Journal of Information Science OnlineFirst, published on December 3, 2007 as doi:10.1177/0165551507084630

Design and development of a concept-based multi-document summarization system for research abstracts

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

Division of Information Studies, School of Communication and Information, Nanyang Technological University, Singapore

Abstract.

This paper describes a new concept-based multi-document summarization system that employs discourse parsing, information extraction and information integration. Dissertation abstracts in the field of sociology were selected as sample documents for this study. The summarization process includes four major steps – (1) parsing dissertation abstracts into five standard sections; (2) extracting research concepts (often operationalized as research variables) and their relationships, the research methods used and the contextual relations from specific sections of the text; (3) integrating similar concepts and relationships across different abstracts; and (4) combining and organizing the different kinds of information using a variable-based framework, and presenting them in an interactive web-based interface. The accuracy of each summarization step was evaluated by comparing the system-generated output against human coding. The user evaluation carried out in the study indicated that the majority of subjects (70%) preferred the concept-based summaries generated using the system to the sentence-based summaries generated using traditional sentence extraction techniques.

Keywords: discourse parsing; information extraction; information integration; multi-document summarization

1. Introduction

Multi-document summarization is regarded as the process of condensing, not just one document, but a set of related documents, into a single summary. This study aimed to develop an automatic method for summarizing sets of research abstracts that may be retrieved by an information retrieval system or web search engine in response to a user query. As an attempt to address the problem of information overload, most information retrieval systems and web search engines rank retrieved records by their likelihood of relevance and display titles and short abstracts to give

Correspondence to: Shiyan Ou, Resarch Group in Computational Linguistics, University of Wolverhampton, WV1 1SB, UK. Email: Shiyan.Ou@wlv.ac.uk

users some indication of the document content. Since the related documents often contain repeated information or share the same background, these single-document summaries (or abstracts) are likely to be similar to each other and thus cannot indicate unique information in individual documents [1]. Moreover, the user has patience to scan only a small number of document titles and abstracts, usually in the range of 10–30 [2]. In such a situation, multi-document summarization for condensing a set of related documents into a summary is likely to be more useful than single-document summarization. A multi-document summary has several potential advantages over a single-document summary. It provides a domain overview of a topic based on a document set – indicating similar information in many documents, unique information in individual documents, and relationships between pieces of information across different documents. It can allow the user to zoom in for more details on particular aspects of interest, and zoom into the individual single-document summaries.

In this study, we selected dissertation abstracts in the sociology domain as source documents. Dissertation abstracts are high quality informative abstracts providing substantial information on the research objectives, research methods and results of dissertation projects. Since most dissertation abstracts have a relatively clear structure and the language is more formal and standardized than in other corpora (e.g. news articles), it is a good corpus for initial development of the techniques for processing research abstracts, before extending them to handle journal article abstracts and other kinds of abstracts. Dissertation abstracts can be viewed as documents in their own right, being relatively long at 300–400 words, or they can be viewed as an intermediate state in a two-stage summarization process – first summarizing documents into single-document abstracts and then combining the single-document abstracts into one multi-document abstract.

The sociology domain was selected for this study partly because many sociological studies adopt the traditional quantitative research paradigm of identifying relationships between concepts operationalized as variables. We take advantage of this research paradigm to provide a variable-based framework for summarizing the research abstracts focusing on research concepts and relationships [35].

Multi-document summarization presents more challenges than single-document summarization in the issues of compression rate, redundancy, cohesion, coherence, temporal dimension, and so on [1]. Traditional single-document summarization approaches do not always work well in a multi-document environment. In a document set, many of the documents are likely to contain similar information and only differ in certain parts. Thus, an ideal multi-document summary should contain similar information repeated in many documents, plus important unique information found in some individual documents [1]. Since much sociological research aims to explore research concepts and relationships [3], multi-document summarization of sociology research abstracts should identify similarities and differences across different studies, focusing on research concepts and the relationships investigated between them.

The summarization method developed in this study is a hybrid method comprising four major steps:

1. *Macro-level discourse parsing:* an automatic discourse parsing method was developed to segment a dissertation abstract into several macro-level sections and identify which sections contain important research information.
2. *Information extraction:* an information extraction method was developed to extract research concepts and relationships as well as other kinds of information from the micro-level structure (within sentences).
3. *Information integration:* an information integration method was developed to integrate similar concepts and relationships extracted from different abstracts.
4. *Summary presentation:* a presentation method was developed to combine and organize the different kinds of information using the variable-based framework, and present them in an interactive web-based interface.

In each step, the accuracy of the system was evaluated by comparing the system-generated output against human coding.

2. Literature review

Summarization approaches can be divided broadly into extractive and abstractive approaches. A commonly used extractive approach is statistics-based sentence extraction. Statistical and linguistic features used in sentence extraction include frequent keywords, title keywords, cue phrases, sentence position, sentence length, and so on [4–6]. Sometimes, cohesive links such as lexical chain, co-reference and word co-occurrence are also used to extract internally linked sentences and thus increase the cohesion and fluency of the summaries [7, 8]. Although extractive approaches are easy to implement, the resulting summaries often contain redundancy and lack cohesion and coherence. These weaknesses become more serious in multi-document summarization because the extracted sentences are from different sources, have different writing styles, often contain repeated information, and lack context. To reduce redundancy in multi-document summaries, some summarization systems, such as MEAD [9], XDoX [10] and MultiGen [11], clustered documents (or sentences) and extracted representative sentences from each cluster as components of the summary. In addition, the Maximal Marginal Relevance (MMR) metric was used by Carbonell and Goldstein [12] to minimize the redundancy and maximize the diversity among the extracted text passages (i.e. phrases, sentences, segments, or paragraphs).

In comparison to extractive approaches, abstractive approaches involve text abstraction and generation to produce more coherent and concise summaries. Thus abstractive approaches seem more appropriate for multi-document summarization [13]. Real abstractive approaches that completely imitate human abstracting behavior are difficult to achieve with current natural language processing techniques [1]. Current abstractive approaches are in reality hybrid approaches involving both extractive and abstractive techniques. Abstractive approaches for multi-document summarization focus mainly on similarities and differences across documents, which can be identified and synthesized using various methods. The MultiGen summarizer identified similar words or phrases across documents through syntactic comparisons and converted them into fluent sentences using natural language generation techniques [11]. Lin [14] identified similar concepts based on a lexical thesaurus WordNet and generalized these concepts using a broader unifying concept. Mckeown and Radev [15] extracted salient information using template-based information extraction and combined the instantiated slots in different templates using various content planning operators (e.g. *agreement* and *contradiction*). Zhang et al. [16] added the sentences that have specific cross-document rhetorical relationships (e.g. *equivalence* and *contradiction*) into a baseline summary generated using a sentence extraction method to improve the quality of the summary. Afantenos et al. [17] created a set of topic-specific templates using an information extraction system and connected these templates according to synchronic rhetorical relations (e.g. *identity*, *elaboration*, *contradiction*, *equivalence*) and diachronic rhetorical relations (e.g. *continuation*, *stability*).

However, most of these studies identified similarities and differences using low-level text analysis, i.e. mainly based on lexical, syntactic and rhetorical relations between text units (e.g. words, phrases, and sentences). It is desirable to identify similarities and differences at a more semantic and contextual level. Thus, this study identified similarities and differences focusing on research concepts and relationships. In sociological studies, the research concepts often represent elements of society and human behavior whereas the relationships are semantic relations between research concepts investigated by researchers. This study adopts a combination of abstractive and extractive approaches – identifying more important sections using discourse parsing, extracting research concepts and relationships using information extraction techniques, integrating concepts and relationships using syntactic analysis, combining the four kinds of information using the variable-based framework, and organizing the integrated concepts using a taxonomy to generate a multi-document summary.

3. Multi-document summarization system

The summarization system has a blackboard architecture with five modules (shown in Figure 1). Each module accomplishes one summarization step. A knowledge base was used as a central repository for

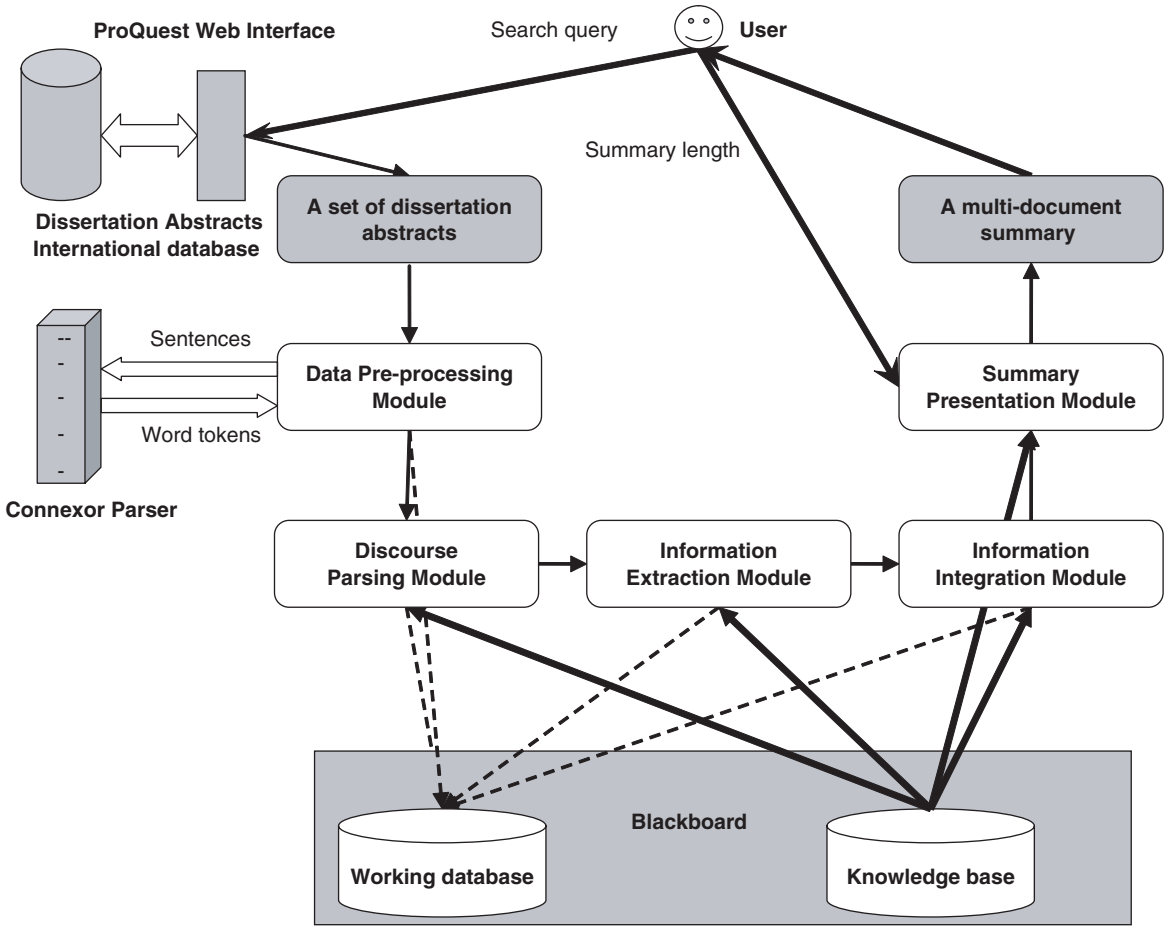

Fig. 1. Diagram of the summarization system architecture.

all shared knowledge needed to support the summarization process. A working database was used to store the output of each module, which becomes the input to the subsequent modules. The system was implemented on the Microsoft Windows platform using the Java 2 programming language and Microsoft Access database. But the system can be migrated easily to a UNIX platform.

3.1. Data pre-processing

The input data are a set of dissertation records on a specific topic retrieved from the Dissertation Abstracts International database indexed under *sociology* subject and *PhD* degree. Each dissertation record is transformed from HTML format into XML format. The abstract text is divided into separate sentences using a simple sentence breaking algorithm. Each sentence is parsed into a sequence of word tokens using the Conexor Parser [18]. 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), base form (lemma) and part-of-speech tag are indicated.

3.2. Macro-level discourse parsing

Most dissertation abstracts (about 85%) have a clear structure containing five standard sections – *background*, *research objectives*, *research methods*, *research results* and *concluding remarks*. Each section

contains one or more sentences. In this study, we treated discourse parsing as a sentence categorization problem, i.e. assigning each sentence in a dissertation abstract to one of the five categories or sections. In previous studies, surface cues have been used for discourse parsing, for example, cue words, synonymous words or phrases, and similarity between two sentences used by Kurohashi and Nagao [19]; lexical frequency and distribution information used by Hearst [20]; and syntactic information, cue phrases and other cohesive devices used by Le and Abeysinghe [21]. However, only some sentences in dissertation abstracts were found to contain a clear cue phrase at the beginning. Thus, we selected a supervised learning method, decision tree induction, which has been used by several researchers, such as Marcu [22] and Nomoto and Matsumoto [23], for discourse parsing. Finally, cue phrases found at the beginning of some sentences were used as a complement to improve their categorization.

To develop a decision tree classifier, a random sample of 300 dissertation abstracts was selected from the set of 3214 PhD dissertation abstracts in sociology, published in the 2001 Dissertation Abstracts International database. The sample abstracts were partitioned into a training set of 200 abstracts to construct the classifier and a test set of 100 abstracts to evaluate the accuracy of the constructed classifier. Each sentence in the sample abstracts was manually assigned to one of the five categories. To simplify the classification problem, each sentence was assigned to only one category, though some sentences could arguably be assigned to multiple categories or no category at all. Some of the abstracts (29 in the training set and 16 in the test set) were found to be unstructured and difficult to code into the five categories and thus removed from the training and test set. A well-known decision tree induction algorithm, C5.0 [24], was used in the study. The decision tree classifier that was developed used high frequency word tokens and normalized sentence position in the abstract as features.

Preliminary experiments were carried out using 10-fold cross-validation to determine the appropriate parameters for constructing the classifier, including the word frequency threshold value for determining the cue words used for categorization and the pruning severity for determining the extent to which the constructed classifier will be pruned. The best classifier was obtained with a word frequency threshold value of 35 and pruning severity of 95%. Finally, the classifier was applied to the test sample and an accuracy rate of 71.6% was obtained.

A set of IF–THEN categorization rules was extracted from the decision tree classifier. An example rule for identifying the *research objectives* section (i.e. Section 2) is as follows:

- **If** N-SENTENCE-POSITION<=0.444444 and STUDY=1 and PARTICIPANT=0 and DATA=0 and CONDUCT=0 and PARTICIPATE=0 and FORM=0 and ANALYSIS=0 and SHOW=0 and COMPLETE=0 and SCALE=0, **then** SECTION=2

In the above rule, '1' indicates that the word appears in the sentence whereas '0' indicates that the word does not appear in the sentence. Thus the rule says that if a sentence contains 'study' but does not contain 'participant', 'data', 'conduct', 'participate', 'form', 'analysis', 'show', 'complete' and 'scale', and it is located in the first half of the document, it is assigned to the research objectives section. In dissertation abstracts, distinctive cue phrases were found at the beginning of some sentences in the research objectives and research results sections. Sentences containing such cue phrases could be categorized more accurately than using the decision tree classifier which makes use of single words as features. For example, 'The purpose of this study was to investigate …' and 'The present study aimed to explore …' indicate research objective sentences, whereas 'The results indicated that …' and 'This research found that …' indicate research result sentences. Thus, the categories of some sentences assigned by the decision tree classifier are improved with a set of cue phrases manually identified from the 300 sample abstracts.

3.3. Information extraction

Four kinds of information were extracted from each dissertation abstract – *research concepts* and *relationships*, *contextual relations* and *research methods*. Relationships were extracted using pattern matching based on a set of manually constructed linguistic patterns. The other three kinds of information appear as nouns or noun phrases, which are extracted using syntactic rules.

In previous studies, rule-based and statistics-based methods have both been used for extracting multi-word terms. Borgigault and Jacquemin [25] extracted noun phrases using shallow grammatical

Table 1
Some part-of-speech patterns for recognizing single-word and multi-word terms

	Part-of-speech tag					
ID	1	2	3	4	5	Example term
1	N					teacher
2	A	N				young teacher
3	N	PREP	N			ability of organization
4	A	N	PREP	N		parental ability of reading
5	N	PREP	A	N	N	effectiveness of early childhood teacher

structure. Nakagawa [26] extracted multi-word terms using statistical associations between a multi-word term and its component single nouns. In the study, we used the rule-based method to extract multi-word terms based on syntactic analysis. Since the language used in dissertation abstracts is formal and regular, the syntactic rules for multi-word terms are easy to construct.

3.3.1. Term extraction

Concepts, expressed as single-word or multi-word terms, usually take the grammatical form of nouns or noun phrases [27]. After data pre-processing, sequences of contiguous words of different lengths are extracted from each sentence to construct *n*-grams ($n = 1, 2, 3, 4,$ and 5). A list of part-of-speech patterns was constructed for recognizing single-word and multi-word terms (see Table 1).

Using the part-of-speech patterns, terms of different numbers of words are extracted from the same part of a sentence. These terms of different lengths represent concepts at different levels of generality (narrower or broader concepts). If two terms have overlapping sentence positions, they are combined to form a full term representing a more specific full concept, e.g.

- 'effectiveness of preschool teacher' + 'preschool teacher of India' → 'effectiveness of preschool teacher of India'

The extracted terms can be research concept terms, research method terms and contextual relation terms. Research method terms and contextual relation terms are selected from the whole text. A list of cue phrases, derived manually from the 300 sample dissertation abstracts, is used to identify the research method terms and contextual relation terms, for example, 'quantitative study', 'interview', 'field work' and 'regression analysis' for research methods, and 'context', 'perception', 'insight' and 'model' for contextual relations. After removing research method and contextual relation terms from the extracted terms, research concept terms are identified as those taken from the *research objectives* and *research results* sections, since these two sections are most likely to contain important research information.

3.3.2. Relationship extraction

There are two kinds of approach for performing relation extraction. One kind of approach makes use of linguistic patterns which indicate the presence of a particular relation in the text. The second makes use of statistics of co-occurrences of two entities (e.g. pointwise mutual information, log-likelihood ratio) to determine whether their co-occurrence is due to chance or an underlying relationship (e.g. [28]). In dissertation abstracts, most of the relationships between research concepts were mentioned explicitly in the text, and thus pattern-based relation extraction was employed. Pattern-based relation extraction involves constructing linguistic patterns of relationships and identifying the text segments that match with the patterns. Patterns can be constructed manually by human experts or learnt automatically from corpora using supervised (for annotated data), semi-supervised (i.e. predefining a small set of seed patterns and bootstrapping from them) or unsupervised (for un-annotated data) methods. In this study, we did not make the effort to construct patterns automatically. Instead, we manually derived 126 relationship patterns from the sample of 300 dissertation abstracts based on the lexical and syntactic information.

The linguistic patterns used in this study are regular expression patterns, each comprising two or more slots and a sequence of tokens. The slots refer to research concepts operationalized as research

Table 2
Example pattern for extracting cause–effect relationship in text

Token	<slot: IV>	have	*	(*)	(*)	(and)	(*)	effect/influence/ impact	on/in	<slot: DV>
Part of speech tag	NP	V	DET	ADV	A	CC	A	V	PREP	NP

IV indicates an independent variable and DV indicates a dependent variable;
( ) indicates an optional cue word;
* indicates a wild card.

variables, whereas the non-slot tokens are cue words which signal the occurrence of a relationship. Each cue word is constrained with a part-of-speech tag. Table 2 gives an example pattern that represents one surface expression of cause–effect relationship in the text.

The pattern matches the following sentences, where the extracted independent variables (IVs) and dependent variables (DVs) are underlined, and their cause–effect relationships are highlighted in bold.

1. Changes in labor productivity **have a positive effect on** directional movement.
2. Medicaid appeared to **have a negative influence on** the proportion of uninsured welfare leaves.
3. Family structure **has a significant impact on** parental attachment and supervision.

A pattern matching algorithm was developed to look for these relationship patterns in the text. Pattern matching was focused on the *research objectives* and *research results* sections to extract relationships and their associated variables. A pattern typically contains one or more slots, and the research concept terms that match the slots in the pattern represent the variables linked by the relationship. Research concept terms had been extracted as nouns or noun phrases in an earlier processing step (see Section 3.3.1).

3.4. Information integration

Information integration includes concept integration and relationship integration. Concept integration involves clustering similar concepts and generalizing them using a broader concept. Relationship integration involves clustering relationships associated with the common concepts, normalizing the different surface expressions for the same type of relationship, and conflating them into a new sentence.

In previous studies, two approaches have been used for concept generalization. The first approach is based on semantic relations among concepts. Lin [14] used *computer* to generalize *mainframe*, *workstation*, *server*, *PC* and *laptop* according to the *is-instance-of* and *is-subclass-of* relations in WordNet. This approach requires a thesaurus, taxonomy, ontology, or knowledge base to provide a meaningful concept hierarchy. The second approach is based on syntactic relations among concepts. Various syntactic variations have been used by researchers to identify term variants which are considered to represent similar concepts. Borgigault and Jacquemin [25] used *internal insertion of modifiers*, *preposition switch* and *determiner insertion* to identify term variants. Ibekwe-SanJuan and SanJuan [29] defined two kinds of variations – the variations that only affected modifier words in a term such as *left expansion*, *insertion*, and *modifier substitution* and the variations that shared the same head words such as *left–right expansion*, *right expansion* and *head substitution*. In this study, we used the second approach to identify and cluster similar concepts based on two kinds of syntactic variations – *subclass modifier substitution* and *facet modifier substitution*.

3.4.1. Concept clustering and generalization

To integrate similar concepts, we analyzed the structure of multi-word terms (concepts) and found that the majority can be divided into the following two parts:

- *Head noun* refers to the noun component that identifies the broader class of things or events to which the term as a whole refers, for example, *cognitive ability*, *educated woman*.

- *Modifier* narrows the denotation of the head noun by specifying a subclass or a facet of the broader concept represented by the head noun. For example, *cognitive ability* (a type of ability), *educated woman* (a subclass of woman), *woman's behavior* (an aspect of woman).

A full term, which represents a specific *full concept* expressed in the text, can be segmented into shorter terms of different numbers of words, e.g. one-, two-, three-, four-, and five-word terms, which are called *component concepts*. There are hierarchical relations among these component concepts, distinguished by their logical roles or functions. A meaningful single noun (excluding stopwords, common words, attribute words and various cue words) can be considered the head noun and represent a broader *main concept*. Two types of sub-level concepts are distinguished – *subclass concepts* and *facet concepts*. A subclass concept represents one of the subclasses of its parent concept. A facet concept specifies one of the facets (aspects or characteristics) of its parent concept. For example, in a full concept 'extent of student participation in extracurricular activities', if 'student' is considered the head noun, the hierarchical relations of the component concepts are expressed as follows:

- [student] –

(facet concept) → [student participation] –

(subclass concept) → [student participation in extracurricular activities] –

(facet concept) → [extent of student participation in extracurricular activities]

The component concepts of different lengths have specific kinds of syntactic variations sharing the same head noun. They are considered a group of term variants representing similar concepts at different levels of generality.

In a set of similar dissertation abstracts, we selected high frequency nouns as the head nouns. Starting from each selected noun, a list of term chains was constructed by linking it level by level with other multi-word terms in which the single noun is used as the head noun. Each chain is constructed top down by linking the short term first, followed by longer terms containing the short term. The shorter terms represent the broader concepts at the higher level, whereas the longer terms represent the narrower concepts at the lower level. The root node of each chain is a noun (or one-word term) representing the main concept, and the leaf node is a full term representing the specific concept occurring in a particular document. The length of the chains can be different by linking different numbers of *n*-word terms but the maximum length is limited to six nodes, i.e. the one-, two-, three-, four-, five-word terms and the full term.

All the chains sharing the same root node (single noun) are combined to form a hierarchical cluster tree (see Figure 2). Each cluster tree uses the root node as its cluster label and contains two concepts at least. The concepts in round boxes represent *subclass concepts* of their parent concepts whereas the concepts in rectangular boxes represent *facet concepts*. The specific concepts occurring in particular documents which are highlighted using shaded boxes are usually at the bottom of the cluster.

In the hierarchical cluster tree, some broader concepts at higher levels are selected to generalize the whole cluster. For example, the main concept at the top level and the second level are used to generalize all the similar concepts related to 'student' and integrated into a summary sentence as follows:

- Student, including college student, undergraduate student, Latino student, …
  Its different aspects are investigated, including characteristics of student, behavior of student, …

The second-level concepts are divided into two groups – *subclass* concepts and *facet* concepts. Thus the summary sentence is divided into two parts – the first part ('including') giving the subclass concepts and the second part ('its different aspects') giving the facet concepts.

3.4.2. Relationship normalization and conflation

To integrate relationships, we identified different types of relationships found in the 300 sample abstracts through manual analysis. Nine types of semantic relationships including five first-order

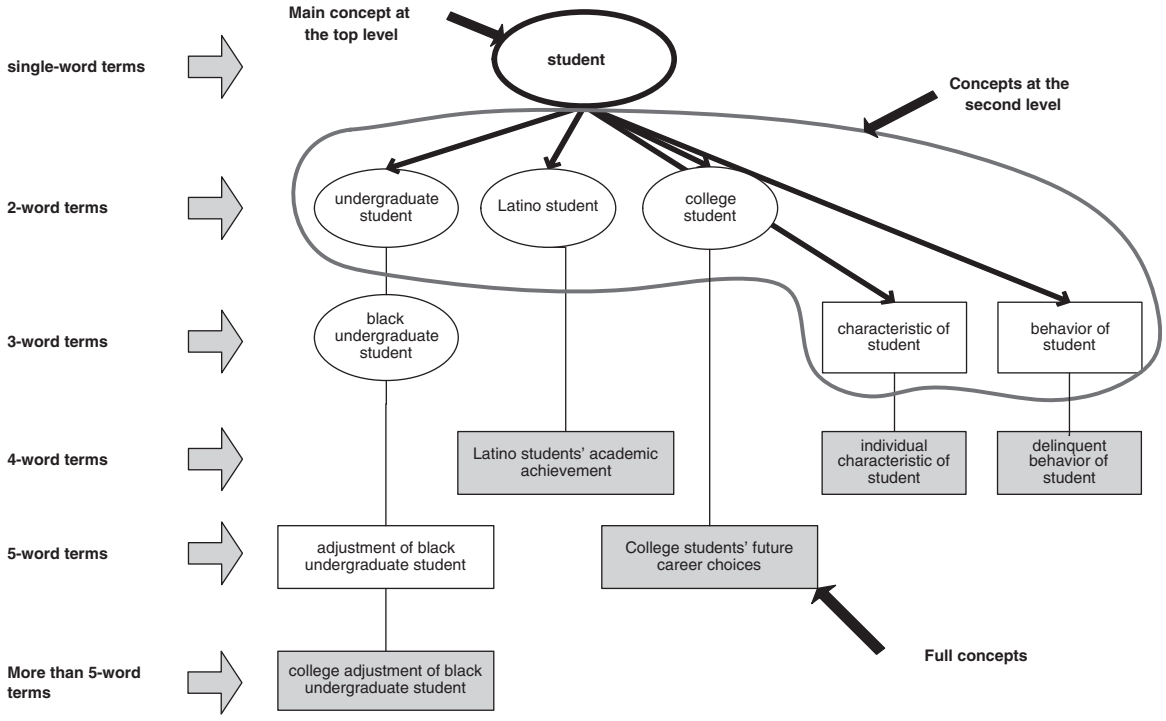

Fig. 2. A cluster tree containing five term chains.

relationships and four second-order relationships were found and listed in Table 3. The second-order relationship refers to the relationship between two or more variables influenced by a third variable. For example, a moderator variable influences the relationship between two variables, whereas a mediator variable occurs between two other variables. A total of 126 relationship patterns were constructed representing different surface expressions of the same types of relationship.

The different surface expressions for the same type of relationship can be normalized using a predefined standard expression. If two variables in a relationship are distinguished in the text as independent variable (IV) and dependent variable (DV), two standard expressions are provided by regarding each of the variables as the main variable. For each standard expression, three modalities are handled – *positive*, *negative* or *hypothesized*. For example, for a cause–effect relationship with the independent variable as the main variable, the three modalities are:

- *Positive:* There was an effect on a <dependent variable>.
- *Negative:* There was no effect on a <dependent variable>.
- *Hypothesized:* There may be an effect on a <dependent variable>.

Table 3
Nine types of semantic relationship

ID	First-order relationship	Second-order relationship
1	Cause–effect relationship	Second-order cause–effect relationship
2	Correlation	Second-order correlation
3	Interactive relationship	Second-order interactive relationship
4	Comparative relationship	Second-order comparative relationship
5	Predictive relationship	–

Some relationship patterns are only for negative relations, e.g. '<slot: variable 1> be *unrelated* with <slot: variable 2>', whereas some are only for hypothesized relations, e.g. '<slot: DV> *may be affected* by <slot: IV>'. However, not every negative relation could be indicated in the patterns. In this study, if a relationship contained a negative cue word (e.g. *no*, *not*, *negative*), it was considered a negative relation.

Similar concepts are identified and clustered as described in the last section. The relationships for similar concepts are clustered together. For example, the following relationships are associated with the main concept 'student':

- Expected economic returns affected the college students' future career choices.
- School socioeconomic composition has an effect on Latino students' academic achievement.
- School discipline can have some effect on the delinquent behavior of students.

In each cluster of relationships, the relationships with the same type and modality are normalized using a standard expression. For example, the above cause–effect relationships associated with 'student' are normalized using the standard expression: '<DV> was affected by <IV>'.

For each cluster of relationships, the normalized relationships using the same expression are conflated by combining the variables with the same roles together. Thus the above relationships associated with 'student' are conflated into a simple summary sentence as follows:

- Different aspects of students were affected by expected economic returns, school socioeconomic composition and school discipline.

Here, 'different aspects of students' refers to 'future career choices', 'academic achievement' and 'delinquent behavior'. The summary sentence provides an overview of all the variables that have a particular type of relationship with the given variable 'student'.

3.5. Summary presentation

In summary presentation, the four kinds of information, i.e. *research concepts* and *relationships*, *contextual relations* and *research methods*, are combined and organized to generate a summary. The summary is presented in an interactive web-based interface rather than traditional plain text so that it not only provides an overview of the topic but also allows the user to zoom in and explore more details of interest.

How to present a multi-document summary in fluent text and in a form that is useful to the user is an important issue. Although sentence-oriented presentation is extensively used in summarization, a few studies have presented concepts (terms) in addition to the important sentences as the components of a summary. Aone et al. [30] 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 and easy browsing. The full text was presented in the right window, in which the sentences identified for generating the summary were highlighted. Ando et al. [31] 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.

In our study, a simple concept-oriented presentation design was adopted for presenting the summary. It is concise and useful for quick information scanning. Figure 3 gives a screen snapshot of a summary. The contextual relations, research methods and research concepts extracted from different dissertation abstracts are presented as concept lists, whereas the normalized and conflated relationships are presented as simple sentences.

As shown in Figure 3, the four kinds of information (i.e. *research concepts* and *relationships*, *contextual relations* and *research methods*) are organized separately in the main window. This design can give users an overview for each kind of information and is also easy to implement.

Fig. 3. A presentation design for the concept-based multi-document summary.

Contextual relations and research methods found in the dissertation abstracts are presented first because these two kinds of information are usually quite short and may be overlooked by users if presented at the bottom of the summary. However, presenting them in this way has the disadvantage that they are presented out of context. Contextual relations and research methods are closely related to specific research concepts and relationships investigated in the dissertations, and provide details of how the concepts and relationships are studied. In future work, new presentation formats that integrate contextual relations and research methods with their corresponding research concepts and relationships can be developed.

Research concepts extracted from the dissertation abstracts are organized into broad subject categories, determined by a semi-automatically constructed taxonomy. Construction and use of the taxonomy has been reported by Ou et al. [32]. A list of subject categories gives users an initial overview of the range of subjects covered in the summary and helps them to locate subjects of interest quickly. Under each subject category, the extracted concepts are presented as concept clusters – each cluster is labeled by a single-word term called *main concept*. For each main concept, a concept list is presented, giving a list of related terms found in the dissertation abstracts. The concept list is divided into two subgroups – one for subclass concepts and another for facet concepts. The important concepts in the sociology domain, determined by the taxonomy, are highlighted in red.

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

4. Evaluation

In this study, the summarization system was evaluated at two levels:

1. *Intermediate component evaluation:* evaluating the accuracy and usefulness of each summarization step;
2. *Final user evaluation:* evaluating the overall quality and usefulness of the generated summaries.

The evaluation for each major summarization step was accomplished by comparing the system-generated output against human coding to address the following questions:

Q1. How accurate is the automatic discourse parsing?

Q2. Is the macro-level discourse parsing useful for identifying the important concepts?

Q3. How accurate is the automatic extraction of research concepts and relationships, contextual relations and research methods?

Q4. How accurate is the automatic concept integration?

Since there is no single 'gold standard', more than one human coding was used. The human coders were social science graduate students at Nanyang Technological University, Singapore.

The generated summaries were evaluated in a user evaluation carried out by researchers in the field of sociology.

4.1. Evaluation of macro-level discourse parsing

To evaluate the accuracy of automatic discourse parsing (i.e. sentence categorization), 50 structured abstracts were selected using a random table from the set of 3214 sociology dissertation abstracts published in 2001. Four human coders were asked to manually assign each sentence to one of the five sections – *background*, *research objectives*, *research methods*, *research results* and *concluding remarks*. The sections assigned by the system were compared against those assigned by the four coders. The percentage agreement between the coders and the percentage agreement between the system and the coders (i.e. system accuracy) were calculated. The accuracy of the system for identifying different sections in the 50 structured abstracts is given in Table 4.

The obtained inter-coder agreement is 79.6%, which is considered satisfactory. However, a lower agreement of 63.4% was obtained between the system and the coders. In the summarization process, only two sections – *research objectives* and *research results* – were used to extract important research information. Thus the identification of these two sections was more important than the other sections. The system worked well in identifying the *research objectives* and *research results* sections with a high accuracy of 90.8%.

4.2. Evaluation of information extraction

The above 50 structured abstracts were also used in the evaluation of information extraction. Three human coders were asked to extract all the *important concepts* manually from the whole text of each

Table 4
Accuracy of the system for identifying different sections in the 50 structured abstracts

Human coder as standard	All five sections	Research objectives (Section 2)	Research results (Section 4)	Research objectives + research results (Sections 2 and 4)
Coder 1	64.2%	71.1%	90.6%	92.7%
Coder 2	61.8%	62.4%	91.0%	90.0%
Coder 3	65.7%	58.8%	92.3%	90.4%
Coder 4	61.9%	58.2%	91.7%	90.1%
Average	63.4%	62.6%	91.4%	90.8%

Table 5
Average precision, recall and *F*-measure for the system-extracted research concepts from the three combinations of sections in the 50 structured abstracts

Importance level		All five sections	Research objectives (Section 2)	Research objectives + research results (Sections 2 and 4)
For the most important concepts	Precision (%)	20.36	31.62	23.60
	Recall (%)	92.26	76.06	87.37
	F-measure (%)	33.15	**43.91***	36.80
For the more important concepts	Precision (%)	31.02	44.51	34.28
	Recall (%)	90.93	59.31	78.81
	F-measure (%)	45.94	**50.27***	47.35
For all the important concepts	Precision (%)	46.18	59.05	49.76
	Recall (%)	89.93	46.65	75.64
	F-measure (%)	**60.44**	51.64	**59.40**

Bold figures indicate the highest values at each importance level.
The more important concepts include the most important concepts.
The important concepts include the more important concepts.
Asterisk indicates that the figure is significantly higher than other figures in the same row.

abstract, and from these to identify the *more important concepts* and then the *most important concepts*, according to the focus of the dissertation research. Meanwhile, we also used the system to extract research concepts automatically from the following three combinations of sections:

- from *research objectives* (section 2) only;
- from *research objectives* + *research results* (sections 2 and 4);
- from the *whole text* (i.e. all five sections).

The system-extracted concepts under the above three combinations were compared against the human-extracted concepts at three importance levels. The average precision, recall and *F*-measure for the system-extracted research concepts from the three combinations of sections in the 50 structured abstracts are given in Table 5.

As shown in Table 5, considering all the *important* concepts, the *F*-measures obtained from the *whole text* (60.4%) and from *research objectives* + *research results* (59.4%) were similar, both of which were higher than that from *research objectives* only (51.6%). This suggests that the *important* concepts were not focused only in *research objectives*, but scattered in the whole text. Therefore, the discourse parsing may not be helpful for identifying the *important* concepts. For the *more important* concepts, the *F*-measure obtained from *research objectives* (50.2%) was significantly higher than those from *research objectives* + *research results* (47.4%) and from the *whole text* (45.9%). This suggests that the *research objectives* section places a bit more emphasis on the *more important* concepts. For the *most important* concepts, the *F*-measure obtained from *research objectives* (43.9%) was significantly higher than those from *research objectives* + *research results* (36.8%) and from the *whole text* (33.2%). This suggests that the *research objectives* section places more emphasis on the *most important* concepts. Moreover, the *F*-measure obtained from *research objectives* + *research results* (36.8%) was significantly higher than that from the *whole text* (33.2%). This suggests that the *research results* section also places more emphasis on the *most important* concepts than the other three sections (i.e. *background*, *research methods* and *concluding remarks*). In conclusion, discourse parsing was helpful in identifying the *more important* and the *most important* concepts in structured abstracts. The *more* and *most* important concepts are more likely to be considered as research concepts.

In addition, the other three kinds of information – *relationships*, *contextual relations* and *research methods* – were extracted manually from the whole text of the 50 abstracts by two of the authors of this paper, who are deemed to be 'experts'. Experts are needed to do this coding because these three kinds of information are difficult to identify without substantial knowledge and training.

Table 6
Precision and recall for the system-extracted contextual relations, research methods and relationships in the 50 structured abstracts

Information piece	Precision	Recall
Relationships	81.02%	54.86%
Contextual relations	85.71%	90.00%
Research methods	97.20%	71.65%

From the two human codings, a 'gold standard' was constructed by taking the agreements in the codings. Differences in the codings were resolved through discussion. The average precision and recall for the system-extracted contextual relations, research methods and relationships in the 50 structured abstracts are given in Table 6.

The system obtained a high precision of 97.2% for extracting research methods and a little lower precision of 85.7% for extracting contextual relations. This indicates that it is effective to use cue phrases to identify these two kinds of information. However, the recall of 90.0% for extracting contextual relations is much higher than that of 71.7% for extracting research methods. This is because research methods can be expressed in various ways. Thus the list of cue phrases for research methods used in the summarization system was not complete since it was only derived from the 300 sample abstracts. Moreover, the research methods expressed in other grammatical forms, such as verb, adverb, and the whole sentence, cannot be identified by the system. In contrast, contextual relations are very specific information. It is easy to derive the cue words for contextual relations exhaustively from the 300 sample abstracts.

The system obtained a high precision of 81.0% for extracting relationships between research concepts, but a low recall of 54.9%. The list of relationship patterns derived from the 300 sample abstracts appears to be incomplete. Moreover, the system can only identify relationships that are located within sentences and with clear cue phrases. Cross-sentence relationships and implied relationships that do not contain clear cue phrases and need inferring cannot be identified with the current pattern matching method.

4.3. Evaluation of information integration

For evaluating the quality of clusters, two types of measures – *internal quality measures* and *external quality measures* – have been used [33]. Internal measures calculate the internal quality of a set of clusters without reference to external knowledge, e.g. overall internal similarity based on the pairwise similarity between members within each cluster. External measures compare how closely a set of clusters matches a set of known reference clusters. In this study, we adopted an external measure – *F*-measure from the field of information retrieval – to calculate the similarity between the set of system-generated clusters and the reference clusters. Two sets of human codings were each used as reference clusters.

In the evaluation, 15 research topics in the sociology domain were haphazardly selected. For each topic, a set of dissertation abstracts was retrieved from the database using the topic as search query. But only five abstracts were selected from the retrieved abstracts to form a document set. In addition, for five of the topics (i.e. document set 11–15), an additional five abstracts were selected for each of them and combined with the previously chosen five abstracts to form a second bigger document set. Thus 20 document sets in total were used in the evaluation. The bigger document sets were used to examine the difference in concept clustering between small sets (five documents) and bigger sets (10 documents). For each abstract, the important concepts were automatically extracted by the system from the *research objectives* and *research results* sections of the abstract.

Human coders were asked to identify similar concepts across abstracts from the list of concepts extracted from each document set and group them into clusters. Each cluster had to contain two concepts at least and was assigned a label by the coders. Thus some of the concepts in the concept

Table 7
Number of concepts used by each of the three clusterings and number of common concepts between any two clusterings

Document set	Total number of concepts for clustering	Number of concepts used by			Number of common concepts between		
		Coder 1	Coder 2	System	Coder 1 and Coder 2	System and Coder 1	System and Coder 2
5-document sets (n = 15)	114.5 (100%)	47.8 (41.7%)	43.4 (39.8%)	68 (59.4%)	29.5 (41.8%)*	35.1 (43.5%)*	35.8 (47.1%)*
10-document sets (n = 5)	225.6 (100%)	89 (39.4%)	80.6 (35.7%)	150.6 (66.8%)	46.6 (37.9%)*	72 (43.0%)*	73 (46.1%)*

* Note: the percentage is calculated by dividing the number of common concepts between two clusterings by the total number of the unique concepts used by the two clusterings, which equals the number of concepts used by clustering 1 plus the number of concepts used by clustering 2 minus the number of common concepts between two clusterings.

list were not selected to form clusters by the coders, presumably because there were no perceived similarities with other concepts. As mentioned earlier, for each document set, two sets of clusters were generated by two coders and one set of clusters was generated by the system.

Table 7 shows the number of concepts used in each of the three clusterings and the number of common concepts between any two clusterings. The system 'worked harder' than the human coders and used a higher number of concepts in the concept list to create clusters. For example, for the five-document sets, the system clustered 59.4% of the given concepts whereas the human coders only clustered 40.8% of the concepts on average. Furthermore, the system clustering had more concepts in common with each of the human clusterings than between two human clusterings. For the five-documents sets, the concepts selected by the system and each of the human coders overlap by 45.3% on average compared to 41.8% for the overlap between two human coders. When the size of document sets increased to 10 documents, the human clustering became more difficult. The percentage of common concepts between two human coders decreased from 41.8 to 37.9%. However, the percentage of common concepts between the system and each of the human coders remained at almost the same level (44.6%). This suggests that the system can handle bigger document sets without degradation.

To measure the similarity between the system-generated clusters and human-generated clusters, we adopted an *F*-measure-based method, as employed by Steinbach et al. [33] and Larsen and Aone [34]. For calculating the *F*-measure, each system-generated cluster is treated as the result of a query and each human-generated cluster as the desired set of concepts for a query. The recall and precision of a system cluster (j) for a given human cluster (i) are calculated as follows:

- precision (i, j) = number of common concepts between a system cluster (j) and a human cluster (i) /Number of concepts in a system cluster (j);
- recall (i, j) = number of common concepts between a system cluster (j) and a human cluster (i)/ number of concepts in a human cluster (i).

Then, the *F*-measure is calculated as the weighted harmonic mean of precision and recall:

- *F*-measure (i, j) = 2* precision (i, j) * recall (i, j) /[precision (i, j) + recall (i, j)].

For a given human cluster (i), the *F*-measure used is the highest one obtained among the entire set of system clusters. Thus, the overall *F*-measure is calculated by taking the weighted average of the F-measures for each human cluster in the entire set:

- Overall *F*-measure = Σ_i[(number of concepts in a human cluster (i) /total number of concepts in the set of human clusters) * max {*F*-measure (i,j)}].

The overall *F*-measures for the sets of system-generated and human-generated clusters are given in Table 8.

Table 8
Overall *F*-measures for the set of system-generated clusters and human-generated clusters

	Human coding 1 as reference clusters		Human coding 2 as reference clusters	
Document set	System	Coder 2	System	Coder 1
5-document sets ($N = 15$)	51.4	47.5	67.8	54.7
10-document sets ($N = 5$)	52.0	43.8	63.2	44.7
Average for all document sets	51.6	46.7	66.7	52.2

Considering human coding 1 as the reference clusters for evaluation, the system obtained higher *F*-measures for 16-document sets than human coder 2. The average *F*-measure obtained by the system (51.6) for the 20-document sets is higher than that obtained by human coder 2 (46.7). With human coding 2 as the reference clusters, the system obtained higher *F*-measures for 15-document sets than human coder 1. The average *F*-measure obtained by the system (66.7) is also higher than that obtained by human coder 1 (52.2). This suggests that the system clustering has a higher similarity score to each of the human codings than that between the human codings! There are two reasons for this:

1. the system used a higher number of concepts to form clusters and the concepts selected by the system have a higher percentage overlap with each human coder than between two human coders;
2. the system created many small clusters with highly similar concepts, whereas the humans had a macro perspective to do clustering and thus created bigger clusters.

4.4. User evaluation

Finally, a user evaluation was carried out to evaluate the overall quality and usefulness of the summaries. Only a summary of the results is given here. The detailed results are reported in a separate paper [35].

Thirty researchers in the field of sociology participated in the user evaluation. Each researcher was asked to submit one research topic that he/she was working on or had worked on. For each topic, a set of sociology dissertation abstracts was retrieved from the Dissertation Abstracts International database using the topic as the search query, and condensed into a summary. Four types of summary were provided for each topic:

1. a variable-based summary generated using our summarization system but *without* the use of the taxonomy (labeled SYSTEM 1);
2. a variable-based summary generated using our summarization system *with* the use of the taxonomy (labeled SYSTEM 2);
3. a sentence-based summary generated by extracting the research objectives sentences of each abstract only (labeled OBJECTIVES); and
4. a sentence-based summary generated by a state-of-the-art summarization system MEAD 3.08 [36], which uses a sentence extraction method (labelled MEAD).

The researchers were asked to rank the four types of summary. The overall ranking obtained was:

1. SYSTEM 2
2. OBJECTIVES
3. SYSTEM 1
4. MEAD.

The researchers were also asked to select one or more summaries that they preferred to use for their research-related work. Some 70% of the researchers indicated preference for the variable-based summaries to the sentence-based summaries. They indicated that the variable-based summaries were efficient in giving an overview of the topic and useful for information scanning. But some users also indicated that the variable-based summaries were too brief to provide accurate information on the topic and had potential to confuse users. Also, 55% of the researchers indicated preference for the research objective summaries, and 25% of the researchers indicated preference for the MEAD summaries. They indicated that the sentence-based summaries could provide more direct information and were easy to understand. On the other hand, complete sentences were more time-consuming to read than concept lists.

5. Conclusion

In this study, we have developed an automatic method for summarizing sets of dissertation abstracts in sociology that might be retrieved by an information retrieval system or web search engine in response to a user query. The summarization process includes four major steps – discourse parsing, information extraction, information integration, and summary presentation. Each of the major steps was evaluated by comparing the system-generated output against human coding.

In discourse parsing, a decision tree classifier was developed to categorize sentences into five standard sections. The system obtained an overall accuracy of 63%, which was rather lower than the inter-coder agreement of 80%. However, the accuracy of 91% obtained for identifying the *research objectives* and *research results* sections was quite high. In the future, other supervised learning techniques such as SVM and Naive Bayes will be investigated. Since supervised learning requires manual assignment of predefined category labels to the training data, which is time consuming, unsupervised learning can also be investigated for parsing the discourse structure of research abstracts and research articles.

In term extraction, we used a rule-based method employing syntactic rules to extract multi-word terms. The system obtained a high recall of 90% for extracting important concepts from dissertation abstracts but the precision of 46% was low. Statistics-based methods can be investigated in the future by examining the statistical associations among the component words in multi-word terms to refine the extracted terms. Among the extracted terms, we selected as research concept terms those extracted from the *research objectives* and *research results* sections. Furthermore, we identified contextual relation terms and research method terms throughout the whole text using cue phrases. The accuracy obtained was good – 86% precision and 90% recall for contextual relations, and 97% precision and 72% recall for research methods. However, this method cannot recognize the contextual relations and research methods expressed in other grammatical forms such as adverbs, verbs and infinitive phrases.

In relationship extraction, we pre-constructed a set of relationship patterns and performed pattern matching to identify the text segments that match with the patterns. This obtained a high precision of 81% but the recall of 55% was low. In the future, relationships across sentences and implied relationships without clear cue phrases will be explored.

In information integration, we performed only syntactic-level generalization since it is easy to realize without the need of an ontology, taxonomy or thesaurus. Although the system clustering is more similar to the human codings (e.g. *F*-measure = 51.6) than between the human codings (e.g. F-measure = 46.7), such generalization is not very accurate without considering the semantic meanings of concepts. Semantic-level generalization using a taxonomy or ontology can be investigated.

In summary presentation, we adopted a simple concept-oriented design to present the summary. However, a well-designed summary presentation is important for end-users. Other presentation designs can be used for operationalizing the variable-based framework. Graphical presentation has been used for single-document summaries in many previous studies, e.g. DimSum [30]. But there are few such studies found for multi-document summaries. Actually, multi-document summaries need a more sophisticated user interface. A multi-document summary is required to provide a

domain-overview of a topic and also allow users to zoom in for more details on aspects of interest. A graphical interface can help users to interact with the summary to locate what they want more rapidly and effectively.

Although there is a large body of literature on how to write good single-document summaries or abstracts, not much is found on how to write good multi-document summaries and literature surveys (summarizing a set of documents is like writing a literature survey). More studies are needed to find out how good literature surveys are written and structured in different situations (e.g. for different purposes and users). More intelligent and useful summarization systems can be developed by following the human cognitive process in summarizing a set of documents and writing a literature survey.

References

[1] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell, Summarizing text documents: sentence selection and evaluation metrics. In: F. Gey et al. (eds), *Proceedings of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval* (ACM, New York, 1999) 121–8.

[2] A. Spink and J.L. Xu, Selected results from a large study of web searching: the Excite study [electronic version], *Information Research* 6(1) (2000). Available at: http://informationr.net/ir/6-1/paper90html (accessed 31 October 2007).

[3] J.J. Macionis, *Sociology* (8th edn) (Prentice Hall, Upper Saddle River, 2000).

[4] H.P. Edmundson, New methods in automatic extracting, *Journal of the ACM* 16(2) (1969) 264–85.

[5] C.D. Paice, Constructing literature abstracts by computer: techniques and prospects, *Information Processing and Management* 26(1) (1990) 171–86.

[6] J. Kupiec, J. Pedersen and F. Chen, A trainable document summarizer. In: E. Fox et al. (eds), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 1995) 68–73.

[7] R. Barzilay and M. Elhadad, Using lexical chains for text summarization (1997). In: U. Hahn et al. (eds), *Proceedings of the ACL 1997 Workshop on Intelligent Scalable Text Summarization*. Available at: http://acl.ldc.upenn.edu/W/W97/W97–0703.pdf (accessed 10 April 2006).

[8] S. Azzam, K. Humphreys and R. Gaizauskas, Using coreference chains for text summarization (1999). In: A. Bagga et al. (eds), *Proceedings of the ACL 1999 Workshop on Coreference and its Applications*. Available at: http://acl.ldc.upenn.edu/W/W99/W99–0211.pdf (accessed 10 April 2006).

[9] R.D. Radev, H. Jing and M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In: E. André et al. (eds), *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization* (ACL, Morristown, NJ, 2000) 21–9.

[10] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. Wise and X. Zhang, Cross-document summarization by concept classification. In: M. Beaulieu et al. (eds), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 2002) 121–8.

[11] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman and S. Teufel, Columbia multi-document summarization: approach and evaluation. In: *Proceedings of the Document Understanding Conference 2001* (NIST, 2001). Available at: www-nlpir.nist.gov/projects/duc/pubs.org.html (accessed 10 May 2006).

[12] J.G. Carbonell and J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: W.B. Croft et al. (eds), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 1998) 335–6.

[13] S. Afantenos, V. Karkaletsis and P. Stamatopoulos, Summarization from medical documents: a survey, *Journal of Artificial Intelligence in Medicine* 33(2) (2005) 157–77.

[14] C.Y. Lin, Topic identification by concept generalization. In: H. Uszkoreit et al. (eds), *Proceedings of the 33rd Annual Meeting of the Association for Computation Linguistics* (ACL, Morristown, NJ, 1995) 308–10.

[15] K. McKeown and D. Radev, Generating summaries of multiple news articles. In: E.A. Fox, P. et al. (eds), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 1995) 74–82.

[16] Z. Zhang, S. Blair-Goldensohn and D. Radev, Towards CST-enhanced summarization. In: R. Dechter et al. (eds), *Proceedings of the 18th National Conference on Artificial Intelligence* (AAAI, Menlo Park, CA, 2002) 439–45.

[17] S. Afantenos, I. Doura, E. Kapellou and V. Karkaletsis, Exploiting cross-document relations for multi-document evolving summarization. In: G.A. Vouros and T. Panayiotopoulos (eds), *Methods and Applications of Artificial Intelligence in Volume 3025 of Lecture Notes in Computer Science*: *Proceedings of the 3rd Helenic Conference on Artificial Intelligence* (Springer, Berlin, 2004) 410–19.
[18] T. Pasi and J. Timo, A non-projective dependency parser. In: R. Grishman et al. (eds), *Proceedings of the 5th Conference on Applied Natural Language Processing* (Morgan Kaufmann, San Francisco, CA, 1997) 64–71.
[19] S. Kurohashi and M. Nagao, Automatic detection of discourse structure by checking surface information in sentences. In: Y. Wilks et al. (eds), *Proceedings of the 15th International Conference on Computational Linguistics* (ACL, Morristown, NJ, 1994) 1123–7.
[20] M. Hearst, Multi-paragraph segmentation of expository text. In: J. Pustejovsky et al. (eds), *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (ACL, Morristown, NJ, 1994) 9–16.
[21] H.T. Le and G. Abeysinghe, A study to improve the efficiency of a discourse parsing system. In: A.F. Gelbukh (ed.), *Volume 2588 of Lecture Notes in Computer Science: Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics* (Springer, Berlin, 2003) 101–14.
[22] D. Marcu, From discourse structure to text summaries. In: I. Mani and M. Maybury (eds), *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization* (ACL, Morristown, NJ, 1997) 82–8.
[23] T. Nomoto and Y. Matsumoto, Discourse parsing: a decision tree approach (1998). In: E. Charniak (ed.), *Proceedings of the 6th Workshop on Very Large Corpora*. Available at: http://acl.ldc.upenn.edu/W/W98/W98–1125.pdf (accessed 10 May 2006).
[24] R. Quanlan, *C5.0: an Informal Tutorial* (Rulequest Research, Sydney, 1998).
[25] D. Borgigault and C. Jacquemin, Term extraction + term clustering: an integrated platform for computer-aided terminology. In: H. Thompson et al. (eds), *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics* (ACL, Morristown, NJ, 1999) 15–22.
[26] H. Nakagawa, Automatic term recognition based on statistics of compound nouns, *Terminology* 6(2) (2000) 195–210.
[27] National Information Standard Organization (NISO), *ANSI/NISO Z39.19–2003: Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (2003). Available at: http://download.www.techstreet.com/cgi-bin/pdf/free/403963/Z39–19_2003.pdf (accessed 10 May 2006).
[28] R. Bunescu, R. Mooney, A. Ramani, E. Marcotte, Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline. In: K. Verspoor et al. (eds), *Proceedings of the HLT-NAACL 2006 Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis* (ACL, Morristown, NJ, 2006) 49–56.
[29] F. Ibekwe-SanJuan and E. SanJuan, Mining textual data through term variant clustering: the TermWatch system. In: C. Fluhr et al. (eds), *Proceedings of RIAO 2004* (C.I.D., Paris, 2004) 487–503.
[30] C. Aone, M. Okurowski, J. Gorlinsky and B. Larsen, A trainable summarizer with knowledge acquired from robust NLP techniques. In: I. Mani and M.T. Maybury (eds), *Advances in Automatic Text Summarization* (MIT Press, Cambridge, MA, 1999) 71–80.
[31] R. Ando, B. Boguraev, R. Byrd and M. Neff, Multi-document summarization by visualizing topic content. In: E. André et al. (eds), *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization* (ACL, Morristown, NJ, 2000) 79–98.
[32] S. Ou, C. Khoo and D. Goh, Constructing a taxonomy to support multi-document summarization of dissertation abstracts, *Journal of Zhejiang University SCIENCE* 6A(11) (2005) 1258–67.
[33] M. Steinbach, G. Karypis and V. Kumar, A comparison of document clustering techniques. In: M. Grobelnik et al. (eds), *Proceedings of KDD 2000 Workshop on Text Mining held with the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2000). Available at: http://rakaposhi.eas.asu.edu/cse494/notes/clustering-doccluster.pdf (accessed 31 October 2007).
[34] B. Larsen and C. Aone, Fast and effective text mining using linear-time document clustering. In: S. Chaudhuri et al. (eds), *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 1999) 16–22.
[35] S. Ou, C. Khoo and D. Goh, Automatic multi-document summarization of research abstracts: design and user evaluation. *Journal of the American Society for Information Science & Technology*, 58(10) (2007) 1–17.
[36] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, E. Drabek, W. Lam, D. Liu, H. Qi, H. Saggion, S. Teufel, M. Topper and A. Winkel, *The MEAD Multidocument Summarizer* (2003). Available at: www.summarization.com/mead/ (accessed 24 May 2006).