<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Multi-document summarization of dissertation abstracts using a variable-based framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Ou, Shiyan; Khoo, Christopher S. G.; Goh, Dion Hoe-Lian</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2003</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/6197">http://hdl.handle.net/10220/6197</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td></td>
</tr>
</tbody>
</table>
This paper reports initial work on developing a method for automatic construction of multi-document summaries of sets of domain-specific dissertation abstracts. A variable-based framework for multi-document summarization of dissertation abstracts in the field of sociology and psychology that makes use of the macro-level and micro-level discourse structure of dissertation abstracts as well as cross-document structure is proposed. The micro-level structure of problem statements found in a sample of 50 dissertation abstracts was analyzed, and the common features found are described. A list of indicator phrases that denote different aspects of the problem statements is provided.

Introduction

Automatic summarization has attracted much attention both in the research community and commercially as a solution for reducing information overload and helping users scan a large number of documents to identify documents of interest. 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). A multi-document summary has several advantages over the single-document summary. It provides an integrated overview of a document set indicating common information across many documents, unique information in each document, and cross-document relationships, and can allow users to zoom in for more details on aspects of interest.

Although multi-document summarization can be seen as an extension of single-document summarization, it 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 (based on a document set) and, if presented in a graphical or visual way, can support user browsing and information visualization.

This study seeks to develop a method for automatic construction of multi-document summaries of sets of dissertation abstracts that might be retrieved by a digital library system or search engine in response to a user’s query. Dissertation abstracts were selected for this study because:

- They contain a wealth of high-quality information;
- An increasing number of universities are setting up digital libraries of students’ dissertations;
• Dissertation abstracts have a relatively clear and standard discourse structure;
• The language is more formal and standardized than in other corpora, e.g. news articles, and is more tractable using current natural language processing techniques.

This initial study focuses on sociology and psychology dissertations because of our greater familiarity with dissertations of this nature.

Approaches for Multi-document Summarization

The main approaches used for multi-document summarization include sentence extraction, information extraction, and identification of similarities and differences between documents. With sentence extraction, documents or sentences across all the documents are clustered, following which, a small number of sentences are selected from each cluster (Radev, Jing & Budzikowska, 2000; Otterbacher, Winkel & Radev, 2002; Boros, Kantor & Neu, 2002). In order to generate more coherent summaries, lexical chains are sometimes considered for extracting internally linked sentences instead of separate sentences (Brunn, Chali & Dufour, 2002). Some multi-document summarizers, such as SUMMONS (McKeown & Radev, 1995), RIPTIDES (White et al, 2001) and GITEXTER (Harabagiu & Lacatusu, 2002), use information extraction techniques to extract pieces of information to fill in a pre-defined template. Another important approach for multi-document summarization is to extract information that is common or repeated in several documents plus selected unique information in individual documents (Mani & Bloedorn, 1999). However, these existing summarization approaches focus more on extracting the salient information from different documents. They use shallow analysis, without paying much attention to higher-level semantic content and semantic relations expressed within and across documents. Another problem is that different users have different information needs. Thus, an ideal multi-document summarization should provide different levels of detail for different aspects of the topic according to the user’s interest. But these approaches usually construct fixed multi-document summaries.

One promising approach to summarize multiple documents is to outline the overall structure of a set of related documents to give users an overview of a specific topic first, and subsequently allow users to zoom in to different areas according to their interest. This is performed by analyzing the cross-document structure, which is a kind of discourse structure covering a set of related documents rather than a single document. A discourse structure is a representation of a text from a linguistic/language perspective, and includes all aspects of the internal organizational structure of a discourse ranging from microstructures, such as lexical items and grammatical structure, to macrostructures, such as topics or themes expressed indirectly in larger stretches of text or in the whole discourse (Bell & Garrett, 1998). In a document set retrieved by a digital library system or search engine, the documents are related to each other and address the same or related topics. This allows us to extend the scope of discourse from one document to a document set, and analyze structure beyond the limitations of document boundaries.

One of the most popular discourse theories is Rhetorical Structure Theory (Mann & Thompson, 1988). The rhetorical parsing algorithm proposed by Marcu (Marcu, 1997) is one way to derive the rhetorical structure of unrestricted texts. This algorithm relies on cue phrases and word co-occurrence to derive a rhetorical structure and present it as a rhetorical structure tree.

Although most previous work on discourse-based summarization follows Rhetorical Structure Theory (Mann & Thompson, 1988), Tuefel and Moens (1997) used a different discourse model, namely a rhetorically defined annotation scheme, to code the text of scientific research articles. In this scheme, the extracted sentences in the text are classified into seven categories according to their argumentative roles, reflecting whether the sentence express a main goal of the text, a shortcoming in someone else’s work, etc. The focus of this model is on rhetorical moves which are global to the argumentation of the paper, whereas the Rhetorical Structure Theory-type relations are more local.

Because Rhetorical Structure Theory is limited to single documents, Radev (Radev, 2000) proposed the Cross-document Structure Theory, which is used to describe the rhetorical structure of sets of related documents. Cross-document Structure makes use of a multi-document graph to represent text simultaneously at different levels of granularity (words, phrase, sentences, paragraphs, and documents). It has two types of links: one type represents inheritance relationships among elements within a single document such as words → sentences → paragraphs → documents; a second type of link represents cross-document semantic relationships among text units, such as equivalence (two text spans have the same information content), cross-reference (the same entity is mentioned), contradiction (conflicting information) and historical background (information that puts current information in context). Finally, different summaries are generated based on the user-specified constraints by preserving some of links in the graph, while other links are removed.

However, Radev’s Cross-document Structure Theory focuses more on rhetorical structure and low-level relations in an arbitrary domain, without considering high-level relations and semantics relevant to the domain or type of text. The cross-document relationships are based on physical elements, such as words, phrases, sentences,
paragraphs, and documents, rather than meaningful entities, such as variables in social science research. Furthermore, Radev does not reorganize selected information in a unified manner.

This study will not use the traditional statistics-based summarization approaches. Instead, the new summarization approach seeks to identify the discourse structure peculiar to dissertation abstracts in sociology and psychology, both at macro-level (between sentences and sections) and micro-level (within sentences). It focuses more on the extraction of variables and semantic relations between them expressed in the text, and then integrates extracted information using the cross-document structure.

in the perception of a group of people or in the context of a framework or model.

• **Document Level:** The structured and summarized document that describes a specific relationship between a pair of variables.

A taxonomy is used to identify similar and related variables. Similar variables can be differentiated by the differences in their attributes and qualifiers. Variables can be linked with relationships at different levels of the taxonomy to provide an overview of a specific topic.

**Discourse Analysis**

In this preliminary study, we downloaded 50 Ph.D dissertation abstracts on sociology and psychology from the Dissertation Abstracts International and manually analyzed their structure at the macro-level and micro-level. These dissertations belonged to five topics: school crime, juvenile crime, domestic violence on children, women’s studies and recreation. These topics were selected haphazardly.

**Macro-level Discourse**

The macro-level discourse structure of dissertation abstracts is well-known to researchers and Ph.D students. Nevertheless we analyzed the macro-level structure of 50 dissertation abstracts and found that they typically have the following sections: background/context, problem statement, research method, research results and concluding remarks. These five categories of information reflect different aspects of a research study. All the information in the abstracts analyzed can be subsumed under these five categories, although not every abstract contains all of the categories.

• **Background/Context.** This section introduces the general area of the research problem, explains why it is an important or interesting problem, and identifies studies that are related to the current study.

• **Problem Statements.** This section includes research objectives, research questions, hypotheses, and the adopted theoretical framework. The expected results are sometimes indicated. This section may also provide definitions or explanations of concepts.

• **Research Method.** This section outlines how the study was carried out. It can be decomposed into three subsections: design, sampling and data analysis.

• **Research Results.** This section reports the results of the data analysis (found in the statistical results

**A Hierarchical Variable-based Framework**

To construct multi-document summaries, we propose analyzing the cross-document structure to identify similar information, unique information, and relationships between pieces of information in different documents so that all the information can be synthesized together. For dissertation abstracts in the sociology and psychology domain, the identification of similarities and differences can focus on the research variables and their relationships because social and psychological research aims mainly to explore relationships between variable concepts and social and psychological phenomena. Therefore, a variable-based framework with a hierarchical structure is proposed to integrate variables and relationships between them extracted from different documents and thus summarize a set of dissertation abstracts on a specific topic (see Figure 1). The hierarchy of the framework has four levels:

• **Dependent Variable Level:** All of the dependent variables identified in a group of related dissertation abstracts;

• **Independent Variable Level:** For each dependent variable, the independent variables investigated in the dissertation projects. The relationship between a specific pair of independent and dependent variables is shown.

• **Contextual Relation Level:** Some studies explore variables and relationships of the dissertation abstracts
subsection) and research conclusion (found in the research findings subsection).

- **Concluding Remarks.** This section presents recommendations, future work, or implications of the research results.

We are currently developing an automatic method for identifying these sections of dissertation abstracts using a decision-tree rule induction technique.

---

**Micro-level Structure of Problem Statements**

After analyzing the macro-level discourse structure, the automatic summarizer can focus on some specific sections to extract different kinds of information. In order to develop the summarizer, we need to analyze the micro-level discourse structure of dissertation abstracts to identify the types of information that can be extracted, and indicator phrases that signal the presence of the type of information. In this preliminary study, we analyzed the micro-level structure of the problem statements in the 50 abstracts.

Problem statements are usually focused on variables. A study might investigate only one variable or the relationship between two or more variables. Social research can be divided into three types depending on what kinds of relationships a study aims to explore (Trochim, 1999):

- **Relational research:** two or more variables are investigated at the same time to see if there is any relationship between them;
- **Causal research:** variables are manipulated by the researcher to see how they affect other variables;
- **Descriptive research:** variables are investigated to identify attributes of interest.

In causal research, one or more variables are designated as the dependent variable (DV) while another group of variables are designated the independent variables (IVs). DVs are the variables the researchers are interested in explaining or predicting, while IVs are variables that affect or are used to predict the DVs. In relational research however, variables are not distinguished as such.

To understand the micro-level structure of problem statements in dissertation abstracts, we analyzed the sample abstracts to identify the main types of semantic relations found in them. These are represented in linear conceptual graph notation (Sowa, 1984) as follows:

- **Relational research:** [variable 1] -> (relation with. *) -> [variable 2]
- **Causal research:** [IV] -> (effect on. *) -> [DV]
- **Descriptive research:** [IV] -> (attribute. *) -> [*]

---

![Figure 1. The multi-document summary on the topic of “school crime” based on the framework](image-url)
In conceptual graph notation, concepts are represented in square brackets and relations are represented in round brackets with arrows indicating the direction of the relation. Concepts and relations can have subtypes. An asterisk indicates that the subtype of the concept or relation is unknown and is to be determined in the research study. Different from relational and causal research, descriptive research does not investigate any relation between variables, but seeks to identify attributes of the variables investigated.

While many studies aim to explore relationships directly, some studies explore relations in the context of a framework, model, theory, hypothesis etc., or in the perception or attitude of a target population. For example, “The purpose of this qualitative, descriptive study was to examine mothers’ perception of how their children are affected by exposure to domestic violence.” We call this a contextual relation. The types of contextual relations identified are given in Table 1.

A variable concept can be of different types:

- an object (such as school, family)
- an abstract object (such as training program, distance)
- an event (such as violence, crime)
- an action (such as participation, recruitment)
- a human agent (subject who executes an action)
- a human patient (object of an action)
- a process (such as change, increase, growth, variation)

All of these may be related to other concepts which may specify an attribute of the variable concepts or qualify the variable concept. An attribute is a quality that is part of the nature of entity denoted by the concept. The attribute may be unknown and needs to be investigated in the study. For example, in “school size”, “size” is the attribute of an object “school”:

\[
\text{[school]} - \text{(attribute)} -> \text{[size]}
\]

A qualifier, on the other hand, restricts or narrows the meaning of the concept. For example, in “serious school crime”, the variable is an event “crime” whose scope is narrowed down by a location qualifier “school” and a degree qualifier “serious”:

\[
\text{[crime]} - \text{ (qualifier. location)} -> \text{[school]}
\]
\[
\text{ (qualifier. degree)} -> \text{[serious]}
\]

In addition, an “action” concept also can be related to an agent or patient. For example, in “student participation”, the action “participation” is executed by the agent “student”, while in “teacher recruitment”, the action “recruitment” is performed on the patient “teacher”:

\[
\text{[participation]} - \text{ (agent)} -> \text{[student]}
\]
\[
\text{[recruitment]} - \text{ (patient)} -> \text{[teacher]}
\]

Table 2 and Table 3 list the concept attributes and qualifiers found in the sample abstracts. A complete example of the micro-level structure of a problem statement is given in Figure 2.

**Indicator Phrases for Information Extraction**

To automatically parse the discourse structure of dissertation abstracts and extract text segments from...
documents corresponding to different parts of the discourse structure, indicator phrases are used, which are associated with and that often signal the type of information in the text. From indicator phrases, patterns can be constructed and used by the summarizer to extract the information from the text. In this study, we focused on problem statements and analyzed their characteristics to derive patterns from them.

In the 50 sample abstracts, almost all the problem statements contain some indicator phrase at the beginning of the sentence such as "The purpose of this study was to investigate ...", "This research was designed to examine...". Although these phrases do not carry any significant content, they are very useful for identifying whether the sentence is a problem statement. Other indicator phrases are given in Table 4.

Table 5 lists indicator phrases that indicate the type of relation between variables.
Table 2. Attributes found in sample abstracts

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>nature</td>
<td>the nature of sibling relationships in families with a history of violence</td>
</tr>
<tr>
<td>characteristics</td>
<td>characteristics of single African-American welfare recipient mothers</td>
</tr>
<tr>
<td>size</td>
<td>school district size</td>
</tr>
<tr>
<td>density</td>
<td>school district density</td>
</tr>
<tr>
<td>rate</td>
<td>rate of school crime</td>
</tr>
<tr>
<td>pattern</td>
<td>the patterns of fathers' touch with low birth weight infants</td>
</tr>
<tr>
<td>quality</td>
<td>the quality of fathers' touch with low birth weight infants</td>
</tr>
<tr>
<td>parameters</td>
<td>parameters of the legal authority of the officers</td>
</tr>
<tr>
<td>type</td>
<td>types of violence</td>
</tr>
<tr>
<td>prevalence</td>
<td>prevalence of violence</td>
</tr>
<tr>
<td>function</td>
<td>if father visitation is more helpful or harmful in domestic violence families</td>
</tr>
<tr>
<td>frequency</td>
<td>frequencies of criminal incidences</td>
</tr>
<tr>
<td>facilitator</td>
<td>facilitators to collaboration</td>
</tr>
<tr>
<td>barrier</td>
<td>barriers to collaboration</td>
</tr>
<tr>
<td>length</td>
<td>length of (children’s) exposure to violence</td>
</tr>
<tr>
<td>category</td>
<td>categories of crimes incidences</td>
</tr>
<tr>
<td>ability</td>
<td>children's ability to manage emotional expressions</td>
</tr>
<tr>
<td>possibility</td>
<td>the possibility of measuring internalizing behaviors in preschool children</td>
</tr>
<tr>
<td>likelihood</td>
<td>the likelihood of engaging in delinquent behavior</td>
</tr>
</tbody>
</table>

Table 3. Qualifiers found in sample abstracts

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree</td>
<td>serious school violence</td>
</tr>
<tr>
<td>location</td>
<td>violence in three high schools in the Camden City Public School District</td>
</tr>
<tr>
<td>scope</td>
<td>officers employed by these programs</td>
</tr>
<tr>
<td>time</td>
<td>performance parameters during 4-h of prolonged steady state cycling followed by a maximal performance time trial</td>
</tr>
<tr>
<td>purpose</td>
<td>a curriculum for leisure transition planning</td>
</tr>
<tr>
<td>aspect</td>
<td>football fans’ attitudes toward televised football</td>
</tr>
</tbody>
</table>

An Example of Extracted Information

This section presents an example of the information that can be extracted from 10 dissertation abstracts on the topic of “school crime” based on our manual analysis.

The extracted dependent variables are:

- criminal incidences
  - frequencies of criminal incidences
  - categories of criminal incidences
  - types of offences
- parameters of legal authority of the officers
- school crime
  - rate of school crime
  - school violence
    - rate of school violence
    - types of school violence
    - prevalence of school violence
  - school delinquency

- serious student crime
- differences in reporting school crime

- school dropout rate
- student extracurricular activity participation
- teacher recruitment and retention
  - teacher recruitment
  - teacher retention
  - districts’ future plans regarding teacher retention and recruitment

The extracted independent variables are:

- areas of the school campus
- attendance
  - student attendance
  - teacher attendance
- different districts
- different schools
- different administrators
Problem statements
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.

Conceptual representation
1: Main relations: the effects of serious school violence on teacher retention and recruitment
   Causal research: [IV] -> (effect on. *) -> [DV]
   [IV. serious school violence] -> (effect on. *) -> [DV. teacher retention] .................... (#1)
   [IV. serious school violence] -> (effect on. *) -> [DV. teacher recruitment] .................... (#2)

2: Contextual relations: the perception of public school districts’ human resources directors
   [*] <- (agent) <- [action] -> (patient) -> [*]
   [ public school districts’ human resources directors] <- (agent) <- [perception] -> (patient) -> (#1)
   -> (#2)

3: Relations in variables
   (1) serious school violence
      [violence] -
      (qualifier) -> [school]
      (qualifier) -> [serious]

   (2) teacher retention
      [retention] -> (patient) -> [teacher]

   (3) teacher recruitment
      [recruitment] -> (patient) -> [teacher]

   (4) public school districts’ human resources directors
      [directors] -
      (qualifier) -> [human resources]
      (qualifier) -> [public school districts]

Figure 2. A complete example of the micro-level structure of a problem statement

Table 4. Indicator phrases for identifying problem statements

<table>
<thead>
<tr>
<th>Indicator Phrase</th>
<th>Problem Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;purpose&gt; * … be to …&lt;investigate&gt; … aim</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>goal examine</td>
</tr>
<tr>
<td></td>
<td>objective answer</td>
</tr>
<tr>
<td></td>
<td>intent discuss</td>
</tr>
<tr>
<td></td>
<td>to propose to</td>
</tr>
<tr>
<td></td>
<td>wish to</td>
</tr>
<tr>
<td></td>
<td>be designed to</td>
</tr>
<tr>
<td></td>
<td>be conducted to</td>
</tr>
</tbody>
</table>

* < > indicates this word or phrase has synonyms and some of synonyms are listed under it.
Table 5. Relations between variables and selected indicator phrases

<table>
<thead>
<tr>
<th>Indicator Phrase</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>descriptive research</td>
<td>This study examines the phenomenon in three contexts: (a) tourism, (b) postmodern Israel, and (c) the journey as a rite of passage.</td>
</tr>
<tr>
<td>… &lt;phenomenon&gt; * … concept experience performance</td>
<td>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. The present study sought to examine the associations between acculturation status and cultural beliefs ……</td>
</tr>
<tr>
<td>relational research</td>
<td>I examine the impact of high school restructuring on school delinquency ……</td>
</tr>
<tr>
<td>… &lt;relationship&gt; &lt;between&gt; … relation with association among link</td>
<td>This study explored factors that influence teachers' attitudes: age, gender, education, ethnicity, and teaching experience. This study examined school administrator differences in reporting school crime based on selected school and administrator variables.</td>
</tr>
<tr>
<td>causal research</td>
<td></td>
</tr>
<tr>
<td>… &lt;impact&gt; … on … effect</td>
<td></td>
</tr>
<tr>
<td>… &lt;influence&gt; … affect</td>
<td></td>
</tr>
<tr>
<td>… &lt;difference&gt; in … &lt;based on&gt; … on the basis of according to</td>
<td></td>
</tr>
</tbody>
</table>

*<>* indicates this word or phrase has synonyms and some of synonyms are listed under it.

- disadvantages
- disorder
- district-wide changes with desegregation, grade structure shifts, and the abolition of corporal punishment
- ethnicity
- expenditure
- level of socioeconomic deprivation in the surrounding community
- minority culture classes
- principal tenure
- parent/student engagement(commitment to school values)
- repeated offenders
- safeguarding the constitutional rights of students as United States citizen
- serious school violence
- student mobility
- school district
  - school district size
  - school district density
  - school district type
- school size
- school decreasing and restructuring
  - school decreasing
  - school restructuring

Dependent and independent variables are grouped hierarchically with the analysis of variable components and the help of the taxonomy. For example,

- “categories” and “frequencies” refer to the different attributes of the variable “criminal incidences”;
- “violence”, “delinquency” and “serious crime” are narrower concepts of the broader concept “crime”. Therefore, “school violence”, “school delinquency” and “serious student crime” are grouped under the variable “school crime”.

The extracted relationships between variables are shown in Table 6.

**Conclusion**

This paper proposes a variable-based framework for summarizing multiple dissertation abstracts through analyzing the micro-level and macro-level discourse structure, extracting variables and relations between them, and integrating the information extracted from different documents. The framework has two advantages: it gives the user a map or overview of a particular topic which the user can explore and zoom in for more details. The results of an analysis of the micro-level structure of the problem
Table 6. Extracted relationship between dependent variables and independent variables

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Relationship</th>
<th>Independent variable</th>
<th>Contextual Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>teacher retention and recruitment</td>
<td>is not affected by serious school violence</td>
<td>serious school violence</td>
<td>in the perception of public school districts’ human resources directors</td>
</tr>
<tr>
<td>teacher recruitment</td>
<td>is not affected by serious school violence</td>
<td>different districts</td>
<td></td>
</tr>
<tr>
<td>districts’ future plans regarding teacher retention and recruitment</td>
<td>is not affected by different districts</td>
<td>different districts</td>
<td></td>
</tr>
<tr>
<td>rate of school crime</td>
<td>is associated with principal tenure student attendance teacher attendance student mobility the district-wide changes</td>
<td>principal tenure student attendance teacher attendance student mobility the district-wide changes</td>
<td></td>
</tr>
<tr>
<td>rate of school crime</td>
<td>is strongly associated with school district size school district density</td>
<td>in a theoretical model</td>
<td></td>
</tr>
<tr>
<td>rate of school crime</td>
<td>is not associated with school size</td>
<td>school size</td>
<td></td>
</tr>
<tr>
<td>school violence</td>
<td>are theft, personal attach, sexual assault, bully behavior, consuming alcohol, graffiti, drugs, and weapons, physical confrontations</td>
<td>as perceived by students, staff, and a Gang Abatement Task Force</td>
<td></td>
</tr>
<tr>
<td>prevalence of school violence</td>
<td>is student tardiness and absenteeism</td>
<td>school size</td>
<td></td>
</tr>
<tr>
<td>school delinquency</td>
<td>is affected by school decreasing</td>
<td>school decreasing</td>
<td>based on a theoretical framework</td>
</tr>
<tr>
<td>school delinquency</td>
<td>is not affected by school restructuring</td>
<td>school restructuring</td>
<td></td>
</tr>
<tr>
<td>school delinquency</td>
<td>is affected by the level of socioeconomic deprivation in the surrounding community</td>
<td>the level of socioeconomic deprivation in the surrounding community</td>
<td></td>
</tr>
<tr>
<td>serious student crime</td>
<td>is not affected by ethnicity minority culture classes expenditure disadvantage</td>
<td>ethnicity minority culture classes expenditure disadvantage</td>
<td></td>
</tr>
<tr>
<td>serious student crime</td>
<td>is strongly associated with disorder parent/student engagement (commitment to school values)</td>
<td>disorder parent/student engagement (commitment to school values)</td>
<td></td>
</tr>
<tr>
<td>differences in reporting school crime</td>
<td>are caused by different schools</td>
<td>different schools</td>
<td></td>
</tr>
<tr>
<td>differences in reporting school crime</td>
<td>are caused by different administrators</td>
<td>different administrators</td>
<td></td>
</tr>
<tr>
<td>legal authority</td>
<td>is balanced with safeguarding the constitutional rights of students as United States citizen</td>
<td>safeguarding the constitutional rights of students as United States citizen</td>
<td></td>
</tr>
<tr>
<td>extracurricular activity</td>
<td>is associated with school size</td>
<td>school size</td>
<td></td>
</tr>
<tr>
<td>school dropout</td>
<td>school dropout rate is not associated with school size</td>
<td>school dropout rate</td>
<td></td>
</tr>
<tr>
<td>criminal incidences</td>
<td>are higher in large school district size</td>
<td>large school district size</td>
<td></td>
</tr>
<tr>
<td>criminal incidences</td>
<td>are higher in urban and suburban school district type</td>
<td>urban and suburban school district type</td>
<td></td>
</tr>
<tr>
<td>criminal incidences</td>
<td>are not associated with areas of the school campus</td>
<td>areas of the school campus</td>
<td></td>
</tr>
<tr>
<td>criminal incidences</td>
<td>types of offences (simple assault, disorderly conduct, intimidation, weapon law violations, aggravated assault, and motor vehicle theft) are more likely committed by repeated offenders</td>
<td>repeated offenders</td>
<td></td>
</tr>
</tbody>
</table>
statements section of dissertation abstracts based on a sample of 50 abstracts are reported. A list of indicator phrases that denote different aspects of the problem statements is provided. These indicator phrases can be used later for automatic extraction of different pieces of information from the abstracts.

We are in the process of analyzing a much bigger sample of sociology abstracts to further develop and validate the macro-level and micro-level structure of dissertation abstracts. We also plan to construct text patterns for automatic parsing and extraction of information. A taxonomy of social science concepts will also be developed to interpret the meaning of concepts and identify similar concepts for the purpose of integrating the variables extracted from different documents. Visualization methods for presenting the synthesized information will also be explored to identify features best suited to represent the variable-based framework and facilitate users’ overall understanding of a set of dissertation abstracts.

References


