Context-Based Similar Words Detection and Its Application in Specialized Search Engines

Hisham Al-Mubaid
University of Houston-Clear Lake
Department of Computer Science
Houston, TX 77058
hisham@cl.uh.edu

Ping Chen
University of Houston - Downtown
Department of Computer Science
Houston, TX 77002
chenp@uhd.edu

ABSTRACT

This paper presents a new context-based method for automatic detection and extraction of similar and related words from texts. Finding similar words is a very important task for many NLP applications including anaphora resolution, document retrieval, text segmentation, and text summarization. Here we use word similarity to improve search quality for search engines in (general and) specific domains. Our method is based on rules for extracting the words in the neighborhood of a target word, then connecting this with the surroundings of other occurrences of the same word in the (training) text corpus. This is an on-going work, and is still under extensive testing. The preliminary results, however, are promising and encouraging more work in this direction.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval – *Information Search and Retrieval*

General Terms

Algorithms, Management, Performance, Design.

Keywords

Word-similarity detection, information retrieval

1. INTRODUCTION

Currently there are huge amounts of documents existing in the *World Wide Web*. Finding information that a user is interested in becomes a critical task. Modern web search engines can cache, index and search several billion of web pages, which only includes a small part of all existing documents in the Web. And even for this small amount, the search quality could not meet a user's requirements in many cases. Many ideas have been proposed to improve the web search quality, which can be measured with the following two metrics: (1) *Precision rate*: the ratio of the number of relevant documents retrieved to the total number of documents retrieved. (2) *Recall rate*: the ratio of the number of relevant documents extracted to the total number of relevant documents in the Web [5]. A search engine usually tries to work with user keywords and a pre-built index for a locally cached web page clusters. Then, the retrieved documents should be ranked before they can be presented to the user. And due to the difficulty of semantic analysis of these web pages, the information on linkage structure among web pages are used to determine the importance of these web pages, and to rank them structure, index building for more efficient accessing, etc. Relatively few techniques have been considered to improve how a search engine processes a user's query and tries to understand it more semantically.

In this research we propose a new approach of finding similar and related words to improve web search quality by expanding the search query.

The reasons we are interested in search engines specialized in a specific domain, not in a general domain, are:

- Using similar words to expand a query has been tested in numerous information retrieval systems, and due to the difficulty of multi-sense problem, query expansion will hurt retrieval quality in many cases. In a specific domain, only one of very few senses of each word will be used.
- In a general domain, lots of documents have very poor quality and are neither useful nor of any interest to users. Such a noisy environment will confuse a retrieval system and make distinguishing of relevant documents from useless ones very difficult.

2. BACKGROUND

Better understanding of user queries will benefit both precision and recall rates. Although search engines usually do not limit the query format, most queries consist of several keywords connected with logical operators. In this paper we only consider this kind of queries since it is used most popularly. In [5] Moldovan proposed to use *WordNet* to extend the search process based on semantic similarity. In this approach a web search engine searches not only the keywords in the user queries, but also the semantically similar words obtained from WordNet. Since more keywords are used during matching process, more documents will be retrieved, and so some new operators are proposed to restrict the number of documents presented to the user. Such an approach raises the question of how precision and recall rates will be affected. Only a few experiments were performed on *AltaVista*, which could not reach a conclusive decision. Moreover, the dependence on a human-built thesaurus, such as WordNet, which is a time-consuming and an ad-hoc process, will generate serious coverage gaps if used for technical domains. Hence automatic discovery of related (no necessarily similar) words directly from the contents of a corpus could encode more semantic information that is especially critical in current web search engines [4]. In summary our main contributions are: (1) Automatic generation of a list of related and similar words from contextual corpus with more encoded semantic information without the need for any external (pre-built) resources. (2) Expand a user query with related words for better search quality.

IUI'05, January 9–12, 2005, San Diego, California, USA.
ACM 1-58113-894-6/05/0001.

3. A CONTEXT-BASED TECHNIQUE

We adopt a context-based text analysis approach that is roughly stemmed from our previous work in context-based word prediction and context-based spelling detection [1,2]. Our approach also is somewhat similar to the one presented in [3]. The main idea is to collect, from a given text corpus, some context-based information on the *target* words. The method requires a (*training*) text to be used for extracting the similar and related words. We collect the information on the target word from the instances of that word in the given text. The information about the target word is then encoded and stored in one or more vectors. Each vector simply relates the contexts of the word for all its occurrences in the text corpus. The context of a word *w* is the (surrounding) *n-1* words around *w*, and *n* is the window size. For example, if the window size is *5*, then the context of the target word *w* in the sentence: $\{p_2 \;\; p_1 \;\; \underline{w} \;\; f_1 \;\; f_2\}$ consists of the two preceding words p_1 and p_2 (p_1 : the word before the target word, p_2: the second preceding word), and the two following words f_1 and f_2 . Initially, all (distinct) corpus words are indexed into a file as shown in Table 1.

Table 1: Word index file

Word	Index
This	1
article	2
presents	3
a	4
study	5
…..	….

Then, for a given target word, the vectors are created , such that, each entry in the vector indicates the index of the word with the highest co-occurrence values with the target word, in one of the context locations around *w*. For example, to continue on the above example, the following vector for *w*:

55	*4*	*206*	*33*

indicates that the word that precedes *w* the most, in the corpus, is the word whose index is 4 (i.e. '*a*'), the word that follows *w* the most is the word whose index is *206*, the highest second preceding word to *w* is the word with index *55*, and the highest second following word to *w* is the word whose index is *33*. Of course, this construction represents window size of 5, which is considered very small, in our work, however, we experiment with window sizes range from 5 to 11. Clearly, this encoding extracts the semantic aspects of the words, and can very well lead to predicting semantically similar words, which we verify in this research through extensive experimentation. In some cases, we construct more than one vector for a target word to accommodate for other very frequently co-occurring context words. For example, a second vector for the word *w* might be:

-	-	125	-

which indicates that, besides 206, *w* is very frequently (and above some threshold) followed by the word with index 125. In our experiments we set the threshold to 25% and a maximum of *three* vectors for each target word. Next, the similarity among vectors is computed using the *cosine* measure or the *L2* measure. The cosine measure has a proven effectiveness in finding vector similarities in many applications, for example in *document classification* [6]. Then the similar words are used to augment and expand the query to improve the search results returned from a search engine.

4. A CASE STUDY

In this case study we build a system to mine financial data from U.S. Securities and Exchanges Commissions (SEC at *www.sec.gov*). We analyzed several 10-K filings from different companies. 10-K filing is an annual financial and transactional report required by SEC to all public companies, and it gives the most comprehensive information on financial information for a public company. At SEC website, 10-K filings of around 10,000 public companies in the last few years are available, and totally there are around 50,000 filings. The size of these documents is around 30GB. The reasons we chose these filings for our experiments are:

- 10-K filings are prepared very carefully and hence are of high quality
- Amount of 10-K filings is sufficient for the purpose of our word similarity analysis
- 10-K filings are specific to financial topics but cover all industries in general, which will provide an ideal semi-close analysis environment for experiments.

The architecture for our system has four components, a web crawler which crawl the SEC website and cache filings to a local repository, a word similarity analysis unit that will analyze these filings, a knowledge base that saves index information and lists of similar words used to expand queries, and a search engine to perform information retrieval. The system architecture is shown in figure 1

So far we have successfully built a web crawler to retrieve SEC filings. We also have implemented our word similarity tool with C++ in a Linux system. Table 2 shows some of the results obtained from the initial experimentations. We analyzed a 10-K filing of a company with CIK 1800 (CIK is the company identification number used by SEC).

Looking at the word pairs and the corresponding similarity values in Table 2 clearly reflects the efficiency of the methods. Also we can find in the table words that are similar or related for other/general domains, for example {best, outstanding}, {info, knowledge}, {address, zip}, {price, valuation} that may benefit the subsequent retrieval step.

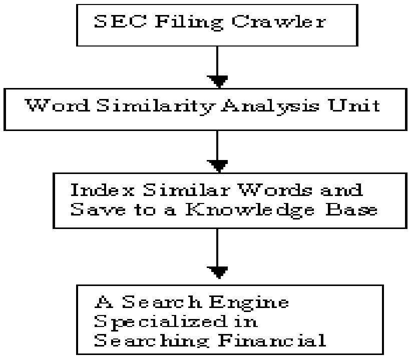
SEC Filing Crawler
Word Similarity Analysis Unit
Index Similar Words and Save to a Knowledge Base
A Search Engine Specialized in Searching Financial

Figure 1: The System Architecture

Table 2: A list of similar word pairs

Word1	Word2	Similarity
accept	accession	0.999
content	message	0.997
info	knowledge	0.977
industrial	enterprises	0.990
filing	form	0.959
address	zip	0.994
phone	telephone	0.919
delinquent	criminal	0.971
best	outstanding	0.965
price	valuation	0.993
consolidated	united	0.919
care	service	0.951
pediatric	children	0.968
consumer	purchasers	0.932
formula	similac	0.923
formula	isomil	0.940

We have conducted some experiments with *Google* search engine (*www.google.com*) using word pairs in Table 2 as keywords. First, we used each word in one word pair as keyword, and evaluated the first 20 retrieved documents and marked each document as "*relevant*" or "*not relevant*". Then we used both words of that word pair as keywords, and also evaluated and marked the first 20 retrieved documents as "relevant" or "not relevant". After comparing the "relevant" percentages we found the searches with both words of a word pair can improve the *Relevance* rate. To be specific, 7 out of 10 searches with both words from a word pair will have higher relevance rate, and on average the relevance rate can be improved by 15%. Also retrieved document quality is improved in the following two aspects:

1. Using one word as keyword returns some documents written in other languages, which happen to have some isolated English words. By using a word pair, the probability of retrieving this kind of documents is reduced.
2. In the retrieved documents, some have a very simple summary entry, such as:

   "At Length Magazine

   www.atlengthmag.com/ -2k –Cached-Similar pages"

   Users rarely click this type of entries due to lack of details. But in retrieving documents using word pairs as keywords, the number these entries is reduced.

In another evaluation, we conducted experiments on specialized search engines, including searching the directory *Business and Economy* in *Yahoo!* and searching the *SEC* site. The searches done using word1 then word2 (Table 2), the retrieved documents were 85% – 95% similar, which verifies the similarity values in the table. The similarity between the two words of each pair is verified in the financial domains because it was computed from financial domain (using *SEC* documents).

5. CONCLUSION AND FUTURE WORK

This research investigates a number of issues including determining and fine-tuning the parameters like the window size, the threshold percentage, and the max number of vectors. Furthermore, we plan to investigate the important issue of using multiple vector similarity models. The approach has been evaluated with an initial set of experiments, and the results showed a great promise. Finally, our approach is computationally effective and more practical as the vector size (*7 on average*) is extremely small compared to 1200 in the method of [3]. Moreover, our method does not require any external resource or pre-built thesaurus.

6. REFERENCES

1. Al-Mubaid, H., Context-Based Word Prediction and Classification. In *Proceedings of 18th Intl Conf on Computers and Their Applications CATA-2003*, Hawaii, USA, 2003.

2. Al-Mubaid, H., and Truemper, K., Learning to Find Context-Based Spelling Errors, in *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques,* E. Triantaphyllou and G. Felici (Editors), Kluwer Academic Publishers, forthcoming 2004.

3. Gauch, S., Wang, J., and Rachakonda, M. A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases. *Proc. of 6th Intl Conf. on Information and Knowledge Management (CIKM'97)*, 1997.

4. Lin, D. Automatic Retrieval and Clustering of Similar Words. *Proceedings of COLING/ACL-98,* Montreal, Canada, 1998, 768-774.

5. Moldovan, D., and Mihalcea, R. A WordNet-based interface to Internet search engines. *In Proceedings of FLAIRS-98* Sanibel Island, FL, May 1998

6. Han, E., and Karypis, G., Centroid-based document classification: Analysis and experimental results. *Proceedings of the 4th PKDD,* France, 2000.