ABSTRACT
A method was developed for automatically aggregating sets of research abstracts in sociology retrieved by an information retrieval (IR) system in response to a user query. The method focused on extracting and integrating research concepts, relationships, research methods and contextual information from different abstracts retrieved by the IR system, and organizing and displaying them in a Web interface. Sociology dissertation abstracts were selected as the sample domain in this study. A user evaluation was carried out to evaluate the effectiveness of the aggregation method. Two concept-based presentations—with or without the use of a taxonomy—were compared against a sentence-based presentation that lists only the research-objective sentences extracted from the abstracts and another sentence-based presentation generated using the MEAD system. A sample of 40 users found the concept-based presentation organized using a taxonomy to be more comprehensible, readable and useful than the sentence-based presentations. The majority of users (64%) also indicated a preference to use the concept-based system.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Clustering. H.3.1 [Content Analysis and Indexing]: Linguistic Processing.

General Terms
Design, Experimentation, Human Factors.

Keywords
Aggregated search, multidocument summarization, social science, information integration, information extraction.

1. INTRODUCTION
With the rapid growth of research literature, information retrieval systems, digital libraries and search engines are retrieving more and more records in response to user queries. Users need better tools to visualize and obtain an overall grasp of the literature retrieved by the system, and quickly zoom in to useful documents. Users also need more tools to help them manage the documents retrieved and to process the content of the documents in various ways, e.g. extracting useful concepts and information from the retrieved documents.

Most information retrieval systems and Web search engines rank documents retrieved by likelihood of relevance, and displaying titles and short abstracts to give users some indication of the document content. Though abstracts are helpful, it is tedious for the user to scan a large number of abstracts.

Related documents retrieved by an information retrieval (IR) system are likely to share some amount of common information as well as contain unique information. There is a need for the IR system to summarize common information and highlight unique or contrasting information in individual documents. An effective presentation of search results should provide an overview of the topic by indicating what is similar and different in different documents and the relationships between pieces of information across documents, and allow the user to zoom in for more details on aspects of interest.

This study sought to develop a method for aggregating search results of research abstracts returned by an IR system. The study focused on informative research abstracts since such abstracts often have a clear structure and provide a summary of the research findings. Sociology dissertation abstracts were selected as the domain for the study because much of sociology research adopts the traditional quantitative research paradigm of looking for relationships between concepts. Sociology dissertation abstracts are well-structured and have the classical research report structure with five standard sections: background, research objectives, research methods, research results and concluding remarks. Many other domains, such as psychology, medicine, crop agriculture and chemistry, adopt this research paradigm and report structure. Although some sociology research is carried out using the qualitative research paradigm, focusing on description, interpretation and explanation, many of these studies also seek to identify...
relationships between concepts representing events, behaviors, attributes, and situations.

A concept-based hierarchical framework was proposed to integrate and organize similarities and differences among the different dissertation abstracts, focusing on research concepts and relationships investigated in the studies. A user evaluation, involving 20 search queries and 40 human users, was carried out to evaluate the aggregated presentation.

A detailed description of the system has been reported in [1] and an in-depth user evaluation reported in [2]. This short paper presents an overview of the method used to generate an aggregated view of the search results and a quick evaluation involving 40 graduate students in an MSc in Information Studies program.

2. PREVIOUS WORK
How to present a set of retrieved documents in fluent text and in a form that is useful to users is an important issue. Most existing Web search engines and information retrieval systems extract some important sentences or create new sentences to indicate the content of each returned record. These extracted sentences and newly created sentences have to be arranged in a particular order that will make sense to the user. Most studies arranged them in the same order as in the document text or in a chronological order to generate a summary. Some studies also organized the sentences in other formats to facilitate user reading and understanding. For example, Farzindar and Lapalme [3] presented the extracted sentences in a tabular format that was divided by such themes as decision data, introduction, context, juridical analysis and conclusion found in legal text. Although sentence-oriented presentation is extensively used in displaying search results by most Web search engines and information retrieval systems, a few studies have presented concepts (terms) in addition to sentences as the summary. Aone et al. [4] presented a summary of a document in multiple dimensions through a graphical user interface. A list of keywords (i.e. person names, entity names, place names and others) was presented in the left window for quick browsing. The full text was presented in the right window, in which the extracted summary sentences were highlighted. Ando et al. [5] identified multiple topics in a set of documents and presented the summary by listing several terms and two sentences that were most closely related to each topic.

3. EXTRACTION AND INTEGRATION OF CONCEPTS AND RELATIONS
Aggregation of search results can be divided into two main phases: 1) selection, extraction and integration of concepts and relations from the retrieved abstracts/documents; and 2) structuring and presentation of aggregated results in the interface. In this section, we describe our approach in selecting, extracting and integrating the concepts and relations from retrieved abstracts.

Much of sociological research aims to explore research concepts and the relationships between them found in social phenomena [6]. In a set of abstracts/studies on a specific topic, some studies might investigate similar concepts but focusing on different attributes of the concept or relationships with other concepts using different research methods and in different contexts. Thus the representation of the research concepts and relationships extracted from the research abstracts should reflect such similarities and differences among the abstracts.

To aggregate the concepts and relations extracted from a set of related abstracts, a hierarchical framework is used. The framework contains four kinds of information:

- **Main concepts**: The common concepts investigated in the studies.
- **Research relationships between concepts**: For each main concept, the attribute values or relationships with other concepts investigated in the studies.
- **Contextual relations**: The context in which the concepts and their relationships are situated—usually the perception of a group of people or a theoretical framework or model.
- **Research methods**: One or more research methods used.

The hierarchical framework represents the summarized information at the top level with more detailed information at the lower levels. Similar concepts extracted from different abstracts are clustered and summarized by a broader concept called main concept. For each concept, its attribute values or relationships with other concept(s) are represented, together with the contextual relations and research methods used. All the relationships involving the same main concept are combined and summarized using a simple, standard expression (i.e. controlled vocabulary). The contextual relations and research methods are also summarized using simple, uniform terms.

Figure 1 illustrates some of the information extracted from 10 dissertation abstracts on the topic of “school crime”. Two main concepts are presented: “school crime” and “school violence”. The concept “school crime” was investigated in five documents whereas “school violence” was investigated in three documents.

“School violence” includes the following four sub-level concepts:
- Serious school violence
- Types of school violence
- Prevalence of school violence
- Rate of school violence

These sub-level concepts may represent a subset (e.g. “serious school violence”) of the main concept “school violence” or specify one of its aspects or characteristics (e.g. type, rate, prevalence). Their relationships with other concepts were investigated in individual studies as follows:

- School violence had no effect on teacher retention and recruitment.
- Types of school violence were theft, personal attack, sexual assault, etc.
- Prevalence of school violence was student tardiness and absenteeism.
- Rate of school violence was not significantly related with school size.

Some relationships were investigated based on a contextual relation and using particular research methods. For example,
- The relationship between “school violence” and “teacher retention and recruitment” was investigated in the perception of human resources directors. The research methods used were:
The framework presents a full map of a specific topic by integrating research concepts and relationships as well as contextual relations and research methods extracted from different dissertation abstracts using a hierarchical structure and organizing them based on the main concepts. It has two advantages: (a) it provides an overview of a subject area by presenting the summarized information at the top level and (b) it allows users to zoom in to more details of interest by exploring the specific information at the lower levels. The framework provides a way to aggregate a set of dissertation abstracts retrieved in response to a search query. This is different from the traditional search result presentation which ranks retrieved records according to their relevance to the query and displays a short, sentence-based abstract to indicate the content of each record.

### 4. PROCESSING OF SEARCH RESULTS

Based on the framework, a set of abstracts retrieved by the information retrieval system is processed in four major steps:

1. Parse each abstract to identify which sections of the abstract contain the information to be extracted.
2. Extract the four kinds of research information (research concepts, relationships, contextual information and research methods) from each abstract.
3. Integrate each kind of information extracted from different abstracts.
4. Combine and organize the four kinds of information, and present them in an interactive Web-based interface.

In discourse parsing, each dissertation abstract was segmented into the five sections—background, research objectives, research methods, research results and concluding remarks. We treated discourse parsing as a text categorization problem — assigning each sentence in a dissertation abstract to one of the five sections. A decision tree classifier that made use of sentence position and indicator words present in the sentence was developed and used in the sentence categorization [7]. In an evaluation involving four
human coders, the system obtained an average of 63.4% agreement with the coders. However, only two sections—the research objectives and research results sections—were used by the system to extract research information for presentation to the user. Thus the identification of these two sections was more important than the other sections. The system had a high accuracy of 90.8% in identifying the research objectives and research results sections.

In information extraction, four kinds of information—research concepts and relationships, contextual relations, and research methods—were extracted from each dissertation abstract. Linguistically, research concepts, contextual relations, and research methods appear as nouns or noun phrases. A list of syntactic rules specifying the possible sequences of part-of-speech tags in a noun phrase was defined and used to identify sequences of contiguous words that were potential noun phrases. Research-concept terms were selected from the research-objectives and research-results sections since these two sections are most likely to contain more important research information. Research-method and contextual-relation terms were identified using indicator phrases. To extract research relationships between concepts, linguistic patterns indicating various kinds of relationships were constructed manually based on a sample of 300 dissertation abstracts. The linguistic patterns used are regular expression patterns, each comprising a sequence of tokens with two or three slots. Each token which is not a slot is constrained with a part-of-speech tag. For example, “[slot: independent variable] have DET (ADJ) effect/influence/impact on/in [slot: dependent variable]” is a pattern describing one way in which the cause–effect relationship can be expressed in the text. A pattern-matching program was developed to identify the text segments in the sentences that match with each relationship pattern. The terms in the text that matches with the slots in a pattern represent the research concepts connected by the relationship.

In information integration, similar concepts and relationships extracted from different abstracts were integrated using concept generalization and relationship conflation. Similar concepts were identified and clustered according to their syntactic variations. Terms of different word lengths which follow specific syntactic variation rules were considered term variants and represented similar concepts at different generalization levels. For example:

- student → undergraduate student → black undergraduate student → adjustment of black undergraduate student → college adjustment of black undergraduate student
- student → behavior of student → delinquent behavior of student → prevention of delinquent behavior of student

These hierarchical term chains were formed by linking shorter term variants to longer term variants, and thus a group of similar concepts from the nodes of the chain was derived. Concepts at the lower level can be generalized by the broader concepts at the higher level. Chains sharing the same root node were combined to form a hierarchical cluster tree which represented a cluster of similar concepts sharing the same cluster label.

5. Presentation of Aggregated Search Results

In the presentation of the aggregated search results, concepts at the higher levels in a cluster were selected and combined in a summary sentence. For example, all the concepts relating to “student” can be presented by selecting the main concept at the top level and the concepts at the second level:

- Student, including college student, undergraduate student, American student, and so on.
- Its different aspects are investigated, including characteristics of student, behavior of student, and so on.

The second-level concepts are divided into two types: subclass concepts and facet concepts. Subclass concepts represent subclasses of the main concept whereas facet concepts represent aspects or characteristics of the main concept. Thus, the sentence is divided into two parts: the first part for subclass concepts (“including …”) and the second part for facet concepts (“its different aspects are investigated, including …”). For a cluster of similar concepts, their relationships with other concepts were integrated together to provide an overview of all associated concepts connected by various types of relationships. Each type of relationship (e.g., correlation and cause–effect relation) was identified using a group of linguistic patterns. For the same type of relationships, the linguistic expressions were normalized and conflated using a standard expression. For example, the following two relationships are associated with the concept “student”:

- Expected economic returns affected the college students’ future career choices.
- School socioeconomic composition has effect on Latino students’ academic achievement.

These two relationships are normalized and conflated into a simple sentence as follows:

- Some facets of students were affected by expected economic returns and school socioeconomic composition.

In the presentation interface, the four kinds of extracted information are combined and organized based on the hierarchical framework and presented in an interactive Web-based interface with three levels:

- The top level—aggregated and summarized information for research concepts and relationships, contextual relations and research methods from different dissertation abstracts;
- The second level—list of document titles with the four kinds of information extracted from each abstract; and
- The third level—the original abstracts.

The three levels are presented in different windows linked by hyperlinks. The summarized information at the top level is presented in the main window to provide an aggregated summary of the retrieved abstracts. The user can click on the hyperlinks to access the more detailed information at the lower levels.

We developed two versions of the aggregated interface: (1) AGGREGATE SYSTEM 1 as described earlier; (2) AGGREGATE SYSTEM 2 that makes use of a taxonomy to structure the aggregated information (see Figure 2). This semi-automatically constructed taxonomy contains lists of important n-word terms (n = 1, 2, 3, 4, and 5) in the sociology domain. Details of how the taxonomy was constructed and used was reported in [8]. Its function is to filter out non-concept terms, specify the important concepts in the sociology domain, and categorize concepts into different subject areas.

As shown in Figure 3, the four kinds of research information are organized to give users an overview for each kind of information. The contextual relations, research methods and research concepts...
extracted from different abstracts are presented as concept clusters, whereas the research relationships are presented as simple sentences. In the display, contextual relations and research methods are presented first followed by research concepts and relationships. This is because contextual relations and research methods are usually quite short and may be overlooked by the user if presented at the bottom of the window.

The concept clusters are categorized into broad subject categories determined by the taxonomy. A list of subject categories can give the user an initial overview of the range of subjects covered by the search results and help the user to locate subjects of interest quickly, thus facilitating the user’s browsing. Furthermore, the important concepts in the sociology domain (as determined by the taxonomy) are highlighted in red.

Each cluster is labeled by a main concept and divided into two subgroups—one for subclass concepts and another for facet concepts. For each concept, the number of documents indicated in parentheses is clickable and is linked to a list of summarized documents in a pop-up window, giving the title, research concepts, contextual relations, and research methods for each document (see Figure 3). The title of the document also is clickable and is linked to the original dissertation abstract in another pop-up window.

6. EVALUATION

To evaluate the effectiveness of the aggregated presentation of search results, a user evaluation was carried out. 20 research topics were obtained from 20 researchers in the field of sociology. For each topic, a set of PhD sociology dissertation abstracts was retrieved from the Dissertation Abstracts International database using the topic as the search query, and at most 200 abstracts were aggregated and presented in one page.

For each topic, the following four types of presentations of aggregated results were provided:

- AGGREGATE SYSTEM 1 as described earlier;
- AGGREGATE SYSTEM 2 described earlier;
- RESEARCH OBJECTIVES interface: a sentence-based presentation generated by displaying the research objective sentences extracted from each abstract;
- MEAD SYSTEM: A sentence-based presentation generated by MEAD 3.08, a multi-document summarization system built by the University of Michigan using a centroid-based cross-document sentence extraction method [9]. The summary consists of sentences ranked as important based on statistical and linguistic features.

The overall readability, comprehensibility, and usefulness were scored on a 7-point scale by 40 students in the MSc Information Studies program at the Nanyang Technological University, Singapore. The average scores for the four types of presentations are shown in Table 1.

AGGREGATE SYSTEM 2 obtained the highest readability score (5.0), the highest comprehensibility score (4.6), and the highest usefulness score (4.6). Its scores were much better than those for...
Table 1. Average evaluation scores for the four types of presentations of aggregated search results (N=40)

<table>
<thead>
<tr>
<th></th>
<th>AGGREGATE 1</th>
<th>AGGREGATE 2</th>
<th>OBJECTIVES</th>
<th>MEAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>3.90</td>
<td>5.00</td>
<td>3.78</td>
<td>3.80</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>3.88</td>
<td>4.58</td>
<td>3.85</td>
<td>3.73</td>
</tr>
<tr>
<td>Usefulness</td>
<td>3.73</td>
<td>4.57</td>
<td>3.73</td>
<td>3.76</td>
</tr>
</tbody>
</table>

AGGREGATE SYSTEM 1, which indicates that with the use of a taxonomy for information filtering and organization, readability and comprehensibility can be substantially improved. The three systems, AGGREGATE SYSTEM 1, RESEARCH OBJECTIVES and MEAD, have about the same evaluation scores.

The 40 users were asked to rank the four presentations. A weighted rank score was calculated for each presentation. A weight of 4 was assigned to the first rank, 3 to the second rank, 2 to the third rank, and 1 to the last. The users were also asked to select one or more presentations that they preferred to use for obtaining general information on a topic. The ranks and the users’ preferences are shown in Table 2.

The overall ranking for the four types of presentations based on the weighted rank scores was:

1- AGGREGATE 2
2- AGGREGATE 1
3- RESEARCH OBJECTIVES
4- MEAD

64% of the users indicated preference for AGGREGATE 2 for obtaining general information on the topic.
7. CONCLUSION AND FUTURE WORK

This study developed a method for automatically aggregating sociology abstracts retrieved by an information retrieval system in response to a user query. The method focused on extracting and integrating research concepts and their research relationships. A simple presentation interface was also developed to give the user an overview of the search results as well as allow the user to zoom in for more detailed information. A taxonomy was used to filter out non-concept terms, highlight important concepts in the domain, and organize concepts into different subjects.

A sample of 40 users found an aggregated search presentation to be more comprehensible, readable and useful than a presentation based on listing the research objective sentences and a sentence-based summary generated by the MEAD system. The majority of users (64%) indicated a preference to use the aggregate system.

In future, other methods for displaying and presenting the concepts and relations extracted from the retrieved abstracts can be investigated. An in-depth qualitative user study needs to be carried out to find out what kinds of functions for manipulating the content of the retrieved abstracts/documents would be useful to the users. Functions for identifying gaps in the literature and for combining research results to hypothesize new relations between concepts are expected to be very useful to researchers.

Though the aggregation method focuses on quantitative research presented in the standard report structure, it can handle qualitative research abstracts to some extent. Many qualitative research studies also seek to identify relationships between concepts. Others seek to describe one or more aspects of the main concept of interest. Different from the concepts operationalized as variables in quantitative research, the concepts in qualitative research usually represent events, phenomena, behaviors, attributes, and situations. These important concepts and their relationships can be identified in the same way as for quantitative research abstracts. However, there are two main challenges. Firstly, many related concepts and mediating and moderating factors are mentioned together with the main concepts. These related concepts interfere with the identification of the important concepts. Secondly, some qualitative research abstracts do not have a clear structure. Although the research objectives section is still discernable in some unstructured abstracts and contains the most important concepts, many important concepts are scattered throughout the whole abstract. Other methods to identify the important information in unstructured abstracts need to be investigated. For example, using sentence extraction to identify the important sentences and then extracting concepts from them, or identifying important concepts based on their term frequency.

Although the aggregation method was developed to handle dissertation abstracts in the field of sociology, it can be applied to dissertation abstracts in other domains, such as psychology, education, medicine, crop agriculture and chemistry, which adopt the same research paradigm of seeking to investigate concepts and their relationships and use a similar research report structure. In the aggregation process, three processing steps—macro-level discourse parsing, information extraction and information integration—are mostly based on the research report structure and linguistic features of dissertation abstracts and are, to a large extent, not domain-specific. These three steps can be applied to dissertation abstracts in other domains with little modification. One exception is the research method terms because different research methods are used in different domains.

In the information presentation step, a taxonomy was constructed for the sociology domain and cannot be applied to other domains. For a different domain, a new taxonomy should be constructed for categorizing concepts by subject and for indicating the important concepts in this domain.

The aggregation method can also be extended to handle full research papers. Research papers are much longer than abstracts. They have the same general sections as abstracts (e.g. introduction, research methods, research findings, etc.) but have more detailed structure in each section. Thus, the aggregation method needs to be improved to handle more detailed and deeper discourse structure. The macro-level discourse structure of full research papers needs to be analyzed to identify which parts contain more important research information. Secondly, the language used in the full papers is probably more complex. Thus more indicator phrases for contextual relations and research methods need to be identified, as well as more relationship patterns.

The aggregation method developed in this study focused more on extraction and integration of semantic content and semantic relations expressed in the text. This idea can also be applied to other corpora. For example, in news stories, the semantic content should be news events, and the information can be organized according to events and different aspects of events (e.g. background, central occurrences, consequences, commentary and follow-up); in medical articles, the semantic content should be diseases, and the information can be organized based on diseases and different aspects of a disease (e.g. pathology, symptom, therapy and medicine).

8. REFERENCES


