

Thesis Proposal Form

Identifying Sets of Related Words from the World Wide Web

Author : Raveendranathan Pratheepan

Advisor: Dr. Ted Pedersen

Proposal:

1. **Provide a brief but complete statement of the research problem to be solved. In a separate paragraph, specify the goals of this research.**

The overall goal of my thesis research is to use the World Wide Web as a source of information to identify sets of words that are related in meaning. Methods have been developed to identify words that are related in meaning in fixed or static corpora of text [2]. However, given the availability of huge amounts of text via the World Wide Web it is important to develop methods that can take advantage of this fact. The Web creates a unique set of challenges, including its ever-changing state, and the presence of repetitive, noisy, or low-quality data.

We will use a commercial search engine such as Google to retrieve text from the Web. Google and many other search engines provide APIs that allow a programmer to interface with their content, and retrieve the data in a more convenient form. Thereafter we will process that data to find sets of related words.

As an example of what we hope to achieve, if we search for words related to *George W. Bush* and *Bill Clinton*, we would hope to find a set of words that include *Ronald Reagan* and *Jimmy Carter*. These names are related to each other as Presidents of the United States. We will also attempt to score the related terms we find, such that a leader of a different country would be scored slightly lower than a former President of the USA.

As another example, if we searched for a word such as *bank* which has multiple senses, we might find two groups of related words. One of the groups can be *money*, *stock*, *state*, *national* and *central* that are

related to the financial sense of *bank*, and the second group can be *become*, *remain*, *seem* as words that are related to the depend sense of *bank*.

In particular, initial our approaches will rely on identifying words that co-occur with each other. For example, in the President example, it might be sufficient to recognize that a set of words occurs frequently with the term *President of the United States [name]*. However, not all sets will have such a clear criteria for membership, so we will need to develop methods that go beyond simple pattern matching and rely on probability models, clustering and learning algorithms.

In order to develop methods to group sets of related words together based on information as found in the World Wide Web, we will need to meet the following specific goals of this research:

- Develop pattern matching techniques that account for the limited matching capabilities of search engines.
- Develop techniques for estimating probabilities of word co-occurrences that account for the instability and noise of data as found on the World Wide Web.
- Develop techniques that cluster words based on the similarity of their surrounding context in a sample of text drawn from the World Wide Web.

2. **Provide a descriptive, one or two paragraph review, which sets the framework for the research and cites recent references which establish the importance of the work or problem to be solved. (Three to five references from recent publications are sufficient.)**

Our research is very much dependant on both the quantity and quality of the Web content that is used to group words that are related in meaning. Hence, we will use a commercial search engine such as Google to gather information about words from the Web. Google has a very effective ranking algorithm called PageRank which attempts to give more important or higher quality web pages a higher ranking [1, 3].

The basic idea behind the PageRank algorithm is to use the number of links to a web page as a source for ranking. Highly linked pages are ranked more important than pages that do not have as many links to them. The links themselves are divided into a set of backlinks and a set of forward links. Backlinks are links that link or point to a certain web page. Forward links are links that a web page points to. For example, as of November 2003, the Google web page has over 300,000 web pages linking to it. Hence, it can be assumed that the Google web site is an important and credible web page. Though this is the basic idea behind the algorithm, the number of backlinks alone cannot guarantee a good ranking. The ranking also depends on the rank of the page that is linking to it. For example, if a web page has only one backlink, but that backlink is from a credible web page such as the Stanford University home page, the web page would then be given a high ranking.

The Pointwise Mutual Information - Information Retrieval (PMI-IR) [4] algorithm shows how to use a commercial search engine (in this case Alta Vista) to exploit the Web as a source of information about words. In this case it is used to determine if words are positive or negative in their sentiment. This information is then used to calculate the Semantic Orientation (SO) of phrases, to decide if a review has an overall positive or negative recommendation. These can be reviews for automobiles, movies, banks and even travel destinations. The algorithm works in three steps. First, a part-of-speech tagger is used to identify words in the review as adjectives, nouns, verbs or adverbs. Next, the algorithm calculates the semantic orientation of each of the adjectives, nouns, verbs and adverbs extracted from the review by forming phrases out of adjacent words. The algorithm then queries a search engine with the combination of positive and negative words with each of these phrases to find the PMI measure, and thereby the Semantic Orientation. The final step is to calculate the average of the semantic orientation calculated thus far, and assign a value to the review as "recommended" or "not recommended". This method gives us some insight on how the Web can be used to find information about words, but in our research we will be focusing on how to group together related words, rather than trying to characterize individual words.

The Clustering By Committee (CBC) algorithm [2] tries to find sets of words related to the different meanings of a given term. This algorithm works by representing each word by a feature vector which corresponds to a context in which a particular word can be used. Take the example *threaten with ...* as a context, and if words such as *handgun* or *pistol* occurred in this context, *threaten with* is a feature of words such as *handgun* and *pistol*. The algorithm has three phases. Phase 1 calculates each elements top k similar elements. Phase 2 constructs a collection of tight clusters. The elements of each cluster form a committee. The CBC algorithm tries to form as many committees as possible, ensuring that each new cluster is not very similar to any of the existing clusters.

The goals of my research are similar in spirit to that of CBC, but would be working with information derived from the World Wide Web rather than a large static corpus of text.

- 3. List the resources (in terms of software and hardware) required to complete the work. If not available within the department, indicate how/where such resources will be obtained. If only one instance of a required resource is available, indicate how you will account for failure of that resource.**

Hardware:

The current facilities provided by the department are adequate for our research.

Software:

- Perl - a freely available programming language that is already installed on all departmental machines.
- Google API: a freely available Java package provided by Google to allow programs to interact with their search engine. This has already been installed on all departmental machines.
- Perl::SOAP::LITE - A CPAN module required to interact with Google API (already installed).
- Perl::LWP - A CPAN module that allows Perl to interact with the Web (already installed).

- Perl::Net - A CPAN module that allows Perl to interact with the Web (already installed).

Data Requirements:

The following online data resources are required to evaluate my approach:

- Macquarie Thesaurus - licensed to Dr. Pedersen and available.
- Longman's Dictionary of Contemporary English (LDOCE) - licensed to Dr. Pedersen and available.
- Roget's Thesaurus - in process of being licensed to Dr. Pedersen.
- WordNet - A freely available lexical database that is already installed on all departmental machines.

4. **In a numbered list below, specify each step required to solve the problem along with a specific target date for the completion of that step. (The last two steps should be the write-up of the thesis and the completion of oral exam and colloquium, respectively, along with their scheduled dates.)**

(a) **Fall 2003**

- Nov 26 : Submit Thesis Proposal
- Dec 19 : Deploy a prototype of a Google Toolkit. This will be a Perl module that supports a variety of methods to obtain information from Google via their API.

(b) **Spring 2004**

- March 15 : Finish developing an approach that identifies related concepts using strict pattern matching (as in the President's example above).
- March 31 : Complete an evaluation of the pattern matching approach relative to WordNet, LDOCE, the Macquarie Thesaurus, and Roget's Thesaurus.

- May 15 : Finish developing an approach that identifies related concepts using more flexible pattern matching. Rather than requiring exact matches of nearby words, allow for more distant words to be factored into the decision.
- May 25 : Complete an evaluation of the flexible pattern matching approach relative to WordNet, LDOCE, the Macquarie Thesaurus, and Roget's Thesaurus.

(c) **Fall 2004:**

- Nov 1 : Finish developing an approach that identifies related concepts using a clustering approach, similar to the CBC algorithm [2] but adapted to the unique characteristics of web data.
- Nov 15 : Complete an evaluation of the clustering approach relative to WordNet, LDOCE, the Macquarie Thesaurus, and Roget's Thesaurus.

(d) **Spring 2005:**

- Feb 15 : Finish developing an approach that identifies related concepts using a machine learning approach such a decision trees or rule based learner.
- Feb 28 : Complete an evaluation of the machine learning approach relative to WordNet, LDOCE, the Macquarie Thesaurus, and Roget's Thesaurus.
- Apr 15 : Submit thesis to committee.
- Early May : Thesis defense.

5. **Specify clearly how the results of the research will be evaluated. What objective measures will be used to establish that the goals of the research have been met?**

The results of this thesis will be evaluated at each stage of the research. We intend to use WordNet, LDOCE, The Macquarie Dictionary and Roget's Thesaurus as sources of sets of related words. We will extract sets of words from each of these resources and compare them to those generated by our method.

6. **State clearly the contribution to research that this work will make.**

This work will develop methods that will group related sets of words together using the World Wide Web as a source of information. This is unique and represents a contribution since existing approaches are based on static corpora such as newspaper articles that are more regular and predictable than Web data. The use of the Web will mean that the approach can be very flexible, and support a wide range of content and even languages.

References

- [1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117, 1998.
- [2] Dekang Lin and Patrick Pantel. Concept discovery from text. In *Proceedings of the Conference on Computational Linguistics*, pages 577 - 583, Taipei, Taiwan, 2002.
- [3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.
- [4] Peter D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pages 417 - 424, Philadelphia, Pennsylvania, 2002.

Signatures:

Student

: _____

Advisor

: _____