Ou, S., Khoo, C. & Goh, D. (2006). Multi-document summarization for digital libraries.

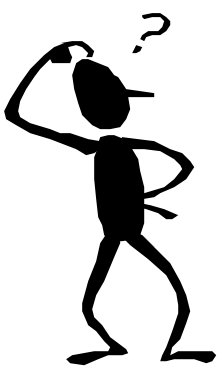
Presented at the Asia-Pacific Conference on Library & Information Education & Practice 2006 (A-LIEP 2006), Singapore, 3-6 April 2006, Nanyang Technological University.

Multi-document Summarization for Digital Libraries

Ou Shiyan, Khoo Christopher, Goh Dion
Division of Information Studies
School of Communication & Information
Nanyang Technological University

1

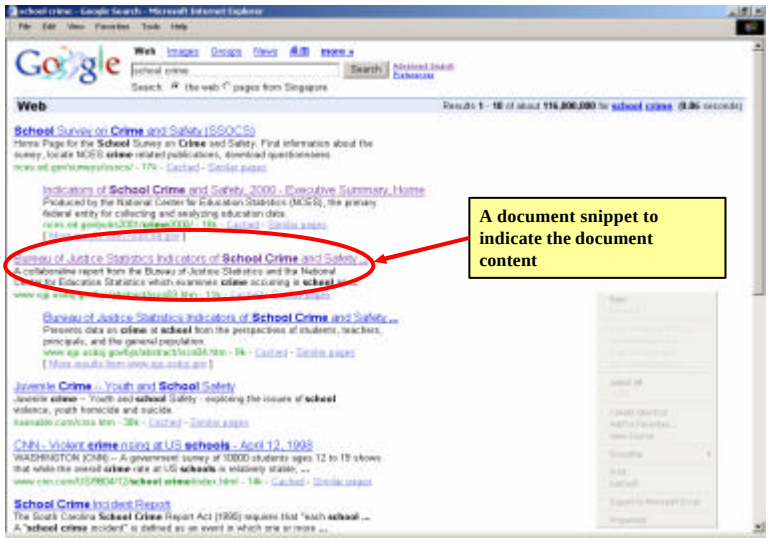
Background



- **How to solve the problem of information overload?**
  - Display a list of titles and short abstracts of retrieved documents (e.g. Google);
  - Group the retrieved documents into folders or categories (e.g. Northern Light search engine);
  - **Construct a multi-document summary of the documents retrieved.**

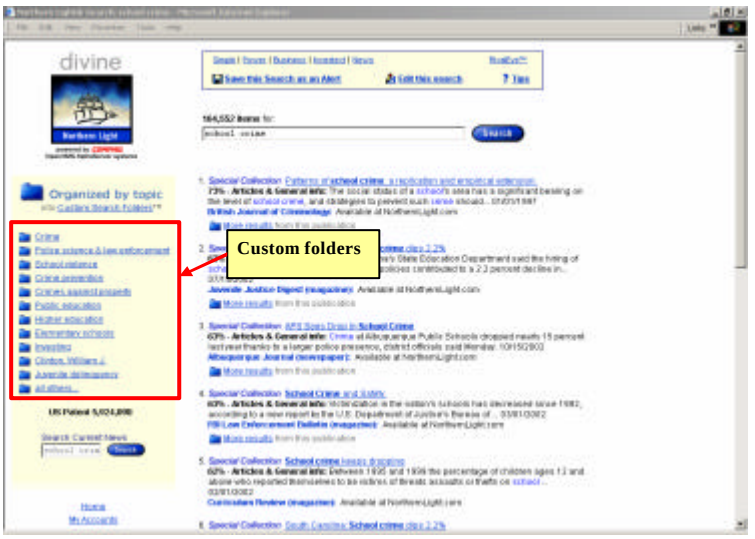
2

A list of document “snippets” returned by Google



3

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



4

Research Objectives

- Develop a method for automatic construction of multi-document summaries of sets of related research abstracts;
- Focus on research concepts (or variables) and relationships;
- Carry out a user evaluation to compare the variable-based summary against two sentence-based summaries.

5

Automatic Summarization

- Process of creating a shorter representation of an original information source;
- Different types of summaries:
  - Extracts vs. Abstracts
  - Indicative vs. Informative summaries
  - Generic vs. User-focused summaries
  - Single-document vs. Multi-document summaries

6

Automatic Multi-document Summarization

- More useful in digital libraries;
- More challenges in the issues of compression, redundancy, cohesion, coherence, and so on;
- New approaches for multi-document summarization
  - Reduce redundancy;
  - Ensure cohesion and coherence;
  - Identify similarities and differences across documents.

7

Summarization Approaches

- Extract important sentences
  - **Strengths:** domain independent; don't need understand the meaning and structure of the text;
  - **Weaknesses:** the resulting extracts are redundant, not coherent and fluent, and hard to read;
- Identify similarities and differences among a set of documents
  - Similarities indicate salient information in a document set;
  - Differences indicate unique information in individual documents;

8

Three Types of Multi-document Summaries

- **MEAD summary:** a sentence-based summary, generated by a state-of-the-art summarization system that extracts the highly ranked sentences across documents;
- **Research objective summary:** a sentence-based summary, generated by extracting research objective sentences from each document;
- **Variable-based summary:** a summary focusing on research concepts often operationalized variables, generated by extracting research concepts and relationships from each document and integrating them across documents.

9

MEAD Summary

- Use a cross-document sentence extraction approach;
- Generated by a state-of-the-art multi-document summarization system, MEAD;
- MEAD, domain public, built by the University of Michigan;
- Cluster similar documents and extract the most important sentences from each cluster using various features;

10

A MEAD summary on the topic of “school crime”

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 affected retention and recruitment in Bellefonte Area School 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 crime reports collected by law enforcement agencies is what many believe may be an underreporting of school crime by school administrators. Many factors exist that may affect 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. [6]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)
- [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. [7]The relationships between school size and school dropout rate and school size and rate of school crime and violence were not statistically significant. (doc-id=5)
- [1]Recent years have seen extensive debate on the multitude of problems plaguing secondary education in the united states, and the problem of school crime and deviance is gaining a sizable share of the attention. [4]I examine the impact of high school restructuring on school delinquency using a broad conception of delinquency that considers both minor and serious juvenile disorders within the school setting. (doc-id=6)
- [1]This study examines factors related to student-initiated serious crime on high school campuses. [3]Using bivariate and multivariate analysis of data drawn from the responses of 889 high school principals, it was found that disorder and lack of parent/student engagement (commitment to school values) displayed the strongest relationship to serious crime. (doc-id=8)
- [1]This study proposes to answer three research questions; namely (1) were the observed frequencies of criminal incidences, per 1,000 students, within expected frequencies based upon school district size and district type (e.g., urban, suburban, non-metropolitan, and rural)? [3](2) Were certain categories of criminal incidences (crimes against persons, crimes against society, crimes against property, and crimes

11

Research Objective Summary

- Use sentence extraction approach;
- Extract research objective sentences from each dissertation abstract;
- Identify research objective sentences using a decision tree classifier and cue phrases found at the beginning of some sentences, such as “*The purpose of the study was to explore*…”, “*This study was to investigate …*”.

12

A research objective summary 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. [2]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. [3]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)
- [4]I examine the impact of high school restructuring on school delinquency using a broad conception of delinquency that considers both minor and serious juvenile disorders within the school setting. [5]My purpose here is to answer the following question: what are the effects of restructuring on school delinquency? [6]The theoretical framework links concepts and variables drawn primarily from social bonding and social disorganization theories of juvenile delinquency to address the problem. (doc-id=6)
- [6]This research posits a theoretical model of school crime based on family, home, economics and demography, in association with minority status and educational elements, exacerbated by drugs, gangs and weapons. [7]Scrupulous attention was paid to the recording of criminal incidents and their associated offenses in the sample districts. [8]A rare opportunity was afforded to analyze empirical data relating to actual criminal offenses on public school campuses. [10]Detailed descriptions are given of numbers of incidents, offenses, ages of offenders, place of offence, sex of offenders, involvement of drugs, gangs and weapons. (doc-id=7)
- [1]This study examines factors related to student-initiated serious crime on high school campuses. (doc-id=8)
- [1]This study proposes to answer three research questions; namely (1) were the observed frequencies of criminal incidences, per 1,000 students, within expected frequencies based upon school district size and district type (e.g., urban, suburban, non-metropolitan, and rural)? [2](2) Were certain categories of criminal incidences (crimes against persons, crimes against society, crimes against property, and crimes against self) more likely to occur in certain areas of the school campus? null [3](3) Were individuals identified as 'repeat offenders' (having committed more than one offense over a three-month period), more likely to commit certain types of crimes? null null (doc-id=9)

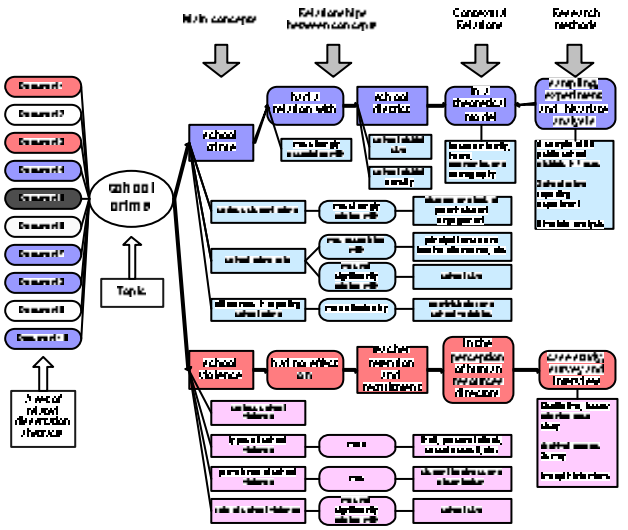
13

Variable-based Summary

- Focus on research concepts and relationships;
- Much of sociology research aims to explore research concepts and relationships (Macionis, 2000).
- The similarities and differences across dissertation abstracts are manly reflected through research concepts and relationships.
- A variable-based framework is proposed for integrating and organizing information extracted from different dissertation abstracts and thus summarizing a set of dissertation abstracts.

14

Variable-based Framework



15

Relationships associated with “school crime” in 5 of 10 dissertation abstracts

- ***School crime*** *was strongly associated with school district size and density;*
- *School size was not significantly related with* ***school crime rate****;*
- ***Serious student crime*** *was strongly related with disorder and lack of parent/student engagement;*
- ***School crime rate*** *was associated with principal tenure and teacher attendance, abolition of corporal punishment etc;*
- ***Differences in reporting school crime*** *was affected by school and administrator variables.*

16

Integrated concepts and relationships associated with “school crime”

- **School crime**, including serious student crime, school crime rate, and so on;

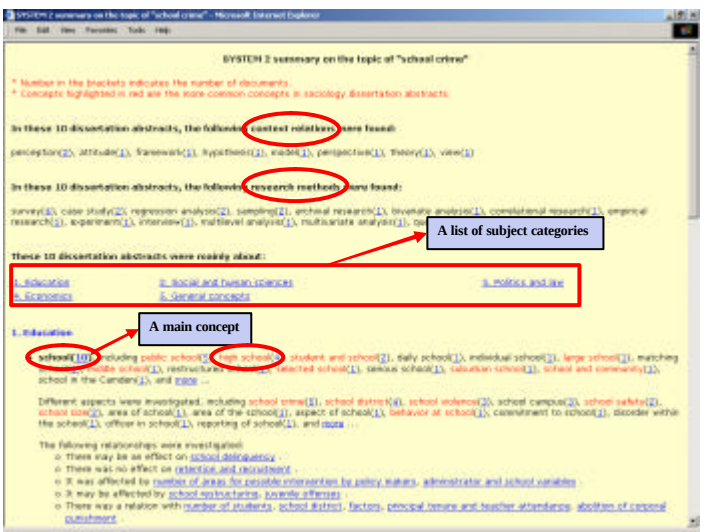
  Its different aspects were investigated, including differences in reporting school crime, and so on.

The following relationships were investigated:

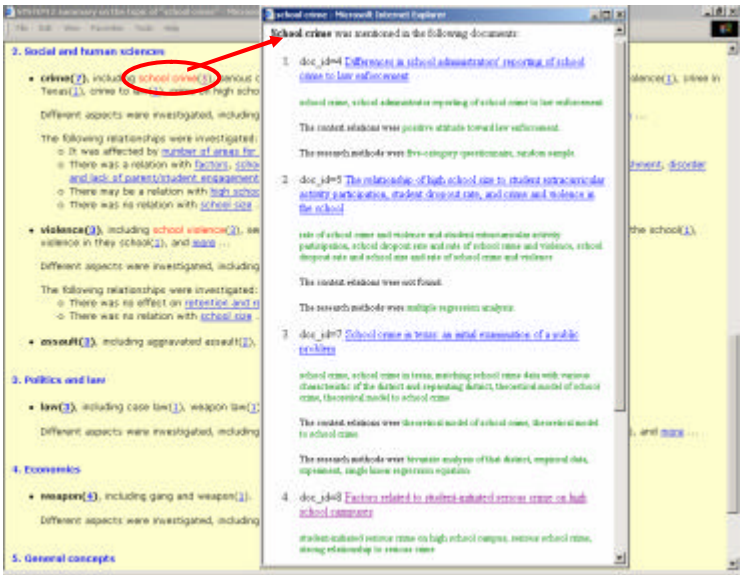
- It was related with school district size and density, principal tenure and teacher attendance, abolition of corporal punishment, disorder and lack of parent/student engagement etc;
- It was not related with school size;
- It was affected by school and administrator variables;

17

A variable-based summary on the topic of “school crime”



18



19

Evaluation Design

- Compare the three types of summaries:
  - Variable-based summary
  - MEAD summary
  - Research objective summary
- 20 researchers in sociology domain as human subjects;
- Three types of summaries for 20 research topics submitted by these 20 researchers;
- Evaluate the three types of summaries on two criteria:
  - Quality: readability and comprehensibility
  - Usefulness for the research-related purpose

20

Evaluation Results (1)

- 14 researchers (70%) indicated their preference for the variable-based summaries;
- 11 researchers (55%) indicated their preference for the research objective summaries;
- 5 researchers (25%) indicated the preference for the MEAD summaries.

21

Evaluation Results (2)

- Variable-based summary:
  - efficient to give an overview of a topic;
  - well-organized and concise;
  - useful for information scanning;
  - too brief to provide accurate information;
  - have the potential to confuse users;
- Research objective summary:
  - more straight indication of the main points of the dissertation;
  - vague and indistinct;
- MEAD summary
  - provide a more complete overview of a topic;
  - hard to read.

22

Evaluation Results (3)

- The variable-based summary was found more useful in the following aspects:
  - Give an overview of a research area;
  - Indicate similarities among previous studies;
  - Indicate important concepts in the research area;
  - Indicate important research methods used in the research area;
- MEAD summary and research objective summary were found to be more useful in the following two aspects:
  - Help to identify the documents of interest easily;
  - Indicate important theories, views, or ideas in the research area.

23

Summary and Conclusion

- Introduce three types of summaries:
  - MEAD summary
  - Research objective summary
  - Variable-based summary
- Evaluate the three types of summaries
  - The majority of users (70%) preferred the variable-based summary to the sentence-based summaries
- Describe a new summarization method
  - Extract research concepts and relationships
  - Integrate similar concepts and same types of relationships
  - Use a concept-oriented presentation design

24

Thank you for your listening

Questions?

25