PAPER

Quality Evaluation for Document Relation Discovery using Citation Information

Kritsada SRIPHAEW[†], ***Student Member***
and **Thanaruk THEERAMUNKONG**[†], ***Regular Member***

SUMMARY Assessment of discovered patterns is an important issue in the field of knowledge discovery. This paper presents an evaluation method that utilizes citation (reference) information to assess the quality of discovered document relations. With the concept of transitivity as direct/indirect citations, a series of evaluation criteria is introduced to define the validity of discovered relations. Two kinds of validity, called soft validity and hard validity, are proposed to express the quality of the discovered relations. For the purpose of impartial comparison, the expected validity is statistically estimated based on the generative probability of each relation pattern. The proposed evaluation is investigated using more than 10,000 documents obtained from a research publication database. With frequent itemset mining as a process to discover document relations, the proposed method was shown to be a powerful way to evaluate the relations in four aspects: soft/hard scoring, direct/indirect citation, relative quality over the expected value, and comparison to human judgment.

key words: *document relations; frequent itemset mining; citation matrix; quality evaluation; document relation evaluation*

1. Introduction

Nowadays, it has become difficult for researchers to follow the state of the art in their area of interest since the number of research publications has increased continuously and quickly. Such a large volume of information brings about serious hindrance for researchers to position their own works against existing works, or to find useful relations between them [1]–[3]. Although the publication of each work may include a list of related articles (documents) as its reference, it is still impossible to include all related works due to either intentional reasons (e.g., limitation of paper length) or unintentional reasons (e.g., naïvely unknown). Enormous meaningful connections that permeate the literatures may remain hidden.

Growing from different fields, known as literature-based discovery, the approach of discovering hidden and significant relations within a bibliographic database has become popular in medical-related fields [4], [5]. As a content-based approach with manual and/or semi-automatic processes, a set of topical words or terms are extracted as concepts and then utilized to find connections among two literatures. Due to the simplicity and practicality of this approach, it was used in several areas by its succeeding works [6]–[8]. Some works proposed citation analysis based on so-called bibliographic coupling [1], and co-citation [2]. While they were successfully applied in several works [9]–[11] to obtain topical related documents, they are not fully automated and have a lot of labor intensive tasks. Based on association rule mining, an automated approach to discover relations among documents in a research publication database was introduced [12]. Mapping a term (a word or a pair of words) to a transaction in a transactional database, the topic-based relations among scientific publications are revealed under various document representations. Although the work expressed the first attempt to find document relations automatically by exploiting terms in documents, it utilized only simple evaluation without elaborate consideration.

There has been little exploration of how to evaluate document relations discovered from text collections. Most works in text mining utilized a dataset, which includes both queries and their corresponding correct answers, as a test collection. They usually defined certain measures and used them for performance assessment on the test collection. For instance, classification accuracy is applied for assessing the class to which a document is assigned in text categorization (TC) [13] while recall and precision are used to evaluate retrieved documents with regard to given query keywords in information retrieval (IR) [14]. As a more naive evaluation method, human judgment have been used in more recent works on mining web documents, such as HITS [15] and PageRank [16], where there is no standard dataset. However, this manual evaluation is a labor intensive task and quite subjective.

Compared to TC and IR, the evaluation of discovered document relations is difficult and complicated. For one reason, the process to prepare correct answers in the test collection is labor-intensive with the exponential number of candidate relations (a relation may involve more than two documents) to be evaluated. Moreover, there is a lack of standard criteria for evaluating document relations. So far, while there have been several benchmark datasets, e.g., UCI Data Reposi-

Manuscript received December 15, 2006.
Manuscript revised March 29, 2007.
Final manuscript received April 24, 2007.

[†]The authors are with School of Information and Computer Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand.

tory*, WebKB**, TREC data***, for TC and IR tasks, there is no standard dataset that is used for the task of document relation discovery.

Toward resolving these issues, this work proposes a method to use citation information in research publications as a source for evaluating the discovered document relations. Conceptually, the relations among documents can be formulated as a subgraph where each node represents a document and each arc represents a relation between two documents. Based on this formulation, a number of scoring methods are introduced for evaluating the discovered document relations in order to reflect their quality. Moreover, this paper also invents a generative probability that is derived from probability theory and uses it to compute an expected score to capture objectively how good evaluation results are.

Section 2 presents a method for discovering document relations using frequent itemset mining. In Section 3, a series of measures called v-validity is defined on direct/indirect citations formulated by so-called order accumulative citation matrices. Soft validity and hard validity as well as the method to estimate the expected validity are also proposed in this section. Section 4 displays the evaluation results on various document representations including the comparison with the statistical generative probability and human judgment. Finally, a conclusion is made in Section 5.

2. Document Relation Discovery using Frequent Itemset Mining

In the past, frequent itemset mining (FIM) was well-known as a process to find co-occurrences (frequent patterns) in a database. As a prominent technique in association rule mining (ARM), it is useful in various applications such as market basket analysis, fraud detection, data classification, etc. ARM was first applied to discover document relations among scientific publications in [12]. By encoding documents as items, and terms in the documents as transactions, a frequent itemset that we can find will be in the form of "a set of documents" which share a large number of terms. Each discovered document set (for short, *docset*) can be assumed as a content-based relation among documents where this relation is introduced by the coincident terms.

A formulation of the ARM task on document relation discovery can be summarized as follows. Let $\mathcal{D}$ be a set of documents (items) where $\mathcal{D} = \{d_1, d_2, ..., d_m\}$, and T be a set of terms (transactions) where $\mathcal{T} = \{t_1, t_2, ..., t_n\}$. Also let $\delta(d_i, t_j)$ represents the existence (0 or 1) of a term t_j in a document d_i. A subset of $\mathcal{D}$ is called a docset whereas a subset of $\mathcal{T}$ is called a termset. Furthermore, a docset $X = \{x_1, x_2, ..., x_k\} \subset \mathcal{D}$ with k documents is called k-docset (or a docset with the length of k). The support of X is defined as follows.

$$sup(X) = \frac{\sum_{j=1}^{n} min_{i=1}^{k} \delta(x_i, t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} \delta(d_i, t_j)}$$

Here, an X that has a support greater than a predefined minimum support is called a *frequent k-docset*. We will use the term "docset" in the meaning of "frequent docset" and "document relation" interchangeably.

In the document-term database, a set of terms used for representing the documents has some effects on the characteristic of discovered docsets. Therefore, it is necessary to investigate document representation that is suitable for discovering high-quality document relations. In several text processing applications, three schemes of term definition, i.e., n-gram, stemming and stopword removal, were successfully applied. Different combinations of these schemes result in different representations for a document-term database. With different representations, one will obtain different sets of document relations. Here, we need some kind of evaluation to assess which document relations are better. Toward resolving this issue, the next section presents the evaluation method which can be applied to measure the quality of any document relations based on some reasonable criteria.

3. Empirical Evaluation using Citation Information

This section presents a method to use citations (references) among technical documents in a scientific publication collection to evaluate the quality of the discovered document relations. Intuitively, two documents are expected to be related under one of the three basic situations: (1) one document cites to the other (direct citation), (2) both documents cite to the same document (bibliographic coupling) [2] and (3) both documents are cited by the same document (co-citation) [1]. An analysis of citation has been applied for several interesting applications [9]–[11].

Besides these basic situations, two documents may be related to each other via a more complicated concept called transitivity. For example, if a document A cites to a document B, and the document B cites to a document C, then one could assume a transitive relation between A and C. In this work, with the transitive property, the concept of order citation is originally proposed to express an indirect connection between two documents. With the assumption that a direct or indirect connection between two documents implies topical relation among them, such connection can be used for evaluating the results of document relation discovery.

In the rest of this section, introductions of the u-th order citation and v-th order accumulative citation

*http://www.ics.uci.edu/∼mlearn/MLRepository.html
**http://www.webkb.org/
***http://trec.nist.gov/data.html

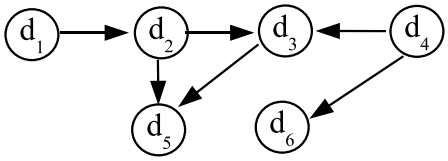


Fig. 1 An example of a citation graph.

doc.	d_1	d_2	d_3	d_4	d_5	d_6
d_1	[1,1,1]	[1,1,1]	[0,1,1]	[0,0,1]	[0,1,1]	[0,0,0]
d_2	[1,1,1]	[1,1,1]	[1,1,1]	[0,1,1]	[1,1,1]	[0,1,1]
d_3	[0,1,1]	[1,1,1]	[1,1,1]	[1,1,1]	[1,1,1]	[0,1,1]
d_4	[0,0,1]	[0,1,1]	[1,1,1]	[1,1,1]	[0,1,1]	[1,1,1]
d_5	[0,1,1]	[1,1,1]	[1,1,1]	[0,1,1]	[1,1,1]	[0,1,1]
d_6	[0,0,0]	[0,0,1]	[0,1,1]	[1,1,1]	[0,0,1]	[1,1,1]

Fig. 2 The 1-, 2- and 3-OACMs: each elements in the table is represented by a set of values $[\delta^1, \delta^2, \delta^3]$.

matrix are given. Then, the so-called validity is proposed as a measure for evaluating discovered docsets using information in the citation matrix. Finally, the expected validity is mathematically defined by exploiting the concept of generative probability and estimation.

3.1 The Citation Graph and Its Matrix Representation

Conceptually citations among documents in a scientific publication collection form a citation graph, where a node corresponds to a document and an arc corresponds to a direct citation of a document to another document. Based on this citation graph, an indirect citation can be defined using the concept of transitivity. The formulation of direct and indirect citations can be given in the terms of the u-th order citation and the v-th order accumulative citation matrix as follows.

Definition 1: (**the u-th order citation**): Let $\mathcal{D}$ be a set of documents (items) in the database. For $x, y \in \mathcal{D}$, y is the u-th order citation of x iff the number of arcs in the shortest path between x to y in the citation graph is u (≥ 1). Conversely, x is also called the u-th order citation of y.

For example, given a set of six documents $d_1, d_2, d_3, d_4, d_5, d_6 \in \mathcal{D}$ and a set of six citations, d_1 to d_2, d_2 to d_3 and d_5, d_3 to d_5, and d_4 to d_3 and d_6, the citation graph can be depicted in Figure 1. In the figure, d_1, d_3 and d_5 is the first, d_4 is the second, and d_6 is the third order citation of the document d_2. Note that although there is a direction for each citation, it is not taken into account since the task is to detect a document relation where the citation direction is not concerned. Moreover, using only textual information without explicit citation or temporal information, it is difficult to find the direction of the citation among any two documents.

Based on the concept of the u-th order citation, the v-th order accumulative citation matrix is introduced to express a set of citation relations stating whether any two documents can be transitively reached by the shortest path shorter than $v + 1$.

Definition 2: (**the v-th order accumulative citation matrix**): Given a set of n distinct documents, the v-th order accumulative citation matrix (for short, v-OACM) is an $n \times n$ matrix, each element of which represents the citation relation δ^v between two documents x, y where $\delta^v(x, y) = 1$ when x is the u-th order citation of y and $u \leq v$, otherwise $\delta^v(x, y) = 0$. Note that $\delta^v(x, y) = \delta^v(y, x)$ and $\delta^v(x, x) = 1$.

For the previous example, the 1-, 2- and 3-OACMs can be created as shown in Figure 2. The 1-OACM can be straightforwardly constructed from the set of the first-order citation (direct citation). The $(v + 1)$-OACM (mathematically denoted by a matrix A^{v+1}) can be recursively created from the operation between v-OACM (A^v) and 1-OACM (A^1) according to the following formula.

$$a_{ij}^{v+1} = \vee_{k=1}^{n}(a_{ik}^{v} \wedge a_{kj}^{1}) \quad (1)$$

where $\vee$ is an OR operator, $\wedge$ is an AND operator, a_{ik}^v is the element at the i-th row and the k-th column of the matrix A^v and a_{kj}^1 is the element at the k-th row and the j-th column of the matrix A^1. Note that any v-OACM is a symmetric matrix.

3.2 Validity: Quality of Document Relations

This section defines the validity which is used as a measure for evaluating the quality of the discovered docsets. The concept of validity calculation is to investigate how documents in a discovered docset are related to each other according to the citation graph. Based on this concept, the most preferable situation is that all documents in a docset directly cite to and/or are cited by at least one document in that docset, and thereafter they form one connected group. Since in practice only few references are given in a document, it is quite rare and unrealistic that all related documents cite to each other. As a generalization, we can assume that all documents in a docset should cite to and/or are cited by each other within a specific range in the citation graph. Here, the shorter the specific range is, the more restrictive the evaluation becomes. With the concept of v-OACM stated in the previous section, we can realize this generalized evaluation by a so-called v-th order validity (for short, *v-validity*), where v corresponds to the specific range mentioned above.

Regarding the criteria of evaluation, two alternative scoring methods can be employed for defining the validity of a docset. As the first method, a score is computed as the ratio of the number of citation relations

in which the most popular document in a docset contains to its maximum. The most popular document is a document that has the most relations with the other documents in the docset. Note that, it is possible to have more than one popular document in a docset. The score calculated by this method is called *soft validity*.

In the second method, a more strict criterion for scoring is applied. The score is set to 1 only when the most popular document connects to all documents in the docset. Otherwise, the score is set to 0. This score is called *hard validity*. The formulation of soft v-validity and hard v-validity of a docset X ($X \subset \mathcal{D}$), denoted by $S_S^v(X)$ and $S_H^v(X)$ respectively, are defined as follows.

$$\mathcal{S}_S^v(X) = \frac{max_{x \in X}(\sum_{y \in X, y \neq x} \delta^v(x,y))}{|X| - 1} \tag{2}$$

For simplicity, we denote a numerator in the above equation with $max^v(X)$. Then,

$$\mathcal{S}_H^v(X) = \begin{cases} 1 & \text{, if } max^v(X) = |X| - 1 \\ 0 & \text{, otherwise} \end{cases} \tag{3}$$

Here, $\delta^v(x,y)$ is the citation relation defined by Definition 2 in Section 3.1. It can be observed that the soft v-validity of a docset is ranging from 0 to 1, i.e., $0 \leq \mathcal{S}_S^v(X) \leq 1$ while the hard v-validity is a binary value of 0 or 1. In both cases, the v-validity achieves the minimum (i.e., 0) when there is no citation relation among any document in the docset. On the other hand, it achieves the maximum (i.e., 1) when there is at least one document that has a citation relation with all documents in a docset. Intuitively, the validity of a bigger docset tends to be lower than a smaller docset since the probability that one document will cite to and/or be cited by other documents in the same docset becomes lower.

In practice, instead of an individual docset, the whole set of discovered docsets needs to be evaluated. The easiest method is to exploit an arithmetic mean. However, it is not fair to directly use the arithmetic mean since a bigger docset tends to have lower validity than a smaller one. We need an aggregation method that reflects docset size in the summation of validities. One of reasonable methods is to use the concept of weighted mean, where each weight reflects the docset size. Therefore, set soft v-validity and set hard v-validity for a set of discovered docsets $\mathcal{F}$, denoted by $\overline{\mathcal{S}_S}^v(\mathcal{F})$ and $\overline{\mathcal{S}_H}^v(\mathcal{F})$, respectively, can be defined as follows.

$$\overline{\mathcal{S}_S}^v(\mathcal{F}) = \frac{\sum_{X \in \mathcal{F}} w_X \times \mathcal{S}_S^v(X)}{\sum_{X \in \mathcal{F}} w_X} \tag{4}$$

$$\overline{\mathcal{S}_H}^v(\mathcal{F}) = \frac{\sum_{X \in \mathcal{F}} w_X \times \mathcal{S}_H^v(X)}{\sum_{X \in \mathcal{F}} w_X} \tag{5}$$

where w_X is the weight of a docset X. In this work, w_X is set to $|X| - 1$, the maximum value that the validity of a docset X can gain. For example calculation, given the 1-OACM in Figure 2 and $\mathcal{F} = \{d_1d_2, d_1d_2d_4\}$, the set soft 1-validity of $\mathcal{F}$ (i.e., $\overline{\mathcal{S}_S}^1(\mathcal{F})$) equals to $\frac{(1\times\frac{1}{1})+(2\times\frac{1}{2})}{1+2} = \frac{2}{3}$ while the set hard 1-validity of $\mathcal{F}$ (i.e., $\overline{\mathcal{S}_H}^1(\mathcal{F})$) is $\frac{(1\times\frac{1}{1})+(2\times 0)}{1+2} = \frac{1}{3}$.

3.3 The Expected Validity

From Equations 2 and 3, the evaluation of discovered docsets will depend on the citation relation (δ^v), which is represented by v-OACMs. As stated in the previous section, the lower v is, the more restrictive the evaluation becomes. Therefore to compare the evaluation based on different v-OACMs, we need to declare a value, regardless of the restriction of evaluation, to represent the expected validity of a given set of docsets under each individual v-OACM. This section describes the method to estimate the theoretical validity of the set of docsets based on probability theory. Towards this estimation, the probability that two documents are related to each other under a v-OACM (later called *base probability*), need to be calculated. This probability is derived by the ratio of the number of existing citation relations to the number of all possible citation relations (i.e., $2 \times \binom{|\mathcal{D}|}{2} = |\mathcal{D}|^2 - |\mathcal{D}|$) as shown in the following equation.

$$p_v = \frac{\sum_{x,y \in \mathcal{D}, x \neq y} \delta^v(x,y)}{|\mathcal{D}|^2 - |\mathcal{D}|} \tag{6}$$

For example, using the citation relation in Figure 2, the base probabilities for 1-, 2-, and 3-OACMs are 0.40 (12/30), 0.73 (22/30) and 0.93 (28/30), respectively. Note that the base probability of a higher-OACM is always higher than or equal to that of a lower-OACM. Using the concept of expectation, the expected set v-validity ($E(\overline{\mathcal{S}}^v(\mathcal{F}))$) can be formulated as follows.

$$E(\overline{\mathcal{S}}^v(\mathcal{F})) = \frac{\sum_{X \in \mathcal{F}} w_X \times E(\mathcal{S}^v(X))}{\sum_{X \in \mathcal{F}} w_X} \tag{7}$$

$$E(\mathcal{S}^v(X)) = \sum_{\forall Y_i, Y_i \in \beta(X)} (\mathcal{S}(Y_i) \times P^v(Y_i)) \tag{8}$$

where $E(\mathcal{S}^v(X))$ is the expected v-validity of a docset X, $\beta(X)$ is the set of all possible citation patterns for X, $S(Y_i)$ is the invariant validity of Y_i, and $P^v(Y_i)$ is the generative probability of the pattern Y_i estimated from the base probability under v-OACM (p_v). Theoretically, finding possible patterns of a docset can be transformed to the set enumeration problem. Given a docset with the length of k (k-docset), there are $2^{\binom{n}{2}}$ possible citation patterns.

With different scoring methods, an invariant validity is individually defined on each criteria regardless of

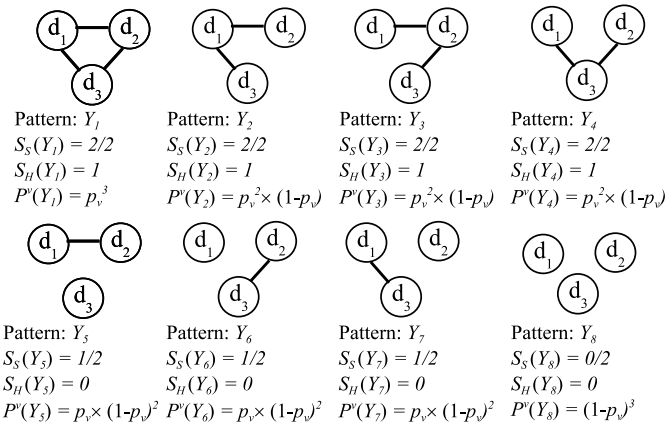


Fig. 3 All possible citation patterns for a 3-docset.

the v-OACM. To simplify this, the notation $\mathcal{S}(Y_i)$ is replaced by $\mathcal{S}_S(Y_i)$ and $\mathcal{S}_H(Y_i)$ for the invariant validity calculated from soft validity and hard validity, respectively. Similar to Equation 2, an invariant validity of Y_i for soft validity is defined as follows:

$$\mathcal{S}_S(Y_i) = \frac{max_{x \in Y_i}(\sum_{y \in Y_i, y \neq x} \delta^{Y_i}(x,y))}{|Y_i| - 1} \quad (9)$$

For simplicity, we denote a numerator in the above equation by $max^{Y_i}(Y_i)$. With another case derived from Equation 3, an invariant validity of Y_i based on hard validity is given by:

$$\mathcal{S}_H(Y_i) = \begin{cases} 1 & \text{, if } max^{Y_i}(Y_i) = |Y_i| - 1 \\ 0 & \text{, otherwise} \end{cases} \quad (10)$$

In the above equations, $\delta^{Y_i}(x,y)$ is the citation relation among two documents x and y in the citation pattern Y_i where $\delta^{Y_i}(x,y) = 1$ when citation relation exists, otherwise $\delta^{Y_i}(x,y) = 0$. Note that all Y_i's have the same docset but represent different citation patterns. The following shows two examples of how to calculate the expected v-validity for 2-docsets and 3-docsets. For simplicity, the expected v-validity based on soft validity is firstly described, and the one based on hard validity is discussed later.

With the simplest case, there are only two possible citation patterns for a 2-docset. Therefore, the expected v-validity based on soft validity of any 2-docset (X) can be calculated as follows.

$$E(\mathcal{S}_S^v(X)) = \frac{1}{1}p_v + \frac{0}{1}(1 - p_v) = p_v \quad (11)$$

In the case of a 3-docset, there are eight possible patterns as shown in Figure 3. From Equation 9, we can calculate the invariant validity based on soft validity ($\mathcal{S}_S$) of each pattern as follows. The first to fourth patterns have the invariant validity of 1 (i.e., $\frac{2}{2}$). The fifth to seventh patterns gain the invariant validity of 0.5 (i.e., $\frac{1}{2}$) while the last pattern occupies the invariant validity of 0 (i.e., $\frac{0}{2}$). The generative probability of the first pattern is p_v^3 since there are three citation relations, and that of the second to the fourth patterns equal to $p_v^2(1-p_v)$ since there are two citation relations and one missing citation relation. Regarding the citation pattern, the generative probabilities of the other patterns can be calculated in the same manner. Applying Equation 8 and the generative probabilities shown in Figure 3, the expected v-validity based on soft validity can be calculated as follows.

$$E(\mathcal{S}_S^v(X)) = 1(\frac{2}{2}p_v^3) + 3(\frac{2}{2}p_v^2(1-p_v)) + 3(\frac{1}{2}p_v(1-p_v)^2) + 1(\frac{0}{2}(1-p_v)^3) \quad (12)$$

Here, the first term comes from the first pattern, the second term is derived from the second to the fourth patterns, the third term is obtained by the fifth to the seventh patterns and the last term is for the eighth pattern.

With another criterion of hard validity, the expected v-validity for a 2-docset is still the same but a difference occurs for a 3-docset. By Equation 10, the invariant validity based on hard validity ($\mathcal{S}_H$) equals to 1 for the first to fourth patterns and becomes 0 for the other patterns. The expected v-validity for a 3-docset based on hard validity is then reduced to

$$E(\mathcal{S}_H^v(X)) = 1(1 \times p_v^3) + 3(1 \times p_v^2(1-p_v)) \quad (13)$$

All above examples illustrate the calculation of the expected validity of only one docset. To calculate the expected v-validity of several docsets in a given set, the weighted mean of their validities can be derived by Equation 7. The outcome will be used as the expected value for evaluating the results obtained from our method for discovering document relations.

4. Experimental Settings and Results

This section presents a set of experimental results when the quality of discovered docsets is investigated under several empirical evaluation criteria. The four main objectives are (1) to investigate characteristic of the evaluation by soft validity and hard validity on docsets discovered from different document representations including their minimum support thresholds and mining time, (2) to study the quality of discovered relations when using either direct citation or indirect citation as the evaluation criteria, (3) to present the relative quality of a discovered relation when it is compared to its statistical expected value, and (4) to show the quality of discovered relations evaluated by human and compare the results with those from the proposed evaluation method.

Towards the first objective, several term definitions are explored in the process of encoding the documents.

To define terms in a document, techniques of n-gram, stemming and stopword removal can be applied. The discovered docsets are ranked by their supports, and then the top-N ranked relations are evaluated using both soft validity and hard validity. Here, the value of N can be varied to observe the characteristic of the discovered docsets. For the second objective, the evaluation is performed based on various v-OACMs, where the 1-OACM considers only direct citation while a higher-OACM also includes indirect citation as shown in Section 3. Intuitively, the evaluation becomes less restrictive when a higher-OACM is applied as the calibration. To fulfill the third objective, the expected set validity for each set of discovered relations is calculated according to the method shown in Section 3.3. Compared to this expected validity, the significance of discovered docsets is investigated. In the last objective, a set of discovered relations is sampled and evaluated by letting a number of experts rate the relatedness of documents in each relation. The result can be used to confirm the potential of our proposed evaluation method.

To implement a mining engine for document relation discovery, the FP-tree algorithm, originally introduced in [17], is modified to mine docsets in a document-term database. In this work, instead of association rules, frequent itemsets are considered. Since a 1-docset contains no relation, it is negligible and then omitted from our evaluation. That is, only the discovered docsets with at least two documents are considered. The experiments were performed on a Pentium IV 2.4GHz Hyper-Threading with 1GB physical memory and 2GB virtual memory running Linux TLE 5.0 as an operating system. The preprocessing steps i.e., n-gram construction, stemming and stopword removal, consume trivial computational time.

4.1 Evaluation Material

There is no gold standard dataset that can be used for evaluating the results of document relation discovery. To solve this problem, an evaluation material is constructed from the scientific research publications in the ACM Digital Library[†]. This dataset was originally used in our previous work [12]. As a seed of constructing the citation graph, 200 publications are retrieved from each of the three computer-related classes, coded by B (Hardware), E (Data) and J (Computer). With the PDF format, each publication is attached with an information page in which citation (i.e., reference) information is provided. The reference publications appearing in these 600 publications are further collected and added into the evaluation dataset. In the same way, the publications referred to by these newly collected publications are also gathered and appended into the dataset. Finally, in total there are 10,817 research publications collected as the evaluation material. After converting these collected publications to ASCII text format, the reference (normally found at the end of each publication text) is removed by a semi-automatic process, such as using clue words of "References" and "Bibliography". With the use of the information page attached to each publication, the 1-OACMs can be constructed and used for evaluating the discovered docsets. Refer to Equation 1, the v-OACM can be constructed from $(v-1)$-OACM and 1-OACM. In our dataset, the average number of citation relations per document is 8 for 1-OACM, 148 for 2-OACM, and 1,008 for 3-OACM. It takes 1.14 seconds for generating 2-OACM from 1-OACM while it takes 15.83 seconds to generate 3-OACM from 2-OACM.

Together with text preprocessing, the BOW library [18] is used as a tool for constructing a document-term database. Using a list of 524 stopwords [19], common words, such as *a*, *an*, *is* and *for*, are discarded. Besides these stopwords, terms with very low frequency are also omitted. These terms are numerous and usually negligible. Moreover, a term occurring less than three times is considered to be insignificant and thus pruned. By this process, the number of terms is dramatically reduced by a factor of 7 to 13. For instance, in case of applying non-stemming, stopword removal and bigram, the number of terms is reduced from 3,866,543 to 283,673. From our observation, the remaining terms in a document still preserve the document contents. In the case of using bigrams as terms, all bigrams are first generated from the original text, and then the bigrams which contain stopwords or have low frequency are pruned. This process will help us to generate pairs of consecutive words, e.g., compound nouns, without the insertion of stopwords.

4.2 Experimental Results

As stated at the beginning of this section, several term definitions can be used as factors to obtain various patterns of document representation. In our experiment, eight distinct patterns are explored. Each pattern is denoted by a 3-digit code. The first digit represents the usage of n-gram, where 'U' stands for unigram and 'B' means bigram. The second digit has a value of either 'O' or 'X', expressing whether the stemming scheme is applied or not. Also the last digit is either 'O' or 'X', telling us whether the stopword removal scheme is applied or not. For example, 'UXO' means document representation generated by unigram, non-stemming and stopword removal. Table 1 expresses the set 1-validity (soft validity/hard validity) of the discovered docsets when various document representations are applied. The minimum support and the execution time of mining for each document representation to discover a specified number of top-N ranked docsets are also given in the table.

[†]http://www.portal.acm.org

Table 1 Set 1-validity for various top-N rankings of discovered docsets, their supports and mining time: soft validity/hard validity (upper: bigram, lower: unigram).
MINSUP: MINIMUM SUPPORT ($\times 10^{-2}$) TIME: MINING TIME (SECONDS)

N	Set Validity (%)			
	BXO	BOO	BXX	BOX
1000	45.47/43.95 MINSUP=0.53,TIME=174.49	46.14/44.33 MINSUP=0.67,TIME=155.92	6.29/6.29 MINSUP=3.94,TIME=442.95	7.09/7.09 MINSUP=4.76,TIME=402.14
5000	29.31/23.88 MINSUP=0.35,TIME=188.88	29.13/27.24 MINSUP=0.47,TIME=166.96	3.83/3.33 MINSUP=3.15,TIME=612.82	3.88/3.59 MINSUP=3.79,TIME=570.65
10000	24.49/19.33 MINSUP=0.32,TIME=189.52	24.40/20.50 MINSUP=0.39,TIME=170.17	3.13/2.33 MINSUP=2.84,TIME=681.40	3.20/2.63 MINSUP=3.42,TIME=627.61
50000	19.29/ 6.36 MINSUP=0.25,TIME=195.39	18.88/ 8.62 MINSUP=0.29,TIME=176.48	2.46/0.98 MINSUP=2.31,TIME=816.43	2.36/1.19 MINSUP=2.71,TIME=767.25
100000	19.51/ 3.67 MINSUP=0.21,TIME=212.14	18.40/ 4.11 MINSUP=0.28,TIME=176.57	2.30/0.63 MINSUP=2.13,TIME=862.84	2.18/0.77 MINSUP=2.48,TIME=832.77
Average	27.61/19.64 MINSUP=0.33,TIME=192.08	27.39/20.96 MINSUP=0.42,TIME=169.22	3.60/2.71 MINSUP=2.87,TIME=683.29	3.74/3.05 MINSUP=3.43,TIME=640.08

N	Set Validity (%)			
	UXO	UOO	UXX	UOX
1000	3.88/3.78 MINSUP=32.72,TIME=122.49	2.36/2.26 MINSUP=46.35,TIME=74.77	2.79/2.79 MINSUP=55.61,TIME=160.98	1.76/1.76 MINSUP=74.78,TIME=89.39
5000	3.77/3.35 MINSUP=26.98,TIME=240.57	2.38/1.99 MINSUP=40.04,TIME=175.72	2.37/2.28 MINSUP=48.46,TIME=359.18	1.55/1.48 MINSUP=66.84,TIME=198.16
10000	3.47/2.63 MINSUP=24.68,TIME=312.69	2.16/1.53 MINSUP=37.63,TIME=231.41	2.09/1.75 MINSUP=45.66,TIME=466.00	1.35/1.11 MINSUP=63.76,TIME=277.67
50000	2.78/1.44 MINSUP=19.95,TIME=478.97	1.75/0.74 MINSUP=32.26,TIME=412.79	1.68/0.84 MINSUP=39.64,TIME=808.61	1.12/0.49 MINSUP=57.08,TIME=539.55
100000	2.71/1.02 MINSUP=18.37,TIME=564.65	1.68/0.48 MINSUP=30.40,TIME=531.10	1.66/0.57 MINSUP=37.40,TIME=1008.38	1.14/0.32 MINSUP=54.55,TIME=691.02
Average	3.32/2.44 MINSUP=24.54,TIME=343.87	2.06/1.40 MINSUP=37.34,TIME=285.16	2.12/1.64 MINSUP=45.35,TIME=560.63	1.38/1.03 MINSUP=63.40,TIME=359.16

From the table, some interesting observations can be made. First, with the same document representation, soft validity is always higher than or equal to hard validity since the former is obtained by less restrictive evaluation than the latter (see Equation 2 and 3). Both validities involve valid relations between any pair of documents in a discovered docset. A relation between two documents is called valid when there is a link between those two documents under the v-OACM (v=1 in this experiment). The evaluation based on soft validity focuses on the probability that any two documents in a docset will occupy a valid relation. On the other hand, the evaluation based on hard validity concentrates on the probability that at least one document must have valid relations with all of the other documents. For example, in the case of top-100000 ranking with the 'BXO' representation (as shown in Table 1), 19.51% of the relations in the discovered docsets are valid while only 3.67% of the discovered docsets are perfect, i.e., there is at least one document that contains valid relations with all of the other documents in the certain docset. Second, in every document representation, both soft validity and hard validity become lower when more ranks (i.e., top-N ranking with a larger N) are considered. As an implication of this result, our proposed evaluation method indicates that better docsets are located at higher ranks. Third, given two representations, say A and B, if the soft validity of A is better than that of B, then the hard validity of A tends to be higher than that of B. Fourth, the results of the bigram cases ('B**') are much better than those of the unigram cases ('U**'). One reason is that the bigrams are quite superior to the unigrams in representing the content of a document. Fifth, in the cases of bigram, the stopword removal process is helpful while the stemming process does not help much. Sixth, in the cases of unigram, non-stemming is preferable while the stopword removal process is not useful. Finally, the performances of 'BXO' and 'BOO' are comparable and much higher than 'BOX' and 'BXX', while the performance of 'UXO' is much higher than the other unigram cases. However, on average, the 'UXX' seems to be the second best case for the unigram. Since the soft validity is more flexible than the hard validity, a higher soft validity is preferable. Although performance of 'BOO' seems to be slightly better than 'BXO' in the higher ranks, 'BXO' performs better on average. In our task, the performance ranks for bigram are 'BXO' > 'BOO' > 'BOX' > 'BXX' and the performance ranks for unigram are 'UXO' > 'UXX' > 'UOO' > 'UOX'.

In terms of minimum support and computational time, we can conclude as follows. First, since a docset discovered from the bigram cases tends to have a lower support than the unigram cases, it is necessary to set a small minimum support in order to obtain the same number of docsets. Second, the cases with stopword removal run faster than ones without stopword removal since they consider fewer words. Moreover, they tend to have a lower minimum support.

As a more detailed exploration of these four best cases, the soft validity and the hard validity as well as the number of discovered docsets for each docset length

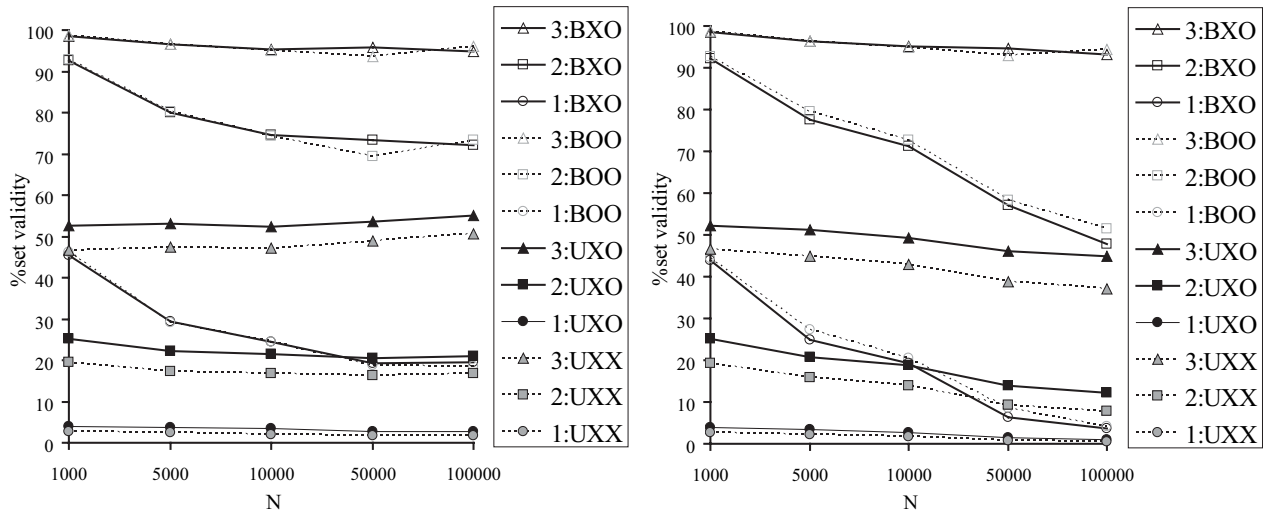


Fig. 4 Set validity based on the 1-, 2- and 3-OACMs when various top-N rankings of discovered docsets are considered: soft validity (left) and hard validity (right).

Table 2 The set 1-validity for each docset length when the top-100000 ranking is considered. Each cell indicates soft validity/hard validity as well as the number of docsets in the bracket.

Docset length	BXO	BOO	UXO	UXX
2	10.86/10.86 (40,870)	11.21/11.21 (38,553)	1.83/1.83 (64,326)	1.31/1.31 (55,262)
3	14.00/4.54 (30,679)	17.35/6.01 (26,174)	3.32/0.35 (33,489)	1.97/0.15 (40,934)
4	20.73/2.05 (10,759)	19.20/1.98 (18,593)	5.09/0.00 (2,181)	1.07/0.00 (3,798)
5	24.40/0.62 (8,004)	21.59/0.66 (13,084)	6.25/0.00 (4)	0.00/0.00 (6)
6	27.07/0.17 (5,266)	24.61/0.09 (3,519)		
7	28.83/0.04 (2,835)	41.31/0.00 (71)		
8	30.60/0.00 (1,168)	45.24/0.00 (6)		
9	32.67/0.00 (347)			
10	35.19/0.00 (66)			
11	38.33/0.00 (6)			
%Set validity	19.51/3.67	18.40/4.11	2.71/1.02	1.66/0.57

are investigated. The result of the top-100000 ranking is shown in Table 2. Due to the space limitation, the results of the other top-N rankings are omitted but they perform in similar manners. From the table, some interesting characteristics are observed: (1) the number of bigger docsets is smaller, (2) compared to the unigram, the bigram produces bigger docsets, (3) in most cases, the soft validity of bigger docsets is higher than that of smaller ones while the hard validity of bigger docsets is lower than that of smaller ones. These observations reflect a good characteristic of the evaluation and match with our expectation.

Besides 1-OACM, the discovered docsets can be evaluated with the criteria of 2-OACM and 3-OACM. In this assessment, only four best representations, two from the unigram cases ('UXO' and 'UXX') and two from the bigram cases ('BXO' and 'BOO'), are taken into consideration. Figure 4 displays the soft validity (the left graph) and the hard validity (the right graph) under 1-, 2-, and 3-OACMs. Since the minimum support and mining time in each case is the same as shown in Table 1, they are omitted from the figure. In the figure, we use the notation to represent the evaluation of docsets under the specified OACM where those docsets are discovered from a specific document representation. For example, '3:BXO' means the evaluation of docsets under 3-OACM where the docsets are discovered by encoding document representation using the BXO scheme (bigram, non-stemming and stopword removal). Being consistent for both soft validity and hard validity, the set 3-validity (one calculated under the 3-OACM) of discovered docsets is higher than the set 2-validity (one calculated under the 2-OACM), and in the same way the set 2-validity is much higher than the set 1-validity (one calculated under the 1-OACM). Compared to the evaluation using only direct citation (1-OACM), more relations in the discovered docsets are valid when both direct and indirect citations (2- and 3-OACMs) are taken into consideration.

Similar to 1-OACM, 'BXO' and 'BOO' are comparable and perform as the best cases for both soft validity and hard validity under the same OACM. Moreover, in the cases of bigram evaluated under the 1- and 2-OACMs, the set validity drops remarkably when top-N rankings with a larger N are focused. The quality of docsets in the higher rank (smaller N) outperforms that of the lower rank. This outcome implies that our evaluation based on direct/indirect citations seems to be a reasonable method for assessing docsets. For all types of document representation, the bigram

Table 3 The actual set validity, the expected set validity and their ratio, for various top-N rankings (soft validity).

Document representation	N	1-OACM			2-OACM			3-OACM		
		actual	expected	ratio	actual	expected	ratio	actual	expected	ratio
BXO	1000	45.47	0.07	676.13	92.56	1.43	64.85	98.47	9.88	9.97
	5000	29.31	0.07	401.43	79.96	1.55	51.64	96.52	10.67	9.04
	10000	24.49	0.07	327.52	74.62	1.59	47.07	95.22	10.89	8.74
	50000	19.29	0.11	180.36	73.40	2.25	32.60	95.77	14.88	6.44
	100000	19.51	0.13	145.03	72.08	2.78	25.96	94.87	16.98	5.59
UXO	1000	3.88	0.06	60.10	25.27	1.37	18.44	52.54	9.49	5.54
	5000	3.77	0.07	56.01	22.32	1.43	15.62	52.98	9.89	5.36
	10000	3.47	0.07	49.92	21.53	1.47	14.61	52.30	10.19	5.13
	50000	2.78	0.08	35.83	20.54	1.65	12.46	53.56	11.38	4.71
	100000	2.71	0.08	32.67	21.03	1.76	11.96	55.11	12.11	4.55

Table 4 The actual set validity, the expected set validity and their ratio, for various top-N rankings (hard validity).

Document representation	N	1-OACM			2-OACM			3-OACM		
		actual	expected	ratio	actual	expected	ratio	actual	expected	ratio
BXO	1000	43.95	0.06	754.31	92.28	1.24	74.33	98.47	8.79	11.20
	5000	24.88	0.05	502.81	77.68	1.06	73.26	96.44	7.75	12.44
	10000	19.33	0.05	402.36	71.30	1.03	69.26	95.10	7.54	12.61
	50000	6.36	0.02	381.96	57.10	0.37	154.06	94.70	3.35	28.30
	100000	3.67	0.01	309.01	47.84	0.27	179.06	93.28	2.49	37.45
UXO	1000	3.78	0.06	59.44	25.07	1.35	18.56	52.24	9.37	5.57
	5000	3.35	0.06	57.67	20.76	1.24	16.76	51.13	8.78	5.83
	10000	2.63	0.05	48.77	18.87	1.15	16.37	49.31	8.32	5.93
	50000	1.44	0.04	37.33	13.97	0.84	16.69	46.00	6.62	6.95
	100000	1.02	0.03	34.05	12.12	0.66	18.28	44.80	5.64	7.95

cases perform better than the unigram cases when they are evaluated under the same v-OACM. Especially the cases under 3-OACM, both two bigram cases ('3:BXO' and '3:BOO') are almost 100% valid while two unigram cases ('3:UXO' and '3:UXX') are approximately 50% valid. This phenomenon shows the advantage of bigram in being a good document representation for document relation discovery and those documents in each docset cite to each other under the specific range within citation graph. Furthermore, the performance gap between bigram and unigram becomes smaller when top-N rankings with a larger N are considered. For a top-N ranking with a larger N, the bigram cases tend to have bigger docsets than the unigram cases and then obtain lower validity since naturally a bigger docset is likely to have lower validity.

In the next experiment, the evaluation is made to investigate the relative quality of discovered docsets against the expected validity. As stated in Section 3.3, to compare the evaluation based on different v-OACMs, the expected validity can be calculated for each individual v-OACM. To do this, the expected set validity is calculated with respect to Equation 7. Using Equation 6, the base probabilities under 1-, 2-, and 3-OACMs (p_1, p_2 and p_3) for our collection are 6.26×10^{-4}, 1.36×10^{-2} and 9.41×10^{-2}, respectively. Due to the space limitation, only the investigation of 'BXO' and 'UXO' are shown here, but the other cases are similar to these two cases. The actual set validity gained from the experiments, the expected set validity calculated from Equation 7 and their ratio are displayed in Table 3 and Table 4, for soft validity and hard validity, respectively. The ratio expresses the quality of the discovered docsets compared to its expected validity.

From the tables, the quality of discovered docsets are significantly high, compared to the expected validity. In principle, the expected validity of a lower-OACM is always lower than or equal to that of a higher-OACM. For our collection, the expected validity of 2-OACM is approximately 20-22 times higher than that of 1-OACM while the expected validity of 3-OACM is about 7-9 times higher than that of 2-OACM. Incidentally this figure is obtained for both soft validity and hard validity. Although it seems that we gain a low set validity for a lower-OACM, but if we compare that validity to its expected validity, we will find out that the ratio is considerably large. That is, the discovered docsets are eligible. For instance, focusing on the top-1000 ranking, although we gained approximately 4% for both soft validity and hard validity under the 1-OACM with the unigram ('UXO'), it corresponds to 60 times over the expected validity. Under the same condition, for the 2- and 3-OACM, we obtained approximately 19 and 6 times over the expected validity, respectively. In the case of bigram ('BXO') and under the 1-, 2- and 3-OACMs, the ratios are approximately 676, 65 and 10 respectively for soft validity, while they raise to approximately 754, 74 and 11, respectively, for hard validity.

By comparing the result to the expected validity, the evaluations under different v-OACMs become comparable with impartial intention. Although the set validity of discovered docsets under a lower-OACM is low, it may be relatively high compared to the expected validity. In Table 3 and 4, although the order of the set validities for different OACMs is 3-OACM > 2-OACM > 1-OACM for given discovered docsets, the order of their

Table 5 Relatedness scores given by experts (BXO vs. UXO).

N	#Samples	% Average Relatedness	
		BXO	UXO
1000	10	77.08	21.25
5000	50	48.25	16.00
10000	100	34.46	12.17

ratios is 1-OACM > 2-OACM > 3-OACM. This result indicates that although the proposed method gains a low value of the set validity for 1-OACM, the result value is quite good compared to the expected value.

In the last experiment, we evaluate the quality of discovered docsets with the answers from human evaluators. Since it is a time consuming task to judge the quality of discovered docsets by hand, some discovered docsets from each top-N ranked docsets are systematically selected as representative samples. One docset from each chunk of one hundred ranked docsets is selected. Then, we get 10, 50 and 100 docsets as the samples for top-1000, top-5000 and top-10000 ranked docsets, respectively. With the limitation of a labor-intensive task, we investigate the docsets discovered from two cases that we focus in the work, i.e., 'BXO' and 'UXO'. Therefore, 320 docsets in total are selected for human judgment. To indicate the relatedness of each docset, four experts holding Ph.D. degrees in computer science or engineering were asked to assign scores for those selected relations in random order and without repetitions. The experts carefully read the documents in a docset one by one and assigned a score for their relatedness. The degree of relatedness is classified into three ordinal scales; 0% for 'not related', 50% for 'somewhat related', and 100% for 'related'. The percentage of average relatedness given by humans are shown in Table 5. This result is consensus with the result from the proposed automatic evaluation method. There are two interesting observations in this table. First, the results from bigram case ('BXO') is better than those from unigram case ('UXO') for any top-N rankings. Second, the results show that better docsets can be discovered in the higher ranks rather than the lower ranks. Although only the average relatedness scores are shown here, the individual evaluation result obtained from each expert also preserves the performance order, i.e. 'BXO' has higher relatedness score than 'UXO' and the higher rank has higher relatedness score than the lower rank. These results support that the proposed evaluation method has high potential to use as an alternative method for evaluating the discovered docsets in order to avoid labor-intensive and time-consuming tasks in human evaluation.

5. Conclusions

This work proposes a method to use citation information in research publications as a source for evaluating the discovered document relations. Three main contributions of this work are as follows. First, soft validity and hard validity are developed to express the quality of document relations, where the former focuses on the probability that any two documents in a docset has a valid relation while the latter concentrates on the probability that at least one document in a docset has valid relations with all of the other documents in that docset. Second, a method to use direct and indirect citations as comparison criteria is proposed to assess the quality of docsets. Third, the expected validity is introduced, using probability theory, to relatively evaluate the quality of discovered docset. By comparing the result to the expected validity, the evaluation becomes impartial, even under different comparison criteria. The manual evaluation was also done for performance comparison. Using more than 10,000 documents obtained from a research publication database and frequent itemset mining as a process to discover document relations, the proposed method was shown to be a powerful way to evaluate the relations in four aspects: soft/hard scoring, direct/indirect citation, relative quality over the expected validity, and comparison to human judgment. As a future work, we plan to explore more on association rules, instead of frequent itemsets. By this, we need to consider the direction of relations between documents. Furthermore, a hybrid approach which utilizes both terms in document and citations among documents for discovering document relations is also valuable for further investigation.

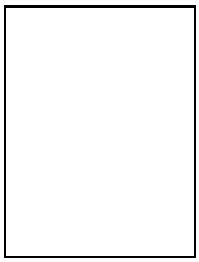
Acknowledgements

This work was supported by Royal Golden Jubilee (RGJ) Ph.D. program of the Thailand Research Fund (TRF) and NECTEC under project number NT-B-22-I4-38-49-05. We also thank to Dr.Cholwich Nattee and Dr.Pakinee Suwannajan on their kind helps in human evaluation.

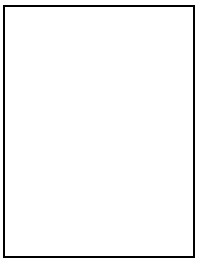
References

[1] M.M. Kessler, "Bibliographic coupling between scientific papers," American Documentation, vol.14, pp.10–25, 1963.
[2] H. Small, "Co-Citation in the scientific literature: a new measure of the relationship between documents," Journal of the American Society for Information Science, vol.42, pp.676–684, 1973.
[3] M. Ganiz, W.M. Pottenger, and C.D. Janneck, "Recent advances in literature based discovery," Journal of the American Society for Information Science and Technology, p.Submitted, 2006.
[4] D. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," Perspectives in Biology and Medicine, vol.30, no.1, pp.7–18, 1986.
[5] D. Swanson, "Medical literature as a potential source of new knowledge," Bulletin of the Medical Library Association, vol.78, no.1, pp.29–37, 1990.
[6] M. Gordon and S. Dumais, "Using latent semantic indexing for literature based discovery," Journal of the American Society for Information Science, vol.49, no.8, pp.674–685,

1998.
[7] R. Lindsay and G. M.D., "Literature-based discovery by lexical statistics," Journal of the American Society for Information Science, vol.50, no.7, pp.574–587, 1999.
[8] W. Pratt, M. Hearst, and L. Fagan, "A knowledge-based approach to organizing retrieved documents," Proc. of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), Orlando, pp.80–85, 1999.
[9] H. Nanba, N. Kando, and M. Okumura, "Classification of research papers using citation links and citation types: Towards automatic review article generation," Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning, Chicago, USA, pp.117–134, Morgan Kaufmann Publishers, San Francisco, US, 2000.
[10] H. White and K. McCain, "Bibliometrics," Annual review on information science and technology, ed. M. Williams, Amsterdam, Netherlands, pp.119–186, Elsevier Science Publishers, 1989.
[11] R. Rousseau and A. Zuccala, "A classification of author co-citations: definitions and search strategies," J. Am. Soc. Inf. Sci. Technol., vol.55, no.6, pp.513–529, 2004.
[12] K. Sriphaew and T. Theeramunkong, "Revealing topic-based relationship among documents using association rule mining.," Artificial Intelligence and Applications, pp.112–117, 2005.
[13] E. Rosch, Principles of Categorization, pp.27–48, John Wiley & Sons Inc, 1978.
[14] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[15] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol.46, no.5, pp.604–632, 1999.
[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," tech. rep., Stanford Digital Library Technologies Project, 1998.
[17] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," 2000 ACM SIGMOD Intl. Conference on Management of Data, ed. W. Chen, J. Naughton, and P.A. Bernstein, pp.1–12, ACM Press, 05 2000.
[18] A.K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," 1996.
[19] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.



Kritsada Sriphaew received his B.Eng. in Computer Engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand in 2000. He is currently a doctoral candidate in combined M.Sc./Ph.D. program, Information and Computer Technology School, Sirindhorn International Institute of Technology, Thailand. His research interests are data mining, information retrieval and computational linguistic



Thanaruk Theeramunkong received a bachelor degree in Electric and Electronics, and master and doctoral degrees in Computer Science from Tokyo Institute of Technology in 1990, 1992 and 1995, respectively. He manages a text data mining project funded by NECTEC, Thailand. His current research interests include data mining, machine learning, natural language processing, and information retrieval.