Automatic Text Browsing Using Vector Space Model

Amit Singhal* Gerard Salton

Department of Computer Science
Cornell University, Ithaca, NY 14853
{singhal, gs}@cs.cornell.edu

Abstract

Vast amounts of text are now available in machine-readable form and can be processed electronically. The vector space model of text processing has been widely used and has consistently produced superior retrieval results for the last thirty years. Traditionally, information retrieval research has concentrated on improving on-demand retrieval of useful textual information. Often a user has no particular information need; instead, he is just interested in *browsing* the available information. Most hypertext systems depend on manually created browsing structures to support text browsing. Manual text linking does not scale up to the amount of electronic information accessible in today's networked computing environment. The vector space model can be used to automatically impose a flexible structure on arbitrary texts. Tools can then be provided for effective browsing of this structured text.

1 Introduction

Advances in optical storage techniques have placed massive amounts of information at our fingertips through CD-ROMs. This explosion of on-line information has made it virtually impossible to find useful items when needed. The research community in the field of information retrieval has substantially improved the on-demand retrieval of textual information in last forty years.[12, 16] With today's information retrieval systems, it is possible to search through several gigabytes of information in just a few seconds.

Traditional approaches to information retrieval help in locating potentially useful items (say several newspaper articles). The burden of reading these articles falls completely upon the user. Ideally, to reduce this information overload, it is desirable to have automatic mechanisms for text summary generation. Since automatic text summarization in unrestricted domains is beyond the current state-of-art, text extraction techniques, based on identification of important pieces within the text and displaying of few *good* pieces to a user, are usually employed in general semantic domain.[8, 6, 9]

Techniques to assign importance to various pieces of a text are needed for successful text extraction. This study concentrates on one such completely automatic technique that is used to assign a measure of goodness of a text piece. This technique is based on the relationship of a particular text piece to various other pieces in the text. Text relationships have been extensively studies by researchers in the field of Hypertext.[4, 3, 5] In a collection of texts (or text pieces), *links* are placed between related pairs of text to indicate relationship between the texts. In such environments, *the importance of a text can judged by the degree of its relatedness to other texts*, or the number of links involving this text.

In current hypertext systems, this linking between texts is done manually by the hypertext authors.

*This study was supported in part by the National Science Foundation under grant IRI 9300124

Manual link creation works well if we have little text, written by a few collaborating authors. In realistic circumstances, when huge amounts of text are present, it is impossible to create hypertext links manually.[2] In such situations, the vector space model of text retrieval can be used to place links between various pieces of text. This generates a rich text structure which can then be used for selective traversal of important pieces.[17, 18] This completely automatic process permits flexible text browsing with no manual intervention.

2 Vector Space Model

The vector space model of text processing [13, 11] is the most widely used information retrieval model. In this model, every information item – including the stored texts and any natural language information request – is stored as a set, or vector, of terms. These terms can, in theory, be selected from a controlled vocabulary, but in view of the difficulty of generating such vocabularies, the terms are, in practice, extracted from the text itself. Usually words are reduced to their morphological roots, or *word stems*, using a well-defined set of rules. Also, non-content words, or *stop words*, like *the, of, an,* ..., are removed from the text vector.

2.1 Term Weighting

Since different terms have different importance in a text, an importance indicator – the *term weight* – is associated with every term.[14, 15] Higher weights are assigned to more important terms. Typically a term that occurs frequently in a text is more important in the text than an infrequent term. This suggests that the number of occurrences of a term, often called the *term frequency* or *tf*, in a text be used as the term weight.

Term importance is not independent of the text corpus being searched. For example, the term *lake* might not be very useful when searching a collection of articles on Minnesota, whereas, it might be very important when searching a database of articles on various deserts of the world. This indicates that the more documents a term occurs in, the less important it may be. So an *inverse document frequency* or *idf* factor is incorporated into term weights. This factor is proportional to $\frac{1}{df}$ where the term occurs in *df* documents in the entire corpus.

Long and verbose documents usually use the same terms repeatedly. As a result, the term frequency factors may be large for long documents. Long documents also have numerous different terms. This increases the number of word matches between a query and a long document, unfairly, increasing the chances of retrieval over shorter documents. To compensate for this effect, *normalization* of term weights is often used. Normalization is a way of imposing some penalty on the term weights for longer documents. *Cosine normalization* is an effective normalization technique. Every document vector is divided by its Euclidean length, $\sqrt{w_1{}^2 + w_2{}^2 + \ldots + w_n{}^2}$, where w_i is the $tf \times idf$ weight of the i-th term in the document. This gives us the final weight for a term as:

$$\frac{tf \times idf\ weight}{Euclidean\ length\ of\ the\ document\ vector}$$

Terms absent from a document are considered to have a zero weight.

2.2 Document Ranking

A vector *inner-product* function can now be used to find the vocabulary overlap between any two text vectors. A natural language query entered by the user is converted into a weighted term vector, and using the inner-product function, a numeric similarity is computed between the query vector and the vector for every document in the collection.

If Q is the query vector and D_i is the vector representation for document-i, the similarity between the query and the document is computed as,

$$Sim(Q, D_i) = \sum_{common\ terms\ t_j} q_j \times d_{ij}$$

Where t_j is a term present in both the query and the document, q_j is the weight of a term t_j in the query

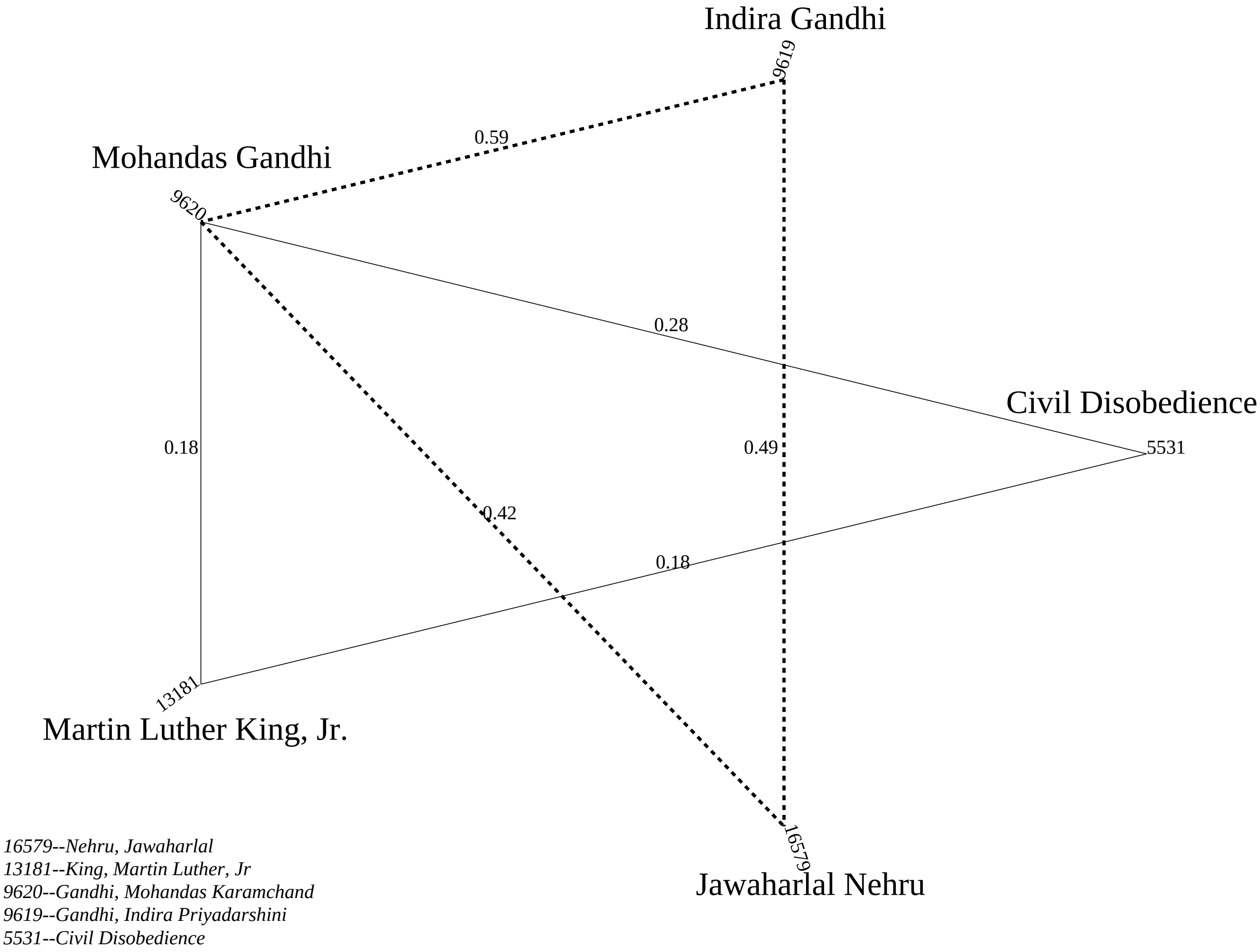

Figure 1: Relationship map for articles related to article Gandhi. The nodes are individual articles. The link weight (number printed on a link) indicates the inner-product vector similarity between two articles.

and d_{ij} is its weight in document-i. The summation is performed over all such terms t_j present in both the query and the document. The inner-product similarity generates a list of documents ranked by their potential usefulness.

2.3 The Smart System

The Smart system is a sophisticated text processing system developed over last thirty years.[10] Based on the vector space model, it has been the single most influential system in information retrieval research. Smart automatically identifies the terms to be associated with documents and generates weighted vectors for both the documents and a user query. Using the inner-product similarity, the system ranks documents in decreasing order of their *goodness* for the user query. All work reported in this study was done using the Smart system.

3 Text Decomposition and Structuring

Analysis of relationships between several texts yields rich connection patterns for the texts at hand. Such connection patterns can in turn be used to analyze the structure of the texts. Text structure can be used to identify the importance of a text in a collection of texts. An important text is related, often strongly,

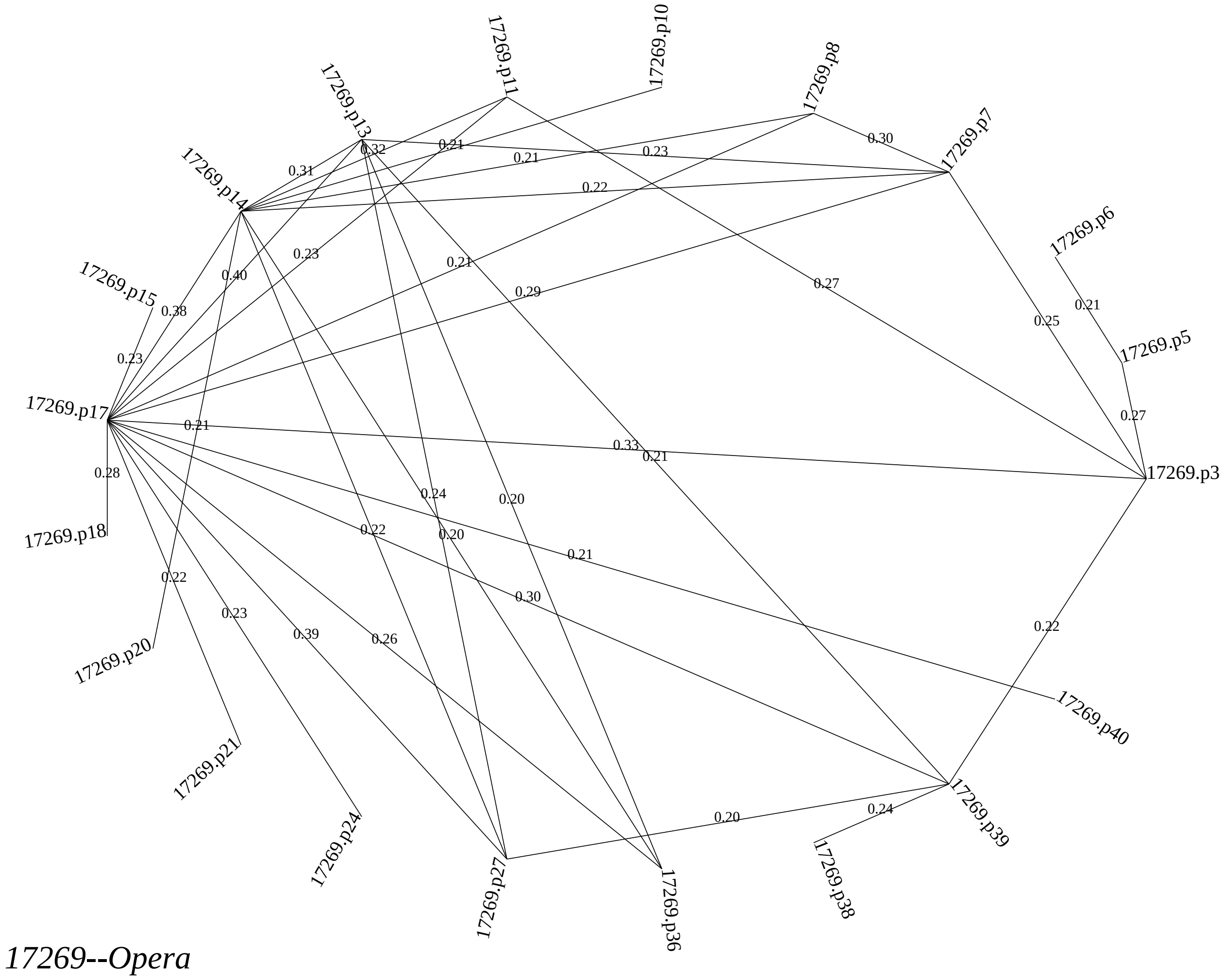

Figure 2: Relationship map for article Opera. The nodes are individual paragraphs and the links indicate the vector similarity between various paragraphs. Links below similarity 0.20 are considered to be weak and are dropped from the map.

to many other texts in the collection. Using document vectors and computing vector similarities, it is possible to obtain the relationships between various documents.[17, 1]

Figure 1 is a graphical view of these relationships for various encyclopedia[1] articles related to the article Gandhi (9620). This relationship map explicitly highlights the following facts:

- The article on Mohandas Gandhi is central to this collection of articles.
- Mohandas Gandhi, Dr. Martin Luther King, Jr., and Civil Disobedience are related to each other. Indeed, Dr. Martin Luther King adopted the principles of non-violent civil disobedience from the teachings of Mohandas Gandhi. This relationship is demonstrated by the following lines from the article on Dr. King:

> *...King's studies at Crozer and Boston led him to explore the works of the Indian nationalist Mohandas K. Gandhi, whose ideas became the core of his own philosophy of nonviolent protest. ...*
>
> *...On a visit to India in 1959 King was able to work out more clearly his under-*

[1]This work was done using the 29 volumes of Funk and Wagnalls encyclopedia.[7]

Paragraph 17269.p3
Opera, drama in which all or part of the dialogue is sung, and which contains instrumental overtures, interludes, and accompaniments. Types of musical theater closely related to opera include musical ...
Paragraph 17269.p13
In the late 17th and early 18th centuries, the infant German style of opera was overwhelmed by Italian opera. The most important German operatic center was Hamburg, ...
Paragraph 17269.p14
Italian opera was extremely popular in England. Nevertheless, two operas written about 1700 by English composers were frequently performed there: Venus and Adonis by John Blow (1649-1708) ...
Paragraph 17269.p17
Several composers in the mid-18th century tried to change operatic practices. They introduced forms other than the da capo for arias and made greater use of choral and instrumental music. The most ...

Figure 3: A 10% central extract for the article Opera. These are the four most central paragraphs in this article of 42 paragraphs. Such bushy extracts are good for highlighting the main topics discussed in a text.

> *standing of Satyagraha, Gandhi's principle of nonviolent persuasion, which he had determined to use as his main instrument of social protest. ...*

- Mohandas Gandhi, Jawaharlal Nehru and Indira Gandhi are mutually related. Nehru fought for the independence of India with Mohandas Gandhi and Indira, like Mohandas Gandhi, was a leader of the nation.

- A study of the relationship map of Figure 1 also shows that there is no relationship between Dr. King and Nehru or Indira Gandhi.

Based on text structure, several inferences can be drawn about the relationships between various texts. These relationships expose properties of text that would otherwise be unclear without reading the text.

The capability of generating weighted vectors for arbitrary texts also makes it possible to decompose individual documents into pieces and explore the relationships between these text pieces. Figure 2 shows a relationship map between paragraphs of the article Opera (17269). Figure 2 shows that several paragraphs of the article 17269 (*e.g.*, paragraphs 14 and 17) are more central to the article than other paragraphs (*e.g.*, paragraphs 6 or 10). Such insights can be used for picking only the "good" parts of the document to be presented to a reader.

4 Text Browsing

It is possible to gather a good deal of information about an article by just skimming the important parts of the article. Such skimming, or selective text traversal, helps one to cope with the large amounts of information present in today's information intensive world. It is possible to develop automatic techniques for selective traversal of automatically generated text relationship maps, thereby reducing the amount of text presented for reading and still capturing the essence of the document.[19]

One such possibility is to select just the central pieces of a text. The centrality of a piece can be judged by its linkage to other pieces of the text. In a relationship map, this refers to the bushiness, or the degree (number of links), of a node. Using this criterion, we can extract the desired number of central text pieces. Such an extract, a *bushy extract*, provides a nice overview of the article with very little text. Figure 3 is such an extract for Figure 2. Here only ten percent of the original text has been extracted. The selected paragraphs give a reasonable overview of the article in very little text.

To improve the coverage of the topics addressed by a document, it is possible to extract, in addition to the bushy nodes, the nodes most resembling the bushy nodes. Such an extract is generally longer than just the bushy extracts and it covers the topics of

Paragraph 17269.p3
Opera, drama in which all or part of the dialogue is sung, and which contains instrumental overtures, interludes, and accompaniments. Types of musical theater closely related to opera include musical ...
Paragraph 17269.p13
In the late 17th and early 18th centuries, the infant German style of opera was overwhelmed by Italian opera. The most important German operatic center was Hamburg, ...
Paragraph 17269.p14
Italian opera was extremely popular in England. Nevertheless, two operas written about 1700 by English composers were frequently performed there: Venus and Adonis by John Blow (1649-1708) ...
Paragraph 17269.p17
Several composers in the mid-18th century tried to change operatic practices. They introduced forms other than the da capo for arias and made greater use of choral and instrumental music. The most ...
Paragraph 17269.p20
France, Germany, and Italy developed characteristic operatic styles during the 19th century, all reflective of the romantic movement. Paris was the birthplace of grand opera ...
Paragraph 17269.p27
Carmen (1875) by the Frenchman Georges Bizet was viewed by the German philosopher Friedrich Nietzsche as bringing a Mediterranean clarity that dispelled "all the fog of the Wagnerian ideal." ...
Paragraph 17269.p36
The first American grand opera was Leonora (1845) by William Henry Fry (1813-64). Most subsequent works, such as The Scarlet Letter (1896) by Walter Damrosch, were European in ...
Paragraph 17269.p39
A growth in multimedia production techniques has paralleled the increase in electronic and synthesizer music by modern composers. Leonard Bernstein's Mass (1971), while not strictly an opera, ...

Figure 4: A 20% detailed extract for the article Opera. These are the four most bushy paragraphs in this article and four more paragraphs that exhibit high similarities with the bushy paragraphs.

the documents in greater detail. Figure 4 shows a more detailed extract of the article Opera (Figure 2) generated using this technique. Clearly, this slightly longer extract provides a more detailed view of the topics covered in the article.

5 Conclusion

Automatic text browsing can be used to remove the information overload posed by today's information systems. A flexible browsing structure can be imposed on arbitrary texts, using the vector space model. This eliminates the need for manual text structuring which becomes infeasible in any realistic information environment. Clever algorithms can be provided to selectively traverse such browsing structures and maximize the information given to a user in minimum possible text. The big advantage of this technique is its domain independence and its ability to do text skimming with no human intervention. Given the infeasibility of deep linguistic analysis needed to do abstract *generation* automatically in unrestricted domains, text *extraction* strategies proposed in this study are the only way to browse large amounts of text.

References

[1] J. Allan, Automatic Hypertext Construction, Doctoral Dissertation, Computer Science Department, Cornell University, November 1994.

[2] M. Bernstein, An apprentice that Discovers Hy-

pertext Links, in *Hypertexts: Concepts, Systems and Applications*, A. Rizk, N. Streitz, and J. Andre, editors, Proc. European Conference on Hypertext, Cambridge University Press, 212–223, November 1990.

[3] J.D. Bolter, *The Writing Space: The Computer, Hypertext, and the History of Writing*, Lawrence Erlbaum, Hillsdale, NJ, 1990.

[4] J. Conklin, Hypertext: An Introduction and Survey, *Computer*, 20:9, 17–41, September 1987.

[5] P. Delaney and G.P. Landow, editors, *Hypermedia and Literary Studies*, MIT Press, Cambridge, MA, 1991.

[6] H.P. Edmundson and R.E. Wyllys, Automatic Abstracting and Indexing — Survey and Recommendations, *Communications of the ACM*, 4:5, 226–234, May 1961.

[7] Funk and Wagnalls New Encyclopedia, 29 volumes, 25,000 encyclopedia articles, Funk and Wagnalls, New York, NY, 1979.

[8] H.P. Luhn, The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 2:2, 159–165, April 1958.

[9] J.E. Rush, R. Salvador and A. Zamora, Automatic Abstracting and Indexing — Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria, *Journal of the ASIS*, 22:4, 260–274, July–August 1964.

[10] G. Salton, editor, *The Smart Retrieval System — Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[11] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, New York, NY, 1983.

[12] G. Salton, *Automatic Text Processing — the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing Co., Reading, MA, 1989.

[13] G. Salton, A. Wong, and C.S. Yang, A vector Space Model for Information Retrieval, *Journal of the ASIS*, 18:11, 613–620, November 1975.

[14] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, *Journal of the ASIS*, 26:1, 33–44, January 1975.

[15] G. Salton and C. Buckley, Term weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24:5, 513–523, 1988.

[16] G. Salton, Developments in automatic text retrieval, *Science*, 253, 974–980, August 1991.

[17] G. Salton and J. Allan, Selective Text Utilization and Text Traversal, Proc. Hypertext-93, *Association for Computing Machinery*, New York, November 1993.

[18] G. Salton, J. Allan and C. Buckley, Automatic Structuring and Retrieval of Large Text Files, *Communications of the ACM*, 37:2, 97–108, February 1994.

[19] G. Salton, J. Allan, C. Buckley, and A. Singhal, Automatic Analysis, Theme Generation and Summarization of Machine Readable Text, *Science*, 264, 1421–1426, June 1994.