A Conception-Based Approach to Automatic Subject Term Assignment for Scientific Journal Articles

EunKyung Chung

School of Library and Information Science, University of North Texas, Denton, TX

echung@lis.admin.unt.edu

Samantha K. Hastings

School of Library and Information Science, University of South Carolina, Columbia, SC

shasting@gwim.sc.edu

This study proposes a conception-based approach to automatic subject term assignment when using Text Classification (TC) techniques. From the perspective of conceptual and theoretical views of subject indexing, this study identifies three conception-based approaches, Domain-Oriented, Document-Oriented, and Content-Oriented, in conjunction with eight semantic sources in typical scientific journal articles. Based on the identification of semantic sources and conception-based approaches, the experiment explores the significance of individual semantic sources and conception-based approaches for the effectiveness of subject term assignment. The results of the experiment demonstrate that some semantic sources and conception-based approaches are better performers than the full text-based approach which has been dominant in TC fields. In fact, this study indicates that subject terms are better assigned by TC techniques when the indexing conceptions are considered in conjunction with semantic sources.

Introduction

Subject indexing through subject term assignment is one of the critical elements to information organization and access, since it provides implicit and connotative access points rather than explicit points provided by keyword-based indexing. In general, the facilitation of implicit subject access to information has been achieved through assigning

subject terms or subject headings to documents in conjunction with appropriate thesauri. Due to the increasing volume of information in digital formats and the perpetual need to organize and give access to information by subject, there have been numerous endeavors to automatically assign subject terms to documents. Text Classification (TC) techniques using supervised learning algorithms have been widely used in various applications to achieve automatic subject term assignment.

However, as Cunningham, Witten, and Littin (1999) pointed out, the models and properties of TC have been approached without reasonably solid conceptual frameworks or backgrounds. Research in TC seems to focus on statistical or probabilistic foundations in terms of document representation, parameter optimizations, and algorithm developments in order to improve the effectiveness, rather than focusing on the understanding of subject indexing frameworks. Although the main function of TC is to identify the patterns and characteristics of a set of documents which are analyzed and then assigned access points by human indexers or subject catalogers, there has been little research reflecting subject indexing frameworks 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 skim texts and then infer the subject terms from specific patterns (Moens, 2002). However, theoretical perspectives and case studies about subject indexing or classification (Foskett, 1996; Hovi, 1988; Jeng, 1996; Mai, 2000; Miksa, 1983) provide some considerations to be reflected in a conceptual framework for TC systems. First, there are specific semantic sources to which human indexers refer in order to capture the subjects of documents, such as titles, keywords, and reference lists. Secondly, some combinations of these document attributes are utilized according to indexers' conceptions of subject indexing (Albrechtsen, 1993; Hjørland, 2002; Mai, 2000; Wilson, 1968). While one of the indexing conceptions emphasizes the objective contents of documents, the other focuses on the author's intentions in creating the documents. In addition, there is an indexing conception to reflect the subject matter of documents within context (Hjørland, 2002; Hjørland & Albrechtsen, 1995; Mai, 2005).

In this sense, the indexing conceptions with corresponding semantic sources are crucial for improving the effectiveness of subject term assignment through TC techniques. By accepting this premise, the purpose of this study investigates these conception-based approaches in conjunction with semantic sources for automatic subject term assignment using TC techniques. In order to accomplish this purpose, this study sets out three objectives: 1) identify semantic sources to which indexers refer during the indexing process, particularly in the indexing of scientific journal articles, 2) identify the conceptions involved in the indexing processes and relate corresponding semantic sources with the conceptions, and 3) evaluate the effectiveness of conception-based approaches compared with the more prevalent full text-based approach.

Subject Term Assignment by Text Classification

Text Classification, also called Text Categorization or topic selecting, explores typical text patterns and characteristics in conjunction with assigned subject terms and then builds a classifier for unknown documents. Since TC utilizes prior human knowledge of subject terms assigned to a certain set of documents (Lewis, 2000), it is well suited to the problem of assigning subject terms to documents. In general, this approach indicates 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 TC procedure consists of training and testing phases. 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 the 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 have been used in TC applications including Neural Networks, Naïve Bayes, Support Vector Machine (SVM), and k-Nearest Neighbors (kNN).

In general, the research involving automatic subject term assignment using TC can be categorized into two types: a generalized approach using the full text and a document-sensitive approach with considering the characteristics and structures of the documents. While the majority of TC research deals with feature selection, document reduction, optimization of specific collections, and effective learning algorithm development from the perspective of a generalized approach (Cunningham, Witten, & Littin, 1999; Sebastiani, 2002; 2005), a document-sensitive approach has emerged with reflecting the understanding of the characteristics and structures of the documents. Consistent with this study, the document-sensitive approach has begun to demonstrate the significance of the various attributes of the documents, instead of focusing on generalized statistical or probabilistic approaches using the full text.

In assigning subject terms to patent documents, Larkey (1999) took into account the significance of document attributes and showed the improvement of effectiveness when using the k-Nearest Neighbor algorithm. Larkey demonstrated that some document attributes such as title, abstract, and the first twenty lines of text characterize the vectors with the best effectiveness instead of the entire documents. In the accuracy measure (described in Section 4.2 Experiment Design), Larkey reported approximately 31% even though with a small amount of training set. On the other hand, in clustering documents by subject headings, Efron, Elsas, Marchionini, and Zhang (2004) demonstrated the effectiveness of document attributes such as keyword and title compared with the full texts of government documents; their results showed 75.08% accuracy in effectiveness

when using SVM. In addition to incorporating document attributes into TC systems, Zhang et al. (2004) included citation information 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. They concluded that the combination of title, abstract, and citation information led to the best results in discovering similar documents and consequently performed well (60.81%) in a test of effectiveness as F measure (described in Section 4.2 Experiment Design). In a more sophisticated approach to subject term assignment, Diaz, Ranilla, Montanes, Fernandez, and Combarro (2004) demonstrated that integrating the contextual information of documents showed improved effectiveness of TC results. When selecting features for document representation, they compared local vocabularies with global vocabularies. The results using local vocabularies showed greater effectiveness than the results using global vocabularies. While local vocabularies refer to the words occurring in documents assigned by specific subject terms, global vocabularies consist of words occurring across all the documents. This study showed that a narrowly defined context of a set of documents can represent more precisely the subjects of a set of documents than a broadly defined context.

While a generalized approach using the full text of documents still makes up the majority of research in TC, a document-sensitive approach have emerged with incorporating the significance of document characteristics and structures into TC systems. Yet, the awareness of the importance of document characteristics and structures reflected in current classification systems is limited to a few document attributes and some degree of contextual information. In fact, this limitation is due to the lack of conceptual understanding underpinning the subject indexing process conducted by human indexers.

Semantic Sources and Conceptions

Semantic Sources for Subject Analysis

From the perspective of the subject term assignment process by human indexers, there are several attributes of documents to be considered for incorporation into the learning process of TC systems. Semantic sources can be defined as document attributes to which human indexers refer in order to analyze the aboutness of a document. Sauperl (2002)'s examination of subject cataloging processes demonstrated that semantic sources such as titles, author's affiliations, and publisher's names led to the determination of subject matter of a document. In addition, Chu and O'Brien (1993) showed the importance of semantic sources for determination of subject matter in the process of subject indexing. In general, there are three types of literature concerning 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 document attributes to use for subject analysis. As Mai (2000) pointed out, the introduction to the Dewey Decimal Classification (DDC, 2004) 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. Consistent with the guidelines provided by the DDC introduction, the ISO standard (ISO, 1985) affirms many of the same semantic sources for subject analysis: title, abstract, list of contents, introduction, illustration/diagram/table with their captions, and words in unusual typeface.

Secondly, textbooks on subject indexing denote semantic sources as well. Originally, Chan (1987) surveyed instructional materials for subject indexing or cataloging. Sauperl (2002) updated new versions of the instructional materials since Chan's survey: Chan (1981), Foskett (1996), and Taylor (2003). Foskett pointed out title, keyword, and citation as subject access points. Moreover, Chan suggested utilizing attributes of a document instead of 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. In addition, Taylor recommended title/subtitle, table of contents, introduction, index terms/words/phrases, and illustrations/diagrams/tables/captions for subject analysis.

Thirdly, subject indexing studies have identified the semantic sources used by indexers or subject catalogers. In order to understand and describe the human indexers or subject catalogers' indexing processes for documents, two types of approaches have been recognized. One attempt is to understand the subject indexing processes using case studies (Jeng, 1996; Sauperl, 2004; Sauperl & Saye, 1998). Jeng related the process of subject catalogers' work to networking and association for subject indexing. Catalogers tend to network and associate semantic sources such as bibliographic information with corresponding subject terms. 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. The other approach of subject indexing studies emphasizes the theoretical perspective of semantic sources for subject indexing (Hjørland, 2001; Mai, 2000). Hjørland identified elements in a typical scientific article for subject indexing as title, abstract, references/citations, full text, and descriptors and

understood them from the theoretical perspective. While Mai investigated the subject indexing process from a semiotic perspective, he identified semantic sources used in classification and indexing for subject indexing.

In sum, identified semantic sources for subject indexing can be categorized as bibliographic information, textual information, and contextual information. The bibliographic information represents value-added elements such as title, abstract, descriptors, and cataloging-in-publication data. While textual information denotes the contents of the full text, some parts of the full text serve to differentiate conceptual parts of the text such as the introduction and the conclusion. An introduction is considered to identify the intention of the authors, and a conclusion is viewed as the objective contents of the document. The contextual information refers to indirect and surrounding information of documents containing information of value in terms of subject indexing. In general, author's affiliation, publisher's name, citation/references, and accompanying materials are considered as semantic sources for contextual information.

Conceptions of Subject Indexing

The conceptions of subject indexing can be defined as the indexers' perceptions or approaches in regard to subject analysis, determination, and indexing. These conceptions of subject indexing have been recognized in various names such as entity-oriented, document-oriented, user-oriented, requirement oriented, content-oriented, and domain-oriented conceptions (Albrechtsen, 1993; Fidel, 1994; Mai 2000; Soergel, 1985; Wilson, 1968). For the purpose of this study, from the perspective of utilizing corresponding semantic sources for each conception, these conceptions for subject indexing can be grouped into three approaches and be named as Content-Oriented, Document-Oriented, and Domain-Oriented.

First, the Content-Oriented conception to subject indexing indicates that subject indexing focuses on the objectivity of subject matters in terms of a prevailing element or a group of 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, instead of extracting keywords from the document. In fact, the Content-Oriented conception lies on the boundary between keyword-based subject indexing and subject term assignment-based indexing. In Wilson's (1968) words, this conception denotes the determination of subject matters based on 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

pointed out that there exists a main element or a group of elements representing the subject matter of a document. In essence, the Content-Oriented conception of subject indexing is an attempt to present the objective subject matters in a document by identifying a dominant element or a group of elements.

Secondly, the Document-Oriented conception endeavors to emphasize the intentions of authors in subject indexing. As Wilson 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 matters for a document. Hjørland (2001) argued that this approach is connected with the theory primarily analyzes the document by studying the author's intentions through the document itself, parts of the document, or related materials for the document. In general, subject indexers are supposed to look for clues from 'introduction' and 'forward' parts in the document in order to identify the author's intentions (Mai, 2000).

Thirdly, there is the Domain-Oriented conception for subject indexing. This conception takes into account users' possible needs and requirements and attempts to incorporate them into subject indexing. 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 information of the document into subject indexing. This conception is consistent with Cooper's (1978) "users' possible utility" and Albrechtsen (1993)'s "requirement-oriented conception". While Cooper's approach attempts to anticipate the user's possible needs for a document, Albrechtsen expresses her approach as making knowledge visible to possible users in the future. In addition, from the perspective of domain analysis for subject indexing, Mai (2005) and Hjørland and Albrechtsen (1995) implied that subject indexing compromises the discourse of a specific document in a context. In this view, the discourse between users and authors in a context can represent the domain of a document. In this way, subject indexing make it possible 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).

For the purpose of this study, the identified conceptions from different perspectives can be viewed with accordingly related semantic sources for automatic subject term assignment using TC techniques. Studies in subject indexing demonstrated that the conceptions such as Content-Oriented, Document-Oriented, and Domain-Oriented, tended to utilize corresponding semantic sources (Hovi, 1988; Jeng, 1996; Sauperl, 2004). Sauperl pointed out that although subject indexers or catalogers are aware of the multiple meanings for different people and different situations, they attempt to limit those meanings within specific boundaries. More specifically, based on the study of 12 professional subject catalogers, Sauperl pointed out that indexers shared the usage of the same semantic

sources based on the indexing conceptions. For instance, while the text itself of the document is generally used for the Content-Oriented conception, references are used for the Domain-Oriented conception. In line of Sauperl's examinations of subject indexers, Hovi demonstrated that indexers and subject catalogers are generally unanimous about the subject matters of documents, although there are differences in representations with respect to subject terms using different controlled vocabularies or thesauri. In addition, Jeng, in her interviews of subject catalogers at the Library of Congress, found that indexers or subject catalogers tend to utilize bibliographic information as a powerful tool for subject cataloging.

An Application to Scientific Journal Articles

For purposes of this empirical experiment, the identified semantic sources and three conceptions in subject indexing will be utilized for a set of scientific journal articles. Typically, a scientific journal article encloses six attributes in 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 also partially to stress one of the conceptions used: the full text, introduction, and conclusion. On the other hand, references of a journal article contain citation information such as author, title, year, source, publisher, etc. Since this study focuses on semantic information from the references rather than citation analysis among 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, keyword, source title, full text, introduction, conclusion, and title of cited works.

Applied to a typical scientific journal article, eight semantic sources are embraced by the three conceptions as shown in Figure 1. In addition, it presents a framework with respect to the three conceptions and the corresponding semantic sources for the experiments of effectiveness. Although the distinction of three conceptions combined with semantic sources are not completely separated from each other, the separation can indicate a way of demonstrating effective approaches to subject indexing with respect to TC. First, in order to obtain the objective subject matters of a document, the Content-Oriented approach mainly considers abstract, conclusion and full text. By nature, an abstract denotes a concise version of the full text (Hjørland & Nielsen, 2001). The conclusion of the full text tends to be a recapitulation of it. It is reasonable that the common characteristics of an abstract and a conclusion are the objective description of the contents in the full text. Secondly, as the Document-Oriented approach is mainly concerned with the intentions of the author, semantic sources such as keywords, title and the introduction part of a document are considered. In general, since keywords in scientific journal articles

are provided by the author(s) of a document when submitting the final draft, it can be considered as an important source to reflect the author(s)'s intentions for the document. On the other hand, the introduction of a document has been recognized as the part reflecting the author's intentions (Wilson, 1968). Although title does not provide comprehensive information, it does contain the intentions of the author when choosing from among many possible alternatives (Hjørland & Nielsen, 2001). Thirdly, the Domain-Oriented approach utilizes source title and title of cited works for subject indexing as one way of incorporating the discourse of a document in a context and then making the document available to possible users' needs. As Hjørland (2002) provides eleven approaches of domain analysis for documents, a bibliometric approach can be considered one way to embrace the discourse surrounding specific documents. On the other hand, from the perspective of subject indexing practices by Sauperl (2004), references indicate the possible users' needs. Therefore, source title and title of cited works are implied as semantic sources emphasizing the Domain-Oriented approach.

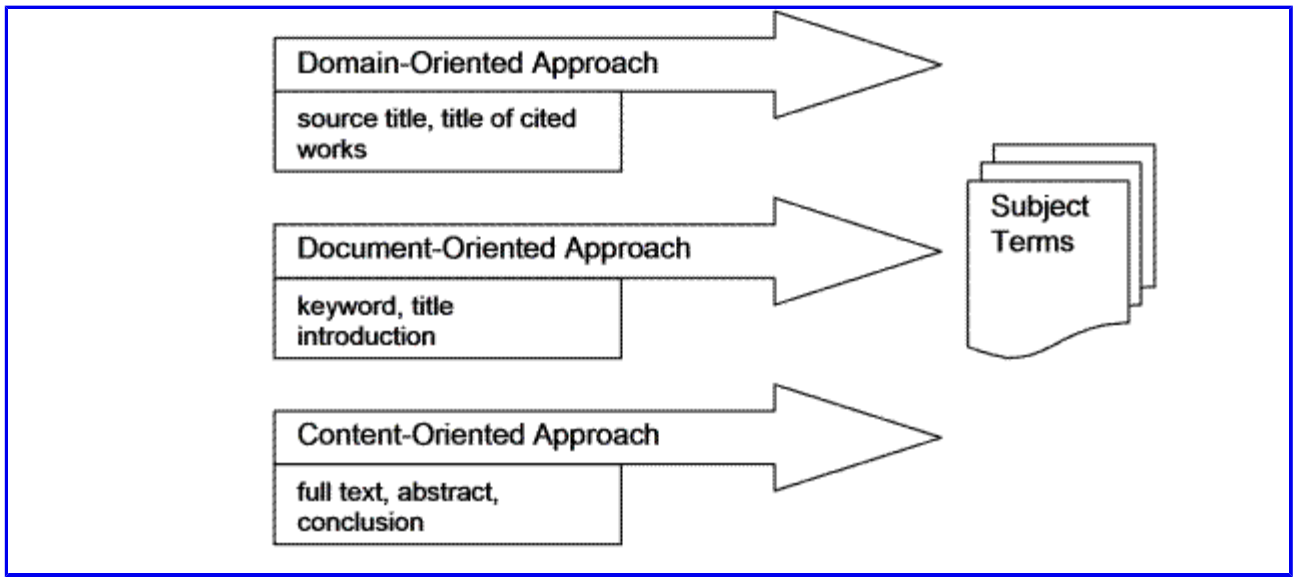

Figure 1. A Conception-based approach with semantic sources to subject term assignment

Research Method

Data Set

Since the purpose of this study is to utilize semantic sources for conception-based approaches, a data set was collected from the INSPEC database during August 2005. 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 (INSPEC, 2005). For this preliminary study, a total of 200 full text articles with bibliographic data were selected according to four subject terms: 'visual databases', 'real-time systems', 'fault

tolerant computing', and 'computational complexity'. The subject terms were chosen based on the balance of the number of corresponding records per subject term within the computer science discipline. This way makes it possible to prevent of one subject term being well-known while other subject terms are unfamiliar to subject indexers. While some articles contain more than one subject term, only the corresponding subject term is considered in order to focus on the purpose of this study and simplify the evaluation process. The four subject terms are from the INSPEC thesaurus which contains controlled vocabulary subject headings. The INSPEC thesaurus is hierarchical in nature and terms are organized by broader (BT: Broader Term), narrower (NT: Narrow Term) or related (RT: Related Term) concepts. The subject terms were assigned to articles in the INSPEC database by human indexers selecting terms from the controlled vocabulary of the INSPEC thesaurus (INSPEC, 2005).

Each article contains bibliographic data such as title, abstract, keyword, and source title in addition to the full text of the article. 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 PDF format to text file format was conducted. Secondly, a semantic source mining procedure was conducted both on the bibliographic information and the full text of the text converted files. Four semantic sources – title, abstract, keyword, and source title – were extracted from the bibliographic information provided by INSPEC. The other four semantic sources – full text, introduction, conclusion, and title of cited works – were extracted from the full text. Whenever 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 parts of each article.

Experiment Design

WEKA (Witten & Frank, 2000), a java-based machine learning implementation, was chosen as a TC system for the experiment with semantic sources and the three approaches because of its reliable effectiveness. Among various learning algorithm implementations, 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). SMO, an SVM implementation in WEKA, was selected for the experiments. For each experiment, the same systematic procedure was followed. Validation methods, stop word removal, stemming, and feature selection for document reduction were fixed for all experiments. 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 used. The method breaks the data into ten equal disjoint subsets and uses one subset as the test data, and the rest as the training data. This is repeated ten times with a different subset as test data for each repetition. The stop word removal procedure was conducted to eliminate unnecessary and insignificant words (e.g., articles) and words were converted to lowercase. In addition, for a term normalization to remove morphological endings from words, Porter's stemming (Porter, 1980) procedure was conducted; this procedure converted plural nouns to singular as well as verbs to their root forms. No feature selection was chosen because this generally deteriorates the effectiveness of SVM. The results of SVM with feature selection were found to deteriorate effectiveness irrespective of the feature selection algorithms used and the chosen reduction factor (Brank, Grobelnik, Milic-Frayling, Mladenic, 2002).

For the quantification of the effectiveness of the semantic sources and the three approaches based on conceptions, the measures of evaluation are defined as shown in Table 1.

Table 1. Contingency table

	Assigned by a human indexer	
Predicted by an automatic classifier	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 TC results (Lewis, 1995; Sebastiani, 2002). The recall refers how good is the classifier at finding positive examples and the precision shows how good are the predictions made by the classifier. While the measure of recall reveals whether the results of trained classifiers are dominated by false positives, precision shows to what extent the results of trained classifiers are subjected by 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 measure F combines the approaches and presents an average of precision and recall. To compute the overall effectiveness of the subject classes, two methods have been primarily used: macroaveraging and

microaveraging. While 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). Since the data set of this study contains a balanced number of documents in each subject class, it seems more reasonable to compute and use macroaveraging for comparison of semantic sources and approaches in the framework.

Experiments

Semantic Sources

While the majority of TC research has focused on utilizing the full text of documents, it is worthwhile to study individual behaviors of semantic sources and to what extent semantic sources are effective for assigning subject terms for documents. With the intention of recognizing the significance of semantic sources, experiments of TC using an individual semantic source one at a time were conducted. The precision, recall, and F measure in each test round were computed as shown in Table 2. The macroaveraged precision, recall, and F measures showed the relative importance of semantic sources for the effectiveness of automatic subject term assignment in the following increasing order: 'conclusion', 'title', 'source title', 'abstract', 'full text', 'introduction', 'title of cited works', and 'keyword'. While 'full text' shows only a moderate result among other semantic sources, 'keyword', 'title of cited works', and 'introduction' each indicate significance for TC. These results are consistent with previously reported results (Larkey, 1999; Zhang et al., 2004) in which better effectivenesss were presented with one or some combination of semantic sources than with the full text.

Table 2. The macroaveraged precision, recall, and F measures

semantic source	precision	recall	F-measure
conclusion	0.63	0.63	0.63
title	0.71	0.65	0.64
source title	0.67	0.67	0.67
abstract	0.69	0.68	0.68
full text	0.71	0.69	0.70
Introduction	0.73	0.72	0.72
title of cited works	0.76	0.73	0.74
keyword	0.82	0.80	0.81

From the results of the eight semantic sources and the characteristics of semantic sources, two comparisons can be noted. One is the comparison between the attributes 'abstract' and 'keyword'. While both 'abstract' and 'keyword' are provided by the authors for representing a concise version of the full text, the difference in effectiveness of 'keyword' and 'abstract' showed a substantial difference. Another comparison can be noted between 'introduction' and 'conclusion'. In general, with almost the same length of data, both were extracted from the full text of the article. However, the results of these two semantic sources again presented a considerable difference in effectiveness. While the effectiveness of 'introduction' shows better effectiveness than 'full text', 'conclusion' is the least effective among the eight semantic sources.

In order to indicate how the effectiveness of semantic sources differs from the effectiveness of 'full text', comparisons were made using t-tests between 'full text' and the remaining semantic sources. The result of TC using 'full text' of documents was selected as the baseline because the majority of current TC research uses full text-based classification. In order to see a significant difference between the baseline and each semantic source, seven pairs of t-tests were applied. Table 3 indicates that while there is no significant differences with the baseline in terms of precision and F-measure, there is a significant difference between 'keyword' and the baseline in recall. In addition, 'title of cited works' presents a nearly significant difference with the baseline. This result indicates that all of the individual semantic sources performed as well as or better than the full text sources in effectiveness in assigning subject terms compared with the full text.

Table 3. *T*-test between each semantic source and baseline

semantic source	precision		recall		F-measure	
	T	*p*	*t*	*p*	*t*	*p*
conclusion	-1.230	0.144	-2.228	0.056	-1.727	0.092
title	-0.041	0.485	-0.408	0.355	-1.739	0.090
source title	-0.529	0.317	-0.305	0.390	-0.461	0.338
abstract	-0.418	0.352	-0.147	0.446	-0.340	0.378
introduction	-0.204	0.426	-0.942	0.208	-0.444	0.344
title of cited works	-0.404	0.357	-2.113	0.063	-0.685	0.272
keyword	-0.971	0.202	-2.828	0.033*	-1.524	0.113

*$p < .05$

Three Conception-Based Approaches

In this experiment, three conception-based approaches were tested and compared with the baseline (full text) in terms of precision, recall, and F-measure. Identified semantic sources for each approach were combined accordingly. For the Domain-Oriented approach, 'source title' and 'title of cited works' were combined for a test. The Document-Oriented approach includes 'introduction', 'title', and 'keyword', and the Content-Oriented approach assembles 'full text', 'conclusion', and 'abstract' for the experiment. Table 4 presents the results of the experiments for the three approaches. With respect to the three measures, the Domain-Oriented and the Document-Oriented approaches present better effectiveness than the Content-Oriented approach. While there are considerable discrepancies between the more effective approaches and the Content-Oriented approach, the difference between the Domain-Oriented and Document-Oriented approach is slight.

Table 4. The microaveraged precision, recall, and F measures for three approaches

approach	precision	recall	F-measure
Domain-Oriented	0.782	0.768	0.773
Document-Oriented	0.790	0.777	0.782
Content-Oriented	0.720	0.697	0.702
Full Text	0.715	0.690	0.696

T-tests between three pairs, Document-Oriented - Domain-Oriented, Document-Oriented - Content-Oriented, and Domain-Oriented - Content-Oriented, were conducted to see if there were significant differences in effectiveness. Table 5 shows that there is a slightly significant difference ($p < .10$) between the Document-Oriented and the Content-Oriented in the recall measure. However, there is no significant difference between the pairs Document-Oriented - Domain-Oriented, and Domain-Oriented - Content-Oriented and no significant difference in the other measures, precision and F-measure.

Table 5. *T*-test between Document - Domain, Document - Content, Domain - Content

approach pair	precision		recall		F-measure	
	T	***p***	***t***	***p***	***t***	***p***
Document – Domain	-.206	.850	-.165	.880	-.190	.861
Document – Content	.556	.617	2.546	.084*	1.087	.357

Domain – Content	.662	.555	1.252	.299	1.312	.281

*$p < .10$

In order to compare the effectiveness of each approach with the baseline, three paired t-tests were applied. Table 6 presents the t-test results between each approach and the baseline in terms of precision, recall, and F-measure. As shown in Table 6, the p-values of the recall measures in Domain-Oriented and Document-Oriented approaches indicate that there are significant differences in the Domain-Oriented approach and the baseline and the Document-Oriented approach and the baseline, respectively.

Table 6. *T*-test between the baseline and three approaches

approach pair	precision		recall		F-measure	
	t	*p*	*t*	*p*	*t*	*p*
Domain-Oriented	0.695	0.269	2.520	0.043*	1.566	0.108
Document-Oriented	0.578	0.302	2.378	0.049*	1.107	0.175
Content-Oriented	0.409	0.355	0.289	0.396	0.398	0.359

*$p < .05$

From the perspective of relative importance among the eight semantic sources and the three approaches, Figure 2 presents a graphical representation in terms of F-measure. With a threshold of .70, two dotted lines shown in Figure 2 confine some of the semantic sources and approaches as a group of better performers for subject term assignment. Among the semantic sources, ‘introduction’, ‘title of cited works’, ‘keyword’ are classified in this group, and the Document-Oriented and the Domain-Oriented approaches are considered in this group as well.

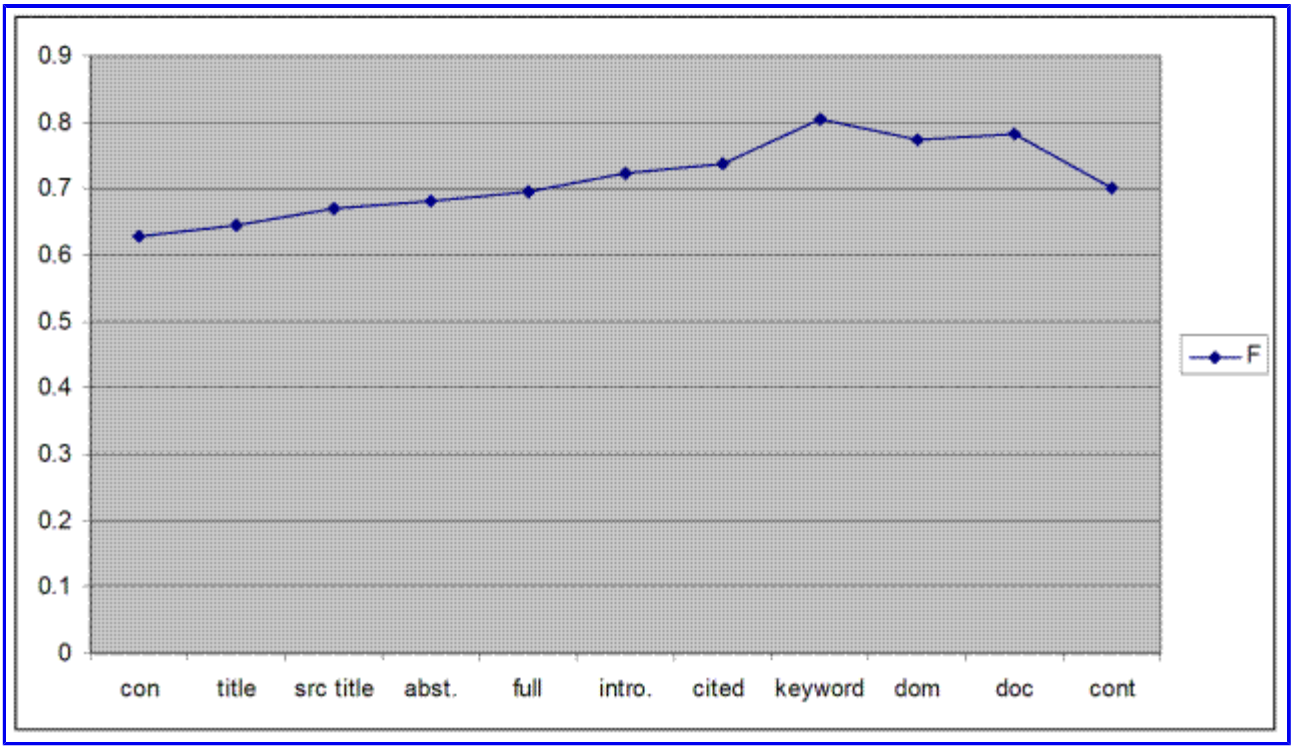

Figure 2. The effectiveness of eight semantic sources and three approaches in F-measure

Conclusion

Based on the premise that subject indexing conceptions in conjunction with semantic sources are important for automatic subject term assignment, this study proposed a conception-based approach. For the purpose of this study, three objectives were set out: 1) identify semantic sources to which indexers refer during the subject term assignment process, 2) identify the conceptions involved in the subject term assignment process and appropriate semantic sources for each conception, and 3) evaluate the effectiveness of conception-based approaches compared with the full text-based approach. 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 are considered as semantic sources. For typical scientific journal articles, eight different semantic sources are identified: title, abstract, keyword, source title, introduction, conclusion, full text, and title of cited works. These semantic sources are diverged into three conception-based approaches: the Domain-Oriented, the Document-Oriented, and the Content-Oriented approaches. While the Domain-Oriented approach uses the combination of the semantic sources 'source title' and 'title of cited works', the Document-Oriented approach uses 'introduction', 'title', and 'keyword', and the Content-Oriented approach utilizes 'full text', 'conclusion', and 'abstract'.

Identified semantic sources in scientific journal articles are utilized for the process of automatic subject term assignment. The full text of documents has traditionally been utilized in TC. However, the results of this study demonstrate that the semantic sources 'keyword', 'title of cited works', and 'introduction' are better performers than 'full text'. These findings indicate the significance of semantic sources and that they can be utilized to improve the effectiveness of TC. Further, three identified conception-based approaches were tested to see the impact of human indexers or subject catalogers' conceptions. Reflections of the authors' intention and contextual understanding of documents, including possible users' needs, are represented in the Document-Oriented and the Domain-Oriented approaches, respectively. The results of conception-based approaches demonstrate that the Document-Oriented and the Domain-Oriented approaches are better performers than the Content-Oriented approach. This indicates that subject terms, products of subject indexing conceptions, can be assigned better by TC techniques when considering the fundamental conceptions in conjunction with semantic sources. Consequently, the findings of this study provide theoretical implications for TC by demonstrating the importance of conception-based approaches with corresponding semantic sources based on the concepts of subject indexing and theoretical views.

Acknowledgements

We would like to express our thanks to Dr. Shawne Miksa for her insightful comments for the earlier versions of this study. We also thank Amy Eklund for editing this paper and the anonymous reviewers for their comments.

References

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 the ICKM-02 Workshop on Text Learning*

Calvo, R. A., Lee, J., & Li, X. (2004) Managing content with automatic document classification *Journal of Digital Information* 52(2)

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

Cooper, W.S. (1978) Indexing documents by gedanken experimentation *Journal of the American Society for Information Science* 29, 107-119

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

DDC. (2004) *Dewey Decimal Classification and Relative Index* Edition 22, 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

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

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

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

ISO (1985) *Documentation-Methods for Examining Documents, Determining Their Subjects and Selecting Indexing Terms* International Standard Organization

INSPEC (2005) *Engineering Village 2* Elsevier Engineering Information Inc., Hoboken, NJ

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 fourth ACM conference on Digital libraries* 179-187

Lewis, D. D. (1995) Evaluating and optimizing autonomous text classification systems *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*

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* Kluwer Academic Publishers

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

Sauperl, A. & Saye, J.D. (1998) Subject determination during cataloging *Proceedings of the 61st American Society of Information Science Annual Meeting*

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

Sebastiani, F. (2005) Text categorization In Alessandro Zanasi (ed.) *Text Mining and its Applications* WIT Press, Southampton, U.K., 109-129

Soergel, D. (1985) *Organizing Information: Principles of Database and Retrieval Systems* NY: Academic Press

Taylor, A. G. (2003) *The Organization of Information* Englewood, CO: Libraries Unlimited

Watters, C., Zheng, W., & Milios, E. (2002) Filtering for medical news items *The Proceedings of the ASIS&T* 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 classification using Support Vector Machine *The 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 classification *Proceedings of the 13th ACM Conference on Information and Knowledge Management* 162-163