Synthesizing correlated RSS news articles based on a fuzzy equivalence relation

Article *in* International Journal of Web Information Systems · April 2009

DOI: 10.1108/17440080910947321 · Source: DBLP

2 authors:

Maria Soledad Pera
Boise State University
110 PUBLICATIONS **804** CITATIONS
SEE PROFILE

Yiu-kai Ng
Brigham Young University - Provo Main Campus
97 PUBLICATIONS **1,566** CITATIONS
SEE PROFILE

Synthesizing Correlated RSS News Articles Based on a Fuzzy Equivalence Relation

Maria Soledad Pera
Yiu-Kai Ng[1]

Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
Email: {mpera@cs.byu.edu, ng@cs.byu.edu}

Abstract

Tens of thousands of news articles are posted on-line each day, covering topics from politics to science to current events. To better cope with this overwhelming volume of information, RSS (news) feeds are used to categorize newly posted articles. Nonetheless, most RSS users must filter through many articles within the same or different RSS feeds to locate articles pertaining to their particular interests. Due to the large number of news articles in individual RSS feeds, there is a need for further organizing articles to aid users in locating non-redundant, informative, and related articles of interest quickly. In this paper, we present a novel approach which uses the word-correlation factors in a fuzzy set information retrieval model to (i) filter out redundant news articles from RSS feeds, (ii) shed less-informative articles from the non-redundant ones, and (iii) cluster the remaining informative articles according to the fuzzy equivalence classes on the news articles. Our clustering approach requires little overhead or computational costs, and experimental results have shown that it outperforms other existing, well-known clustering approaches. The clustering approach as proposed in this paper applies only to RSS news articles; however, it can be extended to other application domains.

Keywords: Information retrieval, clustering, word similarity, Fuzzy equivalence class, RSS news articles

1 Introduction

These days more information is being transmitted through the Internet on a daily basis than any individual could read or process in an entire lifetime. Besides entire libraries of news articles which have been digitally archived and are publicly available, there is a constant influx of new articles. Sorting through this deluge of new articles posted on any Weblogs, news, and other Web sites to find particular items of interest manually and regularly not only is extremely time-consuming and labor-intensive, but it is also an impractical task. In solving this problem, recent efforts have been made on developing RSS[2] (news) feeds, which are XML documents in which new and/or frequently updated news articles are posted, that allow users to subscribe.

Although RSS feeds do save users considerable amounts of search time by classifying newly posted articles according to their subject areas, users are still required to sort through the large number of articles posted within RSS files (of which a single RSS user may subscribe to many) in order to locate articles pertaining to their particular interests. Added to this problem is the presence of *redundant* articles, i.e., articles which contain information already included in other articles in different (or even the same) RSS feeds. In addition, *less-informative* articles, which are not entirely redundant, but which

include significant overlapped information appeared in other articles, are common. For example, a brief, breaking news article would become redundant or less-informative when its updated version is posted, with additional, detailed information on the same event. All of these increase the volume of articles which users must sort through in order to find useful information. Furthermore, there is a need for grouping newly posted articles by *content*, which aids users in seeking related information instantly. In this paper, we present a novel algorithm that employs the fuzzy set information retrieval (IR) model to cluster non-redundant, informative, and related RSS news articles.

While other existing clustering techniques (see details in Section 2) perform well under certain conditions, the constraints imposed on these techniques are often restrictive, which reduce their applicabilities. Some require user feedback, whereas others need a very large data set for analysis before effective clustering may be performed. Our redundant/less-informative article detection and relevant clustering technique has the advantage of (i) being context free, i.e., no preprocessing or user input is required and (ii) computationally effective. This, of course, is an enormous benefit for implementation in real-world systems.

The remainder of this paper is organized as follows. In Section 2 we discuss other works performed which pertain to either detecting replicated information or clustering (news articles). In Section 3 we describe the fuzzy set IR model, which is used as a basis for our informative-document detection approach. In Section 4 we introduce the fuzzy equivalence relation we use for clustering related news articles. In Section 5 we provide the complexity analysis of the proposed clustering method, and in Section 6 we present the experimental results generated by using our clustering approach and a case study to demonstrate the merit of the approach. In Section 7 we address the contribution of our work as presented in the paper, and in Section 8 we give our concluding remarks.

2 Related Work

Among the huge volume of related work in document/text/data analysis, we narrow our discussion on representative work closely related to detecting and/or clustering similar contents (topics) in either RSS news articles or Web documents.

Yang, et al. (Yang, 1998) state that the objective of topic detection and tracking (TDT) is to identify and label stories in several continuous news streams that pertain to new or previously unidentified events to find those that track (or discuss) the same event specified by a user. Bun, et al. (Bun, 2002) suggest using a $TF\text{-}IDF$ algorithm to recognize terms that explain the main topics in a weekly news archive and then cluster the sentences in the news articles with higher average weight according to the topics using a sentence vector. This approach, however, does not deal with the problem of information overload that should be minimized. Broder, et al. (Broder, 1997) introduce a *shingling* approach for detecting similar Web documents. The similarity detection method first represents the content of each document D by the subsequences of words in D, which are called *shingle*. Hereafter, the sets of shingles of size k belonged to any two documents are compared to establish the overlap of the two for clustering purpose.

Nallapati, et al. (Nallapati, 2004) capture the structure of on-line news events that make up different topics and the dependencies among them (i.e., event threading) through different event models. Even though the use of *cosine similarity* and *time-stamps* of news stories in (Nallapati, 2004) produces fairly good results when the events are provided, the performance deteriorates rapidly if the system has to discover the events itself. Khmelev and Teahan (Khmelev, 2003) utilize the R-measure,

which is the normalized sum of the lengths of all *word suffixes* of the text repeated in documents to detect duplicates and plagiarism. Rather than using word suffixes to detect near-duplicate documents, Yang and Callan (Yang, 2006) define *instance-level constraints* that determine to which class a document belongs to identify duplicate and near-duplicate documents. In yet another clustering approach, Li and Chung (Li, 2005) incorporate the user's prior knowledge, which indicate pairs of documents belonged to the same cluster, to obtain the desired cluster structures or to construct accurate clusters. This technique enables users to control the clustering process based on the *prior knowledge* specific to the target data set, which is, however, a constraint.

An incremental hierarchical text document clustering approach for organizing documents from various on-line sources is presented in (Sahoo, 2006). The approach depends on the frequency of occurrence and the contents of the words within documents, which is another *term-frequency* and *word-matching* approach, to determine the topic of a document for clustering. Li, et al. (Li, 2007), who also consider RSS news articles, allow the user to find articles grouped by similar topics. In (Li, 2007), the k- *nearest neighbor* algorithm locates the k nearest stories for each new story S so that the cosine similarity in the 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.

Ordonez (Ordonez, 2003) use three different variations of the k-mean algorithm to find higher quality solutions in less time for clustering binary data streams. Even though the results of the incremental k-mean are good, dependence in initialization, sensitivity to outliers, and skewed distributions could affect the performance of the algorithms. On the contrarily, Cheng, et al. (Cheng, 2006) develop a *divide-and-merge* clustering methodology that combines top-down (divide) and bottom-up (merge) algorithms, which creates a tree structure T whose leaves are documents, and each cluster is a sub-tree rooted at a node of T, whereas Wang and Kitsuregawa (Wang, 2002) describe a clustering algorithm that uses the contents of, as well as linked, Web pages for grouping relevant Web pages into semantically meaningful clusters. In harmony with our clustering approach, the authors of (Wang, 2002) rely on the content of a document to produce clusters of highly related documents. However, the clustering algorithm in (Wang, 2002) has only proved successful for documents containing a single topic, which is a restriction, since a document may cover more than one topic, as in an RSS news article.

Banerjee et al. (Banerjee, 2007) assert that grouping short texts, i.e., texts with one or two sentences, such as the short summaries of RSS news articles, is a difficult task, since there is insufficient content to identify the general topic to which they belong. The authors propose using articles in Wikipedia to enhance the analysis of the content embedded in the abbreviated RSS news articles in RSS feeds, which should facilitate and improve the content-based clustering process of RSS news articles. This approach represents RSS news articles, as well as Wikipedia articles, as weighted term vectors, and for each RSS news article N, the set of highly relevant Wikipedia articles W are retrieved and the words from both (the tiles in) W and N are combined. In other words, given an RSS news article N, instead of simply considering keywords within N, Banerjee et al. use *Wikipedia concepts*, which are words in titles of the top Wikipedia articles retrieved with respect to the words in N, to increase the accuracy of clustering RSS news articles with the same topic. Unlike our clustering approach, this method invokes an extra step, i.e., identifying Wikipedia concepts, in grouping articles, which introduces additional overhead and processing time.

The authors of (Phan, 2008) also address the problem that arises in classifying short texts such as RSS news articles, product reviews, blogs, chat messages, etc. Phan et al. propose using *external*,

large-scale collections, known as ``universal datasets," as well as a subset of previously labeled data, for training a classifier. In extracting topics from the universal datasets, Phan et al. use Latent Dirichlet Allocation (Blei, 2003) and probabilistic latent semantic analysis (Hofmann, 1999), which are known topic analysis algorithms, and Maximum Entropy (Berger, 1996), which is commonly used in natural language processing for classification tasks. Phan et al. claim that short texts ``do not provide enough word co-occurrence or shared context for a good similarity measure." Contrarily, we rely solely on the content, i.e., words within the abbreviated RSS news articles, for clustering news articles without using any external data or requiring any additional training.

Tang, et al. (Tang, 2007) introduce a clustering approach on high-dimensionality data, such as collections of news articles or TREC datasets, which groups the data using a semi-supervised clustering algorithm. As a preprocessing step, a collection of instances and a set of *must-link* and *cannot-link* constraints[3] are given, and the clustering method reduces the number of instances to be further considered by using the cannot-link constraints as dividers and creating a representative instance for each sub-group. Hereafter, based on the representative instances, the spherical *k-means* algorithm is applied to generate the final set of k disjoint clusters. This clustering approach, however, requires previous learning for the clustering algorithm to become effective, and the generated number of clusters is predefined, i.e., k, which is a constraint, since some RSS news articles may not belong to anyone of the predefined k clusters. More important, an incorrect choice of k could yield low-quality clusters.

A method based on the vector space model for clustering and summarizing RSS news articles is presented in (Takeda, 2007), which represents each RSS news article as a term vector such that each term is assigned a weight determined by using the inverse document frequency. Hereafter, the degrees of similarity among vectors are computed using the cosine similarity measure to decide to which topic each RSS news article should be assigned. Since the authors of (Takeda, 2007) only allows RSS news articles to be classified into eight possible categories, i.e., national, politics, international, business, technology, sports, entertainment, and science, which as claimed by the authors are the most commonly-used categories in source news sites, their clustering method cannot be generalized.

Pon, et al. (Pon, 2007) present a method for dealing with the problem of information overload. Using user profiles, they filter RSS news articles based on the user's preference. In this method, both RSS news articles and user profiles are represented as vectors of weighted terms. By computing the cosine similarity between the vector representations, Pon, et al. match news articles according to user's interests, which are later clustered according to the topic the articles share. As suggested by the authors, further experiments with a larger number of users should be conducted before the effectiveness of their approach can be established.

A filtering and duplicate elimination method on RSS news articles is introduced in (Gruhl, 2006). This method relies on user's feedback to establish which RSS news articles are relevant to a particular user's information need. Each user is required to provide information about his/her preference (i) using a simple query interface to begin with and (ii) rating RSS news articles as ``good," ``bad," or ``seen it." Using these information, it is possible to identify the set of relevant RSS news articles with respect to the user's preference, and each one is assigned a *fingerprint*, which is later compared with the fingerprints of RSS news articles previously shown to the user so that only unseen RSS news articles are retrieved for the users, i.e., duplicated news articles are eliminated. What is more, if a date is associated with an RSS news article, the authors of (Gruhl, 2006) can identify the newer, i.e., updated, version of a particular RSS news article and remove outdated content. Although this approach is effective, it requires constant

user participation to determine which RSS news articles are relevant, which is labor-intensive and unintuitive.

Samper, et al. (Samper, 2008) develop another method for filtering RSS news articles according to user's preference. The approach in (Samper, 2008) traces the user's access history on RSS news articles to generate a user profile that represents the user's preference without requiring any user's feedback. By computing the cosine similarity between a user's profile and the headlines of new RSS news articles, Samper, et al. predict and discard the articles that are unlikely matched the user's interests. The proposed method, however, does not address the problem of existing duplicate or near-duplicate RSS news articles, which are abundant on the Web.

3 The Fuzzy Set IR Model and Redundant Articles Detection

Detecting non-redundant and informative RSS news articles is a challenging task, since RSS news feeds are *dynamic* in nature. RSS users subscribe to Web sites that typically add the content of news articles in the machine-readable, XML-formatted file regularly and rapidly. Two of the essential elements in an RSS file are the *title* and *description* of an *item* (i.e., a news article), since the former contains the headline of the article (story), whereas the latter often contains the first couple sentences of the article, and several items can appear in the same RSS file. We concatenate the title and description of each item and treat them as the *content descriptor* of the corresponding article and use a selective clustering approach based on a fuzzy equivalence relation to cluster the articles that possess (majority of the) information which is not included in other articles from either the same or different RSS feeds. This can be done by determining the degrees of similarity of any two articles using the lists of keywords[4] in their respective content descriptors. The degrees of similarity can be computed by using the *correlation factors* among different keywords in the fuzzy set IR model, which are predefined by using a set of more than 930,000 Wikipedia (http://wikipedia.org/) documents to determine the (i) *frequency* of co-occurrence and relative *distance*, i.e., $c_{i,j}$, (ii) normalized value, i.e., $nc_{i,j}$, and (iii) *keyword correlation factor*, i.e., $cf_{i,j}$, of each pair of keywords w_i and w_j as

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

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

$$cf_{i,j} = \frac{\sum_{m=1}^{k} nc_{i,j}^{m}}{k} \qquad (1)$$

where $d(x,y)$ = $|Position(x)$ - $Position(y)|$ + 1 is the distance, i.e., the number of words, between words x and y in a Wikipedia document, $V(w_i)$ ($V(w_j)$, respectively) is the set of non-stop, stemmed words of w_i (w_j, respectively), $|V(w_i)|$ ($|V(w_j)|$, respectively) is the number of words in $V(w_i)$ ($V(w_j)$, respectively), and m is the m^{th} out of the k ($1 \leq m \leq k$) Wikipedia documents in which both w_i and w_j, or their stemmed variances, occur.

According to the precomputed keyword correlation factors, we define a fuzzy association, $\mu_{i,j}$, of the *content descriptors* of an RSS news article pair A_i and A_j, along with their *degree of similarity*, $Sim(i,j)$, as follows:

$$\mu_{k_m,j} = 1 - \prod_{k_l \in A_j} (1 - cf_{m,l}), \forall_{k_m \in A_i}$$

$$Sim(i,j) = \frac{\mu_{k_1,j} + \mu_{k_2,j} + \cdots + \mu_{k_n,j}}{n} \quad (2)$$

where $cf_{m,l}$ is given in Equation 1, $\mu_{k_m,j} \in [0, 1]$ reaches its maximum when $cf_{m,l} = 1$, i.e., $k_m = k_l$ for any $k_l \in A_j$, n is the number of keywords in A_i, and $Sim(i,j) \in [0,1]$.

In general, $Sim(i,j) \neq Sim(j,i)$. If $Sim(i,j)$ = 0, then there are no keywords in A_i that is considered similar to any keyword in A_j. When $Sim(i,j)$ = 1 = $Sim(j,i)$, A_i and A_j are *identical*. If $Sim(i,j)$: 1 and $Sim(j,i) = 1$, where : $1 \equiv \geq 0.93$ and $= 1 \equiv < 0.9$,[5] then A_i is *subsumed* by A_j, i.e., each keyword in A_i is (semantically) the same as some of the keywords in A_j. Using the correlation factors in $Sim(i,j)$, we determine whether A_i or A_j should be treated as (i) *redundant*, i.e., identical or one is `` *subsumed*" by the other, (ii) one is *less-informative* than the other, or (iii) (un-)related.

Example 1 Consider the RSS news articles A_1 and A_2 in Figure 1(a). Both articles were downloaded from the same RSS news feed, i.e., the ABC News, in November 2006, with different publication date and time. Since the keywords in the content descriptor of A_1 and A_2 are the same, $Sim(A_1, A_2)$ = 1 = $Sim(A_2, A_1)$, and A_1 or A_2 is treated as *replicated*. Another two articles A_3, downloaded from the AP International News, and A_4, retrieved from the USA Today, on January 2, 2007 are shown in Figure 1(b). The calculated similarity values between A_3 and A_4 are $Sim(A_3, A_4)$ = 0.69 and $Sim(A_4, A_3)$ = 0.93, which suggests that A_4 is subsumed by A_3.

Shown in Figure 1(c) are two other news articles A_5 (from the Washington Post) and A_6 (from the CBS News) posted on January 16, 2007. Both articles address similar event (i.e., Senator Barack Obama enters the presidential race). Since $Sim(A_5, A_6)$ = 0.36 and $Sim(A_6, A_5)$ = 0.28, the two articles contain some related information. Figure 1(d) shows another two news articles, A_7 from the Boston.com News and A_8 from the ABC News, downloaded on August 1, 2006. A_8 is *less-informative* than A_7, since $Sim(A_8, A_7)$ = 0.86 is high, which means that significant amount of information presented in A_8 is contained in A_7, whereas $Sim(A_7, A_8)$ = 0.42, which indicates that A_7 contains other information that is not available in A_8. During the clustering process of these articles, A_8 is an ideal choice to be eliminated. The corresponding μ-values of the keywords in the two articles are shown in Table I.

<item><title>Move Over, Botox: **Wrinkle 'Filler'** Is **New Option**</title>
<link><![CDATA[http://abcnews.go.com/Health/Cosmetic/story?id=2664045&CMP=...]]></link>
<pubDate>**Thu, 30 Nov 2006 16:41:31** -0500</pubDate>
<description>**Juvederm Injections Said** to **Fill Gaps** in **Aging Faces**</description></item>

<item><title>Move Over, Botox: **Wrinkle 'Filler'** Is **New Option**</title>
<link><![CDATA[http://abcnews.go.com/Health/Cosmetic/story?id=2675805&CMP=...]]></link>
<pubDate>**Thu, 23 Nov 2006 11:21:21** -0500</pubDate>
<description>**Juvederm Injections Said** to **Fill Gaps** in **Aging Faces**</description></item>

(a) A_1 (top) is *identical* to A_2 (bottom)

<title> **Egyptian Ship With 1,300 Aboard Sinks** </title>
<link>http://hosted.ap.org/dynamic/stories </link>
<description>**An Egyptian cruise ship with 1,300 people aboard has sunk in the Red Sea off the Saudi coast during an overnight crossing. At least a dozen people have been rescued, but there are reports of dozens of bodies recovered.** </description>

<title> **Egyptian Ship With 1,300 Aboard Sinks** </title>
<link>http://rssfeeds.usatoday.com/~r/UsatodaycomWorld-TopStories/~3/69844309/2007-... </link>
<pubDate>**Tue, 2 Jan 2007 18:26:28 GMT**</pubDate>
<description> **An Egyptian cruise had 1,300 people aboard when it sank in the Red Sea off the Saudi coast during an overnight crossing.** </description>

(b) A_3 (top) *subsumes* A_4 (bottom)

<title><![CDATA[**Barack Starts 2008 Bid**]]></title>
<link><![CDATA[http://www.washingtonpost.com/...]]></link>
<pubDate><![CDATA[**Tue Jan 16 14:40 EST 2007**]]></pubDate>
<description><![CDATA[**Illinois senator files exploratory committee papers; official announcement is set for Feb. 10.**]]></description></item>

<title>**Barack Obama Jumps Into 2008 Race**</title>
<pubDate>**Tue, 16 Jan 2007 13:31:24 EST**</pubDate>
<link>http://www.cbsnews.com/stories/2007/01/16/politics/main2361354...</link>
<description>**Democratic Sen. Barack Obama of Illinois took the first step in a presidential bid, filing paperwork that will allow** ...</description></item>

(c) *A*5 (top) and *A*6 (bottom) are *related*

<title>**Clean penguins return to sea after spill**</title>
<link>http://www.boston.com/news/science/articles/2006/08/01/clean_penguins_return_... </link>
<description>**Dozens of freshly cleaned Magellanic penguins waddled into the ocean Monday ..., close to 200 goo-covered birds that were rescued and washed after an oil spill.**</description>

<title>**Clean Penguins Return to Sea After Spill**</title>
<link><![CDATA[http://abcnews.go.com/Technology/wireStory?id=2261905&CMP=OTC-RSS..]]></link>
<pubDate>**Tue, 01 Aug 2006 16:20:40** -0400</pubDate>
<description>**Dozens of Cleaned-Up Penguins Return to Sea After Oil Spill Off Argentina's coast**</description>

(d) *A*8 (bottom) is *less-informative* than *A*7 (top)

Figure 1: Examples of identical, subsumed, related, and less-informative news articles

A_8/A_7	Penguin	Waddle	Ocean	Monday	Applause	Onlooker	...	μ-Values
Penguin	1	6.9×10^{-3}	1.9×10^{-3}	1.7×10^{-4}	5.9×10^{-5}	2.3×10^{-4}	...	1
Return	1.1×10^{-3}	1.1×10^{-4}	5.5×10^{-3}	2.1×10^{-3}	2.3×10^{-4}	1.1×10^{-4}	...	1.5×10^{-3}
Sea	2.5×10^{-3}	7.7×10^{-6}	$\mathbf{5.4}\times10^{-2}$	4.6×10^{-4}	3.1×10^{-5}	2.0×10^{-5}	...	9.5×10^{-3}
Oil	1.4×10^{-3}	2.0×10^{-6}	6.1×10^{-3}	7.2×10^{-4}	6.2×10^{-5}	$\mathbf{6.3}\times10^{-6}$	...	1.4×10^{-3}
Spill	2.6×10^{-3}	2.4×10^{-5}	2.5×10^{-3}	4.0×10^{-4}	6.1×10^{-5}	7.6×10^{-5}	...	9.4×10^{-4}
Argentina	7.9×10^{-4}	$\mathbf{7.9}\times10^{-5}$	4.0×10^{-3}	6.8×10^{-4}	7.7×10^{-5}	1.4×10^{-5}	...	9.4×10^{-4}
Coast	1.5×10^{-3}	3.1×10^{-5}	$\mathbf{3.6}\times10^{-2}$	1.2×10^{-3}	1.5×10^{-5}	4.0×10^{-5}	...	6.4×10^{-3}
...	...	...	...	...	...	...	...	...
μ-values	1	1.0×10^{-3}	1.6×10^{-2}	8.1×10^{-4}	7.6×10^{-5}	7.0×10^{-5}	...	

Table I: The correlation factors of (some of) the keywords in the RSS news articles A_7 and A_8 as shown in Figure 1 (d)

4 Fuzzy Equivalence Relation

After discarding *redundant* (i.e., identical or subsumed) RSS news articles based on their Sim values, we proceed to eliminate *less-informative* RSS news articles in a cluster being constructed. This task can be accomplished by first generating clusters of all the non-redundant RSS news articles (both informative and less-informative) that have a certain degree of similarity. A non-redundant cluster of RSS news articles is defined as

$$C_\alpha = \{d \mid Sim(d,e) \geq \alpha, \forall e \in C_\alpha\} \tag{3}$$

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

Clusters of news articles can be created by using a *fuzzy equivalence relation*, which is defined according to the degrees of similarity among different elements in a set, and is used to generate ``crisp'' subsets (i.e., clusters) (Klir, 1997). $R$ is a fuzzy equivalence relation on a fuzzy set $Y$, which defines a ``crisp'' relationship among the elements of Y, if R is *reflexive*, *symmetric*, and $max\text{-}min$ *transitive*, i.e.,

$$R(x,x) = 1, \forall x \in Y \tag{4}$$

$$R(x,y) = R(y,x), \forall x,y \in Y \tag{5}$$

$$R(x,z) \geq \max_{y \in Y}\{\min\{R(x,y), R(y,z)\}\}, \forall x,y,z \in Y \tag{6}$$

The key for establishing a fuzzy equivalence relation is the definition of *transitivity*. The first definition for fuzzy transitivity was proposed by Zadeh (Zadeh, 1971), which is the $max\text{-}min$ transitivity, as defined in Equation 6. However, the $max\text{-}min$ transitivity is known to be a restrictive constraint, and is not applicable to the clustering problem that we deal with in this paper. This is because in order to apply the $max\text{-}min$ transitivity to our clustering problem, it is required that for any two RSS news articles d_x and d_z in a cluster C, there cannot exist another RSS news article d_y in C which has similarities with both d_x and d_z that are greater than the similarity between d_x and d_z. Consider an example of two unrelated RSS news articles (i.e., d_x and d_z) that contain the only sentences S_1 and S_2, respectively, and a third RSS news article (i.e., d_y) that contains only sentence S_3 as shown below.

S_1: *House passes new Medicare drug bill.*
S_2: *On stem cell legislation, a new reprise.*
S_3: *House passes bill relaxing limits on stem cell research.*

Since S_3 has higher similarity with both S_1 and S_2 than S_1 and S_2 have with each other, $max\text{-}min$ transitivity is not suitable to define a fuzzy equivalence relation in this example, as well as in our work.

There exists another fuzzy transitivity relation, the $max\text{-}prod$ transitivity (Zimmermann, 1991) (defined below), which is not as restrictive as the $max\text{-}min$ transitivity and better suits the requirements of our fuzzy equivalence relation.

$$R(x,z) \geq \max_{y \in Y}\{R(x,y) \times R(y,z)\} \tag{7}$$

The $max\text{-}prod$ transitivity can be easily satisfied if the involved values fall in the interval $[0,1]$, such as the μ and Sim functions in our work, since the product of any two numbers $x, y \in [0,1]$ is smaller than x and y, i.e., if $x, y \in [0,1]$, then $x \geq x \times y$ and $y \geq x \times y$. Hence, we adapt the $max\text{-}prod$ transitivity in establishing our fuzzy equivalence relation.

In order to adapt the fuzzy equivalence relation with $max\text{-}prod$ transitivity for eliminating less-informative RSS news articles in a cluster, it is necessary to define a function that ``combines" the similarity measures of any two RSS news articles A_i and A_j, i.e., $Sim(i,j)$ and $Sim(j,i)$, into a single one, to satisfy the conditions of *symmetry* and *transitivity*. We first consider several combination functions and then choose the one that satisfies the $max\text{-}prod$ equivalent relation as the desired combination function.

4.1 Combination Functions

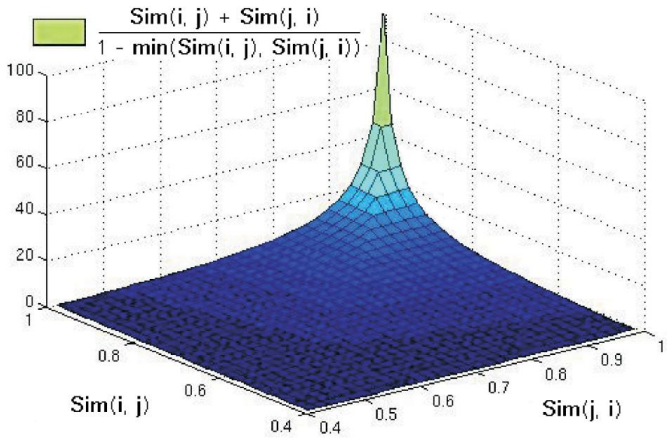
One of the most commonly used combination functions is *average*. However, the average of two pairs of significantly different similarity values, e.g., (0.5, 0.5) and (0.9, 0.1), can yield the same result, e.g., (0.5 + 0.5)/2 = (0.9 + 0.1)/2. Furthermore, although the average function is fuzzy symmetric and reflexive, it is not $max\text{-}prod$ transitive. In (Luger, 2005), two equations, Q and Q', which combine two values, e.g., $Sim(i,j)$ and $Sim(j,i)$, can be defined as follows:

$$Q(Sim(i,j), Sim(j,i)) = \frac{Sim(i,j)+Sim(j,i)}{1-min(Sim(i,j),Sim(j,i))} \tag{8}$$

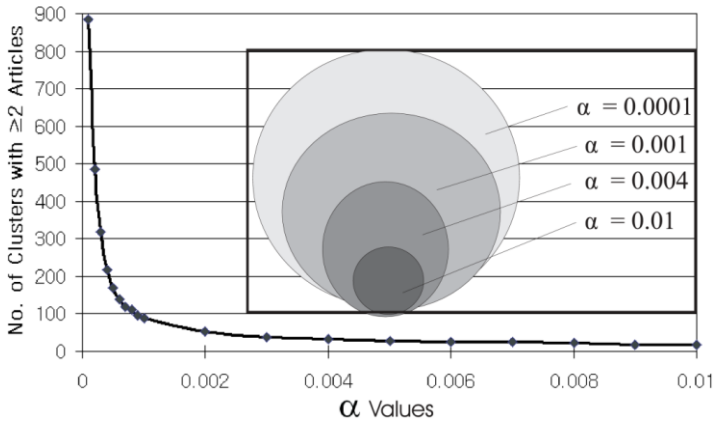
$$Q'(Sim(i,j), Sim(j,i)) = (Sim(i,j) + Sim(j,i)) - (Sim(i,j) \times Sim(j,i)) \tag{9}$$

Both functions, Q and Q', are simple to compute; however, Q' has the same drawback as the average function, i.e., it yields the same result to significantly different pairs of values. For example, both (similarity) value pairs (0.9, 0.9) and (0.99, 0.1) are assigned the same value, i.e., 0.99, by Q'. In contrast, Q assigns a high value only when both similarity measures of $Sim(i,j)$ and $Sim(j,i)$ are high, as shown in Figure 2(a). Furthermore, Q is *fuzzy-symmetric*, however, Q is neither fuzzy-reflexive nor fuzzy-transitive. We modify Q so that the modified Q function, FE, is a fuzzy equivalence relation.

$$FE(d_i,d_j) = \begin{cases} 1 & if\ i = j \\ 0.0001 & if\ Q(d_i,d_j) < 0.0001 \\ \frac{Q(d_i,d_j)}{\max(Q(d_x,d_y))} & otherwise \end{cases} \tag{10}$$



(a) A sample of values computed by using the Q function in Equation 8



(b) Numbers of equivalence classes generated by 429 articles (chart) and subsummed classes (inserted diagram) based on α's

Figure 2: Q function and the numbers of equivalence classes generated by using different α values

The 1^{st} condition in Equation 10 is introduced to satisfy *reflexivity*, whereas the 3^{rd} condition, the normalized Q function, restricts the values of FE to the interval $[0, 1]$. The 2^{nd} condition guarantees $max\text{-}prod$ *transitivity*, since 0.0001 limits the similarity value between any two articles so that the $max\text{-}prod$ transitivity is satisfied[6], a necessary and sufficient condition. In addition, since Q is *symmetric* as mentioned earlier, FE is a fuzzy equivalence relation.

4.2 Clustering with α-Cuts

We apply the fuzzy equivalence relation FE to determine the equivalence classes (clusters) of non-redundant news articles posted under different RSS news feeds by setting the α value (Klir, 1997) of fuzzy sets as shown in Equation 3. An α value is used to generate an α-cut, which is a *set of clusters* such that every pair of news articles in the same cluster has a degree of similarity not less than α. The α-cut analysis, which is widely used in representing uncertainty information, to which the informativeness of RSS news articles belong, restricts degrees of fuzzy members in different classes. We adapt this analysis

approach, instead of other clustering approaches, since α-cut is *seamless* for clustering RSS news articles whose degrees of similarity are determined by the fuzzy set theory.

As α increases, the number of equivalence classes of the corresponding α-cut decreases, and the size of each equivalence class is reduced. (See the chart in Figure 2(b) for an example.) One important property of α-cuts is that the equivalence classes generated by some α-value x will always be subsumed by the equivalence classes generated by some α-value y, if $y < x$ (see the inserted diagram in Figure 2(b) for an example).

By using *smaller* α values, which generate *larger* clusters, we can decrease (i) the probability of eliminating a cluster *entirely* during the process of discarding less-informative news articles, and (ii) the number of singletons (i.e., single-membered clusters) created. However, the clusters generated may be *loosely related* if α is too small. Thus, we often favor *larger* α values, since a larger α value generates *fewer* and *smaller* clusters in which articles are closely related. Table II provides an example of clusters generated using different α values, and Table III shows the corresponding statistics for varying values of α on a sample set of 90 news articles. Notice how the average number of articles per cluster in Table III decreases as α increases. As α increases, clusters in the α-cut are more restricted (i.e., tighter), and consequently, both the total number of clusters generated and the average size of each cluster decrease.

1. Boy in Mo. kidnapping ready for school - Family of boy in Mo. kidnapping case recalls happily receiving word of his recovery

2. Embryo saved after Katrina is born - Parents of embryo rescued after hurricane Katrina celebrate birth of boy Tuesday

3. Abducted boy may have hard time leaving psychologically - He may be the latest of kidnap victims to suffer from Stockholm syndrome

4. Parents of missing boy say his life was threatened - Parents of Shawn Hornbeck, missing four years, say captor threatened his life

RSS News Feeds (Subject Area)	# of Articles
www.english.people.com.cn (World News)	31
www.news.bbc.co.uk (Middle East)	56
www.news.bbc.co.uk (World Edition)	54
www.news.yahoo.com (Entertainment)	36
www.seattletimes.nwsource.com (Sports)	31
www.seattletimes.nwsource.com (Seattle News)	21
www.slashdot.org	21
www.prnewswire.com (Aerospace)	31
www.prnewswire.com (Automotive)	26
www.prnewswire.com (Transportation)	29
www.prnewswire.com (Travel)	25
www.suntimes.com	37

Table II: RSS news feeds used for verifying the $max\text{-}prod$ transitivity

$\alpha = 0.0007$:$\{A_1A_2A_3A_4\}$,...	$\alpha = 0.0018$:$\{A_1A_3A_4\}$,$\{A_2\}$,...	$\alpha = 0.0040$:$\{A_1A_3\}$,$\{A_2\}$,$\{A_4\}$,...

Table III: Different α values and clusters (i.e., α-cuts) generated for a set of 4 news articles. If α is too low, A_2 is clustered with unrelated articles; however, if α is too high, A_4 is isolated from related articles A_1 and A_3. The α-value 0.0018 is effective

Value of α	No. of Clusters	No. of Singletons	Average Number of Clusters an Article is in	Average Number of Articles Per Cluster
0.0001	182	0	8.32	4.38
0.0004	182	0	8.32	4.38
0.0007	182	0	8.32	4.38
0.0010	184	0	7.25	4.06
0.0013	140	1	5.46	4.01
0.0016	120	4	4.22	3.76
0.0019	**93**	**7**	**3.24**	**3.44**
0.0022	72	13	2.84	3.42
0.0025	60	20	2.51	3.50

Table IV: Various α values and their corresponding statistics on a set of 90 news articles. The α value 0.0019 is optimal due to its relatively low number of clusters and singletons

The formula for α has been empirically determined by using eight sample news articles sets (i.e., news feeds) ranging from 34 to 351 articles in size. A linear regression line was calculated using the principle of *least squares* on the test data. Optimal α values vary due to subjectivity, and are a question of both how *tight* clusters should be, and how many *singletons* should be allowed. Typically, an effective α value can be determined by the number of articles in the set, which we define as $\alpha = 15.23e^{-6} \times$ Number of Non-Redundant Articles + $1.49e^{-3}$, and the value of α not only determines what *clusters* will be generated, but also plays a key role, along with the ranking function (defined in Section 4.3), in deciding which news articles (from multiple RSS feeds) are *less-informative*. Moreover, the value of α determines the appropriate *number* of clusters to be generated, which affects the ranking on individual news articles.

4.3 The Ranking Approach

During the process of discarding less-informative news articles generated by an α, we must retain at least one article in each cluster in order to ensure that the content or ``story" of no cluster is lost. Since singleton clusters include the only news article which is dissimilar to other news articles in other clusters, they should not be eliminated.

Since we wish to reduce the total number of (sets of) articles, and the same articles may appear in different clusters[7], we cannot treat every cluster separately while selecting less-informative articles to discard, since we could potentially discard all the articles in a cluster. Consider, for example, a set of articles C = $\{a, b, c, d, e, f\}$ and their FE values: $FE(a, b)$ = 0.055, $FE(c, d)$ = 0.021, $FE(e, f)$ = 0.014,

and the same FE value, 0.0025, for all the other possible pairs of articles in C. If we set α = 0.1, then three clusters C_1 = {a, b}, C_2 = {c, d}, and C_3 = {e, f} are generated. Suppose we need to discard two articles from C. If we rank the articles only in the same cluster based on their similarity values, i.e., without considering the degrees of similarity with other articles in other clusters in which they reside, and discard the two with the highest similarity values, then a and b (which have the highest similarities values) are discarded, and consequently C_1 is eliminated, which is undesirable, since C_1 is the only cluster that contains the story of a and b. Thus, we rank each article A among all the clusters (generated by our α-cut equation) in which A resides, and discard those that have higher rankings (i.e., articles highly similar to others) among all the clusters, which are *less-informative*.

ID	Title; (Portion of the) Description
0	The Warming of Greenland; Arctic melting accelerates, revealing uncharted islands and threatening to raise sea levels all over the world ...
1	Observatory: A Smelly Puzzle, Solved; The world's largest flower, a link between lobsters and whales and where moths get a drink. ...
2	Personal Health: 104 Teenagers Who Are Role Models for Weight Loss; A look at the reasons teenagers give for gaining weight helps in understanding the steps necessary to reverse the process.
3	Q & A: Blood and History; Is there an evolutionary reason for human blood types?
4	Findings: The Voices in My Head Say 'Buy It!' Why Argue?; What is the difference between a tightwad's brain and a spendthrift's brain? ...
5	Ruins in Northern Syria Bear the Scars of a City's Final Battle; Excavation of ruins at Tell Hamoukar reveals ancient weapons of mass destruction ...
6	Baghdad Explosions Kill Dozens, Wound Scores; Series of bomb blasts outside a Baghdad university comes as U.N. report finds that more than 34,000 civilians have been killed during the past year ...
7	Barack Obama Starts 2008 Bid; Illinois senator files exploratory committee papers; official announcement is set for Feb. 10. ...
8	Trial Begins in CIA Leak Case; Defense lawyers argue publicity has damaged I. Lewis "Scooter" Libby's chances for a fair trial. ...
9	Wide Berth Given on Sex Ed; Montgomery County schools invoke codes in defense of lessons on orientation, transgenderism. ...
10	Lawmaker Angers Blacks, Jews; Va. delegate says blacks "should get over" slavery, Christ suggests Jews should "apologize for killing ...
11	More Than 100 Dead In Baghdad Attacks; More than 100 people were killed in attacks in Baghdad, while the U.S. military announced the deaths of four soldiers. The deaths came as the U.N. said more than 34,000 Iraqi civilians were killed ...
12	Barack Obama Jumps Into 2008 Race; Democratic Sen. Barack Obama of Illinois took the first step in a presidential bid, filing paperwork that will allow the newcomer who has rocketed to the top of national politics to raise money for ...
13	Potential Libby Jurors Split On Iraq War; Potential jurors for the trial of a former White House adviser are voicing mixed views about President Bush and the Iraq war. "Scooter" Libby is accused of lying to investigators about what he told ...
14	Chills Across U.S. As Ice Storm Heads East; After weeks of unseasonably mild weather, snow and ice have hit most of the U.S., killing at least 46 people in seven states. Meanwhile, cold temperatures out west have destroyed as much as ...
15	Oil Prices Plummet To 19-Month Low; Oil prices dropped by $2 a barrel to a 19-month low after a report that OPEC powerhouse Saudi Arabia said there's no need for further production cuts ...
16	Iran Buys Surplus U.S. Military Hardware; The Associated Press reports that U.S. military hardware surplus sales have been taken advantage of by arms dealers, who, on several occasions, have sold sensitive bits and pieces on to Iran and China ...

Table V: RSS news articles downloaded from New York Times (www.nytimes.com), Washington Post (www.washingtonpost.com), and CBS News (www.cbsnews.com) on January 16, 2007

We consider the $Rank$ function (given below) to rank articles in all the clusters created by an α-cut that include at least two articles, i.e., articles in singleton clusters are not ranked, since they are not

candidates for elimination. $Rank$ computes the *average* of the *maximum* similarity values of an article d_i with respect to each article d_j in each cluster C_k in which d_i appears.

$$Rank(d_i) = \frac{\sum_{k=1}^{N} \max\{\forall_{d_j \in C_k} Sim(i,j)\}}{N} \quad (11)$$

where N is the total number of non-singleton clusters in which d_i appears. According to the rankings, the top n ($n \geq 1$) ranked (less-informative) articles, are discarded. The value n can be indirectly determined by (i) the *average* number of new articles that are accessed by the user, or (ii) the *number of articles* posted by an individual RSS news feed that the user accesses on a regular basis.

$Rank$ operates in $O(Nm)$ time, where m is the largest number of news articles among all the clusters. The similarity values between each pair of news articles within the same cluster are already known, since they must be computed for all news articles before the α-cut can be created. Furthermore, since the number of keywords in the content descriptor of each news article is relatively small, with an average of 3-4 keywords in the *title* and another 15-20 in the *description*, the computational complexity of the μ-values in $Sim(i,j)$ (as shown in Equation 2) for any two news articles A_i and A_j can be ignored. Computing $Sim(i,j)$ for all pairs of k news articles for an α-cut would require $O(k^2)$ time, and subsequently ranking the k news articles for elimination will require $O(kNm)$ time.

Example 2 Consider the set of 17 non-redundant RSS news articles that were extracted from various RSS news feeds as shown in Table IV. Assume that 30% of the (less-informative) news articles (i.e., a total of five articles) are supposed to be deleted. Table VI shows the non-singleton clusters generated by the α-value 0.013 using Equation 3, which along with the rankings of the 17 articles computed by using Equation 11 on their degrees of similarity, determine which five articles should be discarded.

Original Clusters	Clusters After Elimination	No. of Articles Eliminated	Final Set of Clusters	Ranking
{ 0, 1 }	{ 0, 1 }	0	{ 0, 1 }	**7**
{ 2, 3 }	{ 2 }	1		**8**
{ 0, 5 }	{ 0, 5 }	0	{ 0, 5 }	**3**
{ 4, 6, 14 }	{ 4, 14 }	1	{ 4, 14 }	**6**
{ 4, 8, 13 }	{ 4 }	2		11
{ 8, 9 }	{ 9 }	1	{ 9 }	**13**
{ 6, 10, 11, 14 }	{ 10, 11, 14 }	1	{ 10, 11, 14 }	12
				4
				1
{ 7, 11 }	{ 11 }	1		10
{ 0, 12 }	{ 0, 12 }	0	{ 0, 12 }	0
{ 2, 12 }	{ 2, 12 }	0	{ 2, 12 }	9
{ 7, 12 }	{ 12 }	1		14
{ 12, 13 }	{ 12 }	1		5
{ 6, 15, 16 }	{ 15, 16 }	1	{ 15, 16 }	16
{ 4, 6, 16 }	{ 4, 16 }	1	{ 4, 16 }	2
{ 11, 16 }	{ 11, 16 }	0	{ 11, 16 }	15

Table VI: News articles in each cluster (created by using the α-value 0.013) before and after eliminating less-informative ones, along with the ranking of the 17 articles in Table V

The first article to be eliminated is Article 7, which is less-informative than Article 12 in the same cluster. (Both articles refer to Senator Obama's presidential bid, but Article 7 only mentions the fact, whereas Article 12 provides more detail.) Subsequently, Articles 8, 3, and 6 are eliminated, since each addresses political affairs that are mostly covered by other articles, respectively, e.g., Article 6 is mostly covered in Article 11, and Article 8 is mostly covered in Article 13. These articles were *manually examined* for their relative degrees of similarity. The next news article in the ranking order to be eliminated is Article 11. This article is closely related to Article 6, which covers an accident in Baghdad and has already been eliminated. Since Article 11 is also grouped in a cluster with Article 7, in order to avoid discarding the entire cluster {7, 11} after Article 7 has been removed earlier, we retain Article 11. As a result, we eliminate the next article in the ranking, i.e., Article 13 (on political affair).

After the elimination process is completed, we proceed to eliminate all the *clusters* that are either *duplicated* or *subsumed* by others (as shown in Table VI), which yields only informative clusters.

Each generated cluster C is represented by the top k ($\geq$ 1) keywords among the most frequently-occurred ones in the content descriptors of the news articles in C, and the determination of the ideal k is beyond the scope of this paper.

5 Complexity Analysis of Our Clustering Algorithm and Its Implementation

In this section, we first introduce our clustering algorithm $FEGe$ for generating fuzzy equivalence classes of RSS news articles. (The entire process of $FEGe$ is graphically captured in Figure 3.) Hereafter, we evaluate the overall time complexity of $FEGe$ and discuss its implementation.

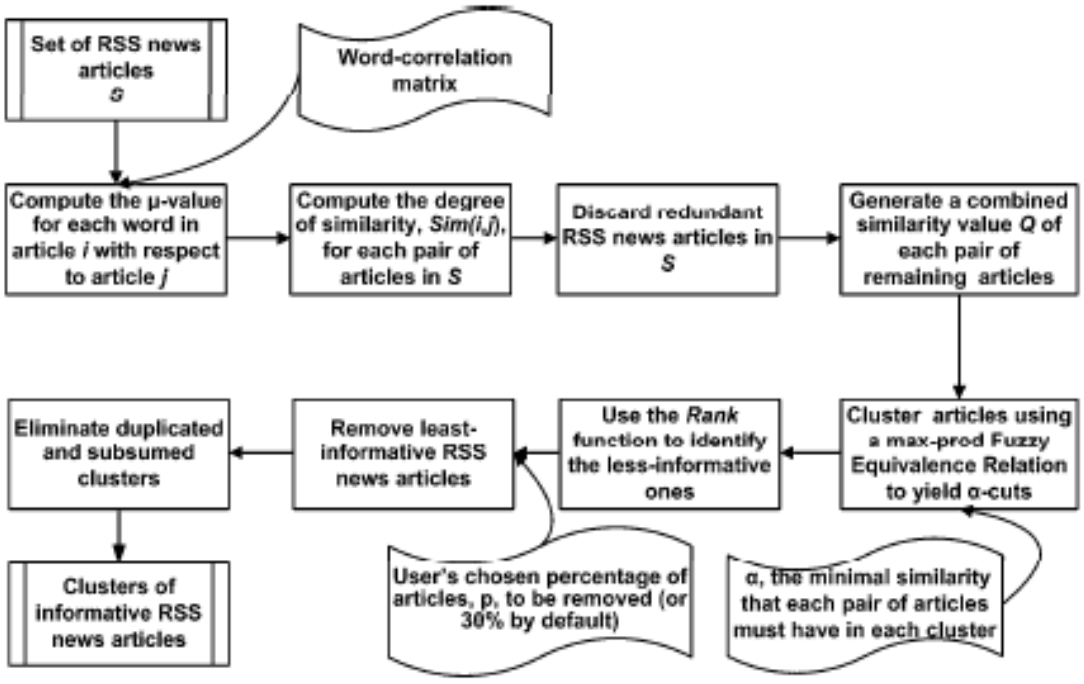


Figure 3: The entire process of our clustering algorithm *(FEGe)*

Algorithm $FEGe$: Fuzzy Equivalence classes Generator /* Cluster RSS news articles */
Input: A set of RSS news articles S, the *Word-Correlation Matrix* WCM, the *percentage* of articles to be eliminated p, the *minimal similarity* α, which can be computed as $15.23e^{-6} \times$ Number of Non-Redundant Articles + $1.49e^{-3}$
Output: A set of clusters, each containing a collection of closely-related, informative, non-redundant RSS news articles in S
/* Calculate the degrees of similarity between articles */
1. Set R := S
2. **WHILE** there exists an article d_i in R
2.1. **WHILE** there exists an article d_j in R such that $d_i \neq d_j$
2.1.1. **FOR** each word $w \in d_i$, **DO**
(a) Compute the similarity μ-value of w with respect to the words in d_j using Equation 2 and WCM
2.1.2. Compute and store $Sim(d_i, d_j)$, the *degree of similarity* of d_i with respect to d_j, using Equation 2 and the computed μ-values in Step 2.1.1
2.1.3. **FOR** each word $w \in d_j$, **DO**
(a) Compute the similarity μ-value of w with respect to the words in d_i using Equation 2 and WCM
2.1.4. Compute and store $Sim(d_j, d_i)$, the *degree of similarity* of d_j with respect to d_i, using Equation 2 and the computed μ-values in Step 2.1.3
2.1.5. **IF** ($Sim(d_j, d_i) \geq$ 0.93 AND $Sim(d_i, d_j) <$ 0.90) **OR**
($Sim(d_i, d_j)$ == 1 AND $Sim(d_j, d_i)$ == 1), **THEN**
/* compare the Sim values with the redundancy thresholds as shown in Section 3 */
(a) Remove d_j from S and R /* eliminate a duplicated/subsumed article */
2.2. Remove d_i from R
3. Set k := |S| /* the number of remaining non-redundant articles in S */
4. **FOR** each pair of articles d_i and d_j in S such that $d_i \neq d_j$, **DO**
4.1. Compute and store $Q(d_i, d_j)$, the *combined degree of similarity* between d_i and d_j, using $Sim(d_i, d_j)$ and $Sim(d_j, d_i)$ calculated in Steps 2.1.2 and 2.1.4, and Equation 8
5. **FOR** each pair of articles d_i and d_j in S, **DO**
5.1. Compute and store $FE(d_i, d_j)$, the *fuzzy equivalence relation value* between d_i and d_j, using $Q(d_i, d_j)$ computed in Step 4.1 and Equation 10
/* Form clusters */
6. Set $Clusters$:= $\emptyset$ /* a set of clusters, initially empty */
7. Add { d } to $Clusters$ /* the initial singleton cluster, where $d \in S$ */
8. Set R := S - { d }
9. **FOR** each d_i in R, **DO** /* d_i is an RSS article */
9.1. Set $newCluster$:= $true$ /* assuming a new cluster should be created */
9.2. **FOR** each cluster c in $Clusters$, **DO**
9.2.1. Set add := $true$ /* assuming d_i should be added to c */
9.2.2. **FOR** each article d_k in c, **DO**
(a) **IF** $FE(d_k, d_i) < \alpha$, **THEN**
(i) add := $false$ /* d_i and at least an article in c should not be assigned to the same cluster */
(ii) Break /* from the FOR Loop of Step 9.2.2 */
9.2.3. **IF** add == true, **THEN**

(a) Add d_i to c
(b) $newCluster$:= $false$
9.3. **IF** $newCluster$ == $true$, **THEN**
Add { d_i } to $Clusters$
/* Rank articles in various clusters */
10. **FOR** i := 1..k, **DO** /* d_i is an RSS article */
10.1. $AcumMaxSim$:= 0 /* sum the *maximum similarity value* of d_i */
10.2. $CountCluster$:= 0 /* counter of *non-singleton* clusters in which d_i appears */
10.3. **For** each non-singleton cluster c in $Clusters$ in which d_i appears, **DO**
10.3.1. $AcumMaxSim$:= $AcumMaxSim$ + $\max\{\forall_{d_j \in c} Sim(d_i, d_j)\}$
10.3.2. $CountCluster$:= $CountCluster$ + 1
10.4. Set $Rank\_arr[i]$:= $\frac{AcumMaxSim}{CountCluster}$ /* Ranks d_i as shown in Equation 11 */
11. Order articles 1..k by $Rank\_arr[]$ using *Quick Sort* in ascending order
/* Eliminate less-informative articles */
12. Cnt := $k \times p$
13. **FOR** i := 1..k, **DO** /* d_1..d_k are ranked articles in $Rank\_arr[]$ */
13.1. **IF** Cnt == 0, **THEN**
Break /* from the FOR Loop of Step 13 */
13.2. **ELSE IF** d_i is in a *non-singleton* cluster, **THEN**
13.2.1. Remove d_i from all clusters in which d_i resides
13.2.2. Set Cnt := Cnt - 1
/* Eliminate duplicated or subsumed clusters */
14. **FOR** each cluster c in $Clusters$
14.1. **IF** c is *subsumed* by another cluster d in $Clusters$, **THEN** /* $c \subseteq d$ */
14.1.1. Remove c from $Clusters$

Among all the inputs of *FEGe*, the *word-correlation factors*, which are computed by using Equation 1, are in the input matrix *WCM*, which is preprocessed, and thus is not involved in the run-time complexity analysis.

Article-to-article similarity measures, which are calculated in Step 2 of *FEGe*, can be performed in $O(k^2W^2)$ time, where k is the number of articles in the given set of RSS news articles S and W is the number of words in an article. After analyzing the various test sets of RSS news articles used in Section 6 (the Experimental Results), we observe that the average number of non-stop, stemmed words in RSS news articles is 16 and the number of articles to cluster usually reaches thousands. As a result, $k >> W$, and W is not a significant factor compared with k. Hence, $O(k^2W^2) \sim O(k^2)$.

After eliminating duplicated and redundant RSS news articles in Step 2.1.5, the fuzzy equivalence class values among the existing pairs of non-redundant RSS news articles are constructed, i.e., Steps 4 and 5, which requires $O(k^2)$. Hereafter, clusters are formed (as shown in Steps 6 through 9 of *FEGe*) such that initially one singleton cluster is created in constant time, which could be further expanded, and the steps require $O(kNs)$, where N is the number of existing clusters and s is the average size of any given cluster. Afterwards, ranking (less-informative) RSS news articles for further article elimination is performed (as shown in Step 10). As discussed in Section 4.3, the Rank function requires $O(Nm)$ time, where m is the largest number of news articles among all the clusters, and thus a complete ranking takes $O(kNm)$ time. The method chosen for sorting, i.e., Step 11, in *FEGe* is quick sort, which runs in $O(k \log k)$.

Eliminating less-informative articles as shown in Steps 12 and 13 of *FEGe* is done progressively, until the quota for less-informative articles to be deleted is met. This step takes $O(kNs)$ time. The final segment of the algorithm, as shown in Step 14 of *FEGe*, involves eliminating subsumed clusters, which can be done simply in $O(N^2)$ time.

In the worst case, the overall complexity of *FEGe* is $O(k^3)$, which occurs in Step 10 when RSS news articles are ranked so that the less-informative ones can be eliminated hereafter.

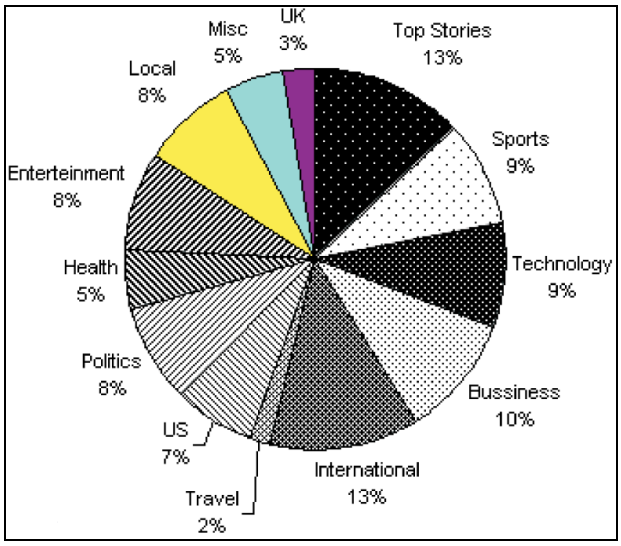
The implementation of *FEGe* was written using the Perl programming language and installed on an Intel Centrino Duo workstation with dual 2.66 GHz processors, 3 GB Ram, and a hard disk of 320 GB, running under the Linux (Ubuntu) operating system. The precomputed word-correlation matrix was implemented using the C programming language and runs on the same workstation as *FEGe*.

6 Experimental Results

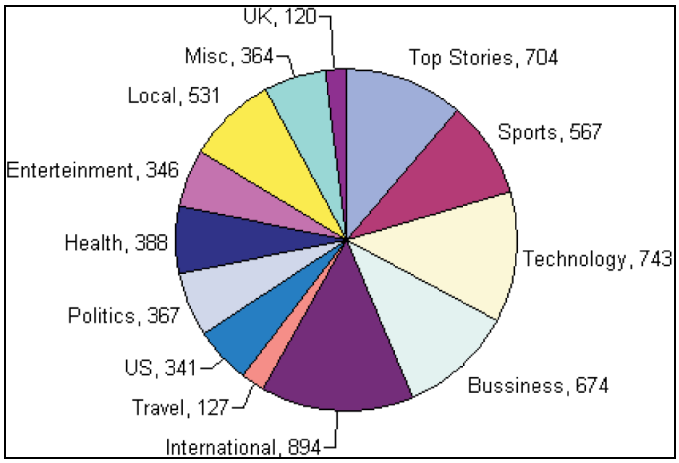
In this section, we present the conducted experiments and based on their results we verify the accuracy of our approach on (i) detecting redundant and less-informative RSS news articles collected from various RSS news feeds and (ii) clustering informative RSS news articles. We analyzed different test sets and determined the number of *false positives*, i.e., news articles that were mistakingly treated as redundant or less-informative and discarded in each set, which was then converted into a percentage of *accuracy*. Table VII and Figure 4 show the wide variation and subject areas of RSS news articles from which each test case was drawn. The collections of articles in each test case was downloaded from multiple RSS feeds, often containing the same (or similar) news articles. We also include a case study in Section 6.5 to demonstrate the merit of our clustering approach.

ID	Sources	RSS Feeds								
		No.	Number of Articles (in the Subject Area)							
1	1115.org	2	10 (TS)	10 (TS)						
2	abcnews.go.com	20	49 (Bs) 17 (Po) 41 (TS)	60 (Bs) 47 (Po) 47 (TS)	23 (ET) 30 (Sp) 65 (US)	32 (ET) 39 (Sp) 28 (US)	68 (Hl) 54 (Te)	115 (Hl) 58 (Te)	26 (IT) 25 (Tr)	43 (IT) 67 (Tr)
3	adn.com	4	6 (Sp)	6 (Sp)	8 (TS)	8 (TS)				
4	blogs.zdnet.com	2	10 (Te)	10 (Te)						
5	bloomberg.com	10	32 each (Bs (3), ET, IT, Lc, Po, Sp, TS (2))							
6	boston.com	8	10 (Lc)	11 (Lc)	15 (Lc)	19 (Lc)	30 (Lc)	38 (Lc)	24 (Mi)	48 (TS)
7	businessweek.com	4	29 (Bs)	29 (Bs)	45 (TS)	45 (TS)				
8	cbsnews.com	8	56 (Bs)	50 (ET)	53 (Hl)	42 (IT)	52 (Po)	51 (Te)	7 (TS)	40 (US)
9	chinaview.cn	6	60 each (Bs, ET, Hl, IT, Mi, Po)							
10	chron.com	6	25 (Bs)	21 (ET)	23 (IT)	25 (Po)	21 (Sp)	20 (TS)		
11	cnn.com	8	8 (ET)	10 (IT)	10 (Mi)	8 (Po)	10 (Po)	10 (TS)	10 (TS)	8 (US)
12	dailymail.co.uk	2	21 (TS)	22 (TS)						
13	dallasnews.com	5	5 (Bs)	5 (ET)	5 (Lc)	5 (Sp)	5 (TS)			
14	english.people.com	12	11 (IT) 11 (Sp)	18 (IT) 11 (Sp)	19 (IT) 15 (TS)	19 (IT) 19 (TS)	11 (Mi)	11 (Mi)	11 (Mi)	16 (Mi)
15	forbes.com	2	10 (Bs)	4 (Bs)						
16	foxnews.com	5	16 (Bs)	15 (ET)	24 (Hl)	23 (IT)	18 (Lc)			
17	guardian.co.uk	3	15 (IT)	15 (IT)	14 (UK)					
18	health.telegraph.co.uk	1	50 (Bs)							
19	hosted.ap.org	23	10 each (Bs(2), ET(2), Hl(2), IT (3), Lc(2), Mi (3), Po(2), Sp(2), Te, TS(2), US(2))							
20	iht.com	9	36 (Bs) 38 (US)	38 (IT)	44 (IT)	32 (IT)	28 (IT)	23 (Sp)	38 (Sp)	68 (Te)
21	latimes.com	11	12 each (ET, Hl, IT(2), Lc, Mi, Po, Sp, TS(2), US)							
22	money.cnn.com	6	10 (Bs)	10 (Bs)	16 (Bs)	50 (Bs)	10 (Te)	20 (TS)		
23	money.telegraph.co.uk	1	50 (Bs)							
24	mybroadband.co.za	2	17 (Te)	10 (Te)						
25	news.bbc.co.uk	4	30 (IT)	16 (Lc)	29 (TS)	17 (UK)				
26	news.com.au	4	7 (IT)	5 (Lc)	12 (Po)	9 (Sp)				
27	news.ft.com	9	15 (Bs) 14 (US)	15 (Hl)	10 (IT)	15 (IT)	14 (IT)	10 (Lc)	15 (Po)	8 (TS)
28	news.telegraph.co.uk	4	27 (IT)	10 (Lc)	7 (TS)	19 (UK)				
29	news.yahoo.com.	3	13 (ET)	11 (Po)	20 (TS)					
30	nytimes.com	9	6 (Hl) 6 (US)	6 (IT)	6 (Lc)	6 (Mi)	6 (Mi)	6 (Po)	6 (Sp)	48 (Te)
31	online.wsj.com	12	4 (Bs) 12 (Te)	7 (Bs) 18 (Te)	9 (Bs) 17 (US)	19 (Bs) 18 (US)	12 (IT)	17 (IT)	21 (IT)	25 (IT)
32	politics.guardian.co.uk	1	15 (Po)							
33	portal.telegraph.co.uk	1	50 (TS)							
34	prnewswire.com	10	17 (Mi) 120 (Te)	20 (Mi) 20 (Tr)	20 (Mi)	20 (Mi)	20 (Mi)	20 (Mi)	20 (Mi)	20 (Mi)
35	seattletimes.nwsource. com	10	14 (Bs) 230 (Sp)	20 (ET) 27 (US)	24 (IT)	24 (Lc)	235 (Lc)	5 (Mi)	23 (Mi)	15 (Po)
36	sfgate.com	4	10 (ET)	25 (Sp)	106 (Te)	29 (TS)				
37	siliconvalley.com	2	20 (Te)	20 (TS)						
38	slashdot.org	1	14 (Te)							
39	sltrib.com	2	7 (Lc)	11 (US)						
40	sportsillustrated.cnn	5	8 (Sp)	9 (Sp)	9 (Sp)	10 (Sp)	10 (Sp)			
41	timesonline.co.uk	4	16 (Bs)	42 (IT)	10 (Sp)	20 (UK)				
42	today.reuters.com	10	10 each (Bs, ET, IT (3), Po, Sp, Te, TS, US)							
43	usatoday.com	11	15 each (Bs, ET, Hl, IT, Lc, Mi (2), Te, Tr, TS (2))							
44	washingtonpost.com	2	3 (Lc)	15 (US)						
45	wired.com	3	24 each (Te (3))							
46	worldpress.org	2	7 (IT)	40 (IT)						
	Total Number of RSS	273								

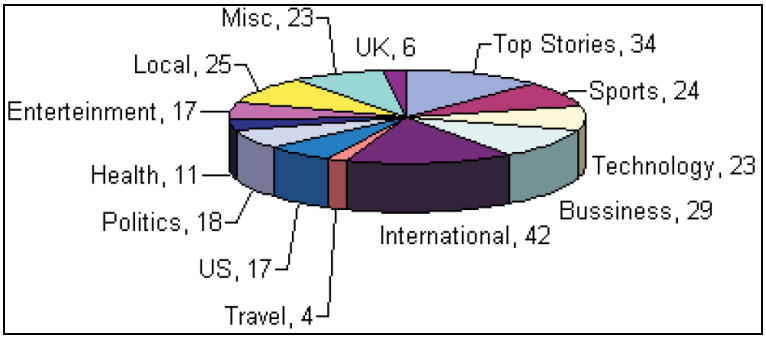
Table VII: A portion of the 273 feeds used for verifying our informative clustering approach with different subject areas: **B**u **s**sines (Bs); **E**n **t**ertainment (ET); **H**ea **l**th (Hl); **I**n **t**ernational (IT); **L**o **c**al (Lc); **Mi**scellaneous (Mi); **Po**litics (Po); **Sp**orts (Sp); **Te**chnology (Te); **Tr**avel (Tr); **T**op **S**tories (TS); **U**nited **K**ingtom (UK); **U**nited **S**tates (US)



(a) The ratios of subject areas



(b) Number of RSS news articles



(c) Number of distinct RSS feeds

Figure 4: Statistical data on subject areas, number of articles, and distinct RSS feeds in Table VII

6.1 Verifying the Accuracy of Detecting Redundant RSS New Articles

Table VIII shows the size of each test case used for identifying redundant articles, along with its result verified by human judges. The results confirm the correctness of our redundancy-detection approach by the 11 test cases. After *manually* examining the articles in various test cases, we conclude that the overall success rate in detecting *redundant* articles (from either the same or different RSS feeds) is *perfect* and have achieved an average accurate rate of 86% in detecting all the *subsumed* articles. (The three missed subsumed articles are caused by our high threshold values of subsumption, i.e., 0.93 and 0.9.)

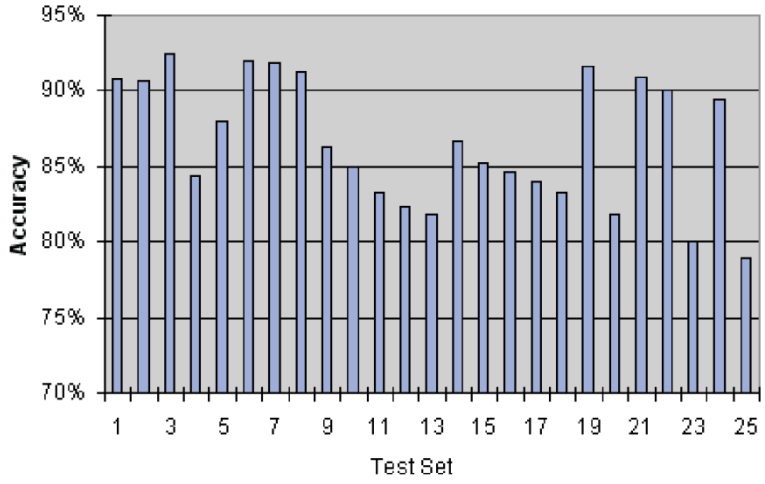
Test Case	Number of Articles	Identical Articles			Subsumed Articles		
		Existed	Eliminated	Accuracy Ratio	Existed	Eliminated	Accuracy Ratio
TC_1	11	0	0	100%	0	0	100%
TC_2	31	2	2	100%	5	4	80%
TC_3	62	3	3	100%	0	0	100%
TC_4	86	3	3	100%	1	1	100%
TC_5	103	8	8	100%	0	0	100%
TC_6	115	5	5	100%	0	0	100%
TC_7	128	7	7	100%	2	2	100%
TC_8	139	14	14	100%	5	4	80%
TC_9	179	2	2	100%	1	1	100%
TC_{10}	194	24	24	100%	3	3	100%
TC_{11}	388	34	34	100%	4	3	75%
AVG	131	9.3	9.3	100%	1.91	1.64	86%

Table VIII: Test cases for verifying the accuracy in detecting redundant (i.e., identical and subsumed) news articles, which were extracted from source data in Table VII

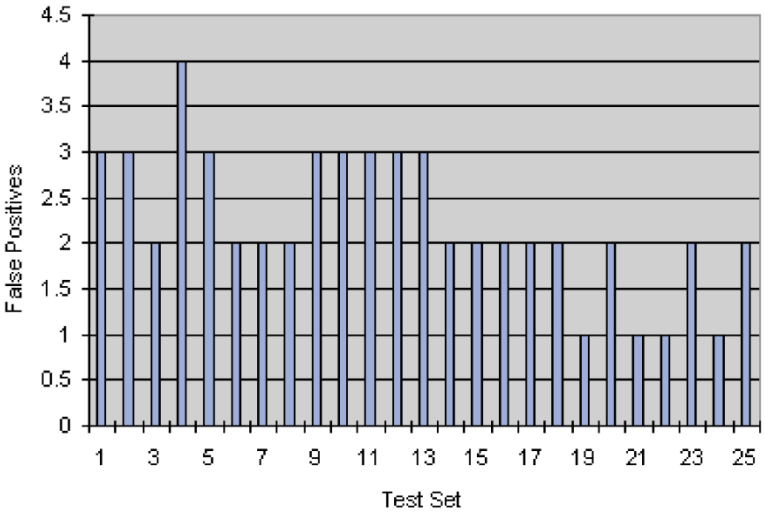
6.2 Accuracy of Our Less-Informative News Articles Elimination Approach

We have considered 25 different test sets, each of which contains multiple non-redundant RSS news articles from various RSS news feeds. We *manually* evaluated the (less-informative) articles chosen automatically to be discarded and counted the number of false positives. We calculated the percentage of *accuracy* as $1 - \frac{FalsePositives}{Discarded}$.

The twenty-five test sets show a high degree of accuracy, even when the percentage of articles discarded is high. The percentages of discarded articles are in the range of 10% to 50%. Figure 5(a) shows the results when 50% of the articles in the 25 test sets are discarded, whereas Table IX shows the accuracy percentage and the corresponding numbers of false positives for each distinct percentage of discarded less-informative news articles. The results show that the overall average of accuracy is 87%.



(a) Percentages of non-redundant news articles that are correctly detected as informative



(b) False positives among all the news articles in test cases

Figure 5: Accuracy percentages for 25 different RSS test sets with 50% discarding rate

Test Set	ToA	SoA	NoA	Percentages of Articles to be shed									
				10%		20%		30%		40%		50%	
				Acc%	FP	Acc%	FP	Acc%	FP	Acc%	FP	Acc%	FP
1	World	3	65	100%	0	92%	1	85%	3	88%	3	91%	3
2	Techno	2	64	100%	0	100%	0	90%	2	88%	3	91%	3
3	USA	3	53	81%	1	91%	1	87%	2	91%	2	92%	2
4	World	2	51	100%	0	100%	0	87%	2	85%	3	84%	4
5	USA	2	50	60%	2	80%	2	87%	2	90%	2	88%	3
6	Sports	2	50	100%	0	90%	1	93%	1	90%	2	92%	2
7	USA	2	49	100%	0	100%	0	93%	1	95%	1	92%	2
8	World	2	46	78%	1	89%	1	93%	1	89%	2	91%	2
9	Sports	2	44	77%	1	89%	1	85%	2	89%	2	86%	3
10	World	2	40	100%	0	88%	1	83%	2	81%	3	85%	3
11	USA	3	36	72%	1	86%	1	91%	1	86%	2	83%	3
12	World	2	34	71%	1	85%	1	90%	1	85%	2	82%	3
13	Entert	3	33	100%	0	85%	1	90%	1	85%	2	82%	3
14	World	1	30	100%	0	100%	0	89%	1	83%	2	87%	2
15	Entert	2	27	100%	0	100%	0	88%	1	91%	1	85%	2
16	World	3	26	100%	0	81%	1	87%	1	81%	2	85%	2
17	USA	1	25	60%	1	80%	1	87%	1	90%	1	84%	2
18	World	1	24	100%	0	79%	1	72%	2	79%	2	83%	2
19	Sports	1	24	100%	0	100%	0	100%	0	90%	1	92%	1
20	USA	2	22	100%	0	100%	0	85%	1	77%	2	82%	2
21	Entert	1	22	100%	0	100%	0	85%	1	89%	1	91%	1
22	USA	1	20	100%	0	100%	0	100%	0	88%	1	90%	1
23	World	1	20	0%	2	50%	2	67%	2	75%	2	80%	2
24	USA	1	19	100%	0	100%	0	82%	1	87%	1	89%	1
25	World	1	19	100%	0	74%	1	82%	1	87%	1	79%	2
Average		1.84	36	**87%**	0.4	**90%**	0.7	**87%**	1.3	**86%**	1.8	**87%**	2.2

T(opic)o(f)A(rticles); S(ources)o(f)A(rticles); N(umber)o(f)A(rticles)

Table IX: A portion of the 25 test cases, where Acc% and *FP* denote the *percentage* of *accuracy* and the number of *false positives* of the corresponding test case, respectively

6.3 The Overall Accuracy and Observations of Our Clustering Approach

Consider the test set of articles as shown in Table VIII and their corresponding sources as shown in Table VII to further verify the accuracy in generating clusters of non-redundant and informative news articles using 0.0019 as the α-cut value and discarding 30% of the less-informative news articles. The empirical results of the conducted experiments are shown in Table X, which indicates an average of 96% accuracy on the test sets.

Using different values of α-cut and elimination percentages of articles in various test cases, we observe that the best clusters of a set of articles are generated when the recommended α-cut value (computed by using the α-equation in Section 4.2) is adapted, since (i) the number of singleton clusters is *low* (and in some cases there are none at all), (ii) a *good* number of clusters (i.e., the number of clusters do not exceed the number of articles) are generated, and (iii) *larger* clusters are reduced. According to the experimental results, the ideal number of clusters (with appropriate number of articles)

is created when 20-30% of the articles are eliminated, since when the numbers of articles to be eliminated is *too small*, many similar articles could remain, whereas if we eliminate more than 30% of the articles, sometimes we not only eliminate similar ones, but also news stories entirely. Furthermore, when α-cut is *low* and the elimination percentage is small (10%), we often obtain clusters that include too many not-closely-related articles. When α-cut is *high* and the elimination percentage is also *high* (50%), we obtain more clusters, and most of them often consist of two articles. Hence, after the less-informative elimination process, all or most of the clusters are reduced to singleton clusters and in some cases, entire clusters are eliminated (due to the requested number of articles to be deleted). However, when the number of articles is *small* and the elimination percentage is *high*, we are forced to delete some informative ones.

Test Case	No. of Articles	No. of Non-Redundant Articles	No. of Removed Less-Informative Articles	No. of Clusters Before Elimination	No. of Clusters After Elimination	Accuary of Our Clustering Approach
TC_1	11	11	3	5	3	66.7%
TC_2	31	25	8	10	9	100%
TC_3	62	59	18	46	31	100%
TC_4	86	82	25	110	81	96.0%
TC_5	103	95	28	100	77	96.4%
TC_6	115	110	33	128	82	97.0%
TC_7	128	120	36	145	114	97.3%
TC_8	139	122	36	268	251	100%
TC_9	179	176	53	274	225	100%
TC_{10}	194	168	51	282	254	100%
TC_{11}	388	351	105	914	751	100%

Table X: Verifying the average accuracy, which is 95.8%, in clustering non-redundant and removing 30% of the less-informative RSS news articles using various test cases in Table VII

6.4 Performance Evaluation Against Other Existing Clustering Approaches

We have verified the merit of our clustering approach by comparing its performance with other well-known clustering approaches, such as k- *means* and *Naive Bayes*. We chose the Reuters corpora (Reuters-21578 distribution 1.0, http://kdd.ics.uci.edu/data bases/reuters 21578/reuters21578.html), which consists of 21,578 news articles manually assigned to one or more 135 established topics, for performance analysis, since the corpora is a popular benchmark data widely used for verifying the merit of any text categorization in IR, machine learning, knowledge discovery, or data mining method. We considered articles in the Reuters corpora that belong to a *unique* topic and ignored topics that have *less* than five articles to facilitate the comparisons, which is the same evaluation strategy used by (Xu, 2004), since it is more accurate to analyze articles that belong to a unique topic. As a result, 9,494 articles and 51 different topics were chosen. To quantify the performance among different clustering methods, including ours, we applied the mutual information (MI) metric (Xu, 2003) (given below), which defines how *similar* or *independent* any two given sets of news article clusters C and C' are, where $p(c_i)$ ($p(c'_j)$, respectively) is the *probability* that a randomly selected article from the reduced Reuters data set belongs to cluster c_i in C (c'_j in C', respectively) and $p(c_i, c'_j)$ is the *probability* that a randomly selected article belongs to both c_i and c'_j.

$$MI(C,C') = \sum_{\forall c_i \in C, \forall c'_j \in C'} p\left(c_i, c_j\right) \times log_2 \frac{p\left(c_i, c_j\right)}{p(c_i) \times p(c'_j)} \quad (12)$$

Since $MI(C,C') \in$ [0..max($H(C), H(C')$)], where $H(C)$ and $H(C')$ denote the *entropy* of C and C', respectively, we normalize the MI metric value as $\widehat{MI}(C,C') = \frac{MI(C,C')}{max(H(C),H(C'))}$, where $\widehat{MI}(C,C') \in$ [0..1], $H(C) = -\sum_{i=1}^{k} p(c_i) log_2 p(c_i)$ measures the *purity* of the set of clusters in C, and $H(C')$ can be defined accordingly.

We evaluate our clustering approach according to the procedure described in (Xu, 2004), which randomly selects news articles from k ($\geq$ 2) different clusters in the reduced Reuters corpora. The set of k clusters and the set of clusters C' generated by using our clustering approach (using α-cut = 0.0019 and discarding 30% of the articles) form the two sets of clusters to obtain their $\widehat{MI}$ values. We repeated this process for different sets of k clusters with randomly selected articles, which are evaluated against C', to yield different test cases and results. Table XI shows some of the test cases used in the evaluation. The test cases include different numbers (14-50) of randomly chosen articles and topics, which generate (3-14) scalable clusters and yield various (0.39-0.88) $\widehat{MI}$s, a similar evaluation procedure as in (Xu, 2004) and are reliable in measuring the accuracy of our clustering approach.

Test Case	Number of Clusters	Number of Articles	$\widehat{MI}$
1	14	50	0.61
2	4	50	0.70
3	7	26	0.88
4	7	14	0.39
5	6	18	0.61
6	5	30	0.64
7	5	20	0.75
8	3	15	0.59
9	6	20	0.52
10	5	20	0.47
Average	6	26	0.66

Table XI: Test cases used for computing the average $\widehat{MI}$ between the clusters in the Reuters corpora and the ones generated by ours

Figure 6 shows the average $\widehat{MI}$ of the clusters generated by using our clustering approach and the other twelve clustering approaches mentioned in (Xu, 2004). Among all the clustering approaches, k- *mean* selects k different seeds (i.e., documents) to determine the centroids of the clusters and iteratively attempts to find new clusters in order to minimize the total intra-cluster variance between the documents and their corresponding cluster centers. The *Gaussian Mixture Model* or *Naive Bayes* method defines a *probabilistic* cluster model and tries to find the model by maximizing the likelihood of data to be clustered. The *Spectral-based clustering* approaches use eigenvectors of a *similarity* matrix derived from the data to be clustered, which establish a measure of similarity between any two points in

the data in order to perform dimensionality reduction to reduce the dimensions in clustering. Xu, et al. (Xu, 2003) use the *Non-Negative matrix*, which captures the topic of a document cluster and represents each document as an additive combination of the base topics, to determine the cluster any document D belongs by finding the base topic with which D has the largest projection value. The *Concept Factorization* technique models each concept as a linear combination of data points and computes the sets of linear coefficients for clustering and labeling the data points.

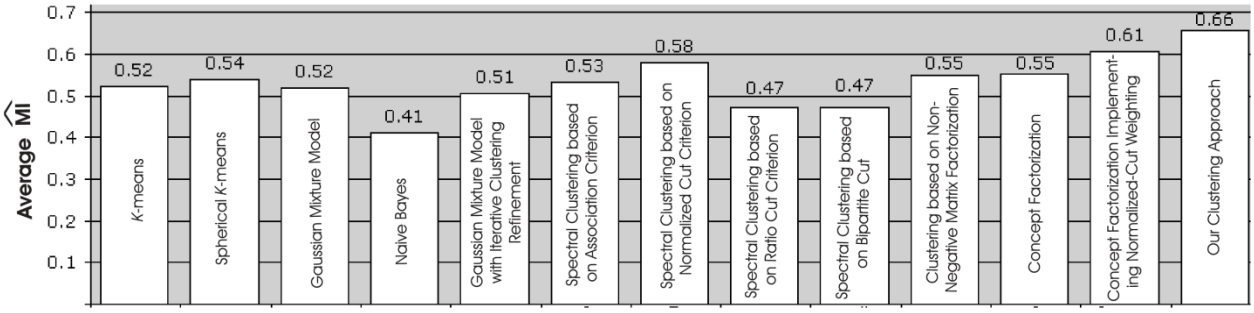


Figure 6: Average normalized Mutual Information ($\widehat{MI}$s) computed for various clustering approaches given in (Xu, 2004) along with ours

As illustrated in Figure 6, our clustering approach outperforms all the other clustering approaches based on averaging $\widehat{MI}$s computed for all the different test cases on the Reuters corpora. In fact, not only our clustering approach achieves a higher average $\widehat{MI}$ value than others, ours is not affected by initialization (i.e., the required training cases) as in other approaches. Furthermore, our clustering approach enhances its performance when the number of articles to be clustered is large, since before clustering we discard redundant and less-informative ones, which yields clusters with higher quality.

6.5 A Case Study

In this section, we use three different news topics, i.e., sports, entertainment, and top news, to show a variety of informative data of interest during the clustering process of RSS news articles from different sources (i.e., RSS feeds). We randomly selected 398 RSS news articles on October 23, 2008, which were downloaded from abcnews.com, news.yahoo.com, hosted.ap.org, foxnews.com, sfgate.com, latimes.com, cnn.com, seattletimes.nwsource.com, cbsnews.com, and www.nytimes.com, and the number of news articles belonged to each one of the three arbitrarily chosen topics are shown in Table XII.

	Sports	Entertaintment	Top News	Total
Original number of articles	173	98	127	398
Number of identical articles	6	7	7	20
Number of subsumed articles	6	2	10	18
Number of remaining articles after redundant ones are eliminated	161	89	110	360
Number of original clusters	148	59	72	279
Number of articles in the largest/ smallest, original clusters	8/1	7/1	8/1	
Average number of articles per cluster	5.6	3.7	4.8	
Total number of less-informative, eliminated/informative articles	48/113	28/61	32/78	108/252
Total number of informative clusters	90	49	59	198
Average number of articles per informative cluster	3.4	2.2	2.6	

Table XII: Statistical data on the articles and clusters generated by FEGe on the initial 398 RSS news articles downloaded from 10 different RSS feeds

The 398 collected RSS news articles generate 158,006 distinct pairs of articles, out of which 20 were identical and 18 were subsumed. As a result, 38 articles were identified as redundant and discarded, and the actual number of non-redundant articles to be clustered was 360. The remaining non-redundant RSS news articles were clustered using the $max\text{-}prod$ fuzzy equivalence relation and α-cuts, where $\alpha = 15.23e^{-6} \times 360 + 1.49e^{-3} = 6.9e^{-3}$, which yielded 279 clusters. Hereafter, the clustered articles were ranked and 30%[8] of the less-informative articles were eliminated. After discarding 108 (30% of 360) less-informative articles, each cluster that became a subset of another cluster was eliminated. The final number of generated clusters is reduced to 198, and the average number of informative articles per cluster belonged to each one of the three distinct topics is shown in Table XII.

While Table XII summarizes the number of articles and clusters generated by $FEGe$, Table XIII provides other informative data on the set of clusters. Note that in Table XIII the numbers of clusters from where articles were deleted represent the number of clusters in the original sets which included at least one less-informative article to be eliminated. Moreover, the numbers of articles in the largest (smallest, respectively) clusters shown in Table XII reflects the size of the largest (smallest, respectively) clusters in the *original* sets of generated clusters, whereas the same values reported in Table XIII reflect the size of the largest (smallest, respectively) clusters in the *final* set of (informative) clusters. In addition, since the similarity values among RSS news articles are within the range of 0 and 1, the average maximum similarity value, 0.47, of the eliminated less-informative articles implies that each eliminated article A shares, on the average, 50% of its content with at least another article in at least one cluster in which A appears, which further verifies that $FEGe$ can identify and discard less-informative RSS news articles.

	Sports	Entertaintment	Top News	Total
Number of original clusters from where articles were deleted	121	39	61	221
Number of informative clusters	90	49	59	198
Number of singleton/non-singleton informative clusters	26/64	12/37	14/45	52/146
Size of the largest/smallest informative cluster	4/1	3/1	3/1	
Average maximum similarity value of the eliminated less-informative articles	0.53	0.42	0.46	0.47 (Average)

Table XIII: Statistical data on the original and final sets of clusters generated by $FEGe$

Figure 7 shows the number of original clusters, the number of singleton clusters, and the final numbers of clusters created by using different α values and percentages of discarded less-informative RSS news articles. Regardless of which α value was used, the total number of less-informative articles to be eliminated was the same for each percentage of discarded articles, i.e., 10% prunes 36 articles, 30% deletes 108 articles, and 50% eliminates 180 articles. However, when the α value increases, the final number of clusters decreases and the final number of singleton cluster increases, since when the minimal similarity value among the articles in a cluster is higher, only more closely related articles remain in the cluster.

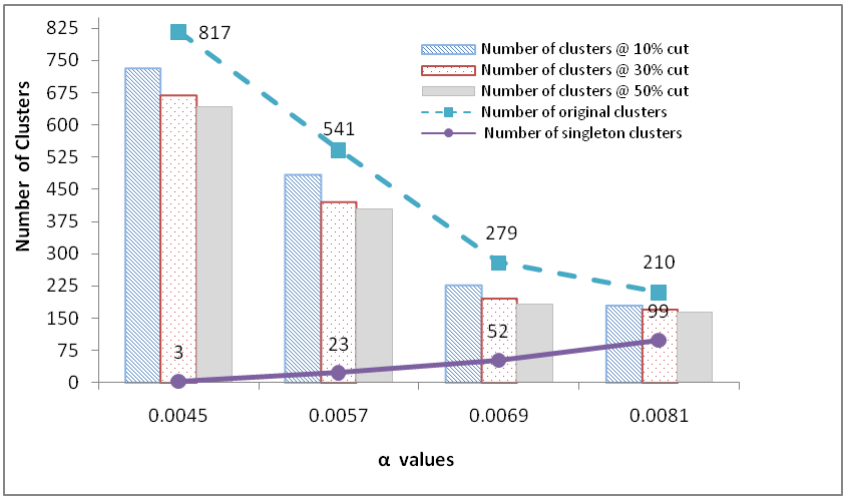


Figure 7: Analysis of the case study for different α values and percentages of discarded less-informative RSS news articles

Based on the results of this case study, we observe that by using the α value, i.e., 0.0069, computed by the equation discussed in Section 4.3, $FEGe$ generates clusters of related RSS news articles without creating an excessive number of singleton clusters (as shown in Figure 7). Furthermore, we manually examined the generated clusters and conclude that (i) the overall intra-cluster similarity is

high, i.e., articles within the same cluster are closely related in terms of their content, and (ii) the overall inter-cluster similarity is *low*, i.e., articles within different clusters include different contents. As shown in Figure 8, on the average RSS news articles within the same cluster share approximately 35% of their content, whereas articles in different clusters share less than 0.1% of their content.

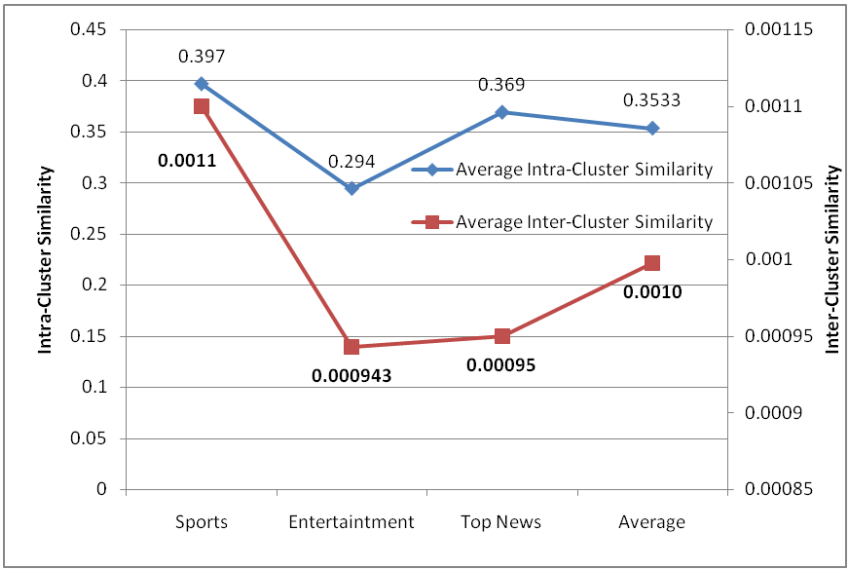


Figure 8: Average inter- and intra- cluster similarity values

7 Contributions and Significant Impacts of Our Clustering Approach

The proposed RSS news article filtering and clustering approach first filters redundant news articles from RSS feeds, shed less-informative, non-redundant ones, and then cluster the remaining informative articles into various fuzzy equivalence classes. As shown in Section 6, our clustering approach is computationally *efficient*, since it requires little overhead because the keyword-correlation factors used for computing the degrees of similarity among different RSS news articles are precomputed, and the time complexity in performing the filtering and clustering tasks is polynominal. This efficience is due to the facts that there are (i) no preprocessing steps involved in filtering RSS news articles, (ii) no user inputs are necessary in determining redundant, less-informative, or closely related RSS news articles, and removing redundant/less-informative news articles and clustering informative ones are computationally inexpensive. Furthermore, our RSS news article filtering and clustering methods achieve high degree of accuracy (in the upper 80% in general), and thus are highly *effective* in eliminating redundant and less-informative RSS news articles and clustering non-redundant and informative RSS news articles. These methods are theoretical sound, using well-established mathematical models such as fuzzy set theory and the α-cut value.

Clusters generated by our clustering approach (i) aids the users in seeking closely related information instantly that avoids the tedious and annoying process of searching for news articles of interest, (ii) minimizes the storage and processing time required for archiving useless information, (iii) can be computed in a timely manner, which satisfies the demand of processing online information in a timely manner. Last but not least, all the implementation details of the proposed filtering and clustering method are hidden from the (naive) users.

8 Conclusions

We have presented a method for clustering non-redundant and informative RSS news articles, which combines well-established techniques, such as α-cuts and fuzzy set logic, with other innovative approaches, such as using *distance-correlation factors*, to filter and classify (i.e., cluster) information in RSS feeds accurately and efficiently. Moreover, we have developed a formula for computing the *optimal value* of α, resulting in $\alpha\text{-}cuts$ which contain an ideal number of clusters and very few singleton ones. We have also analyzed the results of our method for eliminating different percentages of redundant and less-informative RSS news articles, which achieve high accuracy (in the upper 80% in general).

According to the experiments conducted, our clustering approach outperforms other existing clustering approaches. In fact, our clustering approach imposes little overhead. Furthermore, keyword-correlation factors are precomputed, and the algorithms for removing redundant/less-informative news articles and clustering informative ones are computationally inexpensive. Due to the *small overhead*, *high degree of efficiency* and *effectiveness*, and without any *labor-intensive* user feedback, our clustering approach can be implemented in real-world RSS feeds to aid users in locating RSS news articles of interest among tens of thousands posted daily.

References

(Banerjee, 2007) Banerjee, S., Ramanathan, K., Gupta, A. (2007), "Clustering Short Texts Using Wikipedia", In Proceedings of the 30th ACM SIGIR, pp. 787-788.

(Berger, 1996) Berger, A., Pietra, A., Pietra, J. (1996), "A Maximum Entropy Approach to Natural Language Processing". Computational Linguistics, Vol. 22, Issue 1, pp. 39-71.

(Blei, 2003) Blei, D., Ng, A., Jordan M. (2003), "Latent Dirichlet Allocation", Journal of Machine Learning Research. Volume 3, pp. 993-1022.

(Broder, 1997) Broder, A., Glassman, S., Manasse, M., Zweig, G. (1997),"Syntactic Clustering of the Web", Computer Networks and ISDN Systems, Vol. (29), No 8-13, pp. 1157-1166.

(Bun, 2002) Bun, K., Ishizuka, M. (2002), "Topic Extraction from News Archive Using TF*IDF Algorithm", In Proceedings of WISE, pp. 73-82.

(Cheng, 2006) Cheng, D., Kannan, R., Vempala, S., Wang, G. (2006), " A Divide-and-Merge Methodology for Clustering", ACM TODS, Vol. 31, No 4, pp. 1499-1525.

(Gruhl, 2006) Gruhl, D., Meredith, D., Pieper, J., Cozzi, A., Dill, S. (2006), "The Web Beyond Popularity: a Really Simple System for Web Scale RSS", In Proceedings of the 15th International Conference on World Wide Web, pp. 183-192.

(Hofmann, 1999) Hofmann, T. (1999), "Probabilistic LSA", Proceedings of the 22nd ACM SIGIR, pp. 50-57.

(Khmelev, 2003) Khmelev, D., Teahan, W. (2003), "A Repetition-Based Measure for Verification of Text Collections and for Text Categorization", In Proceedings of ACM SIGIR, pp. 104-110.

(Klir, 1997) Klir, G.K., St. Clair, U., Yuan, B. (1997), "Fuzzy Set Theory, Foundations and Applications". Prentice Hall.

(Li, 2005) Li, Y., Chung, S. (2005), "Document Clustering Based on Frequent Word Sequences", In Proceedings of CIKM, pp. 293-294.

(Li, 2007) Li, X., Yan, J., Deng, Z., Ji, L., Fan, W., Zhang, B., Chen, Z. (2007), "A Novel Clustering-Based RSS Aggregator", In Proceedings of World Wide Web, pp. 1309-1310.

(Luger, 2005) Luger, G. (2005), "Artificial Intelligence, Structures and Strategies for Complex Problem Solving", 5^{th} Ed, Addison Wesley.

(Nallapati, 2004) Nallapati, R., Feng, A., Peng, F., Allan, J. (2004), "Event Threading within News Topics",

In Proceedings of CIKM, pp. 446-453.

(Ordonez, 2003) Ordonez, C. (2003), "Clustering Binary Data Streams with K-Means", In Proceedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 10-17.

(Phan, 2008) Phan, X., Nguyen, L., Horiguchi, S. (2008), "Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-Scale Data Collections", In Proceeding of the 17th International Conference on World Wide Web, pp. 91-100.

(Pon, 2007) Pon, R., Cardenas, A., Buttler, D., Critchlow T. (2007), "Tracking Multiple Topics for Finding Interesting Articles", In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560-569.

(Sahoo, 2006) Sahoo, N., Callan, J., Krishnan, R., Duncan, G., Padman, R. (2006), "Incremental Hierarchical Clustering of Text Documents", In Proceedings of ACM CIKM, pp. 357-366.

(Samper, 2008) Samper, J., Castillo, P., Araujo, L., Merelo, J., Cordon, A., Tricas, F. (2008), "NectaRSS, an Intelligent RSS Feed Reader", Journal of Network and Computer Applications. Vol. 31, Issue 4, pp. 793 - 807.

(Takeda, 2007) Takeda T., Takasu A. (2007), "UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing", In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 438-439.

(Tang, 2007) Tang, W., Xiong, H., Zhong, S., Wu, J. (2007), "Enhancing Semi-supervised Clustering: a Feature Projection Perspective", In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 707-716.

(Wang, 2002) Wang, Y., Kitsuregawa, M. (2002), "Evaluating Contents-Link Coupled Web Page Clustering for Web Search Results", In Proceedings of ACM CIKM, pp. 499-506.

(Xu, 2003) Xu, W., Liu, X., Gong, Y. (2003), "News Article Clustering Based on Non-Negative Matrix Factorization", In Proceedings of ACM SIGIR, pp. 267-273.

(Xu, 2004) Xu, W., Gong, Y. (2004), "News Article Clustering by Concept Factorization", In Proceedings of ACM SIGIR, pp. 202-209.

(Yang, 1998) Yang, Y., Pierce, T., Carbonell, J. (1998), "A Study on Retrospective and On-Line Event Detection", In Proceedings of ACM SIGIR, pp. 28-36.

(Yang, 2006) Yang, H., Callan, J. (2006), "Near-Duplicate Detection by Instance-Level Constrained Clustering", In Proceedings of ACM SIGIR, pp. 421-428.

(Zadeh, 1971) Zadeh, L. (1971), "Similarity Relations and Fuzzy Orderings". Information Sciences, Vol. 3, No. 2, pp. 177-200.

(Zimmermann, 1991) Zimmermann, H. (1991), "Fuzzy Set Theory and Its Applications". Kluwer Academic.

[1]Corresponding Author

[2]``RSS" refers to the following standards: Rich Site Summary (RSS 0.91), RDF Site Summary (RSS 0.9 and 1.0), and Really Simple Syndication (RSS 2.0).

[3]A *must-link* constraint determines which pairs of RSS news articles should belong to the same cluster, whereas a *cannot-link* constraint establishes which pairs of RSS news articles should belong to different clusters.

[4]Keywords denote words that are non-stopwords and stemmed. *Stopwords* are very common words, such as prepositions, demonstrative, interrogative, and indefinite pronouns, which do not provide useful information to distinguish the content of different articles. *Stemmed words* are words with common morphological and inflectional endings removed, which minimize the number of (semantically the same) keywords to be compared in different articles.

[5]The threshold values, 0.93 and 0.9, were determined by the conducted empirical study in 2006.

[6]We verified the $max\text{-}prod$ *transitivity* of FE using 400 randomly chosen news articles from 12 different RSS feeds, which include BBC News, PR Newswire, and Slashdot. A portion of the RSS feeds and their corresponding numbers of articles used are shown in Figure 1

[7]A fuzzy equivalence relation based on $max\text{-}prod$ transitivity, which is differed from the $max\text{-}min$ transitivity, does not always yield disjoint equivalence classes, a desirable property, since articles may contain a variety of information, and should be allowed to reside in different classes (clusters) that include news articles with which they share some portion of information.

[8]30% is the percentage of RSS news articles to be discarded by default.