Free-text Medical Document Retrieval Via Phrase-based Vector Space Model*

Wenlei Mao, MS and Wesley W. Chu, PhD
Computer Science Department
University of California, Los Angeles

Many information retrieval systems are based on vector space model (VSM) that represents a document as a vector of index terms. Concepts have been proposed to replace word stems as the index terms to improve retrieval accuracy. However, past research revealed that such systems did not outperform the traditional stem-based systems. Incorporating conceptual similarity derived from knowledge sources should have the potential to improve retrieval accuracy. Yet the incompleteness of the knowledge source precludes significant improvement. To remedy this problem, we propose to represent documents using phrases. A phrase consists of multiple concepts and word stems. The similarity between two phrases is jointly determined by their conceptual similarity and their common word stems. The document similarity can in turn be derived from phrase similarities. Using OHSUMED as a test collection and UMLS as the knowledge source, our experiment results reveal that phrase-based VSM yields a 16% increase of retrieval accuracy compared to the stem-based model.

INTRODUCTION

There is an increasing trend of storing medical documents and literature as free text in digital information systems. Example documents include patient reports in hospital information systems and radiology information systems, medical literatures in medical digital libraries and general medical information on the World Wide Web. As a result, the accurate retrieval of these documents becomes increasingly important.

Indexing is used to facilitate the retrieval of such documents. VSM[1] is widely used to index documents. Under VSM, a document is represented by a vector of terms (*document vector*). The cosine of the angle between two document vectors indicates the similarity between the corresponding documents. A smaller angle corresponds to a larger cosine value and indicates higher document similarity. A query, which describes the information need, is encoded as a vector as well. Retrieval of documents that satisfy the information need is achieved by finding the documents most similar to the query, or equivalently, the document vectors closest to the query vector.

Word stems are widely used as index terms. To improve retrieval accuracy, it is natural to replace word stems with concepts. However, previous research showed not only no improvements, but degradation in retrieval accuracy when concepts were used in document retrieval[2,3,4,5] except when documents were very short[6]. Replacing word stems with multiple word combinations was also studied[7].

In the following sections, we first propose to use phrases instead of word stems as index terms. A *phrase* is a string of words used to represent concepts. Since some concepts are related, we use *conceptual similarity* to describe such relationships. Higher conceptual similarity indicates stronger relationships. Then we introduce phrase-based document similarity measure that takes into account both the concepts represented by and the word stems used in the phrases. Finally, we use OHSUMED test collection to demonstrate that phrase-based VSM yields higher retrieval accuracy than stem-based VSM does.

To facilitate discussion, we shall use the following sample query in this paper: "22 year old with hyperthermia, leukocytosis, increased intracranial pressure, and central herniation. Cerebral edema secondary to infection, diagnosis and treatment." The first part of the query is a brief description of the patient; the second part is the information need.

VECTOR SPACE MODELS

Stem-based VSM

We represent a document as a vector of terms in VSM. The basis of the vector space corresponds to distinct terms in a document collection. Components of the document vector are the weights of the corresponding terms that represent their relative importance in the document. In a naïvest approach, we could treat a word as a term. Yet, morphological variants like "edema" and "edemas" are so closely related that they are usually conflated into a single *word stem*, e.g., "edem," by stemming[8]. Our sample query thus consists of word stems "hypertherm," "leukocytos," "increas," "intracran," "pressur," etc.

Word stems are usually treated as notational,

* This research is supported in part by NIC/NIH Grant #4442511-33780.

rather than conceptual entities. Two word stems are considered unrelated if they are different. For example, the stem of "hyperthermia" and that of "fever" are usually considered unrelated despite their apparent relationship.

In stem-based VSM, word stems constitute the basis of the vector space. The base vectors are orthogonal to each other because different word stems are considered unrelated.

The weight $w^s_{\alpha,u}$ of a word stem u in a document α is determined by the number of times u appears in α (known as the *term frequency*) and the number of documents that contain u (known as the *document frequency*) following the TF-IDF (term frequency, inverse document frequency) scheme[1]. In essence, the more often u appears in α, the more important u is in α. On the other hand, the more documents u belongs to, the less disambiguating power it has, and thus the less important it is.

Concept-based VSM

Using word stems to represent documents results in the inappropriate fragmentation of concepts such as "increased intracranial pressure" into its component stems "increas," "intracran," and "pressur." Clearly, using *concepts* instead of single words or word stems as the vector space basis should produce a VSM that better mimics the human thought processes, and therefore should result in more accurate retrieval.

However, using concepts is more complex than using word stems. First, concepts are usually represented by multi-word phrases such as "increased intracranial pressure." Second, there exist synonymous and polysemous phrases. Two phrases sharing a concept are *synonymous*, and phrases that could represent more than one concept are *polysemous*[9]. For example, "hyperthermia" and "fever" are synonymous because they share the same concept "an abnormal elevation of the body temperature." At the same time, "hyperthermia" is polysemous, because in addition to the above meaning, it also means "a treatment in which body tissue is exposed to high temperature to damage and kill cancer cells." Synonyms can be identified with the help of a dictionary or a thesaurus. Determining which concept a polysemous phrase represents is known as *word sense disambiguation* (WSD)[10]. Third, some concepts are related to one another. Hypernym and hyponym relations are important conceptual relations. If we say "an x is a (kind of) y" then concept x is a hyponym of concept y, and y is a hypernym of x[9]. "Hyperthermia" is a hyponym of "high body temperature;" and "high body temperature" is a hypernym of "hyperthermia."

Concept identifiers are usually used to identify concepts. Using UMLS, our sample query becomes (15967, 203597), (23518), and (151740) etc., representing "hyperthermia," "leukocytosis," and "increased intracranial pressure," etc., respectively.

In concept-based VSM, the basis of the vector space consists of distinct concepts. To model the relationship of such concepts as "hyperthermia" and "elevated body temperature" we remove the orthogonality constraint on base vectors. Base vectors for two related concepts form an acute angle. Only when we cannot find any reasonable relations between two concepts that we treat their corresponding vectors as orthogonal. The cosine of the angle between two concept vectors is defined as the *conceptual similarity* between the corresponding concepts. The conceptual similarity thus ranges from 0 to 1 with 0 indicating unrelated and 1 indicating synonymous concepts.

To study the effects of conceptual similarities, we shall compare two cases. In one, we assume all different concepts are unrelated. Therefore, all base vectors of the vector spaces are orthogonal to one another. In the other, we derive conceptual similarities from knowledge sources. The resulting base vectors are no longer mutually orthogonal.

We derive the weight w^c_{α,x_i} of the i^{th} concept x_i in a document α using a slightly modified version of TF-IDF scheme. The more often a concept x_i appear in the document α, the more important x_i is in α, and the more documents x_i appears in, the less important it is. Furthermore, higher weights are assigned to longer phrases that correspond to more specific concepts. For example, if the term frequencies and document frequencies for "increased intracranial pressure" and "hyperthermia" were identical, the former would obtain a higher weight than the latter.

Phrase-based VSM

Conceptual similarities needed in concept-based VSM are derived from knowledge sources. The quality of such VSM therefore depends heavily on the quality of the knowledge sources. The missing of certain conceptual relations in the knowledge sources potentially degrades retrieval accuracy. For example, treating "cerebral edema" and "cerebral lesion" as unrelated is potentially harmful. Noticing the common component word "cerebral" in the above phrases, we propose phrase-based VSM to remedy the incompleteness of the knowledge sources.

In phrase-based VSM, a document is represented as a set of phrases. Each phrase may correspond to multiple concepts (due to polysemy) and consist of several word stems. Our sample query now becomes [(15967, 203597), ("hypertherm")], [(23518),

("leukocytos")] and [(151740), ("increas", "intracran", "pressur")] etc.

Following the TF-IDF schemes in stem-based and concept-based VSMs, we can derive the stem weight $w^s_{\alpha,u_{i,k}}$ of the k^{th} stem $u_{i,k}$ and the concept weight $w^c_{\alpha,x_{i,m}}$ of the m^{th} concept $x_{i,m}$ in phrase i of α.

Similar to concept-based VSM, we study two cases. In one, different concepts are treated as unrelated; in the other, concepts may be related. In both cases, distinct word stems are assumed to be unrelated.

DOCUMENT SIMILARITY

The similarity of two documents α and β is the cosine of the angle between their corresponding document vectors $\vec{\alpha}$ and $\vec{\beta}$,

$$sim(\alpha,\beta)=\cos(\vec{\alpha},\vec{\beta})=\frac{\vec{\alpha}\bullet\vec{\beta}}{\sqrt{\vec{\alpha}\bullet\vec{\alpha}}\sqrt{\vec{\beta}\bullet\vec{\beta}}}$$

We shall extend the document vector dot product $\vec{\alpha}\bullet\vec{\beta}$ and denote the *extended dot product* (EDP) as $\vec{\alpha}\circ\vec{\beta}$. Using the EDP in place of the dot product, we derive document similarity as,

$$sim(\alpha,\beta)=\frac{\vec{\alpha}\circ\vec{\beta}}{\sqrt{\vec{\alpha}\circ\vec{\alpha}}\sqrt{\vec{\beta}\circ\vec{\beta}}} \quad (1)$$

EDP Derivation

To derive the EDP in the phrase-based VSM, we first consider concepts without polysemy.

$$\vec{\alpha}\circ\vec{\beta}=\sum_{i,j}S^c_{i,j}$$

where $S^c_{i,j}$ is the conceptual contribution of phrase i in α and phrase j in β to the EDP. Assuming that each phrase represents a single concept, we have,

$$S^c_{i,j}=w^c_{\alpha,x_i}w^c_{\beta,y_j}s(x_i,y_j) \quad (2)$$

where $s(x_i, y_j)$ is the conceptual similarity between the i^{th} concept x_i in α and the j^{th} concept y_j in β. When different concepts are treated as unrelated, $s(x,y)$ in (2) is reduced to the Kronecker delta function,

$$\delta(x,y)=\begin{cases}1 & x=y\\ 0 & x\neq y\end{cases}$$

When concepts may be related, we derive conceptual similarities from knowledge sources.

In order to use (2) in the presence of polysemy, we need to disambiguate senses. To avoid WSD cost, we use the most popular concept that a phrase represents as the phrase's meaning. Alternatively, we derive the conceptual contribution to the similarity between two phrases using an aggregation of (2) over all possible concept pairs. Each pair consists of one concept from each phrase.

The contribution of word stems to the EDP is the sum of the weight products for those word stems common to both phrases,

$$S^s_{i,j}=\sum_{k,l}w^s_{\alpha,u_{i,k}}w^s_{\beta,v_{j,l}}\delta(u_{i,k},v_{j,l}) \quad (3)$$

where $u_{i,k}$ and $v_{j,l}$ are the k^{th} word stem in phrase i in α and l^{th} word stem in phrase j in β respectively.

Given the contribution of concepts (2) and stems (3), we select the larger of the two as the contribution of phrase i in α and phrase j in β to the EDP, and get,

$$\vec{\alpha}\circ\vec{\beta}=\sum_{i,j}\max(S^c_{i,j},S^s_{i,j}) \quad (4)$$

Such selection remedies the incompleteness of the knowledge sources. $\vec{\alpha}\circ\vec{\alpha}$ and $\vec{\beta}\circ\vec{\beta}$ can be derived similar to (4). The document similarity can then be computed from (1) using these EDPs.

Conceptual Similarity in Hypernym Hierarchy

Given a hypernym hierarchy, the conceptual similarity $s(x, y)$ between concepts x and y depends on both the distance between them in the hierarchy and their generality. When two concepts are farther apart in the hypernym hierarchy, they are less similar – a concept is less similar to its grandparent than to its parent in the hypernym hierarchy. Thus we define the conceptual similarity to be inversely proportional to the number of "hops" between x and y, $d(x,y)$. The generality of a concept x can be derived from the number of its descendants $D(x)$. The more descendants x has, the more general it is. A general concept like "disease" has much more descendants than a more specific concept like "hyperthermia" has. Because of the exponential growth of the number of descendants when a concept moves up a tree structure, we take the logarithm of the number of descendant in conceptual similarity calculation. The conceptual similarity is therefore defined to be inversely proportional to the logarithm of the number of descendants of the two. A final consideration is the boundary case when we reach the leaves of the hypernym tree. Let us assume we have two concepts x_o and y_o, where x_o is the only direct hypernym of y_o, y_o is the only hyponym of x_o, and y_o has no hyponym of its own. Concepts x_o and y_o are so much alike that we define the conceptual similarity between them to be c close to 1, say 0.9, to represent such closeness. As a result, the conceptual similarity between concepts x and y is,

$$s(x,y)=\frac{c}{d(x,y)\log_2(1+D(x)+D(y))} \quad (5)$$

METHODS

The Test Collection, OHSUMED

OHSUMED[11] is a large test collection used in many information retrieval system evaluations. The test set consists of a reference collection, a query collection, and a set of relevance judgments.

The reference collection is a subset of the MEDLINE database. Each reference contains a title, an optional abstract, a set of MeSH headings, author information, publication type, source, a MEDLINE identifier, and a sequence identifier. The query collection consists of 106 queries. Each query contains a patient description, an information request, and a sequence identifier. The sample query we use in this paper is query 57 in the collection. 14,430 references out of the 348K are judged by human experts to be not relevant, possibly relevant, or definitely relevant to each query.

We use the title, the abstract, and the MeSH headings to represent each document; and the patient description, and the information request to represent each query.

The Knowledge Source, UMLS

UMLS[12] is a medical lexical knowledge source and a set of associated lexical programs. The knowledge source consists of UMLS Metathesaurus, SPECIALIST lexicon, and UMLS semantic network. Especially of interest to us is its central vocabulary component – the Metathesaurus. It contains biomedical phrases from more than 60 vocabularies and classifications. The Metathesaurus contains 1.6M phrases representing over 800K concepts.

A concept unique identifier (CUI) identifies each concept. UMLS tends to assign a smaller CUI to a more popular sense of a phrase. For example, the CUI for the "high body temperature" sense of "hyperthermia" is 15967, while the CUI for its "treatment" sense is 203597. Therefore, we use the concept with the smallest CUI in conceptual contribution calculation. Our experimental results show that this heuristic produces retrieval accuracy comparable to that produced by the aggregation approach.

The Metathesaurus encodes many conceptual relations. We concentrate on hypernym relations. Two relations in UMLS roughly correspond to the hypernym relations: the RB (border than) and the PAR (parent) relations. For example, "hyperthermia" has a parent concept "body temperature change." We combine the 838K RB and 607K PAR relations into a single hypernym hierarchy.

Hypernymy is transitive[13]. For example, "sign and symptom" is a hypernym of "body temperature change" and "body temperature change" is a hypernym of "hyperthermia," so "sign and symptom" is also a hypernym of "hyperthermia." However UMLS Metathesaurus encodes only the direct hypernym relations but not the transitive closure. We derive the transitive closure of the hypernym relation and use (5) to calculate the conceptual similarities.

Phrase Detection

We adopt the Aho-Corasick algorithm[14] for the set-matching problem to detect each occurrences in a set of phrases (1.3M phrases in UMLS) in a set of documents (106 queries and 14K judged documents of OSHUMED):

First, Aho-Corasick algorithm detects *all* occurrences of any phrase in a document. But we only keep the longest, most specific phrase. For example, although both "edema" and "cerebral edema" are detected in the sample query, we keep only the latter and ignore the former.

Second, to detect multi-word phrases, we match stems instead of words in a document with UMLS phrases. We use Lovins stemmer[8] to derive word stems. To avoid conflating different abbreviations into a single stem, we define the stem for a word shorter than four characters to be the original word.

Third, stop-word removal is performed *after* the multi-word phrase detection. In this way, we correctly detect "secondary to" and "infection" from "cerebral edema secondary to infection." We would incorrectly detect "secondary infection" if the stop-words ("to" in this case) were removed before the phrase detection.

Retrieval Accuracy Measurement

To calculate retrieval accuracy using precision-recall[1], we combined the "possibly relevant" and "definitely relevant" judgments in OHSUMED into a single relevant category. Based on the type of VSM, we calculate the document similarity between each of the 14K documents and each of the 105 queries (one query does not have relevant document). For a given VSM and a query, we rank the documents from the most to the least similar to the query. When a certain number of documents are retrieved, *precision* is the percentage of retrieved documents that are relevant; and *recall* is the percentage of the relevant documents that has been retrieved so far. We evaluate the retrieval accuracy by interpolating the precision values at eleven recall points. The overall effectiveness of different VSM is then compared by averaging over the performance of all the 105 queries (Figure 1). The average of the eleven precision values gives an overview of the effectiveness of each VSM[1].

Results

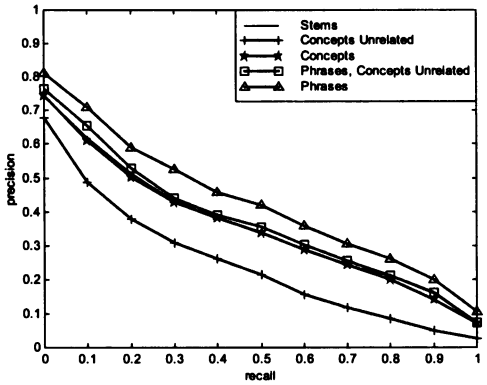

Figure 1. Comparison of the average precision-recall over 105 queries.

1. The baseline (Stems) uses stem-based VSM. Its 11-point average precision is 0.363.
2. Considering the contribution of concepts only, and treating different concepts as unrelated (Concepts Unrelated), we arrive at an 11-point average precision of 0.260, which is a 28% decrease from the baseline.
3. Similar to 2, but taking the concept inter-relationship into consideration (Concepts), we achieve a significant improvement over 2. The average accuracy is similar to that of the baseline.
4. Considering contribution of both concepts and word stems in a phrase, but treating different concepts as unrelated (Phrases, Concepts Unrelated), we arrive at an 11-point average precision of 0.375, a 3% improvement over the baseline.
5. Similar to 4, but taking concept interrelations into consideration (Phrases), we achieve an 11-point average precision of 0.420, which is a significant 16% improvement over the baseline.

Our experiment results reveal that viewing documents as concepts only and treating different concepts as unrelated can cause the retrieval accuracy to deteriorate (case 2). Considering concept inter-relations (case 3) or relating different phrases by their shared word stems (case 4) can both improve retrieval accuracy. The extended dot product combines contributions from the concepts and word stems. The phrase-based VSM utilizes such extended dot product and yields significant improvement in retrieval accuracy.

CONCLUSION

We developed a new vector space model that uses phrases to represent documents. Each phrase consists of multiple concepts and words. Similarity between two phrases is jointly determined by the conceptual similarity and their common word stems. We studied the phrase-based VSM using OHSUMED as the test set and UMLS as the knowledge source. Our experiments show that stem-based VSM performs better than concept-based VSM when different concepts are considered unrelated. When inter-relations between concepts are considered, concept-based VSM yields retrieval accuracy comparable to that of stem-based VSM. Phrase-based VSM yields a 16% increase in the 11-point average retrieval accuracy over the stem-based VSM. This is because in phrase-based VSM, word stems common to phrases can compensate for the inaccuracy in conceptual similarities derived from incomplete knowledge sources.

REFERENCES

[1] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*, 1983

[2] M. Mitra, C. Buckley, A. Singhal and C. Cardie. An Analysis of Statistical and Syntactic Phrases. In *Proc. RIAO97*, 200-214, 1997

[3] R. Richardson and A.F. Smeaton. Using WordNet in a Knowledge-based Approach to Information Retrieval. In *Proc. 17th BCS-IRSG*, 1995

[4] M. Sussna. Text Retrieval using Inference in Semantic Matanetworks. *PhD Thesis*, University of California, San Diego, 1997

[5] E.M. Voorhees. Using WordNet to Disambiguate Word Sense for Text Retrieval. In *Proc. 16th ACM-SIGIR.*, 171-180, 1993

[6] A.F. Smeaton and I. Quigley. Experiments on using Semantic Distances Between Words in Image Caption Retrieval. In *19th Proc. ACM-SIGIR*, 174-180, 1996

[7] D. Johnson, W.W. Chu, J.D. Dionisio, R.K. Taira and H. Kangarloo. Creating and Indexing Teaching Files from Free-text Patient Reports. In *AMIA '99*, 1999

[8] J.B. Lovins. Development of a Stemming Algorithm. In *Mechanical Translation and Computational Linguistics*, 11(1-2), 11-31, 1968

[9] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. Introduction to WordNet: an On-line Lexical Database. In *WordNet: an Electronic Lexical Database*, 1-19, 1998

[10] N. Ide and J. Véronis. Word Sense Disambiguation: the State of the Art. In *Computational Linguistics*, 24(1), 1-40, 1998

[11] W. Hersh, C. Buckley, T.J. Leone and D. Hickam. OHSUMED: an Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proc. 22nd ACM-SIGIR Conf.*, 191-197, 1994

[12] National Library of Medicine. *UMLS Knowledge Sources, 12th edition*, 2001

[13] J. Lyons. *Semantics*, 1977

[14] A.V. Aho and M.J. Corasick. Efficient String Matching: an Aid to Bibliographic Search. In *CACM*, 18(6), 330-340, 1975