Related Word and Phrase Set Generation

Sam Bradley
CS 5762 Introduction to Natural Language Processing
T. Pedersen B. Thomson

Problem: To develop a method that will take a small set of related words and use results from *Google* to find a larger set of words that are also related to the original set in a similar way.

Background: webopedia.com describes a search engine as, "A program that searches documents for specified keywords and returns a list of the documents where the keywords were found." These documents are usually cached, or stored, on a central system so they can be quickly accessed and searched. Keywords are just a set of words that should be present in the document. The simplest concept of a search engine is that the higher the frequency of the keywords in a document, the more relevant that document is to the search.

Google is an advanced search engine. Instead of merely searching documents for the frequency of keywords, *Google* implements a ranking system that considers various features of a document to determine which documents are most relevant to the search terms, or "keywords". The problem with solely using the counts of keywords within a document is that the results are easily skewed. For example, if you wanted your document to be found under certain search terms, all you would need to do would be to make sure your document contained those terms more then any other and it would be counted above any other document. Although this information is important to finding relevant data within a document, *Google* implements a clever way to "weed" out pages that merely contain the search terms in high quantity. Documents on the Internet contain links, or references to each other. If a document contains quality information on it, it will be more likely to be referenced within other documents. Using this method, a document containing very few occurrences of the search terms, but happens to be referenced may times within other documents would be counted above a document containing may occurrences of the terms, but is referenced very few times in other documents. This method is referred to as PageRank, and is one of many part of how *Google* determines which documents are most relevant to the search terms (Brin et al, 1998).

Google Sets is a feature of the Google search engine currently under development. Given at least one, or as many as five words, *Google Sets* will attempt to find other words that are somehow related to the original set of words using information found from searches made by the main Google search feature. An example would be given the set of words, "search, engine, google, sets, keyword", *Google Sets* returns the list "engine, Google, Search, Keyword, Excite, Yahoo, HotBot, Advanced, Additional, Introduction, Browsing, Recalls, Account." Whether or not it's obvious, not all these results fit the general pattern. This is due to the method used to find the set. Some words are not excluded because the method used to find words that relate to the set wants to include these as related words. It's hard to determine why this happens because the method *Google* uses is not public knowledge.

In the paper *Automatic Acquisition of Hyponyms from Large Text Corpora,* Hearst describes how language patterns can be used to determine relations between words or phrases. The example given, "The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string," indicates that the Bambara ndang is a type of bow lute, even though it's not specifically written that way (Hearst 1992). Similarly, even though not written in the previous sentence, it can be inferred that Hearst wrote the paper *Automatic Acquisition of Hyponyms from Large Text Corpora*. This idea can be used in three ways: To find or argue lexicon relations of words, relate noun phrases to a more general meaning, and relating phrases that have similar meaning but no lexical relation (Hearst 1992). Using this information will be useful in finding related sets of words and phrases. The exact method will be discussed in the solution section.

The purpose of this project will be to implement these methods and try to imitate this feature of *Google*.

Solution: For this project it is assumed that related words will appear with each other in corpora, meaning if a document describes the life of a snail, a shell or certain plants it prefers and such would be more likely to be mentioned in the document then a new sports car. The program will first need to search for relevant documents. Instead of forming a unique method to accomplish this, the program will use the standard Google search feature. Form the results of the Google search, sentence fragments containing words from the initial set will be tested to see if they follow certain patterns that indicate that the surrounding words are of a set. If the pattern occurs, then words that fall into certain places within the pattern can be deemed to be hyponyms, words that have an "is a" relation (Hearst 1992). These patterns, given by the Heart paper with examples are as follows:

NP$_0$ such as {NP$_1$, NP$_2$ … , (and | or)} NP$_n$

Ex: The bow lute, *such as* the Bambara ndang, is plucked and has an individual curved neck for each string.

-Hyponym("Bamabra ndang", "bow lute")

such NP as {NP,}* {or | and} NP

Ex: … works by *such* authors *as* Herrick, Goldsmith, and Shakespeare.

-Hyponym("author", "Herrick")
-Hyponym("author", "Goldsmith")
-Hyponym("author", "Shakespeare")

NP {, NP}* {,} or other NP

Ex: Bruises, wounds, broken bones *or other* injuries …

-Hyponym("bruise", "injury")
-Hyponym("wounds", "injury")
-Hyponym("broken bones", "injury")

NP {, NP}* {,} and other NP

Ex: … temples, treasuries, *and other* important civic buildings

-Hyponym("temple", "civic building")
-Hyponym("treasuries", civic building")

NP {,} including {NP,}* {or | and} NP

Ex: All common-law countries, *including* Canada and England …

-Hyponym("Canada", "common-law country")

-Hyponym(“England”, “common-law country”)

NP {,} especially {NP,}* {or | and} NP
Ex: … most European countries, *especially* France, England, and Spain.
-Hyponym(“France”, “European country”)
-Hyponym(“England”, European country”)
-Hyponym(“Spain”, “European country”)

(Hearst 1992)

All the words found and their frequency, or how many times they occur will be kept track of. The end result will be a list of these words in order from the most frequent to the least frequent.

Evaluation: The successfulness of this project will be based on a direct commpaeison to results given by Google Sets. Two list of equal size will be observed one generated by this project, and the other from Google sets. The words will be compared individually to the counterpart of the other list, that is to say the two top results will be compared to each other, the next two top results will be compared with each other and so forth. The words will be compared to each other by comparing the frequency they appear with the original set. For example, if the first result is “x”, and the initial set is “x” and “y”, then the frequency of “x y z” will be observed. The frequency will be determined using the standard Google feature. A ratio will be given of how many of the projects words outranked *Google Sets* to how many of the *Google Sets* words outranked the project. Any repeated words will result in the automatic outrank of the other system, as well as any blank results. Both sets deemed to be strong and week for the project will be tested. These sets can be found intuitively by searching for documents that contain the patterns used in the project, and using sets that are already know to be fund or not found in the pattern, depending if the desired result is a strong or weak set.

Results:
<u>Test 1</u>
Test Data: Timex, Rolex
Relation: Brands of watches
This test data was chosen because of the nature of the internet. The internet is full of stores and other information about brands of merchandise, so the thought is that the results should be fairly good. This set was found by searching for the pattern “such as”, and the set was found as a top result.

Sets.pl	**Google**	**High Freq.**
Rolex	Rolex	Tie
And	Timex	Google by 400 hits
Alba	Citizen	Google by 48520 hits
Seiko	Casio	Sets.pl by 3500 hits
Citizen	Seiko	Google by 7000 hits
Casio	Swatch	Sets.pl bvy 6900 hits
Tag-heuer	Omega	Google by 7600 hits
Hamilton	Breitling	Sets.pl by 8100 hits
Fine	Pulsar	Google by 13400
Such	Cartier	Google by 43180 hits
Â®	Fossil	Google by 43485 hits
Revelation	Concord	Google by 29799 hits
	Longines	Google
Ramblings	Tissot	Google by 43974 hits
	Wenger	Google

Set.pl:3 Google:11

In this test, Google appears to work better.

Test 2

Test Data: Novelists, Playwrights

Relation: Types of authors

This set was chosen by searching for including, and the set was ranked high.

Sets.pl	**Google**	**High Freq.**
Playwrights	Novelists	Tie
Writers	Playwrights	Google by 8000 hits
Novelists	Mystery	Sets.pl by 28210 hits
Playwrights	Humor	Google by repeat
And	Non fiction	Sets.pl by 24060 hits
Lyricists	Romance	Google by 4970 hits
Electronic	Young adult	Google by 90 hits
Or	Science fiction	Sets.p,l by 15740 hits
Screenwriters	Western	Google by 6110 hits
Scanners	Spirituality	Google by 990 hits
Poets	Children's	Sets by 22670
Poems	Poets	Google by 17800 hits
Ebay	Screenwriters	Google by 1478
Authors	Horror	Sets.pl by 9630
Short story writers	Musicians	Sets.pl by 9040

Sets.pl:7 Google:8

In this test both systems seem to work about as well as each other.

Test 3

Test Data: Gibson, Jesus

Relation: The Passion of the Christ

This set was found by searching for the key phrase "especially", and then a set was chosen that ranked low. The idea is this set should test something that might not work so well the way the proram is written.

Sets.pl	**Google**	**High Freq.**
And	Gibson	Google by 200000 hits
Such	Jesus	Google by 535000 hits
Releases	Moses	Sets.pl by 5400 hits
Injustices	Fender	Google by 4820 hits
Upsetting	Judaism	Google by 17630 hits
Shown	Judaism Messianic	Sets.pl by 46430 hits
Conflict	Jehovah's Witnesses	Sets.pl by 34193 hits
Politicians	Who I am	Google by 273700 hits
Terms	John the Baptist	Sets.pl by 97600 hits
Authoritative	God	Google by 299820 hits
Clergy	Graces	Sets.pl by

		15030
Criticism	Paul	Google by 181900 hits
Pleasing	David	Google by 211530 hits
Real	The bible	Sets.pl by 63000 hits
talmud	peter	Google by 186980 hits

Sets.pl:6 Google:9
In this test, google appears to have worked a little better.

Conclusions: All in all, Google Sets does seem to work slightly better, but not too much. Both systems do seem to have their own unique way of working, or Google works similarly, but gets different results for various reasons. These could include additional phrases to match and better searching because Google has better access to their own cached pages. Also, the tests results show that Google seems to put a lot of emphasis on the number of hits, but this could be coincidental. Bottom line of this evaluation is that both Google sets and the project perform fairly similarly, although they may do it in entirely different ways.

Sergey Brin, Rajeev Motwani, Lawrence Page, Terry Winograd (1998).
The PageRank Citation Ranking: Bringing Order to the Web
Retrieved 26 April 2004, from http://citeseer.ist.psu.edu/page98pagerank.html

Marti A. Hearst (1992) Automatic Acquisition of Hyponyms from Large Text Corpora
Retrieved 17 April 2004, from http://citeseer.ist.psu.edu/hearst92automatic.html