Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Constructing a taxonomy to support multi-document summarization of dissertation abstracts

OU Shi-yan, KHOO Christopher S.G., GOH Dion H.

(*Division of Information Studies, School of Communication & Information, Nanyang Technological University, 639798, Singapore*)

E-mail: pg00096125@ntu.edu.sg; assgkhoo@ntu.edu.sg; ashlgoh@ntu.edu.sg

Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: This paper reports part of a study to develop a method for automatic multi-document summarization. The current focus is on dissertation abstracts in the field of sociology. The summarization method uses macro-level and micro-level discourse structure to identify important information that can be extracted from dissertation abstracts, and then uses a variable-based framework to integrate and organize extracted information across dissertation abstracts. This framework focuses more on research concepts and their research relationships found in sociology dissertation abstracts and has a hierarchical structure. A taxonomy is constructed to support the summarization process in two ways: (1) helping to identify important concepts and relations expressed in the text, and (2) providing a structure for linking similar concepts in different abstracts. This paper describes the variable-based framework and the summarization process, and then reports the construction of the taxonomy for supporting the summarization process. An example is provided to show how to use the constructed taxonomy to identify important concepts and integrate the concepts extracted from different abstracts.

Key words: Text summarization, Automatic multi-document summarization, Variable-based framework, Digital library
doi:10.1631/jzus.2005.A1258 **Document code:** A **CLC number:** TP391

INTRODUCTION

Automatic summarization has attracted much attention both in the research community and business community as a solution for reducing information overload and helping users scan a large number of documents to identify documents of interest. 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 (relationships between pieces of information in different documents), and can allow users to zoom in for more details on aspects of interest.

Multi-document summarization is useful in large digital libraries, especially in academic institutions and can be used for knowledge discovery to identify connections between research results that are not obvious, and bridge gaps in the field for future research. Since multi-document summarization combines and integrates information across documents, it performs knowledge synthesis and knowledge discovery, and can be used for knowledge acquisition.

This study is aimed at developing an automatic method for summarizing a set of related 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 started to provide online dissertation abstract services, since they contain a wealth of high-quality information on research objectives, research methods and research 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 main ideas on a specific topic.

The main approaches used for multi-document

summarization include sentence extraction (Radev *et al*., 2000), template-based information extraction (McKeown and Radev, 1995), and identification of similarities and differences across documents (Mani and Bloedorn, 1999). However, these existing summarization approaches focus more on extracting salient information from different documents. They use shallow analysis, without paying much attention to high-level relations and semantics expressed within and across documents. Another problem is that different users have different information needs. But these approaches usually construct fixed multi-document summaries.

This study does not use the traditional statistics-based summarization approaches, but uses the macro-level and micro-level discourse structure of sociology dissertation abstracts to identify and extract desired information and uses the cross-document structure to identify the similarities and differences across the documents. Then a hierarchical framework focusing on research concepts and their research relationships is used to integrate and organize the extracted information across documents and represent the multi-document summary in a Web-based interface. Radev (2000) proposed the Cross-document Structure Theory (CST) and applied it to multi-document summarization in an arbitrary domain. The difference between our work and Radev's is that we pay more attention to higher-level semantic content and semantic relations expressed in the text, whereas Radev focuses more on rhetorical structure and lower-level relations.

This paper reports the use of taxonomy for multi-document summarization of sociology dissertation abstracts. It describes the summarization method based on the variable-based framework and then reports the construction of taxonomy and taxonomy characteristics needed to support the summarization process.

Taxonomy plays an important role in information organization and knowledge organization. It has been used to facilitate information browsing and searching in many previous studies. For example, Wollersheim and Rahayu (2002) created a dynamic taxonomy for navigating medical text databases. Some studies also used the taxonomy to support the summarization work. For example, Endres-Niggemeyer *et al*.(2001) constructed an ontology to support summarization from the WWW for Bone Marrow Transplantation. Hovy and Lin (1999) used a lexical thesaurus WordNet to generalize concepts and thus to identify the topics of the text in a summarization system SUMMARIST. In this study, the taxonomy plays two important roles in the summarization process:

(1) It specifies the important concepts in the domain and the relations between the concepts. This can be used to identify similar concepts in different abstracts and the relations between the concepts. This is critical for delineating the cross-document structure of the set of abstracts.

(2) It provides a structure for linking similar concepts in different abstracts.

SUMMARIAZATION USING A VARIABLE-BASED FRAMEWORK

The summarization method in this study uses the macro-level (between sentences and sections) and micro-level (within sentences) discourse structure of sociology dissertation abstracts to identify which segments of the text contain the desired information and what kinds of information can be extracted from the specific segments, and further uses the cross-document structure to identify similar information, unique information, and relationships between pieces of information in different abstracts. A variable-based framework with a hierarchical structure is used to integrate extracted research concepts and their research relationships to generate a multi-document summary. A taxonomy is constructed to support the summarization by specifying the important concepts in sociology dissertation abstracts, and then identifying and clustering similar concepts. The summarization process involves the following four steps:

1. Parse the macro-level discourse structure

In a previous study, we analyzed the macro-level discourse structure of sociology dissertation abstracts and found that all the information in the abstracts can be subsumbed under five sections—background, research objectives, research methods, research results and concluding remarks (Khoo *et al*., 2002). An automatic discourse parsing method using a decision-tree induction technique was developed to identify these sections of sociology dissertation abstracts

by categorizing each sentence in an abstract into one of the five predefined categories or sections (Ou *et al*., 2004).

2. Extract desired information from the micro-level discourse structure

Among the five sections in dissertation abstracts, research objectives and research results sections are usually focused on research concepts and their research relationships. Through analyzing the micro-level structure of the research objectives and research results sections, four kinds of information can be extracted from these two sections (Ou *et al*., 2003):

(1) Research concepts that are often operationalized as research variables. In descriptive research, one or more research concepts are investigated to identify attributers of interest. In causal research, there are two kinds of variables—dependent variables and independent variables. Dependent variables are the variables that the researchers are interested in explaining or predicting, while independent variables are variables that affect or are used to predict the dependent variables. In relational research, however, variables are not distinguished as such.

(2) Relationships between variables. The relationship between a pair of variables in relational research or the effect of an independent variable on a dependent variable in causal research may be unknown, and determining the relationships may be the research objective of the sociological study. On the other hand, descriptive sociological studies may not be concerned with investigating relationships between variables, but instead, try to identify the attributes of the research concepts, such as entity or phenomenon.

(3) Contextual relations. Some studies do not explore relationships directly, but in the context of a framework, model, theory, hypothesis, etc., or in the perception or attitude of a target population. We call this a contextual relation.

(4) Research methods. In a dissertation study, one or more research methods including research design, sampling, data measure and analysis method are used to explore the relationships between variables or describe the attributes of the research concepts.

3. Identify similar information using a taxonomy

A taxonomy is developed for specifying the important concepts and their relations in the domain of sociology. Therefore, similar concepts can be identified and clustered by examining the different level concepts in the taxonomy to provide an overview of a specific topic.

4. Integrate extracted information using a variable-based framework

A variable-based framework with a hierarchical structure is used to integrate research concepts and their research relationships extracted from different documents and thus summarize a set of related dissertation abstracts. The hierarchy of the framework has four levels—dependent variable level, independent variable level, contextual relation level, and document level. Fig.1 shows some of information extracted from 10 related dissertation abstracts on the topic of "school crime" and integrated together using the variable-based framework.

At the dependent variable level, all research concepts of interest (e.g. often dependent variables in the case of causal research) in a document set are identified and integrated according to the different level concepts in the taxonomy. "School crime" and "school dropout" are two of broad concepts at the first level. "School crime" includes five narrower concepts at the second level, one of which is "school violence". "School violence" includes three lower-level concepts.

The independent variable level lists the corresponding independent variables for each dependent variable. The independent variable for "school crime" is "school district", and the independent variable for "school dropout" is "school size". "School crime" is related with "school district", whereas "school dropout" is not related with "school size". "School district" includes two narrower concepts "school district size" and "school district density". More detailed information on the relationship between "school crime" and "school district" is provided—i.e. it is a strong association. The relationship between "school crime" and "school district" is explored in a theoretical model with more details provided about the model.

Each abstract in the document set can be structured and summarized by dividing the text into the five sections and extracting important information from some specific sections. Currently, only the research objectives and research results sections are considered.

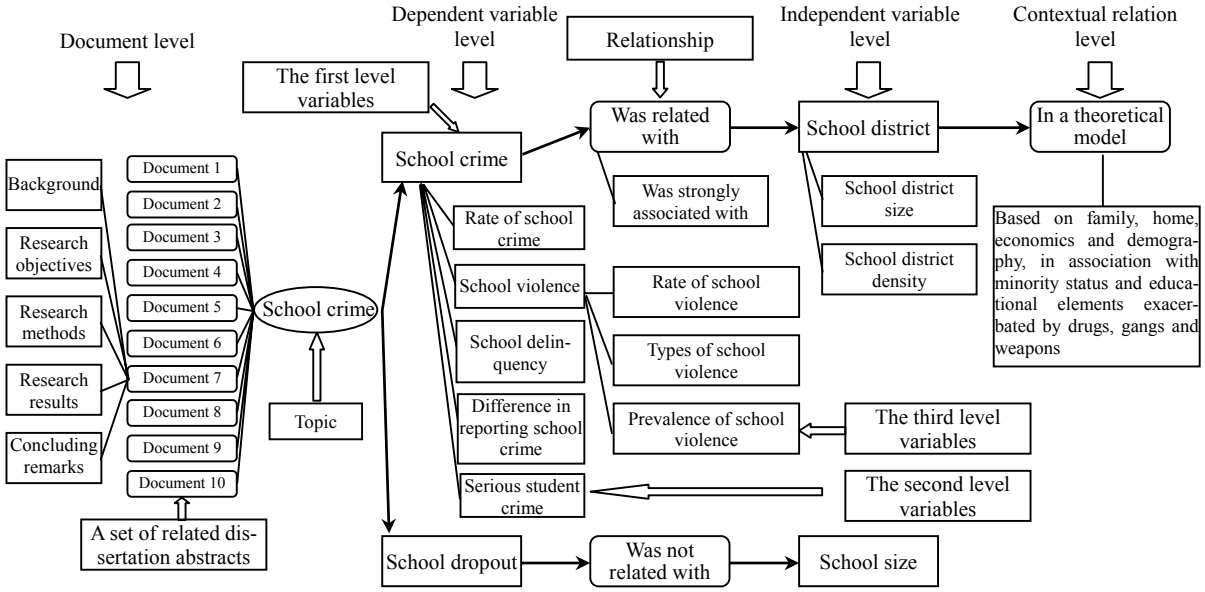

Fig.1 An example of integrated information using a variable-based framework

OVERALL APPROACH FOR CONSTRUCTING THE TAXONOMY OF CONCEPTS

A taxonomy is a set of concepts which are organized using a hierarchical structure to present IS-A (genus-species) and whole-part relations between concepts. A concept may be expressed by a single-word or multi-word term. The grammatical form for a concept should be a noun or noun phrase (NISO, 2003):

(1) Common nouns, such as school, family, culture, history;

(2) Verbal nouns, such as parenting, nursing;

(3) Noun phrases, which are multi-word terms which occur in two forms: Adjectival noun phrases which are generally pre-modified, such as social science, public administration, school crime, adopted children; Prepositional noun phrases which are generally post-modified, such as ability of the organization, history of European.

Adverbs such as "very" or "highly" are deleted. The initial articles are also deleted. For example, "the arts" should be "art", and "the state" should be "state".

Most multi-word terms, including pre- and post-modified noun phrases, can be analyzed into the following two parts:

(1) Head nouns: 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;

(2) Modifier: the part of a multi-word term that refers to a characteristic or logical difference, which narrows the denotation of the head noun by specifying a subclass of the broader concept represented by the head noun. For example, cognitive ability (a type of ability), educated woman (a subclass of woman).

To guide the taxonomy construction, we carried out a conceptual analysis of the concepts that can be extracted from the research objectives and research results sections, and identified the components of concepts and relations that are needed in the taxonomy.

A full concept is often expressed in a phrase composed of several single-word or multi-word terms. The component words or terms often refer to broader concepts than the full concept itself. That is, a full concept is a narrower concept which may be composed of some broader concepts. We call these broader concepts component concepts. They can be identified by segmenting a full term into words or shorter phrases of different lengths. In this study, we segmented a full term into 1-word terms, 2-word terms, 3-word terms, 4-word terms and 5-word terms. The terms with different number of words represent similar concepts at different levels and can be used to

construct a tangled hierarchy of concepts. The following example shows the component concepts of a full concept and their tangled hierarchy.

Here is a full concept extracted from an example research objective sentence (the full concept is underlined):

The purpose of the study was to investigate interaction patterns between young children and their mothers which foster creative thought.

The component concepts of the full concept are identified by segmenting a full term into words or short phrases of different lengths:

(1) 1-word terms: interaction, pattern, child, mother, thought;

(2) 2-word terms: interaction pattern, young child, creative thought;

(3) 3-word terms: –;

(4) 4-word terms: pattern between young child;

(5) 5-word terms: interaction pattern between young child, young child and its mother.

The tangled hierarchy of the component concepts and the full concept is:

[interaction] –
<- [interaction pattern] –
<- [interaction pattern between young child] –
<- [interaction pattern between young child and its mother which foster creative thought]
[child] –
<- [young child]
[thought] –
<- [creative thought]
[pattern] –
<- [interaction pattern] –
<- [pattern between young child] –
<- [interaction pattern between young child] –
<- [interaction pattern between young child and its mother which foster creative thought]

The component concepts of a full concept have relations between them, distinguished by their logical roles or functions. They can be a main component concept, an attribute concept, or a qualifier concept of a full concept. The main component concept can be a concrete object, an abstract object, a person, an organization, a place, an action, an event, a process, and so on. An attribute concept does not represent any specific entity, action or event, but specifies an aspect or a quality whereby the entities, actions, or events can be distinguished and measured quantitatively. The attribute may be unknown and needs to be investigated in the study. For example, "pattern" is the attribute of an action "interaction".

[interaction] –
(attribute) -> [pattern]

A qualifier concept, on the other hand, restricts or narrows the entities, actions, events or processes into a subset. For example, the scope of "interaction" is narrowed down by "between young child and its mother".

[interaction] –
(qualifier) -> [between young child and their mother]

In a previous study, we found a specific list of words used as attribute concepts (Ou *et al*., 2003). In this study, more words were found by analyzing more sample abstracts (300 abstracts). The words frequently used as attribute concepts in sociology dissertation abstracts are shown in Table 1.

A qualifier is usually more complex, and any concept can be a qualifier to restrict or narrow down other concepts. It is hard to find a specific list of words (concepts) for qualifiers.

We identified the single words and multi-word terms which occur frequently in sociology dissertation abstracts as the concepts in this domain to construct a taxonomy. The taxonomy provides a way to identify similar or related concepts and group them in a hierarchical structure. Thus, the similar concepts can also be linked together through clustering their component concepts since the component concepts are often different level boarder concepts of a full concept.

MACHINE AIDED TAXONOMY CONSTRUCTION

In this study, a sample of 3214 abstracts indexed

Table 1 Attribute concepts frequently used in sociology dissertation abstracts

Attribute	Example
Size	Family size, size of organization
Rate	Female suicide rate, rate of crime
Pattern	Cultural pattern, pattern of interaction
Type	Household type, type of social capital
Category	Racial category, social category
Level	Community level, level of academic achievement
Diversity	Cultural diversity, diversity of woman
Dimension	Gender dimension, dimension of personality
Predictor	Predictor of marital satisfaction, predictor of success
Role	Leadership role, role of ethnic identity
Score	Achievement score, depression score
Form	Child form, form of belief
Gap	Gender gap, gap between theory and practice
Difference	Gender difference, difference between family
Status	Class status, status of the student
Quality	Job quality, quality of life
Nature	Nature of woman, social nature
Degree	Degree of conflict, degree of trauma
Performance	Academic performance, performance of expertise
Function	Function of age, family function

under "sociology" subject, "PhD" degree and year of publication "2001" in the Dissertation Abstracts International database was used as the source for constructing the taxonomy. A semi-automatic method based on Microsoft Access database, Java computer programs and manual analysis, was used to construct the taxonomy using the following steps.

Step 1: Segment abstracts into sentences.

All the abstracts were segmented into sentences automatically with a simple computer program. Punctuation marks, such as period, question mark, and exclamation mark, were used to break sentences.

Step 2: Tokenize sentences into single words.

Sentences were tokenized and the words were stemmed using the Conexor parser (Japanainen and Jarvinen, 1997). The document frequency (df) was calculated for each unique word. A stoplist comprising prepositions, articles and auxiliary was used. Only high frequency non-stop words occurring in at least 100 documents (i.e. df≥100) were retained.

Step 3: Select head nouns.

We identified part-of-speech of each high frequency word based on the results of Conexor parser and selected nouns as the candidate head nouns, while the verbs, adjectives, adverbs were used to construct stoplists for identifying meaningful terms in the later step. The common words used in the dissertation abstracts, such as "purpose", "aim", "dissertation", "implication", "conclusion", "result", and indicator words for relationships between variables and contextual relations, such as "relationships", "impact", "associations", were deleted from the head nouns.

Step 4: Construct *n*-grams (*n*=2, 3, 4, or 5) around head nouns.

We extracted different number of continuous words from each sentence to construct *n*-grams. In this study, *n* can be 2, 3, 4, or 5. Only *n*-grams occurring in at least 2 documents (i.e. df≥2) were retained. The *n*-grams containing head nouns were identified using a computer program. The procedures are as follows:

(1) Identify 2-grams containing head nouns from the retained 2-grams;

(2) Identify 3-grams containing the previously found 2-grams from the retained 3-grams;

(3) Identify 4-grams containing the previously found 3-grams from the retained 4-grams;

(4) Identify 5-grams containing the previously found 4-grams from the retained 5-grams.

The 2-grams, 3-grams, 4-grams and 5-grams can be arranged in a hierarchy with different levels. At the top level are the head nouns. Below each head noun are 2-grams, with one word added to the left or the right of the head noun. Below each 2-gram are the 3-grams, with one word added to the right or to the left, and so on.

Step 5: Link *n*-grams (*n*=2, 3, 4, and 5).

The 2-grams, 3-grams, 4-grams and 5-grams were linked together to generate a sketch of a taxonomy (Table 2). The sketch of the taxonomy has a 5-level hierarchy—the first level contains head nouns, the second level contains the 2-grams, the third level contains 3-grams derived from the 2-grams, the fourth level contains the 4-grams derived from the 3-grams, and the fifth level contains the 5-grams derived from the 4-grams. For a high-level gram, if there is no lower-level gram derived from it, the lower level is null. For example, there is no 4-gram and 5-gram derived from the third level 3-gram "ability of some".

Step 6: Select concepts from the *n*-grams.

In this sketch of the taxonomy, there are a large

Table 2 Part of a preliminary taxonomy formed automatically using *n*-grams (*n*=2, 3, 4, 5)

Head noun	2-gram	3-gram	4-gram	5-gram
Ability	Ability of	Ability of some		
		Ability of temporary	The ability of temporary	
		Ability of the	Ability of the organization	The ability of the organization
			The ability of the	The ability of the organization
		Ability of this		
		Ability of traditional		
		Predictive ability of	The predictive ability of	Moderate the predictive ability of
	Language ability	English language ability		
		Language ability and		
	Limited ability			

number of meaningless *n*-grams which cannot be used as concepts. According to the grammatical forms, a concept should be a noun or a noun phrase without verbs, adverbs and initial articles. Thus, an *n*-gram is not a term and is deleted if – it begins or ends with prepositions, articles or auxiliary in a stoplist, such as "the child ability", "child ability to"; it begins with a verb, such as "examine the ability"; it ends with an adjective, such as "child abuse potential"; it ends with an adverb, such as "score be significantly"; it contains a verb in the second location, such as "woman represents a variety"; it contains a finite verb (i.e. not preceded by "to"), such as "partner violence vary", "ethnic identity development be".

However, if an *n*-gram containing an infinite verb, such as "ability to speak English", it is not deleted.

In the above procedures, the verbs, adjectives and adverbs are identified using three previously constructed stoplists of high frequency verbs/adjectives/adverbs (df≥100) which include 180 verbs, 150 adjectives and 25 adverbs. The stoplists are not exhaustive, and thus some *n*-grams containing low frequency verbs/adjectives/adverbs may still not be deleted.

Step 7: Categorize head nouns manually.

On top of the automatically machine-generated five-level hierarchy of concepts, we added a top-most level manually by assigning the concepts into different concept types. Because head nouns represent the broader classes of things or events to which the multi-word terms as a whole refer, we identified the concept types of multi-word terms mainly from the head nouns which they contain.

Some previous studies have done for distinguishing concepts types. National Information Standards Organization (2003) grouped the concepts into the following general types:

(1) Things and their physical parts;

(2) Materials;

(3) Activities or processes;

(4) Events or occurrences;

(5) Properties or states of persons, things, materials, or actions;

(6) Disciplines or subject fields;

(7) Units of measurement.

Medin *et al.*(2000) distinguished concepts according to the structural differences. They are differentiated by:

(1) Nouns vs Verbs;

(2) Count nouns vs Mass Nouns;

(3) Isolated vs Interrelated concepts;

(4) Objects vs Mental events;

(5) Artifacts vs Natural kinds;

(6) Concrete vs Abstract concepts;

(7) Basic level vs Subordinate vs Superordinate concepts;

(8) Hierarchical vs Paradigmatic concepts.

In this study, on the basis of the two kinds of categorization from National Information Standards Organization (2003) and Medin *et al.*(2000), we categorize concepts into the following types:

(1) Concrete objects, such as family, school, center, university, program;

(2) Abstract objects, such as tradition, success, knowledge, justice, culture, life;

(3) Events or occurrences, such as crime, revolution, abuse, violence, conflict;

(4) Actions, such as parenting, participation, action, activity, behavior;

(5) Processes, such as development, change, adjustment, variation, and transition;

(6) Persons, such as student, teacher, child, adult, youth, woman, and man;

(7) Organization, such as institution, community, and organization;

(8) Properties or states of persons, things, materials, or actions, such as gender, and age;

(9) Time, such as year, century, month, period, and decade;

(10) Country or Region, such as United States, America, and California;

(11) Attributes which represent some common measurable aspects or qualities of an entity, action, event, or process, such as size, level, type, form, pattern.

Below the top-most concept level, a second level (i.e. subtype level) was added manually if it is felt that the top-most concept is too broad, or if it will help group head nouns with similar meaning together. For example, under the concept type "person", we added the following subtypes:

(1) General group, such as human, person, people;

(2) Age group, such as youth, adolescent, adult, child;

(3) Gender group, such as man, woman, feminist;

(4) Family group, such as mother, father, couple, parent;

(5) Career group, such as worker, actor, scholar, researcher, staff, student, teacher;

(6) Ethnic group, such as African, Asian, American;

(7) Member group, such as member, partner, peer;

(8) Other group, such as offender, immigrant.

A part of the constructed taxonomy is shown in Fig.2. It contains the concepts which are high frequency nouns or noun phrases in sociology dissertation abstracts. The concepts are organized using a hierarchical structure with 7 levels. The top-most level and the second level are added manually by considering the semantic types and subtypes of the concepts. The other five low levels are generated automatically by considering the grammatical forms of the concepts, i.e. the lower-level concepts (with fewer words) are derived from the high-level concepts (with more words), and are usually the subclasses or facets of the high-level concepts. For example, "young child" is a subclass concept of "child", and "education for young child" is a facet concept of "young child". On the other hand, a facet concept is also a subclass concept of another concept in many cases. For example, "education for young child" is a subclass concept of "education". Therefore, a concept can be assigned into multiple clusters, i.e. it is a subclass concept in one cluster, whereas it is a facet concept in another cluster.

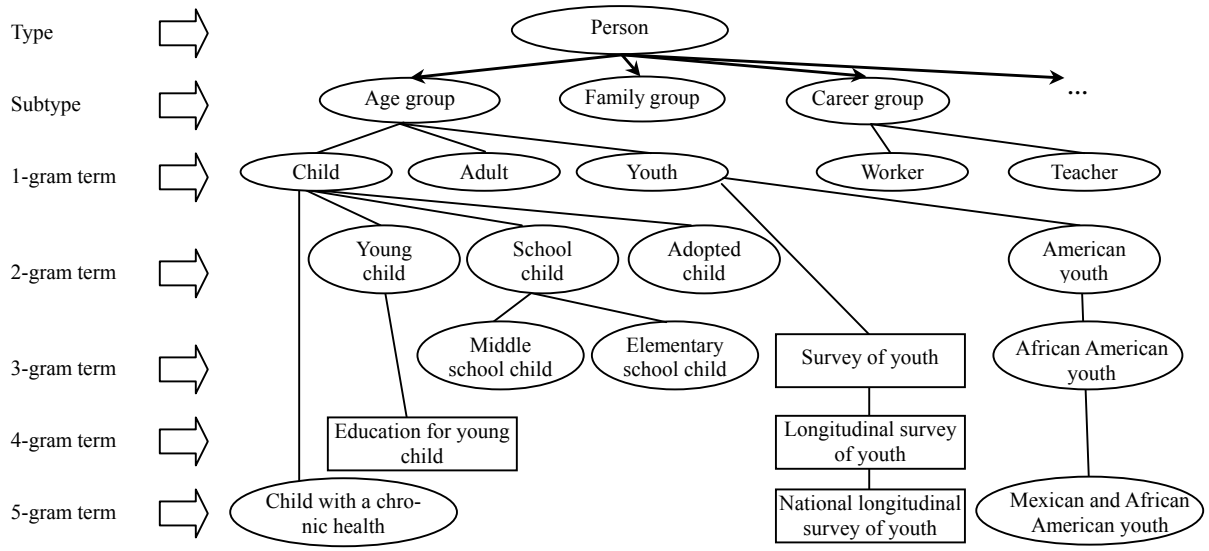

Fig.2 Part of taxonomy (the concepts in round brackets are subclass concepts and those in rectangular brackets are facet concepts)

CONCEPT INTEGRATION USING THE TAXONOMY

This section outlines how the taxonomy was used to integrate extracted information from dissertation abstracts. The dissertation abstracts on four topics from the Dissertation Abstracts International Database were selected to evaluate the effectiveness of the taxonomy. We extracted research concepts from the research objectives and research results sections of the abstracts on each topic automatically using a computer program, segmented the full concepts into shorter terms of different lengths (i.e. 1-word terms, 2-word terms, 3-word terms, 4-word terms and 5-word terms), and then identified the concepts from the terms using the taxonomy. The number of the abstracts and concepts identified from the abstracts on each topic is shown in Table 3.

We used the different level concepts in the taxonomy to integrate the research concepts extracted from the research objectives section of five abstracts on the topic of "teacher training and early childhood". The hierarchy for the integrated concepts related to "teacher" is shown in Fig.3.

We first tried to cluster the concepts using the low-level concepts, if no low-level concept is found, and then clustered them using the higher-level concepts instead. For example, the concept "practicing African American early childhood teacher" and "early childhood teacher input" are clustered using the low-level concept "early childhood teacher", while the concept "classroom teacher" and "teacher's causal belief" are clustered using the broad concept "teacher" directly.

A full concept extracted from a particular dissertation abstract is a single concept or a compound of several component concepts. So a full concept can be assigned into multiple clusters according to its different component concepts. For example, the concept "teacher training" can be clustered into both "person -> teacher-> teacher training" and "action ->training -> teacher training" because it contains two component concepts "teacher' and "training".

In this study, we integrated the concepts only considering its document frequency rather than its context. Actually, a concept in different full concepts

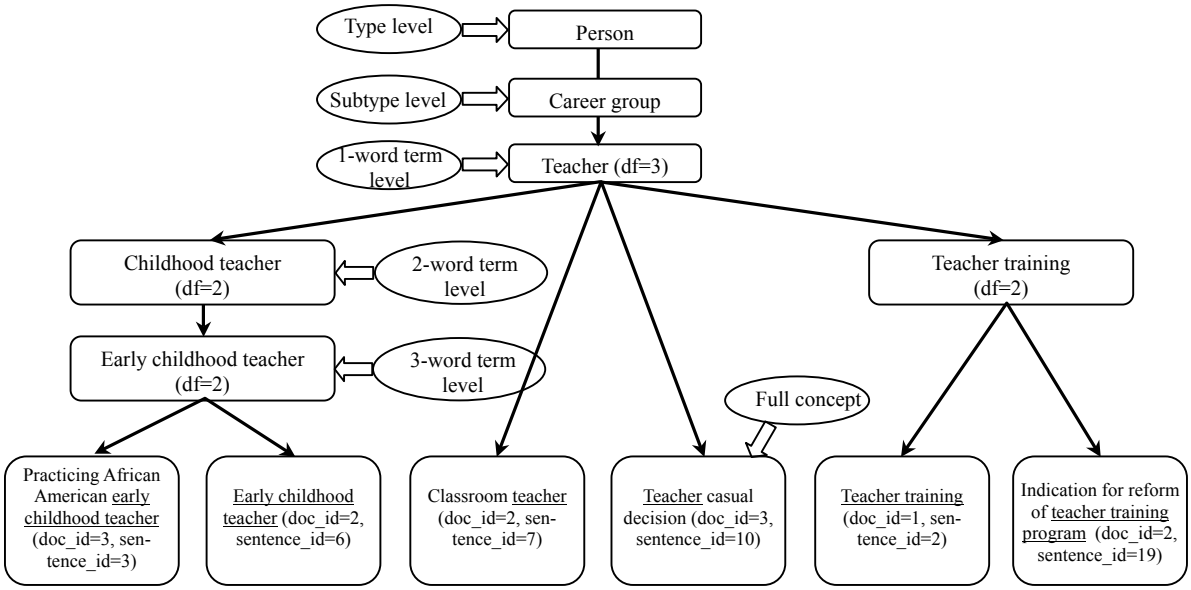

Fig.3 Hierarchy of concepts integrated using different levels of concepts in the taxonomy

Table 3 Number of concepts identified from the dissertation abstracts on each of four topics using the taxonomy

ID	Topic	Number of abstracts	Number of concepts found in the taxonomy	Number of concepts not found in the taxonomy
1	Romantic love	40	745	3392
2	Racial socialization	26	630	2152
3	Bilingual education	56	1009	5421
4	Teacher training and early childhood	5	222	451

has different roles. It can be a main component concept, a qualifier concept, or a human agent/patient of an action. For example, "teacher" is a main component concept in "classroom teacher", a human patient in "practicing African American early childhood teacher", and a qualifier concept in "teacher's causal belief". For a large set of documents, document frequency can reflect the importance of a concept. However, for a specific single document, the role of a concept in a full concept is more important. In future, we will differentiate the roles of the concepts in full concepts and integrate the concepts according to their roles.

CONCLUSION AND FUTURE WORK

This paper describes a new method for automatic construction of multi-document summary of sets of sociology dissertation abstracts. The method uses a hierarchical variable-based framework to integrate four kinds of information—research concepts, relationships between variables, contextual relations, and research methods extracted from different documents, and gives the user a map or overview of a specific topic which the user can explore and zoom in for more details.

A taxonomy is constructed to support the summarization process. It specifies the important concepts in the domain of sociology and the relations between the concepts, and provides a structure for linking similar concepts in different abstracts.

The current taxonomy identifies similar concepts through the grammatical forms of different level concept terms. So the synonymous concepts in the same level cannot be identified using the taxonomy, such as the synonymous terms "woman teacher" and "female teacher", and the synonymous terms "gender inequality" and "gender disparity". In future, we will improve the taxonomy to identify the synonymous concepts in the same level by analyzing the semantic meaning and semantic relations of concept terms. WordNet provides a way to identify the synonyms of single words, and can be used to help identify synonyms of multi-word terms. In addition, the role of a concept in a full concept also needs to be considered so that the similar concepts can be clustered more accurately.

References

Endres-Niggemeyer, B., Hertenstein, B., Villiger, C., Ziegert, C., 2001. Constructing an Ontology for WWW Summarization in Bone Marrow Transplantation (BMT). http://summitbmt.fh-hannover.de/Papers/Washington-Octo11.pdf.

Hovy, E., Lin, C.Y., 1999. Automated Text Summarization in SUMMARIST. *In*: Maybury, M.(Ed.), Advances in Automatic Text Summarization. The MIT Press, p.71-80. http://www.isi.edu/~cyl/papers/ists97.pdf.

Japanainen, P., Jarvinen, T., 1997. A Non-projective Dependency Parser. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP), Washington, DC, p.64-71. http://www.ling.helsinki.fi/~tapanain/dg/doc/anlp97/anlp97.html.

Khoo, C., Ou, S.Y., Goh, D., 2002. A Hierarchical Framework for Multi-document Summarization of Dissertation Abstracts. Proceedings of the 5th International Conference on Asian Digital Libraries, Singapore, p.99-110.

Mani, I., Bloedorn, E., 1999. Summarization similarities and differences among related documents. *Information Retrieval*, **1**(1):1-23.

McKeown, K., Radev, R.D., 1995. Generating Summaries of Multiple News Articles. Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (ACM SIGIR), Seattle, WA, p.74-82.

Medin, D.L., Lynch, E.B., Solomon, K.O., 2000. Are there kinds of concepts? *Annual Review of Psychology*, **51**:149-169.

NISO (National Information Standards Organization), 2003. Guidelines for the Construction, Format, and Management of Monolingual Thesauri. ANSI/NISO Z39.19-1993. NISO Press, Bethesda, Maryland. http://www.niso.org/standards/standard_detail.cfm?std_id=518.

Ou, S.Y., Khoo, C., Goh, D., 2003. Multi-document Summarization of Dissertation Abstracts Using a Variable-based Framework. Proceedings of the 66th Annual Meeting of the American Society for Information Science and Technology, Long Beach, CA, p.230-239.

Ou, S.Y., Khoo C., Goh, D., Heng, H.H., 2004. Discourse Parsing of Sociology Dissertation Abstracts Using Decision Tree Induction. Proceedings of the 14th Annual ASIST SIG CR Workshop, Long Beach, CA.

Radev, R.D., 2000. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure. Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue. http://www.sigdial.org/sigdial-workshop/proceedings/radev.pdf.

Radev, R.D., Jing, H., Budzikowska, M., 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation and User Studies. Workshop Held with Applied Natural Language Processing Conference/Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/ANNCL), p.21-29.

Wollersheim, D., Rahayu, W., 2002. Methodology for Creating a Sample Subset of Dynamic Taxonomy to Use in Navigating Medical Text Databases. Proceedings International Database Engineering and Applications Symposium (IDEAS), Edmonton, Canada, p.276-84.