SAND2005-7449C

Finding Relevant Documents in Enormous Clustered Sets of Scientific Documents

Ron Zacharski‡, Jim Cowie‡, George Davidson† and Kevin W. Boyack†

† Sandia National Laboratories — ‡ New Mexico State University
Sandia National Laboratories, P.O. Box 5800, Albuquerque, New Mexico 87185
New Mexico State University, Las Cruces, New Mexico 88003
{raz, jcowie}@crl.nmsu.edu {gsdavid, kboyack}@sandia.gov

Abstract

When people use an online search engine to find relevant information, often tens of thousands of documents are returned as a result of a single query. For example, when searching for *carbon fixation bacteria* in scholar.google.com over 18,000 documents were returned. Finding germane information in such a large collection of documents is difficult. This paper discusses ways of presenting summarizing views of enormous clustered document sets as a way of helping a person find relevant information. This approach is based on identifying significant words and phrases of a cluster using log likelihood. These significant words and phrases are used in two ways. First, they are highlighted in a visual thumbnail of the documents in a cluster, which enables people to see if a cluster is relevant by quickly viewing the significant words and their distributions. The significant words and phrases are also used to produce a summary of the cluster. This is done by ranking the sentences in a cluster based on the occurrence of these significant terms. The summary is produced by concatenating the highest ranked sentence during each iteration of the ranking process. While this approach was initially developed to summarize clusters of scientific abstracts, it is domain independent.

Introduction

Summarization deals with extracting the most important information from a single document or a collection of documents. The summary produced can be of varying lengths. For example, the title of a research paper can be viewed as an extremely brief summary of an article as shown in example 1 (the results returned from scholar.google.com on the query *carbon fixation bacteria*).

(1) a. Nitrogen fixation in seagrass meadows: Regulation, plant-bacteria interactions and significance to primary productivity

b. Nitrogen-fixation by cyanobacteria associated with Codium fragile (Chlorophyta): Environmental effects and transfer of fixed nitrogen.

c. Microbial microstratification, inorganic carbon photoassimilation and dark carbon fixation at the chemocline of the meromictic Lake Cadagno (Switzerland) and its relevance to the food web.

As can be seen from these examples, titles typically provide a high-quality brief summary of a paper and researchers can quickly scan a short list of titles to quickly identify articles of interest. Another familiar summary is the abstract of a research paper:

(2) The rhizosphere sediments of seagrasses are generally a site of intense nitrogen fixation activity and this can provide a significant source of "new" nitrogen for the growth of the plants. In this paper, I review the data concerning nitrogen fixation in seagrass ecosystems, the of the fixed nitrogen from the bacteria to the plants and its contribution to the overall productivity of seagrasses in different climatic zones.

Abstracts can help a person determine whether the associated article is relevant to that person's interests. While these methods of title skimming and reading abstracts are invaluable, they are not sufficient when faced with finding relevant documents in an enormous collection. For example, searching on *carbon fixation bacteria* in scholar.google.com produces 18,600 results. Even if the research paper titles of the result list were maximally informative, it would take a person an extremely long time to find all the relevant articles. Our research focuses on making this task more manageable. We organize the results into a set of clusters of related documents (Klavans & Boyack, 2006). We then provide the user with different summarizing views of these clusters, ranging from providing a multi-sentence summary of the cluster to graphically illustrating the distribution of words and phrases that make that cluster unique.

A number of factors make this task more difficult than summarizing single documents. The ratio between the size of the initial cluster of documents and the summary produced (the “compression ratio”) for a multi-document

summary is typically significantly smaller than that for a single document summary. Many single-document summarization systems have a compression ratio of 20-25% and the SUMMAC evaluation examined 10% compression summaries (TIPSTER 1998). By comparison, the largest multi-document summaries our system produces have less than a 5% compression ratio. It is more difficult to generate more compressed summaries.

Much of the previous work on multi-document summarization (for example, McKeown and Radev 1999; Goldstein et al. 2001A; D'avanzo et al 2004) has examined the summarization of news articles. One property of a large collection of news articles covering the same event is that there is a high level redundancy in content as (3)-(6) show:

(3) The US spy agency CIA is operating a covert prison system covering eight countries for holding terror suspects, The Washington Post reported Wednesday (Xinhua)

(4) The CIA has been holding and interrogating al-Qaeda captives at a secret facility in Eastern Europe, part of a covert prison system established after the September 11, 2001, attacks, The Washington Post reports. (The Sydney Morning Herald)

(5) The CIA is holding al-Qaida suspects at secret prisons called "black sites" in several Eastern European countries,The Washington Post reported Wednesday. (Science Daily)

(6) A fascinating story in today's Washington Post reveals that the CIA has been maintaining a "hidden global internment network" for the last four years in Thailand, Afghanistan, and some Eastern European countries for the purpose of hiding and interrogating terrorism captives. (The Nashville Scene)

This redundancy can be used to identify important information that should be included in a summary (see, for example, Goldstein 2000a, 2000b). However, redundancy is also a problem in that redundant information should not be included in a summary—sentences (3)-(6) should not be included in the same summary.[1] However, in research articles, the redundancy is not as transparent as that in news articles. News articles often report on the exact same event—possibly from different perspectives. While this type of redundancy does occur in research papers—for example, when the same group of authors publish multiple papers on a particular research effort—it is more common for research papers to highlight and focus on new research while redundant information is treated as background. This makes identifying and exploiting redundant information more difficult.

[1] For approaches to anti-redundancy see Goldstein 2000a, 2000b; Saggion and Gaizauskas 2004.

Related Work

Research on automatic summarization dates back to Luhn's seminal work at IBM in the 1950's (1958). Luhn developed a system that produced summaries by sentence concatenation. Sentences were selected based on the number of significant words they contained. Since that time a large amount of research has been conducted on single-document summarization (see, for example, Mani 2001, and Mani and Maybury 1999). Much of this work on single document summarization has been influenced by research in understanding how people summarize documents. For example, Cremins (1996) found that professional abstractors generally do not seek to fully understand articles they are summarizing and instead rely on surface features such as the position of the sentence in the paragraph, the location of key words, and section headings. All these cues are also used in automatic summarization. For example, Edmundson (1968) developed a summarization system that exploits the rubric that topic sentences generally occur at the beginnings and ends of articles and also occur early in a paragraph. This heuristic has been used in a large number of systems since Edmundson's initial work. In addition, many systems attempt to identify topic or relevant sentences by first locating key words. For example, Kuipec et al. (1999) use the notion of thematic words—identified as the most frequent content words—as a means for scoring sentences. Finally, the use of section headings as a cue to identifying relevant sentences to be included in a summary has been investigated by Teufel and Moens (1999). They use a list of prototypical headers (introduction, discussion, conclusions) to assign a rhetorical feature to each sentence in the article to be summarized.

While there has been significant research on single document summarization, there has been considerably less research on multi-document summarization. McKeown and Radev (1999) describe a system that generates summaries of multiple news articles. The system makes use of pre-existing message understanding systems, which use a set of pre-defined templates to extract information from a corpus (MUC 1992). As an example of such a template consider a terrorist act template that might include fields such as perpetrator, victims, location, and type of event. The message understanding system then uses these templates to produce a set of filled-in forms representing information extracted from the text. McKeown and Radev use these forms as input to their summarization system; the output is a paragraph summary. The system uses a number of heuristics including one that information extracted from multiple documents is more informative than information extracted from just one document. Another example of a heuristic is the agreement heuristic—when 2 news sources agree on a fact, that agreement is explicitly mentioned. An example of the summary produced by this heuristic is shown in (7).

(7) The morning of March 1st 1994. UPI reported that a man was kidnapped in the Bronx. Later, this

was confirmed by Reuters.

These heuristics, or planning operators, produce good quality summaries of news articles. McKeown and Radev suggest that this summarizer—hand-crafted to the domain of news articles—produces superior summaries compared with summarizers based on statistical methods. However, this approach does require manual development of heuristics for each domain.

One technique for multi-document summarization is to run a summarizer on each document in the cluster then combine these summaries. However, if the cluster of documents were retrieved by a particular search term, then there may be a large amount of redundancy in the documents within the cluster (see examples (3)-(6) above) and a summary created from the combining of single document summaries would also be redundant. Goldstein et al. (2000a, 2000b) have proposed multi-document summarization techniques that reduce this redundancy by using a measure of relevant novelty. The approximation to relevant novelty is maximal marginal relevance or MMR (Carbonell and Goldstein 1998). In this approach each document in the cluster is divided into passages. For each passage, an MMR score is obtained. The set of passages with the highest MMR is retained as the summary. The rough idea of the MMR is

MMR = score – penalty

where score includes the cosine similarity metric (Singhal 2001) for query and document and information content, and penalty includes the similarity between the passage and previously selected passages among other factors. Their approach differs from that of McKeown and Radev mentioned above in that it is completely domain independent. Schiffman et al. (2002) describe a multi-document summarizer that produces summaries by extracting the top ranked sentences. This summarizer makes use of information about the structure of news articles. For example, lead sentences of news articles are often brief summaries of the entire article. Another feature the summarizer uses is verb specificity (roughly how informative the verb is—for example, *arrest* has a higher verb specificity score than *be* or *do*). It also makes use of Wordnet Synsets. Rather than measuring aboutness by examining the frequency of words, they use Wordnet to classify words into concepts and then measure the frequency of concepts. In addition to verb specificity and concepts, their system uses a number of heuristics based on other features such as location in document, length of sentences (sentences below 15 words and above 30 are penalized), and pronoun initial sentences. Schiffman et al. note that the application of specific heuristics needs to be based on the cohesiveness of the cluster of documents and also note that correctly assigning words to concepts is difficult.

Lin and Hovy (2001) present a summarizer that is most similar to our approach. However, it is specifically tailored to news article summarization. In their system, they first identify what they call unigram, bigram, and trigram topic signatures (Lin and Hovy 2000, Hovy and Lin 1999) for each cluster of articles using log likelihood (Dunning 1993). These topic signatures are then constructed into a query, which is then used to perform sentence-level information retrieval. Information retrieval returns an ordered set of sentences. The top sentences are then used to create a summary. One novel idea that Lin and Hovy have is to pair each extracted sentence with the initial sentence of the source article of the extracted sentence. The authors thought that this created a more coherent text.

Multidocument Summarization

When a user types in a query to scholar.google.com or another search engine, potentially thousands of documents could be included in the result set. Our interest is in investigating methods of presenting different summarizing views that will assist a user in finding relevant research articles within an enormous set of potentially relevant articles. For example, one method is to cluster related documents in that set and then present concise information about each of those clusters as shown in Figure 1. Part of that view is a bar representing how cohesive each cluster of documents is—the longer the bar the more cohesive, or the more alike, the documents in that cluster are.

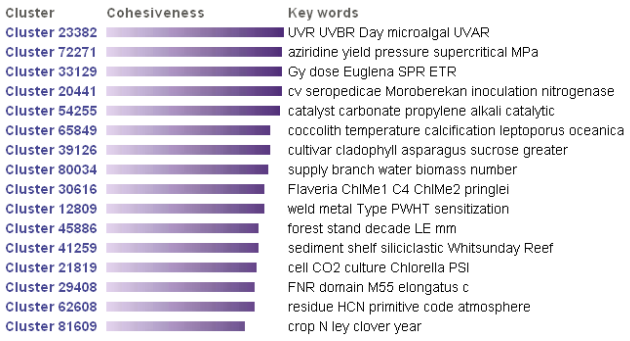

Figure 1: Cohesiveness and keyword view

We hope that this view will assist users in determining whether all documents in a cluster are about the same topic or about different topics. The assumption we make is that if they are about the same topic they would share the same vocabulary, and moreover, the frequency of specific words would be similar. We measure this similarity by using information radius (Dagan et al. 1997).1 Manning and Schütze (2000) describe this measure as answering the question: How much information is lost if we describe the two random variables with their average distribution? First, we compute a word probability vector for all the words in the entire cluster, and a probability word vector for each document. We reduce the size of these vectors by eliminating approximately 150 of the most common English words. First we compute Kulback-Leiber divergence measure where p is the frequency of the word in the document and q is the frequency of that word in the cluster

$$D(p \| q) = \sum_i p_i \log \frac{p_i}{q_i}$$

We do not use KL divergence directly because of several practical problems (Manning and Schütze 1999). One is that if $q_i = 0$ and p_i is not zero then the KL divergence would be ∞. Another problem is that KL divergence is asymmetric. Information radius is defined as:

$$IRad = D\left(p \| \frac{p+q}{2}\right) + D\left(q \| \frac{p+q}{2}\right)$$

Key words

Figure 1 also shows the significant words for each cluster. Significant words are determined by using log likelihood (Dunning 1993). We compute log likelihood as follows. Let w_iC be the number of occurrences of word w_i in the cluster; w_iT the number of occurrences of the word in the reference collection, nC the total number of words in the cluster and nT the number of words in the reference collection. Then the likelihood ratio for that word is

$$e1 = \frac{nC \cdot w_iC}{nC + nT}$$

$$e2 = \frac{nT \cdot w_iC}{nC + nT}$$

$$loglikelihood = 2 \cdot \left(w_iC \log \frac{w_iC}{e1} + w_iT \log \frac{w_iT}{e2}\right)$$

We compute the log likelihood ratio for every word in the cluster and select the 5 words with the highest likelihood value. For example, clicking on a cluster in the view shown in Figure 1, results in a view that shows the distribution of these words in the documents of that cluster, shown here in Figure 2.

Figure 2: Thumbnail view of relevant word distribution

The words are color coded. So, for example, in Figure 2 the word *root* is shown in red and occurrences of that word show up highlighted in red in the document thumbnails. Ogden and Davis (2000) have shown that users are better able to gauge the relevance of a document to their search query using these relevant word thumbnails compared with seeing the titles of the documents.

Key phrases

In addition to showing this thumbnail view of the distribution of relevant words, the distribution of key phrases can also be shown. We evaluated a number of techniques to identifying the key phrases of a cluster. Initially, we directly extracted collocates from a cluster by using log likelihood. The problem with that approach was that the size of a cluster (in number of words) was too small to obtain meaningful results. The method we did employ determines a set of common collocates from a larger reference collection and then identifies the common collocates of a cluster based on frequency. We identify the collocations in the reference collection using log likelihood. The documents in the reference collection were obtained indirectly through a term search on the Science Citation Index. Thus, we have identified collocations that are specific to our domain. A sample of some of the top collocations (and their likelihood values) is shown below:

2629.6	leaf area
2246.8	electron transport
2103.6	amino acid
1965.1	atmospheric CO2
856.5	gas exchange
1827.8	CO2 enrichment
1484.7	e coli
1448.1	CO2 concentration
1086.5	ambient CO2
825.9	unit-cell parameters
804.3	gene expression
780.0	n supply

After the collocations are determined, we count the number of times these collocations occur in a particular cluster. The five most frequently occurring collocations for that cluster are returned as key phrases of the cluster. An example is shown in Figure 3.

Figure 3: Thumbnail showing key phrases

Summaries

Summaries are produced based on the key words and key phrases described above. We rate each sentence in a cluster based on the number of keywords and phrases it contains. The relative value of words and phrases are as follows (from highest to lowest)

1. the first occurrence of a specific key phrase
2. the first occurrence of a specific key word
3. subsequent occurrences of the key phrase
4. subsequent occurrences of the key word.

Sentences are then ordered based on these scores and the highest scoring sentence is selected to be included in the summary. The remaining sentences are re-scored with words and phrases not appearing in the first sentence receiving a higher score than ones that did appear. The idea behind this is a follows. Suppose we have identified *elevated CO2*, *CO2 enrichment*, and *atmospheric CO2* as key phrases for a particular cluster. Suppose that the first sentence of a summary is

(8) The larger biomass accumulation of Q. suber under elevated CO2 is attributable to a higher availability of CO2 coupled to a larger leaf area, with no significant decrease in photosynthetic capacity under CO2 enrichment and elevated N fertilization.

Since *elevated CO2* and *CO2 enrichment* both occur in this sentence adding a sentence to the summary that contains these terms may be less informative than adding a sentence containing a different key phrase such as atmospheric CO2. Example (9) shows an example of the summary produced:

(9) In rainfed ecologies, where deep roots contribute to enhanced drought resistance in rice, the results indicate the possibility of combining drought resistance with higher levels of grain yield. As part of a research programme aimed at using molecular marker technology for the improvement of drought resistance in rice, it is necessary to identify quantitative trait loci (QTLs) associated with root morphology and other drought resistance-related traits. QTLs have been reported for a number of traits potentially related to performance under water deficit, such as improved root morphology and osmotic adjustment. The identification of quantitative trait loci (QTLs) associated with root morphology and other drought resistance-related traits should help breeders produce more drought resistant varieties. The results demonstrate the importance of phenotyping environment and suggest prospects for selection of QTLs for deep root morphology, root thickness, and vigorous seedling growth under anaerobic conditions to improve the constitutive root system of rainfed lowland rice.

Application to Large Document Maps

Sandia National Laboratories has generated clusters of papers in maps containing nearly 1 million scientific documents (Klavans & Boyack, 2006). Each cluster contains an average of 9 papers, with a maximum of about 100 papers per cluster. Although the clusters are formed using citation characteristics, they are very topic focused, and thus are good candidates for multi-document summarization techniques. The examples shown in Figures 1-3 all come from the 2002 map of science.

The keyword, key phrase, and summary forming methods described here have been implemented in two different ways. First, they have been coded as an add-on to the VxInsight visualization software package (Davidson et al, 1998) to summarize local clusters of documents that are being visualized. Second, they have been implemented as scripts that perform these functions sequentially on large numbers (~100,000) of clusters of documents in a batch mode. The output is saved and entered in a database so that descriptions of clusters, rather than documents, can be searched and displayed. Spot-checking of the results of the summarization show that the summaries and key phrases provide good working descriptions of the clusters of documents that are pertinent to the research questions addressed by the clusters of papers.

Acknowledgements

This work was funded in part by the US Department of Energy's Genomes to Life Program under project 'Carbon Sequestration in *Synechococcus* sp: From Molecular Machines to Hierarchical Modeling' (http://www.genomes-to-life.org).

References

Baker, L. Douglas and Andrew K. McCallum. 1998. Distributional clustering of words for text classification. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 96-103.

Klavans, Richard and Kevin W. Boyack. 2006. Quantitative evaluation of large maps of science. Scientometrics, in press.

Bruining, Leon. 2005. Connections between the C, N, and P biogeochemical cycles of system Earth and their relevance for environmental problems. Research Report. Institute of Environmental Sciences, Van Steenisgebouw.

Carbonell, Jaime G. and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reorderings of

documents and producing summaries. Proceedings of SIGIR-98, Melbourne.

Cremins, E. T. 1996. The art of abstracting. Information Resources Press.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. The 35th Annual Conference of the Association for Computational Linguistics 35.56-63.

D'avanzo, Ernesto, B. Magnini, and A. Vallin. 2004. Keyphrase extraction for summarization purposes: The Lake system at DUC-2004. HLT/NAACL Annual Meeting.

Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E., & Wylie, B. N. (1998). Knowledge mining with VxInsight: Discovery through interaction. Journal of Intelligent Information Systems, 11(3), 259-285.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19.61-74.

Edmundson, H. P. 1968. New methods in automatic extraction. Journal of the ACM 16.264-285.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Jamie Callen. 2000a. Creating and evaluating multi-document sentence extraction summaries. CIKM, 165-172.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000b. Multi-document summarization by sentence extraction. Automatic Summarization: ANLP/NAACL 2000 Workshop, ed. by U Hahn, C. Lin, I. Mani, and D. Radev. 40-48.

Hovy, Eduard and Chin-Yew Lin. 2001. NEATS: Automated text summarization in SUMMARIST. Advances in Automatic Text Summarization, ed. by M. Maybury and I. Mani. Cambridge: MIT Press.

Kupiec, Julian, Jan Pederson, and Francine Chen. A trainable document summarizer. Advances in Automatic Text Summarization, ed. by Inderjeet Mani and Mark T. Maybury. Cambridge: MIT Press.

Lee, Lillian. 2001. On the effectiveness of the skew divergence for statistical language analysis. Artificial Intelligence and Statistics 65-72.

Lin, Chin-Yew, and Eduard Hovy. 2001. NEATS: A multidocument summarizer.

Lin, Chin-Yew, and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. Proceedings of the COLING conference.

Luhn, Hans P. 1958. The automatic creation of literature abstracts. IBM Journal 159-165.

Mani, Inderjeet. 2001. Automatic summarization. Amsterdam: John Benjamins.

Mani, Inderjeet, and Mark T. Maybury. 1999. Advances in automatic text summarization. Cambridge: MIT Press.

Manning, Christopher D. and Hinrich Schütze. 1999. Foundations of statistical natural language processing. Cambridge: MIT Press.

McKeown, Kathleen, and Dragomir R. Radev. 1999. Generating summaries of multiple news articles. Advances in Automatic Text Summarization, ed. by Inderjeet Mani and Mark T. Maybury. Cambridge, MA: The MIT Press, 381-390.

MUC. 1992. Proceedings of the Fourth Message Understanding Conference. DARPA Software and Intelligent Systems Technology Office.

Ogden, William C. and Mark W. Davis. 2000. Improving Cross-Language Text Retrieval with Human Interactions. Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33), January 4-7, 2000. Maui, Hawaii (CD-ROM).

Rau, L., R. Brandow, and K. Mitze. 1994. Domain-independent summarization of news. Summarizing Text for Intelligent Communication, 71-75.

Saggion, Horacio and Robert Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. Papers from the Document Understanding Workshop, Boston, Massachusetts.

Schiffman, Barry, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. HLT 2002.

Singhal, A.2001. Modern information retrieval: a brief overview. IEEE Data Engineering Bulletin 24(4).

Teufel, Simone, and Marc Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. Advances in Automatic Text Summarization, ed. by Inderjeet Mani and Mark T. Maybury. 155-171. Cambridge: MIT Press.

TIPSTER. 1998. Tipster text phase 3 18-month workshop notes. May, Fairfax, VA.