iRinformationresearch

Vol. 18 No. 4, December, 2013

Investigation into the existence of the indexer effect in key phrase extraction

Jung Eun Hahm, Su Yeon Kim, Meen Chul Kim and Min Song
Department of Library and Information Science, Yonsei University, Seoul, Korea

Abstract

Introduction. *The indexer effect has been studied in several research studies in the field of information science to reveal intellectual structures. In this study, we bring that concept into document classification to verify whether it also influences the results in key phrase extraction.*
Method. *We employ the well-known key phrase extraction technique called the key phrase extraction algorithm for our study. In particular, we extract key phrases from three different datasets: 1) papers in the same journal, 2) papers from different journals in the same field, and 3) papers from journals in different fields. All of these datasets provide keywords and index terms which we used as training data for the algorithm.*
Analysis. *For evaluation, we compare the difference in the performance of key phrases between two groups of key phrases that were extracted using the algorithm: 1) those that used author-provided keywords for the training set, and 2) those that used indexer-assigned index terms for the training set. We analyse those two groups of extracted key phrases in terms of exact (100%) and fair (70%) matching, which is based on the average number of key phrases extracted correctly per document.*
Results. *We conclude that automatic key phrase extraction based on index terms performs better than its counterpart based on author-provided keywords in most cases. However, it also reveals that indexers tend to assign terms more inconsistently.*
Conclusions. *The findings of the study provide some insights into making use of index terms as training data in key phrase extraction. On the other hand, it should be also noted that automatically extracted key phrases might lead users to irrelevant documents in information retrieval.*

Introduction

Drowning in the flood of data that is overwhelming in this information age, people barely keep their head above water. Fortunately there are lifeguards, the so-called indexers, who help information users to land on the right resources for their needs. They select words to represent documents and thereby make information retrieval more effective and efficient. However, providing proper terms about each document manually is laborious and time-consuming, and it becomes a daunting task due to the rapid increase in electronic data. Thus, many researchers have studied ways of determining the representative words for documents automatically. Two of the popular methods are known as key phrase assignment and key phrase extraction.

key phrase assignment selects the phrases that describe a document from a controlled vocabulary, while key phrase extraction, the approach used in this study, chooses key phrases from the text itself without using a controlled vocabulary. There have been rigorous attempts to find keywords or phrases that are representative of documents based on those methods. A typical approach to key phrase assignment or extraction uses supervised learning, which is a machine learning technique of inferring a document's characteristics from labeled training data. key phrases are groups of keywords, which consist of more than one word. A study shows that indexers prefer to assign index terms in the form of phrases over words (Medelyan *et al.*, 2008) that might be different from an author's preference. It is also worth noting that important terms in articles are usually comprised of more than one word, emphasising the contextual meaning in the reading.

Healey *et al.* (1986) pointed out a possible *indexer effect* in choosing words for documents properly. It indicates that the assignment of index terms represents the analyst's subjective view of the contents rather than the author's original view, and it is expressed in the choice of words. This is because individuals interpret documents differently when they read, thus the words that authors provide sometimes differ from those indexers choose for the index. This also implies that the results of

automatic key phrase extraction based on index terms might be quite distinct from those based on keywords.

Therefore, in this study we investigate whether there is any distinction in automatic key phrase extraction between two different points of view. In addition, we analyse the results to explain what makes it different. Performance is measured by comparing the results of the supervised-learning technique with pre-defined key phrases. We conduct two sets of experiments with the same datasets in the training phrase; one with keywords that authors provide and the other with index terms that indexers assign.

We pose three research hypotheses. First, in the same journal, the results of key phrase extraction based on terms that an indexer assigns would be more effective than author-provided keywords in terms of the number of matched words. Second, in different journals in the same field, the results of key phrase extraction based on indexer-provided terms would be more effective than author-provided keywords in terms of matched words. Third, in journals in different fields, the results of key phrase extraction based on indexer-provided terms would be more effective than author-provided keywords in terms of matched words. The results of the experiments show that key phrase extraction based on indexer-assigned terms, which are brought from a professional index database, performs better in general than those based on author-provided keywords. This implies that the concept of the *indexer effect* does exist in key phrase extraction.

The rest of the paper is organised as follows. The next section discusses related works to our paper. Then we introduce the procedure for key phrase extraction. We describe the research questions and methodology. We compare the results, and these are analysed in the Discussion. Lastly, we conclude the paper with a summary and suggestions for future work.

Related works

The *indexer's effect* has been widely investigated in co-word analysis (Callon *et al*. 1986; Law *et al*. 1988; Whittaker 1989; Marion and McCain 2001), but rarely studied in other topics. Most of the research has empirically confirmed the existence of the *indexer's effect* in analysing intellectual structures, showing the distinct differences between keyword-based and index term-based domain maps. Iivonen and Kivimaki (1998) compared the indexing of forty-nine documents, focusing on three factors: 1) the types of concepts presented in indexing, 2) the degree of concept consistency and concept similarities in indexing, and 3) differences in the indexing of concepts. Based on these qualitative observations, they identified that the indexer's subjectivity and inconsistency could be seen reflected in representing concepts in academic works. Mai (2001) manually examined several existing indexing methods of academic documents. Through exhaustive analysis to understand the indexer's cognitive process, the study revealed that indexing is not a neutral and objective representation of a document's subject matter but an interpretation for future use. He also suggested a semiotic framework for the subject indexing process. Providing proper terms about each document manually, however, is laborious and time-consuming, and it becomes a daunting task due to the rapid increase of information. Based on these empirical and qualitative observations, therefore, we seek to investigate 1) whether the *indexer's effect* exists in the automatic key phrase extraction technique, and 2) whether the indexer's subjectivity and inconsistency can be measured quantitatively. We also evaluate the extraction performance in a rigorous manner which previous studies have not considered.

Previous works on key phrase generation can be categorised into two major approaches: key phrase extraction and key phrase assignment. In key phrase extraction, phrases in a document are identified on the basis of properties such as term frequency and length. key phrase assignment, however, aims at finding appropriate key phrases based on a pre-defined set of terms such as a thesaurus. The latter has an obvious limit because the continuous expansion of controlled vocabularies is inefficient. Against this backdrop, we will briefly discuss what previous research has revealed, especially for key phrase extraction, which serves as a foundation for our experiment.

Key phrase extraction

Key phrase extraction is widely used for the tasks of indexing, summarisation, clustering, categorisation, and more recently, in improving search results and in ontology learning. A number of systems have addressed the task of key phrase extraction, most of which employed supervised machine learning techniques (Turney 1999; Witten *et al.* 1999; Barker *et al.* 2000; Wang *et al.* 2006; Nguyen *et al.* 2007; Wu *et al.* 2008). There have also been studies with unsupervised machine learning techniques (El-Beltagy 2009; Song *et al.* forthcoming).

The first system, GenEx (Generic Extractor), employs machine learning-based key phrase extraction devised by Turney (1999). The GenEx system has two different components: the Extractor algorithm identifies candidate phrases, and scores them depending on their length and position in the text. Next, the Extractor presents the top-ranked key phrases as outcomes to the user, while the Genitor algorithm searches for overlap among the key phrases. Witten *et al.* (1999) suggested a competing key phrase extraction system called the key phrase extraction algorithm, which identifies candidate key phrases using lexical methods, calculates feature values for each candidate and uses a machine-learning algorithm to predict which candidates are proper key phrases. The performance of the algorithm proves to be highly accurate and robust when it is compared to that of any other system. Hence, we used it for our key phrase extraction experiment.

There are some studies which added new features to the existing techniques. Barker *et al.* (2000) presented a system for choosing noun phrases from a document as key phrases. Contrary to Turney (1999) and other works, it restricted key phrases only to noun phrases. However, a limitation when extracting key phrases can be experienced if an important term is not recognised as a noun. Turney (2002) presented a new feature, the concept of a coherence set to measure cohesiveness between candidate phrases. He pointed out that extracted key phrases are sometimes outliers, which means they have no semantic relationships with other key phrases in a document. This new feature is calculated as the statistical association between a candidate and a small group of top candidate phrases.

Other innovative algorithms to extract key phrases also have been devised. Their features and approaches usually prove to be efficient and effective. Nguyen *et al.* (2007) extended the algorithm for extracting key phrases from scientific publications. They added two more features to the algorithm: the positions of phrases in documents and morphological factors found in scientific key phrases, such as whether a candidate key phrase is an acronym or uses terminologically productive suffixes. Wang *et al.* (2006) proposed a key phrase extraction approach based on neural networks. They adopted the following features to extract key phrases from given documents: term frequency and inverted document frequency respectively, and frequency of appearance in the paragraphs of the documents. Wu *et al.* (2008) proposed a *key phrase identification program* algorithm which extracts key phrases based on the composition of noun phrases. In this algorithm, the more keywords a phrase contains, the more likely this phrase is to be a key phrase. The *key phrase identification program* first extracts a list of candidate key phrases, then assigns scores to them based on three factors: frequency of occurrence in a document, composition, and how specific these words and sub-phrases are in the domain of the document. For our experiment, we employ a new feature, index term as training data for key phrase extraction. In addition, we compare the performance of index term-based key phrase extraction to that of keyword-based extraction to identify the existence of the *indexer effect.*

General procedure of the key phrase extraction algorithm

In this section, we explain the key phrase extraction technique based on the key phrase extraction algorithm (Witten *et al.*, 1999) that is used for this study. This algorithm was chosen because it is an open source system written in Java and provides the flexibility of customising the procedure. It is an algorithm for extracting key phrases from the text of a document: all candidate phrases in a document are identified and the features are computed for each candidate phrase. After that, machine learning is used to generate a classifier that determines which candidates should be chosen as key phrases. Figure 1 describes the general procedure of the key phrase extraction system based on supervised learning.

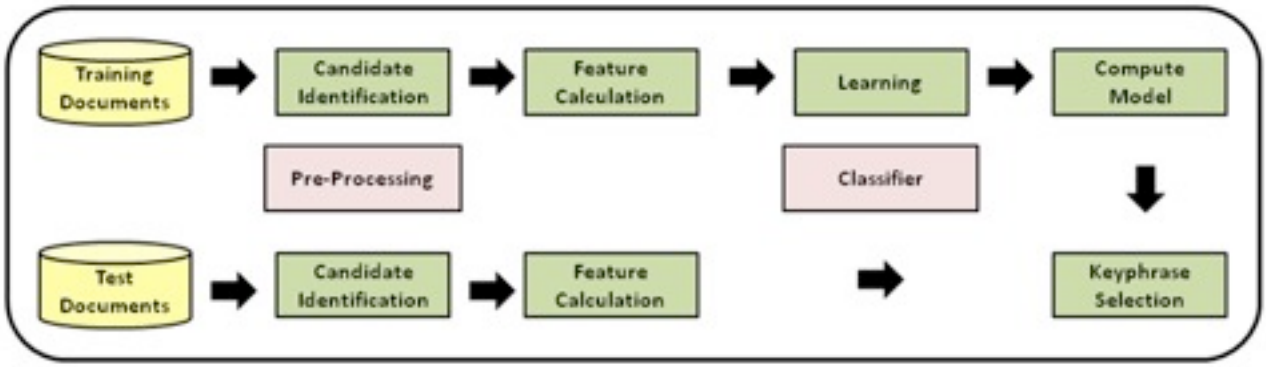

Figure 1: The procedure of the key phrase extraction algorithm

Candidate identification

Each document in the collection is first segmented into tokens on the basis of white space and punctuation. Clues to syntactic boundaries such as punctuation marks, digits and paragraph separators are retained. Next, all word n-grams that do not cross phrase boundaries are extracted and the number of occurrences of each n-gram in the documents is counted. Case folding and stemming is also done to treat different variations on a phrase as the same thing.

Feature calculation

A simple but robust machine-learning scheme is used to determine the final set of key phrases for a document. It uses a set of attributes as input, which is defined for each candidate term. Two features are used in the standard algorithm: a term frequency-inverse document frequency score and the position of first occurrence.

The frequency score compares the frequency of a phrase's use in a particular document with the frequency of that phrase in general use. General usage is identified by the number of documents containing the phrase in the global corpus of documents. By default, the training set is regarded as the global corpus but this can be changed if required. The position of first occurrence of a term is calculated as the distance of a phrase from the beginning of a document in words, normalised by the total number of words in the document. The result represents the proportion of the document preceding the phrase's first appearance. Candidates that have very high or very low values for this feature are more likely to be valid key phrases, because they appear either in the introductory parts of the document such as a title, abstract, table of contents and introduction, or in its final sections such as conclusion and reference lists.

There are more features we can use: length of phrase in words and node degree. The length of a candidate phrase measured in words boosts the probability of two-word candidates being key phrases. The statistical analysis of the experimental data revealed that indexers prefer to assign descriptors consisting of two words whereas most terms in the corresponding thesaurus are one word (Medelyan *et al.*, 2008). The node degree measures how richly the term is connected in the thesaurus graph structure. The degree of a term is the number of semantic links that connect the term to other phrases in the document that have been identified as candidate phrases. Additionally, we use the length of phrases in words feature in our study.

Training: building the model

In order to build the model, a training set of documents with known key phrases is required. For each training document, candidate pseudo-phrases are identified and their feature values described above are calculated. Each phrase is marked as a key phrase or not, using the actual key phrases for those document. The prediction model is then constructed from these training instances with the data mining software, Waikato Envrionment for Knowledge Analysis, which applies a machine learning scheme to learn two sets of numeric weights from the discretised feature values, one set to positive instances and the other to negative ones. In other words, this is binary classification. the key phrase extraction algorithm uses the naïve Bayes technique, which is simple and yields good results. In addition, we also use a decision tree technique to compare the performance with that from naïve Bayes. The final component of the model is the number of positive and negative examples in the training data. To select key phrases from a new document, the algorithm identifies candidate pseudo-phrases and their feature values, and then applies the learned model. Candidate phrases are ranked according to this value, and the first *n* phrases are returned, where *n* is the number of key phrases requested by user.

Research hypotheses

This study investigates the existence of the *indexer effect* when using index terms in place of keywords for automatic key phrase extraction. Because two different people interpret disciplines based on their own understanding, they sometimes use different words in representing a document. This causes inequality in the results of machine-learning, when using different people's document representations for intellectual clues. Regarding this, Marion and McCain (2001) argue that authors view a discipline from the perspective of insiders, whereas indexers do so as outsiders. Therefore, we focus on verifying the indexers' influence in document classification by introducing the concept of the *indexer effect*. We also suggest the possibility of using index terms as training data especially in the situation where keywords are not provided in the articles.

Table 1 gives an example of an article with keywords and index terms. We extracted keywords from the article, and used index terms from *Library and Information Science Abstracts*.

Keywords: given by author	Index terms: given by indexer
Online information retrieval	Online information retrieval
Web databases	
Searching	Searching
User interface	User interface
User satisfaction	User satisfaction
Gender differences	
Retrieval performance measures	Retrieval performance measures
	Evaluation
Experts	**Expert users**
Novices	**Novice users**
Comparative studies	
Web of Science	Web of Science

Table 1: Keywords and index terms for Ahmed, S. Z., McKnight, C. & Oppenheim, C. (2004). A study of users' performance and satisfaction with the Web of Science IR interface. *Journal of Information Science*, 30(5), 459-468.

As shown above, although some words from each perspective are identical, others differ slightly. Among the given words, six of them are exactly the same: *online information retrieval*, *searching*, *user interface*, *user satisfaction*, *retrieval performance measures*, and *Web of Science*. There are several words that somewhat match from each viewpoint. For example, the author assigns *experts* and *novices* to this document, which is very similar to *expert users* and *novice users* used by the indexer. In addition, new words are offered from only one viewpoint. For instance, the author deems *Web databases* and *gender differences* as important terms, which are not included in the index list.

This disparity comes from several distinct conditions related to the circumstance of each group. First, authors are more professional in giving keywords for articles compared to indexers. Authors have knowledge of what people want to find in their field, so they are able to give more appropriate access points for information. In addition, authors provide the words for the articles they write, so they know the precise words for representing their writing. Hence, the keywords authors provide are more specific than words in an index because authors consider their article within the context of their fields.

Secondly, authors want people to cite their articles, so they sometimes exaggerate the contents of their writing. Information users usually decide to continue their reading by looking through the first few paragraphs, so making a good first impression is important for authors. Keywords help people understand reading materials quickly, not only by showing core words indicating major points in the articles but also by offering access points matching people's interests in topics. However, indexers just guide people to the right paths without any profit for themselves. This implies that words based on authors' intellectual structures might be subjective and perhaps even deceptive compared to words in an index, even for the same article.

Finally, indexers usually review various disciplines at the same time, so they tend to remain superficial in offering an index for articles. Even though they are information specialists, indexers may not be deeply knowledgeable in many academic areas. In addition, compared to the small number of indexers, the enormous number of works they are dealing with forces them to complete their work hastily and, perhaps, imprecisely, spending not enough time reading an article in detail. Indexers, therefore, connect more general words to the information resources they handle.

Considering the different situations these two groups face, their intellectual base might also be different, which leads to using different terms for description, even in the same document. We have already confirmed this by comparing the word list for an article based on each point of view above. Depending on this gap, we may be able to conclude there is an *indexers' effect* in the choice of words, which determines access points for people. This also means the possibility of displaying distinct results when automatically extracting words to describe documents, because it uses either keywords or index terms individually. We assume key phrase extraction performance from one particular viewpoint is more efficient than the other, which means selecting the right terms that the group

intends to pick for descriptors. If it isn't so, then we can suggest the utility of index terms as the intellectual base for machine-learning, instead of keywords which are rarely provided. This highlights the importance of our experiments to check whether the *indexer effect* really exists in automatic key phrase extraction or not. The three research hypotheses we set in this study are as follows:

1. In the same journal, the results of key phrase extraction based on indexer-provided terms are more effective than author-provided keywords, in terms of comparing the number of matched words.

2. In different journals in the same field, the results of key phrase extraction based on indexer-provided terms are more effective than author-provided keywords, in terms of comparing the number of matched words.

3. In journals from different fields, the results of key phrase extraction based on indexer-provided terms are more effective than author-provided keywords, in terms of comparing the number of matched words.

Methods

Data collection

To verify the research hypotheses, we collected data from the five different journals shown in Table 2. To evaluate automatic key phrase extraction techniques, articles should have keywords assigned and be found in the index database with appropriate index terms. For journal selection, we first referred to the Web of Science *Journal Citation Reports*. Two important criteria for the journal selection were: 1) whether the subject matter of the journals was suitable for testing our hypotheses and 2) whether the journals had articles to which keywords as well as indexing terms had been assigned. It was important to gather large enough data sets since the precision of key phrase extraction is negatively influenced by a training data set comprised of a small number of observations.

Discipline	Journal title	Index database
Information science	Journal of Information Science, Journal of Informetrics, Scientometrics	LISA
Education	American Educational Research Journal	ERIC
Social science	Sociological Methods and Research	Social Services Abstracts

Table 2. Index database list for journal data

We selected journals in the field of information science as the main training data in our study under the assumption that authors in this discipline might provide keywords in a similar way indexers do. We chose the *Journal of Information Science* for the first research hypothesis. We chose the *Journal of Informetrics* and *Scientometrics* for the second research hypothesis, which addresses different journal data in the same field. Lastly, we included two additional fields for our experiment; one close to information science and the other a little farther from this field, and selected one journal from each discipline. They were the *American Educational Research Journal* and *Sociological Methods and Research*.

From these journals, we collected full-text articles with keywords and converted the full-text articles into text files. To find index data for each article, we referred to an index database such as *Library and Information Science Abstracts*. Then, we retrieved articles that we collected from the index databases and manually matched the respective index terms.

We collected full-text articles from *Journal of Information Science* published between 2004 and 2011. The number of articles available is 330 in total (Table 3).

Journal of Information Science								
2011	2010	2009	2008	2007	2006	2005	2004	Total
44	45	37	46	45	41	34	38	330

Table 3.The number of articles in each year in *Journal of Information Science*

We used them for training as well as test data for research hypothesis 1. Two sets of training were compiled: for the first set, we selected articles published from 2004 to 2010 as training data set 1. We automatically extracted key phrases from articles issued in 2011 based on the model built. For the second training set, we chose those published in 2005-2011 as training data set 2, and used articles in

2004 as test data. This was done because any error from training data can influence the results of key phrases extraction. Hence, we checked if any different occurrence exists when changing the training set. For research hypothesis 2 and 3, we kept using two sets of training data from the *Journal of Information Science.* We additionally collected articles published in 2011, 2010 and 2004 each as test data. The number of articles, index terms, and keywords for each year is as follows in Table 4:

Journal Title	No. of articles	No. of Index terms	No. of Keywords
Journal of Information Science (2011)	44	171	225
Journal of Information Science (2004)	38	363	308
Journal of Informetrics (2011)	46	161	253
Journal of Informetrics (2010)	30	73	146
Scientometrics (2011)	46	202	236
Scientometrics (2010)	45	147	227
American Educational Research Journal (2011)	44	655	215
American Educational Research Journal (2004)	42	427	171
Sociological Methods and Research (2011)	42	463	187
Sociological Methods and Research (2004)	42	345	182

Table 4.The number of articles, index terms and keywords for test data

It should be noted that index terms are not available for *Journal of Informetrics* before 2010, and there are no keywords provided in articles before 2010 for *Scientometrics.* As a result, we collected articles published in 2010 for test data instead of those in 2004 in those journals. There are also several articles which do not contain keywords or are not retrieved in the index database. Thus, we excluded those articles from the test collection that did not meet these conditions.

Performance measure

In the experiment, we compare key phrases extracted by the machine to those assigned by the humans. We consider that there is a match if the stem of the automatically extracted key phrase matches with the stem of either the author-assigned keywords or indexer-provided index terms. In cases of a match, we determine the match to be either 100% (exact) or 70% (fair). Stemming is carried out by the Porter stemmer (1980). Stemming is a process for removing the common morphological and inflectional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up an information retrieval system. Stemming helps minimise term variations. For instance, terms like *retrieval*, *retrieving*', and *retrieved* are stemmed to the stem *retriev*. However, the stemming technique alone cannot resolve the issue of mismatching due to term variation. In our experiment, any partial match is empirically set to 70%. To determine a partial match, we employed the Jaro-Winkler distance technique, which is a measure of similarity between two strings (Winkler, 1990). The higher the Jaro-Winkler distance for two strings, the more similar the strings are. The score is normalised such that 0 equates to no similarity and 1 is an exact match.

For evaluation, inter-indexer consistency and agreement are commonly used in the context of human indexing, while precision and recall are used to evaluate automatic systems (Medelyan *et al.,* 2008). We employed the inter-indexer consistency formula proposed by Rolling (1981) to measure the number of matching words each group provided. Rolling's measure relies on simple formulaic representations of term sets assigned by each indexer as well as the common terms between the two sets. It defines the level of consistency between two indexers as the total number of terms in agreement divided by the total number of distinct terms used by both indexers (Wolfram and Olson, 2007). To calculate the inter-indexer consistency, let A and B be the size of the two indexer's term sets and C be the number of terms in common between the two sets. Then, Rolling's measure is R = 2C/(A+B). Assume that a certain article has *information retrieval*, *ranking model* and *query expansion* as author's keywords, and *information retrieval* and *relevance feedback* as indexer's terms. There are three terms in A and two terms in B, and the number of common terms between A and B is one Then, the inter-indexer consistency of this article is R = 2(1)/(3+2) = 0.4. This evaluation is usually used to compare the inter-indexer consistency among several indexers and to prove the validity of the low performance of a machine learning algorithm; the difference between indexers is similar to that between the human and the machine.

The reason for using this evaluation method in our study is to confirm the performance quality of automatic key phrase extraction by showing that different groups of people assign terms in a highly inconsistent manner.

Additionally, we used the average number of key phrases correctly extracted per document. We compared the extraction results of 5, 10, and 15 key phrases. The KEA package provides this evaluation metric to measure the effectiveness of automatic key phrase extraction systems with manually extracted ones, and we used them for our evaluation.

Results

Measuring the inter-indexer consistency between authors and indexers

In this section, we firstly measured the inter-indexer consistency between authors and indexers for the same articles using Rolling's measure. All the inter-indexer consistency between the two groups is quite low (0.11 – 0.21), except for articles in the Journal of Information Science in 2004, which is 0.49 (Table 5).

Journal Title	No. of index terms	No. of Keywords	No. of correctly extracted words	Inter-consistency
Journal of Information Science (2011)	171	225	36	0.18
Journal of Information Science (2010)	149	198	27	0.16
Journal of Information Science (2009)	109	173	21	0.15
Journal of Information Science (2008)	142	215	37	0.21
Journal of Information Science (2007)	146	241	37	0.19
Journal of Information Science (2006)	284	189	43	0.18
Journal of Information Science (2005)	235	180	37	0.18
Journal of Information Science (2004)	363	308	166	0.49
Journal of Informetrics (2011)	161	253	23	0.11
Journal of Informetrics (2004)	73	146	12	0.11
Scientometrics (2011)	202	236	34	0.16
Scientometrics (2004)	147	227	20	0.11
American Educational Research Journal (2011)	655	215	28	0.06
American Educational Research Journal (2004)	427	171	30	0.10
Sociological Methods and Research (2011)	463	187	30	0.09
Sociological Methods and Research (2004)	345	182	24	0.09

Table 5.The inter-indexer consistency between authors and indexers

Except for the data in 2004, the average of the inter-indexer consistency of documents included in *Journal of Information science* from 2005 to 2011 is 0.18. Other journals maintain similar inter-indexer consistency without being influenced by the temporal variance. Most cases which have lower inter-indexer consistency do not have similar number of words between index terms and keywords; one is extremely higher than the other.

In addition, we used naïve Bayes and decision tree classifiers, and identified that the former performs better than the latter in every case. We, therefore, only reported the results with the naïve Bayes classifier, which is also used as the default algorithm in the key phrase extraction algorithm.

Research hypothesis 1: evaluation for the same journal in the same field

First, we computed the number of exact match (100%) key phrases, i.e. whether the machine learning finds the exact words a human intends to describe, based on each perspective. The results were near to zero but the number of matching words suddenly increased when we applied fair (70%) match key phrases. We found that key phrase extraction results with terms offered by indexers performed better than those from authors in most cases (Table 6).

Journal of Information Science Fair Match (70%)*	5**	10**	15**
2011 keyword	2.3 +/- 2.23	3.91 +/- 3.33	5.37 +/- 3.69
2011 index	2.6 +/- 2.55	4.24 +/- 3.73	5.62 +/- 4.23
2004 keyword	3.32 +/- 2.98	5.92 +/- 4.67	7.37 +/- 5.34
2004 index	3.45 +/- 2.67	5.71 +/- 5.31	7.82 +/- 6.6
* The results of fair matches show both the average number of key phrases extracted correctly and the standard deviation. ** 5, 10, and 15 respectively indicate the number of automatically extracted key phrases.			

Table 6. The performance of the key phrase extraction algorithm with the *Journal of Information Science*

We graphed the average of correctly extracted key phrases divided by the total number of key phrases extracted in Figure 2 and confirmed that the efficiency of the indexers' perspective is higher than that of authors. We marked test data with keywords published in 2004 as 2004 *kyw* and those with index terms in 2004 as 2004 *idx* below. The greater the number of correctly extracted words those two different groups share, the higher the hit rate that machine-extracted phrases show.

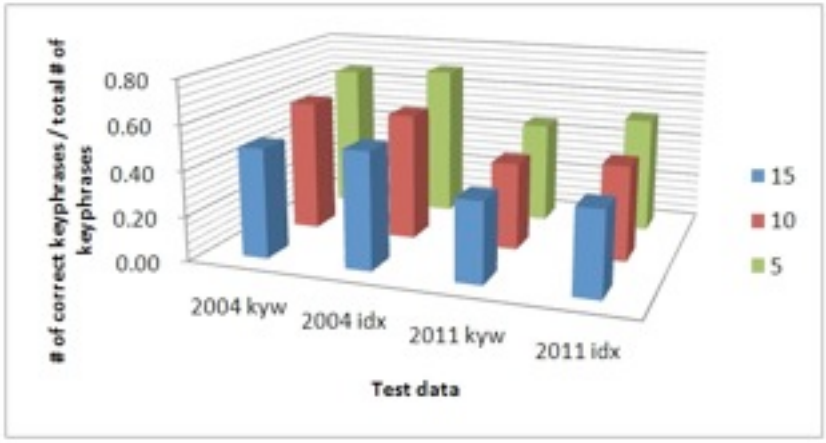

Figure 2.The performance graph with the Journal of Information Science

To test the statistical significance of the two different perspectives, we conducted a t-test. The t-test had a value of 2.354 showing that there is a statistically significant difference of performance (a p-value of 0.046 ($p<0.05$)) between the author's keyword and indexer's term within the same journal published in 2011. That is, any case which employs an indexer's term with a fair (70%) match reveals better performance for training data than an author's keyword in the same journal published in 2011. However, we did not observe statistical significance for the 2004 data.

Research hypothesis 2: evaluation for different journals in the same field

We extracted key phrases from two other journals in the same field with the same training data used in research hypothesis 1. We also presented fair (70%) match key phrases only, as the performance in terms of exact (100%) match key phrases was near to zero.

The result with 2011 data reveals that the performance from the indexers' intellectual base is better than that from the authors' but the cases in 2010 turn out to be the opposite (Table 7). Automatic key phrase extraction based on keywords proved to be more precise in this case.

Journal of Informetrics Fair Match (70%)*	5**	10**	15**
2011 keyword	3.17 +/- 3.56	5.17 +/- 4.37	6.61 +/- 4.75
2011 index	2.85 +/- 5.7	6.33 +/- 11.09	7.96 +/- 13.66
2010 keyword	2.6 +/- 3.7	5.23 +/- 6.45	7.2 +/- 9.33
2010 index	2.13 +/- 3.22	4.53 +/- 5.41	7 +/- 8.67
Scientometrics Fair Match (70%)	**5****	**10****	**15****
2011 keyword	2.78 +/- 2.17	4.43 +/- 3.05	6.57 +/- 4.38

2011 index	3.78 +/- 4.83	6.78 +/- 7.79	11.78 +/- 10.17
2010 keyword	4.09 +/- 3.99	6.38 +/- 5.14	7.56 +/- 5.85
2010 index	3.49 +/- 5.57	5.91 +/- 7.8	8.53 +/- 10.24
* The results of fair matches show both the average number of key phrases extracted correctly and the standard deviation. ** 5, 10, and 15 respectively indicate the number of automatically extracted key phrases.			

Table 7. The performance of the key phrase extraction algorithm with journals in same field

We graphed the average of correctly extracted key phrases divided by the total number of key phrases extracted in Figures 3 and 4, and confirmed again that index term-based extraction mostly performs better than extraction based on author-provided keywords.

In addition, we identified that a gap of performance between the two perspectives exists. Specifically, the gap between extracted results from authors' and indexers' base in the *Journal of Informetrics* is small: however, *Scientometrics* shows a bigger discrepancy.

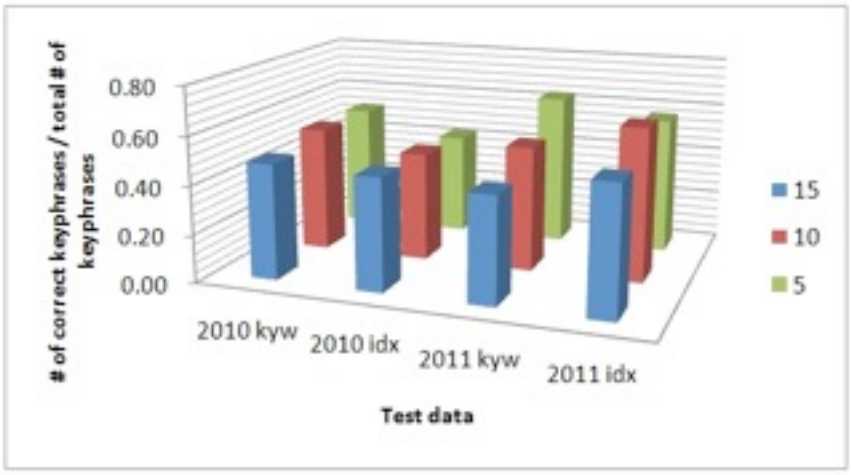

Figure 3: The performance graph with the *Journal of Informetrics*

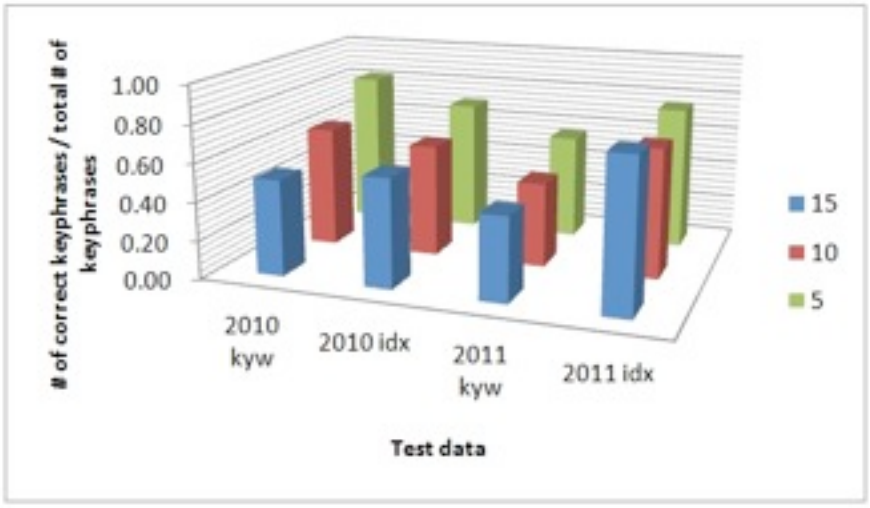

Figure 4: The performance graph with *Scientometrics*

Meanwhile, there was no statistically significant difference in performance between author's keyword and indexer's term for different journals in the same field.

Research hypothesis 3: evaluation for different journals in different fields

With training data in different domains, machine learning by index terms shows better results for finding appropriate key phrases (Table 8).

American Educational Research Journal Fair Match (70%)*	5**	10**	15**
2011 keyword	6.2 +/- 6.24	8.5 +/- 6.62	10.52 +/- 7.09
2011 index	9.7 +/- 10.73	14.09 +/- 13.51	17.75 +/- 14.45
2004 keyword	4.24 +/- 3.64	7.29 +/- 5.65	9.57 +/- 6.93
2004 index	5.12 +/- 5.64	8 +/- 7.02	11.6 +/- 8.52
Sociological Methods and Research Fair Match (70%)*	**5****	**10****	**15****
2011 keyword	2.76 +/- 2.28	3.98 +/- 2.67	5.56 +/- 3.78
2011 index	2.73 +/- 2.56	5.93 +/- 4.81	9.27 +/- 6.37
2004 keyword	1.88 +/- 1.74	3.14 +/- 2.72	4.48 +/- 3.22
2004 index	2.79 +/- 3.22	5.98 +/- 4.09	8.1 +/- 5.32

* The results of fair matches show both the average number of key phrases extracted correctly and the standard deviation.
** 5, 10, and 15 respectively indicate the number of automatically extracted key phrases.

Table 8. The performance of key phrase extraction algorithm with two journals in different fields

We also graphed the results in Figure 5 and 6. In the *American Educational Research Journal*, the performance rate in 2011 *idx* is even greater than 1.0. We also found that within the journals in different domains, there is a subtle distinction among performances. This is due to changes of the training data. To be specific, in the *American Educational Research Journal* which is regarded as publishing studies on very similar topics with the articles of the training journal, the gap between the *kyw* and *idx* is not big.

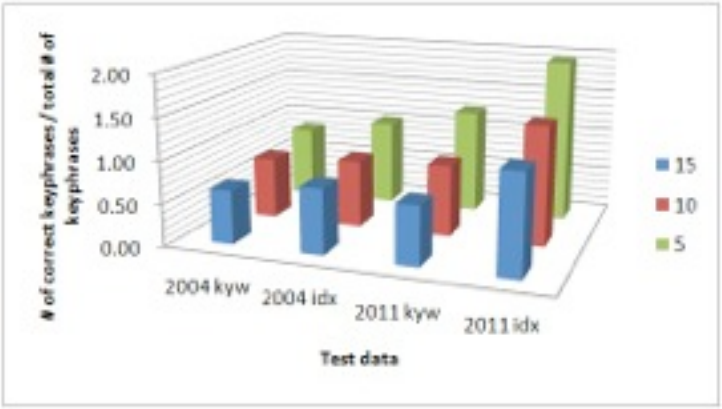

Figure 5: The performance graph with the *American Educational Research Journal*

In *Sociological Methods and Research,* however, it is obvious that the matching rate with index terms shows the better performance. The distinction between the two aspects is quite clear.

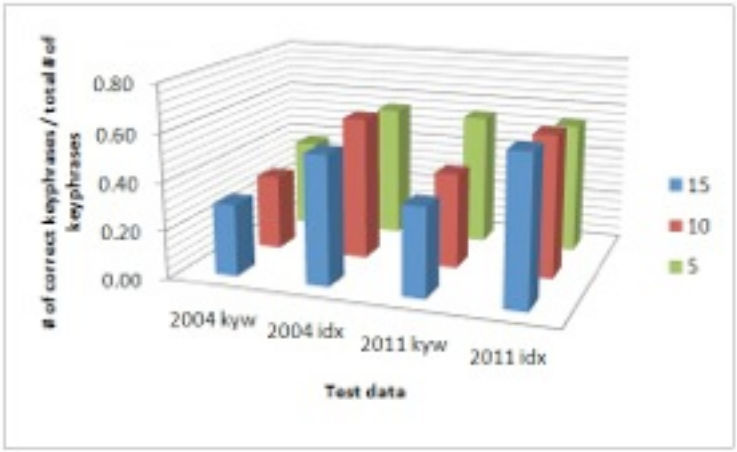

Figure 6: The performance graph with *Sociological Methods and Research*

To test the statistical significance of the two different perspectives, we conducted a t-test. Table 10 below shows a statistically significant difference of performance between author's keyword *(kyw)* and indexer's term *(idx)* exists. That is, any case which employs the indexer's term *(idx)* with a fair (70%) match reveals a better performance as training data than the author's keyword *(kyw)* from journals in different fields.

Test Set	t-value	p-value
2011	-3.348	0.004 (p<0.01)
2004	-4.298	0.000 (p<0.001)

Table 9: T-test of performance between author's keywords and indexers' terms from journals in different fields

Discussion

Findings from the experiment are summarised as follows. First, we found that indexers assign index terms in a more varied form than the keywords authors offer. When we compared the exact (100%) and the fair (70%) matching number of key phrases based on the two intellectual bases, the difference between the two was noticeable when training was done with index terms. We graphed five key phrases extracted from documents issued in 2004 as an example in Figures 7 and 8. As shown below, it is revealed that indexers would select the greater variety of terms for documents while authors choose more consistent words for representing main concepts of the documents. This is especially true when the discipline of the journal between training and test data is not similar.

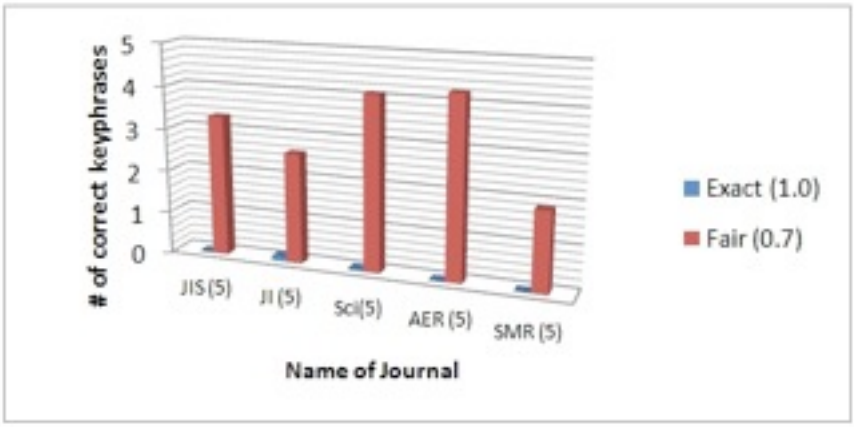

Figure 7: Five key phrase extraction with 2004 documents based on authors

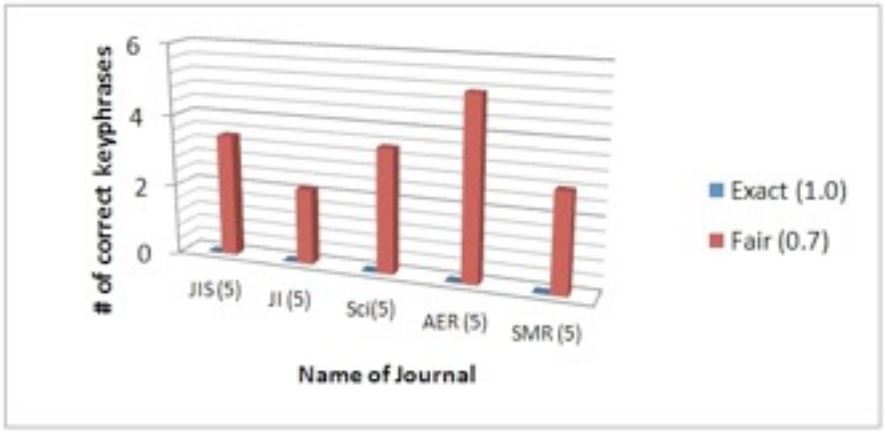

Figure 8: Five key phrase extraction with 2004 documents based on indexers

We found that there is a noticeable discrepancy between author-provided terms and indexer-provided terms. This disparity could be attributed to the way authors view disciplines from the perspective of insiders, while indexers from that of outsiders, as suggested by Marion and McCain (2001). Authors are those who understand their fields in depth and know important concepts in their domain; they are experts in their field and give keywords that have already been widely adopted. Indexers, however, consider words that will represent an article well and thoroughly examine an individual document;

they sometimes select key phrases with greater variation when representing certain concepts, leaving the potential for a wrong decision. For example, in one of the articles used in this study, authors provide a keyword *sliding window*, but indexers give the index term *sliding window method* to indicate that keyword, which might be quite similar in some way but not exactly. Another presents two index terms for one keyword; *fuzzy logic* and *entropy theory*, representing *fuzzy entropy*. This implies that the retrieval results will be different between a keyword-based and an index-term-based search; information users might retrieve less relevant documents from the latter.

We should also note that authors tend to list the exact words they use in their articles for keywords, so the exact matching rate from authors is higher than that from indexers. Although indexers keep an objective attitude while assigning index terms, the exact matching rate from indexers is near to zero, because they determine the access points for articles not from its content but from their indexing rules and judgment.

It is a contradictory finding that even though automatic key phrase extraction with index terms performs better than with keywords, the consistency in giving representative words of the author group is higher than the indexer group. In other words, indexers have greater variance when assigning terms to articles than authors do because indexers use a larger number of terms. We suggest, therefore, with these results that index terms can be utilised efficiently as training data to extract key phrases automatically for a large number of documents. At the same time, it should be noted that information users might not prefer those results, which may bring them to irrelavant resources in the information retrieval process.

Secondly, we proved that the automatic key phrase extraction performance with indexer-assigned terms is more efficient than with those of authors. This result becomes more obvious when training and test data come from different domains. In this case, the supervised learning technique matches a greater number of human-offered words that refer to more general descriptors from the indexers' perspective. We assume that this disparity comes from motivational variance from each group; some authors may assign keywords for their article just because journals ask them to do so. It is also noticeable that key phrase extraction is hardly influenced by the journal from which the training data comes: within the same journal or in different journals in the same field. When it comes to journals in different scholarly areas, however, the characteristics of the training data have an impact on the key phrase extraction outcomes more significantly, resulting in quite a different gap in its performance. This is because the keywords usually come from their own fields, revealing their contextual meaning, so they fail to mirror core words in other domains; index terms might be more efficient, as they take a more general approach to scholarly areas. It also implies that when compiling training data for key phrase extraction, we should carefully look at the characteristics of both training and test data at the outset.

Conclusion and future work

In this study, we investigated whether the *indexer effect* exists in key phrase extraction. We also presented a new key phrase extraction approach based on indexers' terms as training data instead of keywords, which is built based on the key phrase extraction algorithm. For evaluation, we compared the key phrases extracted by the machine with those assigned by authors and by indexers respectively. We analysed extracted key phrases in exact (100%) and fair (70%) matching by the average number of key phrases extracted correctly per document. We extracted key phrases from three different data sources: 1) full-text articles from the same journal, 2) full-text articles from different journals in the same field, and 3) full-text articles from journals in different fields, with keywords and index terms as training data.

To verify the research hypotheses addressed, we selected a variety of journals in the same field and different fields. We were particularly interested in whether the characteristics of the journal affect the performance of key phrase extraction.

The results showed that when articles that have fewer interrelated topics are included in the training data, the results are influenced more by index terms than by keywords. For example, in research hypothesis 2, the *Journal of Informetrics* publishes articles with very similar topics to the articles in *Journal of Information science,* which was used to create the training data, but those in *Scientometrics* do not deal with related topics with those in the training data. Thus, the latter is influenced more strongly by the terms that indexers offer. This causes indexer-based extraction to perform more efficiently in the latter journal, which has more domain-independent characteristics. The same rule also applies to research hypothesis 3. Using a journal in the field of sociology, which is considered less

connected to information science than education, has a greater effect on the extraction performance with index terms because of its more general characteristics.

With the experimental results, we conclude that the automatic key phrase extraction with index terms performs better in most cases but it also reveals that indexers tend to assign terms inconsistently. It implies that it is more appropriate to use index terms as training data in key phrase extraction but automatically extracted key phrases might lead users to less relevant documents in information retrieval.

One of the limitations of our current work is that the training data are constructed with one journal. As a follow-up study, we will explore how training data with multiple journals performs compared to that with one journal. We are also interested in using other kinds of data sets for supervised learning to confirm the *indexer effect* in automatic key phrase extraction. In addition, future work will focus more on adding new features to the algorithm and conducting experiments with other training data.

Acknowledgements

This study was supported by a National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291).

About the authors

Jung Eun Hahm is a MA candidate of Library and Information Science at Yonsei University, Seoul, Korea. Her research interests are text mining, bibliometrics, and information retrieval. She can be contacted at jungeunhahm@yonsei.ac.kr.
Su Yeon Kim is a Ph.D. candidate of Library and Information Science at Yonsei University, Seoul, Korea. Her research interests include text mining, bibliometrics, information retrieval and text categorisation. She can be contacted at suyeon@yonsei.ac.kr.
Meen Chul Kim is an assistant researcher of Library and Information Science at Yonsei University, Seoul, Korea. His research interests include Information retrieval and Text Mining from the Web and Social Media, and Information System Evaluation. He can be contacted at andrewevans@yonsei.ac.kr.
Min Song is an Associate Professor of Library and Information Science at Yonsei University, Seoul, Korea. He is the corresponding author for this paper. His research interests are text mining, social media mining, and bio-literature mining. He can be reached at min.song@yonsei.ac.kr.

References

Ahmed, S. Z., McKnight, C. & Oppenheim, C. (2004). A study of users' performance and satisfaction with the Web of Science IR interface. *Journal of Information Science*, **30**(5), 459-468

Barker, K. & Cornacchia, N. (2000). Using noun phrase heads to extract document key phrases. In: H. J. Hamilton, (Ed.), *Advances in Artificial Intelligence. Proceedings of 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000. 14-17 May 2000, Montreal* . (pp. 40-52). Berlin: Springer.

Callon, M., Law, J. & Rip, A. (1986). *Mapping the dynamics of science and technology: sociology of science in the real world.* London: Macmillan Press Ltd.

El-Beltagy, S. R. & Rafea, A. (2009). KP-Miner: a key phrase extraction system for English and Arabic documents. *Information Systems*, **34**(1), 132-144

Healey, P., Rothman, H. & Hoch, P. (1986). An experiment in science mapping for research planning. *Research Policy*, **15**(5), 233-251

Iivonen, M. & Kivimaki, K. (1998). Common entities and missing properties: similarities and differences in the indexing of concepts. *Knowledge Organization*, **25**(3), 90-102

Law, J., Bauin, S., Courial, J. P. & Whittaker, J. (1988). Policy and the mapping of scientific chance: a co-word analysis of research into environmental acidification. *Scientometrics*, **14**(3), 251-264

Mai, J. E. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation*, **57**(5), 591-622

Marion, L. S. & McCain, K. W. (2001). Contrasting views of software engineering journals: author co-citation choices and indexer vocabulary assignments. *Journal of the American Society for Information Science*, **52**(4), 297-306

Medelyan, O. & Witten, I. H. (2008). Domain-independent automatic key phrases indexing with small training sets. *Journal of the American Society for Information Science and Technology*, **59**(7), 1026-1040

Nguyen, T. D. & Kan, M. Y. (2007). Keyphrase extraction in scientific publications. *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, (pp. 317-326). Berlin: Springer-Verlag. (Lecture Notes in Computer Science, Vol. 4822)

Porter, M. F. (1980). An algorithm for suffix stripping. *program*, **14**(3), 130-137

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, **17**(2), 69-76

Song, M., Yu, H. J. & Han, W. S. (forthcoming). A graph model-driven concept extraction technique for biomedical literatures. *International Journal of Data Mining and Bioinformatics* .

Turney, P. D. (2000). Learning algorithms for key phrase extraction. *Information Retrieval*, **2**(3), 303-336

Turney, P. D. (2003). Coherent key phrase extraction via web mining. In *Proceedings of the 18th international joint conference on Artificial intelligence*, (pp. 434-442). San Francisco, CA: Morgan-Kaufmann.

Wang, H., Peng, H. & Hu, J. S. (2006). Automatic keyphrase extraction from document using neural network. In Daniel S. Yeung, Zhi-Qiang Liu, Xi-Zhao Wang and Hong Yan, (Eds.). In *Advances in Machine Learning and Cybernetics: 4th International Conference, ICMLC 2005, Guangzhou, China, August 2005. Revised Selected Papers*, (pp. 633-641). Berlin: Springer-Verlag. (Lecture Notes in Artificial Intelligence, Volume 3930).

Whittaker, J. (1989). Creativity and conformity in science: titles, keywords, and co-word analysis. *Social Studies of Science*, **19**(3), 473-496

Winkler,W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, (pp. 354-359). Alexandria, VA: American Statistical Association. Retrieved 18 November, 2013 from http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf (Archived by WebCite® at http://www.webcitation.org/6LFH2NSHO)

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C. & Nevill-Manning, C. G. (1999). KEA: practical automatic key phrase extraction. In *Proceedings of the Fourth ACM conference on Digital libraries*, (pp. 254-255). New York, NY: ACM Press.

Wolfram, D. & Olson, H.A. (2007).A method for comparing large scale inter-indexer consistency using IR modeling. In *Proceedings of the 35th Annual Conference of the Canadian Association for Information Science*, Retrieved 4 June, 2012 from http://www.cais-acsi.ca/proceedings/2007/wolfram_2007.pdf. (Archived by WebCite® at http://www.webcitation.org/6HsEtqEFz).

Wu, Y. B. & Li, Q. (2008). Document key phrases as subject metadata: incorporating document key concepts in search results. *Information Retrieval*, **11**(3), 229-249

How to cite this paper

Hahm, J.E., Kim, S.Y., Kim, M.C., Song, M. (2013). Investigation into the existence of the indexer effect in key phrase extraction. *Information Research*, **18**(4) paper 594. [Available at http://InformationR.net/ir/18-4/paper594.html]