Assignment 3

Information Retrieval

Introduction to Expert Systems

COMP 474/6741

Prof: Dr. Sabine Bergler

Angela Gabereau

ID#: 4867815

December 17, 2007

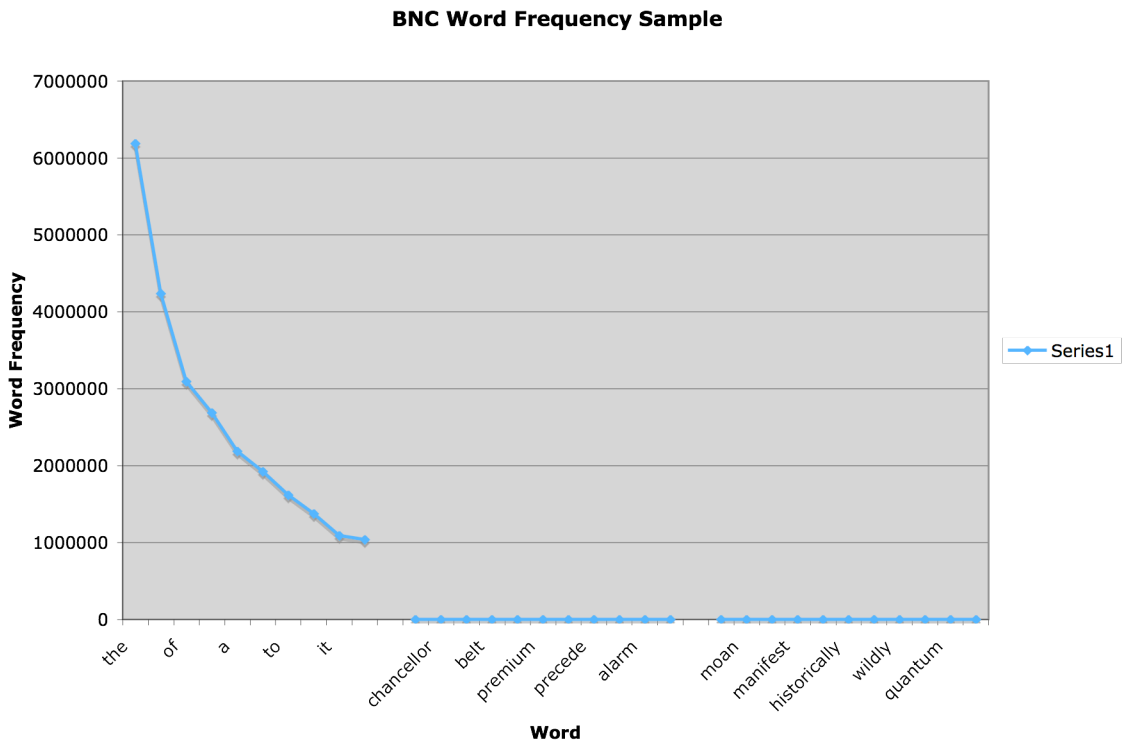
1. Word Frequency

(a) Describe Zipf's Law and how it applies to frequency of words in English. Give some data that you retrieve from the Web.

Zipf's Law is an empirical rule that describes the pattern of word usage across a corpus. It states that a word's frequency within a corpus is inversely proportional to its rank in a frequency table. This Zipfian distribution can be observed in any large sample of language.

On his website Adam Kilgarriff provides a lemmatised frequency list for the 6,138 words in the 100 million word BNC database. From this list, in the table below, I show the ten most frequently occurring words, a sample of the frequency of the middle ranking words and the last ten words on the list. I have plotted the 3 sample segments from Kilgarriff's frequency list on a graph. The curve of the line is characteristic of Zipfian distributions and clearly displays "the long tail" often associated with Zipf's Law.

Top 10 Words			Middle 10 Words			Bottom 10 Words		
Rank	Frequency	Word	Rank	Frequency	Word	Rank	Frequency	Word
1	6187267	the	3160	2335	back	6308	801	skipper
2	4239632	be	3161	2332	chancellor	6309	801	moan
3	3093444	of	3162	2329	crash	6310	801	manpower
4	2687863	and	3163	2328	belt	6311	801	manifest
5	2186369	a	3164	2326	logic	6312	801	incredibly
6	1924315	in	3165	2325	premium	6313	801	historically
7	1620850	to	3166	2325	confront	6314	801	decision-making
8	1375636	have	3167	2324	precede	6315	800	wildly
9	1090186	it	3168	2323	experimental	6316	800	reformer
10	1039323	to	3169	2322	alarm	6317	800	quantum



(b) Describe what is meant by "the long tail" and compare the usage of the term to the usage of the term "Zipf's law".

The Long Tail refers to the property of some statistical distributions that display a high-frequency data followed by low-frequency data that trial off, creating a "long tail." This property can be seen in the above graph where the words that occur at low frequency hug the x-axis.

Chris Anderson first introduced the phrase "The Long Tail" in an article in the October 2004 edition of Wired magazine. In this article Anderson describes a business model that takes advantage of The Long Tail of Zipfian distributions. The model orientates itself towards the demand that occurs in The Long Tail, selling products that are rare and are sold in small volumes. Customers have been shown to prefer the rare products and thus the majority of the sales occur in The Long Tail. The frequency of the sale of these products is low and thus they are represented in The Long Tail, but because the tail is long there is profit to be made. This business model is typical of web-based

business, which, because of reduced overhead when compared to store front business, can offer a wider selection of products to consumers.

Zipf's Law is generally talked about in reference to word frequency of textual documents. The Long Tail popularly refers to the high demand for rare products that is utilized by web commerce, but it can be used to describe Zipfan distributions as well. Zipf's Law can be seen in a broad range of domains and Zipf himself thought that his law could be used to describe patterns across all human activities (Belew, 64).

(c) Discuss what implications of IR strategies are when considering the long tail as a valuable part of the search space. Discuss in particular the suitability of the tf-idf ranking for keywords from the long tail.

The Long Tail of search space provides valuable information about the behavior of users. The Long Tail contains search terms that are rare but nonetheless resulted in a successful search; it contains more specific search terms and search terms never conceived of by other users. Search engines such as Search The Tail utilized The Long Tail to help users refine their searches. Search the Tail shows users possible variations of their queries, from broad to narrow search terms. The narrow search terms are those contained in the tail. By allowing users to refine their queries by utilizing search terms that occur in The Long Tail it increases the likelihood of the recall of relevant documents.

Users are not always able to fully define the answers they seek, there is a knowledge gap between the users' current knowledge and vocabulary and their information need. This knowledge gap is addressed by utilizing The Long Tail of search to aid users in refining their queries.

TF-IDF ranking weights the keywords of a document based on the frequency of the keyword in the document and the frequency of the keyword in the corpus, normalized for the document's length. Queries constructed from the long tail are generally short. According to Belew, in short queries, where multiple occurrences of the same keyword are rare, length normalization is ignored (Belew, 96). Thus for short queries the tf-idf weighting is simply the inverse document frequency. The basis for IDF is that

“infrequently occurring terms have a greater probability of occurring in relevant documents - and should thus be considered as being of greater potential when searching a database”(Jones, 1997). So when users are provided with the opportunity to refine the search by adding more search terms from The Long Tail and then this query is weighted with TF-IDF the likelihood of returning relevant documents increases at both steps. IF-IDF is well suited to queries constructed from The Long Tail because the terms occur at low frequencies and thus will be given a higher weight.

2. Indexing
Using the .html files from http://www-cse.ucsd.edu/~rik/foa/12h/, sections 1,2,3,4, and 6, compile an index for all words UNIX shell commands or your favorite scripting language.

(a) What is it size? How does its size compare to the index of the book?

Before any reduction, the size my index is 5791 words. The index of the book is approximately 1/5th of this size. My index is larger because the noise words have not been removed and the words have not been stemmed.

(b) Reduce the index to a more useful size. Which terms do you suppress? Why? What is the resulting size of your index? You may experiment with different strategies and compare the results. Submit and compare your favorite index with the official book index.

To reduce my index I began by sorting the words by their occurrence frequency. This gave me a Zipfian distribution, with the word "the" at the top of the list with 3369 occurrences in the document. According to Belew, "the most frequently occurring words are not really about anything"(Belew, 73). Thus I removed all the words in my index that had a frequency greater than 150. When I reviewed the list of removed terms it was clear that Belew's concept was correct, but there was some exceptions. For example, foa had a count of 710; document had 473 occurrences, query 225 and search 201. These four words outline the basic materials of the text. They occur so frequently throughout the text as to render them useless in distinguishing one passage in the book from another but they would certainly differentiate FOA from other books in a larger corpus. In other words they would make excellent external keywords but as internal keywords they would be ineffective as noise words.

Belew states that "the best keywords will not be the most ubiquitous, frequently occurring terms, nor those that occur once or twice, but rather those occurring a moderate number of times"(Belew, 77). Thus the next step in reducing my index was to remove the keywords in The Long Tail of my Zipfian distribution. I removed all terms with a frequency of less than 10. I again reviewed the list and found terms that would make important additions to the index. Examples are hyponym and hypernym, which both occurred only once in the text. This shows that although terms in the center of the

Zipfian distribution have a higher probability of expressing about-ness the terms that occur at higher and lower frequencies still may need to be considered.

The final step in reducing my index was to manually go through the list and remove noise words. This process demonstrated the necessity of familiarity with the domain of the text in order to identify what words are noise and which are necessary. To illustrate this point I will use the word “abstract”. It occurs 21 times in the text. In some domains “abstract” is an adjective and would be a noise word, but because I am familiar with the document, I know that in this context it refers to the abstract of a thesis.

Another problem with my index is that it does contain bigrams or trigrams. There is an entry for “search” and an entry for “engine” but none for “search engine”. The same occurs for “inverse document frequency”. To adequately represent the about-ness of FOA the index must contain these terms.

In the end it is clear that the process of creating an index is ambiguous, it is not something that can be easily automated, and it requires knowledge of the domain. In order to create a satisfactory index I think it would be necessary to manually go through all the words a remove those that are deemed inconsequential.

3. The Web as a Super Expert System

Give a critical discussion of the following paragraph taken from Belew: Finding Out About, Section 6.0

Note the ease with which author-as-knowledge engineer can express their knowledge. Hypertext knowledge bases are accessible to every writer. In this view, hypertext solves the key AI problem of the KNOWLEDGE ACQUISITION BOTTLENECK, providing a knowledge representation language with ease, flexibility and expressiveness of natural language – by actually using natural language! The cost paid is the weakness of the inferences that can be made from a textual foundation: contrast the strong theorem-proving notions of inference of Section 6.5.1 with the many confounded associations which arise in Swanson's analysis of latent knowledge in Section 6.5.3

Expert Systems encode the knowledge and expertise of experts in a narrow, specialized domain. A crucial step in development of any expert system is acquisition of the expertise from the expert by the knowledge engineer. This process poses a major difficulty referred to as the knowledge acquisition bottleneck. Knowledge acquisition is inevitably a lengthy and complicated process that requires that the knowledge engineer glean expertise from the expert through conversation. This process is complicated by the fact that the knowledge engineers and the expert have skills specialized in very different domains and must find a common language in order to communicate. This requires that both learn about the domain of other in order to facilitate the process.

Once the knowledge has been acquired the knowledge engineer must convert the knowledge into a representation language that can be input into an expert system shell and used in reasoning. The encoding of knowledge into rules is no easy task. Expertise is any given domain can be inconsistent and imprecise. It can contain weak implications and missing data. The knowledge engineer has many tools to handle these characteristics (Bayesian reasoning, certainty factors, fuzzy reasoning) but it remains a vague practice.

With the advent of the Web the means of storing expert knowledge has dramatically changed. Experts are now able to represent their expertise in natural language. This removes the challenge of knowledge acquisition but it introduces a new set of problems to overcome. Hypertext and natural language, as a knowledge representation, provides very little structure for the expertise. This property restricts the ability to reason over the expertise with any strong inference methods. The sheer quantity of data available on the Web introduces dilemmas that did not exist in Expert

Systems. Expert systems deal with narrow, specialized domains and a knowledge engineer systematically inputs the data. With the web anybody can input data and the knowledge engineer is now concerned with Information Retrieval. IR is concerned with searching for meaningful data within a corpus of documents using probabilistic methods. This requires a means to assess the about-ness of any given document. This, like encoding information into rules in expert systems, is vague.

Natural language is imprecise and, because it is the knowledge representation of the Web, the information itself can be vague. Since experts are free to express their ideas in the format and language that is native to them the same idea can be represented in a multitude of formats. Thus it can be difficult to collate the ideas of many experts into a larger body of expertise that is traceable. Expert systems provides a very structured framework in which the expertise of many experts in a standardized format. This allowed for the integration of the knowledge of many experts in a manner that allows one to draw on the information as a whole.

Traditional expert systems provided a controlled environment for expertise. The authority of the information contained in the expert system was known from the onset. With the Web as a Super Expert System is the issue of the authority of the information becomes a concern. Because anyone can upload information onto the web it is difficult to assess whether the information retrieved is in fact expert knowledge.

The main strength of expert systems is the ability to apply strong inference procedure over expert knowledge. Currently the Web does not provide this functionality.

The development of the Semantic Web addresses many of the shortcomings of the current web when considered as a Super Expert System. The Semantic Web initiative seeks to provide a common framework for data on the Web. With the Semantic Web data is represented in both natural language and in a format that allows machine reasoning. This representation facilitates sharing and integration of information. The Semantic Web uses the Resource Description Framework and represent data as triples of subject-predicate-object. Description Logics are used to represent the semantics of web data. Description Logic "refers, on the one hand, to concept descriptions used to describe a domain and, on the other hand, to the logic-based semantics which can be given by a translation into first-order predicate logic" (Wikimedia). This

representation is thus an unambiguous formal representation that allows inference on the data.

References

Belew, R.K. Finding Out About.Cambridge Univ. Press, 2000

"Description logic." *Wikipedia, The Free Encyclopedia*. 28 Nov 2007, 14:01 UTC. Wikimedia Foundation, Inc. 13 Dec 2007 <http://en.wikipedia.org/w/index.php?title=Description_logic&oldid=174369527>.

Kilgrarriff, Adam. "BNC database and word frequency lists". 3 Nov 1998. British National Corpus (BNC). 13 Dec 2007<http://www.kilgarriff.co.uk/bnc-readme.html>