Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

Measuring the Validity of Document Relations Discovered from Frequent Itemset Mining

Kritsada Sriphaew and Thanaruk Theeramunkong
Sirindhorn International Institute of Technology
131 Moo 5, Tivanond Rd.
Bangkadi, Mueang, Pathumthani, Thailand
Email: kong@siit.tu.ac.th, thanaruk@siit.tu.ac.th

Abstract— **The extension approach of frequent itemset mining can be applied to discover the relations among documents. Several schemes, i.e., n-gram, stemming, stopword removal and term weighting, can be applied to form different document representations for mining. It is necessary to formulate a benchmark for comparing the quality of discovered relations extracted from various document representations. This work proposes a series of evaluation criteria, called order accumulative citation matrix, which is formulated from the citation information in the publications. A new measure, called validity, is presented to reflect the validity (or quality) of discovered relations based on the proposed evaluation criteria. Regarding to the dataset, the expected validity is determined as a baseline for each set of discovered relations. With more than 10,000 documents, the experimental results show that the document document relations using bigram as term definition are more valid than those using unigram with a gap of 13% to 35%. Although the term frequency weighting can improve the validity of discovered document relations when applying unigram as term definition, the binary weighting performs better in the case of bigram. Comparing to the baseline, the results show that the discovered document relations are significantly more valid than the expectation with the factor of 10 to 1,000.**

I. Introduction

In recent years, explosive growth in research publication has made the difficulty for researchers to follow the state of the art in their area of interest. The 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], [2], [3]. Although the publication of each work may include a list of related publications 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., naively unknown). Enormous meaningful connections that permeate the literatures may remain hidden. So far although there have been several approaches to find relations among texts, still very few attempts are made to fully automate the process of discovering relations at the level of document [4]. Some works proposed citation analysis based on so-called bibliographic coupling [1] and co-citation [2]). Although they were successfully applied in several works [5], [6], [7] to obtain topical-related documents, they are not fully automated with a lot of labor intensive tasks.

Towards these problems, The extended approach of frequent itemset mining was proposed in [8] to extract connections of topical-related documents. However, a set of discovered relations is varied according to the scheme of term definition and term weighting used to form document representation. Even if we get a set of relations, it is not an easy task to evaluate which set of relations is better than the others. In this work, we propose a standard evaluation method to measure the validity of discovered relations extracted from various combination of term definition and term weighting schemes. Using citation information in the publications, we can formulate a series of order accumulate citation matrix as the evaluation criteria. Based on this evaluation criteria, the validity is originally proposed to reflect the quality of discovered relations. Moreover, we present an approach to calculate the baseline by expecting the validity of discovered relations regardless to the difficulty level of evaluation criteria.

In the rest, section 2 gives a background of extended frequent itemset mining for document relation discovery. More detailed explanations can be found in [8]. Section 3 proposed an evaluation method which consists of a series of definitions formulated from citation information and the validity measure. A number of experimental results using the proposed validity measure as a measurement are presented, and the comparison of results with the theoretical baseline is given in section 4. Finally, a conclusion is made in section 5.

II. Extended Frequent Itemset Mining for Document Relation Discovery

Unlike most frequent itemset mining (FIM) works on Boolean-valued database, the extended approach addresses to mine frequent itemsets from a real-valued database. Here, the real value indicates a weight of an attribute in the transaction. In the task of mining frequent itemsets, minimum support ($minsup$), a user-specified threshold, is used to filter out the itemsets of which their supports lower than this threshold, considered as infrequent itemsets. 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” (later called *docset*) which share a large number of terms.

Let $\mathcal{D}$ be a set of documents where $\mathcal{D} = \{d_1, d_2, ..., d_m\}$, and $\mathcal{T}$ be a set of terms where $\mathcal{T} = \{t_1, t_2, ..., t_n\}$. Let $w(d_i, t_j)$ represent a weight between a document d_i and a term t_j. A subset of $\mathcal{D}$ is called a docset where a subset of $\mathcal{T}$

1-4244-0705-2/07/$20.00 ©2007 IEEE

is called a termset. A docset $X = \{x_1, x_2, ..., x_k\} \subset \mathcal{D}$ with k documents is called k-docset.

Traditionally, the support of an itemset is defined by a percentage of the number of transactions in which that itemset occurs as a subset to the total number of all transactions in a database. Extended to the real-valued database, the conventional definition of support has to be generalized to take item weights into consideration instead of only item existences as in boolean-valued database. To this end, the equation of support needs to be extended to support the calculation on both boolean-valued and real valued databases as follows.

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

This new equation of support still preserves two closure properties of itemsets [9], i.e., downward closure property ("all subsets of a frequent itemset are also frequent"), and upward closure property ("all supersets of an infrequent itemset are also infrequent"). So far these properties have been applied in most existing FIM algorithms to reduce large computational time. From the extended frequent itemset mining, each discovered docset is in the form of set of documents where all documents in a set are assumed to be related with each other. In the document-term database, various combination of term definition and term weighting schemes, i.e., n-gram, stemming, stopword removal and term weighting, can be applied to formulate many document representations. With those different document representations, different sets of docsets will be discovered. Therefore, it is necessary to evaluate the quality of discovered docsets based on some reliable criteria.

III. The Proposed Evaluation Method

This sections presents a method to use citations (references) among documents in scientific publication collection to evaluate the quality of discovered relations. Intuitively two documents are expected to be related under either of the three basic situations; (1) one document cites to the other (direct citation), (2) both documents cite to the same document (co-citation) [2] and (3) both documents are cited by the same document (bibliographic coupling) [1]. An analysis of citation have been applied for several interesting applications [5], [6], [7].

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 transitively the document B cites to a document C, then one could assume a relation between A and C. With the transitivity property, the concepts of order citation is originally proposed to express both direct and indirect connections 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, an introduction of the u-th order citation and v-th order accumulative citation matrix are given. The so-called validity is proposed as a measure for evaluating discovered docsets using information in the citation matrix. Finally, the baseline is mathematically defined by exploiting the concept of expected validity.

A. The Citation Network and Its Matrix

Conceptually citations among documents in scientific publication collection form a citation network, where a node corresponds to a document and an arc corresponds to the citation of a document to another document. Based on this citation network, the formulation of direct and indirect citations can be defined 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): 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 network is u ($u \geq 1$). Conversely, x is also called the u-th order citation of y.

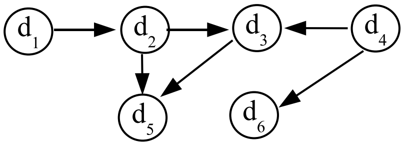

Fig. 1. An example of a citation network.

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 network can be depicted as shown in Figure 1. In the figure, d_2 is the first, d_3 and d_5 is the second, d_4 is the third, and d_6 is the fourth-order citation of the document d_1. As one more example, 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 even there is a direction of citation, it is not taken into account since the task we focus is to detect a sort of document relations where the citation direction is not concerned. Moreover, with only textual information without explicit citation or temporal information, it is hard 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 each other 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.

doc.	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1	1	0	0	0	0
d_2	1	1	1	0	1	0
d_3	0	1	1	1	1	0
d_4	0	0	1	1	0	1
d_5	0	1	1	0	1	0
d_6	0	0	0	1	0	1

1-OACM

doc.	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1	1	1	0	1	0
d_2	1	1	1	1	1	0
d_3	1	1	1	1	1	1
d_4	0	1	1	1	1	1
d_5	1	1	1	1	1	0
d_6	0	0	1	1	0	1

2-OACM

doc.	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1	1	1	1	1	0
d_2	1	1	1	1	1	1
d_3	1	1	1	1	1	1
d_4	1	1	1	1	1	1
d_5	1	1	1	0	1	1
d_6	0	1	1	1	1	1

3-OACM

Fig. 2. 1-OACM, 2-OACM and 3-OACM.

Mapping to the previous example, the 1-OACM, 2-OACM and 3-OACM can be created as shown in Figure 2. The 1-OACM can be straightforwardly constructed from a set of the first-order citation (direct citation). The $(v+1)$-OACM (mathematically denoted by 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.

$$a_{ij}^{v+1} = \vee_{k=1}^{n} (a_{ik}^{v} \wedge a_{kj}^{1}) \tag{2}$$

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

B. 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 much documents in a discovered docset are related to each other in the citation network. 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 the generalization, we can assume that all documents in a discovered docset should cite to and/or are cited by each other within a specific range in the citation network. Here, the shorter the specific range is, the more restrict the evaluation is. 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. The formulation of the *v-validity of a docset* X ($X \subset D$), denoted by $S^v(X)$, is defined as follows.

$$\mathcal{S}^v(X) = \frac{max_{x \in X}(\sum_{y \in X \wedge y \neq x} \delta^v(x,y))}{|X|-1} \tag{3}$$

Here, $\delta^v(x,y)$ is the citation relation defined in by Definition 2. In the equation, we can observe that the v-validity of a docset is ranging from 0 to 1, i.e., $0 \leq \mathcal{S}^v(X) \leq 1$. 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. To get the same value of validity, a bigger docset needs to have more citation relations than a shorter one since it has a larger value of denominator ($|X|-1$).

In practical, instead of an individual docset, the whole set of discovered docsets need 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 a lower validity than a smaller one. We need a consolidation method that reflects the 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. Given a set of discovered docsets $\mathcal{F}$, the v-validity of a set $\mathcal{F}$ (later called *set v-validity*)), denoted by $\overline{\mathcal{S}}^v(\mathcal{F})$, can be defined as follows.

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

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 the 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 1-validity of $\mathcal{F}$ (i.e., $\overline{\mathcal{S}}^1(\mathcal{F})$) equals to $\frac{1\times\frac{1}{1}+2\times\frac{1}{2}}{1+2} = \frac{2}{3}$.

C. The Expected Validity: Baseline

The validity of each docset is varied according to the difficulty level of the evaluation criteria given by v-OACM. As stated in the previous section, the lower v is, the more restrict the evaluation is. This restriction can be assumed as the difficulty level of evaluation criteria. Referred to equation 2 where the higher-OACM is generalized from the lower-OACM, only the citation relation (δ^v) among two documents which is not existing under the lower-OACM may exist under the higher-OACM. According to this fact, the probability that two documents will be related to each other (later called *base probability*) under lower-OACM is higher than such probability under higher-OACM. For example, using the data in Figure 2, the base probability for 1-, 2- and 3-OACMs are 0.40 (6/15), 0.73 (11/15) and 0.93 (14/15), respectively. According to this evidence, the difficulty level of evaluation criteria for 1-OACM is higher than 2-OACM and 3-OACM, and it effects to the low value of set v-validity when using lower-OACM as the evaluation criteria. This is not fair when we want to compare the validity of discovered docsets against different v-OACMs.

Therefore to compare the evaluation based on different v-OACMs, we need to set up a baseline to represent the expected validity of a given set of docsets for each individual v-OACM. Using the concept of expectation, the expected set v-validity, denoted by $E(\overline{\mathcal{S}}^v(\mathcal{F}))$, can be formulated as follows.

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

Since w_X and $\mathcal{S}^v(X)$ are independent, therefore

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

Since w_X is the constant weighting factor of a docset X defined by $|X| - 1$, the formula is then reduced to

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

where $E(\mathcal{S}^v(X))$ is the expected set v-validity of a docset X defined by

$$E(\mathcal{S}^v(X)) = \sum_{\forall Y_i, Y_i \in \beta(X)} (\mathcal{S}^v(Y) \times P^v(Y_i)), \tag{6}$$

$\beta(X)$ is the set of all possible citation patterns for a docset X, and $P^v(Y_i)$ is the generative probability the pattern Y_i estimated from the base probability under v-OACM (p_v). The citation pattern is a form that the documents in a docset cite (or connect) to one another without citation direction. To clarify this, the examples of calculation on 2-docset and 3-docset are described. With the simplest case for a 2-docset, there is only two citation patterns, i.e., all documents is cited and all documents is not cited. The expected set v-validity of any 2-docset can be calculated from the equation 6. Given p_v as the base probability under v-OACM and X as any 2-docset,

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

To generalize for bigger docset, let's assume another example for a 3-docset which contains three documents d_1, d_2 and d_3 as its constituents. All possible citation patterns for a 3-docset are illustrated in Figure 3. Using equation 3 to calculate the validity of each pattern, we get validity 2/2 for patterns 1-4, validity 1/2 for patterns 5-7 and validity 0/2 for pattern 8. Given p_v as the base probability under v-OACM, the probability of citation for each pattern Y_i (denoted by $P^v(Y_i)$) is also shown in the Figure. Given X as any 3-docset,

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

However, these examples show only the calculation on a single docset, but many docsets are discovered in practical. Referred to equation 5, the expected set v-validity represents the weighted mean of validity one expects as the outcome from a set of discovered docsets. This value can be achieved from the summation of expectation on the probability of citation for each docset in a set of discovered docsets, and will be used as a baseline for evaluation.

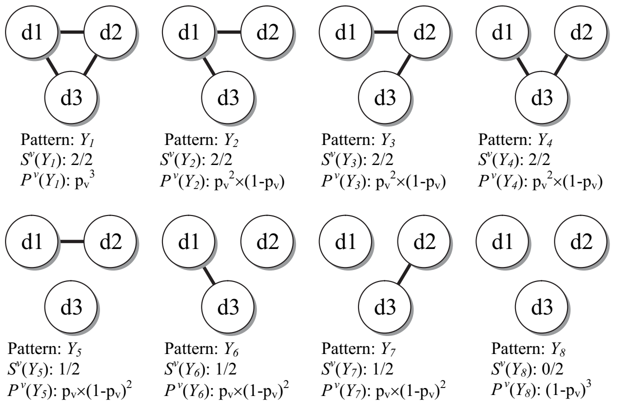

Fig. 3. All possible patterns for citing 3 documents.

IV. Experiments

To study the effect of document representation on the set v-validity, three term definition and one term weighting schemes are applied to form various patterns of document representation for investigation. To define terms in a document, techniques of n-gram, stemming and stopword removal can be applied. For the n-gram scheme, either unigram or bigram is investigated. For the stemming scheme, either stemming or non-stemming is applied. For the stopword removal scheme, either stopword removal or non-stopword removal is applied. For term weighting, either binary weighting or term frequency weighting are taken into consideration. To study the combination of these parameters, sixteen characteristics of document representation for a dataset are generated.

To implement a mining engine, the FP-tree algorithm, originally introduced in [10], is modified to mine docsets in a real-valued database. The discovered frequent itemsets which contain at least two items (discarding 1-docsets) are selected as discovered docsets, and then ranking such docsets by their descending supports. To draw the trends of set v-validity when larger number of knowledge is discovered, the number of discovered docsets used for evaluation is then varied between 1,000 and 100,000 docsets.

A. Test Collection

There is no gold standard corpus, which is coincident with the objective of our approach, available as a benchmark for evaluation. Therefore, a corpus is constructed for this work by the following reasonable method. Possibly to evaluate by the citation relation criteria, the scientific research publications from ACM Digital Library[1] is then retrieved by the following steps. Three classes of CCS (Computing Classification System); B:Hardware, E:Data and J:Computer, are supplied

[1] http://www.portal.acm.org

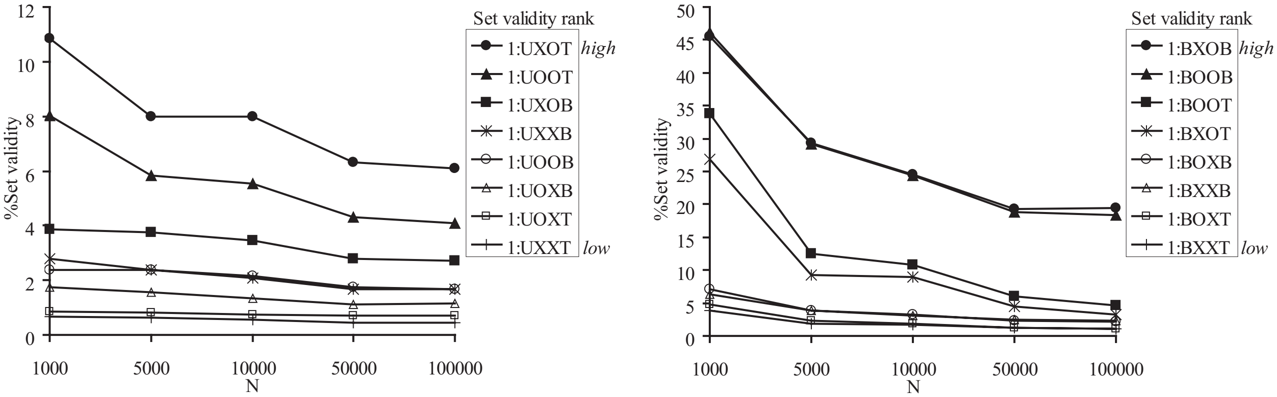

Fig. 4. Set 1-Validity (left: unigram, right: bigram).

as three search keywords. In each class, top 200 related publications in PDF format and their information pages in HTML format are collected as seeds. The links of referenced publications appearing in the seeds are then crawled to gather those publications, and hence augment to the set of seeds. After three iterations, totally 10,817 research publications are gained and used as a corpus for our experiments. After converting all collected publications to ASCII text format, the citation information resides in each text is subsequently removed by using both manual detection and automatic detection in searching lexical cues, such as “References” and “Bibliography”. Moreover, the information pages which were collected during corpus construction are used to construct the v-OACMs which will be further exploited for evaluating the discovered docsets.

For text pre-processing, various characteristics of document representation can be generated by using BOW library [11] as a tool. Although stopword removal scheme can discard meaningless words, some extracted terms are trivial and negligible. To solve this, we assumed that the terms with too low frequencies are unimportant. Using three as a threshold, the terms which have their frequencies lower than this value are considered to be insignificant and thus pruned. The number of terms is dramatically reduced with the factor of 7 to 13. We also apply a trick for pre-processing text by first generating the bigrams and then pruning the bigrams which contain stopwords as their constituents. This will result in getting the real consecutive pairs of words and compound nouns without the insertion of stopwords.

B. Experimental Results

For short reference, each pattern of document representation will be denoted by a 4-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 third digit is either ’O’ or ’X’, telling us whether the stopword removal scheme is applied or not. The last digit indicates which term weighting scheme is applied, where ’B’ means binary weighting and ’T’ means term frequency (*tf*) weighting. For example, ‘UXOT’ means the document representation generated by unigram, non-stemming, stopword removal and *tf* weighting.

Figure 4 shows the set 1-validity using the unigram (left) and bigram (right) as the document representations. Each line in the graph shows the set validity of the given document representation evaluated by v-OACM. For example, 1:UXOT is the percentage of set validity for discovered docsets when applies ‘UXOT’ as document representation and evaluates them under 1-OACM. discovered from and evaluation criteria of each line is encoded by From the figure, some interesting observations can be made. First, applying stopword removal obtains a higher set validity in both unigram and bigram cases. Concretely, the ‘**O*’ gain higher performance than the ‘**X*’ patterns for the same condition. This result is consensus with some reports of the other works in IR and TC areas. Second, stemming is very trivial. There is no dominant difference between ‘*X**’ and ‘*O**’ patterns for the same condition. Third, for term weighting, the unigram cases gain good results when *tf* weighting is applied (‘U**T’) while the binary weighting is useful for the bigram cases (‘B**B’). One possible reason is that *tf* seems to be an important information for the small vocabulary in unigram case while it can be ignored in the bigram case that have rich vocabulary. Fourth, the trend of set validity becomes lower when the number of discovered docsets increases in both unigram and bigram cases. However, the set validity of discovered docsets in the best case when using bigram (‘BXOB’) is dramatically higher than the one in the best case when using unigram (‘UXOT’) with the range of approximately 13%-35%. In the other words, the bigram helps in representing the content of a document more clearer than the unigram. Furthermore, combining both unigram and bigram is also investigated, but the result falls between those of the individual two schemes.

We next evaluate our discovered docsets based on the 2-OACM and 3-OACM by considering on some document representations which perform best set validity. The results

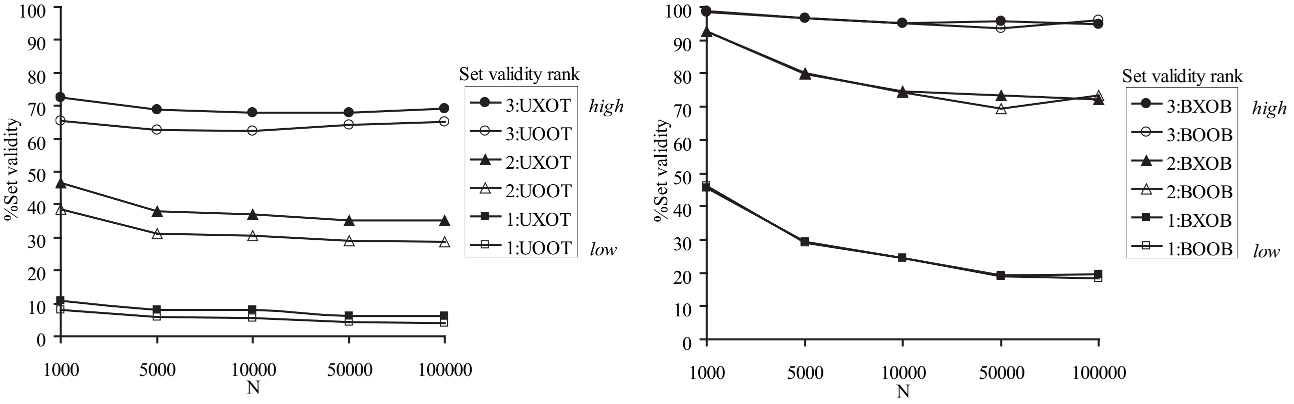

Fig. 5. Set 2-Validity and Set 3-Validity

are shown in Figure 5. The set 2-validity and set 3-validity of discovered docsets are very high since the difficulty level of evaluation criteria is low when the relations among documents need not to be the direct citation as described in Section III-C. Although the set validity evaluated based on 2- or 3-OACMs is extremely high, but the results perform in the same way as previously discussed. Moreover, with the number of discovered docsets increases, the set validity is quite steady in the unigram case but intensely decreasing in the bigram case.

A set of discovered docsets consists of many docsets with different sizes. As previously pointed out, an increasing number of documents in a bigger docset causes the low value of set validity. In practical, the higher number of bigger docset is, the lower value of set validity is. Considering only ‘UXOB’, ‘UXOT’, ‘BXOB’ and ‘BXOT’ as document representations, Figure 6 shows the number of docsets with different sizes whereof total 100,000 discovered docsets are considered. Even the shown graphs are plotted on a specified total number of discovered docsets, but the trend in each case of the different number of discovered docsets is similar to the one shown in Figure 6. In most cases, a set of discovered docsets consists of larger number of smaller docsets than the bigger ones. However, using bigram as document representation is distinctively different from the other cases. In a set of discovered docsets using ‘BXOT’, a fraction of the number of bigger docsets to the total number of discovered docsets is relatively high compared to the other cases. This is another reason why the term frequency weighting performs worser than binary weighting in the bigram case.

From our dataset with 10,817 publications, the base probability that two documents will be related to each other calculated from the 1-, 2- and 3-OACMs are 6.26×10^{-4}, 1.36×10^{-2} and 9.41×10^{-2}, respectively. Using this value of base probability, we further analyze the expected set validity as a baseline to compare the relative quality with the actual set validity. To obviously present the quality of set validity against the baseline, the set validity and the expected set validity for ‘UXOT’ and ‘BXOB’ under each OACM are depicted in the left of Figure 7. In the Figure, the notation ‘E’ is used to represent the baseline or expected validity for individual parameter. For simplifying the illustration, the number of times that the set validity is higher than the expected set validity (relative validity) for each set of discovered docsets is shown in the right of Figure 7. The results are shown in the logarithmic scale over the best document representations for unigram (‘UXOT’) and bigram (‘BXOB’).

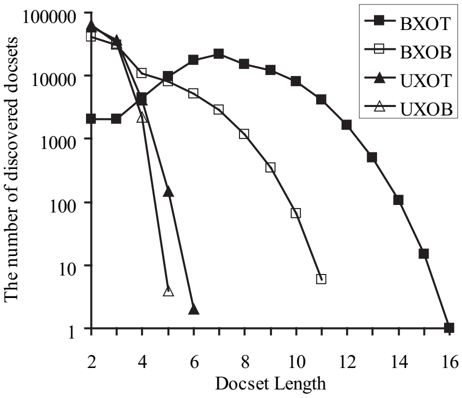

Fig. 6. The number of docsets in each docset length using ‘UXOB’, ‘UXOT’, ‘BXOB’ and ‘BXOT’ as document representations whereof total 100,000 discovered docsets are considered.

As depicted in the left of Figure 7, our approach successes to mine the docsets which are excessively significant according to the baseline for this corpus. Both set validity and baseline tend to be constant in the logarithmic scale when the number of discovered docsets increases. The baseline for the lower-OACM is quite low compared to the baseline for the higher-OACM. Even the set validity (continuous line) for the higher-OACM is high, but the baseline (dashed line) for the higher-OACM is also high. In the right of Figure 7, it shows the relative validity of each scheme in the aspect of the number of times that the set validity is higher than the baseline. In all cases, the relative validity under 1-OACM is higher than 2-OACM and

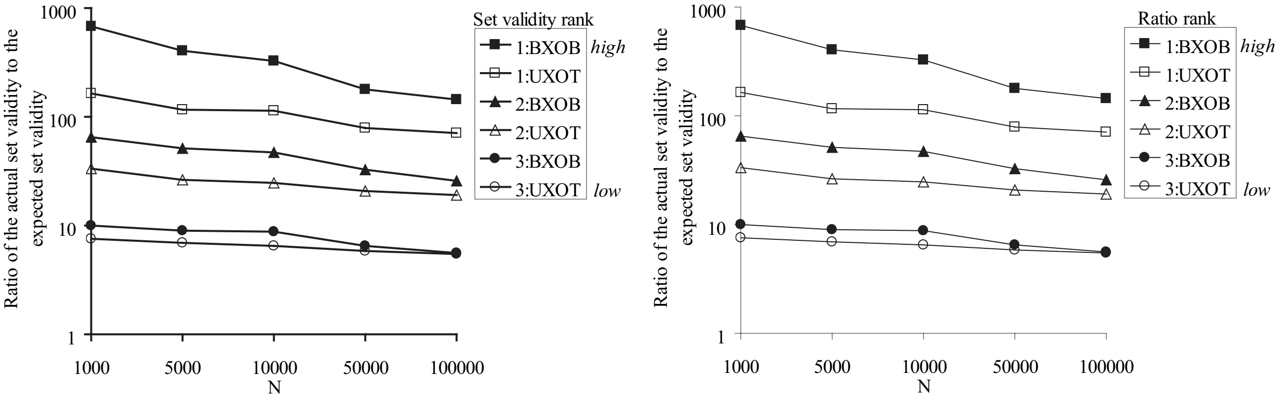

Fig. 7. Set Validity and expected set validity (left), and the ratio of the actual set validity to the expected set validity (right).

3-OACM, respectively. This means that the discovered docsets are highly valid to the direct citations more than the indirect citations. For instance with 1,000 discovered docsets using 'UXOT', the relative quality based on 1-, 2- and 3-OACMs are 165.0, 33.4 and 7.5, respectively. Another example with 1,000 discovered docsets using 'BXOB', the relative quality based on 1-, 2- and 3-OACMs are 685.4, 64.9 and 10.0, respectively. It is noted that the relative quality is exponentially decreasing when evaluating based on the higher-OACM. Based on the same OACM, the relative quality becomes lesser when the number of discovered docsets increases but it is still better than the baseline. Undoubtedly, the relative quality of bigram ('BXOB') is truly better than unigram ('UXOT').

V. Conclusion

This work proposes a standard evaluation method to measure the validity of discovered document relations extracted from the extended frequent itemset mining approach. Specific to the scientific publication domain, the known information of citations resided in the documents is applied to formulate a series of order accumulate citation matrix used as an evaluation criteria. The validity is originally proposed to reflect the quality of discovered document relations. Using the concept of expectation, an approach to calculate the baseline for discovered document relations regardless to the difficulty level of evaluation criteria is also presented in this work. By formulating a collection of documents as document-term database using various combinations of term definition and term weighting schemes, several valid document relations based on their citation information can be discovered. With more than 10,000 documents from a scientific publication collection, the experimental results show that the set validity of discovered docsets from the best case when using bigram is excessively higher than those from the best case when using unigram with the gap of 13%-35%. Furthermore, the term frequency weighting scheme can raise the set validity in the unigram case, while the binary weighting scheme performs better in the bigram case. The approach successes to retrieve the document relations which are excessively significant according to the baseline. Eventually, our proposed evaluation method can reflect the reasonable results of validity for discovered document relations.

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.

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] M. Ganiz, W. Pottenger, and C. Janneck, "Recent advances in literature based discovery," *Journal of the American Society for Information Science*, 2006.

[5] 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.

[6] H. Nanba, N. Kando, and M. Okumura, "Classification of research papers using citation links and citation types: Towards automatic review article generation," in *Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning*. Chicago, USA: Morgan Kaufmann Publishers, San Francisco, US, 2000, pp. 117–134.

[7] H. White and K. McCain, "Bibliometrics," in *Annual review on information science and technology*, M. Williams, Ed. Amsterdam, Netherlands: Elsevier Science Publishers, 1989, pp. 119–186.

[8] K. Sriphaew and T. Theeramunkong, "Revealing topic-based relationship among documents using association rule mining." in *Artificial Intelligence and Applications*, 2005, pp. 112–117.

[9] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," pp. 307–328, 1996.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *2000 ACM SIGMOD Intl. Conference on Management of Data*, W. Chen, J. Naughton, and P. A. Bernstein, Eds. ACM Press, 05 2000, pp. 1–12.

[11] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," 1996.