Ou, S., Khoo, C. & Goh, D. (2006). Automatic multi-document summarization for digital libraries. In C. Khoo, D. Singh & A.S. Chaudhry (Eds.), *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006 (A-LIEP 2006), Singapore, 3-6 April 2006* (pp. 72-82). Singapore: School of Communication & Information, Nanyang Technological University.

AUTOMATIC MULTI-DOCUMENT SUMMARIZATION FOR DIGITAL LIBRARIES

OU SHIYAN, CHRISTOPHER S. G. KHOO & DION H. GOH
School of Communication and Information, Nanyang Technological University
31 Nanyang Link, Singapore, 637718
E-mail: ou_shiyan@pmail.ntu.edu.sg, assgkhoo@ntu.edu.sg ashlgoh@ntu.edu.sg

Abstract. With the rapid growth of the World Wide Web and online information services, more and more information is available and accessible online. Automatic summarization is an indispensable solution to reduce the information overload problem. Multi-document summarization is useful to provide an overview of a topic and allow users to zoom in for more details on aspects of interest. This paper reports three types of multi-document summaries generated for a set of research abstracts, using different summarization approaches: a sentence-based summary generated by a MEAD summarization system that extracts important sentences using various features, another sentence-based summary generated by extracting research objective sentences, and a variable-based summary focusing on research concepts and relationships. A user evaluation was carried out to compare the three types of summaries. The evaluation results indicated that the majority of users (70%) preferred the variable-based summary, while 55% of the users preferred the research objective summary, and only 25% preferred the MEAD summary.

Introduction

With the rapid growth of the World Wide Web and online information services, more and more information is available and accessible online. Automatic summarization has attracted attention both in the research community and commercially commercial organizations as a solution for reducing information overload and helping users to scan a large number of documents to identify documents of interest. It is an important function that should be available in large digital library systems, information retrieval systems and Web search engines, where the retrieval of too many documents and the resulting information overload is a major problem for users.

Digital library systems, Information retrieval systems and Web search engines attempt to address the problem of information overload by ranking documents retrieved by their likelihood of relevance, and displaying titles and short abstracts to give users some indication of the document content. The abstracts may be constructed by humans or automatically generated by extracting the first few lines of the document text (called lead sentences) or extracting the most important sentences in the document. An example application of such summarization is the Google search engine which returns a list of ranked document "snippets" on a search query (see Figure 1).

However, the ranked list presentation is hard for users to find relevant documents from a larger number of records. The user has patience to scan only a small number of document titles and abstracts, usually within the range of 10 to 30 (Jansen, Spink & Saracevic, 2000). To make search results easy to browse, some search engines group the retrieved records into folders or categories, e.g. the Northern Light search engine at http://www.northernlight.com (see Figure 2). These categories may be pre-created, and the records assigned to them by human indexers, or by automatic categorization techniques. The categories may also be constructed dynamically by clustering the documents retrieved, e.g. Grouper, an interface to the results of the HuskySearch meta-search engine (Zamir & Etzioni, 1999).

A related approach is to dynamically construct a multi-document summary of the documents retrieved. While single-document summarization is a well-developed field, especially in the use of sentence extraction techniques, multi-document summarization has begun to attract attention only in the last few years (National Institute of Standards and Technology, 2002). Multi-document summarization is capable of condensing a set of related documents, instead of a single document, into one summary. A multi-document summary has several advantages over the single-document summary. It provides a domain overview of the subject area indicating common information across many documents, unique information in each document, and cross-document relationships (relationships between pieces of information in different documents), and it can allow users to zoom in for more details on aspects of interest.

This paper reports on an initial study to develop a method for summarizing a set of research abstracts that might be retrieved by a digital library in response to a user query. The summarization method that was developed focuses on identifying and extracting research concepts and variables and

relationships between them. The generated summary is presented in an interactive Web-based interface. A user evaluation was carried out to compare this variable-based summary of a set of abstracts with two other types of summaries generated using the traditional sentence extraction approaches: one generated by a state-of-the-art summarization system MEAD that extracts highly ranked sentences based on a variety of criteria; another generated by extracting research objectives from the research abstracts. In this initial study, a database of dissertation abstracts in sociology was used as an example digital library.

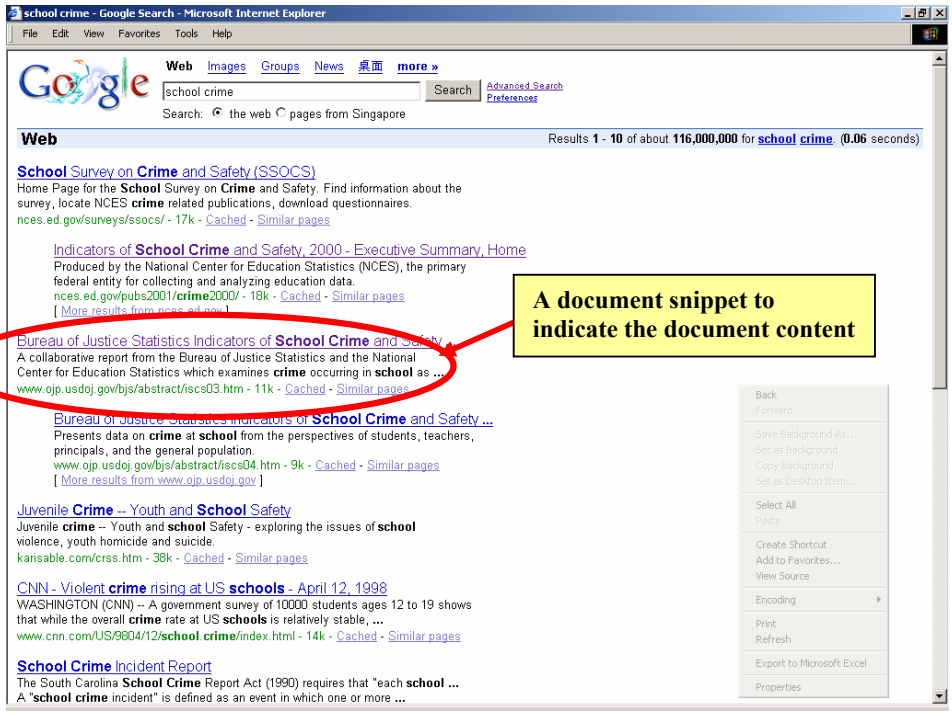

Figure 1. A list of document "snippets" returned by Google

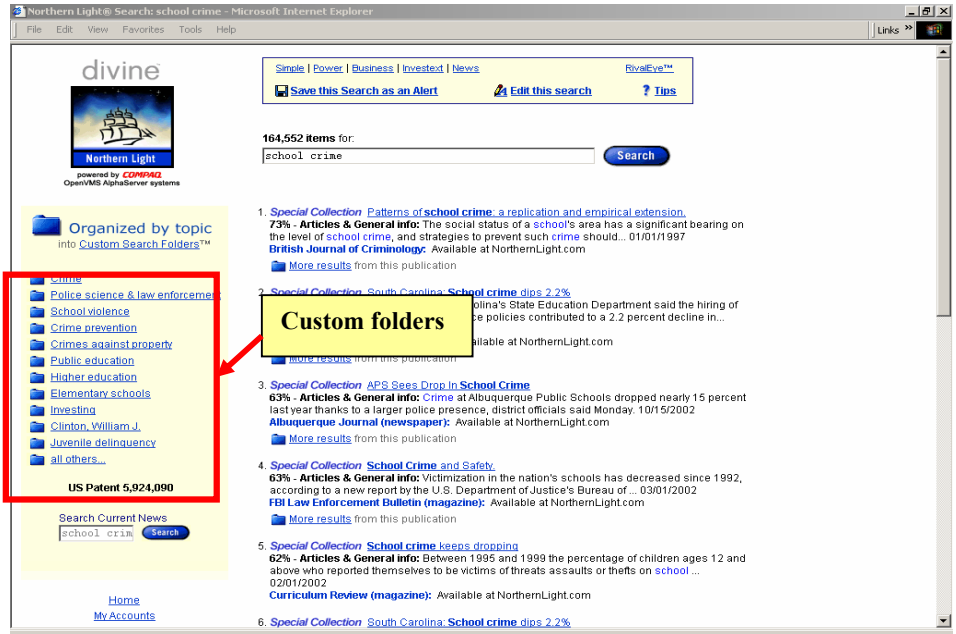

Figure 2. Search results clustered into custom folders by the North Light search engine

Automatic Multi-document Summarization

Mani (2001) characterized automatic summarization as "to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs." Stein, Strzalkowski and Wise (2000) regarded automatic summarization as "the process of creating a shorter representation of an original information source." For different purposes and different users, different types of summaries are generated, for example, *extracts* vs. *abstracts*, *indicative* vs. *informative* summaries, and *generic* vs. *user-focused* summaries (Mani, 2001; Hahn & Mani, 2000).

Depending on the number of documents to be summarized, a summary can be a single-*document summary* and a *multi-document summary* (Hahn & Mani, 2000). Single-document summarization can only condense one document into a shorter representation, whereas multi-document summarization can condense a set of documents into a summary.

The simplest approaches for automatic summarization are sentence extraction approaches. This approach is domain independent and does not attempt to understand the meaning and structure of the text. It makes use of statistical and linguistic features to measure the significance of sentences thereby extracting the highest scoring sentences as the components of the summary. The well-known features used in most summarization work include frequent keywords, indicator phrases, title keywords, sentence position, and so on. The weaknesses of the sentence extraction approaches are that the resulting extracts may be redundant, not coherent and fluent, and hard to read. These weaknesses become more serious in multi-document summarization because the extracted sentences may come from different sources and have different writing styles (Hahn & Mani, 2000).

Multi-document summarization is more useful than single-document summarization in digital libraries. For example, when a user submits a search query to an information retrieval system, digital library system or Web search engine such as Google, thousands of related documents are retrieved, and displayed in decreasing order of probable relevance. Since the related documents are likely to contain repeated information or share the same background, their single-document summaries are likely to be similar to each other and thus cannot indicate unique information in individual documents. Moreover, browsing so many similar single-document summaries is tedious and time consuming, and it is hard to obtain an overview of a subject area. A multi-document summary is likely to be essential in such a situation (Goldstein et al., 1999). It provides an overview of the topic by indicating what is similar and different in different documents, and relationships between pieces of information across documents, and allows people to zoom in for more details on aspects of their interest.

Multi-document summarization can be seen as an extension of single-document summarization, but also can be much more. Since it combines and integrates information across documents, it performs knowledge synthesis and knowledge discovery, and can be used for knowledge acquisition. It provides a domain-overview of a subject area and, if presented in a graphical or visual way, can support user browsing and information visualization. Multi-document summaries are useful in large digital libraries, especially in academic institutions, and can be used for knowledge discovery to identify connections between research results that are not obvious, and gaps in the field for future research.

Multi-document summarization has more challenges than single-document summarization in the issues of compression, redundancy, cohesion, coherence, and temporal dimension, etc. (Goldstein et al., 2000):

- The degree of redundancy becomes significantly higher in multi-document summarization, since some information is repeated in different documents;
- There is a time sequence in a group of related documents, typically in a series of news articles about an unfolding event;
- Since many documents are encapsulated into one summary, a higher compression rate is required;
- Co-reference resolution is more difficult when referents occur across documents;
- Cohesion and coherence is more difficult because the summarized information comes from different sources and have different writing styles;
- In a group of related documents, each document does not only share similarities but also describe differences.

Given the above differences, single-document summarization methods do not work very well in multi-document summarization environment. Therefore, several new approaches were developed for multi-document summarization, focusing on reducing redundancy, ensuring cohesion and coherence, and identifying similarities and differences across documents.

Multi-document summarization usually focuses on similarities and differences among a set of documents. One characteristic of a set of related documents is the repetition of the same information

albeit in different forms. Repeated information is a good indicator of the importance and can be used for the generation of a multi-document summary. However, when there is a substantial amount of differences in individual documents, it is unavoidable that valuable and unique information will be omitted. Thus not only similarities but also differences selected from individual documents need to be considered in multi-document summarization. The advantage of identifying similarities and differences is that the similarities indicate salient information in a document set, whereas the differences indicate unique information about each document.

Three Types of Multi-document Summaries

This section describes the three types of multi-document summaries automatically generated for a set of research abstracts retrieved by a digital library in response to a user query. Dissertation abstracts in sociology were used as source documents. There is increasing interest in constructing digital libraries of dissertations (Moxley, 2001), especially in academic institutions, because there is a ready supply of student dissertations in universities, and dissertations are a rich source of information on new and emerging research fields (Herther, 2000).

Three types of summaries were generated using different summarization methods as follows:

- *MEAD summary* – a sentence-based summary, generated by the MEAD summarization system that extracted highly ranked sentences from different dissertation abstracts using various features;
- *Research objective summary* – a sentence-based summary, generated by extracting research objectives from each dissertation abstract.
- *Variable-based summary* – a summary focusing on research concepts often operationalized as research variables, generated by extracting research concepts and relationships from each dissertation abstracts and integrating them among different abstracts.

MEAD summary

MEAD is a domain-independent multi-document summarization system, built by the University of Michigan. It clusters similar documents in a document set and ranks sentences in each cluster using a linear combination of three features – centroid words (a set of words that are statistically important to a document cluster), sentence position, and first-sentence overlap (Radev et al., 2003). The highest ranked sentences are extracted from each cluster. All the extracted sentences are arranged in the same order as in the document text to form a summary. A sentence-based summary generated by the MEAD system for 10 dissertation abstracts on the topic of "school crime" is given in Figure 3.

Research Objective Summary

Most dissertation abstracts contain a clear research objective section. This section includes research objectives, research questions, hypotheses and the adopted theories and models to give an indication of the main content of the dissertation. Research objective sentences can be identified using cue phrases found at the beginning of sentences, such as "*The purpose of this study was to investigate ...*", "*The present study aimed to explore…*", and "*My purpose here is to answer…*" (Ou et al., 2004). In this approach, research objective sentences were extracted from each dissertation abstract and concatenated to generate a research objective summary (see Figure 4).

Variable-based Summary

The variable-based summary focuses on research concepts and relationships investigated in a set of dissertation abstracts. Much of sociology research adopts the traditional quantitative research paradigm of looking for relationships between concepts operationalized as variables. Although some studies adopt a qualitative research paradigm, many of them also seek to identify relationships between concepts representing events, behaviors, attributes, and situations. This means that research concepts and relationships are the focus of sociology research. A variable-based framework was proposed to integrate and organize the research concepts and relationships among different dissertation abstracts and thus summarize a set of dissertation abstracts on a specific topic (Ou, Khoo & Goh, 2003).

MEAD summary on the topic of "school crime"
(sentences ranked as important extracted from 10 dissertation abstracts)

- [1]The purpose of this research is to assess the perception of public school districts' human resources directors regarding the effects of serious school violence on teacher retention and recruitment. [2]In this study, serious school violence is defined as: homicides, shootings, rape or sexual assaults, robberies, aggravated assaults, and physical attacks or fights with a weapon, perpetrated against a student or teacher on school property. [3]This qualitative, issues-oriented case study was assessed through an analysis of a public school district's human resource department's policies and archival records, and responses on surveys from public school district human resources directors. [6]The human resources directors believe that serious school violence is increasing across the nation, however, 94% believe serious school violence has not effected retention and recruitment in underline>their /underline> district and 92% indicated that their schools were safe to very safe. (doc-id=1)

- [1]The alarming escalation of school crime and violence in recent years is being countered more and more frequently with the establishment of police and security programs within troubled public schools. (doc-id=2)

- [1]The publicity surrounding a number of well publicized and tragic school shootings in the late 1990's has brought a national focus on the school safety issue and has left the impression that america's schools are unsafe. [3]One concern that has emerged from a review of crane reports collected by law enforcement agencies is what many believe may be an underreporting of school crane by school administrators. many factors exist that may effect the accurate reporting of crime to law enforcement by school administrators. [4]This study examined school administrator differences in reporting school crime based upon selected school and administrator variables. [8]Significant differences were found between schools located in cities and towns, schools with and without police or security officers on campus, school administrators who believe reporting crime impacts school safety and those who don't, and those who perceive having district administrative support in reporting school crime and those not perceiving this support. (doc-id=4)

Figure 3. The MEAD summary of 10 dissertation abstracts on the topic of "school crime"

OBJECTIVES summary on the topic of "school crime"
(all the research objectives sentences from 10 dissertation abstracts)

- [1]The purpose of this research is to assess the perception of public school districts' human resources directors regarding the effects of serious school violence on teacher retention and recruitment. [2]In this study, serious school violence is defined as: homicides, shootings, rape or sexual assaults, robberies, aggravated assaults, and physical attacks or fights with a weapon, perpetrated against a student or teacher on school property . (doc-id=1)

- [3]This study addresses these questions through formal analysis of case law on forty-eight relevant cases published from 1974 through 1997 to determine how this legal authority is being defined by federal and state courts across the United States. (doc-id=2)

- [1]The purpose of this study was to determine the types and prevalence of violence in three high schools in the camden city public school district, as perceived by students, staff, and a gang abatement task force. (doc-id=3)

- [4]This study examined school administrator differences in reporting school crime based upon selected school and administrator variables. (doc-id=4)

- [1]The purpose of this study was to investigate the relationship between school size and the variables of school dropout rate, rate of school crime and violence and student extracurricular activity participation. [3]Multiple regression analyses and bivariate correlation procedures were utilized to determine if significant relationships existed between school size and the independent variables of student extracurricular activity participation, school dropout rate and rate of school crime and violence. [4]The additional independent variables of school socioeconomic status, distance from home to school and community size were also entered into the multiple regression analyses. (doc-id=5)

Figure 4. The research objective summary of 10 dissertation abstracts on the topic of "school crime"

This framework contains four kinds of information:

- *Main concepts:* The common research concepts, often operationalized as research variables, investigated by most of the dissertation abstracts in a document set.
- *Research relationships between concepts:* For each main concept, the descriptive attribute values or relationships with other concepts (e.g. correlations and cause-effect relationships) investigated in different dissertation abstracts.
- *Contextual relations:* Concepts and their relationships in the perception, attitude, insight, etc. of a target population, or in the context, framework, model, theory, etc.
- *Research methods:* One or more research methods used to explore the attributes of concepts or their relationships, including research design, sampling, and data measurement & analysis method.

In this framework, each kind of information is integrated across dissertation abstracts and the four kinds of information are combined and organized around the research concepts. This framework has a hierarchical structure in which the summarized information is given at the top level and the more detailed information is at the lower levels. Similar concepts extracted from different dissertation abstracts are clustered and summarized by a broader concept called *main concept*. For a specific concept, its attribute values or research relationships with other concept(s) are given, together with the contextual relations and research methods used in the dissertations. All the relationships involving the same *main concept* are combined and summarized using a simple, standard expression. The contextual relations and research methods are summarized using simple, uniform terms.

Figure 5 shows some of the information which was extracted from 10 dissertation abstracts on the topic of "school crime", and integrated and organized using the variable-based framework.

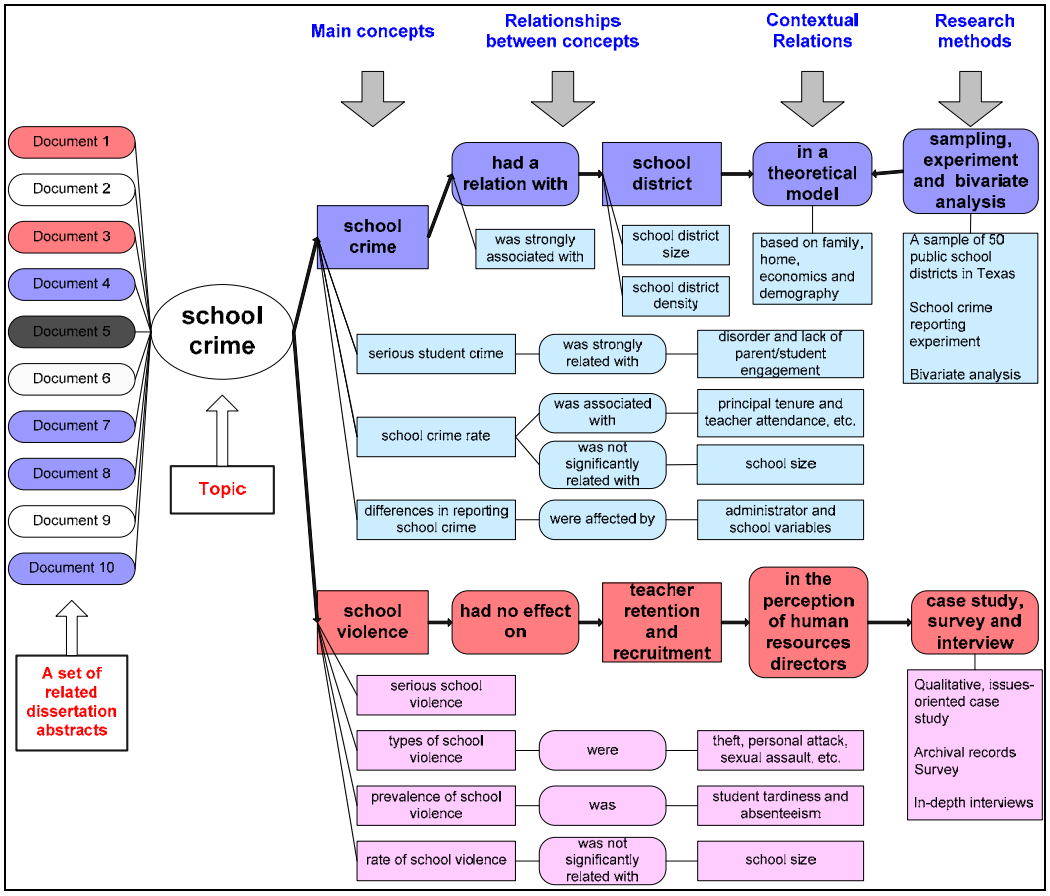

Figure 5. Integrated and organized information across 10 dissertation abstracts on the topic of "school crime" using the variable-based framework

In this study, a new summarization method was developed as one way to operationalize the variable-based framework (Ou, Khoo & Goh, 2005). In terms of the hierarchical structure of the variable-based framework, the generated summary is presented in an interactive Web-based interface with three levels – the summarized information is presented at the top level, the specific information extracted from each dissertation abstracted is presented at the second level, and the original dissertation abstracts are presented at the third level. The three levels are presented on different screens connected by hyperlinks. The summarized information at the top level is presented in the main window and viewed as the main summary (see Figure 6). The user can click on the hyperlinks to access the more detailed information at the lower levels (see Figure 7).

Different presentation formats can be designed and used for the main summary presented in the main window. A simple concept-oriented presentation was adopted for this study. As shown in Figure 6, the four kinds of information are organized separately in the main window. The contextual relations, research methods and research concepts are presented as concept lists, whereas relationships are presented as simple sentences. This design can give users an overview for each kind of information and is also easy to implement.

Contextual relations and research methods found in the dissertation abstracts are presented first because these kinds of information are usually quite short and may be overlooked by users if presented at the bottom of the summary. Research concepts extracted from the dissertation abstracts are organized into broad subject categories (determined by a taxonomy). A list of subject categories can give users an initial overview of the range of subjects covered in the summary and help them locate subjects of interest quickly. Under each subject category, the extracted concepts are presented as concept clusters – each cluster is labeled by a 1-word term called a *main concept*. For each main concept, a concept list is presented, giving a list of related terms found in the dissertation abstracts. The concept list is divided into two subgroups – one for subclass concepts and another for facet concepts. The important concepts in the sociology domain (as determined by a taxonomy) are highlighted in red.

After the concept list, the set of relationships associated with the main concepts are presented as a list of sentences. Each sentence represents a type of relationship that is normalized, conflating different variable concepts found in the dissertation abstracts. When the mouse moves over a variable concept, the original expression of the relationship involving the concept is displayed in a pop-up box.

For each concept, the number of documents is given in parentheses. It is clickable and is linked to a list of summarized documents sharing the given concept in a pop-up window (see Figure 7). For each document, the title, research concepts, contextual relations and research methods are displayed.

Evaluation of the Three Types of Summaries

A user evaluation was carried out to compare the three types of summaries – *the MEAD summary*, *the research objective summary* and *the variable-based summary*. 20 research topics were obtained from 20 researchers in the field of sociology, who were Master's or PhD research students and faculty members at the university. Each researcher submitted one research topic that he/she was working on or had worked on. For each topic, a set of PhD sociology dissertation abstracts were retrieved from the International Dissertation database using the topic as the search query, and at most 200 abstracts were retained for producing the three types of summaries. These researchers (i.e. human subjects) were asked to score and rank the three types of summaries generated for their topic on different criteria (i.e. readability, comprehensibility and usefulness). A questionnaire was used to record their evaluation. Follow-up interviews were carried out to clarify their answers in their questionnaire responses.

Among the 20 researchers, 14 researchers (70%) indicated their preference for the variable-based summaries, 11 researchers (55%) indicated their preference for the research objective summaries, and only 5 researchers (25%) indicated the preference for the MEAD summaries.

According to the researchers' comments, the variable-based summary is more efficient to give an overview of a topic. It is well-organized and concise, and useful for information scanning. But it is too brief to provide accurate information and also has the potential to confuse users. The research objective summary can give more straight indication of the main points of the dissertation. But only research questions are vague and indistinct. The MEAD summary can provide a more complete overview of a topic rather than just research questions. But it contains too many sentences and is hard to read. The detailed comments on the three types of summaries are given in Table 1.

Figure 6. The variable-based summary of 10 dissertation abstracts on the topic of "school crime"

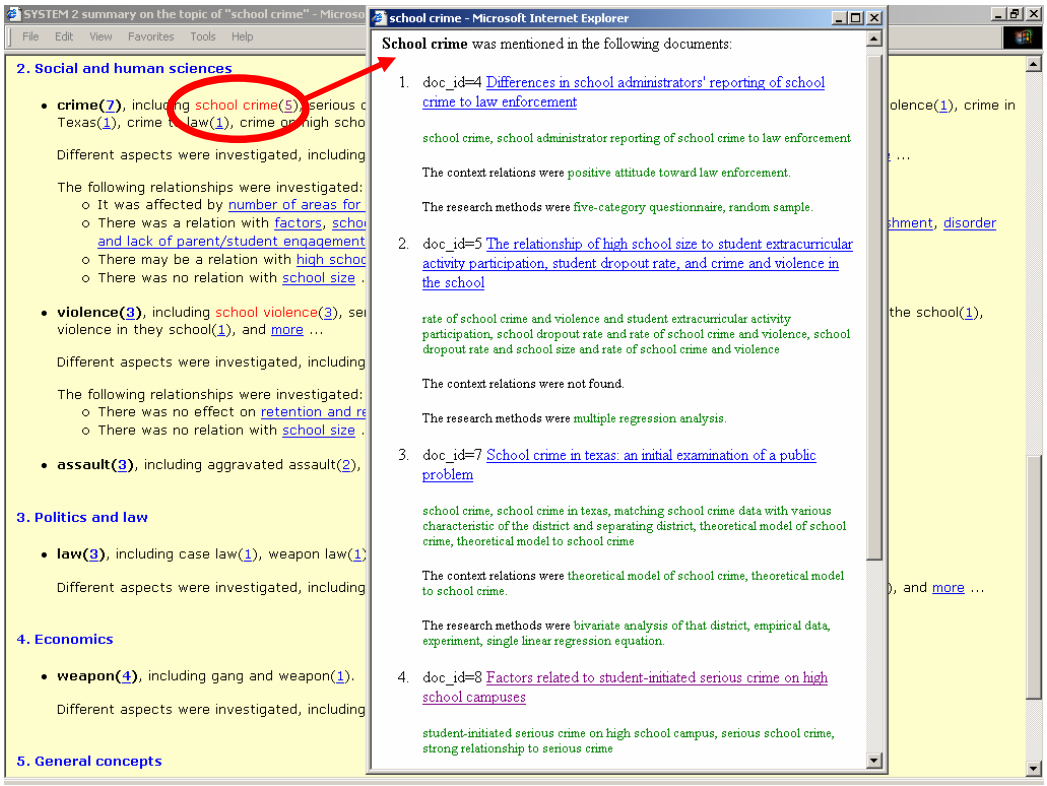

Figure 7. A list of summarized single documents sharing a given concept in a pop-up window

Table 1. Comments by the researchers on the three types of summaries

Comment	Variable-based summary	Research objective summary	MEAD summary
Positive points	- It is more efficient to give an overview of a topic; - It can help researchers find what has been done easily; - It is well-organized and concise. - It makes easier for researchers to find similar information; - It is useful for information scanning; - For quantitative studies which focus on relationships between variables, it is more useful.	- It is more concise than MEAD; - It is much more comprehensible and coherent than MEAD; - It often indicates the most important concepts in the dissertation; - It can give a better indication of research gaps in the subject area and which studies have been done to fill that area; - It can indicate research focus of each dissertation more clearly; - It can help users to identify relevant documents more easily; - The researchers are more interested in the central problems of the research; - It can give a more straight indication of the main points of the dissertation. - It can provide an introduction to the previous studies. For a new topic, it is more useful.	- It can provide more detailed information about the research; - For in-depth studies, important sentences provided by MEAD are more useful; - It can provide a more complete overview of a topic rather than just research questions.
Negative points	- It is too brief to provide accurate information on the topic. - The simple terms in the variable-based structure are easy to make users confused and lost.	- Only research questions are vague and indistinct.	- It is too complicated and hard to read since MEAD contains too many sentences; - Researchers often have their own opinions to determine important sentences, so that those in MEAD can not cater for each person. - MEAD seems more mixed up and confusing - In most of cases, researchers are not interested in looking for "facts" that MEAD can provide.

The three types of summaries were found to be useful in different aspects. The percentage of researchers selecting the different aspects of usefulness for each of them is given in Table 2. The variable-based summary was found to be useful in the following four aspects:

- *Gives an overview of a research area;*
- *Indicates similarities among previous studies;*
- *Indicates important concepts in the research area;*
- *Indicates important research methods used in the research area;*

In contrast, the research objective and MEAD summaries were found to be less useful in the above aspects. However, they were found to be more useful in the following two aspects:

- *Helps to identify the documents of interest easily;*
- *Indicates important theories, views, or ideas in the area.*

Table 2. Percentage of researchers selecting the different aspects of usefulness for the three types of summaries

Usefulness criteria	Variable-based summary	Research objective summary	MEAD summary
1. Gives you an overview of the research area	**14 (75%)**	11(55%)	10 (50%)
2. Helps you identify research gaps in the area easily	6 (30%)	4 (20%)	3 (15%)
3. Helps you identify the documents of interest easily	10 (50%)	**13 (65%)**	**13 (65%)**
4. Indicates research trends in the area	9 (45%)	8 (40%)	8 (40%)
5. Indicates similarities among previous studies	**12 (60%)**	3 (15%)	2 (10%)
6. Indicates differences among previous studies	5 (25%)	5 (25%)	2 (10%)
7. Indicates important concepts in the area	**15 (75%)**	9 (45%)	8 (40%)
8. Indicates important theories, views, or ideas in the area	7 (35%)	**10 (50%)**	**10 (50%)**
9. Indicates important research methods used in the area	**11 (55%)**	6 (30%)	5 (25%)

- *Bold figures are the higher frequency (percentage) of the summaries for each criterion.*

Conclusion

In this study, three types of multi-document summaries were generated for a set of sociology dissertation abstracts using different summarization methods. The MEAD and research objective summaries were generated using the traditional sentence extraction methods. The variable-based summary was generated using a different summarization method which identified and extracted research concepts and relationships from different abstracts.

These three types of summaries were evaluated by 20 researchers (i.e. human subjects) in the field of sociology. The majority of researchers (70%) preferred the variable-based summaries to the other two sentence-based summaries – 55% of the researchers preferred the research objective summary, and only 25% preferred the MEAD summary. The variable-based summary was found to be more useful in giving an overview of a research area, indicating similarities among previous studies, and indicating important research methods used in the research area. The sentence-based MEAD and research objective summaries were found to be more useful in helping identify the documents of interests easily and indicating important theories, views, or ideas in the area.

This study adopted a simple concept-oriented presentation design for the variable-based summary. In the future, sophisticated presentation designs for operationalizing the variable-based framework can be investigated. One suggestion is to combine the research objective sentences with the variable-based summary to provide alternative information presentation, since 55% of users are interested in research objectives. Another suggestion is to integrate contextual relations and research methods with their corresponding research concepts and relationships, since these two kinds of information complement the information on research concepts and relationships, giving more details of how they are studied.

References

Goldstein, J., Kantrowitz, M., Mittal , V., & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 121-128)*. New York: ACM Press.

Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization (pp.40-48)*. Morristown, NJ: Association for Computational Linguistics.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *IEEE Computer*, 33 (11), 29-36.

Herther, K. (2000). Searching dissertation abstracts: Moving into the digital age. *Information Technology Newsletter*, 5.

Jansen, J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2), 207-227.

Mani, I. (2001). *Automatic summarization*. Amsterdam: John Benjamins Publishing Company.

Moxley, M. (2001). Universities should require electronic theses and dissertations. *Educause Quarterly*, 3, 61-63.

National Institute of Standards and Technology. (2002). In *Document Understanding Conferences 2002*. Retrieved April 24, 2005, from http://www-nlpir.nist.gov/projects/duc/index.html

Ou, S., Khoo, C., & Goh, D. (2003). Multi-document summarization of dissertation abstracts using a variable-based framework. In *Proceedings of the 66th Annual Meeting of the American Society for Information Science and Technology (pp.230-239)*.

Ou, S., Khoo, C., & Goh, D., & Heng, Hui-Hing. (2004). Automatic parsing discourse structure of sociology dissertation abstract as sentence categorization. In *Proceedings of the 8th International Conference of the International Society for Knowledge Organization (pp. 345-350)*. Berlin: Ergon-Verlag.

Ou, S., Khoo, G., & Goh, D. (2005). Developing a summarization method focusing on research concepts and their research relationships. In *Proceedings of the 8th ICADL (pp. 283-292)*.

Radev, D. R., Blitzer, J., Winkel, A., Allison, T., & Topper, M. (2003). MEAD Documentation, Version 3.08.

Stein, G. C., Strzalkowski, T., & Wise, G. B. (2000). Interactive, text-based summarization of multiple documents. *Computational Intelligence*, 16(4), 606-613.

Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results. In *Proceedings of the 8th International World Wide Conference (pp.283-296)*.