2008-01-01

Utilizing Phrase-Similarity Measures for Detecting and Clustering Informative RSS News Articles

Yiu-Kai D. Ng
ng@cs.byu.edu

Maria Soledad Pera

Original Publication Citation

Maria Soledad Pera and Yiu-Kai Ng. "Utilizing Phrase-Similarity Measures for Detecting and Clustering Informative RSS News Articles." Journal of Integrated Computer-Aided Engineering (ICAE), Volume 15, Number 4, pp. 331-35, 28, IOS Press. (This is an extended version of the KSEM 27 paper, Finding Similar RSS News Articles Using Correlation-Based Phrase Matching, which was selected and invited to be published by ICAE.)

Utilizing Phrase-Similarity Measures for Detecting and Clustering Informative RSS News Articles

Maria Soledad Pera
Yiu-Kai Ng
Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
Email: ng@cs.byu.edu*, mpera@cs.byu.edu

Abstract

As the number of RSS news feeds continue to increase over the Internet, it becomes necessary to minimize the workload of the user who is otherwise required to scan through huge number of news articles to find related articles of interest, which is a tedious and often an impossible task. In order to solve this problem, we present a novel approach, called *InFRSS*, which consists of a *correlation-based phrase matching* (CPM) model and a *fuzzy compatibility clustering* (FCC) model. CPM can detect RSS news articles containing phrases that are the same as well as semantically alike, and dictate the degrees of similarity of any two articles. FCC identifies and clusters non-redundant, closely related RSS news articles based on their degrees of similarity and a fuzzy compatibility relation. Experimental results show that (i) our CPM model on matching bigrams and trigrams in RSS news articles outperforms other phrase/keyword-matching approaches and (ii) our FCC model generates high quality clusters and outperforms other well-known clustering techniques.

Keywords: Information Search, phrase matching, clustering, fuzzy-set IR model

1 Introduction

Phrase queries are frequently used to retrieve documents from the Web. A phrase, which is often defined as a *sequence of words* [10], can be represented in two folds: (i) the syntactic structure that the words are organized in, and (ii) the semantic content it delivers. Changing either one of the two representations may result in a phrase with a different meaning.

*Corresponding Author

Traditional phrase matching techniques aim to retrieve documents including phrases that match exactly with the query phrase, although some advanced approaches tolerate errors to some extent (e.g., proximity of words, word order, and missing words in a phrase). These inherent characteristics draw restrictions on their potential usages, i.e., they may fail to detect potentially relevant phrases and hence documents. For example, the phrase "heterogeneous node" (on wireless networks) is semantically relevant to "heterogeneous device" and "heterogeneous transport," which could be used along with "heterogeneous node" in retrieving closely related documents.

Neither keyword matching (nor traditional phrase matching as mentioned earlier) can solve the inexact phrase matching problem. Using keywords "heterogeneous" and "node" individually in keyword search could match documents that include either the word "heterogeneous" or "node," but not necessarily both, and thus the content of retrieved documents might be totally unrelated to "heterogeneous node." Some of these documents may address "heterogeneous alloys," whereas others may discuss "homogeneous node." Even though the "matched" documents include both words, they are not necessarily in the same order, which might run into the same "content mismatched" problem. The more sophisticated similarity matching approaches, such as [33], can detect documents that include similar (not necessarily the same) words; they, however, cannot resolve the word-ordering problem. For example, consider the sentences "They *jog* for thirty minutes and *walk* for an hour" and "They *run* for an hour and *stroll* for thirty minutes." Ignoring the word order and simply considering the degrees of (single-)word similarity, i.e., *jog* versus *run* and *walk* versus *stroll*, causes these sentences to be treated as closely related, even though they are semantically different, and filtering out mismatched documents manually is a waste of time.

In this paper, we propose a similarity matching and clustering approach, denoted *InFRSS*, for detecting and clustering informative RSS news articles, which consists of two sub-models: (i) a *correlation-based phrase matching* (CPM) model that can detect RSS news articles containing semantically the same (or similar) phrases, and (ii) a *fuzzy compatibility clustering* (FCC) model that clusters non-redundant, informative RSS news articles based on fuzzy compatibility relation to obtain cohesive clusters. We are interested in

RSS news articles, since there is no precedent in the amazing amount of online news that can be accessed by Internet users these days. Thus, the problem of seeking information in online news articles is not the lack of them but being overwhelmed by them. This brings a huge challenge in finding and grouping related online news with distinct information automatically, instead of manually, which is a labor-intensive and impractical process.

The proposed CPM model measures the degrees of similarity among different RSS news articles using phrase similarity to detect *redundant* and discover *similar* news articles. We call the proposed model *correlation-based*, since we adapt the correlation factors in fuzzy sets to model the similarity relationships among different phrases. For each phrase p, its fuzzy set S is constructed that captures the *degrees of memberships*, i.e., closeness, of p to all the other phrases in S, which are called *phrase correlation factors*. The proposed FCC model, on the other hand, identifies and discards RSS news articles that are considered to be less-informative using the degree of similarity among RSS news articles, and applies the restrictions imposed by the fuzzy compatibility relation for clustering the remaining informative articles.

The rest of the paper is organized as follows. In section 2, we discuss research work in phrase matching and document clustering. In section 3 and section 4, we present the design of InFRSS. In section 5, we verify the accuracy of CPM in detecting related documents and the effectiveness of FCC for clustering RSS news articles using various test cases. In section 6, we include a conclusion and directions for future work for InFRSS.

2 Related Work

Phrase matching has been applied in solving different problems, such as ranking relevant documents, document clustering [10], and Web document retrieval [1]. In [1], a system for matching phrases in XML documents, called PIX, is presented. PIX allows users to specify both (i) tags and annotations in an XML document to ignore and (ii) phrases in the document to be matched. This technique relies on *exact* and *proximity* phrase matching (i.e., words in a phrase that are within a distance of $k(\geq 1)$-words in a document) in retrieving relevant documents.

[10] cluster Web documents based on matched phrases and their levels of significance (e.g., the title and the body) in the documents. This method uses *exact* phrase matching to determine the degrees of overlap among documents, which yield their degrees of similarity.

[30] use phrase matching for ranking medical documents. The similarity of any two phrases in [30] is detected by the number of consecutive three-*letter* triples in common, with various scores assigned to different triples of letters, e.g., uncommon three-letter triples are given a *higher weight*, three-letter triples at the beginning of a word are *more important* than the ones at the end of the word, and long phrases are discounted to avoid bias on their lengths. In [30] phrases are treated as documents and tri-grams of letters are treated as words.

[20], who use phrase and *proximity* terms for Web document retrieval and treat every word in a query as a phrase, show that the usage of phrases and proximity terms is highly beneficial. However, their experimental results show that even though phrases and proximity terms have a positive impact on 2- or 3-word queries, they have less, or even negative, effects on other types of queries.

[8] present a compression method that searches for words and phrases on natural-language text. This method performs an *exact* search for words and phrases on compressed text directly using any sequential pattern-matching algorithm, in addition to a word-based *approximate* for extended search. Thus, searches can be conducted for approximated occurrences of a phrase pattern.

[22] emphasize the importance of phrase extraction, representation, and weighting and claim that phrases obtained by syntactic (instead of statistical) processing often increase the effectiveness of retrieval when proximity and weighting information are adequately attached to a query phrase representation. [28], however, determine the degree of similarity between any two documents by computing the number of common phrases in the documents, and dividing the number of common phrases by the total number of phrases in both, which is intuitively another *exact* phrase matching approach.

In terms of clustering, an incremental hierarchical text document clustering approach used for organizing documents from various online sources is presented in [26]. The cluster-

ing method depends on the frequency of occurrence, as well as the contents, of the words within documents, which is another term-frequency and word-matching approach to determine the topic of a document so that documents on similar topics are clustered. [21] capture the structure of online news events that make up topic and the dependencies among them (i.e., event threading) through different event models. The use of cosine similarity and time-stamps of news stories in [21] produces fairly good results when the events are provided. The performance, however, deteriorates rapidly if the system has to discover the events itself. [15] also consider RSS news articles and allow the user to find articles grouped by similar topics. In [15], the k-nearest neighbor algorithm locates the k nearest stories for each new story S so that the cosine similarity in VSM (Vector Space Model) computed for each of the k stories and S is not lower than the predefined threshold; otherwise, the content of S is treated as a new topic. [4] use three different variations of the k-mean algorithm to find higher quality solutions in less time for clustering binary data streams. This algorithm partitions a dataset into k different clusters using a simple iterative scheme to find a locally minimal solution. Even though the results of the incremental k-mean are good, dependence in initialization, sensitivity to outlier, and skewed distributions could affect the performance of the algorithms in [4]. During the process of document cluster analysis, [16] incorporate the user's prior knowledge, which indicate pairs of documents that are known to belong to the same cluster, to obtain the desired cluster structures or to construct accurate clusters. This technique enables the user to control the clustering process based on the prior knowledge specific to the target data set, which is a constraint.

[6] present an Extended Suffix Tree Clustering algorithm, which uses a scoring function to maximize topic coverage and reduce overlapping among different clusters in order to reduce the number of clustered documents that are presented to the user. Similar to our clustering approach, [6] remove stopwords and stem non-stopwords, use phrases to perform clustering, and allow document overlapping among different clusters. However, [6] do not handle document replication prior to perform clustering and require tree to be first built and then pruned to eliminate low-scoring clusters, which is not cost effective.

[7] propose a method for identifying semantically meaningful groups (i.e., clusters) of

Web pages. [7] construct a base cluster, which is described by a single word, for each of the words that appear in more than 4% of a collection of documents. Hereafter, the distance[1] between the user query and the word describing a cluster is calculated so that clusters with a distance value higher than the threshold are eliminated. The remaining clusters are merged/split accordingly to assure the high quality of the clusters, and using an improved version of the approach in [6] only the high-scoring clusters are returned to the users. The clustering approach in [7], however, is highly depended on user's queries, which may not always be correctly formulated, and thus could affect the quality of the clusters generated.

3 Correlation-Based Phrase Matching

Semantically relevant phrases detected by our CPM model hold the same syntactic features as in other phrase matching approaches, i.e., a phrase is treated as a *sequence* of words and the *order* of words is significant. Unlike existing phrase matching approaches, we develop novel *phrase correlation factors* for the n-gram ($1 \leq n \leq 5$) phrases. Using one of these chosen sets of n-gram phrase correlation factors, the n-gram phrases in an RSS news article are matched against the n-gram phrases in another article to determine their degrees of similarity. We detail the design of our n-gram CPM approach on RSS news articles below.

3.1 Content Descriptors of RSS News Articles

Two of the essential elements in an RSS (XML) news feed file, in which RSS news articles are posted, are the *title* and *description* of an *item* (i.e., a news article), since the former contains the headline and the latter includes the first few sentences of the article. Furthermore, several items can appear in the same RSS feed file. (See, as an example of, an RSS news feed file as shown in Figure 1.) We treat the title and description of each item as the *content descriptor* of the corresponding article and determine its degree of similarity with the *content descriptor* of another item (in the same or a different RSS news feed file) according to the correlation factors of phrases in the two content descriptors.

[1]The distance is computed by using the measure defined in [3].

3.2 Computing the Phrase Correlation Factors

Prior to computing n-gram ($1 \leq n \leq 5$) phrase correlation factors, we first decide at what level the correlation factors are to be calculated, which dictates how the subsequent process of phrase comparison should be conducted.

The major drawback of the phrase-level granularity is its excessive overhead. A phrase may start at any position in a document, and the lengths of phrases vary in practical usage. Thus, the number of possible phrases to be considered could be huge. For example, consider a portion of the paragraph that is randomly chosen from www.cnn.com: "... the organ's unwrinkled surface resembled **that of the** brain of an idiot. ... Researchers contend **that if the** plant-eating beasts ... ." Even only considering trigram phrases, there are 71 trigram phrases in the entire paragraph. However, not all of them, such as the phrases "that of the" and "that if the," are useful in determining the content of the paragraph, or its corresponding document in general. Thus, our CPM model, which pre-computes the correlation factors of any two n-gram phrases, considers only *non-stop, stemmed words*[2] in an RSS news article to form phrases to be matched.

3.2.1 The Unigram Correlation Factors

We construct the unigram (i.e., single-word) correlation factors using the documents in the Wikipedia Database Dump (http://en.wikipedia.org/wiki/Wikipedia:Database_download). We chose the Wikipedia documents for constructing each of the n-gram ($1 \leq n \leq 5$) correlation factors, since the 850,000 Wikipedia documents were written by more than 89,000 authors on various topics. The diversity of the authorships leads to a representative group of documents with different writing styles and a variety of subject areas. Thus, the set of Wikipedia documents is an effective representative set of documents that is appropriate for computing the general correlation factors among unigram, as well as other n-gram ($2 \leq n \leq 5$), phrases. The correlation factors of the unigrams are computed according to the *distance* and *frequency of occurrence* of the unigrams in each Wikipedia document.

[2] *Stopwords* are words that appear very frequently (e.g., "him," "with," "a," etc.), which include articles, conjunctions, prepositions, punctuation marks, numbers, non-alphabetic characters, etc., and are typically not useful for analyzing the informational content of a document. *Stemmed words* are words with the same grammatical root.

Prior to constructing the unigram correlation factors, we first removed all the words in the Wikipedia documents that are *stopwords*. Eliminating stopwords is a common practice, since the process (i) filters the noise within both a query and a document [11, 19] and (ii) enhances the retrieval performance [29], which enriches the quality of our unigram phrase correlation factors. After stopwords were removed, we stemmed the remaining words by using the Porter Stemmer [25], which stems each word to its grammatical root but retains the semantic meaning of the words, e.g., "driven" and "drove" are reduced to their stemmed word "drive." The final count of non-stop, stemmed unigrams is 57,926. The *unigram correlation value* of word w_i with respect to word w_j, which is constructed by using the Wikipedia documents without stop- or non-stemmed words, is defined as

$$c_{i,j} = \sum_{w_i \in V(w_i)} \sum_{w_j \in V(w_j)} \frac{1}{d(w_i, w_j)} \tag{1}$$

where $d(w_i, w_j) = |Position(w_i) \text{ - } Position(w_j)|$ is the distance, i.e., the number of words, between w_i and w_j in a Wikipedia document, and $V(w_i)$ ($V(w_j)$, respectively) denotes the set of stem variations of w_i (w_j, respectively).

Correlation factors among unigrams that co-occur more frequently than others in a document are assigned higher values. To avoid the bias on the frequency of occurrences in a "long" Wikipedia document, we *normalize* $c_{i,j}$ as

$$nc_{i,j} = \frac{c_{i,j}}{|V(w_i)| \times |V(w_j)|} \tag{2}$$

where $|V(w_i)|$ ($|V(w_j)|$, respectively) is the number of words in $V(w_i)$ ($V(w_j)$, respectively).

Given k different $nc_{i,j}$ values, one from each of the Wikipedia documents in which both w_i and w_j occur, the *unigram correlation factor* $cf_{i,j}$ of w_i and w_j is

$$cf_{i,j} = \frac{\sum_{m=1}^{k} nc_{i,j}^{m}}{k} \tag{3}$$

where $nc_{i,j}^{m}$ is the normalized correlation value $nc_{i,j}$ (as defined in Equation 2) of w_i and w_j in the m^{th} ($1 \leq m \leq k$) document in which both w_i and w_j occur, and k is the total number of Wikipedia documents in which w_i and w_j co-occur.

3.2.2 The N-gram Phrase Correlation Factors

The phrase correlation factors of any two n-grams ($2 \leq n \leq 5$) are calculated according to the correlation factors of their corresponding unigrams, since unigram correlation factors are reliable in detecting similar words(see experimental results in section 5.1.) We compute the bigram, trigram, 4-gram, and 5-gram phrase correlation factors using the (prior) Odds (*Odds* for short) [12] that measures the predictive or prospective support according to a hypothesis H by the prior knowledge $p(H)$ alone to determine the strength of a belief, which is the unigram correlation factor in our case.

$$O(H) = \frac{p(H)}{1 - p(H)} \tag{4}$$

Based on the computed unigram correlation factors, $cf_{i,j}$, in Equation 3, we generate the n-gram ($2 \leq n \leq 5$) phrase correlation factors between any n-gram phrases p_1 and p_2 using Equation 4 such that $p(H)$ is defined as the *product of the unigram correlation factors* of the corresponding unigrams in p_1 and p_2, i.e.,

$$pcf_{p_1,p_2} = \frac{\prod_{i=1}^{n} cf_{p_{1_i},p_{2_i}}}{1 - \prod_{i=1}^{n} cf_{p_{1_i},p_{2_i}}} \tag{5}$$

where p_{1_i} and p_{2_i} ($1 \leq i \leq n$) are the i^{th} word in p_1 and p_2, respectively.

3.3 Phrase Comparison

Phrases in the content descriptor of an RSS news article are compared against their counterparts in another RSS news article. CPM can detect phrases in RSS news articles that are semantically relevant (or the same) to phrases in other articles. To accomplish this, the phrase of a chosen length k ($1 \leq k \leq 5$) in a news article A_1 is compared with each phrase of the same length in another article A_2. If there are m (n, respectively) different words in A_1 (A_2, respectively), then there are m-k+1 different phrases in A_1 to be compared with n-k+1 different phrases in A_2, which include overlapped phrases. In computing the degree of similarity of A_1 and A_2, the correlation factors of phrases of the chosen length, i.e., in between 1 and 5, in A_1 and A_2 are used. Since the average content descriptor of an RSS

new article is 25 words in length, the computation time for matching phrases in two news articles is negligible.

3.4 Similarity Ranking of RSS News Articles

In CPM, we use the correlation factors of n-gram ($1 \leq n \leq 5$) phrases of a chosen length to define the degrees of similarity of two RSS news articles A_1 and A_2. The *degree of similarity* of A_1 with respect to A_2 is not necessary the same as the *degree of similarity* of A_2 with respect to A_1, since A_1 and A_2 may share common information but also include information that are unique of their own.

Using the n-gram ($1 \leq n \leq 5$) phrase correlation factors, we define a fuzzy association of each n-gram phrase in A_1 with respect to all the n-gram phrases in A_2. The degree of correlation between a phrase p_i in A_1 and all the phrases in A_2, denoted $\mu_{p_i,2}$, is calculated as the complement of a negated algebraic product of all the correlation factors of p_i and each distinct phrase p_k in A_2, i.e.,

$$\mu_{p_i,2} = 1 - \prod_{p_k \in A_2} (1 - pcf_{i,k}) \quad \text{or} \quad \mu_{p_i,2} = 1 - \prod_{p_k \in A_2} (1 - cf_{i,k}) \tag{6}$$

which is adapted from the fuzzy word-document correlation factor in [23], and the 1^{st} (2^{nd}, respectively) formula in Equation 6 is used for n-gram ($2 \leq n \leq 5$) phrases (unigram phrases, respectively). The correlation value $\mu_{p_i,2}$ falls in the interval [0, 1] and reaches its maximum when $pcf_{i,k}$ ($cf_{i,k}$, respectively) $= 1$, i.e., when p_i ($\in A_1$) $= p_k$ ($\in A_2$).

The *degree of similarity* of A_1 with respect to A_2, denoted $Sim(A_1, A_2)$, using the chosen n-gram phrase correlation factors is calculated as the average of all the values $\mu_{p_i,2}$ for each $p_i \in A_1$ ($1 \leq i \leq m$), and m is the total number of n-gram phrases in A_1.

$$Sim(A_1, A_2) = \frac{\mu_{p_1,2} + \mu_{p_2,2} + \cdots + \mu_{p_m,2}}{m} \tag{7}$$

$Sim(A_1, A_2) \in [0, 1]$. When $Sim(A_1, A_2) = 0$, it indicates that there is no n-gram phrase in A_1 that can be considered similar to any n-gram phrase in A_2. If $Sim(A_1, A_2) = 1$, then either A_1 is (semantically) identical to A_2, or A_1 is *subsumed* by A_2, i.e., all the n-gram phrases in A_1 are (semantically) the same as (some of) the n-gram phrases in A_2, and A_1 is treated as a *redundant* article and can be ignored. $Sim(A_2, A_1)$ can be defined accordingly.

4 Clustering Non-Redundant RSS News Articles

Users, who subscribe to different RSS feeds, expect to receive up-to-the-minute news of their interests from multiple sources. Since RSS news are updated constantly and news feeds are created on a regular bases, the amount of RSS news articles that are made available to the users constantly increase. In processing the constant flow of these information in a timely manner, it is essential not only to eliminate redundant (i.e., replicated or subsumed) RSS news articles, but also to cluster the news articles that share related information and present them as a unit to the users, which should save the users significant amount of time and efforts in browsing through the subscripted RSS feeds for gathering information of interest. In this section, we present FCC an elegant approach for clustering closely related RSS news articles to facilitate information gathering and processing.

4.1 Our Clustering Approach

After redundant RSS news articles are identified and discarded, we proceed to cluster the remaining non-redundant news articles according to the following equation.

$$C_\alpha = \{\ d \mid Sim(d, e) \geq \alpha, \forall e \in C\ \} \tag{8}$$

where α is the *minimum* degree of similarity that any two RSS news articles in the same cluster C_α must hold.

In order to create clusters of RSS news articles according to the clustering criteria specified in Equation 8, we consider a fuzzy equivalence relation on any set of RSS news articles S to generate equivalence classes of S so that (i) each of these classes yields a cluster with closely related RSS news articles and (ii) the degrees of similarity of articles among different clusters are low. The fuzzy equivalence relation is an ideal choice for creating clusters, since as mentioned in [13] a fuzzy equivalence relation is appropriate for specifying the characteristics of the elements (i.e., RSS news articles in this paper) that exist in a certain partition (i.e., set of clusters). A fuzzy equivalence relation R is *reflexive*, *symmetric*, and *max-min transitive* [34] as defined below.

$$R(x, x) = 1, \forall x \in Y \tag{9}$$

$$R(x, y) = R(y, x), \forall x, y \in Y \tag{10}$$

$$R(x, z) \geq max_{y \in Y} min\{R(x, y), R(y, z)\} \tag{11}$$

where Y is a fuzzy set and x, y, and $z \in$ Y.

It is necessary to use a function to combine the similarity measures among any two RSS news articles into a single value so that the max-min transitivity constraint can be applied. As previously mentioned, given any two RSS news articles A_i and A_j, $Sim(A_i, A_j)$ might not be the same as $Sim(A_j, A_i)$. Hence, we combine these two values into one, which reflects the combined degree of similarity between A_i and A_j.

4.1.1 Combined Similarity Values

The combined Sim values of any two RSS news articles A_i and A_j should reflect how closely related the two articles are and yield the relative degree of similarity between the two articles. We apply the *Stanford Certainty Factor* (scf) [18], as defined in Equation 12, on $Sim(A_i, A_j)$ and $Sim(A_j, A_i)$, to obtain the *combined* degree of similarity of A_i and A_j, as shown in Equation 13.

$$CF(C) = \frac{CF(R_1) + CF(R_2)}{1 - MIN(|CF(R_1)|, |CF(R_2)|)} \tag{12}$$

where R_1 and R_2 are two hypothesis that reach the same conclusion C and CF is the *certainty factor* (i.e., confidence measure) associated with C, which is a monotonically increasing (decreasing) function on combined assumptions for creating confidence measures.

$$scf(A_i, A_j) = \frac{Sim(A_i, A_j) + Sim(A_j, A_i)}{1 - MIN(Sim(A_i, A_j), Sim(A_j, A_i))} \tag{13}$$

Given any two RSS news articles A_i and A_j and their Sim values as shown in Table 1, we observe that only when the Sim values of the two articles are both *high*, their scf is also *high* (e.g., A_1 and A_2 in Tables 1 and 2). If one of the Sims is *high* and the other one is *low*, then their scf is *relatively low* (e.g., A_2 and A_5 in Tables 1 and 2). However, if both Sim are *low*, then their scf is *very low* (e.g., A_3 and A_5 in Tables 1 and 2).

We normalized scf so that the normalized scf is bounded between 0 and 1 as

$$nscf(A_i, A_j) = \frac{scf(A_i, A_j)}{max(scf(A_x, A_y))} \tag{14}$$

where A_i and A_j are two RSS news articles in a set of RSS news articles S, and A_x and A_y represent the two articles in S that yield the *highest scf* value among all the pairs of articles in S.

Our *nscf* function, however, does not satisfy the *max-min transitivity* property as defined in Equation 11, which requires that for any two articles A_x and A_z that belong to the same equivalence class C (i.e., cluster), the existence of another article A_y in C which has similarities with A_x, as well as A_z, that are greater than the similarity between A_x and A_z is *disallowed*. In other words, a relation R is not *max-min transitive* if given any two articles A_x and A_z, there exists another article A_y such that $R(A_x, A_z) < R(A_x, A_y)$ and $R(A_x, A_z) < R(A_y, A_z)$, which does not apply to our similarity measure. Consider articles A_1, A_2, and A_5 in Table 3, which are closely related (i.e., on writers strike) and should be clustered by themselves. Assume that A_1 is A_x, A_5 is A_z, and A_2 is A_y. Note that $nscf(A_1, A_5) \ngeq max(min\{nscf(A_1, A_2), nscf(A_2, A_5)\})$ as shown in Table 2, and their *Sim* values are shown in Table 1. We consider another transitivity function, the *max-prod transitivity* [13], which is *less restrictive* than the *max-min transitivity*, as defined below.

$$R(x, z) \geq max_{y \in Y}\{R(x, y) \times R(y, z)\} \tag{15}$$

As mentioned in [13], the max-prod transitivity constraint is easily satisfied by any function with arguments in the interval [0, 1], which are the *nscf* values. This is because the product of any two numbers a and b that are in the interval [0, 1] is smaller than a and b, and hence the max-prod transitivity is easier to satisfy than the max-min transitivity.

Consider the *nscf* values (in Table 2) generated by using the RSS news articles A_1, A_2, and A_5 in Table 3 again. Although the three articles cover the same topic, i.e., the ongoing writers strike in Hollywood at the time the RSS news articles were collected, the restriction imposed by the max-prod transitivity constraint in the fuzzy equivalence relation with A_1, A_2, and A_5 is violated, since $nscf(A_1, A_5)$ $(= 0.53) \ngeq nscf(A_1, A_2) \times nscf(A_2, A_5)$ $(= 0.97 \times 0.75 = 0.73)$. Hence, our *nscf* does not satisfy the max-prod transitivity.

	A_1	A_2	A_3	A_4	A_5
A_1	1.00	0.30	0.20	0.10	0.30
A_2	0.50	1.00	0.13	0.13	0.50
A_3	0.14	0.07	1.00	0.21	0.07
A_4	0.03	0.03	0.08	1.00	0.33
A_5	0.20	0.20	0.05	0.45	1.00

Table 1: Similarity values computed for different pairs of articles using bigram phrases in Table 3.

Pairs of Articles	scf	$nscf$	Pairs of Articles	scf	$nscf$
A_1, A_2	1.14	0.97	A_2, A_4	0.16	0.13
A_1, A_3	0.40	0.34	A_2, A_5	0.88	0.74
A_1, A_4	0.13	0.11	A_3, A_4	0.32	0.28
A_1, A_5	0.63	0.53	A_3, A_5	0.13	0.11
A_2, A_3	0.21	0.18	A_4, A_5	1.18	1.00

Table 2: The combined similarity values, i.e., scf, and normalized combined similarity values, i.e., $nscf$, between the articles in Table 3 with their similarity values as shown in Table 1.

Articles	Title and (Portion of the) Body
A_1	Tonight Show' Staff May Get Stiffed. Network says show staff could be laid off next week if writers strike continues. ...
A_2	Shows Knocked Off Air by Writers Strike. See the complete list of shows going on hiatus. ...
A_3	Late Night Jokes Will Be Old Tonight: Leno, Letterman in Reruns. Tina Fey, creator and star of "30 Rock," was among those picketing in NYC. ...
A_4	Globes Scene Missing Magic and Meaning. A roar went up from the lobby bar of the Beverly Hilton before Sunday night's Golden Globes. ...
A_5	Globes Aren't a Hit With TV Viewers. The Hollywood writers strike took the glitz, the glamour ...

Table 3: RSS news articles A_1, A_2 and A_3 (A_4 and A_5, respectively) downloaded from abcnews.go.com (hosted.ap.org, respectively) on November 2007 (January 2008, respectively).

4.1.2 The Fuzzy Compatibility Relation

Even though neither the max-min nor max-prod transitivity can be used for clustering the RSS news articles based on our $nscf$ function, we consider yet another fuzzy relation, i.e., *fuzzy compatibility relation* [13]. A given relation R is said to be a *fuzzy compatibility relation* if R is *reflexive* (as defined in Equation 9) and symmetric (as defined in Equation 10), i.e., fuzzy transitivity is not imposed on a fuzzy compatibility relation.

According to the fuzzy compatibility relation, we can create clusters of RSS news articles based on their $nscf$ values. Since the $nscf$ of any pair of articles A_1, A_2, and A_5 in Table 3 satisfies the *reflexive* and the *symmetric* properties, the three articles are assigned to the same cluster, which demonstrates that the fuzzy compatibility relation is an ideal choice for clustering related RSS news articles.

4.2 Clustering with α-Cuts

Prior to applying the fuzzy compatibility relation on non-redundant RSS news articles to generate clusters of closely related articles, we must first determine the least degree of similarity that any two articles should have prior to be assigned to the same cluster. Since the use of *bigrams* has been proved to be the most effective approach (as shown in Section 5.1) in determining how closely related any two RSS news articles are, we focus on clustering RSS news articles whose degrees of similarity are computed by using bigrams in the articles.

We first define an α value in [0, 1], which restricts the degrees of membership of elements in a particular cluster, and when applied to a fuzzy set S (i.e., set of RSS news articles) constructs various crisp subsets of S, denoted α-*cut*. In other words, the α-*cut* of S yields a set of clusters that contains all the elements (i.e., RSS news articles) of S so that each cluster C consists of elements in S whose degrees of membership is greater than or equal to the specified α value.

Determining an α value that generates the most ideal α-*cut* for S is essential in FCC, since a *low* α-*cut* value yields *looser* clusters (i.e., elements within each cluster might not be closely related) that are *fewer* in number, whereas a *high* α-*cut* value yields *tighter* clusters,

which as a side-effect might significantly increase the (excessive) number of (singleton) clusters generated.

Example 1 Table 4 shows three different RSS news articles whose Sim and $nscf$ values are shown in Table 5 and Table 6, respectively, whereas Table 7 illustrates how they are clustered together using different α values, which generate different α-*cut*.

As shown in Table 7, when α is too *small* (i.e., 0.004), unrelated articles (i.e., A_3 with respect to A_1 and A_2) are clustered together, whereas when α is too *large* (i.e., 0.02), articles that should belong to the same partition (i.e., A_1 and A_2) are separated, which illustrates the importance of selecting the proper α value for clustering. □

In order to establish the appropriate α-*cut*, we introduce the equation given below, which was determined empirically using various test sets of RSS news articles of various sizes. The test sets, detailed in Table 8, consist of randomly selected groups of RSS news articles collected between January 2007 and January 2008. Since the RSS news articles (i) were collected from many different sources and (ii) covered a variety of topics, they are representative and hence appropriate for determining the correct value of α.

$$\alpha = (25.23 \times 10^{-6} \times \mathrm{N}umberofInformativeRSSNewsArticles) + 1.49 \times 10^{-3} \qquad (16)$$

In determining Equation 16, we used nine different test sets TS, as shown in Table 8, and clustered the articles in each TS using different values of α and determined the α value that yielded *tight* clusters (i.e., the RSS news articles in each generated cluster share a common topic or story), but at the same time the number of *singletons* (clusters that contain only one RSS news article) was kept to a minimum.

To verify the correctness of Equation 16, we determined the appropriate α value for each of the eight verification sets, VS, as shown in Table 9, which consists of groups of different sizes of RSS news articles downloaded from a variety of sources. We manually analyzed the generated clusters, and in each of the verification sets, the generated clusters satisfy the two criteria of ideal clusters using the selected value (i) the number of *singletons* is kept to a minimum and (ii) the articles within the generated clusters are *closely* related, whereas articles among various clusters are *different* in terms of their content.

Articles	Title and (Portion of the) Body
A_1	Osmond Back on 'Dancing With the Stars'. Marie Osmond says the Bible inspired her return to ABC's "Dancing With the Stars" following the death of her father last week. ...
A_2	'Dancing' Waltzes Through Writers Strike. Here's a "Dancing with the Stars" pop quiz: Which of the following performance critiques was delivered by effusive judge Bruno Tonioli before the Hollywood writers strike, and which came after?. ...
A_3	Broadway Strike Goes Into 4th Day. Could the Broadway work stoppage spread to touring companies of major musical hits such as "Wicked" or "Jersey Boys"?. ...

Table 4: RSS news articles downloaded from http://hosted.ap.org, on November 16, 2007.

Articles	A_1	A_2	A_3
A_1	1	0.61	0.17
A_2	0.45	1	0.23
A_3	0.18	0.29	1

Table 5: Similarity values computed for the articles in Table 4 using bigrams.

Articles	A_1	A_2	A_3
A_1	1		
A_2	0.019	1	
A_3	0.004	0.006	1

Table 6: $Nscf$ values computed for the articles in Table 4 using bigrams.

α value	Clusters
0.004	{ A_1, A_2, A_3 }
0.009	{ A_1, A_2 }, { A_3 }
0.020	{ A_1 }, { A_2 }, { A_3 }

Table 7: Clusters generated using different α-*cut*.

Test Cases	Number of Informative RSS Mews Articles	Sources	Topics
TS_1	7	abcnews.go.org	Politics, Local news
TS_2	14	cbsnews.com, fowxnews.com	International, Politics
TS_3	28	nytimes.com, msnbc.msn.com	Bussiness, Top Stories
TS_4	43	prnewswire.com, sfgate.com, usatoday.com	Health, Weather, Politics
TS_5	75	worldpress.org, wired.com	International, Bussiness
TS_6	107	cnn.com, abcnews.go.com, washingtonpost.com	Technology, Entertaintment, Top Stories
TS_7	150	sportillustrated.com, usatoday.com, sltrib.com	Sports, Local news, Politics, International
TS_8	238	abcnews.go.org, timesonline.co.uk, silliconvalley.com, prnewswire.com	Sports, Technology, Bussiness, Top Stories
TS_9	312	english.people.com, hosted.ap.org, seattletimes.nwsource.com, today.reuters.com	Local news, Weather, Entertaintment, Technology, Politics

Table 8: Test cases used for determining the ideal equation of *alpha*.

Verificaiton Sets	Number of Informative RSS News Articles	Sources	Topics
VS_1	7	abcnews.go.org	Politics, Local News
VS_2	19	sltrib.com, cbsnews.com	Bussiness, Top Stories
VS_3	43	abcnews.go.com, msnbc.msn.com	International, Politics, Top Stories
VS_4	89	prnewswire.com, weather.com, nytimes.com	Health, Weather, Travel
VS_5	112	usatodays.org, wired.com	Technology, Bussiness
VS_6	184	hosted.ap.com, abcnews.go.com, washingtonpost.com	Politics, Entertaintment, Top Stories
VS_7	201	money.cnn.com, englishpeople.com, newsbbc.co.uk	Bussiness, Entertaintment, Top Stories
VS_8	294	news.ft.com, news.yahoo.com, seattletimes.nwsource.com	International, Top Stories, Bussiness, Technology

Table 9: Test cases used for verifying the effectiveness of the defined α equation, i.e., Equation 16.

Example 2 Table 10 shows the titles and (a portion) of the bodies of the RSS news articles used in VS_1, Table 11 shows the $nscf$ values for the different pairs of news articles in Table 10, whereas Table 12 shows the generated clusters of RSS news articles using different α-*cut*s. For VS_1, which consists of seven RSS news articles, the α value computed by using Equation 16 is 0.0017. Clearly, as shown in Table 12, when the α value used for generating a particular α-*cut decrease*s, the number of clusters is reduced but the articles within the same clusters are not particularly similar. On the other hand, when α *increases*, singleton clusters are introduced and related articles are assigned to different clusters. Hence, 0.0017 (computed using Equation 16) is the appropriate value for clustering articles in VS_1. □

4.3 Discarding Less-Informative RSS News Articles

As previously stated, prior to cluster a given set of RSS news articles we removed those that are considered to be redundant; however, the number of remaining clustered articles to be presented to a user could still be excessive. Hence, a number of clustered articles, even though non-redundant, should be disposed, especially those articles that include most of the information that are also available in other articles, which we call less-informative news articles. After clusters are created using the α-value equation, i.e., Equation 16, we proceed to eliminate a percentage of the RSS news articles that are *less-informative* according to our *ranking* function (defined below) that will able us to determine which particular articles and in which order they should be eliminated. The ranking function not only affects the content of clusters in which less-informative articles appear, but also prevents (when possible) the elimination of singleton clusters which include an article that is dissimilar to all the other clustered articles, guaranteeing that no news article that reports unique information is lost.

In discarding the less-informative RSS news articles in various clusters generated by the α-*cut* computed by Equation 16, we use the ranking function as defined in Equation 17, which considers (i) the number of clusters (generated by using the α-*cut*) in which a particular RSS news articles appears (since a single RSS news article might cover a variety of topics and as a result it can appear in more than one cluster) and (ii) the degree of

Articles	Title and (Portition of the) Body
A_0	Obama Takes 1st Step in Presidential Bid. Sen. Obama takes step toward presidential bid ...
A_1	Bush Chides Iraq Over Recent Executions. President Bush Chides Al-Maliki's Administration Over ...
A_2	Edwards Supports War Oppostion. As Dems Debate Bush Plan, Ex-Sen. Edwards joins ...
A_3	Pain, Fury Still Rage a Year After Katrina. ABC News Poll Finds Loss, Frustration and Anger Linger ...
A_4	Romney Kicks Off Fundraising Campaign. Romney kicks off Presidential campaign with ...
A_5	Biden Says He Can Run With '08 Rivals. Democratic Sen. Joe Biden says he can hold his own ...
A_6	EXCLUSIVE: Supreme Court Justice Stevens Remembers President Ford. Exclusive Interview: John Paul Stevens Recalls Ford's ...

Table 10: RSS news articles in VS_1 downloaded from http://www.abcnews.go.com on January 2008

Articles	A_0	A_1	A_2	A_3	A_4	A_5	A_6
A_0	1						
A_1	0.0014	1					
A_2	0.0014	0.0017	1				
A_3	0.0082	0.0063	0.0063	1			
A_4	0.0084	0.0019	0.0019	0.0067	1		
A_5	0.0052	0.0017	0.0017	0.0064	0.0020	1	
A_6	0.0012	0.0048	0.0014	0.0059	0.0016	0.0015	1

Table 11: $Nscf$s for different pairs of RSS news articles in Table 10.

α-*cut*	Clusters
0.0011	{ A_0, A_1, A_2, A_3, A_4, A_5, A_6 }
0.0014	{ A_0, A_1, A_2, A_3, A_4, A_5}; { A_1, A_3, A_4, A_5, A_6 }
0.0017	{ A_0, A_2, A_4, A_5 }; { A_3, A_6 }; { A_1, A_2, A_4, A_5 }
0.002	{ A_0, A_2, A_3, A_5 }; { A_0, A_3, A_4, A_5 }; { A_1, A_2, A_3 }; { A_1, A_3, A_6 }
0.0023	{ A_0, A_2, A_3, A_5 }; { A_4 }; { A_1, A_2, A_3 }; { A_1, A_3, A_6 }

Table 12: Clusters of RSS news articles in Table 10, generated by using different α-*cut*s.

similarity of an article with respect to others in the same cluster.

$$D(A_i) = \frac{\sum_{k=1}^{N} max_{C_k}\{nscf(i,j)\}}{N} \tag{17}$$

where N is the number of clusters in which an article A_i appears, and max_{C_k} denotes the maximum $nscf$ value of articles A_i and A_j in cluster C_k $(1 \leq k \leq N)$ in which A_i appears.

Using the ranking function D, we determine the order in which the RSS news articles in each cluster are to be eliminated and proceed to discard the top n $(n \geq 1)$ RSS news articles. The number of articles, n, to be discarded can be manually established and depends exclusively on a given user needs. In fact, n can also be determined automatically for each user based on his/her individual preference and access patterns. By observing the number of articles within the clusters presented to a user that are actually read, this particular information can establish n, since n can be increased or decreased accordingly. In other words, without user's direct involvement, simply recording the number of RSS news articles accessed by a given user over a period of time, e.g. d (≥ 1) days, we can adjust the n value so that the number of RSS news articles presented daily to the user equal the average number of articles read by a user over d days. Hence, the percentage of RSS news articles to be discarded should be closely related to the (i) number of articles that are accessed by a given user, and (ii) the number of articles posted by each individual RSS news feed accessed by a given user on regular basis. As shown in Section 4.3, discarding 30% of the less-informative RSS news articles seems to be optimal, since when the percentage of discarded articles increases there is a chance that certain information covered by the RSS news articles are eliminated entirely, whereas when the percentage of discarded articles decreases, there is a chance that a significant number of articles that cover the same topic remain in the final set of clusters, which should be removed.

Example 3 Table 13 shows nineteen RSS news articles downloaded from http:// abc-news.go.com, http://cbsnews.com and http://usatoday.com on January 2008. The main topics covered in the articles are politics, entertainment, and international news. We first cluster the articles using α-*cut*s and establishing $\alpha = 0.002$ (computed by using Equation 16) as shown in Table 14. Prior to discarding 30% of the RSS news articles that are considered

less-informative, we establish the articles and the order in which they are eliminated using Equation 17 as shown in Table 15.

Our ranking function establishes that out of the six articles in which to be eliminated, four belong to international news (A_1, A_3, A_6 and A_7), one belongs to entertainment (A_{16}), and another belongs to politics (A_{14}), which is appropriate for this particular example, since most of the news in Table 13 cover international news and hence deleting more related articles to this topic, which further reduces the number of related articles presented, is appropriate. What is more, A_9 is not deleted according to the ranking shown in Table 15 i.e., A_9 should be eliminated before A_{16}, since if A_9 were to be eliminated, an entire story (i.e., Colombian hostages freed by Colombian rebels) would disappear. □

After detecting n different articles to be deleted, we further eliminate clusters that are subsumed by (i.e., included in) other clusters to obtain the final set of clusters. As shown in Table 14, news articles covering a particular subject are grouped together. Also, the singleton cluster with A_9 is not removed, as shown in Table 16, since A_9 is closely related to articles A_1 and A_6, which have previously been eliminated.

5 Experimental Results of our InFRSS Approach

In this section, we verify the accuracy of our InFRSS approach in detecting redundant RSS news articles using the CPM model and the quality of clusters generated by using the FCC model. Experimental results, which were generated by using different datasets, were analyzed and conclusions are drawn according to the measures.

5.1 Experimental Results of Phrase Matching

In evaluating the accuracy of using our n-gram ($1 \leq n \leq 5$) CPM approach to determine the degree of similarity of any two RSS news articles, we collected thousands of articles from different sources as partially shown in Table 17. In order to guarantee the impartiality of our experiments, the articles were randomly selected from hundreds of different news feeds, collected between July 2006 and July 2007. The chart in Figure 2 shows the variety of

Articles	Title and (Portion of the) Body
A_0	Israel Rules Out "No Options" Against Iran. Israel's prime minister has told a powerful panel of his nation's ...
A_1	Colombian Hostage Reunited With Her Son. Recently released Colombian hostage Clara Rojas was reunited ...
A_2	Bush Regaled In United Arab Emirates. President Bush got a flavor of the cosmopolitan ...
A_3	Bush Trip Deepens U.S.-Iran Propaganda War. President Bush's trip to the Mideast was to be viewed as a push for Israeli-Palestinian peace ...
A_4	Mideast Negotiators Prep For Tough Topics. Palestinian President Mahmoud Abbas said Sunday negotiating teams ...
A_5	Navy Fired Shots At Iranian Craft In Dec. The U.S. Navy said Friday that one of its ships had fired warning shots ...
A_6	Chavez Sticks Up For Colombian Guerrillas. President Hugo Chavez took the side of leftist rebels in neighboring Colombia's ...
A_7	Bush Leaves Israel, Seeks Arab Support. U.S. President George W. Bush sought Arab support for a U.S.-backed Mideast ...
A_8	Musharraf Warns U.S. Not To Trespass. Pakistan's beleaguered president Pervez Musharraf has again warned the U.S. to keep troops off ...
A_9	Colombian Rebels Free Two Hostages. Colombian rebels freed two women held hostage for more than six years ...
A_{10}	Bush "Confident" Of Mideast Peace In 2008. An optimistic President Bush has again predicted that Mideast peace ...
A_{11}	Racial Tensions Heat Up In Dem Campaign. Hillary Rodham Clinton and Barack Obama have become embroiled in racially tinged disputes ...
A_{12}	Edwards Joins Clinton, Obama Race Dispute. Democratic presidential candidate John Edwards has waded into the minefield racial dispute ...
A_{13}	Racial Tensions Roil Democratic Race. A series of comments from Sen. Hillary Rodham Clinton, her husband, and her supporters are spurring a racial backlash ...
A_{14}	Gender And Race In The Democratic Primary.CBS News' director of surveys Kathy Frankovic says New Hampshire was a fluid race ...
A_{15}	Britney In Desert, Dr. Phil Show A No-Go. The pop star was spotted at a restaurant in Palm Desert ...
A_{16}	Britney's Busy Night Out. Pop star Britney Spears had her car towed away after she left it in the street with flat tire ...
A_{17}	Buzz Briefs: Britney Spears, Madonna. Brit has car trouble again. Posh makes the worst-dressed list ...
A_{18}	Britney's Fight For Visitation Continues. Britney Spears' effort to regain visitation rights with her two small children ...

Table 13: RSS news articles downloaded from http://abcnews.go.com, http://cbsnews.com and http://usatoday.com on January 2008.

Original Clusters	Clusters After Elimination	Final Set of Clusters
$\{ A_0, A_3, A_7 \}$	$\{ A_0 \}$	
$\{ A_2, A_3, A_4, A_6, A_7, A_8, A_{10} \}$	$\{ A_2, A_4, A_8, A_{10} \}$	$\{ A_2, A_4, A_8, A_{10} \}$
$\{ A_3, A_5, A_8 \}$	$\{ A_5, A_8 \}$	$\{ A_5, A_8 \}$
$\{ A_1, A_6, A_9 \}$	$\{ A_9 \}$	$\{ A_9 \}$
$\{ A_6, A_{12} \}$	$\{ A_{12} \}$	
$\{ A_{11}, A_{12}, A_{13}, A_{14} \}$	$\{ A_{11}, A_{12}, A_{13} \}$	$\{ A_{11}, A_{12}, A_{13} \}$
$\{ A_3, A_7, A_8, A_{13} \}$	$\{ A_8, A_{13} \}$	$\{ A_8, A_{13} \}$
$\{ A_8, A_{13}, A_{14} \}$	$\{ A_8, A_{13} \}$	
$\{ A_2, A_{16} \}$	$\{ A_2 \}$	
$\{ A_{15}, A_{16}, A_{17}, A_{18} \}$	$\{ A_{15}, A_{17}, A_{18} \}$	$\{ A_{15}, A_{17}, A_{18} \}$
$\{ A_{14}, A_{17} \}$	$\{ A_{17} \}$	
$\{ A_0, A_8 \}$	$\{ A_0, A_8 \}$	$\{ A_0, A_8 \}$
$\{ A_4, A_8 \}$	$\{ A_4, A_8 \}$	$\{ A_4, A_8 \}$
$\{ A_5, A_8 \}$	$\{ A_5, A_8 \}$	$\{ A_5, A_8 \}$

Table 14: Clusters generated using $\alpha = 0.002$, computed using Equation 16.

Article	D(Article) Value	Article	D(Article) Value
A_1	0.35	A_{17}	0.16
A_{14}	0.32	A_2	0.15
A_7	0.27	A_{12}	0.15
A_3	0.27	A_0	0.15
A_6	0.25	A_8	0.14
A_9	0.25	A_{13}	0.14
A_{16}	0.25	A_5	0.14
A_{11}	0.24	A_{10}	0.13
A_4	0.23	A_{15}	0.09
A_{18}	0.23		

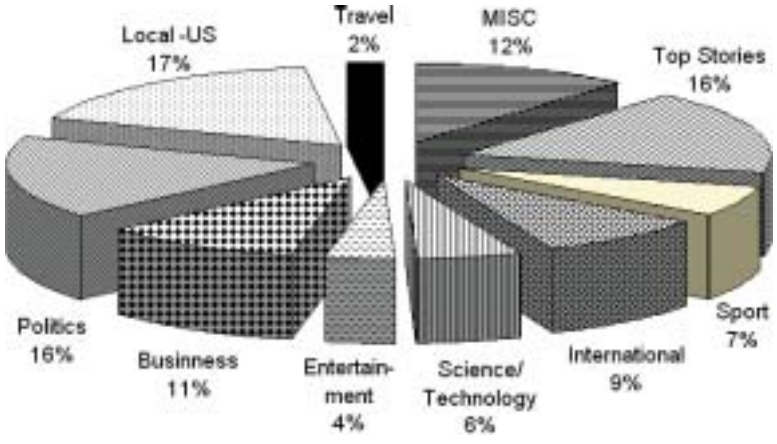
Table 15: Ranking of RSS news articles in Table 13, computed by using Equation 17.

Ranking
A_1
A_{14}
A_7
A_3
A_6
A_{16}

Table 16: The top 30% of RSS news articles that are considered *less informative*, based on the ranking shown in Table 15 and therefore are eliminated.

Figure 1: Portion of an RSS news feed file.

Figure 2: Variety of subject areas covered by the chosen RSS news feeds for the empirical study.



subject areas covered by the chosen RSS news feeds which demonstrate the suitability of our CPM model applied to news articles independent of their content.

We determined the relative degree of similarity of each pair of the 1059 news articles extracted from 200 RSS news feeds using each n-gram ($1 \leq n \leq 5$) phrase matching approach. In this empirical study, we (i) randomly selected 410 pairs, (ii) manually examined each pair to determine their relative degree of similarity, and (iii) compared the manually determined similarity with the automatically computed $nscf$s based on each one of the five n-gram phrase matching approaches. To verify which n-gram CPM is the most accurate in determining related pairs of news articles, we consider the number of *False Positives* (FPs), i.e., *unrelated* pairs with *high* $nscf$s, and *False Negatives* (FNs), i.e., *related* pairs with *low* $nscf$s, generated by using CPM on each type of n-grams as follows:

$$Accuracy = \frac{\text{T}otal\_Number\_of\_Examined\_Pairs - \text{M}isclassified\_Pairs}{\text{T}otal\_Number\_of\_Examined\_Pairs} \tag{18}$$

where $Misclassified\_Pairs$ is the sum of FPs and FNs encountered.

Figure 3(a) shows the number of *correctly* classified pairs, as well as *incorrectly* identified pairs, i.e., the sum of the FPs and FNs, on 410 pairs of randomly chosen news articles, which is a subset of the news articles listed in Table 17, whereas Figure 3(b) shows the accuracy computed by using Equation 18 for each n-gram CPM approach. Clearly, *bigrams* and *trigrams* yield the *lowest* number of *misclassified* pairs of articles, while achieve the *highest* count ($\geq 90\%$) of *correctly* detected pairs among all the n-grams. The use of *4-* and *5-grams* reduces the accuracy to as low as 60%, whereas *unigrams* has an accuracy of 86%.

When using bigrams and trigrams, the misclassified pairs occur, since if they have at least one common bigram or trigram, then their *odds* increase. Also, the degrees of similarity for unigram, bigram, and trigram are relatively higher, and thus their $nscf$s are comparatively higher, whereas the degrees of similarity generated by using 4-grams and 5-grams tend to be much lower, and thus their $nscf$s are often extremely low, which might explain their *low* accuracy in detecting similar pairs of RSS news articles. In general, the *unrelated* pairs of articles detected by using *bigrams* and *trigrams* have a $nscf$ close to the power of E-6, whereas *related* pairs have a $nscf$ above 0.1. Neither FPs nor FNs are

Sources	Number of Feeds	Number of Articles
1115.org	3	19
abcnews.go.com	10	135
adn.com	2	5
blogs.zdnet.com	1	6
boston.com	8	26
businessweek.com	8	41
cbsnews.com	8	24
chron.com	7	21
cnn.com	8	29
dailymail.co.uk	1	4
english.people.com	6	54
forbes.com	2	8
foxnews.com	10	35
guardian.co.uk	4	12
health.telegraph.co.uk	1	4
hosted.ap.org	11	39
iht.com	9	27
latimes.com	3	9
microsoftwatch.com	1	4
money.cnn.com	4	17
money.telegraph.co.uk	1	4
msnbc.com	1	6
news.bbc.co.uk	2	13
news.ft.com	9	47
news.telegraph.co.uk	3	25
news.yahoo.com	3	37
nytimes.com	10	60
online.wsj.com	6	70
politics.guardian.co.uk	1	3
portal.telegraph.	1	3
primezone.com.co.uk	2	8
prnewswire.com	8	24
seattletimes.nwsource.com	8	50
slashdot.org	1	5
sltrib.com	2	6
sportsillustrated.cnn.com	3	9
timesonline.com	3	10
today.reuters.com	12	112
usatoday.com	10	27
washingtonpost.com	2	6
wired.com	3	9
worldpress.org	2	6

Table 17: Sources of RSS news feeds, the number of feeds of each source (200 total), and the number of news articles from each source (1059 total).

desired, and in this study they contribute only 10% of the bigram (trigram) pairs, out of which close to 90% are FP pairs. In fact, FPs are less harmful in our similarity detection approach, since we do not lose many similar pairs.

Based on the empirical study, we conclude that (i) *bigram* and *trigram* outperform others in detecting similar RSS news articles. In most cases, the $nscf$s computed by using bigrams and trigrams on similar RSS news articles are higher than the ones computed by using *unigrams*. (ii) *4-grams* and *5-grams* are not reliable in determining the relevance between any two RSS news articles as explained earlier. Our empirical study further verify the claims made by [20, 22], which state that the use of bigrams and trigrams is often more effective than the use of other n-gram phrases in retrieving information.

5.2 Experimental Results of Our Clustering Approach

In this section, we describe the dataset and present the performance metric used for analyzing the effectiveness of our approach for clustering RSS news articles. We will evaluate the performance of FCC for grouping highly related RSS news articles, as well as comparing FCC with other well-known clustering methods in terms of the quality of clusters generated.

5.2.1 Dataset

The dataset we used for conducting the performance evaluation of FCC is a large set of RSS news articles, denoted $RSSds$. The news articles in $RSSds$ were downloaded from a variety of online sources such as abcnews.com, cbsnews.com, usatoday.com, cnn.com, etc., between January 2007 and January 2008, which cover a wide variety of topics (as shown in Table 18) that allow us to perform an unbiased evaluation of FCC. In addition to $RSSds$, we also used the $Reuters$-21578 (http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html) dataset. This dataset consists of 21,578 news articles grouped into 135 topics; each article has been manually labeled with one or more topics. $Reuters$-21578 is a popular benchmark dataset, appropriate for evaluating the performance of any text retrieval, text categorization, or clustering approach [14].

Sources	Number of RSS News Articles	Topics
abcnwes.go.com	496	Sports, Entertaintment, Bussines, Technology, Health, International, Top stories, Science
blogs.zdnet.com	37	Technology
boston.com	48	Top stories, Local news, International, Technology
cbsnews.com	96	Entertaintment, Politics, International, Local News, Sports, Health, Science
cnn.com	26	Local news, Science
english.people.com	287	Sports, Entertaintment, Top Stories
forbes.com	32	Bussiness
foxnews.com	35	Local news, Top stories, Politics, Entertaintment
health.telegraph.com	21	Health
hosted.ap.org	77	Entertaintment, Politics, International, Top Stories, Local News, Bussiness
iht.com	225	Bussiness, Health, Science, Politics, Interantional, Top stories
latimes.com	27	Top stories
money.cnn.com	28	Bussiness
msnbc.msn.com	12	Local News
news.bbc.com.uk	34	Local News, International
news.ft.com	63	International, Health, Politics, Top Stories
nytimes.com	194	International, Politics, Top Stories, Sports, Entertaintment, Bussiness
online.wsj.com	48	International, Bussiness, Technology, Weather
online.yahoo.com	14	Travel, Sports
politics.guardian.com.uk	12	Politics
prnewswire.com	374	Entertaintment, Health, International, Politics, Bussiness
seattletimes.nwsource.com	71	Politics, Enterteinment, Sports, Travel, Weather, Top stories
sltrib.com	28	Top Stories, Local News
sportsillustrated.com	19	Sports
today.reuters.com	91	Top Stories, Interantional, Politics, Entertaintment, Technology
travel.telegraph.com	17	Travel
usatoday.com	279	Top Stories, International, Politics, Entertaintment, Technology, Sports, Bussiness
washingtonpost.com	37	Top Stories, Sports, Local News
worldpress.org	19	International
Total	2,747	Number of Distinct Topics: 12

Table 18: Sources and their corresponding numbers of RSS news articles in the *RSSds* dataset downloaded between January 2007 and January 2008, which cover various topics for evaluating the performance of FCC.

5.2.2 Performance Measure

We evaluate the performance of FCC on the *RSSds* and *Reuters*-21578 datasets using the mutual information (MI) metric [31], which determines how similar or independent any two sets of clusters C and C' are. C and C' are created using the documents in a given set of documents D, and the mutual information (value) of C and C', denoted $MI(C, C')$, is defined as

$$MI(C, C') = \sum_{c_i \in C, c_j^{'} \in C'} p(c_i, c_j^{'}) \times log_2 \frac{p(c_i, c_j^{'})}{p(c_i) \times p(c_j^{'})} \tag{19}$$

where $p(c_i)$ ($p(c_j^{'})$, respectively) is the *probability* that a document d randomly selected from D belongs to cluster c_i ($c_j^{'}$, respectively) and $p(c_i, c_j^{'})$ is the (*joint*) *probability* that d is in both c_i and $c_j^{'}$, which are formally defined as

$$p(c_i) = \frac{n(c_i)}{N}, \qquad p(c_j^{'}) = \frac{n(c_j^{'})}{N}, \qquad \text{and} \qquad p(c_i, c_j^{'}) = \frac{n(c_i, c_j^{'})}{N} \tag{20}$$

where $n(c_i)$ ($n(c_j^{'})$, respectively) is the *number* of documents in c_i ($c_j^{'}$, respectively), $n(c_i, c_j^{'})$ is the *number* of common documents appeared in both c_i and $c_j^{'}$, and N is the *total number* of documents in D. In computing the mutual information, $log_2 \frac{p(c_i, c_j^{'})}{p(c_i) \times p(c_j^{'})}$ indicates how similar the two clusters c_i ($\in C$) and $c_j^{'}$ ($\in C'$) are, and $log_2 \frac{p(c_i, c_j^{'})}{p(c_i) \times p(c_j^{'})} = 0$ when c and $c^{'}$ are *independent*, i.e., $p(c_i) \times p(c_j^{'}) = p(c_i, c_j^{'})$, which implies that c_i and $c_j^{'}$ do not have any news articles in common.

As stated in [31], $MI(C, C^{'})$ is bounded between zero and $max(H(C),\ H(C'))$ inclusively, where $H(C)$ and $H(C^{'})$ are the *entropies* of C and $C^{'}$, i.e., $H(C)$ and $H(C^{'})$ represent the *purity* of the set of clusters C and $C^{'}$, respectively, and are defined as follows:

$$H(C) = -\sum_{i=1}^{m} p(c_i) log_2 p(c_i) \text{ , and } H(C') = -\sum_{j=1}^{n} p(c_j^{'}) log_2 p(c_j^{'}) \tag{21}$$

where m (n, respectively) is the number of clusters in C ($C^{'}$, respectively) and $MI(C,\ C^{'})$ reaches its maximum value when C and $C^{'}$ are *identical*, whereas $MI(C,\ C^{'})$ is closer to *zero* when C and $C^{'}$ are more *independent*, i.e., they have very few clusters in common.

We normalize MI so that $0 \leq MI \leq 1$ which is defined as follows:

$$MI^*(C, C^{'}) = \frac{MI(C, C^{'})}{max(H(C), H(C'))} \tag{22}$$

As stated in [36], MI^* is impartial to the number of clusters involved in computing the entropy of the two sets of clusters. $MI^*(C, C')$ is a probabilistic measure that determines how much information (news articles in our case) are shared by C and C'; $MI^*(C, C') = 0$ (or close to 0) indicates random *partitioning* on C and C', i.e., the sets of clusters in C and C' do not share any information, whereas $MI^*(C, C') = 1$ implies that the two sets of clusters are *identical*, i.e., the number of clusters, as well as the articles in each cluster, are perfectly matched [36].

MI^* is used for evaluating the performance of FCC (as well as other clustering approaches as described in Section 5.2.4). We compute MI^* in (i) the set of clusters manually created from the set of news articles in $RSSds$ according to the topic(s) they belong to, generating C, and (ii) the clusters of news articles in $RSSds$ using FCC, generating the set of clusters C'. $MI^*(C, C')$ thus yields the degree of similarity among clusters in C and C'.

We perform the same MI^* measure on the *Reuters*-21578 dataset for comparing the performance of FCC with other well-known, existing clustering approaches, except that the clusters in C as provided by Reuters-21578 are predefined in the dataset.

5.2.3 Evaluating Our FCC Model

To determine the overall MI^* value of FCC, we used 67 test cases of different sizes, and each test case consists of randomly selected RSS news articles in the $RSSds$ dataset (as shown in Table 18). We manually labeled each of the RSS news articles in the test cases according to one or more topics to which it belongs. Although a subset of articles in $RSSds$ might share a general topic, some are more closely related than others in terms of the covered stories. Hence, there are news articles considered to be less-informative in the subset, i.e. articles that in a greater or lesser degree cover the same story and can be eliminated.

For the news articles in each test case, we applied FCC to create clusters using the α value determined by Equation 16, and discarded 30% of the less-informative news articles, where 30% is an ideal ratio as discussed in Section 4.3. The number of clusters in C' of each test case T was determined by the α value computed by Equation 16, and news articles in T were manually examined and assigned to various clusters in C according to their actual

contents. Hereafter, we computed the normalized mutual information value of C and C' created from the news articles in T.

Let's consider the following three RSS news articles downloaded from http://seattle times.nwsource.com in January 2008, and the URL link is specified in Table 18.

A_1. Basketball Roundup: Gordon helps Indiana rally to down Illinois. Freshman standout Eric Gordon scored 11 of his 17 ...

A_2. NBA Wire Notes: Teammates extend Noah's suspension. Chicago Bulls center Joakim Noah, the face of Florida's back-to-back basketball national championship teams ...

A_3. Australian Open Tennis: Jankovic outlasts Paszek. Third-seeded Jelena Jankovic of Serbia saved three match points ...

A_1, A_2, and A_3 share a common topic, i.e., sports. Since A_1 and A_2 are basketball sports news and A_3 is a tennis sports news article, which is less closely related than A_1 and A_2, A_1 and A_2 should be assigned to the same clusters, whereas A_3 to a separate one.

Using *RSSds* we evaluated the performance of FCC such that not only the "general" topic of RSS news articles is considered, but also the actual content of the different RSS news articles during the clustering process. Table 19 shows (a subset of) the test cases used for calculating the average MI^* value of the test cases in *RSSds*, which is 0.67. We also evaluated the performance of FCC using the *Reuters*-21578 corpora, but only considering single-topic articles to facilitate the evaluations, obtaining an average mutual information value of 0.61 (see detail in Section 5.2.4).

The average MI^* value, 0.67, obtained by using FCC in *RSSds* is a promising result (see detailed discussion in Section 5.2.5). What is more, by using the *RSSds* dataset, we are able to assess the performance of FCC not only on single-topic news articles, but most importantly for multiple-topic articles, i.e., articles that belong to more than one topic, as (some of the) articles in *RSSds*, which demonstrates the applicability of FCC, since an RSS news article often covers more than one topic.

Test Case	Number of RSS News Articles	Number of Discarded RSS News Articles	Number of Topics	Mutual Information (MI^*)	Sources
TC_1	5	2	1	0.84	abcnews.go.com
TC_2	10	3	1	0.82	prnewswire.com
TC_3	16	5	2	0.78	cbsnews.com, cnn.com
TC_4	28	8	3	0.72	hosted.ap.com, iht.com
TC_5	30	9	3	0.73	online.wsj.com, abcnews.com
TC_6	37	11	4	0.39	money.cnn.com, usatoday.com
TC_7	43	13	5	0.61	worldpress.org, travel.telegraph.com, today.reuters.com
TC_8	45	14	6	0.54	latimes.com, forbes.com, boston.com, english.people.com
TC_9	57	17	6	0.63	news.ft.com, online.yahoo.com, blogs.zdnet.com, sltrib.com
TC_{10}	84	25	7	0.66	politics.guardian.com.uk, seattlenwsource.com, msnbc.com, nytimes.com, cbsnews.com
TC_{11}	119	36	8	0.57	news.bbc.com.uk, abcnews.org.com, washingtonpost.com, travel.telegraph.com, prnewswire.com, hosted.ap.org
TC_{12}	150	45	8	0.49	bonston.com, iht.com, nytimes.com, sportsillustrated.com, sltrib.com, money.cnn.com
...	...	...	...	...	...
Average	87	24	5	0.67	Nnumber of Sources in Each Test Case: 4

Table 19: (A portion of the) Test cases used for evaluating the performance of FCC in terms of the mutual information measure.

5.2.4 Comparing the Performance of FCC with Others

We further assess the performance of FCC by comparing the MI^* values obtained by several well-known clustering techniques (as detailed in [32]) with the MI^* value obtained by FCC, using the *Reuters*-21578 dataset.

In [32] several clustering techniques are presented: (i) the k-means algorithm, which looks for the appropriate k centers (i.e., news documents) to group data according to a predetermined k number of clusters, (ii) the Spherical k-means [9], which models a document collection as a bipartite graph that is used by a spectral algorithm for clustering documents and words simultaneously, (iii) the Gaussian Mixture Model [17], which establishes the most representative features of each cluster and uses them to refine the document clusters based on a majority voting scheme, (iv) an unsupervised version of the well-known Naive Bayes probabilistic model for document clustering [2], (v) the Spectral clustering algorithm based on average association criterion [35], and (vi) the Spectral clustering algorithm based on normalized cut criterion [27]. In both (v) and (vi) a given set of documents is represented as a graph in which each node represents a document, each edge denotes an association between any two documents, and a weight is assigned to each edge to reflect the similarity of the connected documents. As stated in [31], the clusters generated by using the approaches of (v) and (vi) on the *Reuters*-21578 news articles are determined by finding the graph's most suitable cut based on a predefined criterion.

We achieve an unbiased comparison by following the evaluation procedure described in [32], which considers only the news articles in the *Reuters*-21758 dataset that belong to a unique topic and discards topics that had less than five news articles. As a result, 9,494 news articles and 51 topics from the *Reuters*-21758 dataset were used for the experimental and comparison purpose, which are the articles that belonged to a unique topic, including oil, grain, acquisitions, money, trade, cotton, etc. By adopting this evaluation procedure, we are able to objectively determine the quality of FCC compared with other well-known clustering methods in terms of their MI^* measures.

As stated in [32], in order to compute the MI* value, different numbers of clusters k ($2 \leq k \leq 10$) should be considered. For each k, several test runs were performed on

randomly chosen single-topic news articles in the *Reuters*-21758 dataset, and the MI^* value for each k was obtained by averaging the MI^* value obtained in each test run. The overall average MI^* values obtained by a number of existing clustering approaches was reported in [32] and as shown in Figure 4. Clearly, as shown in Figure 4, FCC outperforms the mentioned clustering approaches by an average of 9% in terms of MI*, which shows the quality of clustered documents generated by using FCC.

As stated in [5], the MI^* is a semantic measure that quantifies the statistical information shared between two distributions, i.e., it indicates the shared information between a pair of clusters. Hence, by obtaining an MI^* value higher than the MI^* value generated by other well-known clustering approaches, we can claim that FCC generates higher quality clusters, i.e., more generated clusters are similar to the original set of clusters used for comparison.

5.2.5 Observations

We observe that FCC outperforms known clustering approaches due to the flexibility of the fuzzy compatibility relation that we develop for clustering RSS news articles, since these news articles are grouped in less-restrictive clusters according to their actual content similarity as opposed to the general topic they might belong to.

It is worth to mention that by discarding the less-informative news articles, we provide users with clusters that include the entirety of topics covered in the original set of news articles, while reducing a number of news articles without sacrificing the quality of the clusters. What is more, as previously stated, the user always has the choice of not discarding less-informative news articles (i.e., only redundant and subsumed news articles are eliminated), if desired.

As shown in Section 5.2.3, we achieve a higher MI^* value in using the *RSSds* dataset, which contains multi-topics news articles, than in using the *Reuters*-21578 dataset. The original set of clusters C on *RSSds* includes news articles in multiple clusters, which is the approach we adopt we using FCC, C', i.e., FCC assigns a given article to multiple clusters if necessary, and as a result, the two sets of clusters are more alike, i.e., share more articles in common. However, the original set of clusters C on the *Reuters*-21578 dataset is much

more restrictive, since each article is assigned to only one cluster, whereas the clusters generated by using FCC C', which allows articles to be assigned to more than one clusters, are not as similar to the clusters in C, which yields the lower MI^* value C and C', i.e., 0.61 versus 0.67, the MI^* value obtained by using $RSSds$ dataset.

Furthermore, FCC works with little overhead, since only the combined similarity values, which are computed using the (phrase) correlation factors, are considered during the clustering process, as opposed to some of the existing clustering techniques described in [32] such as Naive Bayes, K-means, and Spherical K-means, which depend on the initialization of the clusters and hence the performance of these approaches is significantly affected by their initial partition. Furthermore, to the best of our knowledge, there are not any benchmark measures available that we could use for evaluating and comparing the performance of FCC in dealing with multiple-topics news articles.

6 Conclusions

We have presented *InFRSS* that combines the correlation phrase (CPM) model with the fuzzy compatibility clustering (FCC) model for finding and clustering informative RSS news articles. In the CPM model, we have considered n-gram ($1 \leq n \leq 5$) phrase matching and verified that bigrams and trigrams outperform other n-grams in detecting similar RSS news articles. We have also verified the accuracy of our CPM approach by analyzing hundreds of pairs of randomly selected RSS news articles from multiple sources and concluded that applying bigram and trigram phrase matching is highly accurate ($\geq 90\%$), which requires little overhead (using predefined correlation factors) in finding related articles. Our CPM can also be used for (i) detecting (similar) junk emails and spam Web pages, (ii) clustering (Web) documents with similar content, and (iii) discovering plagiarism, which form the core future work for our CPM model. We conducted various experiments to measure the accuracy in using each one of the unigram, bigram, trigram, 4-gram, and 5-gram phrase correlation factors to detect similar RSS news articles. Our finding is agreed upon by [24] who show that better results are obtained when applying bigrams and trigrams to retrieve information, whereas [22] state that decomposing long phrases to bigrams is more effective

in retrieving information. Moreover, [20] claim that the use of short queries (2 or 3 terms, i.e., bigram or trigram) has a more positive effect when retrieving information from the Web than the use of longer queries, which can have less, or even negative, effect. We verify the correctness of our CPM model by comparing the detected (non-)similar news articles using various n-gram ($1 \leq n \leq 5$) phrase correlation factors discussed in Section 3.2.

In the FCC model, well-known techniques such as the fuzzy set logic and the α-cuts are combined to generate high quality clusters of RSS news articles. We have defined the mutual information metric to show the effectiveness of our FCC model in clustering RSS news articles and compared the performance of FCC with other well-known clustering techniques. We have verified that on the average our FSS model outperforms existing, well-known clustering techniques by 9%.

We have observed that *InFRSS* performs better when making use of bigrams for computing the degrees of similarity between RSS news articles. We believe that constructing a bigram-correlation matrix using the Wikipedia document collection could further enhance the effectiveness of *InFRSS* in detecting and clustering RSS news articles. Our *InFRSS* could also be used for detecting and clustering other kinds of document collections in different subject areas, such as medical documents, computer science articles, laws, etc.

References

[1] Amer-Yahia, S., Fernandez, M., Srivastava, D., Xu, Y.: PIX: Exact and Approximate Phrase Matching in XML. ACM SIGMOD. (2003) 664-667

[2] Baker, L., McCallum, A.: Distributional Clustering of Words for Text Classification. ACM SIGIR. (1998) 96-103

[3] Cilibrasi, R., Vitanyi, P.: Automatic Meaning Discovery Using Google. www.cwi.nl/paulv/papers/amdug.pdf (2004)

[4] Cordonez, C.: Clustering Binary Data Streams with K-Means. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. (2003) 10-17

[5] Cover, T., Thomas, J.: Elements of Information Theory. (1991) Wiley

[6] Crabtree, D., Gao, X., Andreae, P.: Improving Web Clustering by Cluster Selection. IEEE/WIC/ACM International Conference on Web Intelligence. (2005) pp. 172-178

[7] Crabtree, D., Andreae, P., Gao., X.: Query Directed Web Page Clustering. IEEE/WIC/ACM International Conference on Web Intelligence. (2006) pp. 202-210

[8] de Moura, E., Navarro, G., Ziviani, N., Baeza-Yates, R.: Fast and Flexible Word Searching on Compressed Text. ACM TOIS, vol. 18(2). (2000) 113-139

[9] Dhillon, I.S.: Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. ACM SIGKDD. (2001) 269-274

[10] Hammouda, K., Kamel, M.: Efficient Phrase-Based Document Indexing for Web Document Clustering. IEEE TKDE(16):10. (2004) 1279-1296

[11] Haveliwala, T., Gionis, A., Klein, D., Indyk, P.: Evaluating Strategies for Similarity Search on the Web. World Wide Web Conference. (2002) 432-442

[12] Judea, P.: Probabilistic Reasoning in the Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)

[13] Klir, G.K., St. Clair, U., Yuan, B.: Fuzzy Set Theory, Foundations and Applications. Prentice Hall PTR. (1997)

[14] Lewis, D., Yang, Y., Rose, T., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research(5). (2004) 361-397

[15] Li, X., Yan, J., Deng, Z., Ji, L., Fan, W., Zhang, B., Chen, Z.: A Novel Clustering-Based RSS Aggregator. World Wide Web (2007) 1309-1310

[16] Li, Y., Chung, S.: Document Clustering Based on Frequent Word Sequences. ACM CIKM. (2005) 293-294

[17] Liu, X., Gong, Y., Xu, W., Zhu. S.: Document Clustering with Cluster Refinement and Model Selection Capabilities. ACM SIGIR. (2002) 191-198

[18] Luger, G.: Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 5th Edition. Addison Wesley (2005)

[19] Luo, G., Tang, C., Tian, Y.: Answering Relationship Queries on the Web. World Wide Web Conference. (2007) 561-570

[20] Mishne, G., de Rijke, M.: Boosting Web Retrieval through Query Operations. European Conference on Information Retrieval. (2005) 502-516

[21] Nallapati, R., Feng, A., Peng, F., Allan, J.: Event Threading within News Topics. ACM CIKM. (2004) 446-453

[22] Narita, M., Ogawa Y.: The Use of Phrases from Query Texts in Information Retrieval. ACM SIGIR. (2000) 318-320

[23] Ogawa, Y., Morita, T., Kobayashi, K.: A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. Fuzzy Sets and Systems(39). (1991) 163-179

[24] Pfeifer, U., Poersch, T., Fuhr., N.: Retrieval Effectiveness of Proper Name Search Methods. Journal of Information Processing & Management(32):6. (1996) 667-679

[25] Porter, M.: An Algorithm for Suffix Stripping. Program(14):3. (1980) 130-137

[26] Sahoo, N., Callan, J., Krishnan, R., Duncan, G., Padman, R.: Incremental Hierarchical Clustering of Text Documents. ACM CIKM. (2006) 357-366

[27] Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence(22):8. (2000) 888-905

[28] Toud, S.: Creating a Custom Metrics Tool. MSDN Magazine, http://msdn.microsoft.com/msdnmag/issues/05/04/EndBracket/. April 2005

[29] Tzong-Han, T., Chia-Wei, W.: Enhance Genomic IR with Term Variation and Expansion: Experience of the IASL Group. Text Retrieval Conference. (2005)

[30] Wilbur, W., Kim, W.: Flexible Phrase Based Query Handling Algorithms. American Society for Information Science and Technology. (2001) 438-449

[31] Xu, W., Liu, X., Gong, Y.: News Article Clustering Based on Non-Negative Matrix Factorization. ACM SIGIR. (2003) 267-273

[32] Xu, W., Gong, Y.: News Article Clustering by Concept Factorization. ACM SIGIR. (2004) 202–209

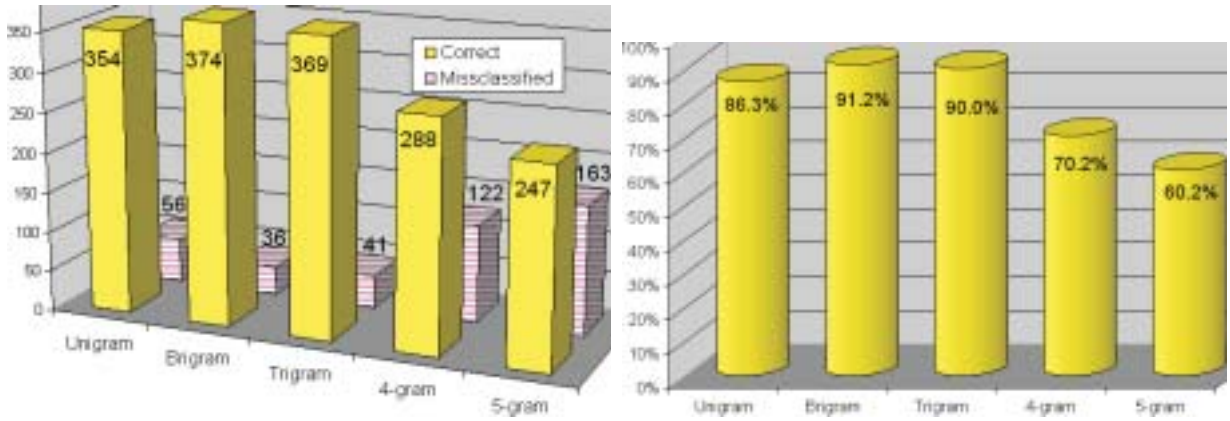
[33] Yerra, R., Ng, Y.-K.: Detecting Similar HTML Documents Using a Fuzzy Set Information Retrieval Approach. IEEE International Conference on Granular Computing (GrC'05). (2005) 693-699

[34] Zadeh, L.A.: Similarity Relations and Fuzzy Orderings. Information Sciences(3). (1970) 177-200

[35] Zha, H., Ding, C., Gu, M., He, X., Simon. H.: Spectral Relaxation for K-Means Clustering. Advances in Neural Information Processing Systems(14) (2001) 1057–1064

[36] Zhong, S., Ghosh, J.: A Comparative Study of Generative Models for Document Clustering. Knowledge and Information Systems(8):3. (2005) 374–384

Figure 3: Classified 410 pairs of news articles and their accuracy.



(a) (In)Correctly classified, related articles using our n-gram CPM.

(b) Accuracy of using our n-gram CPM to detect related articles.

Figure 4: Mutual information values computed by using different clustering algorithms in [24] as well as ours on the *Reuters*-21578 dataset for illustrating the merit of our clustering approach.