Modeling Medical Content for Automated Summarization

DAVID B. JOHNSON,[a] QINGHUA ZOU,[b] JOHN D. DIONISIO,[a]
VICTOR ZHENYU LIU,[b] AND WESLEY W. CHU[b]

[a]*Department of Radiological Sciences, University of California Los Angeles, Los Angeles, California 90024, USA*

[b]*Department of Computer Science, University of California Los Angeles, Los Angeles, California, USA*

ABSTRACT: Medical information is available from a variety of new online resources. Given the number and diversity of sources, methods must be found that will enable users to quickly assimilate and determine the content of a document. Summarization is one such tool that can help users to quickly determine the main points of a document. Previous methods to automatically summarize text documents typically do not attempt to infer or define the content of a document. Rather these systems rely on secondary features or clues that may point to content. This paper describes text summarization techniques that enable users to focus on the key content of a document. The techniques presented here analyze groups of similar documents in order to form a content model. The content model is used to select sentences forming the summary. The technique does not require additional knowledge sources; thus the method should be applicable to any set of text documents.

KEYWORDS: information retrieval; content modeling; automatic summarization

INTRODUCTION

The amount of available information has never been greater. The Internet has fostered the growth and availability of digital text and content. To enable users to make use of online resources, new methods must be found that will enable them to quickly assimilate and make decisions on information from multiple sources in a timely matter. As a method to allow individuals to quickly interpret documents, automatic text summarization has received renewed interest in recent years in hope of enabling users to cope with the growing amount of information.[1]

For most queries, search engines often retrieve a number of relevant documents. Unfortunately, the relevant documents are often heavily outnumbered by irrelevant documents. Thus, the problem becomes one of filtering out irrelevant information to find the desired information. Summarization can be thought of as one such filtering mechanism[2] that can help users alleviate the burden of processing multiple information sources. By filtering information, summarization may enable users to find information that will directly satisfy their particular information need.

Address for correspondence: David B. Johnson, UCLA Telemedicine Division, 924 Westwood Blvd., Suite 420, Los Angeles, CA 90024. Voice: 310-794-3539.
djohnson@itmedicine.net.

Ann. N.Y. Acad. Sci. **980: 247–258 (2002).**

Although expert generated summaries appear superior to automated methods, there are several issues that effect manual methods. First, manual methods are time consuming and costly. In the past, summaries or abstracts have been fixed, long-lived, stand-alone texts. Information access was largely limited to libraries or other static document collections, where abstracts could be methodically implemented for the general user. Today, information access is not limited to controlled collections, rather digital information access has facilitated the creation of dynamic collections often dictated by the interests of each user rather than a specific library collection. Given this dynamic nature of information access, it becomes harder to define the needs of a user as well as control the creation of summaries. Indeed, in a perfect world a battery of experts would be available to quickly compose a tailored summary for any document of interest. The reality is that it is too costly to *manually* provide such summaries. Furthermore, research has shown that human abstractors are often influenced by their own particular background and interests, thus affecting the quality of a given summary.[3] Finally, research has also shown that there is often a wide discrepancy between different human abstractors.[4]

Digital text with automated summarization methods allows a new dynamic type of abstract, that is difficult, if not impossible to create in the physical world. By modeling and inferring the interests of a user, abstracts can be tailored to each users own particular information needs.

Finally, automated summaries are consistent. External factors may influence a human abstractor, but such effects are not found in automated systems. Automatically generated summaries can be customized so that the interests and information needs of a user can be factored into the summarization process. Automated techniques can be readily customized for particular users and their information interests. Thus, automated summarization systems have several desirable features that are difficult to attain with human abstractors.

This paper explores automated methods to summarize documents. Two automatic summarization techniques are described in this paper. Both techniques have been used to summarize documents that may be of interest to patients and their physicians. This paper describes the techniques, as well as presenting summarization results and comparisons to other automated techniques.

BACKGROUND

The process of summarization is often thought as a natural language processing task, requiring in-depth understanding in order to provide a useful summarization.[5] Typically, non-automated summaries are produced by human experts. these experts often include the original author, editors, or others specializing in document summaries.

In general, the summarization process follows a three-step process transforming the source text into a summary output:[6]

1. source analysis,
2. content selection, and
3. summary generation.

Source analysis is a process of determining the features that will drive the summarization process. Depending on the summarization technique, the analysis may take a variety of forms ranging from simple methods that only search for specific key words to complex techniques that employ natural language understanding.

Content selection is a process of determining which of the generated features, provided by the source analysis phase, are to be included in the summary. For example, sentences or paragraphs containing the specific key words may be selected for inclusion for the summarization.

Finally, the summary is generated using the content chosen in the previous step. Using the previous example, the selected sentences or paragraphs may be combined in sentence order to form the summary. The generation phase may also provide some form of processing that may improve the aesthetics of the final summary output.

Documents can be subdivided into natural sections or passages (e.g., document sections, paragraphs, or sentences).[7] Several systems have been developed to extract passages from the original document and combine them to form a summary.[3,7–9] These systems, by selecting and using passages from the original source, greatly simplify the summary generation process.

Previous Work

Word Frequency

In 1958, Luhn proposed a summarization process in which key sentences were selected and combined to form an abstract of the original document. Luhn's method selected sentences based on their use of significant key words.[3] Although the technique could have used a manually compiled dictionary of significant terms, Luhn's method automatically derived the set of significant words directly from the document. Thus, the content of a document was represented by a set of significant words.

The technique determined the significance of a word by measuring its frequency of occurrence within the document. Words that occurred most frequently and least frequently were considered not significant. It is debatable how frequency should be considered in determining the significance of a term, but research continues to support the idea that frequency can be used to select significant terms.[10]

Content Cues

Cue words and cue phrases have been used as a criteria to create summaries. Researchers have noted that writers often use certain words to note important findings or facts. For example, passages containing the words "significant," "impossible," or "hardly" may contain information more important than that found in other passages. Rather than representing content directly, the method relies on *clues* that may or may not actually signify content.

Cue phrases, even more so than cue words, may imply the significance of a sentence in the document.[11] For example, "in conclusion," "in summary," and "the most important" have been used for summary selection. Each of these phrases are often used to signal important content to the reader, thus they can be used by automatic summarization to select passages.

Cue words and phrases can be classified into three different categories, each representing the significance of the word or phrase to the summarization. *Bonus words,*

are those words indicating the positive relevance of a passages to the summary; *stigma words*, words that indicate a negative relevance for a summary, can be used to filter out unimportant passages; and, *null words*, that are irrelevant for selecting summary content.[12] For example, a passage or sentence containing many bonus words and few or no stigma words would be considered highly relevant to the resulting summary. Although cue words and phrases have continued to be used as a selection criteria, some researchers have noted that the approach may work for narrow types of texts, however, given the variability in writing and style, the technique may not be as effective for domain independent summarization.[13]

Content Coverage

Salton *et al.*, proposed a method to summarize documents that uses the intradocument similarity of passages (i.e., sentences or paragraphs) to chose significant text. The content of document passages are represented as vectors of words or word stems. The content model of passages are compared to one another pairwise forming an intradocument similarity matrix. Each pair of passages with a similarity measure greater than a predefined threshold is considered related and a link is formed between them.[4] The threshold defines a minimum amount of content overlap between the two passages necessary to permit one passage to substitute the other. The summary is formed by selecting those passages with the greatest number of links, that is those passages that overlap or cover the content of the other passages.[4,8]

Redundancy Reduction

Carbonell and Goldstein also describe a method to produce summaries by extracting specific passages.[2] Content of each passage is again represented using a simple single word vector space model. Passages are selected for the summary by combining the relevance of the passage with its novelty in the context of previously selected passages. A metric, called marginal relevance, measuring both the relevance of the sentence in question, as well as the degree of redundancy the sentence would add to the summary (i.e., the previously selected sentences), was defined. By discriminating both by relevance and redundancy, the algorithm can increase the overall content of the summary.

METHOD

This section describes two methods to select passages to summarize documents. The first method compares the use of *n*-word combinations, a type of phrase analysis, for document summarization to an existing non-phrase technique. The second method uses *n*-word combinations in conjunction with cluster analysis to summarize a document within the context of several related documents. This second technique can be very useful given that search engines often return many related documents for any search. The information provided by multiple related documents has been ignored by previous methods.

Content Representation

The methods use *n*-word combinations to represent the content of document. An *n*-word combination is defined as an unordered collection of *n* words taken from a document.[14] Previous research has shown that *n*-word combinations can be used to improve the precision of information retrieval tasks.[14] Word combinations can provide needed contextual information missing from single word models. For example, analyzing a document on heart disease and the use of aspirin, the following frequent single words were found: *heart*, *aspirin*, and *patient*. Further examination of the document's sentences found the following word combinations: *aspirin patient*, *aspirin heart*, *aspirin patient take*, *aspirin patient study*, and *aspirin heart regular use*. The word combinations provide additional information and expose the relationships between words in sentences.

Phrase Analysis Summarization

To analyze a particular document, first the document is partitioned into its constituent sentences. The words in each sentence are parsed, stemmed to a common prefix, and combined to form word combinations. A list of the extracted word combinations is maintained with their frequencies.

Word combinations are weighted based on their frequency. Infrequent combinations are considered less significant than frequent combinations. The number of high frequency combinations in each sentence is used to rank and select sentences. The technique is similar in some respects to that proposed by Luhn, although the model used for phrase analysis summarization is richer (*n*-word combinations compared to isolated keywords) and the selection criteria of key content is different. The summary is generated by using the ranked sentences presented in the original sentence order. The summary length can be limited by using only the highest ranking sentences.

Cluster Analysis Summarization

For each document, the source analysis phase follows the same procedure as in the phrase analysis summarization. Each document is partitioned into sentences, the words are parsed and the *n*-word combinations are formed. The *n*-word combinations are used to create a content model for each document.[15]

Cluster analysis summarization uses several related documents to perform summarization. Search engine results frequently return many related relevant as well as irrelevant documents. Clustering is used to group documents together, forming relevant as well as irrelevant groups of documents. To summarize a document, the group containing the document of interest is selected and analyzed.

The technique described in this document relies on document clustering to form subgroups of similar documents. These subgroups are then analyzed for features with high support, called *key features*. The key features are then used to rank sentences. This technique is supported by the *cluster hypothesis*, which states that documents that contain similar features will be relevant to the same queries. Here we interpret the cluster hypothesis as stating that similar documents (i.e., those with similar features) are likely to be about the same thing (i.e., be relevant to similar queries).

These groups of documents are then further analyzed to extract the key features, forming a *cluster signature*, that best characterize each document group. The cluster

signature can be used either directly as a phrasal summary or to select passages for a summary. The summary is generated by matching the cluster signature to each sentence. Both the sentence and the cluster signature are represented using a vector space model and the previously extracted *n*-word combinations.

EXPERIMENTS

Content Model Analysis

Previous work has analyzed *n*-word combinations as a content model for information retrieval. That work compared *n*-word combinations with simpler content models, specifically single words used in a vector space. The experiments supported the idea that *n*-word combinations more precisely modeled the content of the document and queries (i.e., retrieval results using *n*-word combinations were more precise than those based on isolated words). Although *n*-word combinations may better capture content, it is a statistical technique rather than an actual concept based model. Although *n*-word combinations may better model content, they also introduce noise (i.e., combinations that are not easily mapped to a single well defined meaning), that a pure concept based model would not introduce.

To evaluate the effect of noise introduced using *n*-word combinations, a series of information retrieval experiments was conducted comparing a pure concept model with *n*-word combinations. The experiments followed those described elsewhere by two of the authors.[14] The experiments used a collection of thoracic radiology reports. Three queries regarding specific medical findings were used to compare a single word vector model with various length *n*-word combinations.[14]

The results from these earlier experiments were used to compare the differences between 2-word combinations and a concept based model. The same collection was also indexed using a concept based model. Two of the three original queries, "right upper lobe mass" and "left upper lobe mass" were chosen for the evaluation.

Concept indexing uses a version of the UMLS meta-thesaurus to map the words in each document to concepts. The two main anatomy concepts, "right upper lobe" and "left upper lobe" are not defined in the UMLS and were added for the experiment forming the concept map. To identify concepts, words were scanned in order and matched to the concept map. For each sequence of words, the longest matching a concept was used in the document concept list. Matching concepts is restricted by word order, thus the mapping may be incomplete. The scanning process resumed from the word after the last identified concept, and terminated when the last word in the document was processed. The two queries were represented by two concepts each, an anatomy concept and a medical finding concept (e.g., "right upper lobe" and "mass"). The vector distance model was used to compute query-document similarities.

The results from the experiment are shown in FIGURES 1 and 2. The figures show the precision–recall graph for 3-word, 2-word, the concept representation, and an isolated word model. The results show that 3-word combinations that are not restricted to word order better capture the content of the queries than the concept model. As previously reported, the figures also show how multiword combinations better model the content than the single word model.[14]

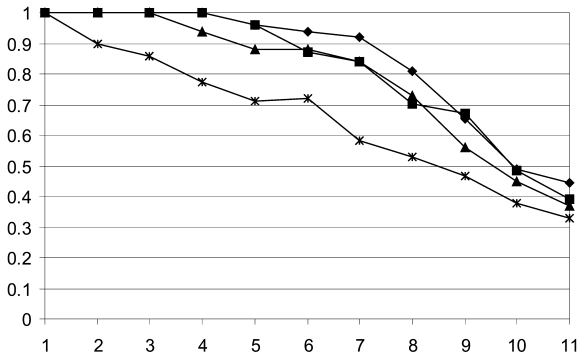

FIGURE 1. Concept and *n*-word—"right upper lobe mass": ◆, 3-word; ■, 2-word; ▲, concept; ×, 1-word.

Frequent Phrase Analysis

The first set of experiments compare the use of multiword combinations and a single word representation. Salton's techniques for summarization are based on the pairwise similarity of sentences via its vector dot product. For a given similarity threshold level, an object (sentence) relational graph can then be constructed in which the nodes represent the sentences and a link between two nodes represents the similarity of the two sentences. The graph is constructed using only those links having values above the specified threshold. The document summarization is derived by selecting a set of these nodes (sentences) that have a higher number of links (bushiness). In general, the threshold level may affect the quality of the summarization.

In the experiments nine different document were summarized using Salton's method using three different thresholds values. For each threshold, the three highest

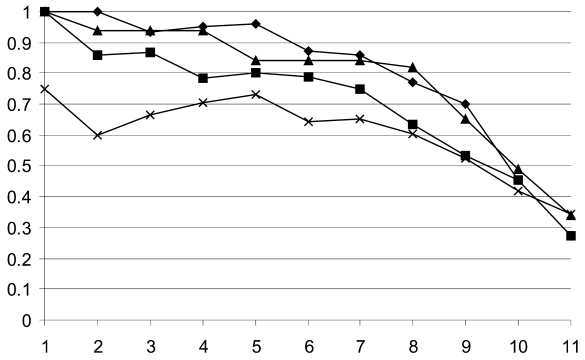

FIGURE 2. Concept and *n*-word—"left upper lobe mass": ◆, 3-word; ■, 2-word; ▲, concept; ×, 1-word.

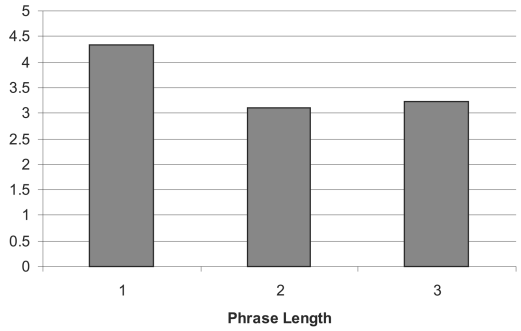

FIGURE 3. Threshold sensitivity.

ranking sentences (S_1, S_2, and S_3) were used as the summary. The results for each phrase length were scored by intersecting the sets of sentences and dividing by the number of documents, $\frac{1}{9}|S_1 \cap S_2 \cap S_3|$, see FIGURE 3. Our results reveal that summarization for single word indexing is very sensitive to the threshold. On average, changing the threshold changed the summary by more than one sentence. However, the algorithm becomes less sensitive to the threshold value when multiword combinations are used (i.e., changing the threshold had little effect on the summary).

A second set of experiments evaluating multiword combinations and single word representations was performed. These experiments used abstracts and full texts from the *Journal of the American Medical Association* (*JAMA*). For a given document and its abstract, the similarity between the abstract and each sentence in the document was calculated. The ten highest ranking sentences was used as a standard set (S) against which to compare the two summarization methods. The precision rate of a summarization A is defined as follow by $P_A = \frac{1}{10}|S \cap A|$ (i.e., the number of sentences in common with the standard set divided by 10).

The test results on five JAMA reports are shown in TABLE 1. The data shows that the average document size is 210 sentences. For the FPA method, the average precision rates are 24%, 42%, and 48% for 1-word, 2-word, 3-word, respectively. For the Salton method, the average precision rates are 8% for 1-word, 24% for 2-word, and

TABLE 1. JAMA sentence precision (*n*-word combinations)

Document name	Sentences	FPA Summary (%)			Salton Summary (%)		
		1-word	2-word	3-word	1-word	2-word	3-word
Joc01517	201	30	40	50	10	30	50
Joc01726	205	20	50	60	0	10	40
Joc01746	283	10	30	30	10	20	20
Joc01942	208	50	60	60	10	10	40
Joc02101	154	10	30	40	10	50	50
Average	210	24	42	48	8	24	40

TABLE 2. JAMA sentence precision (*n*-concepts)

Document name	Sentences	FPA Summary (%)			Salton Summary (%)		
		1-con	2-con	3-con	1-con	2-con	3-con
Joc01517	201	40	40	40	10	20	0
Joc01726	205	60	30	20	0	0	10
Joc01746	283	20	30	30	0	20	0
Joc01942	208	50	40	30	0	60	10
Joc02101	154	70	50	50	0	40	10
Average	210	48	38	34	2	28	6

40% for 3-word. The table shows that for both methods the precision is better for multiword combinations than for 1-word models.

A third set of experiments was performed using a concept representation of content. The experiments were performed as previously, using the same abstracts and full texts from JAMA. A concept model was used rather than the multiword combinations. The UMLS meta-thesaurus was used to map individual words to unique concept identifiers. Both the document and its abstract were modeled using the concept identifiers. For a given document and its abstract, the similarity between the abstract and each sentence in the document was calculated using the concept model. The ten highest ranking sentences was used as a standard set (S) against which to compare the two summarization methods.

The test results on five JAMA reports are shown in TABLE 2. For the FPA method, the average precision rates are 48%, 38%, and 34% for 1-concept, 2-concept, and 3-concept models, respectively. For the Salton method, the average precision rates are 2% for 1-concept, 28% for 2-concept, and 6% for 3-concept. The table shows that the highest precision uses the FPA method using the single concept model. The Salton method performed best for the two concept model.

Cluster Phrase Analysis

Several experiments were performed using various collections compiled from the Internet. The collections were compiled using a search engine and described the effects of aspirin on heart disease or heart attacks. These collections simulated how individuals may use the automated summarization system in conjunction with a search engine. Automated summarization can provide better information than that typically supplied with search engine results.

Heart Disease Collection

Two collections of documents, derived using the Excite and Yahoo search engines, were compiled. Each sentence was compared to the content model and ranked using the cosine similarity measure. The top three ranking sentences (approximately 28% of the original document) presented in sentence order were:

> *DALLAS (AP)—As many as 10,000 American lives per year could be saved if more people who think they're having a heart attack took an aspirin at the onset of chest pains, according to a new report. The American Heart Association*

> *first recommended in 1993 that people take one, full 325-milligram aspirin at the onset of chest pain or other symptoms of a severe heart attack. The 1993 study recommended that those who have already had a heart attack or other serious heart disease take one 50- to 100-milligram baby aspirin per day to prevent a recurrence.*

A second document titled "heart group: take aspirin at first signs of attack" from the same cluster was also summarized using the same technique and content model. The document comprises 19 passages (i.e., sentences, subtitles or titles). The top four ranking sentences (approximately 28% of the original document) presented in sentence order are shown below:

> *Heart group: Take aspirin at first signs of attack. In the latest issue of the journal* Circulation*, the AHA says that as many as 10,000 American lives could be saved every year if everyone followed its recommendation to take a single 325-milligram aspirin tablet at the first signs of a severe heart attack. Four years later, a follow-up report shows that only 20 to 40 percent of heart attack victims are taking the seemingly simple step of taking one aspirin at the onset of symptoms. Doctors already routinely prescribe an aspirin a day for those who have had a heart attack to help prevent another one.*

This summary exposes some of the problems that can arise in ranking and presenting passages directly from the original document. The third sentences starting with "Four year later, a follow-up report shows…" appears to be referring to an unmentioned report. The context becomes clearer when a missing passage is returned:

> *In the latest issue of the journal Circulation, the AHA says that as many as 10,000 American lives could be saved every year if everyone followed its recommendation to take a single 325-milligram aspirin tablet at the first signs of a severe heart attack.* The heart association first made that recommendation in 1993. *Four years later, a follow-up report shows that only 20 to 40 percent of heart attack victims are taking the seemingly simple step of taking one aspirin at the onset of symptoms.*

Although adding this passage does not significantly increase the content of the summary, the passage does improve the readability of the overall summary.

CONCLUSION

As patients and physicians continue to use new online information sources provided by the Internet, automatic summarization can play a larger role in assisting with locating relevant information. The research presented here examines the use of multiword combinations for automatic summarization and describes a unique method to analyze and select content through cluster analysis. The experiments support that multiword combinations can improve the summarization process.

The cluster phrase analysis technique can be used in situation where several related documents are available. The cluster phrase analysis can use this additional information to focus on common key content. It has been hypothesized that documents grouped together based on common features would be relevant to the same

information needs.[16] Recent research has supported this hypothesis, showing that clustering data can help users to find information[17] as well as support their understanding of the document collection.[18] The research described here uses the cluster hypothesis as a foundation and further hypothesizes that it is possible to define features in order to group documents by their content.

A prototype of the system has been developed. The prototype allows groups of documents to be loaded and indexed using *n*-word combinations. Documents can then be clustered into groups. To summarize a document, the cluster of interest can be selected, from which the cluster signature is derived, and the document or documents to summarize. Several experiments presented here show how the cluster signature can be used to automatically rank sentences. The ranked sentences can then be selected and presented to the user as a summary.

Although not seen in the summaries presented here, using a single content model, such as the cluster signature, to rank and select passages can lead to a problem of redundancy. Redundancy can occur if two passages containing similar content are also similar to the content model (e.g., the cluster signature). There are several methods that may be used to significantly decrease redundancy in the resulting summary.

Maximal marginal relevance (MMR) reranking is a method proposed to minimize redundancy in a summarization.[2] MMR reranking takes a ranked set of passages and reranks the passages such that content already included is ranked lower in subsequent passages. As described, the MMR technique did not include an initial ranking algorithm, thus the initial ranking could be performed using the techniques described earlier and then reranked by MMR to minimize redundancy.

REFERENCES

1. LEHMAM, A. 1999. Text structuration leading to an automatic summary system: RAFI. Inform. Process. Manage. **35:** 181–191.
2. CARBONELL, J. & J. GOLDSTEIN. 1998. The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. Proceedings of the ACM SIGIR Conference, Melbourne, Australia. 159–165.
3. Luhn, H.P. 1958. The automatic creation of literature abstracts. IBM J. Res. Develop. **2**(2): 193–207.
4. SALTON, G., A. SINGHAL, M. MITRA & C. BUCKLEY. Automatic text structuring and summarization. Inform. Process. Manage. **33**(2).
5. WILKS, Y. 1998. Information retrieval, extraction and summarisation. IEE Colloquium, Speech and Language Engineering—State of the Art, London, UK, November. IEE.
6. ENDRES-NIGGEMEYER, B. 1994. Summarizing text for intelligent communication – results of the dagstuhl seminar. Knowledge Organiz. **21**(4): 212–223.
7. SALTON, G., J. ALLAN & C. BUCKLEY. 1993. Approaches to passage retrieval in full text information systems. SIGIR 1993, Pittsburgh, PA.
8. SALTON, G., A. SINGHAL & C. BUCKLEY. 1996. Automatic text decomposition using text segments and text themes. Hypertext. 53–65.
9. KUPIEC, J., J. PEDERSEN & F. CHEN. 1995. A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference, Seattle, WA. 68–73.
10. YANG, Y. & J.P. PEDERSEN. 1997. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97).
11. HOVY, E. & C. LIN. 1999. Automated text summarization in SUMMARIST. *In* Advances in Automatic Text Summarization. I. Mani and M.T. Maybury, Eds. The MIT Press.

12. EDMUNDSON, H.P. 1969. New methods in automatic extracting. J. Assoc. Comput. Machinery. 264–285.
13. BRANDOW, R., K. MITZE & L.F. RAU. 1995. Automatic condensation of electronic publications by sentence selection. Inform. Process. Manage. **31**(5): 675–685.
14. JOHNSON, D.B. & W.W. CHU. 1999. Domain specific document retrieval using n-word combination index terms. Proceedings: Fusion '99.
15. JOHNSON, D.B. 2000. Methods for Domain-Specific Information Retrieval. Ph.D. Thesis, University of California, Los Angeles.
16. VAN RIJSBERGEN, C.J. 1979. Information Retrieval. Butterworths.
17. ANICK, P.G. & S. VAITHYANATHAN. 1997. Exploiting clustering and phrases for context-based information retrieval. Proceedings of the Ann. Intl. SIGIR Conference, Philadelphia, PA.
18. LENT, B., A. SWAMI & J. WIDOM. 1997. Clustering association rules. Proceedings 13th International Conference on Data Engineering. IEEE.