# Exploration of Document Relation Quality with Consideration of Term Representation Basis, Term Weighting and Association Measure

Nichnan Kittiphattanabawon, Thanaruk Theeramunkong, and Ekawit Nantajeewarawat

Sirindhorn International Institute of Technology, Thammasat University, Thailand
knichcha@wu.ac.th, {thanaruk,ekawit}@siit.tu.ac.th

**Abstract.** Tracking and relating news articles from several sources can play against misinformation from deceptive news stories since single source can not judge whether the information is a truth or not. Preventing misinformation in a computer system is an interesting research in intelligence and security informatics. For this task, association rule mining has been recently applied due to its performance and scalability. This paper presents an exploration on how term representation basis, term weighting and association measure affect the quality of relations discovered among news articles from several sources. Twenty four combinations initiated by two term representation bases, four term weightings, and three association measures are explored with their results compared to human judgement. A number of evaluations are conducted to compare each combination's performance to the others' with regard to top-k ranks. 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% rank-order mismatch 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 rank-order mismatch of 9.63 %.

**Keywords:** Document Relation, Association Rule Mining, News Relation.

## 1 Introduction

The explosion of the Internet makes it easier for broadcasting news to a large volume of readers. Most of the readers prefer to read news stories from several publishers in order to avoid bias from a single source of information. News portals are the most popular alternatives for the readers because several news portals facilitate news content access by providing linkages among news articles on the web, releasing readers from directly browsing through news publishers' home pages. Reading news stories which come from several sources from the

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 126–139, 2010.
© Springer-Verlag Berlin Heidelberg 2010

services of news portals can prevent the readers from false information since misinformation can make the readers in changing their perception as an attack directed at the mind of the readers of a computer system. Such an attack was defined as cognitive hacking in intelligence and security informatics research [1]. For the success of systems for news provider services, an appropriate organization of news contents is a major requirement. News portals usually organize news into some kinds of relationship structures, e.g., group news articles by category, by recency, or by popularity, summarize news contents, and create relations between news articles. Currently most of these functions require manual arrangement processes. 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 [2,3,4]. Some of them can be adopted for classifying news articles [5,6]. By the classification method, it requires users to provide a number of predefined classes and a large number of training examples. However, the classification approach 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 [7,8]. As a more complicated application, a multidocument summarization can be performed to obtain a shorter description from a cluster of news describing similar events [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] has been applied for discovering document relations in scientific research publications due to its performance and scalability [13]. In Thai language, even there have been several works towards extraction of information on online document, most of them still have limitation in finding document relations. As an early work on relation discovery in multiple Thai documents, Kittiphattanabawon and Theeramunkong [14] 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 support-confidence and support-conviction frameworks under limited environment of top-k ranking evaluation.

In this paper, besides support-confidence and support-conviction, a support-lift framework is investigated and compared with human judgment in a more general environment of up-to top-1100 ranking evaluation. Towards optimal settings, twenty four combinations generated from two term representation bases, four term weightings and three association measures are examined to find suitable combinations for discovering meaningful relations among news articles. In Sect.2, news relation generation is described under the formation of association

rules. The factors for discovering news associations are then presented 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 Relations among News Articles

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 has been applied in finding relations among documents [13,14]. 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 [13], and the generalized conviction in [14] as association measures. A formulation of the ARM task on news article relation discovery can be summarized as follows. Assume that $I = \{i_1, i_2, ..., i_m\}$ is a set of $m$ news articles (items), $T=\{t_1, t_2, ..., t_n\}$ is a set of $n$ terms (transactions), a news itemset $X=\{x_1, x_2, ..., x_k\}$ is a set of $k$ news articles, and a news itemset $Y=\{y_1, y_2, ..., y_l\}$ is a set of $l$ news articles. 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 $conf(X \rightarrow Y)/supp(Y)$, where $conf(X \rightarrow Y)$ is the confidence value of the rule $X \rightarrow Y$ and $supp(Y)$ is the support value of $Y$. The generalized support of X ($sup(X)$), 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 $w(i_a, t_b)$ represents a weight of a term $t_b$ in a news articles $i_a$ and $Z = \{z_1, z_2, ..., z_{k+l}\} \subset I$ with $k + l$ news articles since they are the co-occurrence of terms in both $k$ news articles in the $X$ and $l$ news articles 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 articles. This rule represents that the content overlap among the news articles in the $X$ has a relationship with the content overlap among the news articles in the $Y$. As a special case of one single antecedent and one single consequent, the rule can be interpreted that the news article in the $X$ relates to the news article in the $Y$. Among efficient algorithms such as Apriori [15], CHARM [16,17] and FP-Tree [18], 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.

**Table 1.** Definitions of association measures: (a) generalized support, (b) generalized confidence, (c) generalized conviction, and (d) generalized lift

| | |
|---|---|
| (a) $sup(X) = \frac{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}$ | (b) $conf(X \rightarrow Y) = \frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}$ |
| (c) $conv(X \rightarrow Y) = \frac{1-\frac{\sum_{b=1}^{n} min_{a=1}^{l} w(y_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}}{1-\frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}}$ | (d) $lift(X \rightarrow Y) = \frac{\frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}}{\frac{\sum_{b=1}^{n} min_{a=1}^{l} w(y_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}}$ |

## 3 Term Representation Basis, Term Weighting and Association Measure

In general, the results from the mining process can differ according to setting factors in the process. In this paper, to find an appropriate environment in discovering the news relations, we explore three main factors, (1) term representation basis, (2) term weighting, and (3) association measure. For the term representation basis, unigram (UG) or 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 to handle compound words and then 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 simply indicates the existence or non-existence of a term in a news document while TF indicates the frequency of a term in the document. IDF is often used in complementary with TF, to promote a rare term which occurs in very few documents, as an important word. Although it can be calculated as the total number of documents in the collection ($N$) divided by the number of documents containing the term ($DF$), it is usually used in the logarithm scale. BFIDF and TFIDF of the $i$-th terms are defined as $BF_i \times \log(N/DF_i)$ and $TF_i \times \log(N/DF_i)$ 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. For CONF, it is a well-known rule measure for ARM approach. As for CONV and LIFT, they can result in more interesting relations [19,20], CONV and LIFT are investigated to improve the association of news articles in our work, as shown in the previous section.

## 4 Evaluation Methodology

In this section, we describe an evaluation methodology to investigate the potential combinatorial factors to make a judgement on types of news relations.

### 4.1 Types of News Relations

Most tasks about document relations judged stories to be either relevant or non-relevant, that is, two classes ("yes" and "no") were considered. 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) [14]. The CR relation is detected when two new articles are about an exactly same event. Such a CR relation is always found because every news reporter tries to report daily important events. The same event, therefore, is often published by many publishers in the same time. However, the CR relation may be presented in either different headlines or different writing styles. As a result, this type is often retrieved among different news publishers whose publishing times are the same or quite close to the same. For the SH relation, it is a kind of relation which has only somewhat closely related. The events in both news articles may have similar topics. connect together forming a sequential time series of events. or contain same contents in some parts. However, the contents in news documents whose relationship is SH type are not mentioned exactly the same story. The relation of UR is defined as a relationship of having absolutely nothing related between news articles. 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. The dataset is selected approximately 1,100 news relations mined from 811 Thai news articles of three news online sources (Dailynews (313 articles), Komchadluek (207 articles), and Manager online (291 articles)) [1] during August 14-31, 2007, consisting of three categories (politics, economics, and crime). Each set is comprised of the news relations with their relation types defined in previous section. For the evaluation of human, three assessors who admire reading news provide their judgments on predefined relation types (CR, SH, and UR). After they have been instructed for making a decision about how related two news articles should be identified, they make their decision by comparing the contents in both news articles and then assign only one type for their relation. Note that, every news relation is judged by all three assessors. If there are different opinions on the association, the best judgment is given by voting. However, voting may not be able to guarantee a majority for such an agreement. To decide the answer, an iteration process is performed by asking the assessors to repeat their considerations until the final decision is made. To this end, our dataset contains records of news relations along with their relation types determined by human judgment (65 relations of CR, 571 relations of SH, and 496 relations of UR). Details for evaluation of these types will be described in next section.

[1] `www.dailynews.co.th, www.komchadluek.com, and www.manager.co.th`

### 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 [21] since the paired-wise comparison has been applicably used for counting the mismatches between rankings. For each combination, an evaluation is proceeded by creating a ranking list of relations ordered by its association measure, mapping the resultants from the human judgments to each of these relations in the list, calculating a mismatch score between both of them, and comparing the quality among twenty four combinations with a criterion, so-called rank-order mismatch (ROM) shown in Eq. (1). The ROM value is a calculation of dividing a mismatch score ($M(A, B)$) with the mismatch score of the worst case which all news relations in one method ($A$) are arranged in the reverse order compared to the other method ($B$). In addition, the ROM value expresses, according to the assessment of assessors, the mismatches of relation types of our method to the human suggestions. Note that the ROM value is a value in {0, 100}. If all news relations are found corresponding to the human suggestions, ROM value is equal to 0. $M(A, B)$, the mismatch score, indicates the number of rank mismatches between two methods, say $A$ and $B$, which rank a set of $N$ objects, as shown in Eq. (2), where $r_A(k)$ and $r_B(k)$ are the respective rank of the $k$-th objects based on method $A$ and $B$ respectively. The mismatch score expresses the importance of method $A$ whether corresponding to method $B$ or not. As for our work, since $A$ is the machine ranking method and $B$ is the human ranking method, the ROM($A$, $B$) is denoted by $\text{ROM}_h(A)$. This equation implies, therefore, relation type mismatches between our method and the evaluation method from human.

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

$$M(A, B) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\delta(r_A(i), r_A(j)) - \delta(r_B(i), r_B(j))| \tag{2}$$

A mismatch function, $\delta(a, b)$, returns 1 when $a$ less than $b$, otherwise 0. 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. As stated above, the constructed rank order is arranged by the association measure. It is important to recognize that CONF and CONV are directional while LIFT is not. The direction of rules obtained by LIFT is not taken into account, 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 on the types. CONF and

CONF will be treated to be undirectional by $min()$ function, as presented in following equations.

$$conf(X,Y) = min(conf(X \rightarrow Y), conf(Y \rightarrow X)) \quad (3)$$

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

The reason why we use the $min()$ function is that, the smaller value is make sense to the human judgments because the assessors disregard the direction of news relations. For example, in an evident of vastly different occurrence frequencies between two news articles, 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 of both news articles are definitely dissimilar.

## 5 Experiments

### 5.1 Experimental Setting

To examine how three factors affect the quality of discovered news relations, three experiments are performed using our evaluation dataset. In the first experiment, the effect of each single factor on the relation quality is focused. This experiment includes three comparative studies (UG vs. BG, BF vs. BFIDF vs. TF vs. TFIDF, and 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 (Eq. (1)) of an alternative with that of the other alternative. If the ROM value of the method $A$, $\text{ROM}_h(A)$, is higher than that of the method $B$, $\text{ROM}_h(B)$, the ROM difference between $A$ and $B$ becomes positive, that is, the method $A$ has more mismatches than the method $B$. In other words, the method $B$ provides more similar results to human answers. In the second experiment, for each of twenty four methods, we visualize the ratio of CR, SH, and UR relations for each association measure with respect to top-$k$ intervals 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. The third experiment targets the exploration of the detailed performance (ROM values) of the combinations which perform very well in top-$k$ ranks.

### 5.2 Experimental Results

**Paired Comparative Studies**

*Term Representation Bases.* Figure 1 (a)–(c) show the ROM differences between UG and BG ($\text{ROM}_h(UG)$-$\text{ROM}_h(BG)$) for CONF, CONV, and LIFT respectively. In the figures, the bar graphs are plotted with respect to top-$k$ ranks. As one observation, almost top-$k$ cases give the number of more positive values than negative values, except cases of LIFT (Fig. 1 (c)), after the top-500 mined relations. The results show that BG outperforms UG in almost cases except the

cases of LIFT under top-500 ranking. These results suggest that 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). For lower rank (>500), applying UG with LIFT appears to be effective.

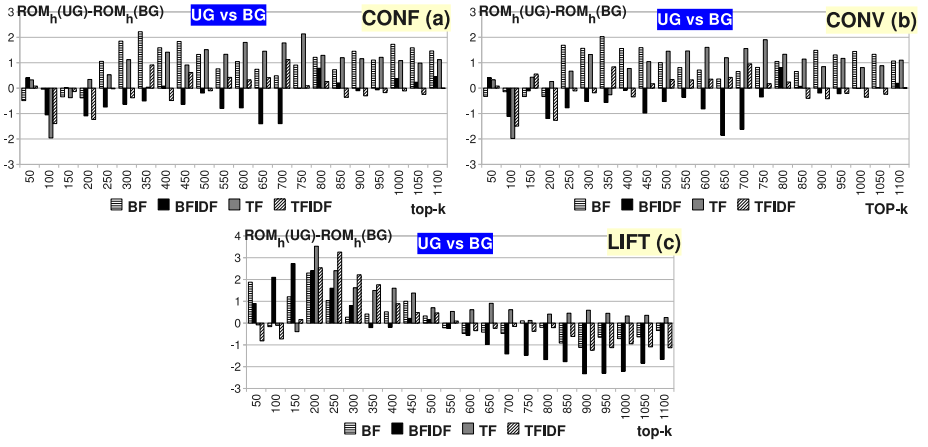


**Fig. 1.** ROM differences between term representation bases: UG vs. BG in the cases of CONF (a), CONV (b), and LIFT (c)

*Term Weightings.* Similar to previous experiment, except that BF, TF, BFIDF, and TFIDF are taken into account instead of UG and BG. We can see that, from Fig. 2 (a) and (d), in almost cases, BFIDF gives lower ROM values than BF and TF respectively, which BF presents higher ROM values than TF, as shown in Fig. 2 (b). However, in Fig. 2 (e), BFIDF outputs higher ROM values than TFIDF in almost cases. Therefore, TFIDF also outperforms BF and TF, as seen in Fig. 2 (c) and (f), in almost cases. These results suggest that TFIDF appears to be the most effective, since TFIDF gives the lowest ROM values.

*Association Measures.* Like previous experiments, the comparisons between association measures are depicted by ROM differences, as shown in Fig. 3 (a)–(c). Figure 3 (a) shows that most of bars are on the positive side of the graph. They indicate that CONV produces lower ROM values than CONF. In both Fig. 3 (b) and (c), the bar graphs appear to be characterized into two groups, upper ranks (< 300) and lower ranks (> 300). With upper ranks, CONF and CONV present higher ROM values than LIFT, while LIFT outputs lower ROM values in lower ranks. However, CONV outperforms CONF as reported above. These suggest that using CONV is effective for the relations placed in upper ranks, but using LIFT is more effective when applied to the relations in lower ranks. For more details, in next experiment, we will investigate types of relations in each rank to determine how effective different combinations produce different relation types.

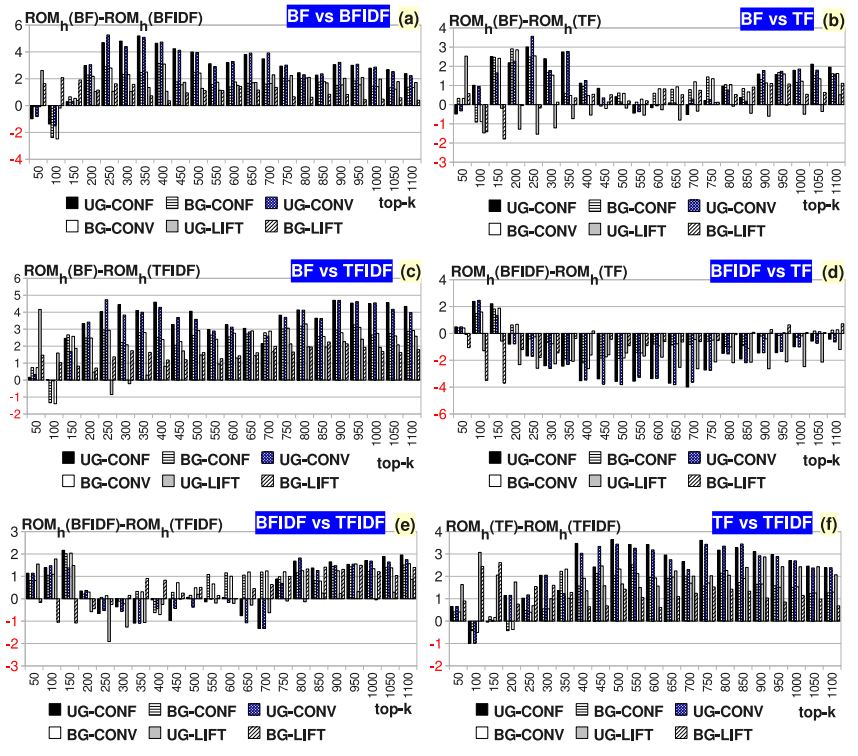


**Fig. 2.** ROM differences between term weightings: (a) BF vs. BFIDF, (b) BF vs. TF, (c) BF vs. TFIDF, (d) BFIDF vs. TF, (e) BFIDF vs. TFIDF, and (f) TF vs. TFIDF

**Rank Analysis on News Relation Types.** The analysis on the number of three relation types under different twenty four combinations and different top-$k$ rank intervals is investigated separately by three association measures (CONF, CONV, and LIFT), as shown in Fig. 4 (a)–(c). Trends of these three graphs can be discussed into three different groups. Such three groups correspond to three types of relations (CR, SH, and UR). In the graphs, the CR and UR relations are represented by the curves with triangle and square symbols respectively. For SH relations, they are plotted by solid lines. Detailed descriptions of them are given here: (1) CR relations are located at upper ranks (say 50-100), (2) SH relations are located next to CR at middle ranks (say 100-350), and (3) UR relations are positioned at lower ranks (say $>$ 350). Obtained results can be stated that, in upper ranks ,the relations 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 position (1-50 interval), the gaps, in the cases of CONF and CONV (Fig. 4 (a) and (b)), are quite large. On the contrary, the cases of LIFT in Fig. 4 (c) trigger smaller gaps between these two types. For the rightmost location (1051-1100 interval), CONF and CONV produce narrow gaps while LIFT causes big gaps between

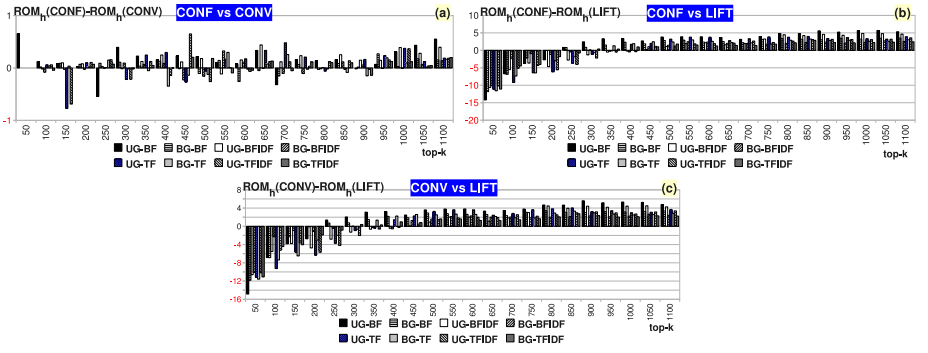


**Fig. 3.** ROM differences between association measures: (a) CONF vs. CONV, (b) CONF vs. LIFT, and (c) CONV vs. LIFT

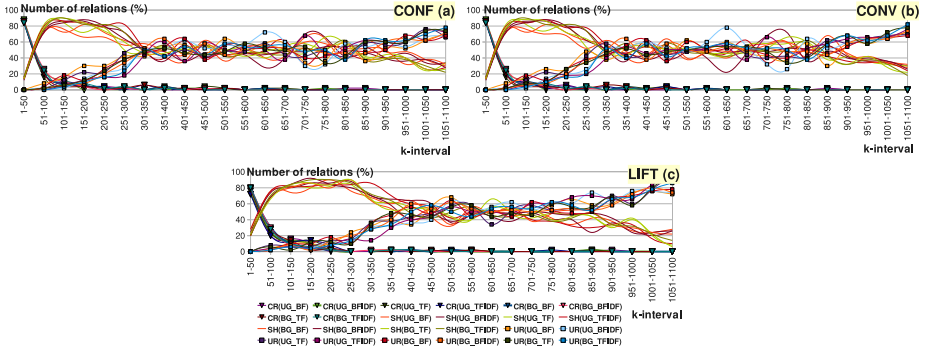


**Fig. 4.** Percentages of the number of relations in the cases of CONF (a), CONV (b), and LIFT (c)

UR and SH relations. We can point out that CONF and CONV are useful for predicting the CR relations, whereas LIFT is fine for separating the SH types from the UR types. The reason is that, for CONF and CONV, the number of SH relations in the leftmost side are significantly different from those of CR relations and are slightly different from those of UR relations in the rightmost side, and vice versa for LIFT.

**Analysis on Combinations of Three Factors.** In the final experiment, the detailed performance on the combinations of three factors is examined. Table 2 shows ROM values of twenty four methods in each top-$k$ rank in order to observe the best combination for using in discovering news relations. For the top-50 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 (0.41%). The method with UG, TFIDF, and LIFT (UG-TFIDF-LIFT) performs the best on the top-1100 rankings by giving the lowest ROM values (9.63%). Such ROM values involve

**Table 2.** ROM values ($ROM_h$) of twenty four combinations of three factors

| Combinations | 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 |

the relation types placing in each rank as investigated in previous experiment (Fig. 4). From Fig. 4, mostly, the types of news relations on top-50 ranks are CRs, while those on top-1100 rankings are URs. The results obtained heretofore can be separated into two cases, (1) the upper ranks, and (2) the lower ranks. In the case of the upper ranks, the best combination is the combination of either BG-TFIDF-CONF or BG-TFIDF-CONV. Another case is the lower ranks that the UG-TFIDF-LIFT combination performs better. The results are in consensus with the three studies of the first experiment. For almost cases, in Fig. 1, BG combined with either CONV or CONF gives better effectiveness than UG, while UG combined with LIFT is more effective when working in the lower rank. Moreover, the results in Fig. 2 suggest that TFIDF is the most effective term weighting. However, CONF and CONV appear to be candidate measures for our method. Possible reasons why these three combinations are the best are given as following. First, BG works well on the related relations at higher ranks because BG can effectively solve the context ambiguity. At lower ranks, the relations are not rather related, therefore the terms counted by the pattern of BG are rarer than those checked by the form of UG. Second, for the case that a news relation (say $news1 \rightarrow news2$) whose document sizes are very different (assume that document size of $news2$ is greater than document size of $news1$), when applying LIFT measure to the relation which is CR, the strength of the relation will be affected by $news2$, due to computing with the equation of LIFT ($lift(X \rightarrow Y) = P(X \cap Y)/P(X)P(Y)$). Thus, LIFT may push the CR relation to the lower ranks which are inappropriate order, but CONF and CONV do not effect them. Consequently, using CONF or CONV for finding semantic relations

is more effective than using LIFT. Conversely, when using LIFT measure on the relation which is UR, because the LIFT equation also uses the weight of $news2$ for measuring its strength, $news2$ has great influence in distinguishing unrelated relation from related relation. This is why the UR relations are located to appropriate orders which are the lower ranks. As a result, applying LIFT on the lower ranks appears to be effectiveness.

### 5.3 Discussion

Once the experimental results obtained by our methods are established, the discussion and error analysis are then given in this section. Our proposed method still misclassified some instances. Some relations assigned by human as CRs are identified by low value of association measures. Causes of this error can be concluded as two cases corresponding to the document size. The first case is when size of two related news documents ($news1$ and $news2$) is very different. In this case, we found out that a news article ($news1$) is a summary of another news article ($news2$), agreeing with the document size ratio which shows that two news articles have very different size. Therefore, the measure value is low because the content overlap between two news articles is not frequent, This leads the proposed method to select the SH relation instead of the CR relation. We plan to overcome this problem in our future works by assigning more weight to headlines of news articles since they comprise a number of the same words. Furthermore, the publishing times of news articles will be considered because the CR relations are rather published in the same time or quite close to the same time. The second case why the relations assigned by human as CRs are identified by low value of association measures is when the document size ratio is not quite different. The main errors of this case are listed as follows. First, although the relations are completely related, two news articles can be written in the different styles due to different publishers who may use various words for representing same meaning words. A synonym detection approach can help for correcting this error. Second, different publishers deliver dissimilar contents for the exactly same event because they may write a news story with either unequal facts or contrasting details. For unequal facts, we plan to give the importance on the news headlines and news contents positioned at the first paragraph of news documents, by weighting them more heavily since the publishers rather bring out the similar contents on these positions. With contrasting details, their different data should be removed before using the proposed method. We plan to project a news difference problem as another research topic. Much more interesting in our discovery is the support-lift framework in performing well in the lower ranks. This is our further study on future works for analyzing on a hybrid method since it considers the number of qualitative criteria.

## 6 Conclusions

In this paper, we have investigated the effects of two term representation bases, four term weighting, and three association measure to discover relations among

news articles. Totally twenty four combinations are explored. To evaluate the quality of discovered news associations, top-$k$ ranked relations are analyzed by rank-order mismatch. By comparing the results to the human judgements rendered by the vote of three assessors, the experimental results show that the ROM values under the combination of BG-TFIDF-CONF, as well as BG-TFIDF-CONV, is suitable to achieve finding related 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%. These results suggest that, for relevant relations like completely related relations, the combination of term frequency with inverse document frequency together with bigram, and either confidence or conviction is applicable, and that together with unigram and lift works well in distinguishing unrelated relations from related relations. For somehow relations, they are still gray areas which need further analysis. As future works, we plan to work on various features by considering more factors and more criteria on each factor for enhancing the quality of document relations. Furthermore, the direction of news relations and deeper analysis on the characteristics of relations will be examined.

## Acknowledgements

This work was granted by Strategic Scholarships Fellowships Frontier Research Networks 2006 from the Commission on Higher Education and Thailand Research Fund under project number BRG50800013.

## References

1. Thompson, P., Cybenko, G., Giani, A.: Cognitive Hacking, ch. 19. Book of Economics of Information Security, pp. 255–287. Springer, US (2004)
2. Ferizis, G., Bailey, P.: Towards practical genre classification of web documents. In: Proc. 15th international conference on World Wide Web, pp. 1013–1014. ACM, New York (2006)
3. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. Coling 2004, Geneva, Switzerland, COLING, August 23-27, pp. 611–617 (2004)
4. Carreira, R., Crato, J.M., Gonçalves, D., Jorge, J.A.: Evaluating adaptive user profiles for news classification. In: Proc. 9th international conference on Intelligent user interfaces, pp. 206–212. ACM, New York (2004)
5. Antonellis, I., Bouras, C., Poulopoulos, V.: Personalized news categorization through scalable text classification. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 391–401. Springer, Heidelberg (2006)
6. Mengle, S., Goharian, N., Platt, A.: Discovering relationships among categories using misclassification information. In: Proc. 2008 ACM symposium on Applied computing, pp. 932–937. ACM, New York (2008)
7. Zhang, N., Watanabe, T., Matsuzaki, D., Koga, H.: A novel document analysis method using compressibility vector. In: Proc. the First International Symposium on Data, Privacy, and E-Commerce, November 2007, pp. 38–40 (2007)

8. Weixin, T., Fuxi, Z.: Text document clustering based on the modifying relations. In: Proc. 2008 International Conf. on Computer Science and Software Engineering, December 2008, vol. 1, pp. 256–259 (2008)
9. Lin, F., Liang, C.: Storyline-based summarization for news topic retrospection. Decision Support Systems 45(3), 473–490 (2008)
10. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report. In: Proc. the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (1998)
11. Papka, R., Allan, J.: Topic Detection and Tracking: Event Clustering as a Basis for First Story Detection, ch. 4. Book of Advances Information Retrieval: Recent Research from the CIIR, pp. 96–126. Kluwer Academic Publishers, Dordrecht (2006)
12. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. International Transactions on Computer Science and Engineering 32(1), 71–82 (2006)
13. Sriphaew, K., Theeramunkong, T.: Quality evaluation for document relation discovery using citation information. IEICE Trans. Inf. Syst. E90-D(8), 1225–1234 (2007)
14. Kittiphattanabawon, N., Theeramunkong, T.: Relation discovery from thai news articles using association rule mining. In: Chen, H., Yang, C.C., Chau, M., Li, S.-H. (eds.) PAISI 2009. LNCS, vol. 5477, pp. 118–129. Springer, Heidelberg (2009)
15. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. the 20th International Conf. on Very Large Data Bases, San Francisco, CA, USA, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994)
16. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed association rule mining. Technical report, Computer Science, Rensselaer Polytechnic Institute (1999)
17. Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. on Knowl. and Data Eng. 17(4), 462–478 (2005)
18. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov. 8(1), 53–87 (2004)
19. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: Measure and statistical validation. In: Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43, pp. 251–275. Springer, Heidelberg (2007)
20. Azevedo, P.J., Jorge, A.M.: Comparing rule measures for predictive association rules. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 510–517. Springer, Heidelberg (2007)
21. David, H.: The Method of Paired Comparisons. Oxford University Press, Oxford (1988)