A Framework of Automatic Subject Term Assignment for Text Categorization: An Indexing Conception-Based Approach

EunKyung Chung
Ewha Womans University, Library and Information Science, 11–1 Seodaemun-Gu Daehyun-Dong, Seoul, Korea 120–750. E-mail: echung@ewha.ac.kr

Shawne Miksa
University of North Texas, College of Information, Department of Library and Information Sciences, 1155 Union Circle 311068, Denton, TX 76203. E-mail: Shawne.Miksa@unt.edu

Samantha K. Hastings
University of South Carolina, School of Library and Information Science, 1501 Greene Street, Columbia, SC 29208. E-mail: hastings@sc.edu

The purpose of this study 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 dataset. 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, two research questions were explored: To what extent are semantic sources effective? To what extent are indexing conceptions effective? The experiments were conducted using a Support Vector Machine implementation in WEKA (I.H. Witten & E. Frank, 2000). Using F-measure, the experiment results showed that cited works, source title, and title were as effective as the full text while a 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. The content-oriented and the document-oriented indexing approaches especially 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 dataset, the objective contents and authors' intentions were more desirable for automatic subject term assignment via TC than the possible users' needs. The 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.

Received August 28, 2009; revised October 9, 2009; accepted October 12, 2009

 DOI: 10.1002/asi.21272

Introduction

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. 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. However, as Cunningham, Witten, and Littin (1999) noted, the models and properties of TC have been approached without reasonably solid understandings of how human indexers approach subject indexing. More specifically, research in TC focuses on statistical and probabilistic foundations with respect to document representation, parameter optimizations, and algorithm developments 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, most studies do not concentrate on how indexing processes have been conducted by human indexers. One of plausible assumptions used is that "human indexers simply skim texts and then infer the subject terms from specific patterns" (Moens, 2002, p. 111).

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) noted, 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 interwoven 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. One line of research in subject indexing has 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 primarily on the possible users' needs. Another line of research has shown that different sets of document attributes are utilized by indexers depending on the nature of indexers' approaches to subject indexing (Foskett, 1996; Jeng, 1996; Mai, 2000; Miksa, 1983). For example, 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 TC. 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. Based on the general background and problem area, this study is guided by two research questions that address automatic subject term assignment through human indexers' indexing approaches in conjunction with semantic sources. First, this study investigates the significance and characteristics of semantic sources, or document attributes, in terms of improving the effectiveness of TC. Some studies have demonstrated some improvement in effectiveness when weights are assigned to specific document attributes for TC (Diaz, Ranilla, Montanes, Fernandez, & Combarro, 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 attempts to show the characteristics and importance of semantic sources in the context of subject-indexing frameworks for TC. Second, this study investigates the characteristics and importance of an indexing conception-based framework in terms of improving the effectiveness of TC. This study explores the effectiveness of three indexing conception-based approaches as compared to the results of the full-text-based approach. Automatic subject term assignment through TC 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 TC systems. Given the importance of automatic subject term assignment through TC techniques in this digital age, this study is important because of the significance of incorporating an understanding of the characteristics and structures of information entities into the design of TC systems.

Literature Review

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. Since TC in general utilizes prior human knowledge of subject terms assigned to a certain set of documents, it is well suited to the problem of automatic subject term assignment to documents (Lewis, 1992, 2000). In this sense, automatic subject term assignment employs supervised learning algorithms which exploit given subject terms and a set of documents to assign terms to new and unknown documents. While TC research primarily focuses on system-oriented aspects such as feature selection, dimension reduction, optimization of specific collections, and effective learning-algorithm development (Cunningham et al., 1999; Sebastiani, 2002, 2005), there is an emerging line of research which reflects an understanding of the characteristics and structures of the documents in terms of TC applications. In a sense, the line of research reflecting the document characteristics can be stemmed from indexing approaches and conceptions practiced by human indexers. Human indexers who assign subject terms or headings are likely to index target documents utilizing document characteristics with underlying indexing approaches or conceptions. For details, two aspects have been explored: (a) research utilizing document attributes for TC and (b) research utilizing document attributes in conjunction with indexing conceptions.

Document Attributes for Text Categorization

Since patent documents consist of distinguishable document attributes such as claim, purpose, and application, one research area that has employed the intrinsic features and

DOI: 10.1002/asi

characteristics of a document is for patent documents. Several studies have been conducted to classify patent documents using TC applications in terms of improving the effectiveness of TC. Kim and Choi (2007) identified that document attributes unique to patent documents such as claim, purpose, and application showed considerably positive impact on the effectiveness of TC. More specifically, they demonstrated a 74% improvement with patent documents attributes over an approach solely with the main text. Similarly, Koster, Seutter, and Beney (2003) investigated only the abstract of patent documents using the Winnow algorithm. The result of this investigation showed that there was no improvement over the full text of patent documents. On the other hand, Fall, Torcsvari, Benzineb, and Karetka (2003) investigated various parts of patent documents as well as claim, purpose, and title. Similar to the results of Kim and Choi, Fall et al. showed that using document attributes of patents improved the effectiveness of TC compared to using only the main text. In addition, they showed that the results were consistent over various TC algorithms in terms of Support Vector Machine (SVM), Naïve Bayes, k-NN, and SnoW. Larkey (1999) supported that incorporating the document attributes into TC applications had positive impact on the effectiveness of TC. Larkey identified that document attributes such as title, abstract, and the first 20 lines of text were more effective than was the full text of documents. In addition, Larkey reported approximately 31% improved results over the full text, although a small training dataset was used. Larkey noted that this was because document attributes better characterize the vectors in terms of effectiveness measurement.

On the other hand, various types of documents were investigated regarding whether specific attributes unique to the types of documents had a positive impact on the effectiveness of TC. For government documents, Efron et al. (2004) demonstrated that document attributes such as the keyword and title were more effective than the full text. Using Support Vector Machine, this study showed 73% improved effectiveness over the full text. Moreover, for research articles, Zhang et al. (2004) included citation information to discover the most similar documents using the k-NN algorithm. Results of this study demonstrated that the combination of title, abstract, and citation information led to the best results. Consistent with Zhang et al.'s results, Slattery (2002), using SVM, identified that link information in hypertext documents played a role in increasing the effectiveness of TC. Fujino, Ueda, and Saito (2007) explored document attributes such as title, link, and author for Web pages and technical papers in terms of TC applications and concluded that these selected document attributes were effective compared to solely full texts.

Document Attributes in Conjunction With Indexing Conceptions

In general, three indexing conceptions have been discussed in terms of human indexers' approaches when indexing. First, the content-oriented indexing conception focuses on the objectivity of subject matters of document. Albrechtsen (1993) recognized the "content-oriented conception" (p. 220) as the practice of assigning objective subject terms implied by indexers' interpretations of a principle subject element. Wilson (1968) noted that there exists a main element or a group of elements because not all elements demonstrate the same amount of weight in terms of the relative dominance in subject. Accordingly, the document attributes associated with the content-oriented indexing conception can consider a conclusion part in a document, an abstract, and chapter headings in a book. Second, the document-oriented conceptions endeavor 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 authors' intentions for a document are subject matters for a document. To identify the authors' intentions for a document, indexers are supposed to look for document attributes such as the introduction, forward, preface, and title (Mai, 2000). Third, the domain-oriented conception for subject indexing takes into account the domain of knowledge surrounding a document when representing users' possible needs and requirement. To achieve the purpose of the domain-oriented indexing, Mai (2005) and Hjørland and Albrechtsen (1995) implied that subject indexing compromises the discourse of a specific document in 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 particular documents for potential use instead of exclusively focusing on the contents of documents (Blair, 1990; Hjørland, 1992; Weinberg, 1988). Associated document attributes with the domain-oriented indexing conception can be considered references in a document.

According to the supports of document attributes and related fundamental indexing conceptions, a preliminary framework is proposed with the context of typical scientific journal articles and associated document attributes. Among typical document attributes, six attributes relevant to the subject matter are selected as semantic sources: title, abstract, keyword (author-provided), source title (e.g., journal title or conference proceeding title), references, and full text. As shown in Figure 1, these six semantic sources of documents are depicted in conjunction with individual indexing conceptions. First, in terms of the content-oriented indexing conception, abstract, conclusion, and the full text are considered. Second, document attributes such as keyword, title, and introduction are regarded as the document-oriented indexing conception. Third, for the domain-oriented indexing conception, source title and title of cited works are considered.

Research Method

Dataset

The INSPEC database was chosen because it contains bibliographic information, the full text, and subject terms assigned by indexers and because it covers the scientific literature in the fields of electrical engineering, electronics,

DOI: 10.1002/asi

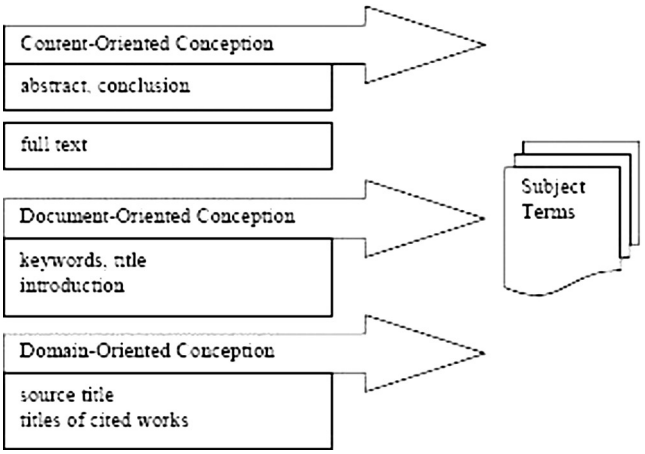

FIG. 1. A preliminary framework of conception-based approaches applied in typical scientific journal articles.

physics, control engineering, information technology, communications, computers, computing, and manufacturing and production engineering (Engineering Village 2, n.d.). This database contains over 8-million bibliographic records that represent 3,500 scientific and technical journals and 1,500 conference proceedings (Engineering Village 2, n.d.). Typical bibliographic records contain 24 elements including title, keywords (uncontrolled terms), abstract, and INSPEC controlled terms. Although not true for all records, most current records are likely to provide links to the full text of the documents. The INSPEC controlled terms are assigned by human indexers after analyzing the contents of each document (Engineering Village 2, n.d.). The INSPEC thesaurus (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.

For this study, 1,000 documents, each with full text and bibliographic information, were collected from the INSPEC database according to specified subject terms. The dataset was then divided into sets of 50, which were then associated with 20 different subject classes. According to previous research, no standard was identified for the number of instances or words per each subject class during the training phase of TC. However, a pilot study (Chung & Hastings, 2006) conducted earlier showed satisfactory results when using 50 documents per subject class; therefore, the same strategy was employed when assigning a subject class to each set.

Additionally, the dataset was divided into two subdatasets: a homogeneous set and a heterogeneous set. While a homogeneous dataset contains more semantically related documents, a heterogeneous dataset 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 subdatasets 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 dataset environments. Furthermore, the experiment results can be validated as to whether the results were consistent with the diverse nature of datasets.

To fulfill the objectives of building a dataset and separating it into two subsets, two considerations were applied: (a) The selected terms in the set have the same or a similar hierarchical depth with respect to the INSPEC thesaurus, and (b) 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. Second, by leveling the numbers of records per subject term, selected terms were evenly familiar to indexers without being too well-known or too underrecognized. In addition, computer science and information technology areas were chosen for collection of a dataset 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 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 dataset, 10 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, 20 narrower terms with the same hierarchical depth (Level 2) were considered candidate subject terms for a homogeneous dataset. However, based on the balance of numbers of records for those 20 terms, 10 of 20 subject terms were selected for the collection of a homogeneous dataset. On the other hand, within the heterogeneous dataset, another 10 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 dataset. Based on the number of records per subject term, 10 of 12 subject terms were selected. Therefore, a total of 20 subject terms are selected for homogeneous and heterogeneous datasets as shown in the Appendix.

Data Preprocessing

The search process for collecting 50 full-text articles for the dataset according to the selected 20 terms was specified as follows.

1. SELECT DATABASE was set as INSPEC database.
2. Twenty subject terms for homogeneous and heterogeneous datasets were typed in SEARCH FOR.

DOI: 10.1002/asi

3. SEARCH IN was specified as the Controlled Term from the drop-down lists.
4. The search process was LIMITED BY the English language.
5. The search process was LIMITED BY Years 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 6 years. A 6-year time span was arbitrarily determined as a sufficient length of time, and it was hypothesized that cataloging/indexing practices, as well as terminology, do not change much in this length of time; 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 to focus on the purpose of this study and to simplify the evaluation process.

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

First, a procedure for converting a PDF format to a text-file format was conducted using an Adobe Acrobat Capture[1] software program. Next, a semantic source mining procedure was conducted both on the bibliographic information and on 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. In instances where no indications or subtitles such as "introduction" and "conclusion" were found, the first or last 50 lines of each full text from the beginning and from the end were used to represent the introduction and the conclusion, respectively, of each article.

After eight semantic sources were constructed using the two procedures, information that was potentially distracting was removed or changed as follows:

(a) Case: Uppercase letters were converted to lower case.
(b) Stopwords:[2]: *the*, *and*, *a*(*n*), *from*, and so on were removed, including punctuation and numbers.
(c) Word normalization:[3] Words are reduced to a standard form which ignores endings for plurals and tense by the Porter stemming algorithm (Porter, 1980).

[1]http://www.adobe.com/products/acrcapture/capfullfeature.html

[2]Used the implementation with a stop words corpus in the Natural Language Toolkit for Python (http://nltk.org)

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

Text Representation

As the raw text of a document cannot be used for TC 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 TC systems including SVM (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. (2002) demonstrated that feature selection tends to be detrimental to the performance of SVM. This leads to the use of Bag of Words representation of the full text. As represented in the Bag of Words, the text was transformed to the WEKA file format (*.arff) for the input of the experiments. This sparse format is able to accelerate the processing time because it skips data with a value of "0."

Text-Categorization System

WEKA (Witten & Frank, 2000), a java-based machine learning implementation, was chosen as a TC 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 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 10-fold cross-validation method, was followed. Since the average classification error over the 10 trials is a good estimate of the overall classification error of the learning method (Watters, Zheng, & Milios, 2002), a 10-fold cross-validation method was chosen. The validation method breaks the data into 10 equal disjointed subsets and uses one subset as the test data and the rest as the training data. This was repeated 10 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 1. These metrics primarily measure how well TC classify given documents.

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

Three effectiveness measures, *recall*, *precision*, and *F*, are common metrics for evaluating TC results (Lewis, 1995;

DOI: 10.1002/asi

TABLE 1. Contingency table.

	Assigned term	
Predicted term	Correct	Incorrect
Correct	*a*	*b*
Incorrect	*c*	*d*

Sebastiani, 2002). The recall refers to the classifier's ability to automatically assign subject terms to the documents among positive examples (i.e., correctly classified instances), and the precision shows the classifier's ability to automatically assign subject terms to the documents among positive and negative examples. 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 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 (harmonic mean) of precision and recall. In addition, to simplify the measures, accuracy also was used. Accuracy refers to the number of correct predictions in the classification results.

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 dataset in this study contains a balanced number of documents ($n = 50$ per subject class), with each class viewed as being equally important. Therefore, it is reasonable to compute and use macroaveraging for comparison of semantic sources and approaches in the proposed framework (Lewis, 1992).

TABLE 2. Macroaveraged precision, recall, and F-measure for the full dataset.

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

TABLE 3. Macroaveraged precision, recall, and F-measure for the homogeneous dataset.

Semantic source	Precision	Recall	F-measure
Abstract	0.283	0.275	0.262
Cited works	0.345	0.310	0.322
Conclusion	0.249	0.244	0.243
Full text	0.287	0.258	0.261
Introduction	0.286	0.234	0.247
Keyword	0.402	0.382	0.384
Source title	0.267	0.269	0.262
Title	0.310	0.292	0.295

TABLE 4. Macroaveraged precision, recall, and F-measure for the heterogeneous dataset.

Semantic source	Precision	Recall	F-measure
Abstract	0.425	0.393	0.389
Cited works	0.493	0.459	0.469
Conclusion	0.371	0.349	0.348
Full text	0.564	0.505	0.520
Introduction	0.444	0.426	0.427
Keyword	0.587	0.568	0.569
Source title	0.461	0.420	0.432
Title	0.487	0.422	0.439

Results

Analysis of Semantic Source Effectiveness

TC experiments using individual semantic sources one at a time were conducted with the intention of recognizing the significance of semantic sources. For the full dataset, the precision, recall, and F-measure in each test round were computed as shown in Table 2. In terms of F-measure, the keyword, source title, title, cited works, and full text show relatively high effectiveness. 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.

For the homogeneous dataset, Table 3 presents the effectiveness of semantic sources in terms of the three measures: precision, recall, and F-measure. The relative effectiveness of full text (0.3 in F-measure) decreases compared to full text (0.261 in F-measure) in the full dataset. While cited works and keyword still demonstrated high effectiveness, the conclusion, introduction, and abstract present low effectiveness among semantic sources. In general, the overall effectiveness of semantic sources is consistent with the effectiveness of the full dataset.

Table 4 presents the effectiveness of each semantic source for the heterogeneous dataset. The effectiveness of semantic sources in this set generally increases compared to the full dataset and the homogeneous dataset. The effectiveness of the full text increases following the effectiveness of keyword when compared to the full dataset and the homogeneous dataset. The semantic sources' effectiveness is consistent with the full dataset and the heterogeneous dataset except for the high effectiveness of the full text in the heterogeneous dataset. While cited works and keyword were considerably effective for TC, semantic sources such as conclusion, abstract, and introduction were less effective.

DOI: 10.1002/asi

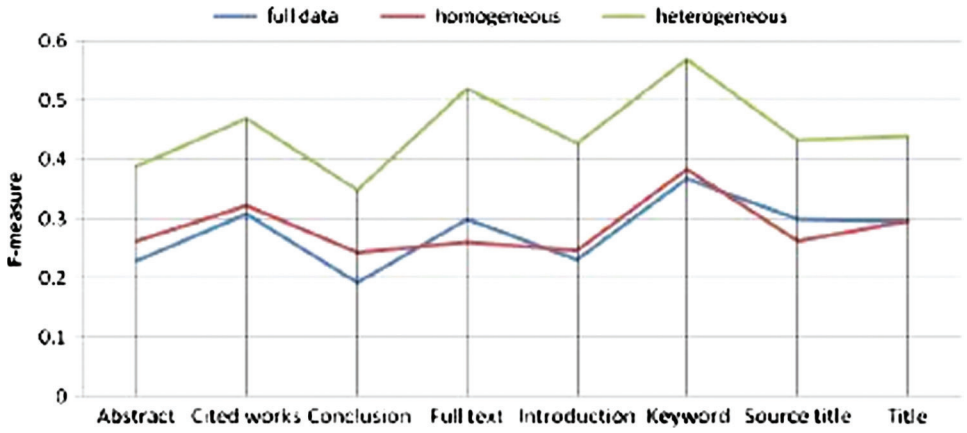

FIG. 2. Semantic sources from three datasets in F-measure.

TABLE 5. *T*-tests between the baseline and individual semantic sources in terms of F-measure.

	Full dataset		Homogeneous dataset		Heterogeneous dataset	
Semantic source	*t*	*p*	*t*	*p*	*t*	*p*
Abstract	3.446	**0.003***	−0.025	0.981	3.273	**0.010***
Cited works	−0.249	0.806	−0.857	0.414	1.252	0.242
Conclusion	4.481	**0.000***	0.500	0.629	4.333	**0.002***
Introduction	2.587	**0.018***	0.310	0.763	2.254	0.051
Keyword	−2.231	**0.038***	−5.212	**0.001***	−0.753	0.471
Source title	0.043	0.966	−0.015	0.988	1.402	0.194
Title	0.110	0.913	−1.267	0.237	1.321	0.219

*$p < 0.05$.

Although some similarities were found across the three datasets in the previous analysis, Figure 2 presents the F-measures to examine the similarity of three datasets more clearly. Figure 2 confirms that the heterogeneous dataset generally shows better effectiveness than the full dataset and the homogeneous dataset both in terms of F-measure. Each of the semantic sources demonstrates nearly the same behaviors among the three datasets. Keyword and cited works were found effective while conclusion and abstract were not as effective as the other semantic sources.

In sum, keyword, cited works, source title, and title show high effectiveness and are fairly consistent. However, two comparisons should be noted when taking into account the effectiveness results of semantic sources. First, there is a substantial gap between the abstract (0.389 in F-measure, Heterogeneous) and keyword (0.569 in F-measure, Heterogeneous) even though both are provided by the authors of the articles and represent a concise version of the full text. Second, there is a considerable difference in the effectiveness results between the introduction (0.427 in F-measure, Heterogeneous) and the conclusion (0.348 in F-measure, Heterogeneous). The introduction shows greater effectiveness than does the conclusion, even though they are extracted from the full text of the articles for different purposes. Specifically, the number of words associated with the introduction is approximately twice as large as the number of words associated with the conclusion.

Comparisons of semantic sources with the baseline. To investigate the differences between individual semantic sources and the baseline, *t* tests of individual semantic sources compared to the baseline (full text) are presented in Table 5 using F-measure. When the alpha value is set to 0.05, using F-measure as shown in Table 5, 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 full dataset and the heterogeneous dataset indicate more distinctive results among semantic sources than does the homogeneous dataset. Keyword, in the homogeneous dataset, is the only one that demonstrates significant differences. In the heterogeneous dataset, 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 datasets, significantly different semantic sources such as abstract, conclusion, introduction, and keyword are summarized as follows. First, keyword shows greater effectiveness when compared to the baseline. Second, title, source title, and cited works show no significant difference when 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. Of seven

DOI: 10.1002/asi

TABLE 6. Macroaveraged precision, recall, and F-measure for the full dataset.

Indexing conception	Precision	Recall	F-measure
Content-oriented	0.450	0.394	0.409
Document-oriented	0.541	0.253	0.312
Domain-oriented	0.270	0.155	0.168

TABLE 7. Macroaveraged precision, recall, and F-measure for the homogeneous dataset.

Indexing conception	Precision	Recall	F-measure
Content-oriented	0.407	0.402	0.399
Document-oriented	0.485	0.173	0.194
Domain-oriented	0.310	0.195	0.199

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 in effectiveness measures were presented with a combination of one or more document attributes rather than just the full text alone. When utilizing individual semantic sources instead of the full text, the data-analysis results indicate that one semantic source (keyword) demonstrated consistently better effectiveness in F-measure 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 only the full text is desirable for automatic subject term assignment through TC, especially when considering the computing time and resources combined with the effectiveness of the individual semantic sources.

Analysis of Three Indexing Conception-Based Approaches

Tables 6, 7, and 8 present the results of the experiments for three approaches in terms of precision, recall, and F measure, respectively, for the full, the homogeneous, and the heterogeneous datasets. With respect to the three measures, in general, the content-oriented and the document-oriented approaches show greater effectiveness when compared to the domain-oriented approach.

In addition, the comparison in Figure 3 is presented to examine whether the behaviors of different datasets are consistent with the results in terms of F-measure. The figure indicates that the heterogeneous dataset showed better effectiveness in F-measure than do the full dataset and the homogeneous dataset. Of the three indexing conceptions, the domain-oriented approach demonstrated less

TABLE 8. Macroaveraged precision, recall, and F-measure for the heterogeneous dataset.

Indexing conception	Precision	Recall	F-measure
Content-oriented	0.623	0.562	0.576
Document-oriented	0.676	0.590	0.608
Domain-oriented	0.412	0.296	0.309

effectiveness compared to the content-oriented and the document-oriented approaches. These results are consistent with three datasets. In sum, the effectiveness of the three indexing conceptions is consistent with respect to different datasets in term of precision, recall, and F-measure. The heterogeneous dataset shows better effectiveness in F-measure than do the homogeneous and the full datasets. From the perspective of indexing conceptions, the content-oriented and the document-oriented approaches indicate better effectiveness in precision, recall, and F-measure, especially for the heterogeneous dataset, than do the domain-oriented indexing conceptions within the context of three datasets. In addition, the homogeneous dataset and the full dataset are similar in terms of results.

Comparisons of indexing conceptions with the baseline. In this experiment, three indexing conception-based approaches were conducted and compared with the baseline (full text) in terms of precision, recall, and F-measure to investigate the differences between the indexing conceptions and the baseline. As shown in Tables 9, 10, and 11, the experiments were conducted for the full dataset, the homogeneous dataset, and the heterogeneous dataset, respectively. For this purpose, three paired t tests were conducted.

Table 9 presents t-tests results comparing the F-measures between the baseline and each of the indexing conceptions for the full dataset, the homogeneous dataset, and the heterogeneous dataset. The full dataset 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 dataset, significant differences are presented between both the document-oriented and the domain-oriented approaches and the baseline in the heterogeneous dataset. In terms of the effectiveness values of the three indexing conceptions and the baseline, the content-oriented and the document-oriented indexing conceptions have a positive impact on TC effectiveness while the domain-oriented indexing conception has a negative impact on TC effectiveness.

Precision measures for the three datasets (Table 10) show that there are significant differences for all three approaches in the full dataset, and this is consistent with the F-measure results. The homogeneous dataset also is 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

DOI: 10.1002/asi

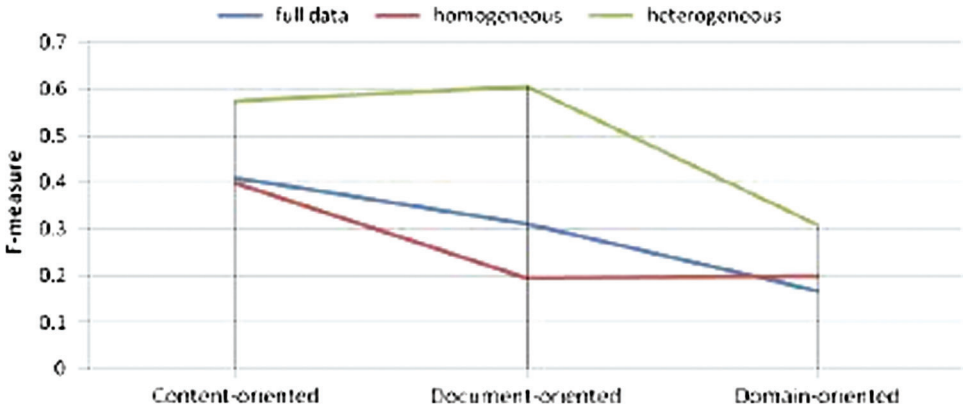

FIG. 3. Three indexing conceptions in terms of F-measure.

TABLE 9. *T*-tests between the baseline and each of the indexing conceptions in terms of F-measure.

	Full dataset		Homogeneous dataset		Heterogeneous dataset	
Indexing conception	*t*	*p*	*t*	*p*	*t*	*p*
Content-oriented	5.492	**0.000***	5.552	**.000***	1.822	0.102
Document-oriented	4.266	**0.000***	4.947	**.001***	2.310	**0.046***
Domain-oriented	−5.505	**0.000***	−1.346	.211	−5.791	**0.000***

*$p < 0.05$.

TABLE 10. *T*-tests between the baseline and each of the indexing conceptions in precision.

	Full dataset		Homogeneous dataset		Heterogeneous dataset	
Indexing conception	*t*	*p*	*t*	*p*	*t*	*p*
Content-oriented	−3.575	**0.002***	4.623	**.001***	0.857	0.414
Document-oriented	4.950	**0.000***	3.754	**.005***	1.570	0.151
Domain-oriented	6.980	**0.000***	0.735	.481	−2.379	**0.041***

*$p < 0.05$.

TABLE 11. *T*-tests between the baseline and each of the indexing conceptions in terms of recall.

	Full dataset		Homogeneous dataset		Heterogeneous dataset	
Indexing conception	*t*	*p*	*t*	*p*	*t*	*p*
Content-oriented	4.550	**0.000***	3.935	**0.003***	1.493	0.170
Document-oriented	2.145	**0.045***	4.229	**0.002***	1.639	0.136
Domain-oriented	−2.708	**0.014***	−0.663	0.524	−3.151	**0.012***

*$p < 0.05$.

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.

As shown in Table 11, recall measures demonstrate that there are significant differences for all three approaches in the full dataset. The *t* test results of both the full dataset and the homogeneous dataset are consistent with F-measure and precision results. The *t*-test results of the heterogeneous dataset show a similar pattern 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.

Discussion

The preliminary framework described 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, the

DOI: 10.1002/asi

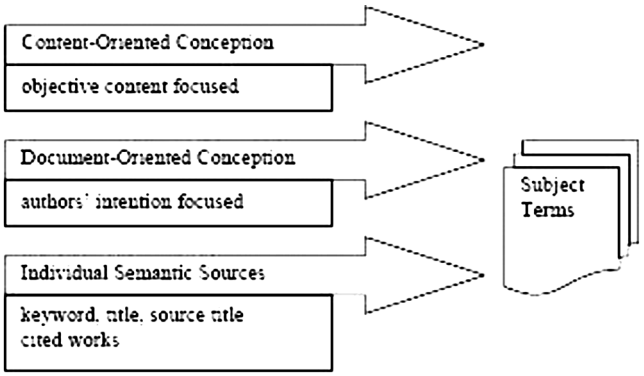

FIG. 4. A revised framework for automatic subject term assignment through text categorization.

preliminary framework was revised and presented in Figure 4. 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, the three indexing conceptions were found to incorporate the understandings of human indexers' indexing practices into TC. The results of the experiment reveal that the content-oriented and the document-oriented indexing conceptions are desirable to be integrated in TC because they both performed better in effectiveness measures. While the content-oriented and the document-oriented indexing conceptions show positive impacts on the effectiveness of TC, the domain-oriented indexing conception does not present such benefits. Although the domain-oriented indexing conception show less effectiveness compared to the content-oriented and concept-oriented indexing conceptions, it might be due to the limitation of resources. Since the semantic source used in this study is simply source title and title of cited works, there might be different when integrating rich sources for the domain-oriented indexing conception. Within the scientific journal article datasets and limited semantic sources, indexing conceptions that orient around objective content or authors' intention are more important than are possible users' needs.

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 TC. The research findings in this study have implications for both theoretical and practical perspectives.

In terms of theoretical implications, the findings demonstrate 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 datasets, the findings indicate that the content-oriented and the document-oriented indexing conceptions are more effective than is the full text. In a sense, subject indexing of scientific journal articles has tended to focus on the objective contents of a document and authors' intentions rather than on possible users' needs. This study suggests that the paradigm of TC research should be changed accordingly. TC research has focused on the statistical and probabilistic foundations utilizing the full text to improve the effectiveness of automatic subject term assignment (Moens, 2002). However, this study sheds 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 a 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 datasets, system designers of TC are able to choose various semantic sources and indexing conceptions by applying them to specific system requirements. In addition, the findings provide the flexibility to select semantic sources and indexing conceptions in terms of the three measures and the various weights of the measures depending on the domain areas.

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 datasets, it was found that the content-oriented and the document-oriented indexing conceptions were more effective than were the domain-oriented indexing conceptions.

Conclusion

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.

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 dataset, 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 datasets (the full data, the homogeneous dataset, and the heterogeneous dataset) were utilized for automatic subject term assignment

DOI: 10.1002/asi

through TC. The experiment results indicate keyword is more effective as a semantic source than is 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.

Second, the research revealed that an indexing conception-based framework is more effective than is 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 is 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 is the domain-oriented indexing conception. In other words, in the context of the scientific journal article dataset of this study, the objective content-oriented indexing conception and the authors' intentions-oriented indexing conception are considered more effective than is the possible users' needs-oriented indexing conception. These findings can be explained as a consequence of the types of datasets, such as 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 the social sciences have an effect on different weights of the three indexing conceptions. Since the dataset for this study is composed of typical scientific journal articles, the objective contents and authors' intentions are identified as more effective than are possible users' needs.

Future 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 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 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 in the future.

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.

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models. Proceedings of the 19th Conference on Machine Learning, Workshop on Text Learning (ICML-02) (pp. 234–241). San Francisco: Morgan Kaufmann.

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

Chung, E., & Hastings, S. (2006). A conception-based approach to automatic subject term assignment for scientific journal articles. In Proceedings of the American Society for Information Science and Technology, Vol. 43 (pp. 1–21). Medford, NJ: Information Today.

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.

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

Engineering Village 2. (n.d.). INSPEC online database. Retrieved November 11, 2006, from http://www.theiet.org/publishing/inspec/

Fall, C.J., Torcsvari, A., Benzineb, K., & Karetka, G. (2003). Automated categorization in the international patent classification. ACM SIGIR Forum, 37(1), 10–25.

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.

Fujino, A., Ueda, N., & Saito, K. (2007). A hybrid generative/discriminative approach to text classification with additional information. Information Processing & Management, 43, 379–392.

Hjørland, B. (1992). The concept of "subject" in information science. Journal of Documentation, 48(2), 172–200.

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.

INSPEC. (2004). Thesaurus. London: Institution of Electrical Engineers.

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 (pp. 137–142).

Kim, J., & Choi, K. (2007). Patent document categorization based on semantic structural information. Information Processing & Management, 43, 1200–1215.

Koster, C.H.A., Seutter, M., & Beney, J. (2003). Multi-classification of patent applications with Winnow. Lecture Notes in Computer Science: Perspectives of System Informatics, 2890, 545–554.

Larkey, L.S. (1999). A patent search and classification system. Proceedings of the 4th ACM Conference on Digital Libraries (pp. 179–187).

Lewis, D.D. (1992). Representation and learning in information retrieval. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Lewis, D.D. (1995). Evaluating and optimizing autonomous text categorization systems. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 246–254). New York: ACM Press.

Lewis, D.D. (2000). Machine learning for text categorization: Background and characteristics. Proceedings of the 21st National Online Meeting (pp. 221–226).

DOI: 10.1002/asi

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 & Management, 41, 599–611.
Miksa, F. (1983). The subject in the dictionary catalog from cutter to the present. Chicago: American Library Association.
Moens, M.F. (2002). Automatic indexing and abstracting of document texts. Norwell, MS: Kluwer.
Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14, 130–137.
Sebastiani, F. (2002). Machine learning in automated categorization. ACM Computing Surveys, 34(1), 1–47.
Sebastiani, F. (2005). Text categorization. In A. Zanasi (Ed.), Text mining and its applications (pp. 109–129). Southampton, United Kingdom: WIT Press.
Slattery, S. (2002). Hypertext categorization. Unpublished doctoral dissertation, Carnegie Mellon University, School of Computer Science, Pittsburgh, PA.
Watters, C., Zheng, W., & Milios, E. (2002). Filtering for medical news items. Proceedings of the 65th annual meeting of the American Society for Information Science and Technology (pp. 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: University of California Press.
Witten, I.H., & Frank, E. (2000). Data mining: Practical machine learning tools and techniques with JAVA implementations. San Diego, CA: Academic Press.
Xu, Z., Yu, K., Tresp, V., Xu, X., & Wang, J. (2003). Representative sampling for text categorization using support vector machine. Proceedings of the 25th European Conference on Information Retrieval Research (pp. 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 (pp. 162–163).

Appendix

Table A1. Subject terms for homogeneous and heterogeneous datasets.

Term for homogeneous dataset	Term for heterogeneous dataset
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

DOI: 10.1002/asi