A FRAMEWORK OF AUTOMATIC SUBJECT TERM ASSIGNMENT:

AN INDEXING CONCEPTION-BASED APPROACH

EunKyung Chung, B.A., M.A., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2006

APPROVED:

Samantha K. Hastings, Major Professor
Shawne Miksa, Co-Major Professor
Rada Mihalcea, Committee Member
Brian O'Connor, Associate Director of the
Interdisciplinary Ph.D. Program
Herman Totten, Dean of the School of Library
and Information Science
Sandra L. Terrell, Dean of the Robert B.
Toulouse School of Graduate Studies

Chung, EunKyung, *A Framework of Automatic Subject Term Assignment: An Indexing Conception-Based Approach*. Doctor of Philosophy (Information Science), December 2006, 115 pp., 26 tables, 22 figures, references, 49 titles.

The purpose of dissertation is to examine whether the understandings of subject indexing processes conducted by human indexers have a positive impact on the effectiveness of automatic subject term assignment through text categorization (TC). More specifically, human indexers' subject indexing approaches or conceptions in conjunction with semantic sources were explored in the context of a typical scientific journal article data set.

Based on the premise that subject indexing approaches or conceptions with semantic sources are important for automatic subject term assignment through TC, this study proposed an indexing conception-based framework. For the purpose of this study, three hypotheses were tested: 1) the effectiveness of semantic sources, 2) the effectiveness of an indexing conception-based framework, and 3) the effectiveness of each of three indexing conception-based approaches (the content-oriented, the document-oriented, and the domain-oriented approaches). The experiments were conducted using a support vector machine implementation in WEKA (Witten, & Frank, 2000).

The experiment results pointed out that cited works, source title, and title were as effective as the full text, while keyword was found more effective than the full text. In addition, the findings showed that an indexing conception-based framework was more effective than the full text. Especially, the content-oriented and the document-oriented indexing approaches were found more effective than the full text. Among three indexing

conception-based approaches, the content-oriented approach and the document-oriented approach were more effective than the domain-oriented approach. In other words, in the context of a typical scientific journal article data set, the objective contents and authors’ intentions were more focused that the possible users’ needs. The research findings of this study support that incorporation of human indexers’ indexing approaches or conception in conjunction with semantic sources has a positive impact on the effectiveness of automatic subject term assignment.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors, Dr. Samantha Hastings and Dr. Shawne Miksa. Dr. Sam Hastings has been my advisor, mentor, and friend and I believe she always will be for me. I was so much inspired from her during classes and research projects. I am truly thankful to Dr. Shawne Miksa for all the guidance and insights she provided for me. My special thanks go to Dr. Rada Mihalcea in the Computer Science and Engineering Department for her constructive comments and feedback. I am thankful to Dr. William Moen for his generous supports through the Texas Center for Digital Knowledge Research Fellowship and particularly with providing extra memory (2G) to my computer for the experiments. I am also grateful to Dr. Brian O'Connor who is incredibly supportive during my entire doctoral program. I am thankful to Dr. Elizabeth Figa for her support and guidance during the Virtual/Digital Storytelling Project.

My special thanks go to my friends and my family. Without their intellectual and emotional supports in every way possible, I could not come this far. Specially, I am grateful to JungWon Yoon who has shared intellectually and personally important moments with me in Denton. I am truly thankful to my sister, Ok-Kyung Chung, brother-in-law, Soo-Jin Lee, my brother, YoungRak Chung, sister-in-law, Eun-Young Choi for their persistent and heartfelt support. With my unfathomable love, I dedicate this dissertation to my parents in the heaven, my mom, Soon-Joo Yoo, and my dad, Jung-Mo Chung.

TABLE OF CONTENTS

Page

Appendices

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

General Background

Subject representation of information entities through the use of subject indexing has been a practice in information organization for centuries. Subject terms or headings serve as subject access points of value to information users when searching information retrieval systems. Subject indexing has been regarded as one of the most critical elements of information organization and access (Taylor, 2003). While extract-based (e.g. keyword-based) subject indexing does not always provide satisfactory subject representation (O'Connor, 1996), subject terms assigned through controlled vocabularies or thesauri provide meaningful subject representation of information and allow for the collocation of information entities by subject within a collection. The organization of information by subject has not only been manifested in the traditional information entities such as monographs and journals, but has also affected digital information entities, especially in networked information environments. For instance, internet sites categorized or classified by subject (e.g. internet search engine directory services) have played a key role in successful information search engines and portal services.

Traditionally, the facilitation of subject access to information has been achieved by human indexers' assignment of subject terms to documents utilizing appropriate controlled vocabularies or thesauri. However, due to the increasing volume of information and the perpetual need to organize and give access to information by subject, there have been numerous endeavors to automatically assign subject terms to the documents by using the full-text of the document. One way to assign subject terms automatically is through the use of Text categorization (TC) using supervised machine learning algorithms. A computer application that is implemented using a machine learning algorithm is likely to predict appropriate subject terms for unknown and new documents after learning the patterns and rules from a training data set with assigned subject terms.

However, as Cunningham, Witten, and Littin (1999) pointed out, the models and properties of TC have been approached without reasonably solid understandings of how human indexers approach to subject indexing. More specifically, research in TC focuses on statistical and probabilistic foundations with respect to document representation, parameter optimizations, and algorithm developments in order to improve effectiveness, rather than basing it on understandings of subject indexing as a conceptual framework. Consequently, there has been little research reflecting the understandings and theoretical backgrounds of subject indexing in the context of TC systems. In fact, with a limited understanding of subject indexing as an underlying framework, the assumption used in most studies is that human indexers simply skim texts and then infer the subject terms from specific patterns (Moens, 2002. p. 111).

Definition of Terms

Subject indexing

Conceptually, subject indexing refers to the activity of representing the intellectual contents of an information entity. In the context of this study, subject indexing is operationalized to represent the *aboutness* of a document by assigning subject terms from pre-coordinated controlled vocabularies or thesauri to information entities. For instance, subject headings or terms are assigned for the collection of most libraries in the United States and Canada using Library Congress Subject Heading (LCSH).

Document

In general, a document denotes an object such as a physical book, printed page(s) or a virtual document in electronic/digital format containing textual information, although it can be manifested in various formats (Buckland, 1997). In the context of this study, an operationalized definition of a document is a journal article in an electronic/digital format.

Automatic subject term assignment

In the context of this study, an operationalized definition of automatic subject term assignment refers to Text categorization (TC) using supervised machine learning techniques (i.e. the machine assigns subject headings or terms, not a human indexer).

Document attribute

A document attribute is defined as a specific part of the document such as title, and references, which are based on the specific characteristics and structures of the type of document.

Semantic source

In the context of this study, semantic sources include specific parts of a full text as well as document attributes. Depending on particular indexing conceptions, semantic sources are defined as sources to which human indexers refer during subject indexing processes and generally refer to document attributes or pieces of bibliographic information such as title, keyword, abstract, etc.

Indexing conception

As an operationalized definition, an indexing conception is defined as an approach, viewpoint, or perception of human indexers concerning the analyses of the subject matters and choice of the subject terms for a document.

Three indexing conceptions

In the context of this study, various indexing conceptions are combined into three approaches: content-oriented, document-oriented, and domain-oriented conceptions. The content-oriented conception refers to the endeavors by indexers to focus on the objective subject matters of the documents. The document-oriented conception emphasizes the focus on reflecting of the author's intentions on the subject matters of the document. Finally, the Domain-Oriented conception focuses on the possible users' needs and requirements by incorporating the contextual information into the subject matters of the documents.

Statement of Problem

Text categorization (TC) using supervised machine learning algorithms is an effective method of automatically assigning subject terms for documents. In general, TC approaches to assigning subject terms focus on the statistical and probabilistic analyses stemming from keyword-based indexing approaches which primarily utilize the full text. However, as Fidel (1994) points out, the conceptions or approaches of subject indexing are more complex and theoretically demanding compared to extract/keyword-based indexing. Despite the fact that subject indexing is complicated and interweaved with various approaches from human indexers, subject indexing, cataloging and classification research is not often consulted in the construction of underlying frameworks for TC systems.

A line of the research in subject indexing demonstrated that indexers have various approaches when indexing documents by subject (Albrechtsen, 1993; Hjørland, 2002; Mai, 2000; Wilson, 1968). For instance, when assigning subject terms to information entities, some indexers may focus on the objective contents, while others may emphasize the author's intentions. Alternatively, they may focus solely on the possible users' needs. Another line of research pointed out that different sets of document attributes are utilized by indexers depending on the nature of indexers' approaches to subject indexing (Foskett, 1996; Hovi, 1988; Jeng, 1996; Mai, 2000;

Miksa, 1983). For instance, while the approaches of human indexers during the process of assigning subject terms for the information entities in conjunction with associated sets of document attributes may have the potential for improving automatic subject term assignment, they have seldom been employed for TC activities.

The purpose of this study is to provide information on the value of human indexers' approaches to subject indexing in terms of improving automatic subject term assignment through text categorization. This purpose is met by creating a conceptual framework for TC that employs the approaches taken by human indexers, when performing subject analysis and indexing, and the utilization of specific document attributes based on the approaches. There are specific semantic sources, such as titles, keywords, and reference lists, to which human indexers refer in order to capture the subject matter of documents. Some combinations of these semantic sources are utilized according to indexers' approaches (Albrechtsen, 1993; Hjørland, 2002; Mai, 2000; Wilson, 1968). While one of the indexing conceptions emphasizes the objective contents of documents, others focus on the author's intentions in creating the documents. In addition, an indexing conception may concentrate on revealing the potential users' needs by reflecting the subject matter of documents within a context (Hjørland, 2002; Hjørland & Albrechtsen, 1995; Mai, 2005).

Research Questions and Hypotheses

Based on the general background and problem area, this study is guided by two research questions, and associated hypotheses, that address automatic subject term assignment through human indexers' indexing approaches in conjunction with semantic sources.

First, this study will investigate the significance and characteristics of semantic sources, or document attributes, in terms of improving the effectiveness of text categorization. A line of research demonstrated some improvement in effectiveness when weights are assigned to specific document attributes for text categorization (Diaz, Ranilla, Montanes, Fernandez, & Comarro, 2004; Efron, Elsas, Marchionini, & Zhang, 2004; Larkey, 1999; Slattery, 2002; Zhang et al., 2004), but failed to explain the results within the context of subject indexing frameworks and conceptual understandings of subject indexing. Accordingly, the first research question and related hypothesis attempt to show the characteristics and importance of semantic sources in the context of subject indexing frameworks for text categorization.

Research Question 1:
Can the use of semantic sources (document attributes or the resources to which human indexers refer during the indexing process) improve the effectiveness of text categorization compared to the full text-based text categorization?

Hypothesis for RQ1:
1. Automatic subject term assignment via semantic sources improve the effectiveness of automatic subject term assignment in terms of the measures such as recall, precision, and F measure.

Secondly, this study investigates the characteristics and importance of an indexing conception-based framework in terms of improving the effectiveness of text categorization. Furthermore, this study explores the effectiveness of three indexing conception-based approaches as compared to the results of the full text based approach. The second question and the related hypotheses examine whether an indexing conception-based framework is effective and whether there is a significant difference among the three approaches.

> Research Question 2:
> Can the indexing conceptions (approaches of human indexers to subject analysis and subject indexing for documents) in conjunction with semantic sources, improve the effectiveness of text categorization compared to the full text-based text categorization?

> Hypotheses for RQ2:
> 1. Automatic subject term assignment via an indexing conception-based framework is more effective than the full text-based approach in terms of the measure such as recall, precision, and F measure.
>
> 2. Automatic subject term assignment via three indexing conception-based approaches (the content-oriented, the document-oriented, and the domain-oriented) improve the effectiveness in terms of recall, precision, and F measure and the improvement rate will differ among the three approaches.

Purpose of the Study

In an attempt to improve the effectiveness of automatic subject term assignment using text categorization, this study proposes a framework based on indexing conception-based approaches. In general, the proposed framework is designed to improve the effectiveness of automatic subject term assignment. By employing the proposed framework, different types of indexing conceptions used by indexers, combined with specific semantic sources, are incorporated into the process of automatic subject term assignment.

This study sets out three objectives: (1) identify semantic sources to which indexers refer during the indexing process, (2) identify the indexing conceptions involved in the indexing processes and relate corresponding semantic sources with the conceptions, and (3) evaluate the effectiveness of indexing conception-based approaches in conjunction with semantic sources compared to the effectiveness of the full text-based approach. The operationalized environment for this study consists of the typical indexing of scientific journal articles through the utilization of a scientific thesaurus.

In order to accomplish the objectives, this study defines a framework based on indexing conceptions in conjunction with semantic sources, and conducts experiments using the framework within the operationalized context. First, this study will gather and synthesize subject indexing, cataloging, and classification studies that demonstrate the utilization of semantic sources, indexing conceptions, and the relationships between these sources and conceptions. Secondly, by conducting experiments, this study will examine whether the proposed framework can improve the effectiveness on automatic subject term assignment using text categorization compared to a full-text based approach.

Significance of the Study

Automatic subject term assignment has received increased attention within the context of the volume of published information entities and the need for organization of these objects by subject. Despite the fact that subject indexing practices are complex and theoretically demanding compared to keyword-based indexing, there is little research reflecting conceptual understandings and theoretical backgrounds of subject indexing on text categorization systems. Given the importance of automatic subject term assignment through text categorization techniques in this digital age, this study is significant because it is among only a handful of studies (Diaz, Ranilla, Montanes, Fernandez, & Comarro, 2004; Efron, Elsas, Marchionini, & Zhang, 2004; Larkey, 1999; Slattery, 2002; Zhang et al., 2004) that incorporate an understanding of the characteristics and structures of information entities into the design of text categorization systems. Furthermore, this study appears to be the first to examine theoretical and conceptual understandings of subject indexing as a framework in order to improve the effectiveness of text categorization.

From the theoretical perspective, this study examines an indexing conception-based framework that reflects conceptual understandings of subject indexing practices. From the practical perspective, the framework may be used to design text categorization systems that utilize indexing conceptions with semantic sources. It is possible to customize text categorization systems depending on the focal point of various indexing conceptions of collections or databases. The results and findings of this study will serve as an underpinning framework for many application areas such as automatic metadata generation in digital libraries, automatic information organization, information filtering, genre classification, and recommending systems (Sebastiani, 2005).

In addition, this study is significant because it provides data sets that are of value to the text categorization/classification research communities. These communities have demanded good quality data sets for various experiments for quality text categorization research and implementations (Lewis, 2000).

Scope and Limitation

The scope of this study lies in determining whether a framework of indexing conception-based approaches in conjunction with semantic sources has a positive impact in order to improve the effectiveness of automatic subject term assignment. In order to investigate the effectiveness of the experiments, the evaluations are restricted only to three measures which are primarily used in text categorization communities: recall, precision, and F-measure.

One of the major limitations of this study is that the data set is not a standard test set such as Reuters, Ohsumed, or 20Newsgroups from the Text Retrieval Conference (TREC) document sets. In light of this limitation, the results of this study cannot be compared to related research results and may not be easily generalized. However, two factors which reconcile this limitation can be considered. First, taking Lewis (2000) plea for better test collections for text categorization research in consideration, this study constructs a new data set for text categorization with respect to the purpose of this study. The constructed data set is expected to investigate the problem areas of text categorization in a more realistic environment. Secondly, in order to generalize the results of this study as much as possible, the data set is designed to resemble various cases of data environments such as more semantically related and less semantically related data sets. More semantically related data set contains homogeneous subject terms, for instance, from the same top term, while less semantically related data set includes heterogeneous subject terms.

Summary

This chapter provided the general and theoretical background as well as the problem area of this study. Based on the background and research problem area, two research questions and three related hypotheses were stated. The research questions and associated hypotheses address whether an indexing conception-based framework in conjunction with corresponding semantic sources may improve the effectiveness of automatic subject term assignment. Additionally, the main purposes, significance, definition of terms, and scope and limitation of this study were outlined.

CHAPTER 2

REVIEW OF LITERATURE

Introduction

This chapter presents a review of the literature in two subject areas relevant to this study: text categorization and subject indexing. More specifically, this chapter explores text categorization literature focusing on supervised machine learning techniques for automatic subject term assignment and explores the literature surrounding the theoretical framework for subject indexing. As one of the supervised machine learning techniques, the Support Vector Machine algorithm is described with respect to its geometrical and theoretical definitions. The review of the theoretical framework literature for subject indexing is divided into three sections: Semantic Sources for Subject Indexing, Conceptions of Subject Indexing, and Relationships between Conceptions and Semantic Sources. Finally, a preliminary framework of subject indexing for automatic subject term assignment is proposed in the context of a set of typical scientific journal articles. The detailed reviews of both bodies of literature are gathered and combined to support the basic premises of this study.

Automatic Subject Term Assignment

Text categorization, also called text classification or topic selecting, explores typical text patterns and characteristics in conjunction with assigned subject terms and then builds a classifier for unknown documents (Lewis, 1992). Since text categorization in general utilizes prior human knowledge of subject terms assigned to a certain set of documents (Lewis, 2000), it is well suited to the problem of automatic assignment of subject terms to documents. In this sense, automatic subject term assignment employs supervised learning algorithms which exploit given subject terms and a set of documents in order to assign terms to new and unknown documents. A typical text categorization procedure consists of training phase and testing phase. In the training phase, the learner, through an inductive process, observes the patterns and characteristics of documents with pre-assigned subject terms. By observing and identifying these patterns in a set of documents with specific subject terms, the learner is able to build a classifier. Then, in the testing phase, the classifier is able to predict subject terms for unknown documents. Various learning algorithms, including neural networks, naïve bayes, support vector machine (SVM), and *k*-nearest neighbors (kNN), have been used in text categorization applications. SVM was introduced by Joachims (1998) and subsequently used in many text categorization problems (Sebastiani, 2005).

In terms of geometry, the goal of SVM is to maximize the margin between positive examples and negative examples by identifying support vectors in each example as shown in Figure 1. The hyperplane (a solid line in Figure 1) separates the positives from the negatives by the widest possible margin and results in minimizing the generalization errors, i.e., the error of the resulting classifier on unknown testing sets. Consistent with the geometrical terms of the SVM algorithm, the theoretical definitions of SVM begin from computational learning theory. The aim of SVM is to find a hypothesis that guarantees the lowest true error, i.e., where the true error is the probability that the hypothesis will make an error on an unseen and randomly selected text example (Joachims, 1998).

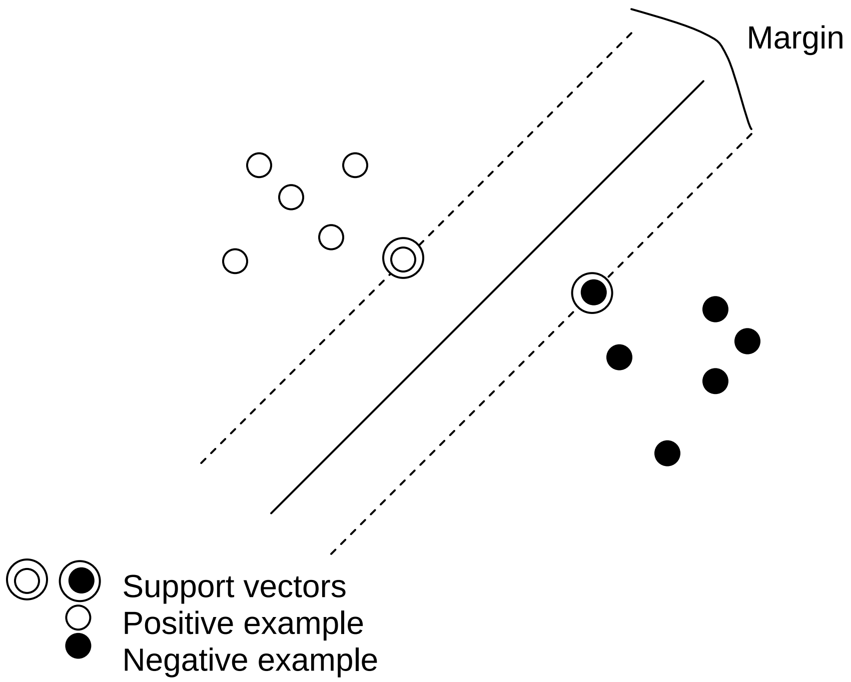


Figure 1. Visualization of support vector machine (SVM) algorithm.

In general, the research involving automatic subject term assignment using text categorization techniques falls into two categories: generalized approaches using the full text and document-sensitive approaches utilizing specific document attributes. The majority of text categorization research deals with feature selection, document reduction, optimization of specific collections, and effective learning algorithm development from the perspective of generalized approaches (Cunningham, Witten, & Littin, 1999; Sebastiani, 2002; 2005). However, document-sensitive approaches are emerging as a line of research which reflects an understanding of the importance of the characteristics and structures of the documents in the development of text categorization systems.

These document-sensitive approaches are beginning to demonstrate the significance of the various attributes of documents for text categorization systems, instead of focusing on generalized statistical or probabilistic approaches using the full text.

The document-sensitive approaches to text categorization are more relevant to this study, since these approaches utilize the significance of document attributes to improve the effectiveness (Diaz, Ranilla, Montanes, Fernandez, and Combarro, 2004; Efron, Elsas, Marchionini, and Zhang, 2004; Larkey, 1999; Slattery, 2002; Zhang et al., 2004). In assigning subject terms to patent documents, Larkey (1999) took into account the significance of document attributes and demonstrated the improvement of effectiveness when using the *k*-Nearest Neighbor algorithm. Larkey supported that document attributes such as title, abstract, and the first twenty lines of text characterize the vectors with the best effectiveness and that text categorization results were more effective using these attributes than when using the full text of the documents. In terms of the accuracy measure (described in Chapter 3, Effectiveness Evaluation Section), Larkey reported approximately 31% accuracy, even though there is a small training data set. In a similar study, Efron, Elsas, Marchionini, and Zhang (2004) demonstrated that when clustering government documents by subject headings, document attributes such as keyword and title were more effective than the full texts: their results showed 73% accuracy in effectiveness when using SVM. In addition to incorporating document attributes into text categorization systems, Zhang et al. (2004) included citation information in order to discover the most similar documents using the *k*-Nearest Neighbor algorithm. In general, the k-Nearest-Neighbor algorithm assigns a class to a document by computing a distance (similarity measure) between an unknown document and a corpus of documents assigned a set of subject terms. The researchers concluded that the combination of title, abstract, and citation information, led to the best results when discovering similar documents and consequently performed well (60.81%) in a test of effectiveness as F measure (described in Chapter 3, Effectiveness Evaluation Section). Consistent with Zhang et al.'s results, Slattery (2002), using SVM, identified that hyperlink patterns in hypertext documents have a positive impact on the effectiveness of text categorization. In a more sophisticated approach to subject term assignment, Diaz, Ranilla, Montanes, Fernandez, and Combarro (2004) demonstrated that integrating the contextual information represented by the term localities showed improved effectiveness in text categorization. When selecting features for document representation, they compared the effectiveness of local terms with global terms. While local terms refer to the words occurring in documents assigned by specific subject terms, global terms consist of words occurring across all the documents. The results using local terms showed greater effectiveness than the results using global terms. This study clearly showed that a narrowly defined context for a set of documents can more precisely represent the subject matters of a set of documents than a broadly defined context.

While generalized approaches using the full text of documents still makes up the majority of research in text categorization, document-sensitive approaches are emerging that incorporate the significance of document characteristics and structures into text categorization systems. Yet, the awareness of the importance of document characteristics and structures reflected in current text categorization systems has been limited to only a select group of document attributes and a limited degree of contextual information. In fact, this limitation is due to the lack of a underpinning framework and an understanding of the conceptual subject indexing process conducted by human indexers.

Semantic Sources for Subject Indexing

From the perspective of the subject term assignment process by human indexers, there are several attributes of a document which could be incorporated into the learning process of text categorization systems. Semantic sources are defined as document attributes or a set of document attributes to which human indexers refer in order to analyze the *aboutness*, or intellectual content, of a document. In general, there are three types of literature which discuss the identification of semantic sources for subject indexing: subject indexing schemes and guidelines, textbooks for subject indexing or cataloging, and empirical or theoretical studies undertaken to understand the process of subject indexing.

First, subject indexing schemes and guidelines recommend some attributes of a document to use for subject analysis. As Mai (2000) pointed out, the introduction to *Dewey Decimal Classification and Relative Index* (Dewey, 2003) states some attributes of a document for subject determination; title, table of contents, chapter headings, preface, introduction, foreword, book jacket, accompanying materials, the text itself, bibliographic references, index entries, cataloging-in-publication data, and reviews. Similar to the guidelines provided by the DDC introduction, the ISO standard (ISO 5963:1985, 1985) affirms many of the same semantic sources for subject analysis: title, abstract, list of contents, introduction, illustrations/diagrams/tables with their captions, and words in unusual typeface.

Secondly, textbooks on subject indexing and subject cataloging denote semantic sources as well. Originally, Chan (1987) surveyed instructional materials for subject indexing or cataloging. Sauperl (2002) updated new versions of instructional materials since Chan’s survey including Chan (1981), Foskett (1996), and Taylor (2003). Foskett pointed out title, keyword, and citation as subject access points. In addition, Taylor recommended title/subtitle, table of contents, introduction, index terms/words/phrases, and illustrations/diagrams/tables/captions for subject analysis. More importantly, Chan emphasized utilizing the attributes of a document, rather than the full text for subject analysis. These attributes include title, abstract, table of contents, chapter headings, preface, introduction, book jacket, slipcase, and other accompanying descriptive materials. For external sources, Chan recommended bibliographies, catalogs, review media, and other reference sources.

Table 1

Semantic Sources Recommended in Guidelines and Textbooks for Subject Indexing and Cataloging

Type	DDC (2003)	ISO 5963: 1985 (1985)	Chan (1981)	Foskett (1996)	Taylor (2003)
Bibliographic Information	- title	-title	- title	- title	-title/subtitle
				- keyword	
	-cataloging-in-publication				
Entity Information		- abstract	- abstract		
	- table of contents - chapter headings	- list of contents	- table of contents - chapter headings		-table of contents
	- preface - introduction - forward	- introduction	- preface - introduction		-introduction
	-the text itself -index entries				-index terms
		- illustration, diagram, table with their captions - words in unusual typeface			-illustration, diagram, table with their captions
	-book jacket		- book jacket - slip case		
	-accompanying materials		- accompanying materials		
Contextual Information	-reviews		- catalogs - other reference sources - reviews		
	-bibliographic references		- bibliographies	-citation	

As shown in Table 1, these suggested and recommended semantic sources for subject indexing are various depending on the document types and situations. Given the diversity of the attributes, the semantic sources are divided into three categories: bibliographic information, entity information, and contextual information. The bibliographic information represents information needed to locate an item including title/subtitle, keywords, and cataloging-in-publication data. The entity information refers to the information dealing with the contents of the entire document or parts of the document. It can be the text itself, the summarization of the document (i.e. abstract), or certain parts of the document such as introduction, preface, forward, index entries, table of contents, chapter headings/subheadings, or book jacket/slip case. On the other hand, the contextual information refers to indirect and surrounding information of value in terms of subject indexing. For the contextual information, Table 1 indicates bibliographic references/citation, accompanying materials, catalogs, and reviews.

Thirdly, as shown in Table 2, a line of research indicates that indexers and subject catalogers utilize specific semantic sources for subject indexing or subject cataloging in practices (Chu & O’Brien, 1993; Jeng, 1996; Sauperl, 2002; 2004). In order to understand and describe the process used by the human indexers and subject catalogers, researchers used primarily qualitative case study methodologies. Jeng (1996) demonstrated that subject catalogers approach the subject matter of a document through networking and association techniques. In terms of subject cataloging in practice, Jeng identified that catalogers tend to network and associate semantic sources such as bibliographic information with corresponding subject terms. On the other hand, Sauperl synthesized a hypothetical subject cataloger based on the results of the case study of twelve expert catalogers in practice. The hypothetical cataloger is likely to examine semantic sources such as title, author's name, publisher's name, and author's affiliation for subject analysis. Similarly, Chu and O'Brien pointed out the importance of semantic sources such as title, subtitle, abstract, paragraph headings, and initial paragraphs for determination of subject matter in the process of subject indexing.

Table 2 indicates three types of information that are consistent with the semantic sources recommended by subject indexing and cataloging textbooks and guidelines: bibliographic, entity, and contextual information. While utilization of semantic sources for the bibliographic information type is affirmed by case studies, semantic sources both in entity and contextual information types is relatively less focused.

Table 2

Semantic Sources Identified in Case Studies

Type	Jeng (1996)	Sauperl (2002; 2004)	Chu & O'Brien (1993)
Bibliographic Information	Bibliographic information	Title	Title, subtitle
Entity Information			Abstract, headings, initial paragraphs
Contextual Information		Author's name, publisher's name, author's affiliations	

Conceptions of Subject Indexing

The conceptions of subject indexing are defined as the indexers' perceptions, viewpoints, or approaches with regard to subject analysis, determination and indexing. These conceptions of subject indexing have been recognized from various perspectives (Albrechtsen, 1993; Fidel, 1994; Hjørland, 1992; Mai, 2000; Soergel, 1985; Wilson, 1968). More specifically, as shown in Figure 2, the conceptions of subject indexing have been identified from three perspectives: *dual*, *detailed*, and *convergent*.

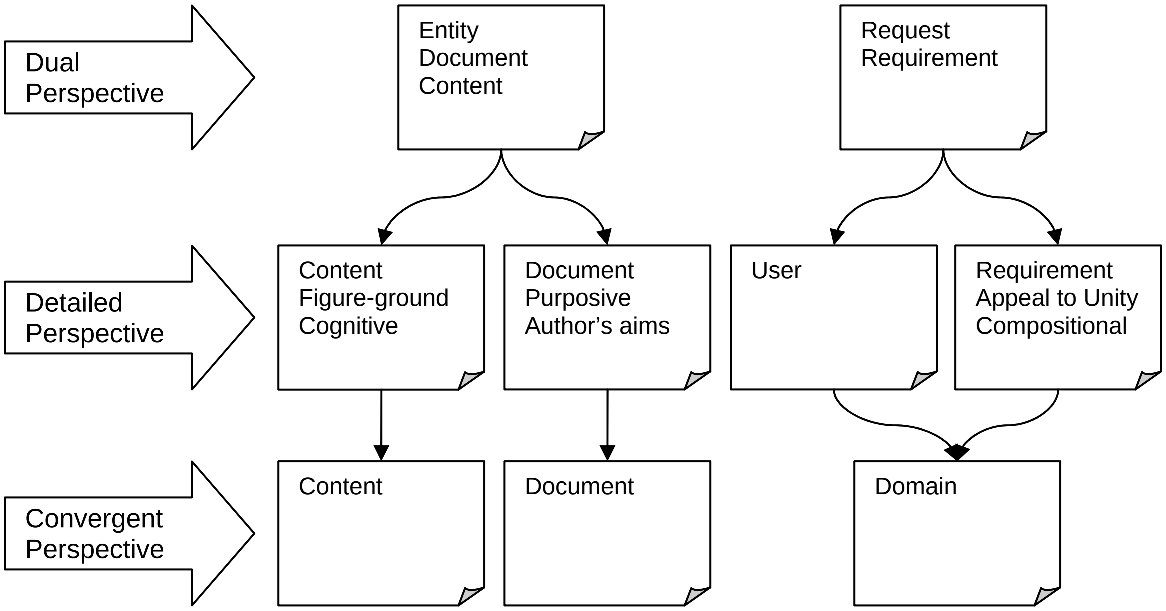


Figure 2. Subject indexing conceptions from dual, detailed, and convergent perspectives.

In the dual perspective, the conceptions of subject indexing have been divided into two approaches: the entity/document-oriented approach and the user/requirement-oriented approach (Albrechtsen, 1992, 1993; Fidel, 1994; Soergel, 1985). The primary discussion in the literature centers on the differences between these two approaches. In general, while the entity/document-oriented approach focuses on the document itself, the user/requirement approach centers on the needs of users. Fidel (1994) specified that while the document-oriented approach focuses on the objectivity of the subject matter of a document, the request-oriented approach is mostly related to users’ needs and prior indexing before users actually use the document. Albrechtsen (1992; 1993), in addition to identifying the simplistic conception of subject indexing represented by keyword-based indexing, she identified content-oriented and requirement-oriented conceptions providing implicit information to users through subject term assignment. Consistent with Fidel’s argument, Albrechtsen specified that content-oriented conception is related to the abstraction of the objective contents of a document, while the requirement-oriented conception focuses on mediating and rendering the information visible to possible users. In line with Fidel and Albrechtsen’s research, Soergel (1985) discussed entity-oriented indexing and request-oriented indexing. He posited that entity-oriented indexing focuses on the representative contents of the documents, while request-oriented indexing focuses on incorporating possible users’ queries into the subject indexing terms. In a sense, entity-oriented indexing is similar with pre-coordinated indexing, while request-oriented indexing is compatible with post-coordinated indexing.

By contrast, the detailed perspective as shown in Figure 2 provides divergent conceptions of subject indexing by examining more closely the two conceptions of the dual perspective. This perspective presents more detailed and divergent conceptions of subject indexing (Hjørland, 2001; Mai, 2000; Wilson, 1968). Wilson identified the four methods of subject analysis as purposive, figure-ground, objective, and the appeal to unity or the rule of selection/rejection. The purposive method is a system in which indexers try to grasp the author’s objects, aims, or purposes. The figure-ground method centers on the indexers’ interpretation of a central figure, or group of central figures in a document. The Objective method is a system of simply counting references to an element, or a group of elements, in a document. The appeal to unity or the rule of selection/rejection method is related to the formation of a set of rules for selection or rejection based on the subject of a document. Hjørland (2001) further argued for these four methods in terms of theories of meanings, interpretation, and epistemology. The purposive method is explained as the connection with hermeneutics theory primarily analyzing documents by studying the author's intentions, personality, and biography. The Figure-ground method is related to psychological and cognitive approaches. The Objective method utilizes positivistic, bibliometric, and statistical ways of analyzing documents. The Appeal to unity method is interpreted using text linguistics and compositional methods for analyzing documents (Hjørland, 2001). Based on the epistemological positions of these different conceptions, Mai (2000) pointed out five conceptions of subject indexing including document-oriented, content-oriented, user-oriented, and requirement-oriented as well as an automatic keyword-based approach as the simplistic conception. While the document-oriented conception centers on the information presented in the document, the content-oriented conception is an objectivist conception and, in the extreme, would support the claim that there is only one correct analysis of a given document. On the other hand, Mai noted that while the user-oriented and the requirement-oriented conceptions both focus on users, both conceptions base the subject analysis on the future potential use of the documents. More importantly, by utilizing the fact that both user-oriented and requirement-oriented conceptions have a common focus on the potential future use of the document, the convergent perspective can emerge from four detailed conceptions.

From the convergent perspective, shown in Figure 2, three conceptions such as content-oriented, document-oriented, and domain-oriented are identified in the context of this study. These three conceptions are considered to be of value in terms of the relationship between semantic sources and indexing conceptions. Although the distinctions between different conceptions are not entirely exclusive, the differences in corresponding semantic sources used are considered critical factors. That is, with the purpose of identifying the associated semantic sources for the three indexing conceptions, three convergent conceptions are more appropriate with respect to identifying clear relationships, instead of two and four conceptions from the dual and detailed perspectives, respectively. For instance, while four conceptions from the detailed perspective are likely to contain the same semantic sources in different indexing conceptions, two conceptions from the dual perspective do not provide enough discriminating contrast to relate the document attributes to appropriate indexing conceptions.

First, the content-oriented conception of subject indexing indicates that subject indexing focuses on the objectivity of subject matters in terms of a prevailing element or a group of prevailing elements from the document. Albrechtsen (1993) recognized the “content-oriented conception” (p.220) as the practice of assigning objective subject terms to a document implied by the human indexer’s interpretations of a principal subject element, instead of automatically extracting keywords from a document. In fact, the content-oriented conception lies on the boundary between keyword-based subject indexing and term assignment-based subject indexing. In Wilson’s (1968) words, this conception denotes the determination of subject matters based on the indexer’s interpretation of the “figure-ground” (p.81) of a document. The “figure” refers to the relative dominance and subordination of various elements in the document. Since not all elements in a document demonstrate the same amount of weight to readers, Wilson posited that there exists a main element or a group of elements. Hence, an element or a group of elements can represent the subjects of a document. This conception is related to the cognitive approaches of analyzing documents by weighing the relative dominance and subordination of different elements revealed by reading the document (Hjørland, 2001). In essence, the content-oriented conception of subject indexing is an attempt to present the objective subject matter of a document by identifying a dominant element or a group of elements in terms of indexers' interpretations.

Secondly, the document-oriented conception endeavors to emphasize the intentions of authors in subject indexing. Just as Wilson (1968) recognized it as "the purposive way" (p.78), this conception is based on the approach that the authors' intentions for a document are the subject matter for a document. Hjørland (2001) argued that this approach is connected with the theory of classical hermeneutics which primarily analyzes the document by studying the author's intentions, personality and biography.

The Domain-oriented conception for subject indexing takes into account the domain of knowledge surrounding a document when representing users' possible needs and requirements. While the content-oriented and the document-oriented approaches view a document as an isolated-entity (Soergel, 1985), the domain-oriented approach incorporates the surrounding and connected information of a specific document into subject indexing, and focuses on users' possible needs, requests, and requirements. In order to achieve the purpose of Domain-oriented indexing conception, Mai (2005), Hjørland and Albrechtsen (1995) implied that subject indexing compromises the discourse of a specific document within a context. In this sense, the discourse between users and authors in a context can represent the domain of a document; then, subject indexing is able to anticipate the impact and value of a particular document for potential use, instead of exclusively focusing on the contents of documents (Blair, 1990; Hjørland, 1992; Soergel, 1985; Weinberg, 1988).

Relationships between Conceptions and Semantic Sources

In the context of this study, which utilizes conceptions of subject indexing in conjunction with corresponding semantic sources for automatic subject term assignment, it is important to reveal the relationships between the three identified conceptions and the corresponding semantic sources. Through empirical studies of subject indexing processes undertaken by human indexers, two premises have been identified for indicating the relationships between semantic sources and conceptions of subject indexing. One premise is that subject indexers can agree on the subject matter of a document. The second premise is that human indexers utilize semantic sources for subject indexing. First, arguably, human indexers can agree on the subject matter of a document, despite the fact that there are widely recognized inconsistencies between indexers. Hovi (1988) demonstrated that indexers and subject catalogers are generally unanimous about the subject matter of a document, but there may be differences in representations with respect to subject terms chosen from different controlled vocabularies or thesauri. Hovi’s study indicated that agreed-upon subject matter for a document exists among indexers, despite the differing representations of these documents due to different indexing schemes. For the second premise, as described Semantic Source for Subject Indexing section of Chapter 2 and based on the investigations of subject indexing guidelines and schemes, textbooks on subject indexing, and case studies on subject indexing processes, semantic sources are shown to be critical elements utilized by human indexers for subject indexing.

By accepting the two premises suggested by empirical and case studies and by investigating various literature sources, Table 3 presents archetypical semantic sources, provided by Mai (2005), with corresponding types of information and conception areas.

Table 3

Semantic Sources as Archetype and Corresponding Types of Information and Conceptions

Type	Archetype	Conception area
Bibliographic information	Title	Document-oriented
	Abstract	Content-oriented
Entity Information	The full text itself	Content-oriented
	Index entries	Content-oriented
	Table of contents	Content-oriented
	Chapter headings	Content-oriented
	Chapter subheadings	Content-oriented
	Illustration/diagram captions	Content-oriented
	Tables and captions	Content-oriented
	Introduction	Document-oriented
	Forward	Document-oriented
	Preface	Document-oriented
Contextual Information	References	Domain-oriented

While the contextual information corresponds to the domain-oriented approach, document attributes from the bibliographic and entity information are divided into the content-oriented and the document-oriented conceptions, as shown in Figure 3.

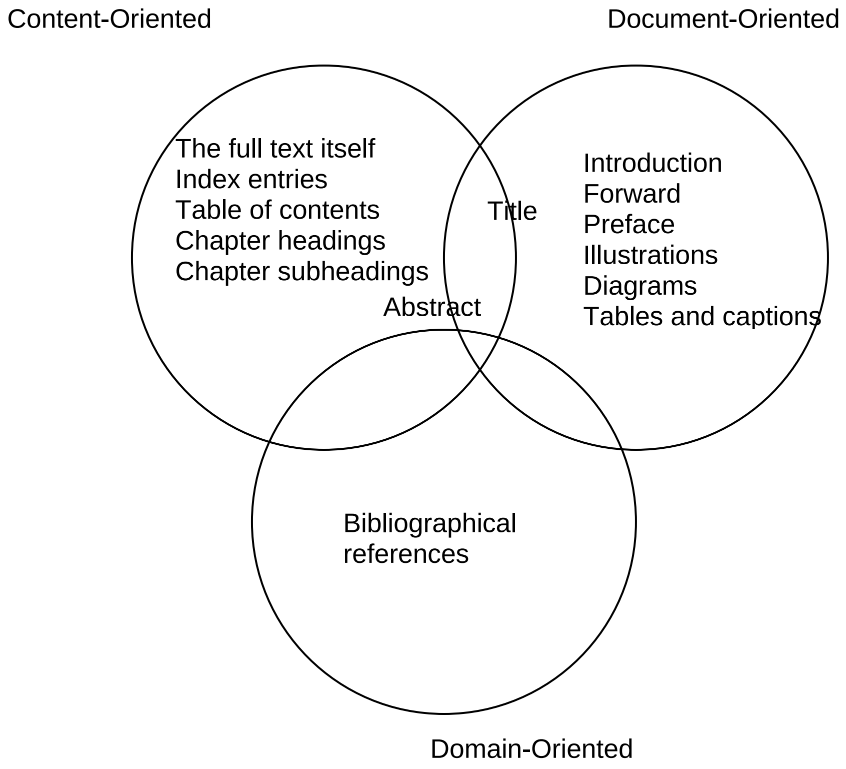


Figure 3. Semantic sources and corresponding conception areas.

Figure 3 presents the content-oriented conception area with a set of document attributes such as the full text, index entries, table of contents, chapter headings, chapter subheadings, and abstract based on the archetypical document attributes. Because the fundamental and common function of these document attributes is to represent the content of the document objectively or represent the document itself, it is rational to consider this set of document attributes as the semantic sources for the content-oriented conception. Among these semantic sources, the abstract is ambiguous as a semantic source and intersected by the content-oriented and the document-oriented conceptions in *Figure 3*, because it is likely to be influenced by author's or authors' intentions, especially if the abstract is provided by the author(s). However, as Hjørland and Nielsen (2001) point out, that because an abstract generally denotes a concise version of the full text and is more likely to contain the objective content of the document, it is reasonable to include it in the set of attributes for the content-oriented conception.

Secondly, the document-oriented conception area contains another set of document attributes such as title, introduction, forward, preface, illustrations, diagrams, and tables with their captions. Mai (2000) identified that indexers are supposed to look for clues from the 'introduction', 'forward', and 'preface' parts of the document in order to identify the author's purpose. In addition, Wilson (1968) pointed out that illustrations, diagrams and tables with their captions represent the author's or authors' intentions or purposes. In this sense, the introduction, forward, preface, illustrations, diagrams, and captioned tables are considered semantic sources for the document-oriented conception. However, as with the abstract in the content-oriented conception area, title could be considered for both conceptions and is intersected by the content-oriented and the document-oriented conceptions. Even though the title shares the function of providing a relevant description of the document being represented, it does contain the intentions of the author(s) when choosing from among many possible alternatives and therefore can be considered biased based on the author's intentions (Hjørland & Nielsen, 2001).

Thirdly, in Figure 3, the Domain-oriented conception area presents the contextual information, including bibliographical references. The reference list of a document is a considerable source for representing the domain knowledge of the document being represented (Hjørland, 2002; Mai, 2000; Sauperl, 2004). A bibliometric approach, one of eleven approaches to domain analysis provided by Hjørland (2002), is considered one way to embrace the discourse surrounding a specific document. Specifically, the contextual information of a document, such as references, represents the potential future use of the documents (Mai, 2000). On the other hand, from Sauperl (2004)'s perspective of subject indexing practices, the indexers consider references as a possible source of the potential users' needs. In essence, while the contextual information of a document has been identified as a representation of domain knowledge and the discourse between users and author(s), it is also considered a source for representing possible users' needs with respect to the Domain-oriented conception.

A Preliminary Framework Applied in Typical Scientific Journal Articles

A preliminary framework is proposed within the context of typical scientific journal articles and utilizes indexing conceptions with associated semantic sources for automatic subject term assignment. The identified semantic sources and three conceptions in subject indexing are employed in a data set of typical scientific journal articles. Typically, a scientific journal article includes bibliographical identification (journal name, volume, pages), title, author(s), corporate affiliation and address, author abstract, author keywords, introduction, apparatus/materials/method/results/discussion, conclusion, acknowledgement, and references (Hjørland & Nielsen, 2001). Among those elements, a typical article presents six attributes relevant to the subject matter of a document: title, abstract, keyword, source title (e.g. journal title or conference proceeding title), full text, and references.

Among these six attributes, the full text is utilized in its entirety and simultaneously, partially to stress one of the conceptions used, i.e., the full text itself, introduction, and conclusion. On the other hand, references cited in the article contain information such as author, title, year, source, publisher, etc. Since this study focuses on semantic information sources from reference list rather than from citation analysis of cited and citing articles, the titles of cited works are considered sufficient. Therefore, six attributes in a typical scientific journal article become eight semantic sources for subject term assignment: title, abstract, keywords, source title, full text, introduction, conclusion, and titles of cited works.

Applied to a data set of typical scientific journal articles, eight semantic sources are embraced in Figure 4 and present a framework with respect to the three conceptions and the corresponding semantic sources for automatic subject term assignment. Although the three conceptions combined with the semantic sources are not completely distinct, the separation can indicate a way of demonstrating effective approaches to subject indexing with respect to text categorization.

First, in order to obtain the objective subject matters of a document, the content-oriented approach considers abstract, conclusion and full text. However, Figure 4 shows that two categories of semantic sources are distinguished; full text and document attributes. This distinction is necessary to see if there are differences between the content-oriented approach using the full text and using the document attributes because most text categorization systems focus on the full text alone when assigning subject terms for the documents. As document attributes, abstract and conclusion are considered the semantic sources for the content-oriented conception. The conclusion of the full text tends to be a recapitulation of it. Therefore, it is reasonable to assume that the common characteristics of an abstract and a conclusion are the objective description of the contents of a specific document.

Secondly, semantic sources such as keywords, title and the introduction are considered because the document-oriented approach is mainly concerned with the intentions of the author. In general, an important source in scientific journals that reflect the author's intentions are the keywords they provide when they submit the final draft of the article for publication.

Thirdly, the Domain-Oriented approach utilizes source title and title of cited works for subject indexing. This is one way of incorporating the discourse of a document within a context when then makes the document available for possible users' needs. Therefore, source title and titles of cited works are implied as semantic sources emphasizing the Domain-Oriented approach.

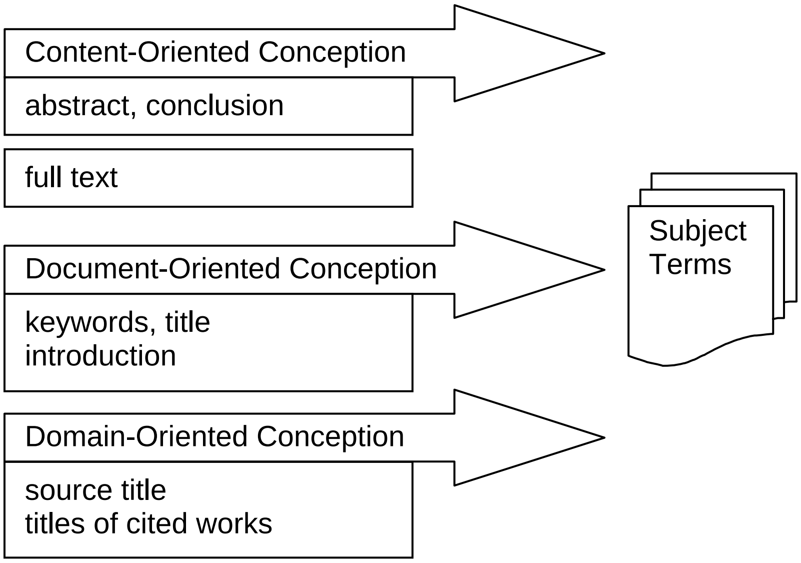


Figure 4. A preliminary framework of conception-based approaches applied in typical scientific journal articles.

Summary

This chapter presented a synthesized and critical review of the literature and a preliminary framework for investigating the impact of conceptions of subject indexing in conjunction with semantic sources for automatic subject term assignment using text categorization techniques. The literature review was organized into five sections. The first section discussed text categorization techniques in general and specifically the Support Vector Machine (SVM) algorithm. In particular, the literature relevant to automatic subject term assignment with document-sensitive approaches was reviewed and criticized. The second section presented semantic sources, document attributes to which human indexers refer in order to analyze the subject matter of a document. The types of literature mentioned in this section were subject indexing schemes/guidelines, textbooks on subject indexing and empirical or case studies of subject indexing. The third section discussed the conceptions of subject indexing with a detailed review of the literature. While the section presented various perspectives, or conceptions, of subject indexing, three conceptions were deemed relevant for this study: content-oriented, document-oriented, and domain-oriented. It was stated that semantic sources will be combined with the conceptions to improve the effectiveness of text categorization. The fourth section illustrated the relationships between the conception of subject indexing and the corresponding semantic sources. The last section presented a preliminary model of a conception-based approach to automatic subject term assignment utilizing typical scientific journal articles.

CHAPTER 3

METHODOLOGY

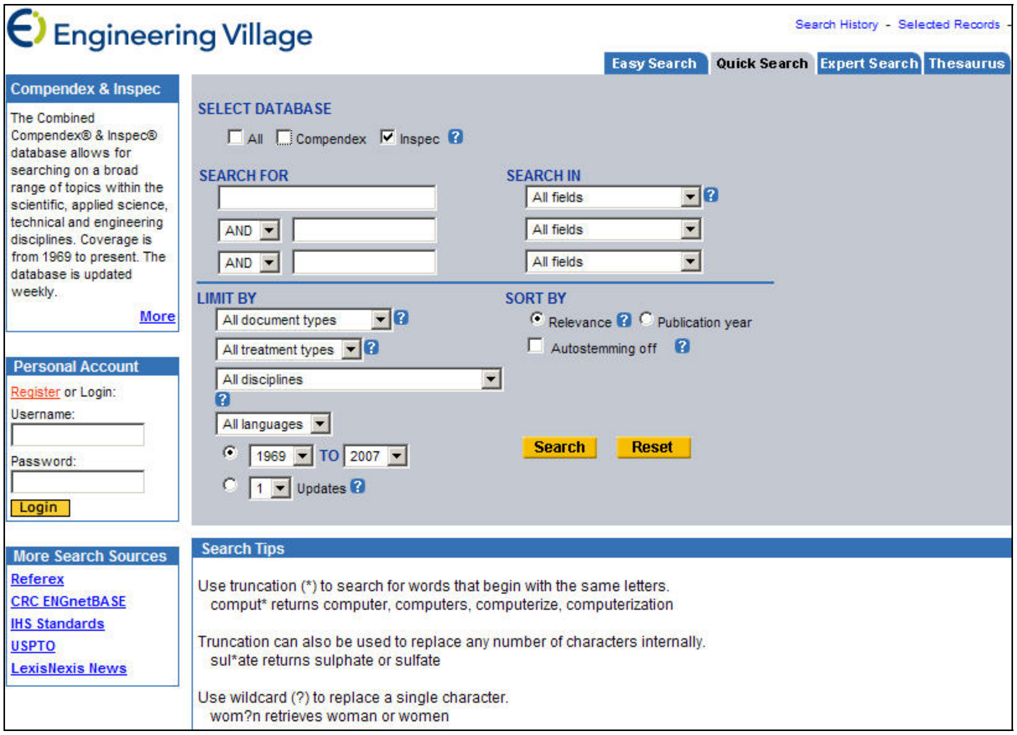
Introduction

A preliminary model applied to typical scientific journal articles was proposed In chapter 2 to improve the effectiveness of automatic subject term assignment by utilizing indexing conceptions in conjunction with semantic sources. The underlying framework of the investigation focused on three indexing conceptions, content-oriented, document-oriented, and domain-oriented, along with corresponding semantic sources. This chapter presents research methods for investigating the research questions and related hypotheses defined in chapter 1. More specifically, this chapter discusses the following: the data source, the data preprocessing, text representation, the text categorization system, effectiveness evaluation, data analysis, and the pilot study.

Data Source

The purpose of this study is to utilize corresponding semantic sources applied in typical scientific journal articles for three conception-based approaches. To accomplish this, the following three requirements are needed for an appropriate data source: 1) bibliographic information for the documents, 2) the full text of the documents, and 3) subject terms assigned by human indexers.

The INSPEC® database was chosen because it met the three requirements listed above. The INSPEC database covers the scientific literature in the fields of electrical engineering, electronics, physics, control engineering, information technology, communications, computers, computing, and manufacturing and production engineering (Engineering Village^{TM}, n.d.). Figure 5 presents the interface of the INSPEC database through Engineering Village ^{TM} 2.



Engineering Village

Search History - Selected Records

Easy Search | Quick Search | Expert Search | Thesaurus

Compendex & Inspec

The Combined Compendex® & Inspec® database allows for searching on a broad range of topics within the scientific, applied science, technical and engineering disciplines. Coverage is from 1969 to present. The database is updated weekly.

More

Personal Account

Register or Login:

Username:

Password:

Login

More Search Sources

Referex

CRC ENGnetBASE

IHS Standards

USPTO

LexisNexis News

SELECT DATABASE

□ All □ Compendex ☑ Inspec ?

SEARCH FOR | SEARCH IN

All fields ?

AND | All fields

AND | All fields

LIMIT BY

All document types ?

All treatment types ?

All disciplines ?

All languages

● 1969 TO 2007

○ 1 Updates ?

SORT BY

● Relevance ? ○ Publication year

□ Autostemming off ?

Search Reset

Search Tips

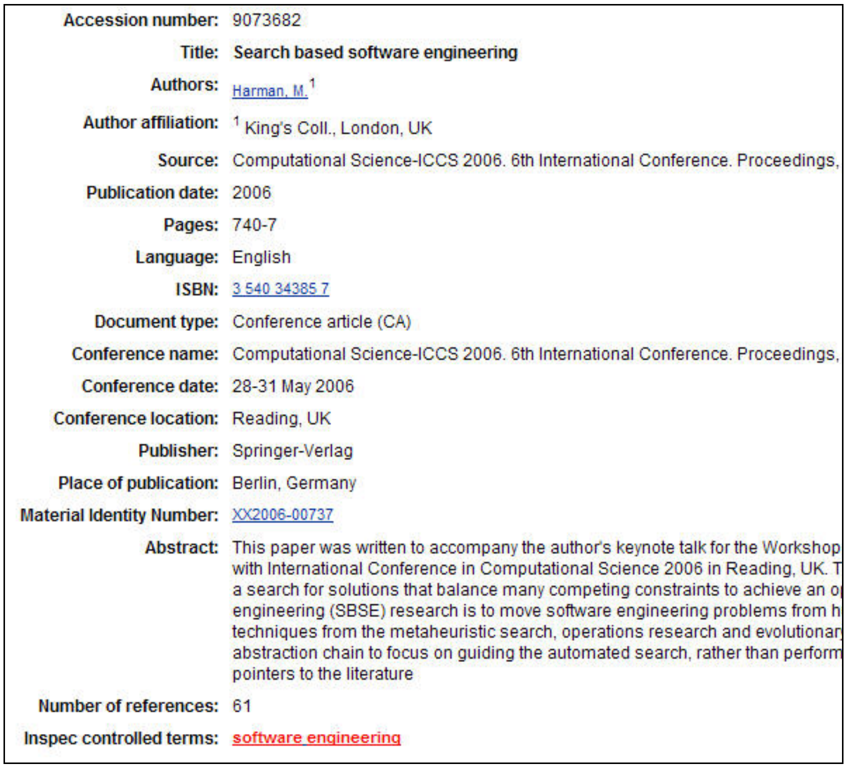
Use truncation (*) to search for words that begin with the same letters.
comput* returns computer, computers, computerize, computerization

Truncation can also be used to replace any number of characters internally.
sul*ate returns sulphate or sulfate

Use wildcard (?) to replace a single character.
wom?n retrieves woman or women

Figure 5. INSPEC database interface through Engineering Village ^{TM} 2.

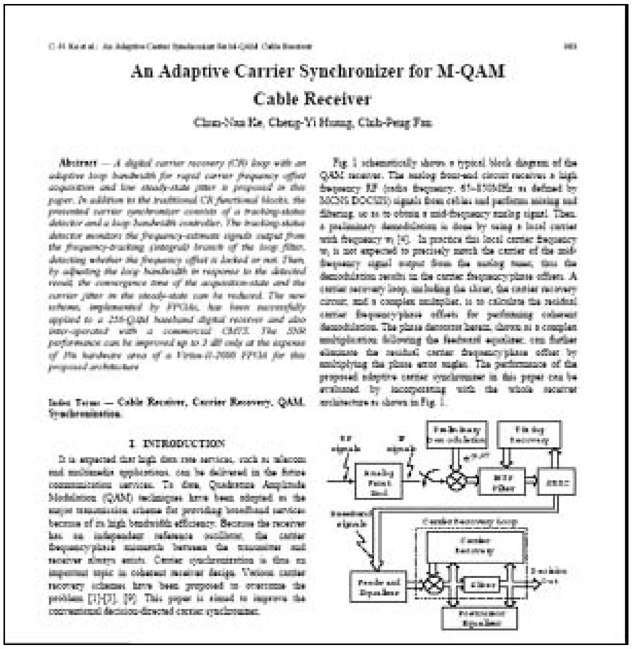
The INSPEC database contains over eight million bibliographic records that represent 3,500 scientific and technical journals and 1,500 conference proceedings (Engineering village ^{TM} 2, n.d.). Figure 6 shows that typical bibliographic records contain 24 elements including title, keywords (uncontrolled terms), abstract, and INSPEC® controlled terms.



Accession number:	9073682
Title:	**Search based software engineering**
Authors:	Harman, M.[1]
Author affiliation:	[1] King's Coll., London, UK
Source:	Computational Science-ICCS 2006. 6th International Conference. Proceedings,
Publication date:	2006
Pages:	740-7
Language:	English
ISBN:	3 540 34385 7
Document type:	Conference article (CA)
Conference name:	Computational Science-ICCS 2006. 6th International Conference. Proceedings,
Conference date:	28-31 May 2006
Conference location:	Reading, UK
Publisher:	Springer-Verlag
Place of publication:	Berlin, Germany
Material Identity Number:	XX2006-00737
Abstract:	This paper was written to accompany the author's keynote talk for the Workshop with International Conference in Computational Science 2006 in Reading, UK. T a search for solutions that balance many competing constraints to achieve an o engineering (SBSE) research is to move software engineering problems from h techniques from the metaheuristic search, operations research and evolutionar abstraction chain to focus on guiding the automated search, rather than perform pointers to the literature
Number of references:	61
Inspec controlled terms:	software engineering

Figure 6. An example of a bibliographic record in the INSPEC® database.

Although not true for all records, most current records are likely to provide links to the full text of the documents, as shown in an example in Figure 7.



C.-N. Ke et al.: An Adaptive Carrier Synchronizer for M-QAM Cable Receiver [illegible]

An Adaptive Carrier Synchronizer for M-QAM Cable Receiver

Chun-Nan Ke, Cheng-Yi Huang, Chih-Peng Fan

Abstract — *A digital carrier recovery (CR) loop with an adaptive loop bandwidth for rapid carrier frequency offset acquisition and low steady-state jitter is proposed in this paper. In addition to the traditional CR functional blocks, the presented carrier synchronizer consists of a tracking-status detector and a loop bandwidth controller. The tracking-status detector monitors the frequency-estimate signals output from the frequency-tracking (integral) branch of the loop filter, detecting whether the frequency offset is locked or not. Then, by adjusting the loop bandwidth in response to the detected result, the convergence time of the acquisition-state and the carrier jitter in the steady-state can be reduced. The new scheme, implemented by FPGAs, has been successfully applied to a 256-QAM baseband digital receiver and also inter-operated with a commercial CMTS. The SNR performance can be improved up to 3 dB only at the expense of 3% hardware area of a Virtex-II-2000 FPGA for this proposed architecture*

Index Terms — Cable Receiver, Carrier Recovery, QAM, Synchronization.

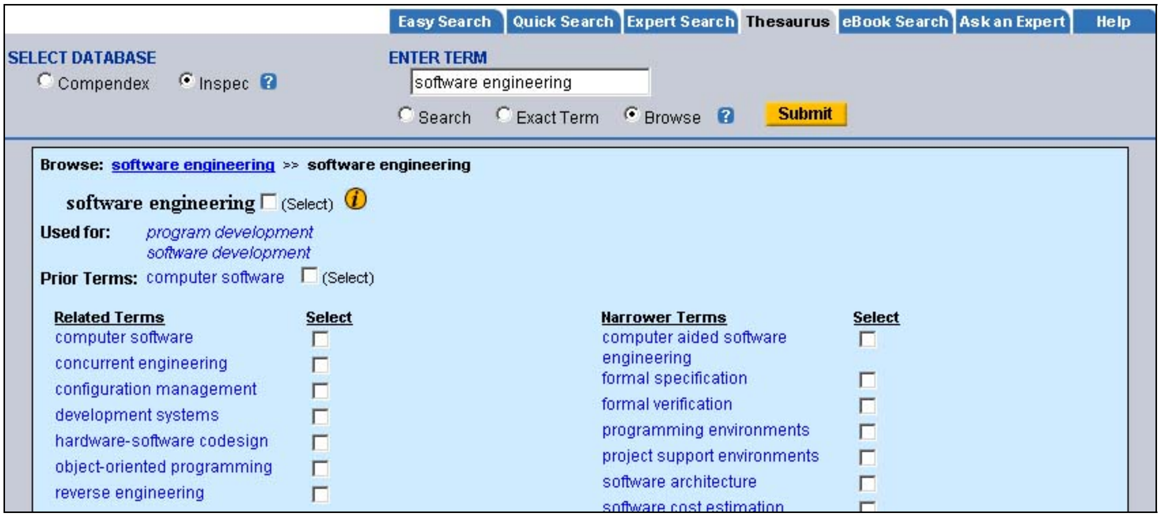
I. INTRODUCTION

It is expected that high data rate services, such as telecom and multimedia applications, can be delivered in the future communication services. To date, Quadrature Amplitude Modulation (QAM) techniques have been adopted as the major transmission scheme for providing broadband services because of its high bandwidth efficiency. Because the receiver has an independent reference oscillator, the carrier frequency/phase mismatch between the transmitter and receiver always exists. Carrier synchronization is thus an important topic in coherent receiver design. Various carrier recovery schemes have been proposed to overcome the problem [1]-[3], [9]. This paper is aimed to improve the conventional decision-directed carrier synchronizer.

Fig. 1 schematically shows a typical block diagram of the QAM receiver. The analog front-end circuit receives a high frequency RF (radio frequency, 85–850MHz as defined by MCNS DOCSIS) signals from cables and performs mixing and filtering, so as to obtain a mid-frequency analog signal. Then, a preliminary demodulation is done by using a local carrier with frequency w_l [4]. In practice this local carrier frequency w_l is not expected to precisely match the carrier of the mid-frequency signal output from the analog tuner, thus the demodulation results in the carrier frequency/phase offsets. A carrier recovery loop, including the slicer, the carrier recovery circuit, and a complex multiplier, is to calculate the residual carrier frequency/phase offsets for performing coherent demodulation. The phase detector herein, shown as a complex multiplication following the feedforward equalizer, can further eliminate the residual carrier frequency/phase offset by multiplying the phase error angles. The performance of the proposed adaptive carrier synchronizer in this paper can be evaluated by incorporating with the whole receiver architecture as shown in Fig. 1.

Figure 7. An example of a full text document in the INSPEC® database.

The INSPEC® controlled terms are assigned by human indexers after analyzing the contents of each document (Engineering village TM 2, n.d.). The INSPEC® thesaurus (INSPEC® thesaurus 2004, 2004) is hierarchical in structure: terms are organized by top (TT: Top Term), broader (BT: Broader Term), narrower (NT: Narrow Term) or related (RT: Related Term) concepts.



SELECT DATABASE
○ Compendex ● Inspec

ENTER TERM
software engineering
○ Search ○ Exact Term ● Browse

Browse: software engineering >> software engineering

software engineering □ (Select)

Used for: program development
software development

Prior Terms: computer software □ (Select)

Related Terms	Select	Narrower Terms	Select
computer software	□	computer aided software engineering	□
concurrent engineering	□	formal specification	□
configuration management	□	formal verification	□
development systems	□	programming environments	□
hardware-software codesign	□	project support environments	□
object-oriented programming	□	software architecture	□
reverse engineering	□	software cost estimation	□

Figure 8. INSPEC® thesaurus Interface.

Data Set

For this study, a total of 1,000 documents with the full text and bibliographic information was collected from the INSPEC® database according to specified subject terms. The 1,000 documents were divided into 20 subject classes. Therefore, the data set for this study will contain 50 full text documents per each of 20 subject classes. For the training phase of text categorization, no standard was identified for the number of instances or words per each subject class in previous research. However, a pilot study conducted earlier in this study showed satisfactory results (at least .696 in F-measure) when using 50 documents in each subject class, and therefore the same strategy was employed when choosing 50 documents in each subject class.

Additionally, the data set was divided into two sub data sets: a homogeneous set and a heterogeneous set. While a homogeneous data set contains more semantically related documents, a heterogeneous data set is composed of more semantically detached documents. For instance, while subject terms within the hierarchy under the same top term are considered as a homogeneous set, subject terms from across multiple top terms are identified as a heterogeneous set. By utilizing two sub-data sets for experimental investigations, this study was able to provide the information on the value of the proposed indexing conception-based framework in conjunction with semantic sources in diverse data set environments. In addition, the experiment results can be validated as to whether or not the results were consistent with diverse nature of data sets.

In order to fulfill the objectives of building a data set and separating it into two sub-sets, two considerations were applied to the sets: 1) the selected terms in the set have the same or a similar hierarchical depth with respect to the INSPEC® thesaurus, and 2) the selected term set has the same or a similar number of records per subject term. First, by ensuring hierarchical depths with similar levels between subject terms, selected terms were prevented from being too specific or too general in comparison to each other. Secondly, by leveling the numbers of records per subject term, selected terms were evenly familiar to indexers without being too well-known or too under-recognized. In addition, computer science and information technology areas were chosen for collection of a data set from among the multiple disciplines of electrical engineering, electronics, physics, control engineering, information technology, communications, computers, computing, and manufacturing and production engineering.

Since the researcher’s expertise lies in computer and information science, the selection process of subject terms was more reliable with respect to ensuring the semantic distance between subject terms.

The INSPEC® thesaurus lists 590 top terms across the previously mentioned scientific disciplines (INSPEC® thesaurus 2004, 2004; Engineering village ^{TM} 2, n.d.). Among the 590 top terms, 32 top terms with hierarchies were identified as related to the disciplines of computer science and information technology. For a homogeneous data set, ten subject terms were selected from one specific top term hierarchy, *software engineering*, within computer engineering and information technology disciplines. Under the *software engineering* top term*,* twenty narrower terms with the same hierarchical depth (the 2nd level) were considered candidate subject terms for a homogeneous data set. However, based on the balance of numbers of records for those twenty terms (see Appendix B), ten terms out of 20 subject terms were selected for the collection of a homogeneous data set. On the other hand, for a heterogeneous data set, another ten subject terms were selected from the computer engineering and information technology disciplines. Twelve top terms with similar hierarchical depths were chosen as candidate subject terms for a heterogeneous data set. Based on the number of records per subject term (see Appendix B), ten subject terms out of twelve were selected. Therefore, a total of 20 subject terms are selected for homogeneous and heterogeneous data sets as shown in Table 4.

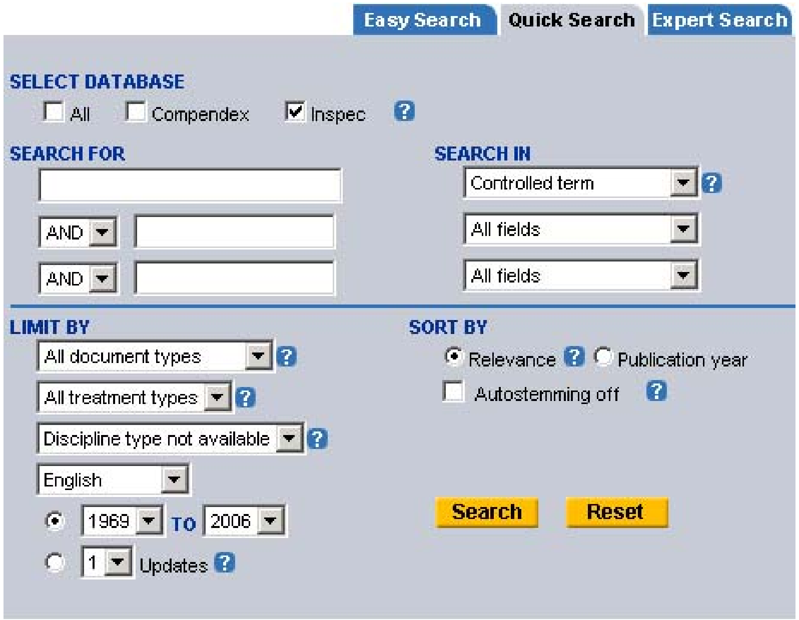
Table 4

Subject Terms for Homogeneous and Heterogeneous Data Sets

Term for homogeneous data set	Term for heterogeneous data set
software architecture	computer architecture
software development management	computer graphics
software libraries	computer interfaces
software maintenance	discrete systems
software metrics	Information management
software portability	knowledge based systems
software prototyping	pattern recognition
software quality	Reliability
software reliability	software engineering
software reusability	user interfaces

Searching Process

As shown in Figure 9, the search process for collecting 50 full text articles for the data set according to the selected 20 terms was specified as follows. 1) SELECT DATABASE was set as INSPEC® database, 2) 20 subject terms for homogeneous and heterogeneous data sets were typed in SEARCH FOR, 3) SEARCH IN was specified as Controlled Term from the drop down lists, 4) the search process was LIMITED BY English language, and 5) the search process was LIMITED BY years of 2000 TO 2006. Since the INSPEC® database contains articles and bibliographic information from 1969 to 2006, a method was needed to limit the affect of changes in subject terms because of the cataloging/indexing practices and policies and terminology. Therefore, search processes for the full text with appropriate bibliographic information was limited to the past six years. A six year time span was arbitrarily determined as a sufficient length of time and it is hypothesized that cataloging/indexing practices, as well as terminology, have not changed that much in this length of time, and current documentation has been applied to these records. In addition, although some articles contain more than one subject term, consideration was limited to a corresponding subject term in order to focus on the purpose of this study and simplify the evaluation process.



SELECT DATABASE
□ All □ Compendex ☑ Inspec

SEARCH FOR

AND

AND

SEARCH IN
Controlled term
All fields
All fields

LIMIT BY
All document types
All treatment types
Discipline type not available
English
● 1969 TO 2006
○ 1 Updates

SORT BY
● Relevance ○ Publication year
□ Autostemming off

Figure 9. Search screen for collecting the documents.

Data Preprocessing

A typical document in the data set contains bibliographic data such as title, abstract, keyword, and source title in addition to the full text of the document. In order to extract eight semantic sources, two procedures were executed: a converting procedure and a semantic source mining procedure.

First, since the full texts of the journal articles are in PDF format, a procedure of converting a PDF format to a text file format was conducted using an Adobe Acrobat Capture[1] software program. Secondly, a semantic source mining procedure[2] was

[1] http://www.adobe.com/products/acrcapture/capfullfeature.html
[2] A Python program written by the researcher

conducted both on the bibliographic information and the full text in the text files. Four semantic sources -- title, abstract, keywords, and source title -- were extracted from the bibliographic information provided by the INSPEC® database. The other four semantic sources -- full text, introduction, conclusion, and titles of cited works -- were extracted from the full text. When there were no indications or subtitles like ‘introduction’ and ‘conclusion’, the first or last 50 lines of each full text from the beginning and from the end, respectively, were used to represent the introduction and the conclusion of each article.

After eight semantic sources were constructed using the two procedures, potentially distracting information was removed as follows.

1) Case: cases in words are converted to lower case
2) Stopwords[3]: stopwords such as the, and, a(n), from, and etc. are removed. Both punctuation and numbers are removed as well.
3) Word normalization[4]: Words are reduced to a standard form which ignores endings for plurals and tense by the Porter stemming algorithm (Porter, 1980).

Text Representation

As the raw text of a document cannot be used for text categorization systems as the input format, the text was converted to an appropriate format for the particular learning algorithm after the preprocessing procedure. The bag of words representation is widely used by text categorization systems including Support Vector Machine

[3] Used the implementation with a stopwords corpus in Natural Language Toolkit for Python (http://nltk.sourceforge.net)

[4] Used the Porter stemming algorithm implementation in Natural Language Toolkit for Python (http://nltk.sourceforge.net)

(Slattery, 2002). The Bag of Words representation reduces each preprocessed document to a list of the unique words in the document and the number of occurrences of each of those words. In addition, since the SVM is fairy robust and scales up to considerable dimensionalities, dimension reduction is generally not needed (Brank, Grobelnik, Milic-Frayling, & Mladenic, 2002). Brank et al. demonstrated that feature selection tends to be detrimental to the performance of SVM. This leads to the bag of words representation of the full text as shown in Table 5.

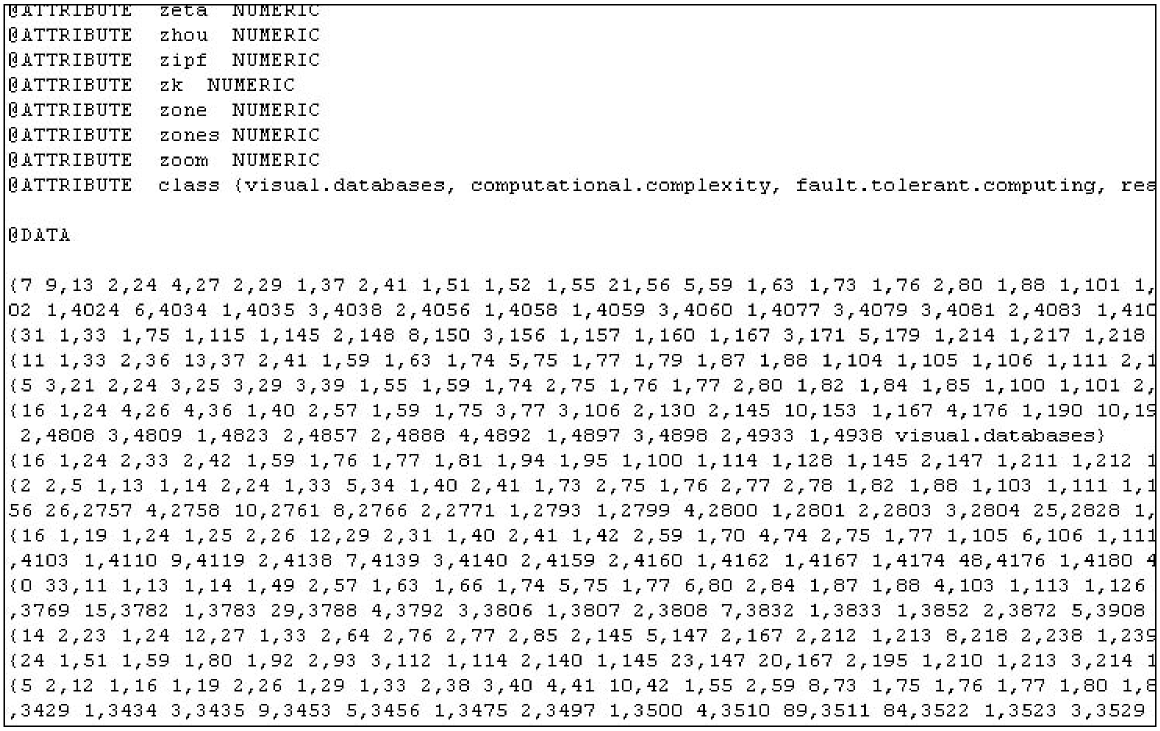
Table 5

Bag of Words Representation of the Full Text of a Document

Word	Frequency	Word	Frequency	Word	Frequency
abnormality	2	bayesian	2	cellular	1
absence	7	beigel	3	chemical	1
acceleration	9	benchmark	1	chernoff	2
Access	1	benefit	1	circle	1
accuracy	12	boolean	5	circuit	0
Acyclic	0	border	6	classify	2
backlog	3	calculation	1	cluster	3
backward	1	camera	1	collect	1
baseline	1	catastrophic	0	....	……

As represented in the Bag of Words, the text was transformed[5] to the WEKA file format (*.arff) for the input of the experiments. More specifically, the sparse arff format was selected as shown in Figure 10. The sparse arff format is able to accelerate the processing time because this format skips data with a value of ‘0’.

[5] A Python program written by the researcher



```
@ATTRIBUTE  zeta  NUMERIC
@ATTRIBUTE  zhou  NUMERIC
@ATTRIBUTE  zipf  NUMERIC
@ATTRIBUTE  zk  NUMERIC
@ATTRIBUTE  zone  NUMERIC
@ATTRIBUTE  zones NUMERIC
@ATTRIBUTE  zoom  NUMERIC
@ATTRIBUTE  class {visual.databases, computational.complexity, fault.tolerant.computing, res

@DATA

{7 9,13 2,24 4,27 2,29 1,37 2,41 1,51 1,52 1,55 21,56 5,59 1,63 1,73 1,76 2,80 1,88 1,101 1,
02 1,4024 6,4034 1,4035 3,4038 2,4056 1,4058 1,4059 3,4060 1,4077 3,4079 3,4081 2,4083 1,410
{31 1,33 1,75 1,115 1,145 2,148 8,150 3,156 1,157 1,160 1,167 3,171 5,179 1,214 1,217 1,218
{11 1,33 2,36 13,37 2,41 1,59 1,63 1,74 5,75 1,77 1,79 1,87 1,88 1,104 1,105 1,106 1,111 2,1
{5 3,21 2,24 3,25 3,29 3,39 1,55 1,59 1,74 2,75 1,76 1,77 2,80 1,82 1,84 1,85 1,100 1,101 2,
{16 1,24 4,26 4,36 1,40 2,57 1,59 1,75 3,77 3,106 2,130 2,145 10,153 1,167 4,176 1,190 10,19
 2,4808 3,4809 1,4823 2,4857 2,4888 4,4892 1,4897 3,4898 2,4933 1,4938 visual.databases}
{16 1,24 2,33 2,42 1,59 1,76 1,77 1,81 1,94 1,95 1,100 1,114 1,128 1,145 2,147 1,211 1,212 1
{2 2,5 1,13 1,14 2,24 1,33 5,34 1,40 2,41 1,73 2,75 1,76 2,77 2,78 1,82 1,88 1,103 1,111 1,1
56 26,2757 4,2758 10,2761 8,2766 2,2771 1,2793 1,2799 4,2800 1,2801 2,2803 3,2804 25,2828 1,
{16 1,19 1,24 1,25 2,26 12,29 2,31 1,40 2,41 1,42 2,59 1,70 4,74 2,75 1,77 1,105 6,106 1,111
,4103 1,4110 9,4119 2,4138 7,4139 3,4140 2,4159 2,4160 1,4162 1,4167 1,4174 48,4176 1,4180 4
{0 33,11 1,13 1,14 1,49 2,57 1,63 1,66 1,74 5,75 1,77 6,80 2,84 1,87 1,88 4,103 1,113 1,126
,3769 15,3782 1,3783 29,3788 4,3792 3,3806 1,3807 2,3808 7,3832 1,3833 1,3852 2,3872 5,3908
{14 2,23 1,24 12,27 1,33 2,64 2,76 2,77 2,85 2,145 5,147 2,167 2,212 1,213 8,218 2,238 1,239
{24 1,51 1,59 1,80 1,92 2,93 3,112 1,114 2,140 1,145 23,147 20,167 2,195 1,210 1,213 3,214 1
{5 2,12 1,16 1,19 2,26 1,29 1,33 2,38 3,40 4,41 10,42 1,55 2,59 8,73 1,75 1,76 1,77 1,80 1,8
,3429 1,3434 3,3435 9,3453 5,3456 1,3475 2,3497 1,3500 4,3510 89,3511 84,3522 1,3523 3,3529
```

Figure 10. Sparse ARFF format for WEKA input.

Text Categorization System

WEKA (Witten & Frank, 2000), a java-based machine learning implementation, was chosen as a text categorization system for investigating the effectiveness of an indexing conception-based framework with semantic sources because of its reliable performance. Among various learning algorithm implementations, the Support Vector Machine (SVM) has been recognized as one of the most successful classification methods (Joachims, 1998) and has been used extensively because of its strong computational learning theory and successes in comparative experiments (Xu, Yu, Tresp, Xu, & Wang, 2003). In WEKA, SVM is implemented as “weka.classifiers.functions.SMO” and selected for the experiments of this study. For each experiment, the same validation method, a ten-fold cross validation method, was followed. Since the average classification error over the ten trials is a good estimate of the overall classification error of the learning method (Watters, Zheng, & Milios, 2002), a ten-fold cross validation method was chosen. The validation method breaks the data into ten equal disjointed subsets and uses one subset as the test data, and the rest as the training data. This is repeated ten times, with each repetition using a different subset.

Effectiveness Evaluation

For the quantification of the performance of the semantic sources and the three approaches based on conceptions, the measures of evaluation were defined as shown in Table 6.

Table 6
Contingency Ttable

	Assigned Term	
Predicted Term	Correct	Incorrect
Correct	a	b
Incorrect	c	d

Recall $= R = \frac{a}{a+c}$, Precision $= P = \frac{a}{a+b}$, and $F = \frac{2PR}{P+R}$

Three effectiveness measures, *recall*, *precision*, and *F*, are common metrics for evaluating text categorization results (Lewis, 1995; Sebastiani, 2002). The recall refers to the ability of the classifier, automatically assigning subject terms to the documents among positive examples and the precision shows the ability of the classifier, automatically assigning subject terms to the documents among positive and negative examples. While the measure of recall reveals whether or not the results of trained classifiers are dominated by false positives, precision shows to what extent the results of trained classifiers are subjected to false negatives (Calvo, Lee, & Li, 2004). Since there is a trade-off between precision and recall as a metric, an approach of combining both has been widely used (Diaz et al., 2004). The F-measure combines the approaches and presents an average of precision and recall. In addition, to simplify the measures, accuracy was also used. It refers to the number of correct predictions in the classification results.

In order to compute the overall performance of the subject classes, two methods were primarily used: macroaveraging and microaveraging. Macroaveraging computes the average precision or recall over all the subject classes. Microaveraging computes the number of documents in each subject class and computes the average in proportion to the number of documents (Diaz et al., 2004). The data set of this study contains a balanced number of documents (50 per subject class) with each class viewed as being equally important. Therefore, it seems reasonable to compute and use macroaveraging for comparison of semantic sources and approaches in the proposed framework (Lewis, 1992).

Data Analysis

This study tests the impact of an indexing conception-based framework on effectiveness when classifying typical scientific journal articles using text categorization techniques. Data analysis focused on four components of this study: (1) the characteristics and importance of semantic sources; (2) the characteristics and importance of an indexing conception-based framework in conjunction with semantic sources; (3) the comparison of the effectiveness of three indexing conceptions; and (4) the repetition of experiments both in the homogeneous and heterogeneous data set environments.

For this study, independent variables contain eight semantic sources including the full text and three indexing conception-based approaches. The eight semantic sources were specified as title, abstract, keywords, introduction, conclusion, source title, titles of cited works, and the full text. The three indexing conception-based approaches were the content-oriented approach, the document-oriented approach, and the domain-oriented approach. The dependent variables in this experiment were the effectiveness of text categorization operationalized by precision, recall and F-measure.

Characteristics and Importance of Semantic Sources

The characteristics and importance of semantic sources are discussed with respect to improving the effectiveness when utilizing semantic sources for text categorization. Related research question and hypothesis described in chapter 1 are as follows.

> Research Question 1:
>
> Can the use of semantic sources, document attributes or the resources to which human indexers refer during the indexing process, improve the effectiveness of text categorization compared to the full text-based text categorization?
>
> Hypothesis for RQ1:
>
> 1. Automatic subject term assignment via semantic sources is effective for automatic subject term assignment in terms of the measures such as recall, precision, and F measure.

In order to investigate the characteristics and significances of semantic sources presented from RQ1 and Hypothesis 1, five steps were followed: 1) the number of words in eight semantic sources was determined, 2) the effectiveness of the eight semantic sources was measured, 3) eight semantic sources were compared with each other and 4) the full text and seven semantic sources were compared.

First, including the full text of the documents, the number of words in eight semantic sources was determined to provide the information on critical relationships between the number of words and the effectiveness. Secondly, the effectiveness of the eight semantic sources was measured by precision, recall, and F-measure in order to provide the information on the relative significance of each semantic source for improving the results of text categorization. Thirdly, the effectiveness in terms of precision, recall, and F-measure was compared in order to obtain information on the comparative importance of the semantic sources using a *t*-test. Finally, the effectiveness of the semantic sources was compared with the effectiveness of the full text as the baseline using *t*-tests. The comparison between the baseline (the full text) and seven semantic sources provided information on the significance of the semantic sources in contrast to the majority of current text categorization studies.

Characteristics and Impact of an Indexing Conception-Based Framework

The characteristics and impact of an indexing conception-based framework are discussed by utilizing three indexing conceptions in conjunction with corresponding semantic sources. The related research question and hypotheses from chapter 1 are as follows.

> Research Question 2:
>
> Can the indexing conceptions, approaches of human indexers to subject analysis and subject indexing for documents, in conjunction with semantic sources, improve the effectiveness of text categorization compared to the full text-based text categorization?

> Hypotheses for RQ2:
>
> 1. Automatic subject term assignment via an indexing conception-based framework is more effective than the full text-based approach in terms of the measure such as recall, precision, and F measure.

In order to investigate the characteristics and impact of an indexing conception-based framework, three steps were conducted as follow: 1) the number of words in three indexing conceptions was determined, 2) the effectiveness of each of the three indexing conceptions was measured, 3) the effectiveness of each of the three indexing conceptions was compared, and 4) the full text and three indexing conceptions were compared.

First, the number of words in each indexing conception was determined in order to provide the information on the relationships between the length of a training set and effectiveness. This revealed whether or not the number of words in each indexing conceptions was significantly correlated with the three indexing conceptions. Secondly, the effectiveness of each of the three indexing conceptions in terms of precision, recall, and F-measure was measured in order to demonstrate the impact of each of the three indexing conception-based approaches. Thirdly, the effectiveness of each of the three indexing conceptions was compared using *t*-tests in order to determine if there were significant differences among three indexing conceptions. Finally, the effectiveness of each of the three indexing conception was compared to the full text as the baseline using a *t*-test in order to reveal the impact of an indexing conception-based framework for text categorization.

Comparison of Three Indexing Conception-Based Approaches

The relative importance and focused conceptions for this specified data set was discussed by comparing the three indexing conceptions within the proposed framework.

> Hypothesis for RQ2:
>
> 2. Automatic subject term assignment via three indexing conception-based approaches, the content-oriented, the document-oriented, and the domain-oriented, is effective in terms of recall, precision, and F measure and the effectiveness will differ among the three approaches.

In order to investigate the relative importance and focused conceptions of this data set, two steps were specified as follows: 1) the effectiveness of three indexing conception-based approaches was measured, and 2) the comparisons between three indexing conception-based approaches were compared.

First, the effectiveness of each of the three indexing conception-based approaches was calculated in terms of precision, recall, and F-measure by utilizing corresponding semantic sources. Secondly, the effectiveness of three indexing conceptions was compared with each other in order to see whether there are significant differences between them.

Repetition of experiments in both homogeneous and heterogeneous data sets

One limitation of this study is the inability to utilize the standard data sets as described in Chapter 1. As a result of this limitation, this study may lead to a limited generalization of the results. As one way of overcoming this limitation, the experiments described in previous sections (Characteristics and Importance of Semantic Sources, Characteristics and Impact of an Indexing Conception-Based Framework, and Comparison of Three Indexing Conception-Based Approaches) were repeated in the homogeneous and heterogeneous data set environments. In order to investigate whether the proposed framework is valid in both homogeneous and heterogeneous data sets, nine steps for each data set were conducted as follows: 1) the number of words in eight semantic sources was determined, 2) the effectiveness of eight semantic sources was measured, 3) the eight semantic sources were compared, 4) the full text and seven semantic sources were compared, 5) the number of words in three indexing conceptions was determined, 6) the effectiveness of three indexing conceptions was measured, 7) the full text and three indexing conceptions were compared, 8) the effectiveness of three indexing conception-based approaches was measured, and 9) the three indexing conception-based approaches were compared.

Pilot Study

A pilot study was conducted to explore the feasibility of utilizing an indexing conception-based framework in conjunction with corresponding semantic sources for automatic subject term assignment. The pilot study was performed in part to refine the study’s data collection and preprocessing and strengthen the overall design. The operationalized pilot study environments are as follows.

- Data Set: 200 full text articles with bibliographic information from INSPEC® database
- Four Subject Terms: ‘visual database’, ‘real-time systems’, ‘fault tolerant computing’, ‘computational complexity’
- Semantic Sources: ‘title’, ‘abstract’, ‘keywords’, ‘source title’, ‘introduction’, ‘conclusion’, ‘titles of cited works’
- Three indexing conceptions: content-oriented, document-oriented, domain-oriented
- Support vector machine algorithm was chosen in WEKA implementations

The pilot study showed that some of the semantic sources were more effective than the full text for automatic subject term assignment. For instance, ‘introduction’, ‘titles of cited works’, and ‘keywords’ demonstrated better effectiveness than the full text in terms of F-measure. The three indexing conception-based approaches in conjunction with corresponding semantic sources were tested via precision, recall, and F measures. From the perspective of relative importance of the three indexing conception-based approaches, the document-oriented and domain-oriented approaches were more effective than the baseline (the full text) for automatic subject term assignment. The findings of this pilot study indicated that subject terms were assigned more effectively by text categorization when incorporating the indexing conceptions and corresponding semantic sources into text categorization applications (Chung & Hastings, 2006).

Summary

This chapter presented the research approach for investigating the impact of an indexing conception-based framework in conjunction with semantic sources for this study. Detailed descriptions were provided for the research methods including the data source, preprocessing methods, text representation for text categorization systems, a selected text categorization system, and the methods of effectiveness evaluation. The data analysis included four components: 1) the characteristics and significance of semantic sources, 2) the characteristics and impact of an indexing conception-based framework, 3) the comparison of three indexing conception-based approaches, and 4) the repetition of experiments in homogeneous and heterogeneous data set environments. Finally, the results of a pilot study found that indexing conceptions combined with semantic sources were more effective than full text-based text categorization.

CHAPTER 4

DATA ANALYSIS

Introduction

The main purpose of this study was to investigate whether indexing conception-based approaches in conjunction with semantic sources are effective for text categorization (TC). The results of data analysis are provided in light of two research questions and three related hypotheses: 1) the effectiveness of semantic sources, 2) the effectiveness of indexing conception-based approaches, and 3) the effectiveness of each of the indexing conception-based approaches. This chapter is composed of three sections: the data sets, the effectiveness of semantic sources, and the effectiveness of indexing conceptions. The first section describes the composition of the data sets, the characteristics of the data sets, and the number of words associated with each data set. The second section, the effectiveness of semantic sources, presents the results of experiments utilizing seven individual semantic sources for TC; this section contains a general description of effectiveness using individual semantic sources, a comparative analysis of semantic sources, and a comparative analysis with the baseline. Finally, the effectiveness of indexing conceptions section provides the results of experiments exploiting indexing conception-based approaches for TC in order to assign subject terms to the documents.

Data Set

A total of 1,000 records containing bibliographic information were searched and obtained using 20 subject terms in Table 4 from the INSPEC® database via the Engineering Village ^{TM} 2 (electronic resource) at the University of North Texas Libraries[6]. The data set used in this study is composed of three data sets, the full data set and two sub-data sets, one of which is a homogeneous data set and the other is a heterogeneous data set. While the homogeneous data set contains semantically-related subject classes, the heterogeneous data set represents less semantically-related subject classes. The homogeneous data set was searched using ten subject terms from one specific Top Term of the INSPEC® Thesaurus (INSPEC® thesaurus 2004, 2004) in order to build a semantically-related data set. The heterogeneous data set was built using another ten subject terms from diverse areas of Top Terms for a less semantically-related data set. The full data set is a combination of the both sets. In addition, these three data sets are orthogonal to the three indexing conceptions and the eight semantic sources as shown in Figure 11. In other words, each of the semantic sources and indexing conceptions includes three data sets - the full data set, the homogeneous data set, and the heterogeneous data set.

[6] http://www.library.unt.edu/

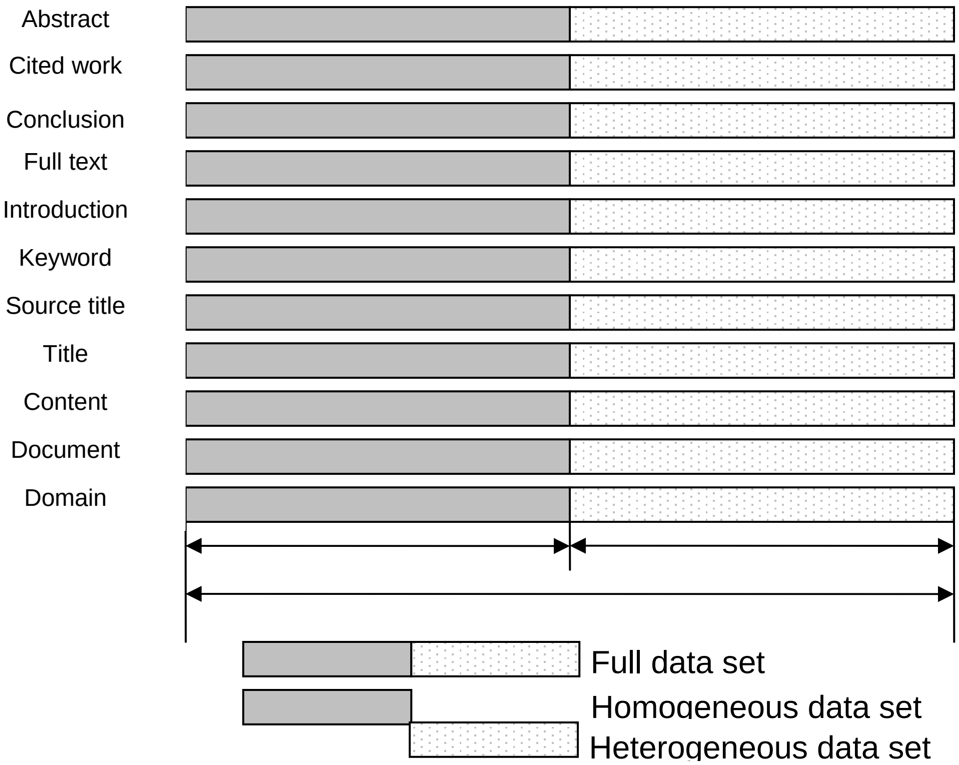


Figure 11. The composition of three data sets.

Ideally, the full data set for each semantic source and indexing conception should contain 1,000 records using twenty subject terms. The homogeneous and heterogeneous data sets for each semantic source and indexing conception should each have 500 records using ten subject terms. However, some events caused missing data. For example, conversion from a PDF file format to a text file format caused problems because of unusual PDF file formats; exclusion of citations, conclusions, abstracts, introduction, and keywords. As a result, Table 7 presents the actual numbers of each data set used in this study in conjunction with the semantic sources and the indexing conceptions. There are 10,536 records in the full data set, 5,266 records for the homogeneous data set, and 5,270 records for the heterogeneous data set, respectively.

Table 7

Actual Number of Data Associated with Three Data Sets

Semantic source and Indexing conception	Full data set	Homogeneous Data set	Heterogeneous Data set
Abstract	959	479	480
Cited works	946	486	460
Conclusion	915	446	469
Full text	955	477	478
Introduction	955	477	478
Keyword	959	479	480
Source title	959	479	480
Title	959	479	480
Content-oriented	957	472	485
Document-oriented	979	494	485
Domain-oriented	993	498	495
Total	10,536	5,266	5,270

As described in the methodology, these three data sets were preprocessed in terms of case, stopwords, and word normalization. First, upper cases words were converted to lower case. Secondly, stopwords (such as ‘the’, ‘a(n)’, ‘from’, ‘to’), punctuation, and numbers were removed as well. Finally, words were reduced to a standard form by ignoring endings for plurals and tenses. After the three data sets were normalized, the number of words associated with the semantic sources and indexing conceptions were computed as shown in Table 8.

Table 8

Number of Words Associated with Three Data Sets

	Full data	Homogeneous Data	Heterogeneous Data
Abstract	121,699	61,023	60,676
cited works	472,985	244,502	228,483
Conclusion	275,253	149,797	125,456
Full text	6,602,440	3,507,097	3,095,343
Introduction	517,197	275,728	241,469
Keyword	32,861	15,625	17,237
Source title	4,074	2,026	2,048
Title	7,553	3,690	3,863
content-oriented	396,952	210,820	186,132
document-oriented	557,611	295,043	262,569
domain-oriented	477,059	246,528	230,531
Total	9,465,684	5,011,879	4,453,807

The number of words associated with each of the semantic sources and indexing conceptions are graphically demonstrated in Figure 12. This figure illustrates the differences between the baseline (i.e., the full text) and the individual semantic sources and indexing conceptions. The number of associated words demonstrates that the baseline contains 11 times more than the greatest number of words (document-oriented) and 1,620 times more than the least number of words (source title).

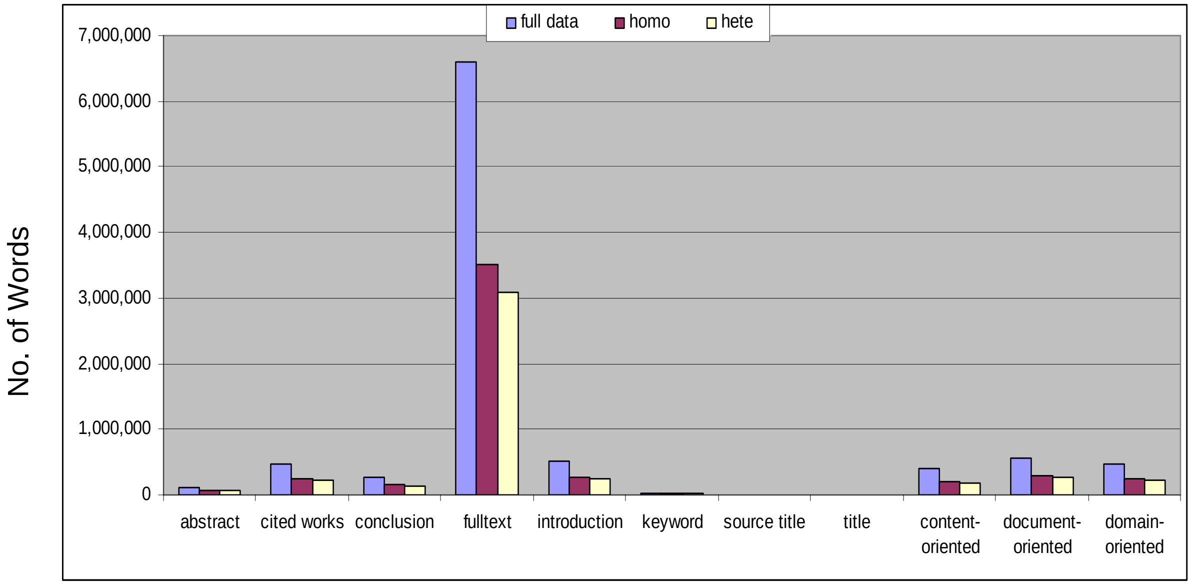


Figure 12. The number of words associated with datasets.

Analysis of Semantic Source Effectiveness

There are several attributes of documents that can be considered for incorporation into the learning process of TC when taking the perspective of the subject term assignment process by human indexers. Semantic sources are defined as document attributes to which human indexers refer in order to analyze the *aboutness* of a document. While the majority of TC research has focused on utilizing the full text of documents without considering the importance of document attributes, it is worthwhile to study the individual values of semantic sources in terms of automatic subject term assignment through TC. The investigations of these sources can reveal to what extent they are effective, when compared to the full text, for assigning subject terms through TC techniques.

General Analysis of Semantic Sources

TC experiments using individual semantic sources one at a time were conducted with the intention of recognizing the significance of semantic sources. In this section, the results are shown in terms of effectiveness for all the three data sets.

For the full data set, the precision, recall, and F-measure in each test round were computed as shown in as shown in Table 9 and Figure 13.

In

Table 9, the macroaveraged precision, recall, and F-measure are given for the full data set. In terms of F-measure, keyword, source title, title, cited works, and full text show relatively high effectiveness. In Figure 13, the effectiveness of individual semantic sources for the full data set is graphically shown in terms of precision, recall, and F-measure. An increasing order of effectiveness, by sources such as conclusion, abstract, introduction, title, source title, full text, cited works, and keyword, is revealed through the use of the F-measure.

Table 9

Macroaveraged Precision, Recall and F-measure for the Full Data Set

Semantic source	Precision	Recall	F-measure
Abstract	.277	.234	.230
Cited works	.344	.290	.308
Conclusion	.206	.193	.193
Full text	.349	.283	.300
Introduction	.283	.213	.231
Keyword	.387	.366	.368
Source title	.323	.304	.299
Title	.319	.293	.296

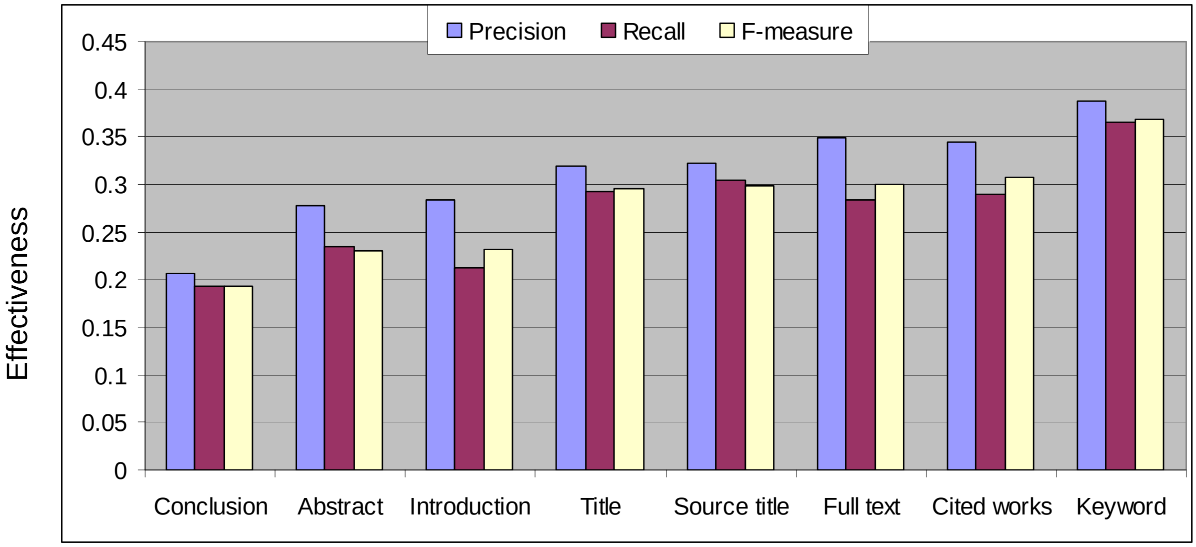


Figure 13. The effectiveness of semantic sources for the full data set.

For the homogeneous data set, Table 10 and Figure 14 present the effectiveness of semantic sources in terms of the three measures (precision, recall, and F-measure). The relative effectiveness of full text (.3 in F-measure) decreases compared to full text (.261 in F-measure) in the full data set. While cited works and keyword still demonstrated high effectiveness, the conclusion, introduction, and abstract are the least effectiveness among the semantic sources. In general, the overall effectiveness of semantic sources is considered as consistent with the effectiveness of the full data set.

Table 10

Macroaveraged Precision, Recall and F-measure for the Homogeneous Data Set

Semantic source	Precision	Recall	F-measure
Abstract	.283	.275	.262
Cited works	.345	.310	.322
Conclusion	.249	.244	.243
Full text	.287	.258	.261
Introduction	.286	.234	.247
Keyword	.402	.382	.384
Source title	.267	.269	.262
Title	.310	.292	.295

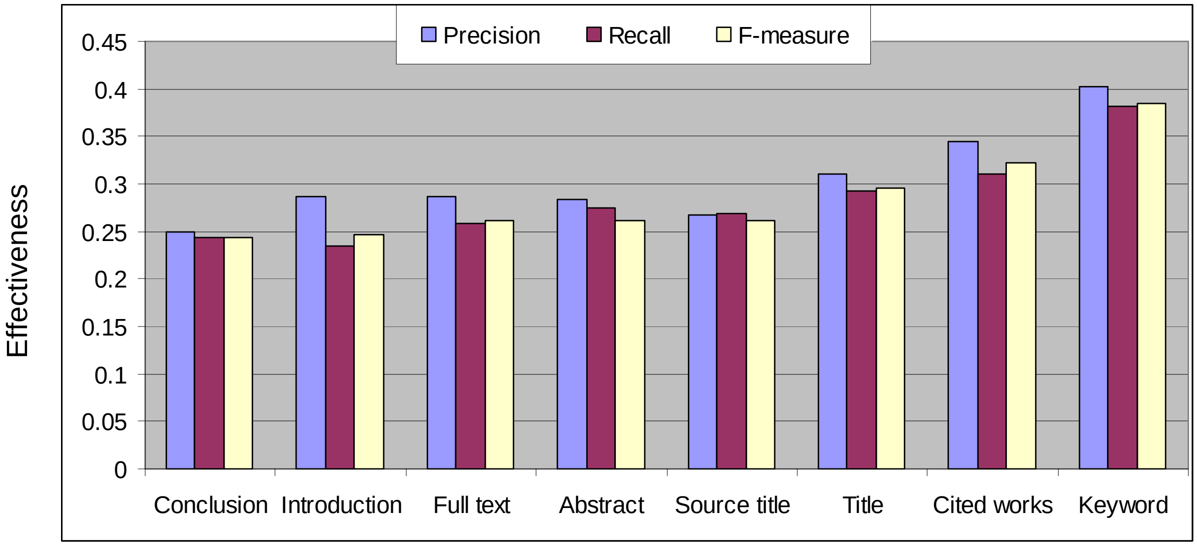


Figure 14. The effectiveness of semantic sources for the homogeneous data set.

Table 11 and Figure 15 present the effectiveness of each semantic source for the heterogeneous data set. The effectiveness of semantic sources in this set generally increase compared to the full data set and the homogeneous data set. As shown in Figure 15, the effectiveness of the full text increases following the effectiveness of keyword when compared to the full data set and the homogeneous data set. Except for the high effectiveness of the full text in the heterogeneous data set, the semantic sources' effectiveness are consistent with the full data set and the heterogeneous data set. While cited works and keyword were considerably effective for TC, semantic sources such conclusion, abstract, and introduction are shown to be less effective.

Table 11

Macroaveraged Precision, Recall and F-measure for the Heterogeneous Data Set

Semantic source	Precision	Recall	F-measure
Abstract	.425	.393	.389
Cited works	.493	.459	.469
Conclusion	.371	.349	.348
Full text	.564	.505	.520
Introduction	.444	.426	.427
Keyword	.587	.568	.569
Source title	.461	.420	.432
Title	.487	.422	.439

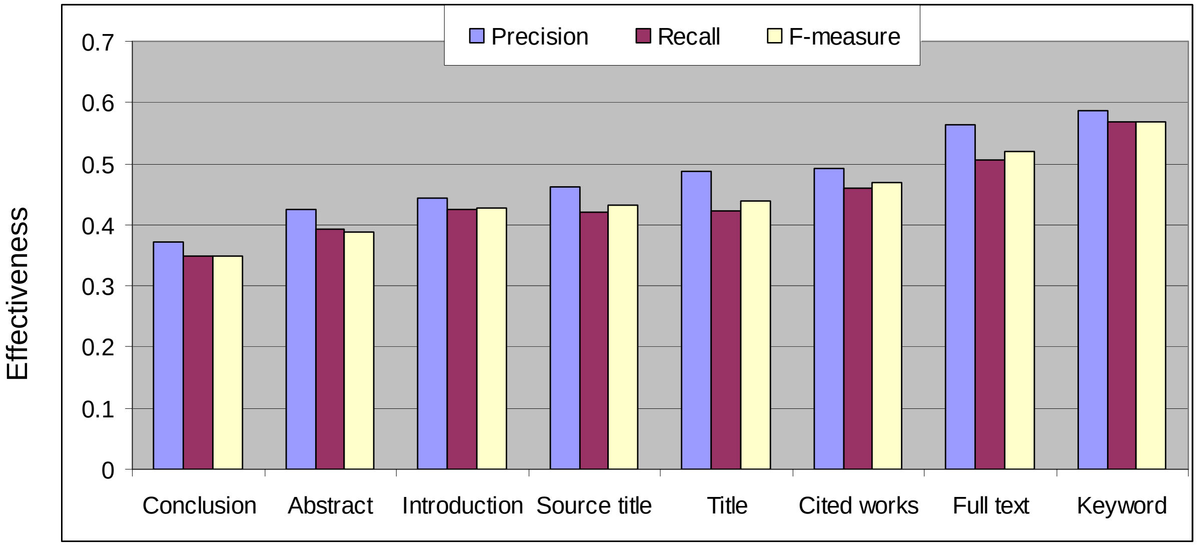


Figure 15. The effectiveness of semantic sources for the heterogeneous data set.

Although some similarities are found among the three data sets in the previous analysis, Figure 16, Figure 17, and Figure 18 present the F-measures, precision, and recall, respectively, in order to examine the relationships more clearly.

Figure 16 confirms that the heterogeneous data set generally shows better effectiveness than the full data set and the homogeneous data set in terms of F-measure. Each of the semantic sources demonstrates nearly the same behaviors among the three data sets. Keyword and cited works were found effective, while conclusion and abstract were not as effective as the other semantic sources.

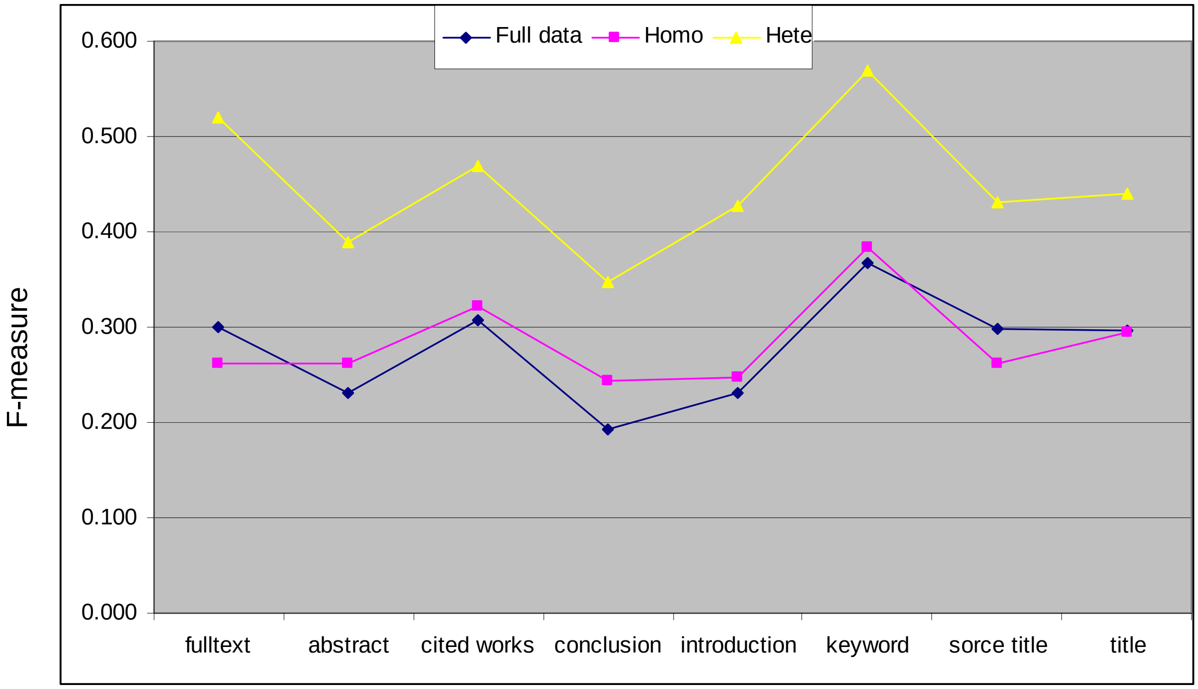


Figure 16. Semantic sources from three data sets in F-measure.

As shown in Figure 17, precision measures are consistent with the F-measure in that the heterogeneous data set shows greater effectiveness than the full data set and homogeneous data set. Cited works is notable because both the homogeneous data set and the heterogeneous data set have lower effectiveness compared to the results of F-measures.

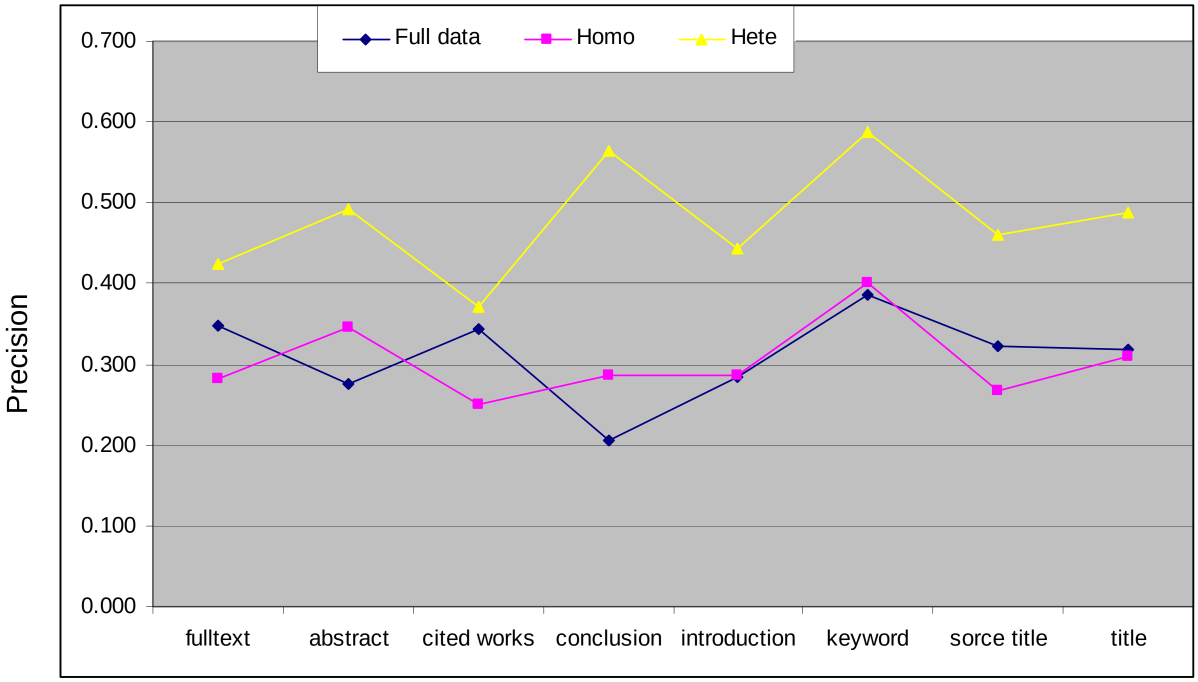


Figure 17. Semantic sources from three data sets in precision.

In terms of recall, Figure 18 demonstrates that the heterogeneous data set shows greater effectiveness than the full data set and the homogeneous data set.

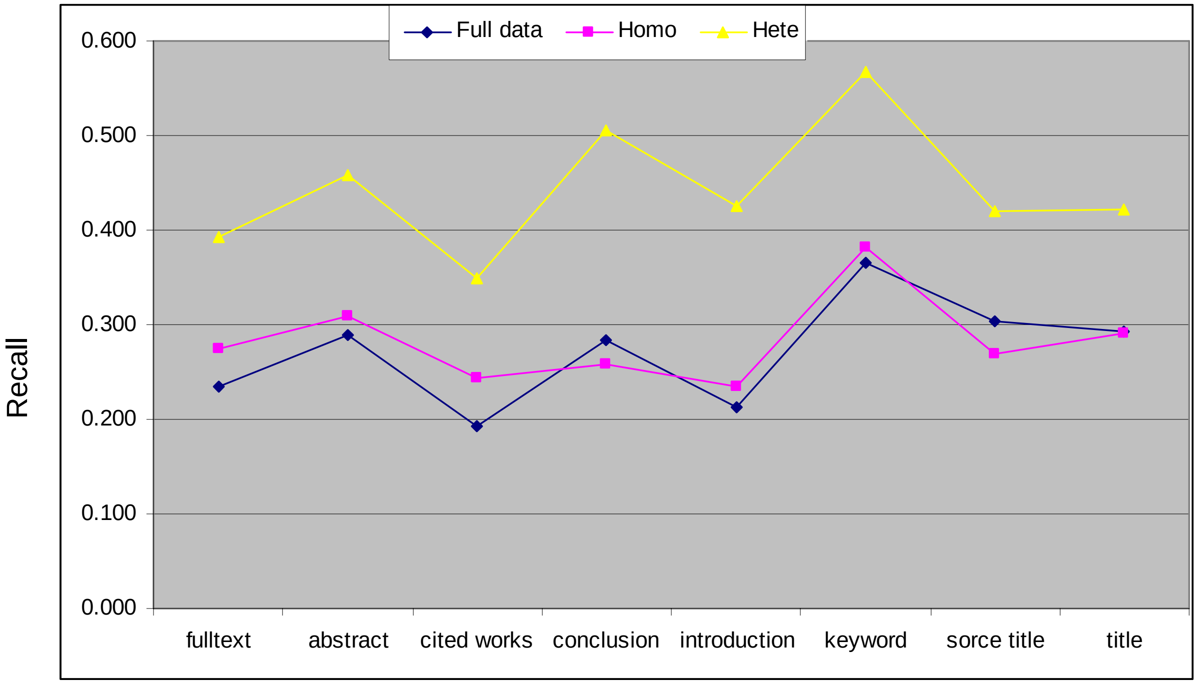


Figure 18. Semantic sources from three data sets in recall.

In conclusion, keyword, cited works, source title, and title show high effectiveness and are fairly consistent. Taking into account both the effectiveness results of semantic sources, two comparisons can be noted. One is the characteristic comparison between abstract (.389 in F-measure, Heterogeneous) and keyword (.569 in F-measure, Heterogeneous). While both abstract and keyword are provided by the authors of the articles and represent a concise version of the full text, there is a substantial gap between the two, even though each is provided by the author(s) of the articles and represent a concise version of the full text. Another worthy comparison is that between introduction (.427 in F-measure, Heterogeneous) and conclusion (.348 in F-measure, Heterogeneous). Although both semantic sources were extracted from the full text of the articles for different purposes, there is a considerable difference in the effectiveness results between the two. While introduction shows greater effectiveness than conclusion, the number of words associated (Table 8) with introduction is approximately twice as large as the number of words associated with conclusion.

Comparisons of semantic sources

Comparisons were made using *t*-tests between each of the semantic sources In order to indicate how the effectiveness of each semantic source differs from the effectiveness of other semantic sources.

Table **12** demonstrates that there are significant differences between the thirteen pairs when setting alpha value to .05. In terms of a positive impact on the effectiveness of TC, keyword is significantly different from abstract, conclusion, title, source title, and introduction. In addition, cited works show significantly greater effectiveness in comparison to abstract, conclusion, and introduction. While the effectiveness of title is significantly different than abstract, conclusion, and introduction, the effectiveness of source title is significantly different than that of abstract or conclusion.

In the homogeneous data set, five pairs of semantic sources were found to be significantly different as shown in Table 13. Keyword was identified to be significantly different when paired with abstract (.000), conclusion (.001), introduction (.044), source title (.001), and title (.010).

Table 12

t-*tests with each semantic sources in terms of F-measure (Full Data Set)*

	Abstract	Cited-works	Conclusion	Introduction	Keyword	Source Title	Title
Abstract		**.035***	.079	.975	**.000***	**.014***	**.023***
Cited-works			**.003***	**.045***	.059	.748	.744
Conclusion				.126	**.000***	**.000***	**.001***
Introduction					**.000***	.054	**.036***
Keyword						**.010***	**.003***
Source Title							.931
Title							

*p<.05

Table 13

t-*tests with each semantic source in terms of F-measure (Homogeneous Data Set)*

	Abstract	Cited-works	Conclusion	Introduction	Keyword	Source Title	Title
Abstract		.357	.520	.783	**.000***	.993	.431
Cited-Works			.243	.416	.283	.369	.692
Conclusion				.940	**.001***	.648	.221
Introduction					**.044***	.811	.357
Keyword						**.001***	**.010***
Source Title							.381

Title							

**p<.05*

For the heterogeneous data set, Table 14 illustrates that eight pairs of semantic sources are identified as significantly different. Keyword was found to be significantly different when paired with abstract (.006), conclusion (.001), introduction (.001), and title (.001). In addition, conclusion, when paired with cited works (.001), introduction (.002), and title (.001) show significant differences.

Table 14

t-*tests with each semantic source in terms of F-measure (Heterogeneous Data Set)*

	Abstract	Cited-works	Conclusion	Introduction	Keyword	Source Title	Title
Abstract		.148	.335	.270	**.006***	.441	.346
Cited-Works			**.001***	.170	.069	.526	.544
Conclusion				**.002***	**.001***	.120	**.047***
Introduction					**.001***	.914	.732
Keyword						**.004***	**.001***
Source Title							.879
Title							

**p<.05*

In sum, the homogeneous data set demonstrates the most rigorous comparison of results. From the homogeneous data set, the comparisons between each of semantic sources indicate that keyword achieved significantly greater effectiveness when compared to abstract, conclusion, introduction, source title, and title. The full data set and the heterogeneous data sets support the excellence of keyword as well. The two data sets also demonstrate that cited works was found to be more effective than abstract, conclusion, and introduction. In addition, source title and title were found to be significantly more effective than abstract, conclusion, and introduction.

Comparisons of semantic sources with the baseline

In order to investigate the differences between individual semantic sources and the baseline, *t*-tests of individual semantic sources compared to the baseline (the full text) are presented in Table 15, Table 16, and Table 17 using F-measure, recall, and precision. Each table presents seven pairs of applied *t*-tests.

Table 15

t-*tests between the baseline and individual semantic sources in terms of F-measure*

Semantic source	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Abstract	3.446	**.003***	-.025	.981	3.273	**.010***
Cited works	-.249	.806	-.857	.414	1.252	.242
Conclusion	4.481	**.000***	.500	.629	4.333	**.002***
Introduction	2.587	**.018***	.310	.763	2.254	.051
Keyword	-2.231	**.038***	-5.212	**.001***	-.753	.471
Source title	.043	.966	-.015	.988	1.402	.194
Title	.110	.913	-1.267	.237	1.321	.219

*p<.05

When the alpha value is set to .05, using F-measure as shown in Table 15, significant differences are found between the baseline (full text) and the abstract, introduction, and keyword for the full data. There are no significant differences between the baseline and cited works, source title, and title. On the other hand, the homogenous data set and the heterogeneous data set indicate more rigorous results than the full data set. Keyword, in the homogeneous data set, is the only one that demonstrates significant differences. In the heterogeneous data set, abstract and conclusion demonstrate significant differences compared to the baseline. In the same set, cited works, source title, and title show no significant difference with the baseline.

From three data sets, significantly different semantic sources such as abstract, conclusion, introduction, and keyword are summarized as follows. First, keyword show greater effectiveness when compared to the baseline. Secondly, title, source title, and cited works show no significant difference compared to the full text. Finally, abstract, conclusion, and introduction indicate less effectiveness when compared with the full text. These results can guide the utilization of individual semantic sources for more efficient automatic subject term assignment through TC.

Table 16

t-*tests between the baseline and individual semantic sources in terms of Recall*

Semantic source	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Abstract	1.443	.165	-.441	.669	1.763	.112
Cited works	-.160	.875	-.601	.563	1.398	.196
Conclusion	3.432	**.003***	.300	.771	4.176	**.002***
Introduction	1.787	.090	.339	.742	1.619	.140
Keyword	-2.935	**.008***	-3.945	**.003***	-1.057	.318

Source title	-.852	.405	-.484	.640	1.677	.128
Title	-.354	.727	-1.076	.310	1.733	.117

*p<.05

Table 17

t-tests between the baseline and individual semantic sources in terms of Precision

Semantic source	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Abstract	2.356	**.029***	.120	.907	2.306	**.047***
Cited works	.131	.897	-.925	.379	1.095	.302
Conclusion	4.523	**.000***	1.086	.306	2.611	**.028***
Introduction	2.389	**.027***	.010	.992	2.034	.073
Keyword	-.924	.367	-4.757	**.001***	-.286	.781
Source title	.704	.490	.476	.645	1.235	.248
Title	.644	.527	-.510	.623	.938	.373

*p<.05

In the full data set, abstract, conclusion, and introduction are significantly different from the baseline in terms of precision as shown in Table 17. For the homogeneous data set, only keyword is found significantly different from the baseline, while abstract and conclusion were found to be significantly different in the heterogeneous data set. The positive impact of keyword is found only in the homogeneous data set. On the other hand, abstract, conclusion, and introduction were found to be significantly different from the baseline in a negative way.

The results of comparison of semantic sources and the baseline are summarized as follows. Out of seven comparisons with the baseline, keyword was found to be significantly different from the baseline in a positive way, while abstract, conclusion, and introduction were found significantly different in a negative way. In addition, cited works, title, and source title indicate no significant differences from the baseline. Cited works, title, and source title can be considered practical alternatives for TC without considerable processing procedures when dealing with the full text, in addition to the demonstrated excellent results of keyword. In general, these results are consistent with previously reported results (Larkey, 1999; Zhang et al., 2004) in which better performances were presented with a combination of one or more document attributes, rather than just the full text alone.

Therefore, the results of the data analysis using semantic sources support Hypothesis 1 set forth in Chapter 3.

> H1: Automatic subject term assignment via semantic sources is effective for automatic subject term assignment in terms of recall, precision, and F measure.

When utilizing individual semantic sources instead of the full text, the data analysis results indicate that one semantic source (keyword) demonstrated consistently better effectiveness than the full text and that three other semantic sources (cited works, title, and source title) are just as effective as the full text. Utilization of individual semantic sources instead of just the full text is desirable for automatic subject term assignment through TC, especially when considering the computing time and resources combined with the good effectiveness of the individual semantic sources,

Analysis of Three Indexing Conception-Based Approaches

The conceptions of subject indexing are defined as the indexers' perceptions or approaches in regard to subject analysis determination, and indexing. Based on the purpose of this study, human indexers' subject indexing conceptions, the content-oriented, the document-oriented, and the domain-oriented indexing conceptions, were tested to see if utilization of subject indexing conceptions in conjunction with corresponding semantic sources is effective for automatic subject term assignment through TC. Identified semantic sources for each approach were combined according to the preliminary framework as shown in Figure 4. For the domain-oriented approach, source title and cited works were used for the experiments. The document-oriented approach includes introduction, title, and keyword, and the content-oriented approach includes conclusion and abstract for the experiments.

General Analysis of Indexing Conception-Based Approaches

Table 18, Table 19 and Table 20 present the results of the experiments for three approaches in terms of precision, recall, and F-measure for the full data set, the homogeneous data set, and the heterogeneous data set. With respect to the three measures, in general, the content-oriented and the document-oriented approaches show greater effectiveness compared to the domain-oriented approach.

Table 18
Macroaveraged Precision, Recall, and F measure for the Full Data Set

Indexing conception	Precision	Recall	F-measure
Content-oriented	.450	.394	.409
Document-oriented	.541	.253	.312
Domain-oriented	.270	.155	.168

Table 19

Macroaveraged Precision, Recall and F-measure for the Homogeneous Data Set

Indexing conception	Precision	Recall	F-measure
Content-oriented	.407	.402	.399
Document-oriented	.485	.173	.194
Domain-oriented	.310	.195	.199

Table 20

Macroaveraged Precision, Recall, and F-measure for the Heterogeneous Data Set

Indexing conception	Precision	Recall	F-measure
Content-oriented	.623	.562	.576
Document-oriented	.676	.590	.608
Domain-oriented	.412	.296	.309

In addition, Figure 19, Figure 20, and Figure 21 are presented in order to examine whether the behaviors of different data sets are consistent with the results in terms of recall, precision, and F-measure. In general, the graphical demonstration (Figure 19) indicates that the heterogeneous data set showed better effectiveness than the full data set and the homogeneous data set. Of the three indexing conceptions, the domain-oriented approach demonstrated less effectiveness compared to the content-oriented and the document-oriented. These results are consistent with three data sets, the full data set, the homogeneous data set, and the heterogeneous data set.

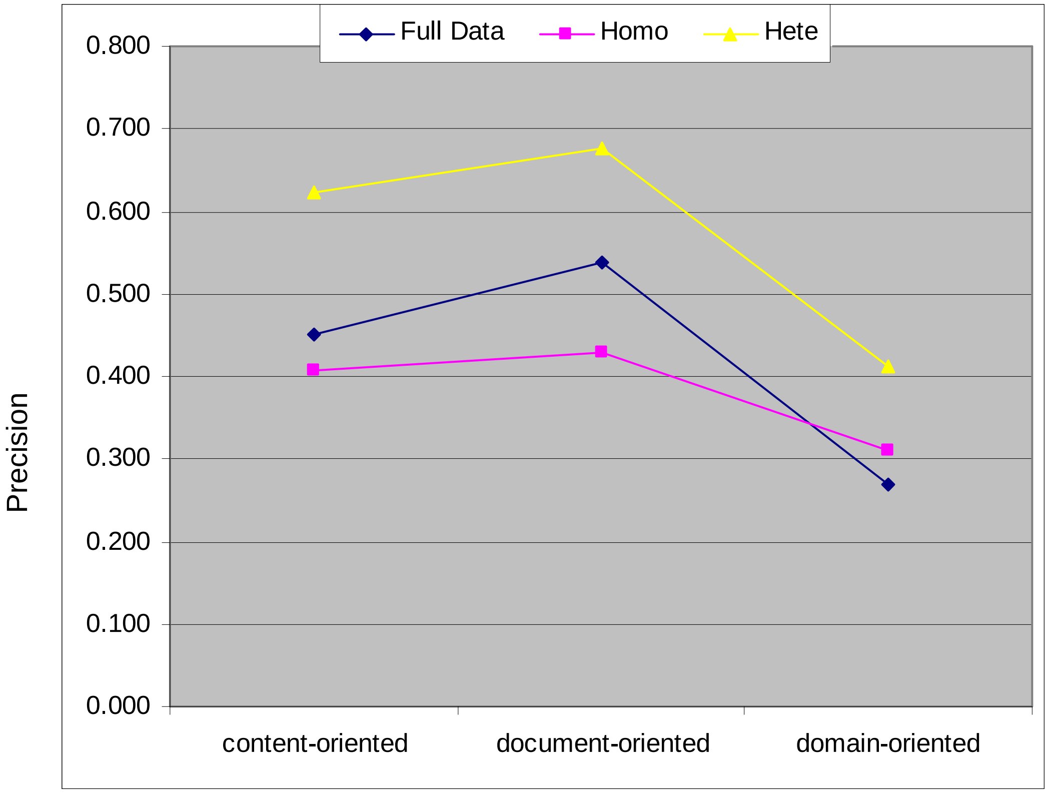


Figure 19. Three indexing conceptions in terms of precision.

In terms of recall, as shown in Figure 20*,* the heterogeneous data set shows general better effectiveness than the homogeneous data set and the full data set and this is consistent with the results using the F-measure. It is worth noting that the behaviors of the three indexing approaches are very closely related with respect to the homogeneous data set and the full data set. The content-oriented and the document-oriented approaches, of the three indexing conceptions, are more effective than the domain-oriented indexing approach.

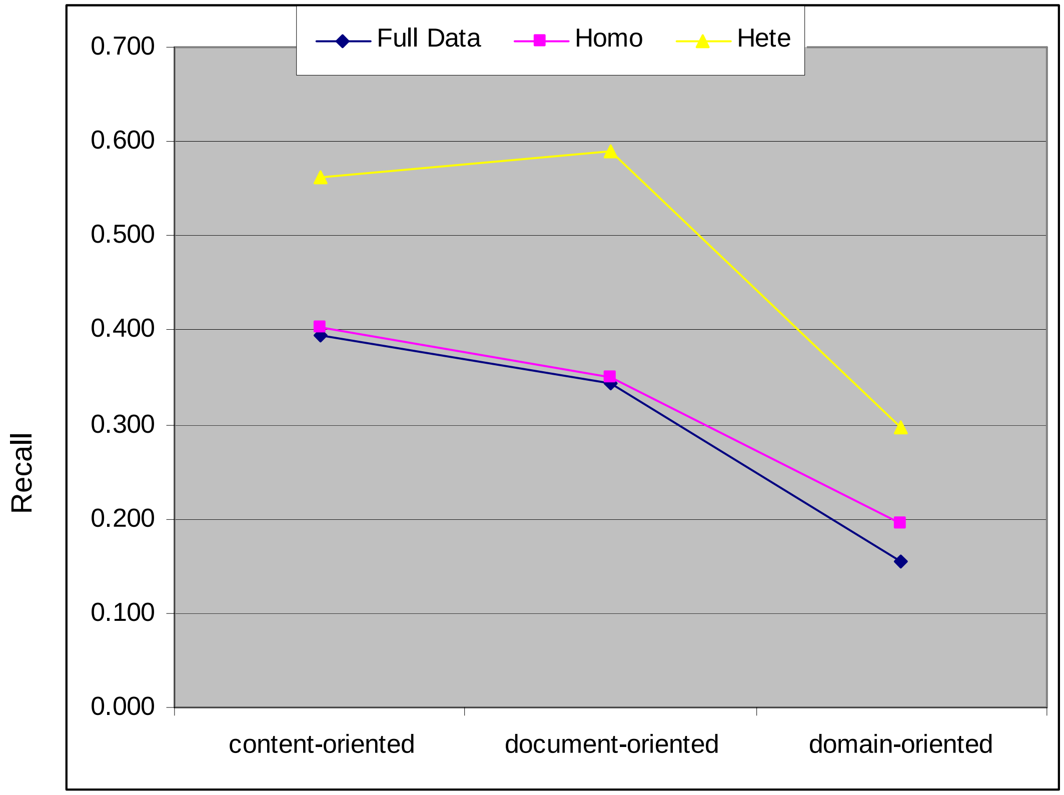


Figure 20. Three indexing conceptions in terms of recall.

Figure 21 shows the F-measure, the balanced average values of precision and recall (refer to Table 6 in Chapter 3). In general, the heterogeneous data set shows better effectiveness than the full data set and the homogeneous data set. This is consistent with measure of recall and precision. While the document-oriented approach in the heterogeneous data set shows the best effectiveness, the content-oriented approach in the full data set and the homogeneous data set shows even greater effectiveness. The behaviors of the full data set and the homogeneous data set are closely related, which is consistent with the results of recall.

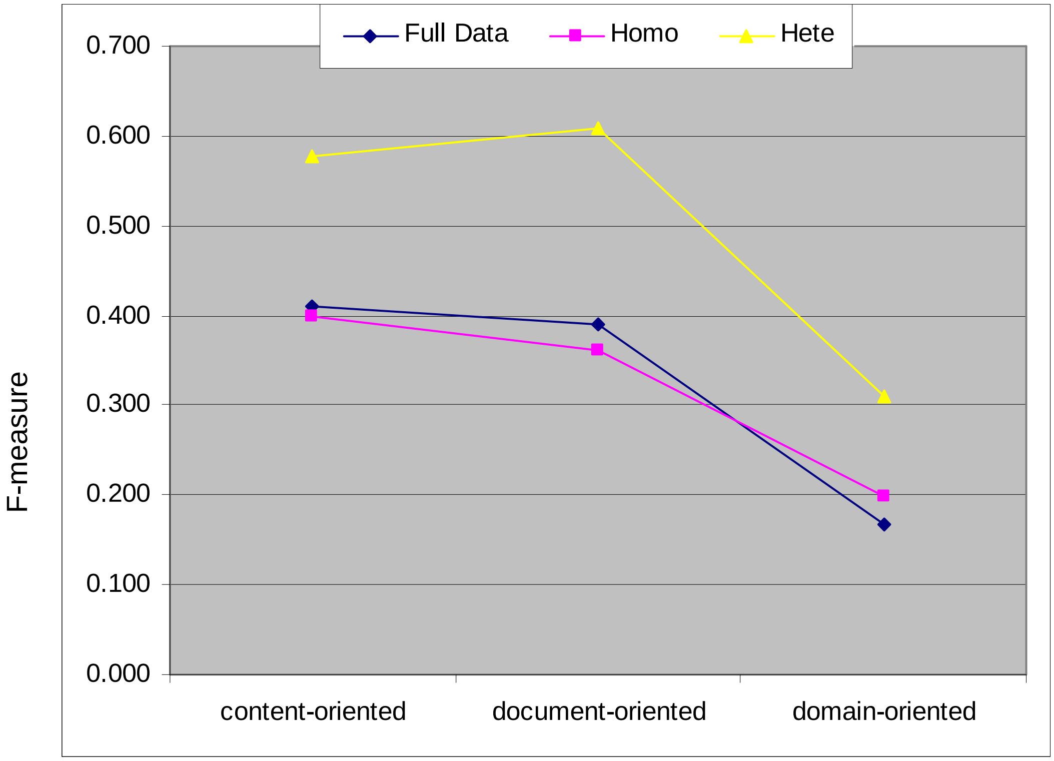


Figure 21. Three indexing conceptions in terms of F-measure.

In sum, the effectiveness of the three indexing conceptions is consistent with respect to different data sets in term of precision, recall, and F-measure. The heterogeneous data set shows better effectiveness than the homogeneous and the full data sets. From the perspective of indexing conceptions, the content-oriented and the document-oriented indicate better effectiveness than the domain-oriented indexing conceptions within the context of three data sets. In addition, the homogeneous data set and the full data set are closely related in terms of results.

Comparisons of each of indexing conceptions

The *t*-tests of three pairs, the content-oriented vs. the document-oriented, the document-oriented vs. the domain-oriented, and the content-oriented vs. the domain-oriented, were conducted to see if there were statistically significant differences in effectiveness. The experiment results are shown in Table 21, Table 22, and Table 23.

Table 21

t-tests between each of the indexing conceptions in terms of F-measure

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	P	t	p
Content vs. Document	1.368	.187	1.401	.195	-1.225	.252
Content vs. Domain	8.642	**.000***	4.003	**.003***	7.869	**.000***
Document vs. Domain	8.825	**.000***	3.902	**.004***	10.871	**.000***

*p<.05

In terms of F-measure as shown in Table 21, all of three data sets indicate that there are significant differences between the content-oriented and the document-oriented approaches showing *p* values less than .05. However, there is no significant difference between the content-oriented and the document-oriented approaches indicating no *p* values less than .05.

Table 22

t-*tests between each of the indexing conceptions in terms of Precision*

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	T	p	t	p
Content vs. Document	2.836	.011	-.478	.644	-1.335	.215
Content vs. Domain	4.863	**.000***	2.325	**.045***	5.261	**.001***
Document vs. Domain	-3.111	**.006***	2.261	**.050***	5.415	**.000***

*p<.05

Table 22 presents the precision measure among the three data sets. As with the results of the F-measure, all three data sets indicate that there are significant differences between the content-oriented and the document-oriented approaches. In addition, no significant difference was found between the content-oriented and the document-oriented approaches.

Table 23

t-*tests between each of indexing conceptions in Recall*

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	T	p	t	p
Content vs. Document	1.884	.075	1.374	.203	-.872	.406
Content vs. Domain	4.819	**.000**	2.158	.059	5.992	**.000***
Document vs. Domain	4.025	**.001**	1.811	.104	9.766	**.000***

*p<.05

On the other hand, recall measures (Table 23) show slightly different results when compared to F-measure and precision. The full data set and the heterogeneous data set present consistent results with F-measure (Table 21) and precision (Table 22). However, no significant differences are found between the three pairs of *t*-tests. The homogeneous data set does not distinguish the three indexing conception-based approaches in terms of significant differences.

In sum, the domain-oriented indexing conception is significantly different with the content-oriented and the document-oriented indexing conceptions within the context of the three data sets. When taking into consideration the nominal numbers of effectiveness using the three indexing conceptions, the effectiveness of the domain-oriented indexing approach is regarded less effective than either the document-oriented or the content-oriented indexing conceptions.

Comparisons of indexing conceptions with the baseline

In this experiment, three indexing conception-based approaches were tested and compared with the baseline (full text) in terms of precision, recall, and F-measure, in order to investigate the differences between the indexing conceptions and the baseline.

As shown in Table 24, Table 25, and Table 26, the experiments were conducted for the full data set, the homogeneous data set, and the heterogeneous data set, respectively. For this purpose, three paired *t*-tests were applied.

Table 24

t-*tests between the baseline and each of indexing conceptions in terms of F-measure*

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Content-oriented	5.492	**.000***	5.552	**.000***	1.822	.102
Document-oriented	4.266	**.000***	4.947	**.001***	2.310	**.046***
Domain-oriented	-5.505	**.000***	-1.346	.211	-5.791	**.000***

*p<.05

Table 24 presents *t*-tests results comparing the F-measures between the baseline and each of the indexing conceptions for the full data set, the homogeneous data set, and the heterogeneous data set. The full data set shows that there are significant differences between each of the three indexing conceptions and the baseline. While there are significant differences between both the content-oriented and the document-oriented approaches and the baseline in the homogeneous data set, significant differences are presented between both the document-oriented and the domain-oriented approaches and the baseline in the heterogeneous data set. In terms of the effectiveness values of the three indexing conception and the baseline, the content-oriented and the document-oriented indexing conceptions have positive impact on TC effectiveness, while the domain-oriented indexing conception has a negative impact on TC effectiveness.

Table 25

t-*tests between the baseline and each of indexing conceptions in Precision*

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Content-oriented	-3.575	**.002***	4.623	**.001***	.857	.414
Document-oriented	4.950	**.000***	3.754	**.005***	1.570	.151
Domain-oriented	6.980	**.000***	.735	.481	-2.379	**.041***

*p<.05

Precision measures for the three data sets (Table 25) show that there are significant differences for all three approaches in the full data set, and this is consistent with the F-measure results. The homogeneous data set is also consistent with the F-measure results in that significant differences were found between both the content-oriented and the document-oriented approaches and the baseline. On the other hand, only one significant difference was found between the domain-oriented approach and the baseline. Taking into account the effectiveness of the three indexing conceptions and the baseline, the content-oriented and the document-oriented indexing conceptions are more effective than the baseline, but the domain-oriented indexing conception had a negative impact on the effectiveness of TC.

Table 26

t-*tests between the baseline and each of indexing conceptions in terms of Recall*

Indexing conception	Full data set		Homogeneous data set		Heterogeneous data set	
	t	p	t	p	t	p
Content-oriented	4.550	**.000***	3.935	**.003***	1.493	.170
Document-oriented	2.145	**.045***	4.229	**.002***	1.639	.136
Domain-oriented	-2.708	**.014***	-.663	.524	-3.151	**.012***

*p<.05

As shown in Table 26, recall measures demonstrate there are significant differences for all three approaches in the full data set. The *t*-test results of both the full data set and the homogeneous data set are consistent with F-measure and precision results. The *t*-test results of the heterogeneous data set are constant with precision results. When examining three indexing conceptions, there are significant differences in recall when compared with the baseline. However, the only significant difference in recall measure was found between the domain-oriented indexing approach and the baseline.

In sum, the full data set indicates that the three indexing conceptions are significantly different with the baseline in terms of the three measures. While the content-oriented and the document-oriented indexing conceptions are more effective than the baseline, the domain-oriented indexing conception is less effective. The homogeneous data set demonstrates there are significant differences between the content-oriented and the document-oriented indexing conceptions and the baseline. However, the data set results do not suggest there is any significant difference between the domain-oriented and the baseline. On the other hand, the heterogeneous data set indicates that there are significant differences between the document-oriented indexing conception and the baseline in a positive way, but between the domain-oriented indexing conception and the baseline in a negative way.

These results address Hypothesis 2 and Hypothesis 3, set forth in Chapter 3.

> H2: Automatic subject term assignment via an indexing conception-based framework is more effective than the full text-based approach in terms of recall, precision, and F measure.

> H3: Automatic subject term assignment via three indexing conception-based approaches, content-oriented, document-oriented, and domain-oriented, is effective in terms of recall, precision, and F measure and the effectiveness will differ among the three approaches.

For H2, the results of data analysis partially supported H2 in that an indexing conception framework is effective for automatic subject term assignment. While two indexing conceptions (the content-oriented and the document oriented) were found to be more effective than the baseline, the effectiveness of the domain-oriented indexing conception was less effective compared to the full text.

In addition, H3 was tested to ascertain if there are differences in effectiveness among the three indexing conceptions. From the data analysis results, H3 is supported by the findings of this study. The three indexing conceptions demonstrated different effectiveness for automatic subject term assignment. The differences between both the content-oriented and the document-oriented and the domain-oriented indexing conceptions were significant, whereas there was no significant difference between the content-oriented and the document-oriented. Taking into account the effectiveness values of the three indexing conceptions, both the content-oriented and the document-oriented indexing conceptions are more effective than the domain-oriented indexing conception. In the context of the data set for this study, the findings of this study suggest that indexing focuses on objective contents and authors' intentions, rather than on possible users' needs. Consequently, according to data set characteristics such as document formats and areas of discipline, identification and utilization of focal indexing conceptions is more desirable.

Summary

This chapter presented the results of data analysis including the effectiveness of semantic sources, the effectiveness of the indexing conception-based framework, and the effectiveness of each of three indexing conceptions. All three research questions and related hypotheses set forth in Chapter 3 were explored. The data set environment was composed of a full data set, a homogeneous data set, and a heterogeneous data set in order to ensure the experiment results were reliable.

Hypothesis 1 addressed whether semantic sources were effective for TC. The results of data analysis for the effectiveness of semantic sources indicate that keyword is more effective for TC than the full text within the context of three data sets. In addition, the data analysis demonstrated that cited works, source title, and title are not significantly different from the full text. Abstract, conclusion, and introduction were significantly different with the full text in a negative way. Taking into consideration computing time and resources, utilization of individual and effective semantic sources can be an efficient alternative for automatic subject term assignment through TC, instead of relying solely on the full text. Consequently, the findings of this study reveal the importance of semantic sources for automatic subject term assignment through TC, which contradicts to the emphasis on utilizing the full text in most TC research.

Hypothesis 2 was tested to see if indexing conception-based approaches as a framework was effective for automatic subject term assignment through TC. The data analysis of the indexing conception-based framework indicated that the content-oriented and the document-oriented indexing conceptions were more effective than utilizing just the full text.

Finally, Hypothesis 3 was tested to see if there were differences in the effectiveness of the three indexing conceptions. In the context of scientific journal article data sets, the content-oriented and the document-oriented indexing approaches are more focal indexing conceptions than is the domain-oriented indexing conception. In other words, indexing conceptions used in this data set are more focused on objective content-oriented and authors' intention-oriented than possible users' needs-oriented approaches. As a result, an improvement in effectiveness was observed when incorporating into Text Categorization the understandings of subject indexing as conducted by human indexers.

CHAPTER 5

DISCUSSION AND IMPLICATIONS

Introduction

This study examined whether the understandings of subject indexing processes, as conducted by human indexers, had a positive impact on the effectiveness of automatic subject term assignment through Text Categorization (TC). More specifically, the conjunction of human indexers' subject indexing approaches or conceptions and semantic sources was explored. This chapter presents four sections: Summary of the Findings, Framework Revisited, Implications, and Future Study. The Summary of the Findings presents the abstract of the experiment results and the consequent findings. The Framework Revisited returns to the preliminary framework in order to reflect on the findings of this study. Implication's section suggests theoretical and practical suggestions and recommendations of this study. The final section presents three research directions for future exploration.

Summary of the Findings

This study proposed an indexing conception-based framework based on the premise that subject indexing conceptions in conjunction with semantic sources are important for automatic subject term assignment through TC. For the purpose of this study, three research questions and related hypotheses were tested: 1) the effectiveness of semantic sources, 2) the effectiveness of an indexing conception-based framework, and 3) the effectiveness of each of three indexing conception-based approaches -- the content-oriented, the document-oriented, and the domain-oriented approaches.

First, semantic sources were defined as attributes of documents to which indexers refer while indexing the subject matters of documents. Various document attributes such as title, keyword, abstract, citation and specific parts of the full text were considered as semantic sources. For a typical scientific journal article data set, eight semantic sources were identified: abstract, cited works, conclusion, full text, introduction, keyword, source title, and title. The identified semantic sources in the context of three types of data sets (the full data, the homogeneous data set, and the heterogeneous data set) were utilized for automatic subject term assignment through TC. The experiment results indicate keyword is more effective as a semantic source than the full text, while cited works, source title, and title are just as effective as the full text. Consequently, utilizing individual semantic sources for automatic subject term assignment has practical benefits in terms of computing time and resources in contrast to the time and expenses associated with utilizing just the full text for TC.

Secondly, the research revealed that an indexing conception-based framework is more effective than the full text for automatic subject term assignment through TC. More specifically, the content-oriented and the document-oriented indexing conceptions in the proposed framework are more effective than the full text. Since indexing conceptions utilize small portions of the full text or document attributes, utilization of indexing conceptions has practical benefits in terms of computing time and resources in contrast to the time and expenses associated with utilizing the full text for TC.

Finally, it was found that the content-oriented and the document-oriented indexing conceptions are more effective than the domain-oriented indexing conception. In other words, in the context of the scientific journal article data set of this study, the objective content-oriented indexing conception and the authors' intentions-oriented indexing conception are considered more effective than the possible users' needs-oriented indexing conception. These findings can be explained as a consequence of the types of data sets. For example, the influence that physical types of documents such as monographs and journal articles have on the focus of the indexing approaches. In addition, the disciplinary areas such as Science, Technology, the Humanities, and Social Science have an effect on different weights of the three indexing conceptions. Since the data set for this study is composed of typical scientific journal articles, the objective contents and authors' intentions are identified as more effective than possible users' needs.

Framework Revisited

The preliminary framework described in Chapter 2 was proposed to incorporate human indexers' indexing approaches into the automatic subject term assignment process through TC. Based on the findings and the experiment results of this study presented in Chapter 4, the preliminary framework was revised and presented in Figure 22.

Based on the support of the experiment results and findings (Figure 22), individual semantic sources support the framework, thereby, improving the effectiveness of TC. Semantic sources such as abstract, cited works, conclusion, introduction, keyword, title, and source title are identified in the context of typical scientific journal articles. While keyword was shown to be more effective than the full text, it was found that cited works, title, and source title were as effective as the full text. In terms of practical benefits such as computing time and resources, it is desirable to use individual semantic sources over the full text. In contrast, Chapter 3 demonstrated that the three indexing conceptions were found to incorporate the understandings of human indexers' indexing practices into TC. The results of the experiment reveal that that the content-oriented and the document-oriented indexing conceptions should be integrated in TC because they both performed better. As a result, within the scientific journal article data sets, indexing conceptions that orient around objective content or authors' intention are more important than possible users' needs.

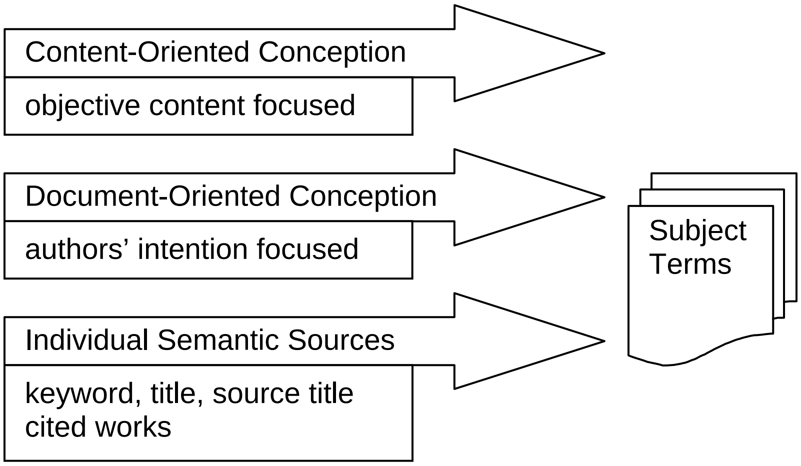

Content-Oriented Conception
objective content focused
Document-Oriented Conception
authors' intention focused
Individual Semantic Sources
keyword, title, source title
cited works
Subject Terms

Figure 22. A revised framework for automatic subject term assignment through TC.

Implications

The main purpose of this study was to investigate whether human indexers' subject indexing approaches in conjunction with corresponding semantic sources are effective for automatic subject term assignment through text categorization (TC). The research findings in this study have implications from both theoretical and practical perspectives.

In terms of theoretical implications, the findings of this study that those who employ TC should have a strong understanding of subject indexing as performed by human indexers, particularly when utilizing TC to improve the effectiveness of subject term assignment. More specifically, the subject indexing approaches or conceptions used by human indexers' during subject analysis (e.g., subject determination, and subject term assignment processes) are very effective for TC. In the context of typical scientific journal article data sets, the findings of this study indicate that the content-oriented and the document-oriented indexing conceptions are more effective than the full text. In a sense, subject indexing of scientific journal articles has focused on the objective contents of a document and authors' intentions, rather than possible users' needs. This study suggests that the paradigm of TC research should be changed accordingly. Currently, TC research has focused on the statistical and probabilistic foundations utilizing the full text to improve the effectiveness of automatic subject term assignment. However, this study shed light on TC from the perspective of subject indexing conducted by human indexers. In this sense, the findings of this study have significant implications for a new theoretical approach to automatic subject term assignment through TC.

From the practical implications perspective, the findings of this study provide a framework for TC system designers. Based on the availability and the characteristics of specific collections or data sets, system designers of TC are able to choose various semantic sources and indexing conceptions by applying them to specific system requirements. In addition, considering various weights of the three measures depending on the domain areas, the findings of this study provide the flexibility to select semantic sources and indexing conceptions in terms of the three measures.

Future Study

This study examined whether understandings of subject indexing conducted by human indexers can be utilized to improve the effectiveness of automatic subject term assignment through TC. The results of this study indicate that inherent indexing conceptions of human indexers in conjunction with semantic sources are effective for TC. In the context of scientific journal article data sets, it was found that the content-oriented and the document-oriented indexing conception were more effective than the domain-oriented indexing conceptions.

For the future, the research using the proposed framework can take three directions. One direction involves exploring the diverse types of information entities. This study focused on textual information entities, but other information entities such as images, video or audio information entities with associated textual information are good candidates for the future study. Another direction involves focusing a user-oriented approach, rather than an indexer-oriented approach. The current study emphasized the utilization of an indexer-oriented approach for assigning subject terms to the documents, but it is worthwhile to examine the experiments using user-provided information for the framework. A third direction to explore for the future is to incorporate different types of applications into the proposed framework. The current study focused on automatic subject term assignment through TC, but automatic metadata generation systems and recommending systems can be explored for the future.

Summary

This chapter presented a summary of experiment results and research findings. The research findings support that incorporation of human indexers' indexing approaches with semantic sources has a positive impact on the effectiveness of automatic subject term assignment. In particular, within the context of scientific journal article data sets, the content-oriented and the document-oriented indexing conceptions were more effective than the full text. The research findings of this study have both theoretical and practical implications for TC research and system development communities. Research directions to be explored for the future include utilizing diverse information entities, user-provided information, and incorporating various applications.

APPENDIX A

CANDIDATE TERMS AND THE NUMBER OF RETRIEVED RECORDS FROM THE INSPEC® DATABASE

1. Possible top terms with similar hierarchical depths for a heterogeneous set

Term	No. of records
computer architecture	22,572
computer graphics	42,473
computer interfaces	33,789
data handling	9,449
discrete systems	27,971
Information management	28,278
knowledge based systems	23,977
pattern recognition	34,923
personal computing	8,694
programming	192,921
reliability	75,614
software engineering	35,034
user interfaces	32,786

2. Possible terms for a homogeneous set (the same hierarchical subject terms under "software engineering" top term)

Term	No. of records
computer aided software engineering	3,876
formal specification	23,536
formal verification	11,716
programming environments	10,655
project support environments	923
software architecture	8,152
software cost estimation	1,356
software development management	6,066
software libraries	5,605
software maintenance	6,139
software metrics	4,125
software performance evaluation	9,943
software portability	4,096
software process improvement	1,480
software prototyping	4,102
software quality	6,989
software reliability	7,163
software reusability	7,709
software tools	19,915
Vienna development method	156

APPENDIX B

A TYPICAL BIBLIOGRAPHY INFORMATION OF INSPEC® ARTICLES

Accession number
Title
Authors
Author affiliation
Serial title
Abbreviated serial title
Volume
Issue
Publication date
Pages
Language
ISSN
CODEN
Document type
Publisher
Country of publication
Material Identity number
Abstract
Number of references
Inspect controlled terms
Uncontrolled terms
Inspec classification codes
Treatment
Discipline
DOI
Database

APPENDIX C

A PYTHON CODE FOR EXTRACTING SEMANTIC SOURCES AND INDEXING CONCEPTIONS (KEYWORD CASE)

```
from nltk.corpus import stopwords
from nltk.probability import FreqDist
from nltk.tokenizer import *
import os
import re
import string

class inspec_nofull:

    data = []

    def __init__(self):
        clsList=os.listdir('C:\Documents and Settings\echung\My
Documents\Dissertation_Data\\keyword')
        for item in clsList:
            itemList=os.listdir(
                'C:\Documents and Settings\echung\My
Documents\Dissertation_Data\\keyword'+'\\'+item)
            for each in itemList:
                self.data.append(item+'\\'+each)

    def items(self):
        print self.data

    def item(self,c):
        lenStr=len(c)
        firstIndex=0
        lastIndex=0

        for each in self.data:
            if each[0:lenStr] == c:
                firstIndex = self.data.index(each)
                break

        for each in self.data:
            if each[0:lenStr] == c:
                lastIndex +=1

        lastIndex = firstIndex+lastIndex

        subData = self.data[firstIndex:lastIndex]
        subD = tuple(subData)
        return subD

    def read(self, item):
        item.split('\\')
        length = len(item)
        dir = item[0:length-13]
        file = item[length-12:length]

        input = open('C:\Documents and Settings\echung\My
Documents\Dissertation_Data\\keyword'+'\\'+dir+'\\'+file, 'r')
        line = input.read()

        words = line.split()

        return words
        text_token = Token(TEXT=line)
        WhitespaceTokenizer(SUBTOKENS='WORDS').tokenize(text_token)
        return text_token

############################
###   GLOBAL VARIABLES     ###
############################
stopwordsDict  = {}   # list of stopwords
featurePattern = ''   # how a potential feature should look like
TEST_CNT       = 2    # number of documents left for testing
x=inspec_nofull()

##############################################
###   NAME    : pull_att()                                  ###
##############################################

def pull_keyword():
    clsSet = ['computer_architecture', 'computer_graphics', 'computer_interfaces',
'discrete_systems',
    'information_management', 'knowledge_based_systems', 'pattern_recognition',
'reliability', 'software_architecture',
    'software_development_management', 'software_engineering', 'software_libraries',
'software_maintenance', 'software_metrics',
    'software_portability', 'software_prototyping', 'software_quality', 'software_reliability',
'software_reusability',
    'user_interfaces']

    for cls in clsSet:
           dirName = cls
           os.mkdir ('C:\Documents and Settings\echung\My
Documents\Dissertation_Data\_keyword'+'\\'+dirName)
        clsDocs = x.item(cls)

        for each in clsDocs:
```

```
    token=x.read(each)
    ind = each.find('\\')
    length = len(each)
    dirName = each[0:length-13]
    fileName = each[length-12:length]

    output = open('C:\Documents and Settings\echung\My
Documents\Dissertation_Data\_keyword'+'\\'+dirName+'\\'+fileName, "w")

        keywordstartInd=0
        keywordendInd=0

        for item in token:
                if item == 'Uncontrolled':
                    keywordstartInd=token.index(item)+2
                index=token.index(item)+1
                length = len(token[index])
                output.write(token[index][6:length])
                output.write(" ")
                    for it in token:
                  if it == 'classification':
                    next = token.index(it)+1
                    if token[next][0:6] == 'codes:':
                            keywordendInd=token.index(it)-2
keywordendInd

      while keywordstartInd <= keywordendInd:
            output.write(token[keywordstartInd])
            output.write(" ")
            keywordstartInd+=1
```

APPENDIX D

A PYTHON CODE FOR CONVERTING TO WEKA INPUT FORMAT (KEYWORD CASE)

```
from nltk.corpus import stopwords
from nltk.probability import FreqDist
from nltk.tokenizer import *
import os
import re
import string


class inspec_nofull:

    data = []

    def __init__(self):
        clsList=os.listdir(
            'C:\Documents and Settings\echung\My
Documents\Dissertation_Data\_source_title')
        for item in clsList:
            itemList=os.listdir(
                'C:\Documents and Settings\echung\My
Documents\Dissertation_Data\_source_title'+'\\'+item)
            for each in itemList:
                self.data.append(item+'\\'+each)

    def items(self):
        print self.data

    def item(self,c):
        lenStr=len(c)
        firstIndex=0
        lastIndex=0

        for each in self.data:
            if each[0:lenStr] == c:
                firstIndex = self.data.index(each)
                break

        for each in self.data:
            if each[0:lenStr] == c:
                lastIndex +=1

        lastIndex = firstIndex+lastIndex

        subData = self.data[firstIndex:lastIndex]
        subD = tuple(subData)
        return subD
    def read(self, item):

        ind = item.find('\\')
        dir = item[0:ind]
        ind +=1
        file = item[ind:]
        input = open('C:\Documents and Settings\echung\My
Documents\Dissertation_Data\_source_title'+'//'+dir+'//'+file, 'r')
        line = input.read()
        text_token = Token(TEXT=line)
        WhitespaceTokenizer(SUBTOKENS='WORDS').tokenize(text_token)
        return text_token

############################
###   GLOBAL VARIABLES   ###
############################
stopwordsDict  = {}
featurePattern = ''
TEST_CNT       = 2
x=inspec_nofull()


def build_stoplist():
    stopwordsDict = {}
    for stopword in stopwords.read('english')['WORDS']:
        stopwordsDict[stopword['TEXT']] = 0
    return stopwordsDict

def is_feature_good (candidateFeature):
    if featurePattern.match(candidateFeature):
        if stopwordsDict.get(candidateFeature.lower(),1) > 0:
            return 1
    return 0


def extract_features_and_freqs(tokens, convertToLowerCase):
    features = {}
    for token in tokens['WORDS']:
        tokenText = token['TEXT']
        if convertToLowerCase:
            tokenText = tokenText.lower()
        if is_feature_good(tokenText):
            features[tokenText] = features.get(tokenText,0) + 1
    return features
```

```
def extract_features_and_freqs_forall(classes, convertToLowerCase):
    globalFeatureFreq = {}
    for newsgroup in classes:
        for item in classes[newsgroup]:
            tokens = x.read(item)
            featureFreq = extract_features_and_freqs(tokens,
convertToLowerCase)
            for feature in featureFreq:
                globalFeatureFreq[feature] =
globalFeatureFreq.get(feature,0) + featureFreq[feature]
    return globalFeatureFreq


def filter_infrequent_features (featureDict, minFeatureFreq):
    newDict = {}
    for feature in featureDict:
        if featureDict[feature] >= minFeatureFreq:
            newDict[feature] = featureDict[feature]
    return newDict


def write_WEKA_input(fileName, featureDict, relationName, classes, writeSparseARFF,
convertToLowerCase):

    sortedFeatures = featureDict.keys()
    sortedFeatures.sort()

    outFile = open(fileName, "w")
    outFile.write("@RELATION " + relationName + "\n\n")
    for feature in sortedFeatures:
        outFile.write("@ATTRIBUTE\t" + feature + "\tNUMERIC\n")
    outFile.write("@ATTRIBUTE\tclass\t{" + string.join(classes,', ') + "}\n")
    outFile.write("\n@DATA\n\n")
    for newsgroup in classes:
        for item in classes[newsgroup]:
            tokens = x.read(item)
            freqs = extract_features_and_freqs(tokens, convertToLowerCase)
            if writeSparseARFF:
                outFile.write('{')
                featIndex = 0
                for feature in sortedFeatures:
        if freqs.get(feature,0) > 0:
                        outFile.write(str(featIndex) + " " +
str(freqs[feature]) + ",")
                    featIndex = featIndex + 1
                outFile.write(str(featIndex) + " " + newsgroup + "}\n");
            else:
                for feature in sortedFeatures:
        outFile.write(str(freqs[feature]) + ",")
                outFile.write(newsgroup + "\n");
    outFile.close()


def write_ARFF (minFeatureFreq,
        removeStopWords,
        featurePattrn,
        convertToLowerCase,
        writeSparseARFF,
        arffRelationName,
        clsTraining,
        outputFileNameTrain,
        clsTesting = [],
        outputFileNameTest = []):
    global featurePattern
    global stopwordsDict
    featurePattern = re.compile(featurePattrn)
    if removeStopWords: stopwordsDict = build_stoplist()
    featureDictTrain = extract_features_and_freqs_forall(clsTraining,
convertToLowerCase)
    featureDictFilteredTrain = filter_infrequent_features(featureDictTrain,
minFeatureFreq)
    write_WEKA_input(outputFileNameTrain, featureDictFilteredTrain,
arffRelationName, clsTraining, writeSparseARFF, convertToLowerCase)

    if clsTesting != []:
        write_WEKA_input(outputFileNameTest, featureDictFilteredTrain,
arffRelationName, clsTesting, writeSparseARFF, convertToLowerCase)


def get_classes_all (clsSet):
    clsTraining = {}   # training sets
    clsTesting  = {}   # testing sets
    for cls in clsSet:
    clsDocs = x.item(cls)
    clsDocsCnt = len(clsDocs)
     trainCnt = len(clsDocs) - TEST_CNT
     clsTraining[cls] = clsDocs[0:trainCnt]
     clsTesting[cls]  = clsDocs[trainCnt:(trainCnt+TEST_CNT)]
    return (clsTraining, clsTesting)

def train_test ():
        homogeneousSet = ['computer_architecture', 'computer_graphics',
'computer_interfaces', 'discrete_systems',
   'information_management', 'knowledge_based_systems', 'pattern_recognition',
'reliability', 'software_architecture',
   'software_development_management', 'software_engineering', 'software_libraries',
'software_maintenance', 'software_metrics',
   'software_portability', 'software_prototyping', 'software_quality', 'software_reliability',
'software_reusability',
   'user_interfaces']

        (clsTraining, clsTesting) = get_classes_all(homogeneousSet)

        write_ARFF (5,                                          1,
                        "^[a-zA-Z]+$",                          1,
                        1,
                        "Keyword",
                        clsTraining,
                        "C:\Documents and Settings\echung\My
                Documents\Dissertation_Data\WEKA_source_title\\Source_Title_train.arff",
                        clsTesting,
                        "C:\Documents and Settings\echung\My
Documents\Dissertation_Data\WEKA_source_title\\Source_Title_test.arff")
```

REFERENCES

Albrechtsen, H. (1992). PRESS: A Thesaurus-based Information system for software reuse. In N. J. Williamson & M. Hudon. (Eds.), *Classification research for knowledge representation and organization* (pp.137-144). Amsterdam: Elsevier Science Publishers.

Albrechtsen, H. (1993). Subject analysis and indexing: from automated indexing to domain analysis. *The Indexer, 18*(4), 219-224.

Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers.

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models. *Proceedings of ICML-02, 19th Conference on Machine Learning, Workshop on Text Learning*, Sydney, Australia.

Buckland, M. (1997). What is a “document”? *Journal of the American Society for Information Science, 48*(9), 804-809.

Calvo, R. A., Lee, J., & Li, X. (2004). Managing content with automatic document classification. *Journal of Digital Information, 52*(2). Retrieved November 14, 2006. from http://jodi.tamu.edu/Articles/v05/i02/Calvo/

Chan, L.M. (1981). *Cataloging and classification: An introduction*. New York City, NY: McGraw-Hill.

Chan, L.M. (1987). Instructional materials used in teaching cataloging and classification. *Cataloging and Classification. 7,* 131-144.

Chu, C.M. & O'Brien, A. (1993). Subject analysis: The critical first stage in indexing. *Journal of Information Science. 19*, 439-454.

Chung, E. & Hastings, S.K. (2006). A conception-based approach to automatic subject term assignment for scientific journal articles. *The Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology.*

Cunningham, S.J., Witten, I. H., & Littin, J. (1999). Applications of machine learning in information retrieval. *Annual Review of Information Science and Technology, 34,* 341-384.

Dewey, M. (2003). *Dewey decimal classification and relative index*. Edited by Joan S. Mitchell [et. al]. Dublin, OH: OCLC Online Computer Library Center, Inc.

Diaz, I., Ranilla, J., Montanes, E., Fernandez, J., & Combarro, E. (2004). Improving performance of text categorization by combining filtering and support vector machines. *Journal of the American Society for Information Science and Technology, 55*(7), 579-592.

Efron, M., Marchionini, G., Elsas, J., & Zhang, J. (2004). Machine learning for information architecture in a large governmental Website. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries,* 151-159.

ENGINEERING VILLAGE TM 2. (n.d.). Retrieved November 11, 2006, from http://www.engineeringvillage2.org/controller/servlet/Controller?EISESSION=1_12bf89210ebdc84b45369fses2&CID=quickSearch&database=1.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, *45*(8), 572-576.

Foskett, A.C. (1996). *The subject approach to information*. London: Library Association Publishing.

Hjørland, B. (1992). The concept of ‘subject’ in information science. *Journal of Documentation, 48*(2), 172-200.

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content… and relevance. *Journal of the American Society for Information Science and Technology, 52*(9), 774-778.

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches-traditional as well as innovative. *Journal of Documentation, 58*(4), 422-462.

Hjørland, B. & Albrechtsen, H. (1995). Toward a new horizon in information science: domain-analysis. *Journal of the American Society for Information Science, 46*(6), 400-425.

Hjørland, B. & Nielsen, L. K. (2001). Subject access points in electronical retrieval. *Annual Review of Information Science and Technology, 35*, 249-298.

Hovi, I. (1988). The cognitive structure of classification work. *Proceedings of 44th FID Conference and Congress*, Finland, 225-236.

INSPEC thesaurus 2004. (2004). London, UK: Institution of Electrical Engineers.

ISO 5963:1985. (1985). *Documentation-methods for examining documents: Determining their subjects and selecting indexing terms*. International Standard Organization.

Jeng, L.H. (1996). Using verbal reports to understand cataloging expertise: Two cases. *Library Resources and Technical Services, 40*(4), 343-358.

Joachims, T. (1998). Text categorization with support vector machine: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning,* 137-142.

Larkey, L. S. (1999). A patent search and classification system. *Proceedings of the 4th ACM conference on digital libraries*, 179-187.

Lewis, D. D. (1992). *Representation and learning in information retrieval*. Unpublished Ph.D. Dissertation, University of Massachusetts, Massachusetts.

Lewis, D. D. (1995). Evaluating and optimizing autonomous text categorization systems. In E.A. Fox, P. Ingwersen, & R. Fidel, *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval,* 246-254.

Lewis, D. D. (2000). Machine learning for text categorization: background and characteristics. *Proceedings of the twenty-first national online meeting*, 221-226.

Mai, J.E. (2000). Deconstructing the indexing process. *Advances in Librarianship. 23*, 269-298.

Mai, J.E. (2005). Analysis in indexing: document and domain centered approaches. *Information Processing and Management, 41*, 599-611.

Miksa, F. (1983). *The subject in the dictionary catalog from Cutter to the present*. Chicago, IL: American Library Association.

Moens, M.F. (2002). *Automatic indexing and abstracting of document texts*. Norwell, MS: Kluwer Academic Publishers.

O’Connor, B. C. (1996). *Explorations in indexing and abstracting: pointing, virtue, and power.* CO: Libraries Unlimited.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program, 14*, 130-137.

Sauperl, A. (2002). *Subject determination during the cataloging process*. Lanham, MD: Scarecrow Press.

Sauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology, 55*(1), 55-63.

Sebastiani, F. (2002). Machine learning in automated categorization. *ACM Computing Surveys, 34*(1), 1-47.

Sebastiani, F. (2005). Text categorization. In A. Zanasi (Eds.), *Text mining and its applications* (pp. 109-129), Southampton, U.K.: WIT Press.

Slattery, S. (2002). *Hypertext categorization*. Unpublished Doctoral Dissertation. School of Computer Science. Carnegie Mellon University. Pittsburgh, PA.

Soergel, D. (1985). *Organizing information: Principles of database and retrieval systems*. NY: Academic Press.

Taylor, A. G. (2003). *The organization of information* (2nd ed.). Englewood, CO: Libraries Unlimited.

Watters, C., Zheng, W., & Milios, E. (2002). Filtering for medical news items. *The Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology,* 284-291.

Weinberg, B.H. (1988). Why indexing fails the researcher. *The Indexer, 16*(1), 3-6.

Wilson, P. (1968). *Two kinds of power: An essay on bibliographic control.* Berkeley, CA: University of California Press.

Witten, I.H. & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with JAVA implementations*. CA: San Diego, Academic Press.

Xu, Z., Yu, K., Tresp, V., Xu, X., & Wang, J. (2003). Representative sampling for text categorization using support vector machine. *Proceedings of 25th European Conference on Information Retrieval Research,* 393-407.

Zhang, B., Goncalves, M. A., Fan, W., Chen, Y., Fox, E.A., Calado, P. Cristo, M. (2004). Combining structural and citation-based evidence for text categorization. *Proceedings of the 13th ACM Conference on Information and Knowledge Management,* 162-163.