Exploiting Semantic Proximity for Information Retrieval

Sanjeet Khaitan*
IBM India Research Lab
Block-I, IIT Delhi, Hauz Khas,
New Delhi-110016 INDIA
Email: skhaitan@in.ibm.com

Kamaljeet Verma
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: kamal@cse.iitb.ac.in

Rajat Kumar Mohanty
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: rkm@cse.iitb.ac.in

Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: pb@cse.iitb.ac.in

* The work was done when the author was at IIT Bombay

Abstract

In this paper, we propose a method which exploits the semantic proximity of words in unrestricted natural language text to retrieve relevant documents. In order to facilitate this functionality, the system represents the documents and the query in the form of semantically relatable sets (SRS), which are a group of entities demanding semantic relations when the semantic representation of the sentence is ultimately produced. We also devise a method to augment the SRSs to further boost the performance. WordNet is used to deal with different forms of divergence between the query and the documents. In a series of experiments on TREC data, our semantic proximity based retrieval technique yields high precision with improved mean-average-precision in comparison to conventional retrieval techniques.

1 Introduction

Information retrieval is an important application area of natural language processing where one encounters the challenge of processing unrestricted natural language text. However, the real challenge is to understand and represent appropriately the content of a document and query so that the relevance decision can be made effectively. It is now understood that injecting semantics is the key to improve the performance of search engines [Croft, 1995].

In this paper we describe a search strategy which is based on semantically relatable sets (SRS) [Mohanty et al., 2005] where words are in semantic proximity. The strategy exploits the semantics of the queries and the information by representing both the query and the documents in SRS form. Since both the query and the search base are in a semantic net like structure, we get results with high precision and improved Mean Average Precision (MAP) compared to conventional information retrieval techniques.

The rest of the paper proceeds as follows. In Section 2, the work related to this project is outlined. Section 3 describes the basic approach of our work. Experimental setup is detailed in Section 4 and Section 5 presents the initial results in comparison with the conventional retrieval technique *tfidf*. Section 6 describes the various pitfalls and Section 7 illustrates the enhancements applied. Implementation details are given in Section 8, which is followed by the final results in Section 9. We conclude in Section 10 with some further discussion of this approach and directions for future research.

2 Related Work

This section gives a brief review of work related to this project. This allows us to put our model in perspective.

[Corley, 2005] presents a knowledge-based method for measuring the semantic similarity of texts. A method is introduced that combines word-to-word similarity metrics into a text-to-text metric.

[Guha et al., 2003] present an application called *Semantic Search* which is built on the technologies including Web services and Semantic Web, which are creating a web of machine understandable data. They provide an overview of TAP, the application framework upon which the Semantic Search is built. They describe two implemented Semantic Search systems which, based on the denotation of the search query, augment traditional search results with relevant data aggregated from distributed sources.

[Mayfield and Finn, 2003] describe an approach to retrieval of documents that containing both free text and semantically enriched markup. They present a prototype of a framework in which documents and queries can be

marked up with statements in the DAML+OIL semantic web language. These statements provide them with both structured and semi-structured information about the documents and their content.

[Guarino et al., 1999] use linguistic ontology for content matching in information retrieval. Their approach applies only to the field of the search to a relevant class of information repositories-online yellow pages and product catalogs.

[Evans and Zhai, 1996] report on the application of a few noun-phrase analysis techniques to create indexing phrases for information retrieval. They describe a hybrid approach to the extraction of meaningful (continuous or discontinuous) sub-compounds from complex noun phrases using both corpus statistics and linguistic heuristics.

3 Search using Semantically Relatable Sets

The novelty of our approach is in representing the query and the documents in a form which captures the meaning contained in them. Universal Networking Language (UNL) [Uchida and Zhu, 2001] is the computers language to represent meaning contained in natural language sentences. But obtaining UNL expressions is an expensive operation, as is the task of creating multi level indexing on them. An intermediate step towards obtaining UNL expressions is semantically relatable sets (SRS) [Mohanty et al., 2005]. SRSs are created using parse tree for the sentences with the use of several heuristics and lexical resources like WordNet, Oxford Advance Learners' Dictionary. Our representation of the information is on SRSs containing bag of words in ***semantic proximity***. In what follows we explain SRSs and the initial results we obtain from applying the SRS based method.

3.1 Semantically Relatable Sets

Semantically Relatable Sets comprise of sets of words which are semantically related. The following example gives an insight.

Consider the sentence:

(1) *The man bought a new car in June.*

This sentence contains five content words - *man, bought, new, car, June* - and three function words - *the, a, in*. In order to obtain the semantic representation of (1), we need the following sets:

(2) *a. {man, bought}*
b. {bought, car}
c. {bought, in, June}
d. {new, car}
e. {the, man}
f. {a, car}

The words within these sets have to be related and the sets themselves need linking. This is depicted in Fig 1.

It has been postulated that a sentence needs to be broken into sets of at most three forms, as shown in (3).

(3) *a. {CW, CW}*
b. {CW, FW, CW}
c. {FW, CW}

Where FW stands for *function words* and CW stands either for a *content word* or for a *clause*. These sets are called *Semantically Relatable Sets (SRS)* and are defined below.

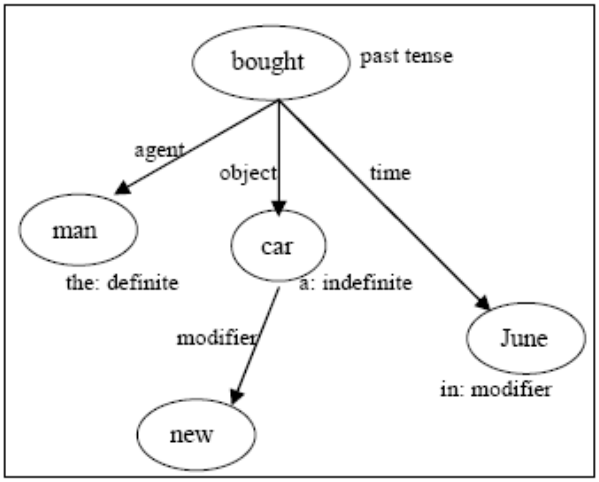

Fig. 1: Semantic graph of the sentence (1)

Definition: A semantically relatable set (SRS) of a sentence is a group of unordered words in the sentence (not necessarily consecutive) that appear in the semantic graph of the sentence as linked nodes or nodes with speech act labels.

SRSs can be used to represent different kinds of constituents as illustrated below.

Consider the sentence:

(4) *The boy saw the girl in the office.*

The sets, *{The, boy}*, *{boy, saw}* and *{the, office}* are three SRSs which are generated from semantically connected words in the sentence. The sets *{saw, girl}* and *{saw, in, office}* illustrate the fact that SRSs can span across the sentence to bring together semantically related non-consecutive entities like *"saw"* and *"office"*.

Now consider the following sentence:

(5) *The boy said that he was reading a novel.*

In the above sentence, the embedded clause *"he was reading a novel"* is denoted in the SRS representation by the term *SCOPE*. A *SCOPE* provides an umbrella for the words occurring in a clause or involved in compounding. The SRS for the clause words such as *{he, reading}* are marked being under *SCOPE*, as illustrated in (6). The semantic relation between the embedded clause and the words in the main clause is depicted through the SRS *{said, that, SCOPE}*.

(6) *a. {the, boy}*
b. {boy, said}
c. {said, that, SCOPE}
d. SCOPE:{he, reading}
e. SCOPE:{reading, novel}
f. SCOPE:{a, novel}

g. SCOPE:{was, reading}

The phrase *"John and Mary"* in sentence (7) shown below, represents a compound concept and is hence, marked under *SCOPE*.

(7) John and Mary went to school.

The linking of this phrase to the rest of the sentence is indicated by (8a).

(8) a. {SCOPE, went}
b. SCOPE:{John, and, Mary}
c. {went, to, school}

These examples illustrate different cases of SRS construction leading to the semantics of a sentence. Thus a search on SRS representation of documents and query helps us in retrieving documents based on semantics. The SRS generator module used generates the SRSs from the parse tree of the sentence. The parse tree is traversed in a breadth first manner and each node of the tree is processed according to its tag, head word and neighbors to generate the SRSs. The interested user can find more details on SRS in [Mohanty et al., 2005].

3.2 SRS Based Search

The relevance score for a document *d* is evaluated as follows:

$$R_q(d) = \frac{\Sigma_{s \in Sd}\, r_q(s)}{|S_d|} \qquad (i)$$

where,

$R_q(d)$ = Relevance of the document *d* to the query *q*
$|S_d|$ = Number of sentences in the document *d*
$r_q(s)$ = Relevance of sentence *s* to the query *q*

The relevance of the sentence *s* to the query *q* is calculated as:

$$r_q(s) = \frac{\sum_{srs \in q} weight(srs) * pres_s(srs)}{\sum_{srs \in q} weight(srs)} \qquad (ii)$$

where,

weight(srs) = weight of the SRS *srs*. It depends on the type of the SRS (see (3) above for SRS types).
$pres_s(srs)$ = It is a boolean function which returns true if *srs* is present in sentence *s*, false otherwise.

4 Experimental Setup

To test the performance of the SRS based search we used the Text Retrieval Conference (TREC) data. We chose *1919* documents at random from the AP newswire, Wall Street Journal and the Ziff data, and *250* queries (title field from topics *1-200* and *251-300*. These title fields are quite succinct, typically having not more than 4 words). Queries 201 to 250 were for question answering task and were not considered. The search for these queries was made on the SRS search engine and compared with the results of *tfidf* scores. *Lucene* [Cutting, 1998] was used for getting t*fidf.*

5 Initial Results

Fig. 2 shows the recall, precision and Mean Average Precision (MAP) comparison between *tfidf* and the SRS based search method. We see that even though the precision of the SRS based method is very high compared to *tfidf*, the method suffers with poor recall.

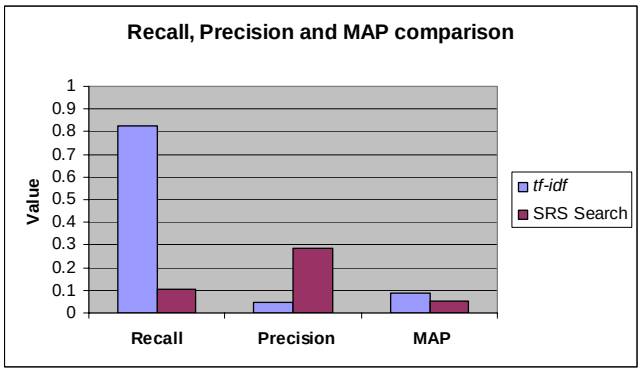

Fig. 2: Results of SRS search in comparison with *tfidf*

Because of the poor recall, the mean-average-precision was also not able to score well. But the dominating precision encouraged us to further explore and enhance the method. In the following section we discuss the various reasons for low recall and present the enhancement schemes.

6 Reasons for Low Recall

The initial results showed the potential of SRS based search technique in retrieving precise results. But the major drawback was the low recall of the technique. Here we describe the various reasons that were responsible for low recall.

6.1 Morphological Divergence

Morphological divergences occur between the query and document words. Consider the following example.

Query: "child abuse"
Query SRS: (child, abuse)
Sentence: "children are abused"
Sentence SRS: (children, abused)

Here the sentence contains terms "children" and "abused", while the query contains "child" and "abuse" respectively. Thus, we get two different SRS tuples.

6.2 Synonymy/Hypernymy/Hyponymy Divergence

The next dominant reason for low recall was the synonymy/hypernymy/hyponymy problem where the synonyms, hypernyms or the hyponyms of the query keywords were present in the documents. Examples for different cases are presented below.

Synonymy Case
Consider the following example:

Query: "antitrust cases"
Query SRS: (antitrust, cases)
Sentence: "An antitrust lawsuit was charged today."
Sentence SRS: (antitrust, lawsuit)

As we can see that here "cases" and "lawsuit" are synonyms of each other. But since SRS search engine just compares the query SRS tuples with sentence SRS tuples, the relevance will not be found in this case. It has to be noted that *tfidf* will fetch this document because of the keyword "antitrust".

Hypernymy Case
Suppose the query has keyword "mammal", while the document has keyword "animal". We can see that "animal" is the hypernym of "mammal". Thus due to similar reasons as stated above, the document will not be fetched.

Hyponymy Case
The query can be "mammal" whereas the document might contain "dog", where "dog" is the hyponym of "mammal". Again, these divergences restrict relevant documents to be fetched.

6.3 Physical Separation Divergence

There were problems when words in query were found physically separated in the document sentence. Consider the following example.

Query: "antitrust lawsuit"
Query SRS: (antitrust, lawsuit)
Sentence: "The federal lawsuit represents the largest antitrust action"
Sentence SRSs: (lawsuit, represents), (represents, action), (antitrust, action)

Here we see that although the words "antitrust" and "lawsuit" are semantically related, they fall in different SRS tuples and hence the document is not retrieved.

6.4 Other divergences

There were problems due to other divergences too. Few examples are listed below:

Query: "debt rescheduling"
Query SRS: (debt, rescheduling)
Sentence: "rescheduling of debt"
Sentence SRS: (rescheduling, of, debt)

Query: "polluted water"
Query SRS: (polluted, water)
Sentence: "water pollution has increased in the city"
Sentence SRS: (water, pollution)

Query: "sheet charges"
Query SRS: (sheet, charges)
Sentence: "charges on a sheet"
Sentence SRS: (charges, on, sheet)

Here we see that the query SRSs don't match with the sentence SRSs even though they have similar meaning.

7 Enhancements to Handle Divergences

7.1 Stemming

Words in the document and the query were stemmed before matching. The stemmer used is based on WordNet and gets the base form, while keeping the tag of the word unchanged. E.g., the word "children_NN" will be stemmed to "child_NN", but the word "childish_JJ" will not be stemmed to "child_NN", since the word "childish" is an adjective, whereas "child" is a noun. This stemming takes care of the morphological divergence problem discussed in section 6.1.

7.2 Using Word Similarity

The synonymy/hypernymy/hyponymy problem motivated us to incorporate the notion of "semantic similarity" between two paradigmatic words in the SRS based search. So, we incorporated the semantic similarity approach (*path*) using WordNet proposed by [Pedersen et al., 2004].

To affect the use of word similarity, the formulation of the sentence relevance measure $r_q(s)$ was changed to

$$r_q(s) = \frac{\sum_{srs \in q} weight(srs) * max_{srs' \in s}(t(srs,srs'))}{\sum_{srs \in q} weight(srs)} \quad (iii)$$

where,
weight(srs) = weight of the SRS *srs*. It depends on the type of the SRS (see (3) above for SRS types).
t() is the SRS similarity measure given by,

$$t(srs,srs') = t(cw1,cw1') * equal(fw,fw') * t(cw2,cw2') \quad (iv)$$

For *(FW,CW)* matching, *t(cw1,cw1')* is set to one and for *(CW,CW)* matching, *equal(fw,fw')* is set to one. In all other cases, *t(w1,w2)* (*equal(w1,w2)* for function words) gives the relatedness measure of *w1* and *w2* calculated using the baseline similarity measure "path" discussed in [Pedersen et al., 2004].

7.3 SRS Augmentation

To deal with the "Other divergences" discussed in section 6.4, numerous rules were developed to augment the SRSs in the documents as well as the query. Some example rules with their explanation are as follows:

Rule: *(noun1, in, noun2) => (noun2, noun1)*
Example: (defeat, in, election) will create an augmented SRS as (election, defeat)

Rule: *(noun1, on, noun2) => (noun2, noun1)*
Example: (charges, on, sheet) will create the SRS (sheet, charges)

Rule: *(adjective, noun) => (noun, adjective_in_noun_form)*
Example: (polluted, water) will augment (water, pollution)

Some rules have restrictions e.g., consider the following example:

Rule: *(adjective, with, noun–(ANIMATE)) => (noun, adjective_in_noun_form)*

Here the noun *noun* after *with* should not be of ANIMATE class for the rule to be applicable. E.g., (angry, with, result) will augment (result, anger), whereas (angry, with, John) will not augment (John, anger).

8 Implementation

The JWNL API [Didion, 2004] was used for stemming discussed in section 7.1. To obtain the similarity measure between two words as discussed in section 7.2, a PERL package by [Pedersen et al., 2004] was used. Online computation of the word similarity using WordNet was not feasible since it would have decreased the retrieval speed dramatically. To avoid this situation, similarity measures were required to be pre-computed. We could not compute similarity for all word pairs since the number of such pairs was very high. Instead for a given word (noun/verb), we calculated its similarity with those words only which were related to it through WordNet (synonymy/hypernymy/hyponymy) up to depth 2. An average of 200 related words for every word were found and similarity measures were computed.

For SRS augmentation discussed in section 7.3, we required the functionalities of obtaining derived forms e.g., childish_JJ to child_NN. Most of the derived forms were directly obtained from WordNet. But for some cases there are no derived forms directly linked in WordNet. E.g., nouns are linked towards derived verb forms only. Therefore to obtain the adjective form of a noun some other method was needed. In these exceptional cases, we used the Porter Stemmer [Porter, 1980] to get the stem of the source word and then searched for that stem in the WordNet with the required target form. Among the various stem-matched words found, the word with the same stem and largest lexicographical match with the source word was considered as the derived form.

For the retrieval purpose, 200 top ranked documents by *tfidf* and the documents retrieved by the SRS based search method of equation *(ii)* were merged and then the relevance of these documents were calculated by the enhanced SRS based search engine of equation *(iii)*. SRS weights, *weight(srs)* were chosen empirically as *40, 50* and *10* for the *srs* types *(CW,CW)*, *(CW, FW, CW)* and *(FW, CW)* respectively. Since *(CW, FW, CW)* has more information compared to the other two *srs* types, it was assigned the highest weight value. Similarly, since *(CW,CW)* has more information compared to *(FW, CW)*, it was assigned the next highest weight value.

9 Final Results

The final experiment was carried out after applying all the enhancements discussed in section 7. The experimental setup here was same as discussed in section 4.

Fig 3 shows the curves of three metrics (recall, precision and MAP) with varying cutoff values. These cutoff values are minimum relevance scores for a document to be qualified as *retrieved*. The cutoffs can be set according to the retrieval requirement of a system.

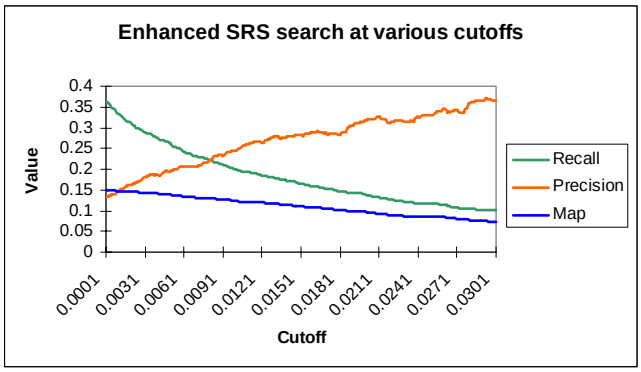

Fig. 3: Recall, Precision and MAP with varying cutoff for enhanced SRS based search

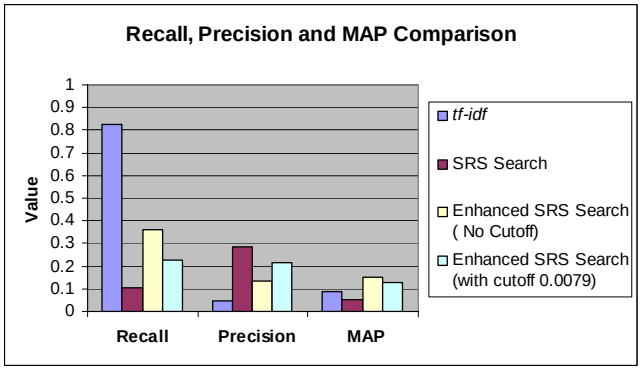

Fig 4. Comparison between *tfidf*, SRS based and enhanced SRS based methods

In Fig. 4 we see the comparison of recall, precision and MAP between *tfidf*, SRS search and the enhanced SRS based search methods. Clearly, the recall of the enhanced system has dramatically improved (0.362 from 0.102) with significant rise in MAP (0.149 from 0.054) as well. We see that our enhanced SRS based search method dominates the *tfidf* method with a high precision (0.131 compared to 0.049) and an improved MAP (0.149 compared to 0.086).

A fall in precision has come into picture because of the boost in recall, but still the overall precision is consistently much better than *tfidf*. Also, the cutoffs can be used to maintain a tradeoff between recall and precision. Measures with a sample cutoff are also shown in Fig 4.

10 Conclusion

We have presented a search strategy based on semantically relatable sets which combine words in semantic proximity. This method avoids the full semantics extraction from sentences which is a costly operation. The experimental results on TREC show that our semantic proximity based search is more effective than conventional *tfidf* based search. The system filters out non-sense documents and provides high precision in the retrieval. The high MAP value signifies the overall quality of the method, since MAP contains both precision and recall elements and is also sensitive to ranking.

The future work consists of automatically determining various parameters of the system e.g., weight parameter for different SRS types. Currently these parameter values are determined experimentally. The physical separation divergence problem discussed in section 6.3 also needs to be addressed.

References

[Corley, 2005] C. Corley and R. Mihalcea. Measuring the Semantic Similarity of Texts. *In the Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 1318, Ann Arbor, June 2005.

[Croft, 1995] W. Bruce Croft. What Do People Want from Information Retrieval?. *D-Lib Magazine*, 1995.

[Cutting, 1998] D. Cutting. The *Jakarta Lucene* project, 1998, http://jakarta.apache.org/lucene

[Didion, 2004] John Didion. Java WordNet Library (JWNL), 2004, http://sourceforge.net/projects / jwordnet

[Evans and Zhai, 1996] D. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for information Retrieval. *In the Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 17-24,* 1996.

[Guarino et al., 1999] N. Guarino, C. Masolo, and G. Vetere. OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems*, Vol. 14, No. 3, May/June 1999.

[Guha et al., 2003] R. Guha , Rob McCool, and Eric Miller. Semantic Search. *In the Proceedings of the 12th international conference on World Wide Web*, May 20-24, Budapest, Hungary, 2003.

[Mayfield and Finn, 2003] J. Mayfield and T. Finin. Information retrieval on the semantic web: Integrating inference and retrieval. *In Proceedings of the SIGIR 2003 Semantic Web Workshop*, 2003.

[Mohanty et al., 2005] Rajat Kumar Mohanty, Anupama Dutta, and Pushpak Bhattacharyya. Semantically Relatable Sets: Building Blocks for representing semantics. *In the Machine Translation Summit,* 2005.

[Pedersen et al., 2004] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. *In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, July 25-29, 2004

[Porter, 1980] M.F. Porter, An algorithm for suffix stripping. *Program*, 14(3) pp 130−137, 1980.

[TREC] The Second Text Retrieval Conference (TREC-2). Edited by D K Harman. Gaithersburg, MD: NIST 1994.

[Uchida and Zhu, 2001] Hiroshi Uchida and Meiying Zhu. The Universal Networking Language beyond Machine Translation. *UNL Foundation*, 2001.