

Generating Hierarchical Summaries for Web Searches

Dawn J. Lawrie and W. Bruce Croft
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

ABSTRACT

Hierarchies provide a means of organizing, summarizing and accessing information. We describe a method for automatically generating hierarchies from small collections of text, and then apply this technique to summarizing the documents retrieved by a search engine.

Categories and Subject Descriptors

H.5 [Information Systems]: Information Interfaces and Presentations

General Terms

Design

1. INTRODUCTION

A topic hierarchy is a description of a body of text which both summarizes the text and provides a method of navigating through it. The popularity of hierarchies as a method of organization indicates that this type of summary is relatively easy to understand. For example, the Yahoo hierarchies[7] and MeSH headings[4] have been used for many years. In contrast to clustering algorithms, the hierarchy is comprised of monothetic groups of documents, which are described by the feature(s) that the group has in common. Such groupings are easy for a user to understand because a single word or phrase is used to describe the feature.

In order to construct the hierarchies, we build statistical models of language to identify topic terms in a document set. These statistical models can be built recursively to identify subtopics of a main topic, thus creating a hierarchy that summarizes the collection. Documents are attached to the hierarchy if they include the topic terms, thereby creating a means of navigating the collection. When searching for web documents, hierarchies can give the user an alternative to the organization provided by a ranked list.

2. EXAMPLES

Search engines have become very good at ranking relevant web-pages for certain types of queries. Unlike 5 to 10 years ago, when a search for a company web page would often rank personal home pages above a company's own page, today the company page is likely to be the top ranked page. However, there are other types of queries where web search engines do not perform nearly as well. This occurs frequently when the query is a search for general information on a topic. Such topical searches are where hierarchies can help a user the most. The hierarchy provides an alternative to

browsing a ranked list, and can be a more effective organization of the documents that were retrieved, thus allowing users to find the relevant information.

Figure 1 shows an example hierarchy generated from the snippets of documents retrieved by the Google search engine[1] for the query: "Hubble Telescope Achievements". The top part of the figure is a portion of the hierarchy that the user would see when interacting with our system. The ranked list that follows is the portion of the ranked list that would be seen by navigating to "achievements→Hubble→Hubble's achievements" and selecting the topic "Hubble's achievement". The order of the documents is the same order that would be seen in the ranked list and the rank numbers correspond to the ranks assigned by the search engine.

This topic includes eleven documents, although only the first six appear in the figure. Of the eleven documents, three of them are about Hubble telescope achievements including the ones shown in Figure 1 that rank at 100 and 140. Because of the sites' low ranks, it is unlikely a user would have found them without using a tool such as the hierarchy.

3. FRAMEWORK FOR HIERARCHIES

The main challenge to creating topic hierarchies is selecting the correct terms that will accurately describe the document set to the user. We propose that the best topical summarization terms are those which are both about the topic and predictive of other terms. We model the set of topical summary terms T as the maximization of the joint probability of topicality and predictiveness, given by the Term Selection Formula[2]:

$$\arg \max_T \mathcal{P}(A_T, B)$$

where A_T refers to topicality with respect to topic T and B refers to predictiveness.

Topical terms are the content-bearing terms with respect to a particular topic. These are not necessarily the most frequent terms, since they may only be mentioned once in a document; however, the information conveyed to the reader by content-bearing terms is crucial to the reader's understanding of the document. Predictive terms are those whose occurrence is a precondition for many other terms. This quality captures the presence of subtopics that are associated with a topic. Previous work on topic hierarchies [6] has shown that this is an important aspect of topic terms, because these are the terms that are frequently used to discuss the topic at different levels of generality. Included in this set of terms are most stopwords, which due to their frequent occurrence in a document make most terms dependent on them. Also in the set of predictive terms are general terms related to a topic. For example, in a retrieved set about Endangered Mammals, "endangered species" is likely to be a predictive term because many other words only occur when it is present.

TREC Query 303: Hubble Telescope Achievements

Hubble Space Telescope - 363	Hubble - 553	Hubble's achievements - 11
Hubble telescope - 249	space - 104	Latest Hubble Telescope - 2
telescope - 231	Measuring Hubble's constant - 4	Space Telescope - 19
achievements - 565	memorable achievements - 6	Hubble Telescope Achievements - 3
	NASA's Hubble Space Telescope - 7	crowning intellectual achievements - 2
	Hubble Space Telescope Science - 4	Generation Space Telescope - 4
	pioneering achievements - 4	scientific achievements - 7
	deep space - 7	major achievements - 5
	Hubble Space - 2	radio telescope - 4
		technical achievements - 4

achievements→Hubble→Hubble's achievements

Rank 19: *Books with Pictures From Space (Science U)*

... of Our Cosmos, by Simon Goodwin A gallery of the most significant photographs taken by the Hubble telescope explains what Hubble's achievements can ...

Rank 34: *Amazon.COM: buying info: Hubble's Universe: A Portrait of Our ...*

... Ingram A gallery of the most significant photographs of space as taken by the Hubble telescope explains what Hubble's achievements can tell us about the ...

Rank 63: *ESA Portal - Press Releases - HST's 10th anniversary, ESA and ...*

... A public conference will take place in the afternoon to celebrate Hubble's achievements midway through its ... Notes for editors. The Hubble Space Telescope ...

Rank 100: *FirstScience.com - The Hubble Decade*

... astronauts' first view of the Earth from the Moon - and the Hubble Space Telescope's ... View from the top. On the scientific front, Hubble's achievements ...

Rank 111: *The Hindu : Discoverer of expanding universe*

... Hubble's achievements were recognised during his lifetime by the many honours conferred upon him In 1948 he was elected an ... 'The Hubble Space Telescope ...

Rank 140: *HubbleSite — Science*

... farther and sharper than any optical/ultraviolet/infrared telescope ... very specific goal (like the Cosmic Background Explorer), Hubble's achievements ...

Figure 1: The figure shows a portion of the hierarchy created for the TREC Topic 303: Hubble Telescope Achievements, created from snippets of documents retrieved by Google. The portion of the ranked list displayed corresponds to documents that contain the terms “achievements”, “Hubble”, and “Hubble’s achievements”. The snippets indicated that all of the documents have to do with the Hubble Telescope, but one of the best sites for finding out about the Hubble Telescope’s achievements is the HubbleSite — Science, ranked 140th by Google.

By combining these two properties, one finds a set of words that will maximize a user’s understanding of the information contained in the documents. Since there are some terms that are predictive and not topical (i.e. stopwords), and other terms that are topical but not predictive (i.e. single occurrences of content-bearing terms), we assume the two probabilities are independent, and so $\mathcal{P}(A_{\mathcal{T}}, B) = \mathcal{P}(A_{\mathcal{T}})\mathcal{P}(B)$. We use statistical language models[5] to estimate $\mathcal{P}(A_{\mathcal{T}})$, topicality, and $\mathcal{P}(B)$, predictiveness. In order to estimate topicality, a unigram language model is computed, and then the Kullback-Leibler divergence contribution of each term is calculated. In order to estimate predictiveness, a language model is necessary to show how terms relate to each other.

4. EVALUATION

We use two different metrics to evaluate the hierarchies. The first evaluation measures how good a summary the hierarchy is. The second measures the percentage of documents a user can find. For our evaluations, we created a test bed of 50 different hierarchies for the TREC Topics 301 to 350. We used the TREC Topic titles as queries to retrieve web documents using the Google Search Engine. The text used to create the hierarchies came from up to 1000 snippets.

Since the hierarchy is intended to be viewed as a summary, it is important to determine how well it summarizes the text in the documents. This can be done using automated techniques because the hierarchy is a predictive summary, which means the terms that occur in the hierarchies should predict the text. If one were to remove the structure of the hierarchy, a bag-of-words is all that would remain. By treating the documents as a bag-of-words, we can compare the distribution of terms found in the hierarchy to the distribution of all terms. To do this we calculate EMIM, which measures the extent to which the distributions of the two sets deviate from

stochastic independence as described in Lawrie and Croft[3]. The greater the dependence between the two distributions, the better the hierarchy is a summary of the text. We use this evaluation to compare the choices of words made by our algorithm to the top ranked TF.IDF words. We found that the topics selected for the hierarchy were significantly better at summarizing the documents than the top TF.IDF terms.

Another important attribute of the hierarchy is the ability to find documents within it, which we refer to as the reachability of the hierarchy. Because of the statistical nature of the hierarchies, there is no guarantee that there will exist a path to all documents used to create the hierarchy. This evaluation imposes different cut-offs as the maximum size group a user would explore, and then calculates the percent of documents that are reachable. We compared the effects of different policies with hierarchies and with ranked lists. Specifically, we compared the percentage of documents that can be discovered with different policies for both the hierarchy and the ranked list. The performance of a hierarchy is dependent on the maximum size of the levels in the hierarchy. As expected, the hierarchy with the most number of topics in a level allows access to the most number of documents. Also, the policy that looks at the largest size document groups accesses the most number of distinct documents. A policy of looking at document groups no larger than 20 allows access to 52.6% of the documents with a standard deviation of 11.6% over all queries for hierarchies with a level of size 20. This is equivalent to examining the top 400 in the ranked list. With levels of 5, the average number of documents found is 18.8% with a standard deviation of 9.8%. Even with only 5 topics per level of the hierarchy a user can reach about 150 documents, given that an average number of documents are returned by the search engine. This number of documents is still more than the typical user is likely to see using a ranked list.

5. CONCLUSION

This paper briefly describes our design of hierarchies for web documents. In our preliminary evaluation, we show examples of the usefulness of the hierarchies. We show that the terms selected to be part of the hierarchy are better summary terms than the top TF.IDF terms, and that the hierarchy provides users with more access to the documents retrieved than using a ranked list alone.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

7. REFERENCES

- [1] GOOGLE. Google. www.google.com.
- [2] D. Lawrie. *Language Models for Hierarchical Summarization*. PhD thesis, University of Massachusetts, Amherst, to appear.
- [3] D. Lawrie and W. Croft. Finding topic words for hierarchical summarization. In *Proceedings on the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 349–357, 2001.
- [4] H. Lowe and G. Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Journal of the American Medical Association*, 271(4):1103–1108, 1994.
- [5] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
- [6] M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 206–213, 1999.
- [7] YAHOO. Yahoo. www.yahoo.com.