A Multi-document Summarization System for Sociology Dissertation Abstracts: Design, Implementation and Evaluation

Shiyan Ou, Christopher S.G. Khoo, and Dion H. Goh

Division of Information Studies,
School of Communication & Information,
Nanyang Technological University,
Singapore, 637718
{pg00096125, assgkhoo, ashlgoh}@ntu.edu.sg

Abstract. The design, implementation and evaluation of a multi-document summarization system for sociology dissertation abstracts are described. The system focuses on extracting variables and their relationships from different documents, integrating the extracted information, and presenting the integrated information using a variable-based framework. Two important summarization steps – information extraction and information integration were evaluated by comparing system-generated output against human-generated output. Results indicate that the system-generated output achieves good precision and recall while extracting important concepts from each document, as well as good clusters of similar concepts from the set of documents.

1 Introduction

Multi-document summarization has begun to attract much attention in the last few years [6]. 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 allows users to zoom in for more details on aspects of interest.

Our present work aims to develop a method for automatic summarization of sets of sociology dissertation abstracts that may be retrieved by a digital library system or search engine in response to a user query. Recently, many digital libraries have begun to provide online dissertation abstract services, since they contain a wealth of high-quality information by specifying research objectives, research methods and results of dissertation projects. However, a dissertation abstract is relatively long about 300~400 words and browsing too many of such abstracts results in information overload. Therefore, it would be helpful to summarize a set of dissertation abstracts to assist users in grasping the main ideas on a specific topic.

The main approaches used for multi-document summarization include sentence extraction, template-based 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

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 450 – 461, 2005.

are selected from each cluster [1,7,12]. Some multi-document summarizers, such as SUMMONS [5], RIPTIDES [14] and GITEXTER [2], 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 to generate the summaries [4]. In addition, cross-document rhetorical relationships are used to create multi-document summaries [11, 15]. However, these existing summarization approaches focus more on physical granularities (words, phrases, sentences and paragraphs) and rhetorical relations based on 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.

In out work, we do not use the traditional summarization approaches. Instead, our work focuses on semantic-level research variables and their relationships. A variable is a specific concept explored in a particular research whose value changes from case to case. For example, gender can be considered a variable because it can take two values "*male*" and "*female*". A relationship refers to the correspondence between two variables [13]. In a set of related sociology dissertation abstracts, similar concepts across documents are usually investigated in various projects in different contexts or from different perspectives. This means that the similarities and differences across dissertation abstracts are mainly transferred through variables and their relationships. Therefore, a variable-based framework is developed to integrate variables and their relationships extracted from different abstracts and thus summarize a set of dissertation abstracts on a specific topic [8]. The framework has a hierarchical structure in which the summarized information is at the top level and the more detailed information is found at lower levels. It integrates four kinds of information:

- *Main variables:* The main variables are usually common concepts investigated by most dissertation abstracts in a document set.
- *Relationships between variables:* For each main variable, the descriptive values or the relationships with other variables are investigated in different dissertations.
- *Contextual relations:* Some studies explore variables and relationships through the perception of a group of people or in the context of a framework or model.
- *Research methods:* To explore the attributes of a variable or relationships between a pair of variables, one or more research methods are used.

The framework not only provides an overview of the subject area but also allows users to explore details according to their interest. Based on the framework, an automatic summarization method for sociology dissertation abstracts is developed. The method extracts variables and their relationships from different documents, integrates the extracted information across documents, and presents the integrated information using the variable-based framework. Although the summarization method was developed based on sociology dissertation abstracts, it is also applicable to other domains, such as psychology and medicine, which adopt the same research paradigm of seeking to investigate concepts and variables and their relationships and use a similar research report structure.

2 Overview of the Multi-document Summarization System

The summarization system follows a pipeline architecture with five modules as shown in Figure 1. Each module accomplishes one summarization process. In data preprocessing, the input dissertation abstracts in HTML format are parsed into sentences and further into a sequence of word tokens with part-of-speech tags. In macro-level discourse parsing, each sentence is categorized into one of five predefined sections using a decision tree classifier. In information extraction, important concepts are extracted from specific sections and their relationships are extracted using pattern matching. Further, research methods and contextual relations are identified using a list of identified indicator phrases. In information integration, a hierarchical concept clustering method is used to group similar concepts extracted across documents hierarchically and summarize them using a broader concept. The same types of relationships linking similar concepts are integrated and normalized using uniform expressions. Finally, the integrated information is formatted based on the variable-based framework to generate the final summaries. The system is implemented using Java and the final summaries are presented in a Web-based interface.

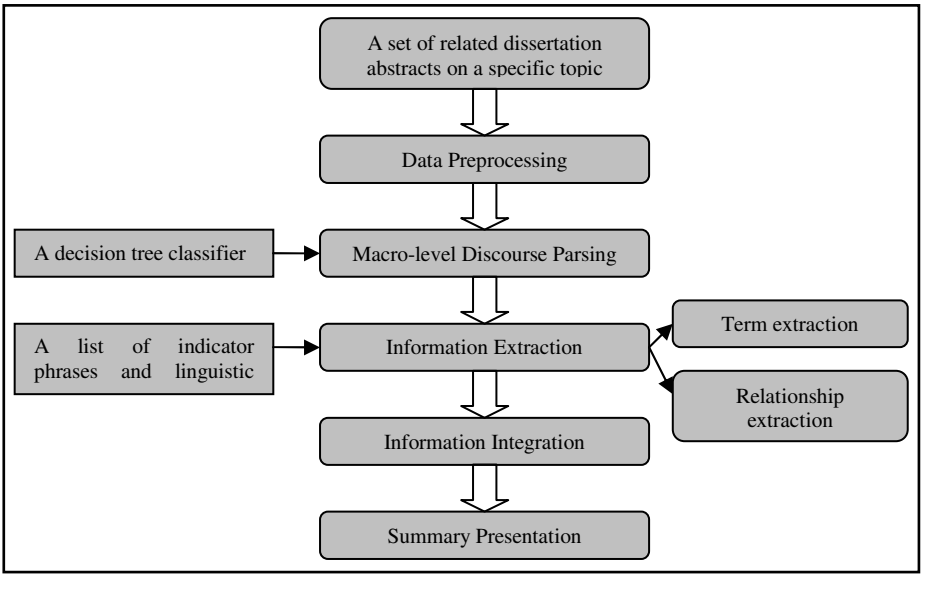


Fig. 1. Summarization system architecture

2.1 Data Preprocessing

The input files are a set of related dissertation abstracts on a specific topic retrieved from the Dissertation Abstracts International database indexed under *Sociology* and *PhD degree*. Each file contains one dissertation abstract in HTML format. First, each file was transformed into a uniform XML representation. Next, the text was segmented into sentences using a short list of end-of-sentence punctuation marks,

such as periods, question marks and exclamation points. The exclamation point and question mark are less ambiguous as end-of-sentence indicators. However, since a period is not used exclusively to indicate sentence breaks (e.g., it can be used to indicate an abbreviation, a decimal point and parts of an e-mail address), a list of common abbreviations (e.g., "i.e.", "u.s.") and an algorithm for detecting decimal, e-mail address and ellipsis was used to ensure more reliable identification of sentence boundaries. Finally, each sentence was parsed into a sequence of word tokens using the Conexor Parser [10]. For each word token, its document_id, sentence_id, token_id (word position in the sentence), word form (the real form used in the text), lemma (base word), and part-of-speech tag were indicated.

2.2 Macro-level Discourse Parsing

In previous work, an automatic method for parsing the macro-level discourse structure of sociology dissertation abstracts was developed by using decision tree induction to categorize each sentence into one of the five predefined sections or categories – *background*, *research objectives*, *research methods*, *research results* and *concluding remarks* [9]. The decision tree classifier made use of the normalized sentence position and single indicator words to identify the categories and obtained an accuracy rate of about 72% when applied to structured dissertation abstracts.

It was observed however that some sentences in the *research objectives* and *results methods* sections contained clear indicator phrases at the beginning of the sentences. For example, "*The purpose of this study was to investigate …*" and "*The present study aims to explore …*" often appeared at the beginning of sentences in the *research objectives* section, whereas "*The results indicated...*" and "*The study suggested…*" often appeared in sentences in the *research results* section. These indicator phrases can identify *research objectives* and *research results* sentences more precisely than the single indicator words used by the decision tree classifier.

Therefore, the sentence categories assigned by the decision tree classifier were adjusted further using the indicator phrases to improve the accuracy of identifying *research objectives* and *research results* sentences.

2.3 Information Extraction

Four kinds of information were extracted from the dissertation abstracts – *research variables*, *relationships between variables*, *research methods* and *contextual relations*, using indicator phrases or linguistic patterns. Research variables, research methods and contextual relations are concepts which can be extracted using term extraction. The research methods and contextual relations were identified from the extracted terms using a list of indicator phrases, whereas the variables were identified by focusing on the *research objectives* and *research results* sentences To extract the relationships between variables expressed in the text, pattern matching was performed to identify the segments of the sentence that match with each pattern.

2.3.1 Term Extraction

After data preprocessing, a sequence of word tokens was obtained for each sentence in each document. Sequences of contiguous words of different lengths (i.e. 2, 3, 4, 5 words) were extracted from each sentence to construct n-grams (where n is the

number of words). Surface syntactic patterns were used to differentiate between terms and non-terms among the n-grams. A list of surface syntactic patterns was constructed for recognizing 1, 2, 3, 4 and 5-words terms (see Table 1 for examples).

Table 1. Some of surface syntactic patterns used for identifying terms

ID	1	2	3	4	5	Example Term
1	N					teacher
2	A	N				young child
3	N	PREP	N			ability of organization
4	N	PREP	A	N		effectiveness of early childhood
5	N	PREP	A	N	N	effectiveness of early childhood teacher

Using the surface syntactic patterns, terms of different numbers of words were extracted from the same position in the sentences. These terms represent concepts at different levels (narrow and broad concepts). For example, in the sentence "*The present study assessed the effectiveness of preschool teachers of India with respect to their interactions with young children and parents*", one set of extracted terms were:

- 1-word terms: effectiveness, preschool, teacher, India, child, parent
- 2-word terms: preschool teacher, young child
- 3-word terms: effectiveness of preschool, teachers of India, child and parent
- 4-word terms: effectiveness of preschool teacher, preschool teachers of India, young child and parent
- 5-word terms: -

To identify the full terms in the sentences, the terms of different lengths extracted from the same position are compared and terms which cannot be covered by other terms are retained, e.g. "*effectiveness of preschool teacher" and "preschool teachers of India"*. Then, the terms which have overlapping tokens are connected to form a full term representing a specific full concept in the text, e.g. "*effectiveness of preschool teacher of India*".

The concepts relating to research methods and contextual relations were identified from the full terms using a list of indicator phrases derived manually from 300 sample documents. Some of the indicator words and phrases for research methods and contextual relations are given in Table 2.

To identify the variable concepts, the full terms extracted from the *research objectives* and *research results* sentences were selected, since these two sections focus more on variables and their relationships.

Table 2. Some of indicator words and phrases for research methods and contextual relations

Types	Subtypes	Indicator words and phrases
Research methods	Research design	*interview, field work, survey, qualitative study*
	Sampling	*convenience sampling, snowball sampling*
	Data analysis	*univariate analysis, time series analysis*
Contextual relations		*perception, attitude, insight, perspective ,view, thought, model, hypothesis, assumption, context*

2.3.2 Relationship Extraction

Extraction of relationships between variables involves looking for certain linguistic patterns that indicate the presence of a particular relationship. In this study, we used linear regular expression patterns. A linear pattern is a sequence of tokens each representing one or more of the following:

- A literal word in the text which has been converted to a base form (i.e. lemma);
- A wildcard which can match with one or more words in the sentence;
- A part-of-speech tag, e.g. N, V, A, ADV;
- A slot to be filled in by words or phrases in the text.

Some tokens are compulsory whereas others are optional. In this way, the patterns are more general and flexible enough to match more relationship expressions used in the text. The following is an example of a linear pattern that describes one way that cause-effect relationship can be expressed in the text:

- *[slot:IV] have <DET> (<A>) effect/influence/impact on [slot:DV]*

The tokens within square brackets represent slots to be filled by words or phrases in the text. The slots indicate which part of a sentence represents the independent variable (IV) and which part represents the dependent variable (DV) in a cause-effect relationship. The tokens within round brackets represent optional words. For example, the above pattern can match with the following sentence:

- *Changes in labor productivity have a positive effect on directional movement.*

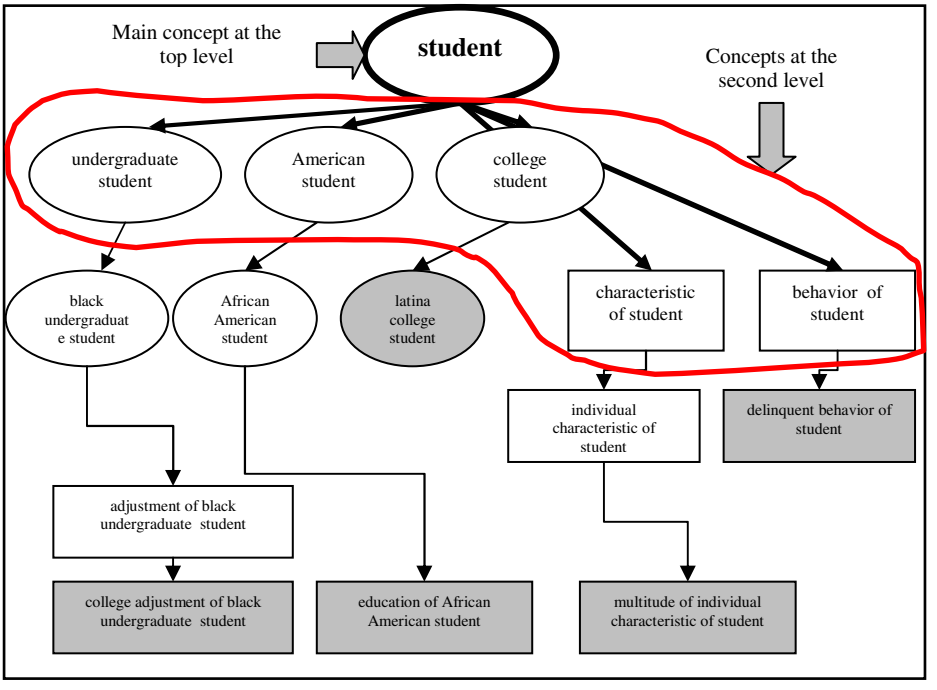
We derived 126 relationship patterns as a result of analyzing 300 sample dissertation abstracts. These patterns belong to five specific types of relationships which are often investigated in sociological research:

1. *Cause-effect relationships:* one variable causes a change or effect in another variable;
2. *Correlations:* a change in one variable is accompanied by a change in another;
3. *Comparative relationships:* There are differences between two or more variables;
4. *Predictive relationships:* One variable predicts another one;
5. *Second-order relationships:* The relationship between two or more variables is influenced by a third variable.

A pattern matching program was developed to identify the segments of the text that match with each pattern. If a text segment matches with a pattern, then the text segment is identified to contain the relationship associated with the pattern. A pattern typically contains one or more slots, and the parts of the text that match with the slots in the pattern represent the variables related by the relationship.

2.4 Information Integration

A hierarchical concept clustering method was used to group similar concepts into a tree or a hierarchy. However, concept integration is more than a simple clustering. It involves summarizing a group of similar concepts with a broader concept. Therefore, traditional clustering methods such as hierarchical agglomerative clustering were not used. Instead, we used a clustering method that links similar concepts into a hierarchical structure. It includes three phases:



- *The highlighted concepts are full concepts occurring in the text.*

Fig. 2. A cluster of similar concepts linked in a hierarchical structure

1. *Segment the full terms:* Each full term occurring in the text was segmented into 1, 2, 3, 4, 5-word terms and only high frequency 1-word terms above a specific threshold value were retained. Stop words and all indicator words were removed.
2. *Construct term chains:* For each frequent 1-word term, a list of concept chains was constructed by linking it with other multi-word terms in which the single word occurred as a head noun. Each chain was constructed top-down by linking the short terms first followed by the longer terms containing the short term. The root node of each chain is the 1-word term (main concept), and the leaf node is the full term (full concept). The length of the chains can be different but the maximum length is six nodes – the 1, 2, 3, 4, 5-word terms and the full terms.
3. *Build cluster tree:* All chains sharing the same root node were combined to form a cluster tree. Each cluster tree uses the root node (1-word term) as its cluster label. At the root node, two types of sublevel concepts exist – subclass concepts and facet concepts. Subclass concepts are subclasses of the main concept which are restricted or narrowed down by one or more qualifiers, while facet concepts specify various aspects (characteristics or facets) of the main concept.

In this way, similar concepts at different levels are clustered automatically. In Figure 2, the concepts in round boxes represent subclass concepts, whereas the concepts in rectangular boxes represent facet concepts. The full concepts (more specific concepts occurring in the text) are at the bottom of the cluster. In the

hierarchical cluster, concepts at the lower levels can be summarized by broader concepts at the higher levels.

The relationships between variables can be grouped according to the variable concepts they refer to. Next, the same types of relationships are converted to a uniform representation. The concepts relating to research methods and contextual relations are grouped and summarized using the indicator phrases used for identifying them. Synonyms are replaced by uniform words or phrases.

2.5 Summary Presentation

We generated interactive summaries to allow users to explore details of interest by clicking on hyperlinks rather than viewing traditional plain text summaries. Hence, the final summaries are presented in an HTML format viewable on a Web browser. It includes three hierarchies – the main summary, lists of summarized single documents sharing the same concepts, and the original dissertation abstracts. The main summary is displayed in one main window while the other two hierarchies are displayed separately in pop-up windows. In the main window, the grouped and summarized research methods, contextual relations, research variables and their relationships extracted from different documents, are integrated based on the variable-based framework (see Figure 3). For each concept, the number of documents is given in parenthesis. This is clickable and links to a list of summarized single documents sharing the given concept are displayed in a pop-up window. For each document, the title, research variables (full concepts), research methods and contextual relations are displayed. The title of the document is also clickable and links to the original dissertation abstract in a separate pop-up window.

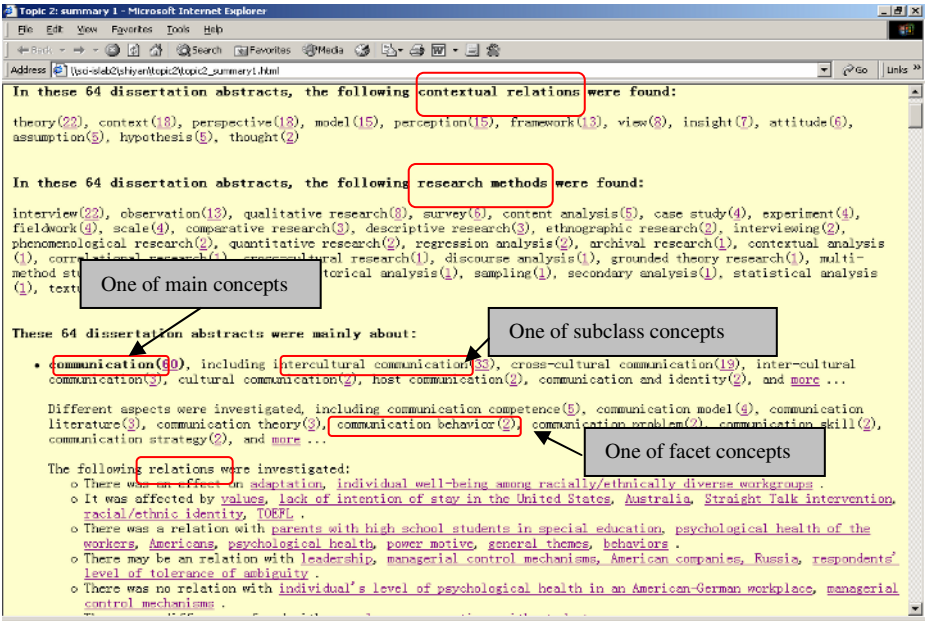


Fig. 3. The main summary on the topic of "intercultural communication"

3 Evaluation

The purpose of the evaluation was to assess the accuracy and effectiveness of two important summarization steps – information extraction and information integration, since they influence the final output of the summarization system. In this study, we assumed that the human-generated output was the "gold standard" to be measured against. The human coders were social science graduate students at the Nanyang Technological University, Singapore.

For the evaluation of information extraction, 50 PhD sociology dissertation abstracts were selected systematically. Three human coders were asked to extract all the *important concepts* from each abstract, and from among these to identify the *more important concepts* and then the *most important concepts*, according to the focus of the dissertation research. The human-extracted concepts were used as the "*ideal*" concepts to compare against the concepts extracted automatically in three situations – concepts from *research objectives* (section 2) only, from both *research objectives* and *research results* (section 2 & 4), and from all sections.

In the evaluation of information integration, 15 topics contributed by sociology researchers were used. For each topic, a set of PhD sociology dissertation abstracts were retrieved from the Dissertation Abstracts International database by using the topic as the query and five abstracts were selected to form a document set. Moreover, another five abstracts were selected for each of five topics, including the previously chosen five abstracts, to construct a second bigger document set. From each abstract, the important concepts were extracted from the *research objectives* and *research results* automatically using our system. The human coders were asked to identify similar concepts, cluster and summarize them by assigning a category label to each cluster. Each document set was coded by two human coders. The concept clusters generated by human coders were used as the "*ideal"* clusters to compare against those generated automatically by our system.

3.1 Evaluation Results for Information Extraction

Three human coders had extracted concepts at three levels of importance. Table 3 shows the average precision, recall and F-values among the three coders. Note that the

Table 3. Average precision, recall and F-value among the three coders

Importance level		All sections	Section 2	Section 2 & 4
The *most important* concepts	Precision (%)	20.36	31.62	23.60
	Recall (%)	92.26	76.06	87.37
	F-value (%)	33.15	**43.91**	36.80
The *more important* concepts	Precision (%)	31.02	44.51	34.28
	Recall (%)	90.93	59.31	78.81
	F-value (%)	45.94	**50.27**	47.35
The *important* concepts	Precision (%)	46.12	58.84	49.68
	Recall (%)	89.99	46.63	75.66
	F-value (%)	**60.40**	51.53	**59.34**

** Section 2 refers to research objectives, whereas section 4 refers to research results.*

set of *important* concepts include the set of *more important* and *most important* concepts. Similarly, the set of *more important* concepts include the set of *most important* concepts.

For all *important concepts*, the F-value obtained for *all sections* (60%) and for *section 2 & 4* (59%) were similar, and both were much better than the F-value obtained for *section 2* (52%). This suggests that important concepts are not focused only in *section 2*, but scattered in all sections. Therefore, our macro-level discourse parsing for identifying different sections of the dissertation abstract may not be helpful for identifying the important concepts.

For the *more important concepts*, the F-value obtained for *section 2* (50%) was a little higher than those for *section 2 & 4* (47%) and for *all sections* (46%). This suggests that *section 2* places a bit more emphasis on the more important concepts.

However, for the *most important concepts*, the F-value (44%) obtained for *section 2* was much higher than for *section 2 & 4* (37%) and for *all sections* (33%). This suggests that *section 2* focuses on the most important concepts, and *section 2 & 4* also can contribute to identifying the most important concepts to a less extent.

In conclusion, our macro-level discourse parsing should be helpful in identifying the more important and most important concepts. Concepts are the main elements of our system-generated summaries and accurate information extraction can result in more concise summaries which focus on the more important or most important concepts.

3.2 Evaluation Results for Information Integration

For each document set, we calculated the inter-coder similarity among the two sets of clusters created by the two human coders using a similarity measure employed by Macskassy et al. [3] as follows:

Overall similarity between coding 1 and coding 2

$$= \frac{\text{number of common same-cluster-pairs between coding 1 and coding 2}}{\text{total number of unique same-cluster-pairs obtained from coding 1 and coding 2}}$$

Similarity calculation involves first identifying all the possible pairs of terms within the same cluster (same-cluster-pairs). If the two human coders created the same clusters, then the pairs of terms obtained for both codings will be the same, and the similarity value obtained will be 1. The average inter-coder similarity obtained for the 20 document sets was a low 0.19. The value ranged from 0.04 to 0.43 across the document sets. This means that clustering is a very subjective operation.

We calculated the similarity between the sets of clusters created by our system and each of the two human coders, and obtained a higher average similarity of 0.26. This indicates that the system's clustering is more similar to each of the human coders than between the two human coders! The accurate clustering can result in a clear identification of similarities across documents in the final summaries.

4 Conclusion

This paper has outlined the design, implementation and evaluation of a multi-document summarization method for sociology dissertation abstracts. Our system focuses on extracting variables and their relationships, integrating the extracted information, and presenting the integrated information using a variable-based framework with an interactive Web interface. The summarization method employs term extraction, pattern matching, discourse parsing, and a kind of concept clustering.

Two important summarization steps – information extraction and information integration were evaluated by comparing our system's generated output against human-generated output. The results indicate that the macro-level discourse parsing was helpful in identifying the more important concepts, and the accuracy of the automatic information extraction was acceptable (46% precision and 90% recall). Information integration using the hierarchical concept clustering method generated reasonably good clusters compared to human clustering. User evaluation of the summarization method is in progress.

References

1. Boros, E., Kanto, P.B., & Neu, D.J.: A clustering based approach to creating multi-document summaries. *Document Understanding Conferences* (2002). Available at http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/rutgers_final.pdf
2. Harabagiu, S.M., & Lacatusu, F.: Generating single and multi-document summaries with GISTEXTER. *Document Understanding Conferences* (2002). Available at http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/utdallas_sanda.pdf
3. Macskassy, S.A., Banerjee, A., Davison, B.D., & Hirsh, H.: Human performance on clustering Web pages: A preliminary study. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Presss (1998) 264-268
4. Mani, I., & Bloedorn, E.: Summarization similarities and differences among related documents. Information Retrieval, 1(1) (1999) 1-23
5. Mckeown, K., & Radev, D.: Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (ACM SIGIR)*. Seattle, WA (1995) 74-82
6. National Institute of Standards and Technology. : *Document Understanding Conferences* (2002). Available at http://www-nlpir.nist.gov/projects/duc/index.html
7. Otterbacher, J.C., Winkel, A.J., & Radev, D.R.: The Michigan single and multi-document summarizer for DUC 2002. *Document Understanding Conferences* (2002) Available at http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/umich_otter.pdf
8. Ou, S., Khoo, C., & Goh, D.: 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 (ASIST)*. Long Beach, CA, 19-23 October (2003) 230-239
9. Ou, S., Khoo, C., Goh, D., & Heng, Hui-Hing. : Automatic discourse parsing of sociology dissertation abstracts as sentence categorization. In *Proceedings of the 8th International ISKO Conference*. London, UK, 13-16 July (2004) 345-350

10. Pasi, J. & Timo, J.: A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington, DD: Association for Computational Linguistics(1997) 64-71
11. Radev, D.: A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue* (2000). Available at http://www.sigdial.org/sigdialworkshop/proceedings/radev.pdf
12. Radev, D., Jing, H., & Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Workshop held with Applied Natural Language Processing Conference / Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/ANNCL)* (2000) 21-29
13. Trochim, W.: *The research methods knowledge base*. Cincinnati, OH: Atomic Dog Publishing (1999)
14. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., & Wagstaff, K.: Multi-document summarization via information extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT-01)* (2001)
15. Zhang, Z., Blair-Goldensohn, S., & Radev, D.: Towards CST-enhanced summarization. *In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002).* Edmonton , Canada, August (2002)