminated intravascular coagulation". If that concept did not exist, we would assign the more general concepts of "intravascular" and "coagulation". After a concept is assigned, the process is repeated on the remaining query words.

The algorithm for extracting UMLS concepts from the OHSUMED documents is adapted from [12]. It is similar to the query classification process, but necessarily more automated due to the magnitude of text. We break the document fields *title*, *keywords*, and *abstract* into noun phrases using the Brill part of speech tagger [13]. The noun phrases are then normalised using the UMLS provided lexical variant generator, using the options: l, t, g, p, B and w. This process removes stop words and punctuation, and stems the resulting words, putting them into a canonical form that is ready for lookup in the UMLS database table MRXNS_ENG (This table contains the English language normalised forms of all UMLS conceptual variations). All combinations of words from the normalised phrases are matched against the preferred forms of the UMLS terms, favouring matches of maximum length consecutive words. Figures 3-6 shows the process of mapping a phrase to a set of concepts.

Using the SMART v 11.0 IR system (available at ftp.cs.cornell.edu/pub/smart), we find the document-query word-statistical retrieval rankings. We use the SMART relevance weighting found to be most effective with the OHSUMED collection [9].

Finally, we determine the closest connection between each concept from the queries and each relevant documents. We find all the expansions of each extracted query concept, within a certain semantic distance, that will connect that query term to every relevant document. Distance is calculated according the sequence of relations that connect the query terms and the document terms. We give ancestor/descendent relationships are distance of 1, and sibling relationships were excluded. We permute concepts to a maximum distance of 2 units. The relationship distance calculation is outlined in detail in [14].

3. RESULTS

The OHSUMED collection and the UMLS database were imported into Oracle 9i database. Where not otherwise stated, data manipulation was done with the Perl programming language.

The OHSUMED collection consists of 355013 documents. The test collection contains 106 queries. Of these queries, 2394 documents are judged relevant to them. This gives an average of 23.3 relevant documents per query, and a range of 0 to 116 relevant documents per query. From the relevant documents, we extracted 190,619 UMLS concepts, while there were 634 concepts assigned against the queries. The ratio of an average of 80 concepts per document compared to 6 per query, is consistent with the relative size of the queries compared to documents, and also the more precise method of assigning concepts to queries.

There were 13339 possible connections from a query concept to a distinct relevant document. Of these, 5680 relationships were discovered. This means that 42.6% of the concepts that were extracted from the queries found matches in their relevant documents.

From the set of matched concepts, there was an even spread over the 3 distances: 32.3% of matches were at distance 0 (meaning that these concepts had exact matches in the documents). 23.0% of query concepts were matched at distance 1, i.e. one relationship away from the document, while 44.5% were found at distance 2. Of the relationships at distance 1, 62% of the matches were parent type links (either parent or relation-broader), while the remaining matches were child type (either child or relation-narrower) links. The matches of distance 1 between query concept and document concepts having the top 10 number of occurrences can be seen in table 1.

Matches of distance 2 are not so straightforward. They tend to include relationship paths consisting of the more general UMLS concepts, the overarching categories. For example, the 'joining' categories with the top 3 number of occurrences are "Thesaurus of Psychological Index Terms". "general adjectival modifiers", and "Attribute". These relationships are to broad to be of much semantic value. To solve this problem, we must discount such terms. This could be done on a term frequency basis, weighting terms higher if they have fewer connections. This weighting can provide feedback to the concept expansion algorithm, where we can spend more energy exploring the higher value links.

One problem with this algorithm is that we treat document concepts as boolean values, ignoring the frequency of words within a document. An example of this is that, of the 1839 documents that were classified as having the same UMLS concepts as their relevant queries (semantic distance = 0), only 36% of them were ranked as being relevant by the SMART retrieval system. This implies that, even though the relevant words were present in the document, they were either 1) not frequent enough in the document, or 2) too frequent in the collection, to ensure that the document was deemed relevant. This could be solved by looking at term frequency when judging the value of a found relationship. It is not useful to add terms to a document which will not increase relevance.

It will be interesting to perform this process on a domain outside medicine. An example would be to use the

Some patients converted from ventricular fibrillation to organized rhythms by defibrillation-trained ambulance technicians (EMT-Ds) will refibrillate before hospital arrival.

Figure 3. Example phrase from the abstract of an Ohsumed document.

▪some patients
▪from ventricular fibrillation to
▪rhythms by defibrillation
▪ambulance technicians emt ds will
▪before hospital arrival

Figure 4. Noun phrases extracted from input

•some patient
•from ventricular fibrillation
•rhythm defibrillation
•ambulance technician emt d will
•before hospital arrival

Figure 5. Noun phrases normalised with UMLS provided program lvg, using options l, t, g, p, B and w.

C0205392	some
C0030705	patient
C0042510	fibrillation ventricular
C0871269	rhythm
C0002422	ambulance
C0043177	will
C0600109	will
C0019994	hospital

Figure 6. Concepts identifiers and matching phrases from among the preferred terms of UMLS concepts

Query Concept	Matching Document Concept	# OF Occurences	Relat-ionship
Primary carcinoma of the liver cells	Liver neoplasms	69	RB
Lupus Erythematosus	Lupus Nephritis	58	RN
Differential Diagnosis	Diagnosis	44	PAR
Lupus Coagulation Inhibitor	Antibodies, Antiphospholipid	41	PAR
Outpatients	Patients	40	PAR
hypothermia, natural	Body Temperature	37	PAR
hypothermia, induced	Body Temperature	35	RB
Mass in breast	Mass, NOS	29	RB
Malignant neoplasm of breast	Breast	29	RB
Anorexia	Anorexia Nervosa	23	RN

Table 1: Successful expansions between query concepts and relevant document concepts with only one relationship between query concept and document concept. RB=Relation Broader, RN=Relation Narrower, PAR=Parent

thesaurus Wordnet in combination with general purpose test collection. Another extension of this work would be to feed the results back into a document retrieval system, evaluating the performance of query expansions found.

4. CONCLUSION

This paper describes a methodology for using a combination of document test collections and ontologies to find maximum value query expansions. This is done by identifying the ontological concepts from the test collection's documents, and from the queries. The relationships between the resulting concepts are explored, and used to suggest directions for query expansion.

REFERENCES

1. Lindberg, D.A.B., B.L. Humphreys, and A.T. McCray, *The Unified Medical Language.* System. Meth. Inform. Med., 1993. **32**: p. 281-291.

2. Campbell, K.E., et al., *Representing thoughts, words, and things in the UMLS.* Journal of the American Medical Informatics Association, 1998. **5**(5): p. 421-31.

3. Hersh, W., S. Price, and L. Donohoe, *Assessing thesaurus-based query expansion using the UMLS metathesaurus.* Journal of Americian Medical Informatics Association, 2000. **Suppl. S 2000**: p. 344-348.

4. Aronson, A.R. and T.C. Rindflesch, *Query expansion using the UMLS(R) Metathesaurus(R).* 1997.

5. Srinivasan, P., *Query expansion and MEDLINE.* Information Processing & Management, 1996. **32**(4): p. 431-43.

6. Mandala, R., et al. *Ad hoc retrieval experiments using WordNet and automatically constructed thesauri.* in *Seventh Text REtrieval Conference (TREC-7).* 1999. Gaithersburg, MD, USA.: Nat. Inst. Standards & Technol.

7. Srinivasan, P., *Exploring the UMLS: A Rough Sets Based Theoretical Framework.* Proceedings / AMIA Annual Symposium, 1999.

8. Srinivasan, P., et al., *Vocabulary mining for information retrieval: rough sets and fuzzy sets.* Information Processing & Management, 2001. **37**(1): p. 15-38.

9. Hersh, W.R., et al., *A performance and failure analysis of saphire with a medline test collection.* Journal of the American Medical Informatics Association, 1994. **1**(1): p. 51-60.

10. Lenat, D.B., *Cyc: A Large-Scale Investment in Knowledge Infrastructure.* Communications of the ACM, 1995. **38**(11).

11. McCray, A.T., et al., *UMLS Knowledge for Biomedical Language Processing.* Bulletin of the Medical Library Association, 1993. **81**(2): p. 184-194.

12. Nadkarni, P., R. Chen, and C. Brandt, *UMLS Concept Indexing for Production Databases: A Feasability Study.* Journal of American Medical Informatics Association, 2001. **8**: p. 80-91.

13. Brill, E., *Some Advances In Rule-Based Part of Speech Tagging.* AAAI, 1994.

14. Wollersheim, D. *Methodology for Retrieval of Medical Guideline Fragments using Domain Ontology Concept based Query Expansion.* in *ISCIS XVIII - Eighteenth International Symposium on Computer and Information Sciences.* 2003 (Submitted). Antalya, Turkey.