PAPER *Special Section on Knowledge Discovery, Data Mining and Creativity Support System*

News Relation Discovery Based on Association Rule Mining with Combining Factors

Nichnan KITTIPHATTANABAWON[†a)], ***Student Member***, **Thanaruk THEERAMUNKONG**[†b)], ***and*** **Ekawit NANTAJEEWARAWAT**[†c)], ***Members***

SUMMARY Recently, to track and relate news documents from several sources, association rule mining has been applied due to its performance and scalability. This paper presents an empirical investigation on how term representation basis, term weighting, and association measure affects the quality of relations discovered among news documents. Twenty four combinations initiated by two term representation bases, four term weightings, and three association measures are explored with their results compared to human judgment of three-level relations: completely related, somehow related, and unrelated relations. The performance evaluation is conducted by comparing the top-*k* results of each combination to those of the others using so-called rank-order mismatch (ROM). The experimental results indicate that a combination of bigram (BG), term frequency with inverse document frequency (TFIDF) and confidence (CONF), as well as a combination of BG, TFIDF and conviction (CONV), achieves the best performance to find the related documents by placing them in upper ranks with 0.41% ROM on top-50 mined relations. However, a combination of unigram (UG), TFIDF and lift (LIFT) performs the best by locating irrelevant relations in lower ranks (top-1100) with 9.63% ROM. A detailed analysis on the number of the three-level relations with regard to their rankings is also performed in order to examine the characteristic of the resultant relations. Finally, a discussion and an error analysis are given.

key words: *news relations, news relation discovery, association rule mining, combining factors*

1. Introduction

Recently, most news publishers have provided their electronic news on the web in order to increase the number of their readers. To allow gaining access on these contents efficiently, they usually organize news contents with appropriate structures. In many cases, the readers often avoid bias from a single source of information by reading news from several publishers. As one facilitation, a number of news portals are constructed on the web to provide linkages among news documents from multiple sources. Most of them usually organize news into some kinds of relationship structures, e.g., grouping news documents by category, by recency, or by popularity, summarizing news contents, personalizing news access based on readers' interests, and creating relations between news documents. However, most of these operations are performed manually with a lot of tedious efforts. Towards automated content organization, while classification techniques can be applied to assign a category label to each document based on a number of criteria, such as text genre, text style, and users' interest [1]–[3]. Some of them can be adopted for classifying news documents [4], [5]. By the classification method, it requires users to provide a number of predefined classes and a large number of training examples. Releasing from these requirements, clustering can be used to group documents according to their similar characteristics [6], [7]. As a more complicated task, a multidocument summarization can be performed to obtain a shorter description from a cluster of news describing similar events [8], [9]. For the past several years, event-based topics of news stories has been investigated by Topic Detection and Tracking (TDT) research [10], [11]. Event clustering and first story detection are two main problems in TDT. Normally, by the way of event clustering, news stories that include several events can be grouped into a number of clusters, each of which is about a single news topic. On the other hand, the task of first story detection is to identify whether a news story includes new events which are never seen. Recently, an association rule mining approach [12], [13] has been introduced for discovering document relations in scientific research publications due to its performance and scalability [14], [15]. Given N documents, cluster-based association discovery requires the calculation of all possible combination of documents, that is the complexity of $O(2^N)$. In contrast, the association rule mining approach utilizes a criterion of minimum support to control the searching space, resulting in the complexity of $O(N^L)$, where L is the length of the longest pattern. As for related tasks in Thai, even there exist several works towards extraction of information from online documents, there are very few works on finding document relations. An event classification on Thai news using a similarity measure based on Term Frequency-Inverse Document Frequency was given in [16]. This work attempted to classify news documents by selecting the category using the largest similarity scores. A method for clustering a non-segmented Thai document was proposed in [17]. Their concept combined self-organizing map and frequent max substring technique to generate document clusters. Automatic Thai text summarization technique for single document was presented in [18]. This technique applied content-based feature and graph-based approach to divide a document into a set of segments and to construct a document graph. Their text summary was ex-

Manuscript received May 6, 2010.

Manuscript revised September 29, 2010.

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

a) E-mail: knichcha@wu.ac.th

b) E-mail: thanaruk@siit.tu.ac.th

c) E-mail: ekawit@siit.tu.ac.th

DOI: 10.1587/transinf.E94.D.404

tracted by a graph search technique, identifying a set of segments that represent the content of a document whose similarity score is the most significant. In [19], another approach to Thai news text summarization was proposed to extract the most relevant paragraphs (important portions) from the original document and then reform a summary. As an early work on relation discovery in multiple Thai documents, Kittiphattanabawon and Theeramunkong [20] have proposed a method based on association rule mining to find the relations among Thai news documents. The work gave a preliminary exploration on the performance of mining a pair of relevant news documents by support-confidence and support-conviction measurements under limited environment of top-k ranking evaluation.

In this paper, in addition to support-confidence and support-conviction, support-lift measurement is investigated and compared with human judgment in a more general environment of up-to top-1100 ranking evaluation. Twenty four combinations generated from two term representation bases, four term weightings and three association measures are examined to find optimal combinations for discovering meaningful relations among news documents. In Sect. 2, news relation generation is presented under the formation of association rules. For discovering news association, sets of the combination of factors for mining process are then described in Sect. 3. The generalized association measures are also defined in this section. Section 4 presents evaluation methods including a description of types of news relations, a construction of evaluation dataset, and criteria for evaluation. A number of experimental results and discussion are given in Sect. 5. Finally, a conclusion and future works are made in Sect. 6.

2. Association Rule Mining for Discovering News Relations

Association rule mining (ARM) is well-known as a process to find frequent patterns in the form of rules from a database. Recently ARM or its derivatives have been applied in find relations among documents [15], [20]. By encoding documents as items, and terms in the documents as transactions, we mine a set of frequent patterns, each of which is in the form of a set of documents sharing common terms more than a threshold, called support. Thereafter, as a further step, a set of frequent rules can be found based on these frequent patterns with another threshold, namely confidence. In this work, in order to work with non-binary data, we adopt the generalized support and generalized confidence in [15], and the generalized conviction in [20] as association measures. A formulation of the ARM task on news document relation discovery can be summarized as follows. Assume that $D = \{d_1, d_2, \ldots, d_m\}$ is a set of m news documents (items), T=$\{t_1, t_2, \ldots, t_n\}$ is a set of n terms (transactions), a news itemset X=$\{x_1, x_2, \ldots, x_k\} \subset D$ is a set of k news documents, and a news itemset Y=$\{y_1, y_2, \ldots, y_l\} \subset D$ is a set of l news documents. As an alternative to confidence and conviction, a measure called lift is introduced in this work. Conventionally, the lift of an association rule $X \rightarrow Y$ is defined as $conf(X \rightarrow Y)/sup(Y)$, where $conf(X \rightarrow Y)$ is the confidence value of the rule $X \rightarrow Y$ and $sup(Y)$ is the support value of Y. The generalized support of $X \rightarrow Y$ ($sup(X \rightarrow Y)$), the generalized confidence of $X \rightarrow Y$ ($conf(X \rightarrow Y)$), the generalized conviction of $X \rightarrow Y$ ($conv(X \rightarrow Y)$), and the generalized lift of $X \rightarrow Y$ ($lift(X \rightarrow Y)$) are shown in Table 1, where $w(d_i, t_j)$ repre-

Table 1 Definitions of generalized association measures: (a) generalized support, (b) generalized confidence, (c) generalized conviction, and (d) generalized lift. Here, $D = \{d_1, d_2, \ldots, d_m\}$ is a set of m news documents (items), T=$\{t_1, t_2, \ldots, t_n\}$ is a set of n terms (transactions), a news itemset X=$\{x_1, x_2, \ldots, x_k\} \subset D$ is a set of k news documents, a news itemset Y=$\{y_1, y_2, \ldots, y_l\} \subset D$ is a set of l news documents, and $Z = X \cup Y = \{z_1, z_2, \ldots, z_{k+l}\} \subset D$ with $k + l$ news documents.

(a)

$$\begin{aligned} sup(X,Y) &= sup(X \rightarrow Y) \\ &= sup(Y \rightarrow X) \\ &= \frac{\sum_{j=1}^{n} min_{i=1}^{k+l} w(z_i, t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} w(i_i, t_j)} \end{aligned}$$

(b)

$$conf(X \rightarrow Y) = \frac{\sum_{j=1}^{n} min_{i=1}^{k+l} w(z_i, t_j)}{\sum_{j=1}^{n} min_{i=1}^{k} w(x_i, t_j)}$$

(c)

$$conv(X \rightarrow Y) = \frac{1 - \frac{\sum_{j=1}^{n} min_{i=1}^{l} w(y_i,t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} w(i_i,t_j)}}{1 - \frac{\sum_{j=1}^{n} min_{i=1}^{k+l} w(z_i,t_j)}{\sum_{j=1}^{n} min_{i=1}^{k} w(x_i,t_j)}}$$

(d)

$$\begin{aligned} lift(X,Y) &= lift(X \rightarrow Y) \\ &= lift(Y \rightarrow X) \\ &= \frac{\frac{\sum_{j=1}^{n} min_{i=1}^{k+l} w(z_i,t_j)}{\sum_{j=1}^{n} min_{i=1}^{k} w(x_i,t_j)}}{\frac{\sum_{j=1}^{n} min_{i=1}^{l} w(y_i,t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} w(i_i,t_j)}} \end{aligned}$$

sents a weight of a term t_j in a news document d_i and $Z = X \cup Y = \{z_1, z_2, \ldots, z_{k+l}\} \subset D$ is a set of $k+l$ news documents, composed k news documents in the X and l news documents in the Y. By this method, the discovered relations are in the form of "$X \rightarrow Y$", where X as well as Y is a set of news documents. The relation "$X \rightarrow Y$" with a high association value (high confidence, high conviction, or high lift) indicates that the news documents in the X has a relationship with the news documents in the Y with considerable content overlap among them. In this work, considering a special case of one single antecedent and one single consequent, the rule implies that the news document in the X relates to the news document in the Y. In the traditional ARM [12], [21], minimum support and minimum confidence are defined to filter out trivial rules. Among efficient algorithms such as Apriori [21], CHARM [22], [23] and FP-Tree [24], in this work we select FP-Tree since it is the most efficient mining algorithm that can generate conventional frequent itemsets, not closed frequent itemsets.

3. Association Rules with Combining Factors

In general, the results from the mining process can differ according to the setting factors in the process. In this paper, to find an appropriate environment in discovering the news relations, we explore three main factors for generating association rules, i.e., (1) term representation basis, (2) term weighting, and (3) association measure. For the term representation basis, unigram (UG) and bigram (BG) are investigated as the term representation for the content of news documents. Intuitively, UG may be not sufficient for representing the content of a news document since there exists term ambiguity in the context. As an alternative, BG considers two neighboring terms as a unit in order to handle compound words and then to partially solve the ambiguity of words. For term weighting, binary term frequency weighting (BF), term frequency weighting (TF), and their modification with inverse document frequency weighting (BFIDF, TFIDF) are explored. BF_{ij} simply indicates the existence or non-existence of the j-th term in the i-th news document while TF_{ij} indicates the frequency of the j-th term in the i-th news document. IDF is often used in complementary with BF and TF, to promote a rare term which occurs in very few documents, as an important word. Although IDF_j can be calculated as the total number of documents in the collection (N) divided by the number of documents containing the j-th term (DF), it is usually used in the logarithm scale. Therefore, $BFIDF_j$ and $TFIDF_j$ of the j-th term are defined as $BF_j \times \log(N/DF_j)$ and $TF_j \times \log(N/DF_j)$ respectively. To measure the appropriateness of relations, quantitative measure is another factor, which needs to be carefully selected. In this work, to find a suitable measure, we consider confidence (CONF), conviction (CONV), and lift (LIFT) as association measures. CONF is a well-known rule measure for ARM approach. Some literatures [25]–[28] showed that CONV and LIFT can result in more interesting relations. In our work, CONV and LIFT are investigated to find the most suitable measurement for the association of news documents. The generalized CONF, generalized CONV, and generalized LIFT are summarized in Table 1. From two types of term representation bases, four types of term weightings, and three types of association measures, twenty four combinations are investigated by using an association rule discovery approach.

4. Evaluation Methodology

In this section, we describe an evaluation methodology to investigate the potential of combinations of factors. We first explain a description of news relation types, then we give the details of dataset construction for evaluation. A criterion for evaluation is then given for assessing the quality of discovered news relations.

4.1 News Relation Types

Basically, for document relation discovery, two types of relation, "relevant" and "non-relevant" are always considered to judge a relationship between events. In this work, three main types of news relations are classified based on the relevance of news events: (1) "completely related" (CR), (2) "somehow related" (SH) and (3) "unrelated" (UR) [20]. A CR relation is detected when two new documents mention a same story. Such a CR relation is usually reported by several publishers at the (almost) same time period as a daily newspaper. However, news documents with the CR-type may be presented in different headlines, different phrases, and different writing styles. For SH relation, it is a kind of relation which has only somewhat closely related. The SH relation may exist with either one of three following types. As the first type, namely "similar theme (ST)", two news documents carry similar topics or themes. Any two news documents with the ST-type may involve two different events but they indicate the same topic. For the second type, called "series (SE)", describes the situation that two news documents connect together by forming a sequential time series of events. The last type, named "subnews (SN)", two news documents contain the same event but one may have more details than the others. The relation of UR is defined as a relationship of having absolutely unrelated in their events between news documents. In other words, It could be considered as a non-relevant story.

4.2 Evaluation Dataset

As there is no standard dataset for news relations in Thai available as a benchmark for assessing performance of our approach, we construct our own dataset based on an evaluation of human from 811 Thai news documents of three news online sources (Dailynews (313 documents), Komchadluek (207 documents), and Manager online (291 documents)) during August 14-31, 2007, consisting of three categories (politics (266 documents), economics (250 documents), and crime (295 documents)). The statistics of the constructed

Table 2 Numbers of news documents, classified by publishing source, news category, and document size (in words, excluding stopwords).

Source	Category	Number of Words						Total
		1-100	101-200	201-300	301-400	401-500	>500	
Dailynews	Politics	50	35	0	0	1	5	**91**
	Economics	27	63	15	1	1	0	**107**
	Crime	43	50	15	7	0	0	**115**
	Subtotal	120	148	30	8	2	5	**313**
Komchadluek	Politics	5	34	32	8	6	2	**87**
	Economics	2	34	5	0	0	0	**41**
	Crime	2	21	34	15	3	4	**79**
	Subtotal	9	89	71	23	9	6	**207**
Manager online	Politics	6	52	19	4	5	2	**88**
	Economics	5	36	41	11	5	4	**102**
	Crime	21	32	32	9	4	3	**101**
	Subtotal	32	120	92	24	14	9	**291**
	Total	161	357	193	55	25	20	**811**

news collection characterized by their categories and sizes (the number of words in a news document after applying stopword removal) are shown in Table 2. Our stopword list includes words which may not contribute much to semantics of news documents, but frequently appear in the documents, i.e., conjunction, person title, organization title, and so on. All sources (publishers) publish news documents with fewer than 300 words, and Dailynews tend to write a news story with a shorter text than those from the other two sources. From this news collection, a set of 1,132 news relations (a pair of news documents) with high association values (high CONF, high CONV, or high LIFT) are selected from each top-k relation sets of twenty four combinations (described in Sect. 3). In this work, k is set to 1,000. In the evaluation of human, for each of 1,132 news relations, three assessors who often read news are chosen to judge whether two news documents are related with each other by one of the predefined relation types, i.e., CR, ST, SE, SN, and UR, as explained in Sect. 4.1. Through this judgment, the assessors have been conducted to understand the relationship between news documents before starting to do the task. They are assigned to read the news relation (two news documents), compare the content in both news documents, and set the type (CR, ST, SE, SN, or UR) to this news relation. Here, every news relation is judged by all three assessors. If there are different opinions on the judgments, the final decision is done by voting. However, sometimes voting may not be able to guarantee a majority for such a decision. To decide the answer, an iteration process is performed by asking the assessors to reconsider their judgments until the final decision is made. Finally, the relation types of 1,132 news relations are determined as shown in Table 3, i.e., 65 relations of CR, 571 relations of SH (297 ST, 199 SE, and 75 SN relations), and 496 relations of UR. In the table, the number of news relations are grouped according to whether they are obtained from same or different publishers (sources). The final form of our dataset can be represented in tabular. Each row in the table corresponds to each mined relation while the columns display the values of the association measures (i.e., CONF, CONV, and LIFT) corresponding to each of the combinations, and the relation types judged by human with its corresponding score (i.e., 0.0 for UR-type, 0.5 for SH-type, and 1.0 for CR-type). As our preliminary study, we have focused on three main types (CR, SH, and UR) by ignoring the subtypes of SH-type (ST, SE, and SN).

Table 3 Numbers of news relations grouped by their types based on human judgment.

News Relation	Number of Relations		
Type	Total	Same Source	Different Sources
CR	**65**	20	45
SH	**571**	263	308
- ST	297	149	148
- SE	199	95	104
- SN	75	19	56
UR	**496**	245	251
Total	**1132**	528	604

4.3 Evaluation Criterion

The quality of twenty four combinations in discovering news relations is evaluated by comparing the results generated by each of them to those from human judgments. The evaluation method is applied from a paired-wise comparison technique [29] since the paired-wise comparison has been applicably used for counting the mismatches between rankings. In this work, for each combination, an evaluation is proceeded by creating a ranked list of relations ordered by its association measure (a score suggested by the system), and the other ranked list of relations suggested by human judgment, (i.e., 0.0 for UR, 0.5 for SH, and 1.0 for CR relations). Then, by mapping the resultant from the human judgment list to each of the relations in the system ranked list, a mismatch score between these two ranked lists is calculated by Eq. (1). The quality among twenty four combinations is compared by a criterion, so-called rank-order mismatch (ROM), as shown in Eq. (3).

$$M(A,B) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\delta(r_A(i), r_A(j)) - \delta(r_B(i), r_B(j))| \quad (1)$$

$$\delta(a,b) = \begin{cases} 1 & \text{if } a < b \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$ROM(A,B) = \frac{2 \times M(A,B)}{N(N-1)} \times 100 \quad (3)$$

In Eq. (1), $M(A, B)$, the mismatch score, indicates the number of rank mismatches between two ranked list, say A and B, given a set of N objects to be ranked. The mismatch score expresses how much conflict the ranked list A has with the ranked list B. The $r_A(k)$ and $r_B(k)$ are the respective ranks of the k-th objects based on the ranked lists A and B respectively. A mismatch function, $\delta(a, b)$, returns 1 when a less than b, otherwise 0, as shown in Eq. (2). Such a function indicates that a relation in an upper rank (a) which has a score lower than one in a lower rank (b) presents in a mismatch order. The so-called rank-order mismatch (ROM) in Eq. (3), ranging between 0 and 100, is a calculation of dividing a mismatch score ($M(A, B)$) with the mismatch score of the worst case, the case of a ranked list A are arranged in the reverse order compared to the other ranked list B. As for our work, the ROM is calculated by setting A to the ranked list produced by the system (our proposed method) and B to the ranked list suggested by human (h). For compactness, the $ROM(A, B)$ is denoted by $ROM_h(A)$. If all news relations are ranked by the proposed method (A) in the same order with the human judgment, the $ROM_h(A)$ becomes 0. Implicitly, the ROM value reflects the amount of incorrect relation types suggested by the system, compared to the human judgment.

As stated above, the constructed rank order is arranged by the association measure. In this work, CONF, CONV and LIFT are considered as the association measures. We can observe that CONF and CONV are directional measures while LIFT is not. The direction of rules obtained by LIFT is trivial, i.e. $lift(X \rightarrow Y)$ is equal to $lift(Y \rightarrow X)$ but $conf(X \rightarrow Y)$ is not equal to $conf(Y \rightarrow X)$, and also $conv(X \rightarrow Y)$ is different from $conv(Y \rightarrow X)$. Through our work with three types of news relations, we do not account for the direction of the rules because it does not perceive meaningful differences. Therefore, CONF and CONF will be treated to be undirectional by $min()$ function, as presented in the following equations.

$$conf(X,Y) = min(conf(X \rightarrow Y), conf(Y \rightarrow X)) \quad (4)$$

$$conv(X,Y) = min(conv(X \rightarrow Y), conv(Y \rightarrow X)) \quad (5)$$

The reason why we use the $min()$ function is that, in news relation discovery, the smaller value is make sense to the human judgments since the assessors disregard the direction of news relations. For example, in an evident of vastly different occurrence frequencies between two news documents, if the relation $news1 \rightarrow news2$ has very high confidence of 90% and $news2 \rightarrow news1$ has very low confidence of 10%, the judgments of assessors will be made on the UR relation rather than the CR relation because the contents between $news1$ and $news2$ are definitely dissimilar.

5. Experiments

5.1 Experimental Setting

To examine how three factors (term representation bases, term weightings, and association measures) affect the quality of discovered news relations, three experiments are performed using our evaluation dataset (comprised of 1,132 relations) described in Sect. 4.2. The rank-order mismatch (ROM) in Eq. (3) is used for evaluating performance quality. In the first experiment, the effect of each single factor on the relation quality is focused at different k's (k = 50, 100, …, 1100) of the top-k ranks in a ranked list generated from each individual combination. This experiment includes three comparative studies. The first one stands for comparing two term representation bases (UG vs. BG). The second one investigates four term weightings (BF vs. BFIDF vs. TF vs. TFIDF). The third one aims to contrast three association measures (CONF vs. CONV vs. LIFT). Here, any pair of possible alternatives for each factor is compared by calculating the difference of their ROM values, i.e., subtraction of the ROM value of an alternative with that of the other alternative. If the ROM value of the method A, $ROM_h(A)$, is higher than that of the method B, $ROM_h(B)$, the ROM difference between A and B ($ROM_h(A)$-$ROM_h(B)$) becomes positive, that is, the method A has more mismatches than the method B. In other words, the method B performs better than the method A since the method B provides more similar results to human expert answers. In contrast with the pairwise comparison in the first experiment, the second experiment targets the exploration of the detailed performances (ROM values) of all twenty four combinations. As detailed investigation, we select a number of best combinations and then investigate their performances on the top-k ranks. In the third experiment, for each of twenty four methods, we visualize the ratio of CR, SH, and UR relations with respect to top-k intervals, instead of top-k ranks, in order to investigate whether CR relations can be located at higher ranks followed by SH, and UR can be placed at lower ranks, or not. In this experiment, we group the results by each of the best combinations discovered by previous experiment in order to grasp the distribution of relation types (CR, SH, and UR) obtained from each association measure. In addition to these three experiments, we also conduct an analysis by figuring out the number of relations for each relation type according to document size and content overlap.

5.2 Experimental Results

This section presents three experimental results and their discussions.

5.2.1 Paired Comparative Studies

(1) Term Representation Bases: UG vs. BG

Figure 1 (a)–(c) show the ROM differences between UG and

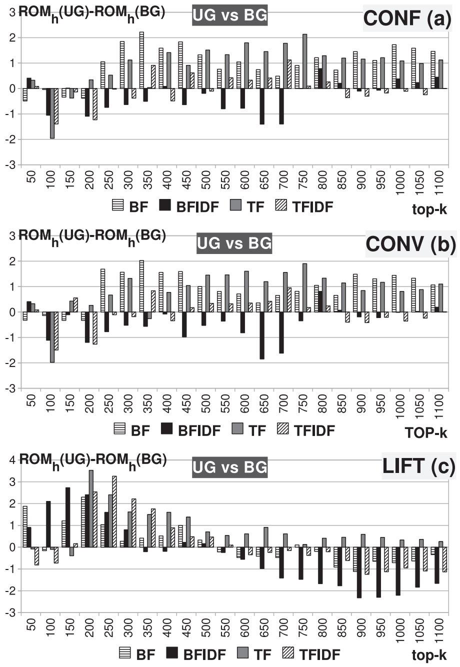


Fig. 1 ROM differences between term representation bases: UG vs. BG in the cases of (a) CONF, (b) CONV, and (c) LIFT.

BG (ROM_h(UG)-ROM_h(BG)) for CONF, CONV, and LIFT respectively. In the figures, the bar graphs are plotted with respect to top-k ranks. As one observation, the bar graphs of most top-k's locate in the positive area, except those of the lower-ranks of LIFT (Fig. 1 (c)). The result implies that BG outperforms UG in most cases except the lower-ranks of LIFT. As a summary, using either BG with CONF or BG with CONV is effective in all ranks, and using BG with LIFT is effective in upper rank (≤ 500). However, it seems that applying UG with LIFT gains good performance in lower ranks (>500).

(2) Term Weightings: BF vs. BFIDF vs. TF vs. TFIDF

Following the same setting of the previous experiment, the performances of BF, TF, BFIDF, and TFIDF are investigated in place of the comparison between UG and BG. From the results in Fig. 2 (a) and (d), we observe that BFIDF gives lower ROM values than BF and TF, in most top-k ranks while TF obtains lower ROM values than BF, as shown in Fig. 2 (b). Figure 2 (c) and (f) shows that TFIDF performs better than BF and TF, in most ranks. Figure 2 (e) indicates that TFIDF outperforms BFIDF in most ranks as it gains lower ROM values. By these results, TFIDF is the most effective in total with its lowest ROM values, compared to BF, TF and BFIDF. In summary, the performance order seems to be TFIDF > BFIDF > TF > BF. The existence of IDF can be recognized as an important component to improve the performance.

(3) Association Measures: CONF vs. CONV vs. LIFT

Like previous experiments, the comparison among association measures are investigated, as shown in Fig. 3 (a)–(c).

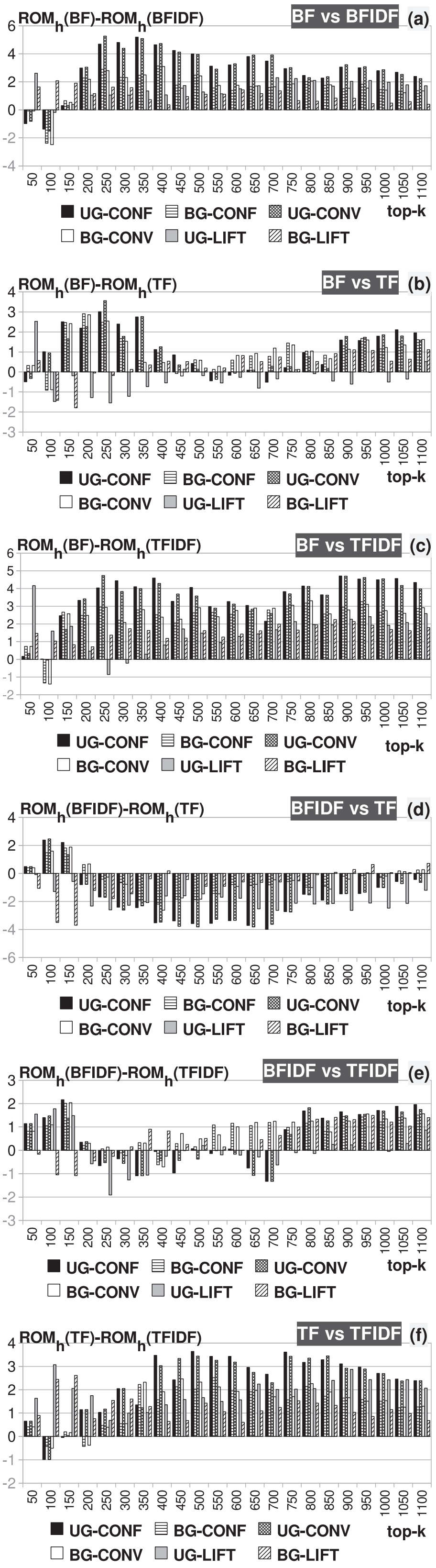


Fig. 2 ROM differences between term weightings: (a) BF vs. BFIDF, (b) BF vs. TF, (c) BF vs. TFIDF, (d) BFIDF vs. TF, (e) BFIDF vs. TFIDF, and (f) TF vs. TFIDF.

Table 4 ROM values (ROM_h) gained from twenty four combinations of three factors.

Combination	top-*k*											
	50	100	200	300	400	500	600	700	800	900	1000	1100
UG-BF-CONF	0.65	4.08	7.97	10.35	11.41	12.79	13.47	14.27	16.87	17.98	18.15	17.55
UG-TF-CONF	1.14	3.07	5.78	7.95	10.29	12.37	13.63	14.78	15.91	16.38	16.36	15.60
UG-BFIDF-CONF	1.63	5.45	4.98	5.54	6.77	8.80	10.26	10.79	14.42	14.92	15.36	15.17
UG-TFIDF-CONF	0.49	4.06	4.64	5.90	6.82	8.73	10.21	12.12	12.74	13.27	13.66	13.21
UG-BF-CONV	0.82	3.96	7.95	9.95	11.25	12.57	13.39	14.59	16.74	17.99	17.84	17.00
UG-TF-CONV	1.14	3.01	5.68	8.17	9.99	12.43	13.46	14.30	15.97	16.22	15.98	15.41
UG-BFIDF-CONV	1.63	5.47	4.90	5.56	6.52	8.62	10.11	10.67	14.44	14.77	14.97	14.77
UG-TFIDF-CONV	0.49	4.00	4.53	6.12	6.96	8.99	10.27	12.00	12.62	13.30	13.29	13.02
UG-BF-LIFT	14.86	10.81	10.69	7.90	8.02	8.97	9.58	11.12	12.06	12.38	12.51	12.22
UG-TF-LIFT	12.33	12.28	11.97	9.11	8.56	9.15	9.85	11.46	12.14	12.99	13.01	11.70
UG-BFIDF-LIFT	12.24	10.99	9.65	6.85	6.95	7.69	8.07	8.83	9.96	10.35	10.53	10.50
UG-TFIDF-LIFT	10.69	9.21	10.23	8.11	7.21	7.49	8.28	9.45	10.09	10.12	10.58	**9.63**
BG-BF-CONF	1.14	4.12	8.35	8.49	9.82	11.47	12.42	13.79	15.66	16.52	16.42	16.09
BG-TF-CONF	0.82	5.03	5.44	6.83	8.88	10.86	11.83	13.00	14.62	15.22	15.27	14.47
BG-BFIDF-CONF	1.22	6.51	6.08	6.17	6.69	8.99	11.04	12.19	13.64	15.02	14.98	14.72
BG-TFIDF-CONF	**0.41**	5.45	5.87	6.28	7.31	8.84	9.88	11.00	12.48	13.58	13.76	13.21
BG-BF-CONV	1.14	4.10	8.28	8.39	9.70	11.57	12.68	13.94	15.69	16.51	16.40	15.93
BG-TF-CONV	0.82	4.99	5.42	6.85	9.22	10.98	11.85	12.75	14.64	15.37	15.17	14.31
BG-BFIDF-CONV	1.22	6.59	6.10	6.08	6.60	9.15	10.93	12.29	13.63	14.97	14.99	14.58
BG-TFIDF-CONV	**0.41**	5.49	5.80	6.30	7.31	8.65	9.92	11.05	12.38	13.71	13.64	13.01
BG-BF-LIFT	12.98	10.97	8.40	7.63	7.50	8.64	10.06	11.60	12.25	13.50	13.22	12.56
BG-TF-LIFT	12.41	12.38	8.45	7.50	6.96	8.45	9.23	10.85	11.73	12.40	12.68	11.45
BG-BFIDF-LIFT	11.35	8.89	7.24	6.05	7.15	7.53	8.63	10.24	11.63	12.68	12.74	12.16
BG-TFIDF-LIFT	11.51	9.94	7.69	5.90	6.32	7.03	8.63	9.61	10.30	11.36	11.53	10.77

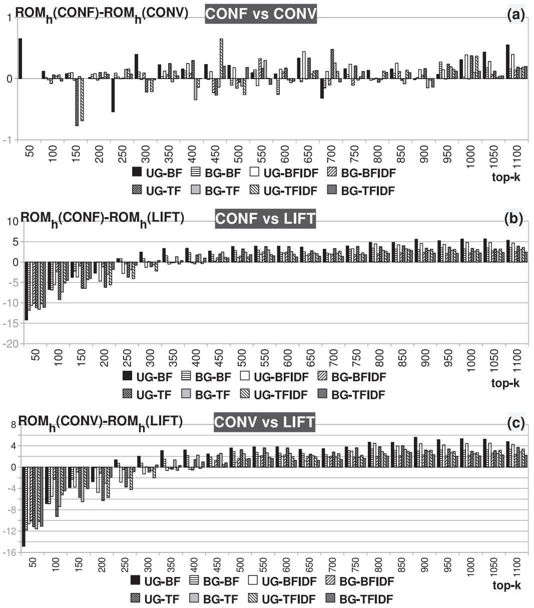


Fig. 3 ROM differences between association measures: (a) CONF vs. CONV, (b) CONF vs. LIFT, and (c) CONV vs. LIFT.

Figure 3 (a) shows trivial ROM differences between CONF and CONV, more specifically less than 1%. It implies that both CONF and CONV obtain discovered relations of comparable quality. In Fig. 3 (b) and (c), the bar graphs can be characterized into two groups, upper ranks (≤ 300) and lower ranks (> 300). With upper ranks, CONF and CONV present lower ROM values than LIFT, while LIFT outputs lower ROM values in lower ranks. These results suggest that CONF as well as CONV is effective to rank the relations in upper ranks, but LIFT is more effective to rank relations in lower ranks. In the next experiment, we plot the numbers of relation types (CR, SH and UR) in each top-*k* interval to examine how effective each of twenty four factor combinations is.

5.2.2 Analysis on All Combinations of Three Factors

As the second experiment, the detailed performance of each combination of three factors is examined in order to find the optimal factor combination in discovering meaningful news relations. Table 4 shows the ROM values of twenty four methods in each top-*k* rank. For the top-50 to top-200 mined relations, the combination of BG, TFIDF, and CONF (BG-TFIDF-CONF), as well as BG, TFIDF, and CONV (BG-TFIDF-CONV), appears to be the most effective since it has the lowest ROM values of 0.41% for top-50 and 5.80%-5.87% for top-200. In the same range, UG-TFIDF-CONF and UG-TFIDF-CONV also perform well with the low ROM values of 0.49% for top-50 and 4.53%-4.64% for top-200. This result indicates that the combinations which include TFIDF are effective for upper ranks. On the other hand, in the upper ranks (top-50 to top-200), the methods with LIFT as their measures obtain more than 10% ROM value, indicate that LIFT is not good at finding good relations at these upper ranks. Moreover, if the methods with LIFT are compared with themselves, they obtain improved performance in the middle ranks (top-200 to top-500) but become worse in the lower ranks. Compared to the methods with CONF or CONV, the methods with LIFT are not good at the upper ranks but perform well in the lower ranks (> top-500). For example, the method with UG, TFIDF, and LIFT (UG-TFIDF-LIFT) performs the best on the top-

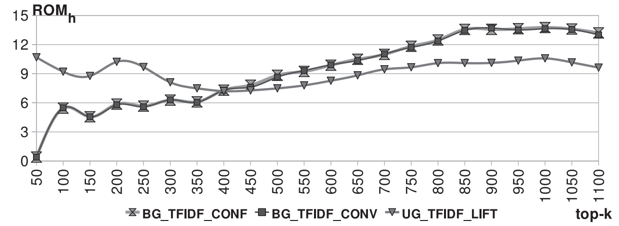


Fig. 4 ROM values (ROM$_h$) in the cases of BG-TFIDF-CONF, BG-TFIDF-CONV and UG-TFIDF-LIFT.

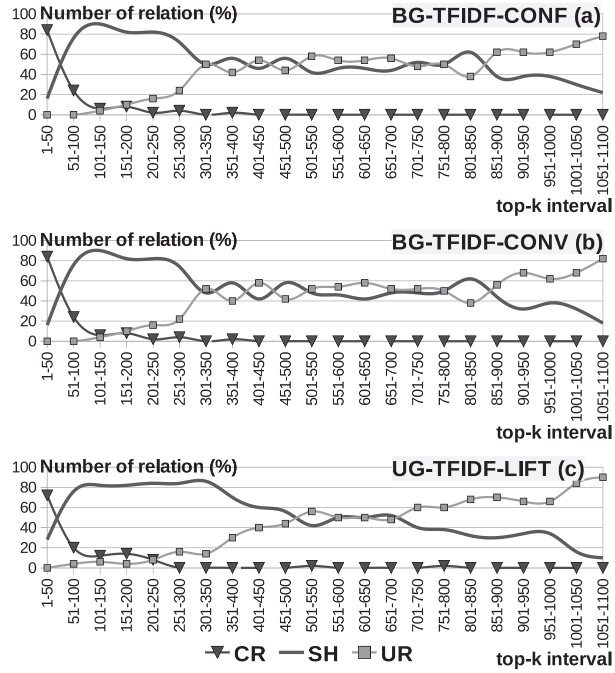


Fig. 5 Percentages of the number of relations in the cases of (a) BG-TFIDF-CONF, (b) BG-TFIDF-CONV, and (c) UG-TFIDF-LIFT.

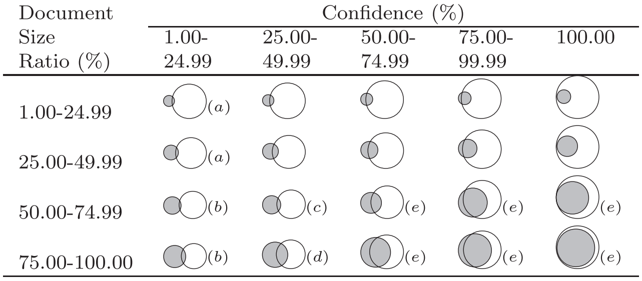
1100 rankings by giving the lowest ROM values (9.63%). These results are consensus with the results in the first experiment, where BG with either CONV or CONF is effective in the upper ranks while UG with LIFT has a good performance in the lower rank, and TFIDF is the most effective term weighting for both cases. To examine the performance trend in the top-k ranks, we select two best combinations for the upper ranks and one best combination for the lower ranks, i.e., (1) BG, TFIDF, and CONF (BG-TFIDF-CONF), (2) BG, TFIDF, and CONV (BG-TFIDF-CONV), and (3) UG, TFIDF, and LIFT (UG-TFIDF-LIFT). The results of these three combinations are shown in Fig. 4. In the next experiment, we have made an experiment to analyse the performance of these three combinations, with respect to the type of relations.

5.2.3 Investigation of Relation Types in Each Rank Interval

This section investigates the performance of the best three combinations by analyzing the numbers of CR, SH and UR relation types in each top-k rank interval. The results for BG-TFIDF-CONF, BG-TFIDF-CONV and UG-TFIDF-LIFT are shown as the graphs in Fig. 5 (a)–(c), respectively. Plotted in each graph are three curves with the symbols of triangle, square and none, representing the ratio of the CR, UR and SH relations, respectively. For every method, CR relations are located at upper ranks (say ≤ 100), SH relations next to CR at middle ranks (say 101-350), and UR relations at lower ranks (say > 350). This result states that, the relations in upper ranks are judged to either CR or SH without UR relations while no CR relations are available in lower ranks. When observing gaps between the CR lines and the SH lines in the leftmost rank interval (1-50 interval), the gaps in the cases of BG-TFIDF-CONF and BG-TFIDF-CONV (Fig. 5 (a) and (b)), are relatively larger, compared to UG-TFIDF-LIFT (Fig. 5 (c)). In the rightmost rank interval (1051-1100 interval), BG-TFIDF-CONF and BG-TFIDF-CONV have narrow gaps while UG-TFIDF-LIFT possesses a wide gap between UR and SH relations. A number of observations can be made as follows. Firstly, for BG-TFIDF-CONF and BG-TFIDF-CONV, SH relations may not be found in the upper ranks and most relations in the upper ranks are of the CR type. Secondly, on the other hand, the BG-TFIDF-CONF and BG-TFIDF-CONV may not be good at the lower ranks since it frequently ranks the SH relations in lower ranks and then mixes with the UR relations. Thirdly, the UG-TFIDF-LIFT obtains the opposite result with the BG-TFIDF-CONF and BG-TFIDF-CONV by ranking relations better in the lower ranks but worse in the upper ranks. As a summary, this result implies that BG-TFIDF-CONF and BG-TFIDF-CONV are good at separating the CR relations from the SH relations whereas UG-TFIDF-LIFT is effective to distinguish the UR relations from the SH relations. The explanation behind the above conclusions can be done as follows. Firstly, since the completely-related news documents (CR relations) usually contain many identical compound words, BG (bigram) seems effective to grasp this characteristic. Secondly, TFIDF is the most effective for weighting terms in discovering news relations. As known in the field of information retrieval and text classification, TF can be used to trigger the frequent terms in a document as representatives for that document and IDF is used in complementary with TF, to promote a rare term, which occurs in very few documents as an important word. Thirdly, while the CONF and CONV can distinguish well between CR relations and SH relations, the LIFT cannot. To elaborate this property, let's consider the following two scenarios of a pair of news documents (X and Y) and we are going to evaluate the relation between X and Y. The first scenario is that the two news documents are (almost) identical (a CR relation), i.e., $|X \cap Y| \approx |X| \approx |Y|$ while the second one is that one news document subsumes the other news document (a SH relation), i.e., $X \subset Y$ or $Y \subset X$. According to Eq. 4, CONF will give different values for these two cases, i.e., 1 vs. $min(|X|, |Y|)/max(|X|, |Y|)$ and the CONV also provides different values for these cases, i.e., infinite vs. approx. $max(|X|, |Y|)/max(|(Y \cap \neg X)|, |(X \cap \neg Y)|$, while the methods with LIFT will provide the same value, i.e., $N/(max(|X|, |Y|)$, where the N is the total number of distinct terms in the corpus. This elaboration implies that LIFT cannot distinguish CR and SH relations and may cause misranking among them. For LIFT, some SH relations may mistakenly be promoted to upper ranks and then mix with CR relations. This situation will not occur when we use CONF and CONV as the association measure. Concludingly, CONF as well as CONV is more effective than LIFT in finding relations of identical news (CR relations).

5.3 Discussion and Error Analysis

This section gives a detailed discussion and an error analysis of the experimental results. For this purpose, we analyze the result of BG-TFIDF-CONF or BG-TFIDF-CONV, which is the best combination in discovering meaningful relations. However, since the results of BG-TFIDF-CONF and BG-TFIDF-CONV are very similar, for sake, we select BG-TFIDF-CONF as the model in our error analysis. Towards this analysis, we figure out how many CR, SH or UR relations are located in each range of association measures, with respect to the document-size ratio. Figure 6 visualizes the patterns of relations between two news documents, categorized by document-size ratio ($(min(|X|, |Y|)/max(|X|, |Y|))$ and BG-TFIDF-based confidence ($|X \cap Y|/min(|X|, |Y|)$). In Table 5, the number of relations in each relation type is counted with the consideration of document-size ratio and BG-TFIDF-based confidence, categorized by the patterns in Fig. 6. The investigation on document-size ratio and BG-TFIDF-based confidence is done on 25-percent intervals, i.e., < 25.00%, 25.00% – 49.99%, 50.00% – 74.99%, and ≥ 75.00%. More precisely, the document size is calculated by excluding stopwords when counting. The document-size ratio shows the relative size of two news documents, calculated by $min(|X|, |Y|)/max(|X|, |Y|)$. This ratio is close to 100% when the two news documents (X and Y) have nearly the same size ($|X| \approx |Y|$). The BG-TFIDF-based confidence implies how many overlapping terms exist between two news documents, computed by $|X \cap Y|/min(|X|, |Y|)$. A higher the BG-TFIDF-based confidence implicitly expresses the situation that more overlapping terms are shared between two news documents. As the extreme case, the con-



Document Size Ratio (%)	Confidence (%) 1.00-24.99	25.00-49.99	50.00-74.99	75.00-99.99	100.00
1.00-24.99	(a)				
25.00-49.99	(a)				
50.00-74.99	(b)	(c)	(e)	(e)	(e)
75.00-100.00	(b)	(d)	(e)	(e)	(e)

Fig. 6 Patterns of news relations categorized by document-size ratio ($min(|X|, |Y|)/max(|X|, |Y|)$) and BG-TFIDF-based confidence ($|X \cap Y|/min(|X|, |Y|)$).

Table 5 Numbers of news relations ($X \rightarrow Y$) on the consideration of document-size ratio ($(min(|X|, |Y|)/max(|X|, |Y|)$) and BG-TFIDF-based confidence ($|X \cap Y|/min(|X|, |Y|)$) weighted by BG-TFIDF, with regard to relation types (CR, SH, and UR).

Relation Type (Total #Relations)	Document-Size Ratio (%) (Total #Relations)	Confidence (%) 1.00-24.99	 25.00-49.99	 50.00-74.99	 75.00-99.99	 100.00
CR (65)		20	17	15	10	3
	1.00-24.99 (2)	2	-	-	-	-
	25.00-49.99 (12)	11	1	-	-	-
	50.00-74.99 (11)	3	6	2	-	-
	75.00-100.00 (40)	4	10	13	10	3
SH (571)		559	12	-	-	-
	ST (297)	294	3	-	-	-
1.00-24.99 (74)	1.00-24.99 (45)	45	-	-	-	-
25.00-49.99 (191)	25.00-49.99 (78)	78	-	-	-	-
50.00-74.99 (177)	50.00-74.99 (94)	92	2	-	-	-
75.00-100.00 (129)	75.00-100.00 (80)	79	1	-	-	-
	SE (199)	195	4	-	-	-
	1.00-24.99 (19)	19	-	-	-	-
	25.00-49.99 (85)	85	-	-	-	-
	50.00-74.99 (60)	58	2	-	-	-
	75.00-100.00 (35)	33	2	-	-	-
	SN (75)	70	5	-	-	-
	1.00-24.99 (10)	10	-	-	-	-
	25.00-49.99 (28)	27	1	-	-	-
	50.00-74.99 (23)	21	2	-	-	-
	75.00-100.00 (14)	12	2	-	-	-
UR (496)		496	-	-	-	-
	1.00-24.99 (123)	123	-	-	-	-
	25.00-49.99 (130)	130	-	-	-	-
	50.00-74.99 (134)	134	-	-	-	-
	75.00-100.00 (109)	109	-	-	-	-

fidence of 100% indicates that two news documents (X and Y) are completely identical ($|X \cap Y| \approx |X| \approx |Y|$).

As observed in Table 5, most CR relations (40 out of 65 CR relations) have a high document-size ratio (≥ 75.00%). It implies that two highly related news documents (CR relation) usually have the same figure of their sizes. However, some pairs of completely related news, i.e., 25 out of 65 CR relations, have quite different sizes (< 75.00%). With manual exploration, we found out that most of these cases are triggered when one news document is a summary of the other. The SH relations (368 out of 571 SH relations) are mostly found in the average document-size ratio (25.00%-74.99%). That is, their size difference is not too small and not too large. The possible reason is that the SH relation does not refer to exactly the same story but somewhat related to the same topic, therefore they always have both same and different contents. The document-size ratio can be varied for the UR relation since two unrelated news documents have no any relationship. In order to analyze the errors obtained in the proposed method, we investigate the relations which some instances are misclassified. From the table, we can observe that relations with high BG-TFIDF-based confidence (≥ 50.00%) can be recognized as the CR relation. They are also found in high document-size ratio (50%-100%), i.e., (e) in Fig. 6. Relations with the BG-TFIDF-based confidence of 25.00%-49.99% can be recognized as CR or SH relation. In this range, most CR relations (10 out of 17 CR relations) have high document-size ratio (75.00%-100.00%), i.e., (d) in Fig. 6. Moreover, there are a small number of SH relations (12 out of 571 SH relations), which are placed in the range of 50%-100% document-size ratio, i.e., (c) and (d) in Fig. 6. It can be implied that there are mixtures of CR relations and SH relations. As observing the BG-TFIDF-based confidence which is less than 25.00%, all relation types (CR, SH, and UR) are located in this range, i.e., (a) and (b) in Fig. 6. By manually investigating, misclassified types occur between CR relations and SH relations, and SH relations and UR relations while the misclassification between CR relations and UR relations is not found. To identify possible reasons why CR relations are existed in low BG-TFIDF-based confidence (< 50%) and mixed with the SH relations, let's consider the following two cases of the document-size ratio, i.e., (1) < 50% and (2) ≥ 50%. For the first case, the CR relations, which have low document-size ratio (quite different size, < 50%) and low BG-TFIDF-based confidence (low related, < 50%), i.e., (a) in Fig. 6 may have different levels of event details. That is, two completely related news documents which mention the same event may have unequal facts, triggering different size and few overlapping terms among those documents. For the second case, the CR relations have low BG-TFIDF-based confidence (low related, < 50%) but a high document-size ratio (close size, ≥ 50%), i.e., (b), (c), and (d) in Fig. 6. Such CR relations may exist with SH relations in two manners as following. Firstly, any two completely-related news documents, which may be written by using synonyms or in different styles, initiate few overlapping terms between these news documents, and usually they occupy similar sizes. Secondly, while two news documents mention the same event, they may have contrast details with almost the same size. On the other hand, SH relations are located in the same BG-TFIDF-based confidence range as CR relations because of the following error. Any two somehow-related news documents may share several terms, such as person name and place names, but refer to different events. Therefore, they may be classified as related news relations. Analyzing misclassification between SH relations and UR relations, we manually observe that some SH relations are mixed with UR relations and vice versa. The reasons why some SH relations may be placed in the same BG-TFIDF-based confidence range as UR relations are as follow. Any two somehow related news documents may be identified as a UR relation when they do not share enough overlapping terms. Conversely, some UR relations may be recognized as SH relations when any two unrelated news documents are in the same category, such as politics news, but share many common terms used in that category, such as "government", "law", and "election." These terms are likely to trigger high overlapping among the two documents, and they are not identified as stopwords. To overcome all the problems above, a number of solutions include improving weighting schemes, especially adding more weight to news headline, considering news metadata, such as publishing time and publisher information, and handling synonym.

6. Conclusions

This paper investigates the effect of combining factors for discovering relations among news documents using association rule mining. We focus on three factors, i.e., two term representation bases, four term weightings, and three association measures. Totally twenty four combinations are explored. To evaluate the quality of discovered news relations, by comparing the results to the human judgments, top-k ranked relations are analyzed by rank-order mismatch (ROM). The experimental results show that the ROM value under the combination of BG-TFIDF-CONF as well as BG-TFIDF-CONV is suitable to achieve finding semantic relations with 0.41% ROM on top-50 mined relations while UG-TFIDF-LIFT performs well, up to top-1100, with ROM of 9.63%. Our results suggest that BG-TFIDF-CONF as well as BG-TFIDF-CONV is effective for separating the CR relations (relevant relations) from the SH relations whereas UG-TFIDF-LIFT is applicable in distinguishing the UR relations (irrelevant relations) from the SH relations.

As future works, we plan to examine the changing point from the relevant area up to the irrelevant area in order to apply a suitable combination for each area. Towards this, the improvement of ranking mechanism for each area may be needed. Moreover, a hybrid method which selectively uses different criteria for different areas may be beneficial in news document relation discovery. Furthermore, the direction of the relations is necessary for efficient discovery of news relations. Structured data of news (news metadata), such as publishing time, publisher information and news category is also one resource towards the enhancement of news discovery process. The detailed analysis, especially, an analysis of semantic relation among news documents, such as synonym handling and discourse analysis can possess the potential to provide higher performance of news relation discovery.

Acknowledgements

This work was supported by the grant of Strategic Scholarships Fellowships Frontier Research Networks 2006 from the Commission on Higher Education, NECTEC Project Grant No.NT-B-22-KE-54-01, and National Research University Project via Thammasat University.

References

[1] G. Ferizis and P. Bailey, "Towards practical genre classification of web documents," Proc. 15th International Conference on World Wide Web, pp.1013–1014, New York, NY, USA, 2006.

[2] M. Gamon, "Linguistic correlates of style: Authorship classification with deep linguistic analysis features," Proc. Coling 2004, pp.611–617, COLING, Switzerland, Aug. 2004.

[3] R. Carreira, J.M. Crato, D. Gon, calves, and J.A. Jorge, "Evaluating adaptive user profiles for news classification," Proc. 9th International Conference on Intelligent User Interfaces, pp.206–212, New York, NY, USA, 2004.

[4] I. Antonellis, C. Bouras, and V. Poulopoulos, "Personalized news categorization through scalable text classification," Frontiers of WWW Research and DevelopmentAPWeb 2006, Lect. Notes Comput. Sci., vol.3841, pp.391–401, 2006.

[5] S. Mengle, N. Goharian, and A. Platt, "Discovering relationships among categories using misclassification information," Proc. 2008 ACM Symposium on Applied Computing, pp.932–937, New York, NY, USA, 2008.

[6] N. Zhang, T. Watanabe, D. Matsuzaki, and H. Koga, "A novel document analysis method using compressibility vector," Proc. First International Symposium on Data, Privacy, and E-Commerce, pp.38–40, Nov. 2007.

[7] T. Weixin and Z. Fuxi, "Text document clustering based on the modifying relations," Proc. 2008 International Conf. on Computer Science and Software Engineering, pp.256–259, Dec. 2008.

[8] F. Lin and C. Liang, "Storyline-based summarization for news topic retrospection," Decision Support Systems, vol.45, no.3, pp.473–490, 2008.

[9] J. Kuo and H. Chen, "Multidocument summary generation: Using informative and event words," ACM TALIP, vol.7, no.1, pp.1–23, 2008.

[10] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.194–218, 1998.

[11] R. Papka and J. Allan, "Topic detection and tracking: Event clustering as a basis for first story detection, ir 4," in Book of Advances Information Retrieval: Recent Research from the CIIR, pp.96–126, Kluwer Academic Publishers, 2006.

[12] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," Proc. 1993 ACM SIGMOD International Conf. on Management of Data, pp.207–216, ACM, New York, NY, USA, 1993.

[13] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," International Transactions on Computer Science and Engineering, vol.32, no.1, pp.71–82, 2006.

[14] K. Sriphaew and T. Theeramunkong, "Revealing topicbased relationship among documents using association rule mining," in Artificial Intelligence and Applications, ed. M.H. Hamza, pp.112–117, IASTED/ACTA Press, 2005.

[15] K. Sriphaew and T. Theeramunkong, "Quality evaluation for document relation discovery using citation information," IEICE Trans. Inf. & Syst., vol.E90-D, no.8, pp.1225–1234, Aug. 2007.

[16] U. Inyaem, P. Meesad, and C. Haruechaiyasak, "Namedentity techniques for terrorism event extraction and classification," Proc. Eighth International Symposium on Natural Language Processing, 2009, pp.175–179, Oct. 2009.

[17] T. Chumwatana, K.W. Wong, and H. Xie, "Non-segmented document clustering using self-organizing map and frequent max substring technique," Proc. ICONIP (2), pp.691–698, 2009.

[18] O. Sornil and K. Gree-ut, "An automatic text summarization approach using content-based and graph-based characteristics," Proc. 2006 IEEE Conf. on Cybernetics and Intelligent Systems, pp.1–6, June 2006.

[19] C. Jaruskulchai and C. Kruengkrai, "A practical text summarizer by paragraph extraction for thai," Proc. Sixth International Workshop on Information Retrieval with Asian Languages, pp.9–16, Morristown, NJ, USA, 2003.

[20] N. Kittiphattanabawon and T. Theeramunkong, "Relation discovery from thai news articles using association rule mining," Pacific AsiaWorkshop on Intelligence and Security Informatics, Lect. Notes Comput. Sci., vol.5477, pp.118–129, 2009.

[21] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," Proc. 20th International Conf. on Very Large Data Bases, pp.487–499, San Francisco, CA, USA, 1994.

[22] M.J. Zaki and C.J. Hsiao, "Charm: An efficient algorithm for closed association rule mining," Tech. Rep., Computer Science, Rensselaer Polytechnic Institute, 1999.

[23] M.J. Zaki and C.J. Hsiao, "Efficient algorithms for mining closed itemsets and their lattice structure," IEEE Trans. Knowl. Data Eng., vol.17, no.4, pp.462–478, 2005.

[24] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Min. Knowl. Discov., vol.8, no.1, pp.53–87, 2004.

[25] A. Merceron and K. Yacef, "Interestingness measures for association rules in educational data," Proc. Educational Data Mining Conference, pp.57–66, 2008.

[26] S. Lallich, O. Teytaud, and E. Prudhomme, "Association rule interestingness: Measure and statistical validation," Quality Measures in Data Mining, Studies in Computational Intelligence, vol.43, pp.251–275, Springer Berlin/Heidelberg, 2007.

[27] P.J. Azevedo and A.M. Jorge, "Comparing rule measures for predictive association rules," Machine Learning: ECML 2007, Lect. Notes Comput. Sci., vol.4701, pp.510–517, 2007.

[28] A. Jorge and P. Azevedo, "An experiment with association rules and classification: Post-bagging and conviction.," Discovery Science, Lecture Notes in Computer Science, pp.137–149, 2005.

[29] H. David, The Method of Paired Comparisons, Oxford University Press, New York, 1988.

Nichnan Kittiphattanabawon received a bachelor degree in Computer Science from Rangsit University, Thailand in 1993, and master degree in Computer Science from Prince of Songkla University, Thailand in 1999. She is currently a student in Ph.D. program, School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasart University, Thailand. Her research interests are data mining and knowledge discovery.

Thanaruk Theeramunkong received a bachelor degree in Electric and Electronics Engineering, and master and doctoral degrees in Computer Science from Tokyo Institute of Technology in 1990, 1992 and 1995, respectively. Now he is reserved as an associate professor at Sirindhorn International Institute of Technology, Thammasart University, Thailand. He is currently a member of ACM and ECTI association. His current research interests include data mining, machine learning, natural language processing, and knowledge engineering.

Ekawit Nantajeewarawat received the B.Eng. degree in Computer Engineering from Chulalongkorn University, Thailand, in 1987; and the M.Eng. and D.Eng. degrees in Computer Science from the Asian Institute of Technology, Thailand, in 1991 and 1997, respectively. He is currently an Associate Professor of Computer Science at Sirindhorn International Institute of Technology, Thammasat University, Thailand. His research interests include knowledge representation, automated reasoning, rule-based equivalent transformation, program synthesis, formal ontologies, and object-oriented modeling. He is a member of the Association for Computing Machinery.