2007-01-10

Eliminating Redundant and Less-informative RSS News Articles Based on Word Similarity and A Fuzzy Equivalence Relation

Ian Garcia
Brigham Young University - Provo

BYU ScholarsArchive Citation

Garcia, Ian, "Eliminating Redundant and Less-informative RSS News Articles Based on Word Similarity and A Fuzzy Equivalence Relation" (2007). *All Theses and Dissertations*. 1054.
https://scholarsarchive.byu.edu/etd/1054

ELIMINATING REDUNDANT AND LESS-INFORMATIVE RSS NEWS ARTICLES BASED ON WORD SIMILARITY AND A FUZZY EQUIVALENCE RELATION

by

Ian Garcia

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree

Master of Science

Department of Computer Science

Brigham Young University

April 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Ian Garcia

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date Yiu-Kai Dennis Ng, Chair

Date Phillip Windley

Date Christophe Giraud-Carrier

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Ian Garcia in its final form and have found that (1) its format, citations and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

____________________ ____________________

Date

Yiu-Kai Dennis Ng

Chair, Graduate Committee

Accepted for the Department

____

Parris Egbert

Graduate Coordinator

Accepted for the College

____

Thomas W. Sederberg

Associate Dean, College of Physical and

Mathematical Sciences

ABSTRACT

ELIMINATING REDUNDANT AND LESS-INFORMATIVE RSS NEWS ARTICLES BASED ON WORD SIMILARITY AND A FUZZY EQUIVALENCE RELATION

Ian Garcia

Department of Computer Science

Master of Science

The Internet has marked this era as the information age. There is no precedent in the amazing amount of information, especially network news, that can be accessed by Internet users these days. As a result, the problem of seeking information in online news articles is not the lack of them but being overwhelmed by them. This brings huge challenges regarding processing of online news feeds, i.e., how to determine which news article is important, how to determine the quality of each news article, and how to filter irrelevant and redundant information. In this thesis, we propose a method for filtering redundant and less-informative RSS news articles that solves the problem of excessive number of news feeds observed in RSS news aggregators. Our filtering approach measures similarity among RSS news entries by using the

Fuzzy-Set Information Retrieval model and a fuzzy equivalent relation for computing word/sentence similarity to detect redundant and less-informative news articles.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

During the past decades, besides abundant amounts of information on the Web, the way information is released has also been changed. “Just a decade ago, large-scale flows of information, such as news feeds, were owned, monitored, and filtered by organizations specializing in the provision of news. The Web has brought the challenges and opportunities of managing and absorbing news feeds to all interested users” [GDH 04]. The traditional way in which Internet users access news is by visiting a Web site, and revisiting the Web site to check for updates on any information. Typical Internet users are not interested in the information of only one Web site, instead they often visit several Web sites, diversified for various sources of information. If an Internet user wants to remain current with updates posted by those Web sites, (s)he would have to visit all the selected Web sites several times a day, which is a tedious and inefficient process. Since accessing online news is one of the favorite activities of Internet users[1], eliminating redundant and less-informative online news articles could assist Internet users in terms of saving time in locating useful information.

[1] According to the Newspaper Audience Database (NADbase), http:// www.naa.org/nadbase, one in three Internet users (55 million) visited a news portal over the course of a month, which was incremented 21% from the year of 2005 to the first quarter of the year of 2006.

In March 1999, Dan Libby created a method for syndicating information (originated by UserLand Software in 1997) for use on the *My Netscape* portal, called RSS[2]. RSS is a series of XML formats for Web syndication which has become the de-facto standard for news portals and has also been widely adopted to release other type of information, such as Weblogs, commercial Web sites, job listings, bug reports, and government information on the Web. As it has been claimed in *Feedster*[3], "Everything that is timely and valuable on the Web will be available as an RSS feed."

With the development of RSS, Internet users can (i) personalize the news they are interested in by including the headlines from the Web sites they access on a regular basis, and (ii) retrieve an RSS file containing not only the headlines, but a short description of each news and a link to the source of the news. The users can have their favorite sites summarized on one page, which is very useful. Due to the large number of online news portals and huge amount of RSS news feeds, the challenge now is to minimize the excessive amount of RSS news feeds users have to process informative news articles in a timely manner. A possible solution to this problem is to deliver personalized news by removing redundant or less-informative news articles.

In this thesis, we propose a filtering strategy for detecting and eliminating redundant and less-informative RSS news articles among the excessive number of news feeds entries observed in RSS news aggregators. Our filtering approach measures similarity among RSS news entries by using the Fuzzy-Set Information Retrieval (IR) method and a word cluster for computing word/sentence similarity to detect redundant and less-informative news articles.

We proceed to present our results as follows. In Chapter 2, we discuss related

[2] The abbreviation is used to refer 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] Feedster (http://www.feedster.com) is one of the first Web sites to search, crawl, and index Weblogs.

works in detecting similar documents. In Chapter 3, we introduce different word clusters and the Fuzzy-Set IR approach for detecting similar and redundant RSS news articles. In Chapter 4, we propose the detailed design on using a fuzzy equivalence relation for filtering less-informative RSS news articles. In Chapter 5, we present the experimental results, which justify the accuracy of our approach in detecting less-informative and redundant RSS news articles. In Chapter 6, we give a concluding remark.

Chapter 2

Related Work in Similarity Measures

Computing the similarity between documents has been extensively studied as an essential tool for applications such as text document searching [CHI 05], document clustering [ZK 05], copy or plagiarism detection [SGM 95, NH 96], text document retrieval, filtering, and categorization. Determining whether two documents are similar and to what extent they are similar is a non-trivial problem. Due to its complexity, automatic detection of document similarity is a difficult task. The accuracy of similarity detection between two documents d_1 and d_2 relies heavily on computing the degree of similarity between d_1 and d_2, which can be determined by (i) the degree of lexical overlap in terms of the contents of d_1 and d_2, or (ii) the semantic contents of d_1 and d_2, i.e., words/sentences in d_1 and d_2 that should be treated as (semantically) the same or different. The semantic content approach goes beyond counting the number of words that appear in both d_1 and d_2, and the ability to assess the degree of semantic similarity between d_1 and d_2 automatically, scalably, and accurately is a key factor for justifying the effectiveness of most information handling and decision support systems that detect similar text documents.

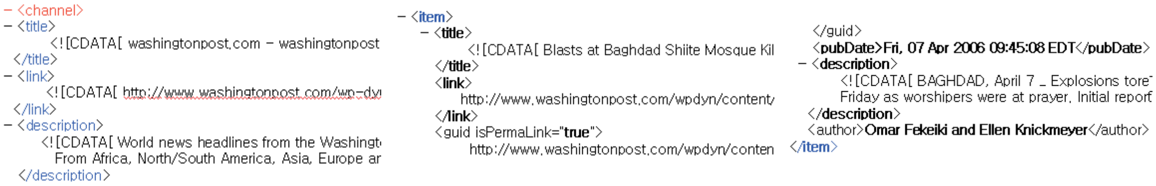
Many efforts have been made for computing the degree of similarity between two documents. From relatively simple programs, such as the widely used *diff* command in UNIX/LINUX, which compares any two text documents in a line-by-line fashion and displays the contents and the line numbers where the two documents differ, through other more complex systems, such as COPS (COpy Protection System) [BDG 95], which is designed for detecting plagiarism. SIF [MAN 94] is one of the first copy-detection systems developed for detecting similarity among documents, which was intended not only for detecting similar text documents but binary documents as well. SIF, which considers the checksum of a file as its *fingerprint*, identifies similar files in a file system and its approach is completely syntactic. COPS and SCAM (Stanford Copy Analysis Mechanism) [SGM 95] are two other copy-detection systems, which index a collection of documents by assigning hash values to sentences and paragraphs and comparing the hash values to determine the similarity among the corresponding documents. The main drawback of these copy detection approaches is the creation of a large number of collisions, same as other approaches that use hashing.

While SCAM is designed for comparing only small documents in a word-based fashion, document index graph (DIG) [HKK 04] is a document clustering system that uses a phrased-based matching model and an index model to detect similarity among documents. Even though *Diff, SIF, DIG, COPS*, and *SCAM* are different systems with different design goals, they all adopt lexical comparison approaches and thus fail to consider lexically different but semantically the same or similar documents. We adopt a semantic document comparison method that has been developed in [OMK 91, YNG 05], called Fuzzy-Set Information Retrieval (IR) model, for detecting similar, but not necessary the same, documents.

Chapter 3

Word Clusters and the Fuzzy-Set IR Model

Detecting redundant and less-informative RSS news articles is a challenging task, since RSS news feeds are dynamic in nature. The technology of RSS allows Internet users to subscribe to Web sites that typically add or modify content regularly and rapidly. To use this technology, site owners create or obtain specialized software (such as a content management system), which is in the machine-readable XML format, and present new articles in a list, including a line or two of each article and a link to the full article. (See as an example of an RSS feed file as shown in Figure 3.1.) One of the essential elements in an RSS file is *Channel*, which contains several sub-elements that describe the information contained in the file. Sub-elements of *Channel* include (i) *title*, which is similar to the *title* tag in an HTML file and is used for identifying the RSS news feeds. Usually, the content of *title* is the name of the Web site that provides the RSS file. (ii) *Link*, which is the URL of the Web site from where the RSS file can be retrieved. (iii) *Description*, which includes a sentence that briefly describes what the "stories," i.e., news articles, contained in the file are about, such as politics, economics, sports, etc. (iv) *Item*, which is a story identified by an <item>



```
- <channel>
- <title>
      <![CDATA[ washingtonpost.com - washingtonpost
  </title>
- <link>
      <![CDATA[ http://www.washingtonpost.com/wp-dyi
  </link>
- <description>
     <![CDATA[ World news headlines from the Washingt
       From Africa, North/South America, Asia, Europe ar
  </description>
```

```
- <item>
  - <title>
        <![CDATA[ Blasts at Baghdad Shiite Mosque Kil
    </title>
    <link>
       http://www.washingtonpost.com/wpdyn/content/
    </link>
    <guid isPermaLink="true">
         http://www.washingtonpost.com/wpdyn/conten
```

```
    </guid>
    <pubDate>Fri, 07 Apr 2006 09:45:08 EDT</pubDate>
  - <description>
       <![CDATA[ BAGHDAD, April 7 _ Explosions tore
       Friday as worshipers were at prayer. Initial report
    </description>
    <author>Omar Fekeiki and Ellen Knickmeyer</author>
  </item>
```

Figure 3.1: Portion of a RSS news feed file.

tag. Several *items* can be specified in the *Channel* element. The most important sub-elements of an *item* are (a) *title*, which contains the headline of the story, (b) *link*, which is the URL where the story in full can be retrieved, (c) *description*, which contains a few lines about the story and many times it is the first sentences of the story, and (d) *pubDate*, which is the date and time when the story is posted. We treat an *item* as a tuple of an RSS news feed, which contains the elements in the <item> tag, i.e., *title*, *link*, *description*, and *pubDate*.

We propose a selective filtering approach using the Fuzzy-Set IR model, where RSS news entries to be filtered are the ones that provide less-information or possess information which is already included in other RSS news entries, i.e., redundant, from a different or even the same RSS news feed. This can be achieved by comparing the RSS news entries, i.e., tuples, and determining the similarity (in terms of content) among them. Details of our filtering approach are discussed below.

3.1 The Fuzzy-Set IR Model

The fuzzy set theory relies on two main concepts: (i) sets are not well-defined, and (ii) an element (e.g., a word) has a degree of membership to a set which falls within the range of the real interval [0, 1] [KSY 04]. In [YNG 05], the Fuzzy-Set IR model is adopted to determine whether a keyword in a sentence belongs to a fuzzy set that contains different words, which have certain degrees of similarity among themselves.

The degrees of similarity, also refereed as the *correlation factors* among words, are given by a function which assigns a value in the range [0, 1] to any two words. There are several methods to define the correlation factors among different words. The *keyword-connection* correlation factor has been used as the correlation metric in [YNG 05, OMK 91], and the *association* and *metric* metrics [BYR 99] have also been defined.

Using the correlation factors among different words, the fuzzy set IR model computes the degrees of similarity of words contained in two documents, which has been proven to provide good similarity measures of text documents [OMK 91]. The Fuzzy-Set IR model in [OMK 91] assigns correlation factors among words based on the number of documents where the words appear together; however, the frequency of *co-occurrences* of different words and their *distances* within each document are not considered at all, whereas [BYR 99] only define the co-occurrence (i.e., *association*) and distance (i.e., *metric*) matrices, but do not further illustrate how they could be adopted in different applications. The Fuzzy-Set IR model [YNG 05] measures the associativity between a word w ($\in d_1$) and a document d_2 by using the correlation values between w and all the words in d_2. If d_2 includes w or contains a word with a high degree of similarity with w, then w is considered highly related to d_2. Instead of adopting the keyword-connection approach to define the correlation factors among distinct words in [YNG 05], we consider all three matrices, i.e., the *keyword-connection*, *co-occurrence*, and *distance* matrices (which are discussed in details in subsequent sections), and choose the one to be used in the Fuzzy-Set IR model, which provides the most accurate similarity measures among different documents. Using these similarity measures, we can compare the news articles in the same or different RSS news feeds to discard news articles that are either redundant or less-informative than others.

3.2 The Correlation Factor

The Fuzzy-Set IR model makes use of *correlation factors*, each of which is a similarity measure of any two words w_1 and w_2, i.e., the degree of similarity between w_1 and w_2. The correlation factor between w_i and w_j $(i, j \geq 1)$ is given in a symmetric, correlation matrix M, where each entry $m_{i,j} \in M$ is the correlation factor between w_i and w_j (also called keywords in [BYR 99]) and $m_{i,j} \in [0, 1]$.

In order to generate a correlation matrix of correlation factors, we must first collect a large number of "representative" English documents, which should be *unbiased* in terms of writing styles and *diversed* in contents, to calculate the correlation factors among distinct words according to their numbers of occurrences within each document of the collection. Some of the most popular sets of (archive) documents used in related projects are the TREC collection (http://trec.nist.gov/) and the Gutenberg project (http://www.gutenberg.org) These sets of documents, however, have major drawbacks. The Gutenberg project, which is a collection of books that is periodically augmented with new books, lacks a variety of topics, especially in science and technology. Even though the TREC collection includes a wide variety of topics, its public version has not been updated for several years. For these reasons, we have chosen the Wikipedia collection. Wikipedia [WIKI 05] is a free online encyclopedia, which contains more than 930,000 articles and approximately 340 million words. The collection of Wikipedia articles, which are written by more than 89,000 volunteers, overcomes the drawbacks of the TREC and Gutenberg collections, since Wikipedia contains almost all possible topics in different areas of study, is constantly updated, and is unbiased in terms of writing styles and authorship. With the use of the Wikipedia documents to generate word-correlation factors, we can obtain a reliable similarity measures of different words according to their occurrences in various documents.

The Wikipedia articles are comprised in a single XML file of approximately 4.6Gb in size. We obtain the frequencies of co-occurrence among different words and their relative distances within each document by "filtering" the contents in the "title" and "text" tags of each Wikipedia article. The filtering process requires two processing steps: (i) removing stopwords [BYR 99] and (ii) stemming remaining words [Porter 80]. Stopwords, which are words that are very common, such as prepositions, demonstrative, interrogative, and indefinite pronouns, do not provide useful information to distinguish the content of different documents. Stopword removal is accomplished by verifying if a word is contained in the stopwords hash table, which has been constructed by using several widely used stopword lists. Once the stopwords are removed, we proceed to stem all the remaining words using the Porter algorithm [Porter 80]. Quoting Martin Porter himself [Porter 80]: "The Porter stemming algorithm (or Porter stemmer) is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. Hereafter, we processed the remaining non-stop stemmed words[1] and created a *word-frequency-location file* in which each record consists of (i) one of the non-stop, stemmed words w, (ii) the document D (identified by document number) where w appears, (iii) the frequency of co-occurrence of w in D, and all the positions where w is present in D. (See Table 3.1 for portion of the word-frequency-location file.) The generated word-frequency-location file contains 144,048,788 records, and the number of non-stop, stemmed words in the file is 57,926, which means that the matrix contains 1,677,681,775 = 57,926 × (57,925 - 1) / 2 entries, since it is a symmetric matrix, and takes up 6.3Gb of disk space where each entry is stored as a

[1] From now on, unless stated otherwise, whenever we use the term "word," we really mean "non-stop, stemmed word".

Keyword	Document (D)	Frequency	Positions in D
computer	5	5	3, 21, 40, 32, 88
computer	12	3	14, 45, 100
computer	65	2	120, 176
⋮	⋮	⋮	⋮
comrade	1	2	30, 64
comrade	21	3	8, 30, 58

Table 3.1: Some keyword Frequency and Positions in the word-frequency-location file

32-bit floating point number. Once the word-frequency-location file is generated, we can calculate a word-word correlation matrix by considering either (i) the number of documents (i.e., Wikipedia articles) in which two distinct term w_1 and w_2 appear, which yields the keyword-connection matrix, (ii) the frequency of co-occurrence of w_1 and w_2 in each document, which yields the co-occurrence matrix, and (iii) the distance between w_1 and w_2 in each document, which yields the distance matrix.

3.2.1 Keyword Connection

The *(key)word-connection* correlation factor, which has been used for comparing similarity among documents, calculates the correlation of any two words w_1 and w_2 by computing the number of documents in a collection C where both w_1 and w_2 appear together [OMK 91, BYR 99]. In the keyword-connection matrix [OMK 91], each entry $m_{i,j}$, i.e., the *correlation factor*, for w_i and w_j is calculated as

$$c_{i,j} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}} \tag{3.1}$$

where n_i (n_l, respectively) is the number of documents in C in which the keyword w_i (w_l, respectively) appears, and $n_{i,l}$ is the number of documents in C in where both keywords appear. The correlation factors of different keywords computed by using the *keyword-connection* method follow the conjuncture that "The more documents in which two keywords occur, the more they relate to each other" [OMK 91].

The keyword-connection matrix is simpler to compute than the co-occurrence and the distance matrices; however, a major drawback of this simplicity is accuracy, since keyword-connection correlation factors do not consider other factors among different words, which include (i) the *frequency* of co-occurrence of any two words within a document and (ii) how *close* any two words appear together in a document. These factors are further considered by the co-occurrence frequency matrix (i.e., *association cluster* in [BYR 99]) and the distance matrix (i.e., *metric cluster* in [BYR 99]), respectively in computing the correlation factors of any two words.

3.2.2 Co-Occurrence Frequency

The *co-occurrence correlation factor* not only considers the number of documents in a collection where both words w_1 and w_2 appear, but it also considers the frequency of co-occurrence of both w_1 and w_2 in a document. In order to compute the co-occurrence factors, we obtain a frequency matrix m where each entry f_{i,d_u} is the frequency of word i in document d_u. The composition of m and its transpose, i.e., m^t, yields the matrix $c = m \cdot m^t$, where each entry $c_{i,j}$ of the matrix is the co-occurrence frequency factor of words w_i and w_j, i.e.,

$$c_{i,j} = \sum_{n=1}^{l} (f_{i,d_n} \times f_{j,d_n}) \tag{3.2}$$

where l is the total number of documents in a collection. We can normalize each correlation factor to limit the values in the interval [0, 1] as

$$c_{i,j_{norm}} = \frac{c_{i,j}}{c_{i,i} + c_{j,j} - c_{i,j}} \tag{3.3}$$

The frequency of co-occurrence of any two words in a document yields a better accurate correlation factor than the keyword-connection approach. Consider in an Wikipedia article about Shakespeare (http://en.wikipedia.org/wiki/ Shakespeare) in which *Shakespeare* appears sixty times, *english* appears twenty times, and *french* appears only one time. If we use the keyword-connection method, the correlation factor of *Shakespeare* and *english* will be the same as the correlation factor between *Shakespeare* and *french*. However, it is clear that the correlation factor between *Shakespeare* and *english* should be higher. Although the co-occurrence frequency matrix considers the co-occurrence frequency of any two words, it does not consider how close any two words are as they appear in a document, which is another important factor in providing an accurate correlation factor of any two words.

3.2.3 The Distance Matrix

The *distance correlation factor* between any two words w_1 and w_2 considers the frequency of occurrence, as well as the “distance” that is measured by the number of words, between w_1 and w_2 within a document as an additional factor to calculate the degree of correlation factor of w_1 and w_2. In the distance matrix approach, it is assumed that keywords which appear closer together are more likely related than those that appear far apart in the same document. For example, in a document that discusses computer architecture, the two words “computer” and “architecture” are likely to appear closer most of the times, and their correlation factor computed by using their distance correlation factor would be higher if their positions are considered, along with their frequency of co-occurrence. On the other hand, if a document discusses the usage of computers by architects for creating the blue prints of home designs, the keyword connection or co-occurrence correlation factor, could give the same correlation factor to “computer” and “architecture” in both documents, which is inaccurate, since it is clear that in the document talking about computer architecture, the correlation factor is higher between the two words, whereas in the document about architects using computers the correlation factor of the same two words should be smaller.

As mentioned earlier, the distance $d(w_i, w_j)$ between any two words w_i and w_j is defined by the (absolute) difference of the positions of any occurrence of w_i and w_j in a document, i.e., $d(w_i, w_j) = |Position(w_i) - Position(w_j)|$, and $d(w_i, w_j) = \infty$ when keywords w_i and w_j do not appear in the same document. For each document d_u in a collection C, the distance among each occurrence of keyword w_i and each occurrence of keyword w_j is calculated, and the correlation factor of w_i and w_j, i.e., $c_{i,j}$, is computed as the sum of the inverse of the distances between any occurrence of w_i and w_j:

$$c_{i,j} = \sum_{w_i \in V(s_i)} \sum_{w_j \in V(s_j)} \frac{1}{d(w_i, w_j)} \tag{3.4}$$

where $V(S_i)$ ($V(S_j)$, respectively), denotes the sets of words that include s_u (s_v, respectively) as its respective stemmed words. In order to normalize the distance correlation factors in the interval [0, 1] as in the other correlation matrices, $c_{i,j}$ is defined as

$$s_{i,j} = \frac{c_{i,j}}{|V(S_i)| \times |V(S_j)|} \tag{3.5}$$

Event though the calculation of the distance matrix is more complex than the keyword-connection and the co-occurrence matrices, the distance matrix is computed only once and captures the correlation factors of two words w_i and w_j more accurately than the other two, which will be verified in Section 3.3. We adopt the distance matrix

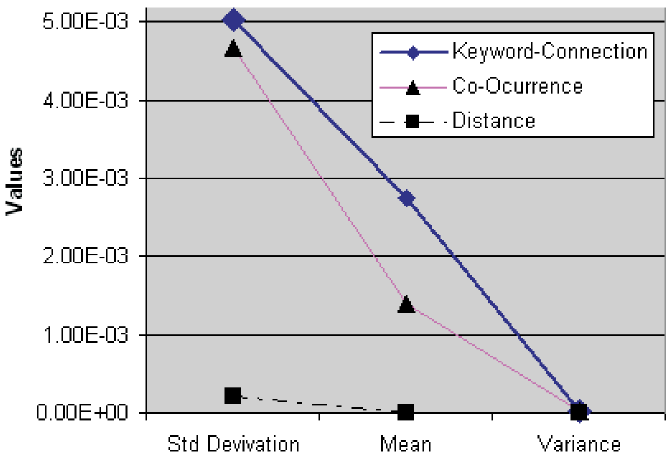


Figure 3.2: Derivation of the correlation factors in the three correlation matrices

.

approach in measuring the degrees of similarity among different words, which is used in the Fuzzy-Set IR method for detecting redundant or less-informative RSS news entries.

Although the three correlation factor matrices are calculated from the same set of Wikipedia documents, the correlation factors produced by each approach are not comparable, since the ranges of their correlation factors vary significantly as shown in Figure 3.2. The differences in their correlation factors ensures that each method yields different degrees of similarity S between any two RSS news articles when it is used in the Fuzzy-Set IR method to calculate S.

3.3 Word-Sentence-Document Fuzzy Association

Once a word-to-word correlation matrix is calculated, we can define a fuzzy set association to each word and a sentence, paragraph, or document itself. Since the news articles in RSS news feeds include only a brief summary, i.e., the 3-4 line summary, of an article in the Title and Description sections, we treat the Title and Description sections as the content of an RSS news article such that the degree of similarity between any two RSS news articles is determined by the correlation factors among the words in the respective Title and Description sections of the two articles.

The association between a keyword k_i and a document d_j (i.e., an RSS news article in this paper) referred in [OMK 91] as the word-sentence-document correlation factor $\mu_{i,j}$, is calculated as the complement of a negated algebraic product of all the correlations of the (key)word k_i and each (key)word k_l, where $k_l \in d_j$, i.e.,

$$\mu_{i,j} = 1 - \prod_{w_k \in d_j} (1 - c_{i,k}) \tag{3.6}$$

The correlation value $\mu_{i,j}$ falls in the interval [0, 1] and reaches its maximum when $c_{i,k} = 1$, for any $k \in d_j$. In [YNG 05], the $\mu_{i,j}$ factor is modified, which is different from the $c_{i,j}$ defined in [OMK 91], to compute the *word-sentence* correlation factor between word i and sentence j. Since each RSS news article contains no more than three short sentences (on the average in its RSS file as explained earlier), we treat them as a *single* sentence.

The degree of similarity of sentence S_i (in an RSS news article) with respect to sentence S_j (in another RSS news article), denoted as $Sim(i, j)$, is calculated as the average of all the values $\mu_{i,j}$, where $w_i \in S_i$, with respect to (the words in) S_j as

$$Sim(i, j) = \frac{\mu_{w_1,j} + \mu_{w_2,j} + \cdots + \mu_{w_n,j}}{n} \tag{3.7}$$

and $Sim(i, j) \in [0, 1]$. It is important to note that in general, $Sim(i, j) \neq Sim(j, i)$. When $Sim(i, j) = 0$, it indicates that there are no words in sentence S_i that can be considered similar to any word in sentence S_j. If $Sim(i, j) = 1$, then either sentence S_i is (semantically) identical to sentence S_j, or S_i is subsumed by sentence S_j, i.e., all the words in S_i are (semantically) the same as (some of) the words in S_j.

An *EQ* function of $Sim(i, j)$ and $Sim(j, i)$ is defined (below) in order to determine whether sentences S_i and S_j should be considered as (semantically) the same.

$$EQ(S_i, S_j) = \begin{cases} 1 & \text{if } Min(Sim(i,j), Sim(i,j)) \geq \alpha \\ & \quad \wedge \mid Sim(i,j) - Sim(i,j) \mid \leq \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

The parameter α is called the *permission threshold value*, and it determines the *minimum similarity* value for which S_i and S_j can be considered to be (semantically) the same. The second parameter ϵ is called the *variation threshold value*, which defines the *maximum difference* in terms of the degree of similarity, between S_i and S_j, for which S_i and S_j can be considered to be (semantically) the same. The values for α and ϵ were empirically obtained through experiments using a data set consisting of twenty-six RSS news articles obtained from different RSS news feeds, which include Associated Press (http://www.ap.com), Yahoo (http://news.yahoo.com), and Boston Globe (http://www.bostonglobe.com). The number of possible combinations of the matching news articles pairs is $26 \times 25 \div 2 = 325$. (We manually evaluated them and counted a total of 17 matching pairs among all the articles.) We considered the word-correlation factors in each of the different matrices, i.e., the keyword-connection, co-occurrence, and distance matrices that were previously created using the Wikipedia articles and tested for the degrees of similarity among the twenty-six articles. We counted (i) the number of pairs of RSS news entries that were considered *redundant* or *less-informative* when in fact they are not (i.e. *false positives*) and (ii) the number of pairs that were considered *different* but are in fact *redundant* or *less-informative* (i.e., *false negatives*).

The results we obtained by using each of the three matrices are shown in Table 3.2, which demonstrates that the correlation factors in the distance matrix provide the most accurate results among all the three matrices. The distance metric yields 3 false

Correlation Matrix	False Positive	False Negative	Matches	Accuracy %
Keyword-connection	2	9	8	47
Co-occurrence	1	8	9	52
Distance	3	1	16	**94**

Table 3.2: Experimental Results on twenty-six RSS news using the three different matrices

positives, 1 false negative, and 16 (out of 17) correctly matched pairs of news feeds, with the accuracy rate of 94%. These experimental results confirm the adoption of the distance matrix and the *EQ* function for detecting redundant or less-informative news articles in our proposed filtering method.

Chapter 4

A Fuzzy Equivalence Relation

By adopting the Fuzzy-Set IR model and the distance matrix, we can discard *redundant* RSS news articles, i.e., articles that do not provide new information. This task can be accomplished by eliminating any RSS news article s such that $Sim(s, t) = 1$, where t is another RSS news article, which implies that the information contained in article s is entirely subsumed by article t. Hereafter, we proceed to eliminate less-informative RSS news articles. This task can be accomplished by first generating clusters of all the RSS news articles that have certain degree of similarity. Clusters of news articles can be created by adopting a fuzzy equivalence relation that applies to the news articles to generate "crisp" subsets (i.e., clusters). News articles in each cluster that are "less-informative" are discarded. A cluster of RSS news articles is defined as $C_\alpha = \{d \mid Sim(d, e) \geq \alpha, \forall e \in C_\alpha\}$, where α is the *minimum degree of similarity* such that any two articles in C_α must hold. We generate clusters of non-redundant RSS news articles collected from various RSS news feeds by considering a fuzzy equivalence relation as given in [KSY 04]. A fuzzy equivalence relation defines a "crisp" equivalence relation among the elements of a set, which have been widely studied to measure the degree of similarity among different elements in a set. A (similarity) relation R is a *fuzzy equivalence relation* if it is reflexive, symmetric, and *max-min* transitive, i.e.,

$$R(x,x) = 1, \quad \forall x \in R \tag{4.1}$$

$$R(x,y) = R(y,x), \quad \forall x, y \in R \tag{4.2}$$

$$R(x,z) \geq \max_{y \in Y} \min\{R(x,y), R(y,z)\} \tag{4.3}$$

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

The key for establishing a fuzzy equivalence relation is the definition of transitivity. The first definition for fuzzy transitivity was proposed by Zadeh [ZAD 70], which is called the *max-min* transitivity, as defined in Equation 4.3. However, the *max-min* transitivity is known to be a restrictive constraint, which is not applicable to the similarity relation problem that we deal with. This is because in order to apply the *max-min* transitivity to our similarity problem, it is required that for any two articles (or documents) d_x and d_z, there cannot exist another article d_y whose similarities with both d_x and d_z is greater than the similarity between d_x and d_z, i.e., the relation R is not *max-min* transitive if given d_x and d_z, there exists d_y such that $R(d_x, d_z) < R(d_x, d_y)$ and $R(d_x, d_z) < R(d_y, d_z)$. Consider an example with the following sentences and the degrees of similarity $Sim_{1,2} = 0.20$, $Sim_{1,3} = 0.80$, $Sim_{2,1} = 0.33$, $Sim_{2,3} = 1.00$, $Sim_{3,1} = 0.50$, and $Sim_{3,2} = 0.37$:

S_1: *Bush's proposal will benefit illegal immigrants.*

S_2: *The war in Iraq is not over said Bush.*

S_3: *Bush proposed a bill that will benefit immigrants and he said the war in Iraq will continue.*

In order to establish a fuzzy equivalence relation that satisfies the *max-min* transitivity for S_1, S_2, and S_3, it is necessary to define a relation (function) R of the similarity values among S_1, S_2, and S_3 such that

$$f(0.20, 0.33) \geq \min\{f(0.80, 0.50), f(1.0, 0.37)\} \tag{4.4}$$

$$f(0.80, 0.50) \geq \min\{f(0.20, 0.33), f(1.0, 0.37)\} \tag{4.5}$$

$$f(1.00, 0.37) \geq \min\{f(0.80, 0.50), f(1.0, 0.37)\} \tag{4.6}$$

We consider a function that takes two input values, which are the similarity measures between two documents i and j and returns a high value if the two input values are high. The function f as shown in Equations 4.4, 4.5, and 4.6 does not satisfy this criteria, since f in each of the inequality equations yields a high value when the given values are low.

There exists another fuzzy transitivity relation, the *max-prod* transitivity [KSY 04] (as defined below), which can be adopted to establish a fuzzy equivalence relation.

$$R(x, z) \geq \max_{y \in Y}\{R(x, y) \times R(y, z)\} \tag{4.7}$$

where x, y, z, and Y are as defined in Equation 4.3.

The *max-prod* transitivity is not as restrictive as the *max-min* transitivity and can be more easily satisfied by a function R whose values fall in the interval $[0, 1]$, since the product of two numbers $x, y \in [0, 1]$ in the *max-prod* transitivity is smaller than x and y, i.e., if $x, y \in [0, 1]$, then $x \geq x \times y$ and $y \geq x \times y$. For this reason, we consider the *max-prod* transitivity instead of the *max-min* transitivity.

In order to adopt the fuzzy equivalence relation with *max-prod* transitivity for eliminating less-informative RSS news articles, it is necessary to define a function that combines the similarity measures of two news articles d_i and d_j, i.e., $Sim_{i,j}$, and $Sim_{j,i}$, into a single one. We consider several functions and choose the one that satisfies the *max-prod* transitivity property as the desired function.

4.1 Combination Functions

One of the most commonly used combination equations is *average*. However, the average of two pairs of different similarity values, e.g., (0.5, 0.5) and (0.9, 0.9), can yield the same result, e.g., (0.5 + 0.5)/2 = (0.9 + 0.1)/2. In addition, the average function is fuzzy symmetric and reflexive, but not *max-prod* transitive.

In [YNG 05], an $EQ(S_i, S_j)$ function of the similarity values of two documents d_i and d_j, i.e., $Sim_{i,j}$ and $Sim_{j,i}$, is defined (as given in Equation 3.8), which determines whether d_i and d_j should be considered as the same. The EQ function, which is a discrete function, assigns the values 1 or 0, and is symmetric and reflexive; however, EQ is neither *max-min* transitive nor *max-prod* transitive. Furthermore, one of the drawbacks of EQ is that two estimated threshold parameter values, i.e., α and ϵ, must be established before EQ can be adopted.

In [LUG 97], two combination equations that combine two values, e.g., $Sim_{i,j}$ and $Sim_{j,i}$, are defined as follows:

$$Q(Sim_{i,j}, Sim_{j,i}) = \frac{Sim_{i,j} + Sim_{j,i}}{1 - min(Sim_{i,j}, Sim_{j,i})} \tag{4.8}$$

$$Q_1(Sim_{i,j}, Sim_{j,i}) = (Sim_{i,j} + Sim_{j,i}) - (Sim_{i,j} \times Sim_{j,i}) \tag{4.9}$$

Both functions Q and Q_1 are simple to compute; however, Q_1 has the similar

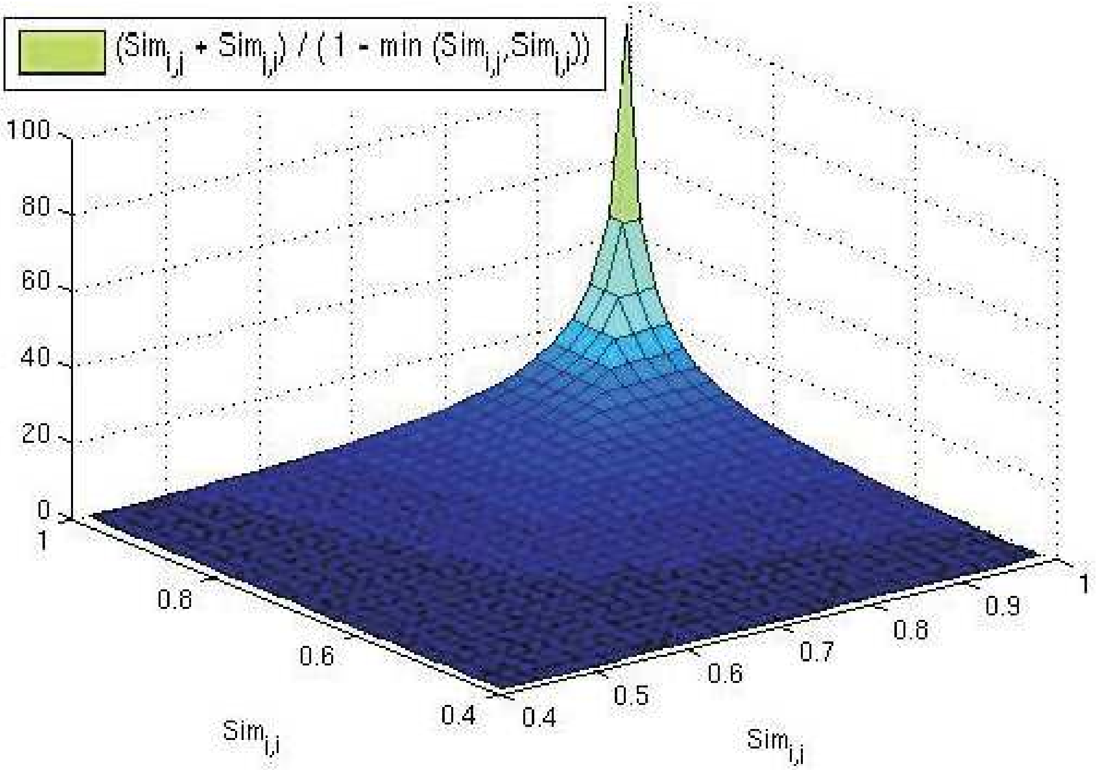


Figure 4.1: A sample of values computed by using the Q function in Equation 4.8.

drawback as the average function, i.e., it yields the same result to different pairs of values. For example, both (similarity) value pairs (0.9, 0.9) and (0.99, 0.1) are assigned the same value 0.99 by Q_1. In contrast, Q assigns a high value only when both similarity measures of two documents d_i and d_j are high, as shown Figure 4.1. Furthermore, function Q in Equation 4.8 is fuzzy-symmetric, however, Q is neither fuzzy-reflexive nor fuzzy-transitive. We modify Q so that the modified Q function, i.e., E, is a fuzzy equivalence relation.

$$E(d_i, d_j) = \begin{cases} 1 & \text{if } i = j \\ 0.0001 & \text{if } Q(d_i, d_j) < 0.0001 \\ \frac{Q(d_i,d_j)}{\max(Q)} & \text{otherwise} \end{cases} \tag{4.10}$$

The first condition in Equation 4.10 is introduced to satisfy *reflexivity*, whereas the third condition is the normalized Q function, which restricts the values of E to the interval [0, 1]. The second condition guarantees *max-prod transitivity*.

We have experimentally verified that the E function is *max-prod* transitive, using twelve sets of news articles from different RSS news feeds (see details in Table 4.1) and evaluating the *max-prod* transitivity inequality (as given in Equation 4.7) for each set of news articles. We consider every possible pair of news articles d_x and d_z in each RSS news feed and compute every possible $E(d_x, d_y) \times E(d_y, d_z)$ value to verify that $E(d_x, d_z) \geq E(d_x, d_y) \times E(d_y, d_z)$, for every d_y. The paramenter value 0.0001, which was obtained empirically using the twelve sets of RSS news articles, limits the similarity value between any two news articles so that the *max-prod* transitivity is satisfied. It is essential to understand that the value 0.0001 is not simply set up to satisfy the *max-prod* transitivity. In fact, given any two documents d_1 and d_2 that are similar, if there exists another document d_3 such that d_1 and d_3, as well as d_2 and d_3, are similar, then the *max-prod* transitivity using 0.0001 always holds for d_1, d_2, and d_3.

4.2 Clustering and Discarding Less-Informative News Articles

After the fuzzy equivalence relation E is established, we can apply E to determine the equivalence classes (clusters) of news articles from different RSS news feeds by setting an α-cut value [KSY 04]. As the value of α increases, the number of equivalence classes of the α-cut also increases, whereas the size of each equivalence class is reduced. (See Table 4.2 and Figure 4.2, as well as Figure 4.3 for an example.) An α-cut value guarantees that every pair of news articles in the same cluster has the degree of similarity no less than α. We discard one or more (but not all) news articles

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

Table 4.1: RSS news feeds used to verify the *max-prod* transitivity.

from each cluster and retain at least one news article in each cluster. We do not consider clusters with only one article, since they include the only news article which is dissimilar to other news articles in other clusters, assuming that our equivalence class generation approach is correct.

Since the same news articles may appear in different clusters[1], we cannot treat every cluster separately while selecting less-informative news articles to discard. Instead, we rank the news articles in different clusters generated by an α-cut value and discard those that have higher rankings, i.e., articles that are highly similar to others and thus are "less-informative," in the same cluster. While discarding news articles from different clusters, we cannot rank the news articles simply based on their sim-

[1] A fuzzy equivalence relation based on the *max-min* transitivity does not always yield disjoint equivalence classes, which is different from the fuzzy equivalence relation based on the *max-min* transitivity.

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1	0.2	1	0.6	0.2	0.6
x_2	0.2	1	0.2	0.2	0.8	0.2
x_3	1	0.2	1	0.6	0.2	0.6
x_4	0.6	0.2	0.6	1	0.2	0.8
x_5	0.2	0.8	0.2	0.2	1	0.2
x_6	0.6	0.2	0.6	0.8	0.2	1

Table 4.2: A fuzzy equivalence relation

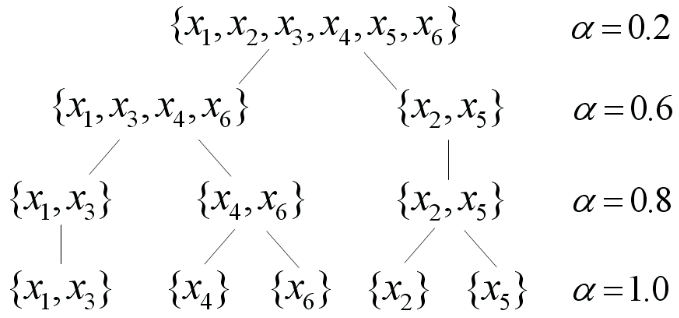


Figure 4.2: Different α-cuts applied to a set of elements that yield different number of clusters (i.e., equivalence classes).

ilarity values, since we could discard all the news articles in a cluster. Consider the similarity matrix for a set of documents $C = \{a, b, c, d, e, f\}$ as shown in Table 4.3 and the E values computed by using Equation 4.10 on the similarity values in the matrix, i.e., $E(a, b) = 0.055$, $E(c, d) = 0.021$, $E(e, f) = 0.014$, and the same E value, 0.0025, for all the other possible pairs. If we set $\alpha = 0.1$, then three clusters $C_1 = \{a, b\}$, $C_2 = \{c, d\}$, and $C_3 = \{e, f\}$ are generated. Suppose we need to discard two news articles. If we rank the news articles simply based on the similarity values and discard the first two with highest similarity values, then we would discard articles a

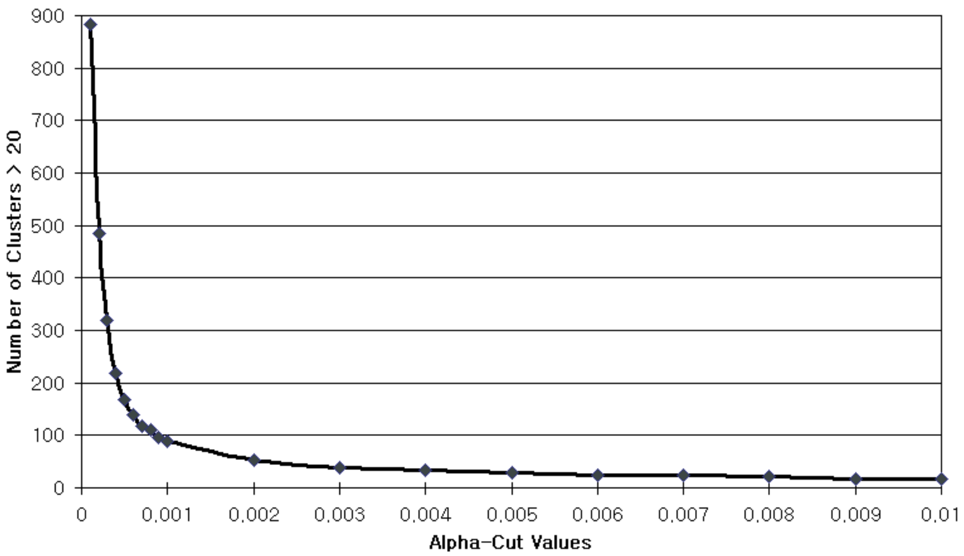


Figure 4.3: Numbers of equivalence classes generated by using different α-cut values on a sample data set.

and b in C_1, which have the highest similarities values as shown in Table 4.3.

4.3 Different Ranking Approaches

We compare three different approaches in ranking the news articles in different clusters generated by an α-cut value. In the first approach, we consider the *entropy value* of each news article. The entropy in information theory is an alternative way to describe how much information is carried. We measure the entropy of each news article by using

$$H(x) = -\sum_{i=1}^{N} p(i) \log_2 p(i) \tag{4.11}$$

where x represents a news article and N is the number of clusters to where x belongs.

We have chosen the Q function in Equation 4.8 to compute probability p in Equation 4.11 for each news article. Table 4.4 shows the ranking according to the entropy values of a set of news articles (collected from various RSS news feeds) in different

	a	b	c	d	e	f
a	1	**0.80**	0.2	0.2	0.2	0.2
b	**0.85**	1	0.2	0.2	0.2	0.2
c	0.2	0.2	1	**0.60**	0.2	0.2
d	0.2	0.2	**0.70**	1	0.2	0.2
e	0.2	0.2	0.2	0.2	1	**0.61**
f	0.2	0.2	0.2	0.2	**0.50**	1

Table 4.3: Similarity among a set of documents.

clusters generated by an α-cut value. By using the number of occurrences of each news article in different clusters, we realize that there is a clear bias towards those articles that have higher frequency of occurrences, which is not a desired behavior, since favoring news articles that appear several times in different clusters are not necessarily the ones that have the higher degree of similarity with a particular news article than the remaining news articles. We have considered another ranking, which is the *average* of the E values of news articles. The average includes all the E values obtained between a document d_i and each of the other documents d_j that appear in each of the clusters to where d_i belongs. However, the average of the E values does not yield a good ranking. The average of the E values assigns a low ranking to articles 41 and 36. Article 41 should have a high ranking and be discarded since it is very similar to article 28, and article 36 should also receive a high ranking, since (i) it appears in several clusters and (ii) has high similarity values with those news articles in the clusters. Instead of adopting the entropy or the average E values, we use the function M in Equation 4.12 to rank news articles in all the clusters that include at least two news articles generated by an α-cut value. M provides the *average* of the maximum similarity values of a news article d_i with another news article d_j in each

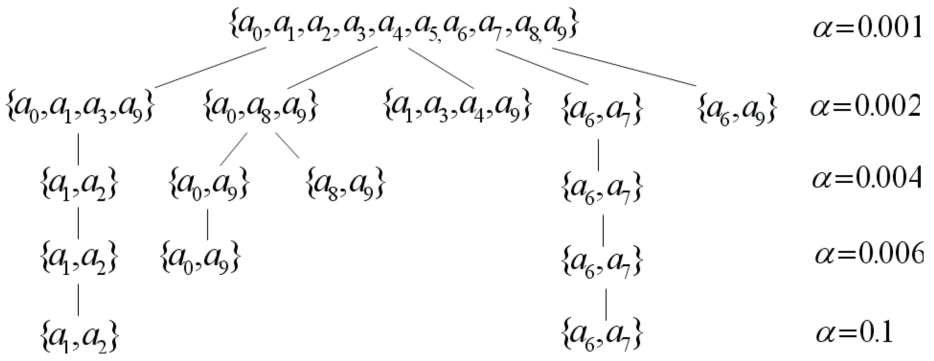


Figure 4.4: Clusters generated on a set of 10 RSS news articles as shown in Table 4.5 according to different α-cut values.

cluster where d_i appears.

$$M(d_i) = \frac{\sum_{k=1}^{N} \max_{C_k}\{Sim(i,j)\}}{N} \tag{4.12}$$

The ranking based on the average of H values considers the trade off between the (i) the frequency of a news article in clusters generated by an α-cut value and (ii) the similarity values with respect to other news articles in various clusters, which yields the most accurate ranking on news articles according to the conducted experiments (e.g., Table 4.4). According to the rankings, the top n ($n \geq 1$) ranked (i.e., less-informative) news articles, along with all the redundant ones that have been detected earlier, are discarded, assuming that their frequencies of occurrence is greater than 1, i.e., they occur in non-singleton clusters.

Example 1 Consider a set of ten RSS news articles that were extracted from various RSS news feeds as shown in Table 4.5 and the non-singleton clusters that were generated as shown in Figure 4.4. The clusters and the computed rankings of the ten articles can be used to determine which one(s) of the ten articles should be discarded.

It is important to point out that the number of "less-informative" news articles to be discarded can be determined by (i) the average number of new articles that are accessed by an individual user, or (ii) the number of articles posted by each individual RSS news feed that the user accesses on a regular basis.

Average M			**Entropy**			**Average** E		
Article#	Freq.	Rank	Article#	Freq.	Rank	Article#	Freq.	Rank
204	1	0.93	36	4	0.27	222	1	0.06
41	1	0.93	67	3	0.24	218	1	0.06
324	1	0.91	139	5	0.23	415	1	0.04
218	1	0.88	271	3	0.19	15	1	0.04
299	2	0.87	137	1	0.15	204	1	0.02
101	1	0.87	101	1	0.14	276	1	0.02
116	1	0.86	299	2	0.13	324	1	0.02
58	1	0.85	112	2	0.12	180	1	0.02
160	1	0.85	27	2	0.12	281	1	0.02
415	1	0.85	222	1	0.12	99	1	0.01
281	1	0.83	218	1	0.12	129	1	0.01
179	1	0.80	204	1	0.11	132	1	0.01
40	1	0.80	276	1	0.11	137	1	0.01
132	1	0.80	6	2	0.11	101	1	0.01
70	1	0.80	49	1	0.10	350	1	0.01
85	1	0.77	172	2	0.10	179	1	0.01
36	4	0.76	89	2	0.10	89	2	0.01
106	1	0.73	185	2	0.10	185	2	0.01
271	3	0.72	324	1	0.10	49	1	0.01
67	3	0.71	415	1	0.09	299	2	0.01
78	1	0.70	15	1	0.09	51	1	0.01
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.4: Article rankings using Equations 4.12, 4.11, and 4.10.

ID	Title and URL
a_0	*House Considers Bill to Boost Refineries* http://abcnews.go.com/Business/wireStory?id=1191725&amp
a_1	*Intruder Gains Front Lawn of White House* http://abcnews.go.com/Politics/wireStory?id=1823761
a_2	*Screaming Intruder Jumps White House Fence* http://www.examiner.com/a-73142
a_3	*Saudi Ambassador: Iraq Invasion Helped Spread Terrorism* http://www.foxnews.com/story/0,2933,178188,00.html
a_4	*Bush says Iraqi parliamentary elections "a major step forward"* http://english.people.com.cn/200512/16/eng20051216
a_5	*Powerball Winners Donate to Homeless* http://www.foxnews.com/story/0,2933,190670,00.html
a_6	*mSleep disorders affect millions of Americans* http://english.people.com.cn/200604/05/eng20060405_256141.html
a_7	*Millions of Americans suffer sleep disorders* http://news.xinhuanet.com/english/2006-04/05/content_4386829.htm
a_8	*Gasoline price drop may only be temporary* http://www.boston.com/news/nation/washington/articles/2005/09/19/
a_9	*House considers bill to boost refineries* http://abcnews.go.com/Business/wireStory?id=1191983&page=1

Table 4.5: Ten RSS news articles downloaded from various RSS news feeds and their URLs

Chapter 5

Experimental Results

In this chapter we present experimental results to verify the accuracy of our approach for detecting less-informative and redundant RSS news articles collected from one or more RSS news feeds. We analyzed 25 test sets each of which includes one or more RSS news feeds. The accuracy of our detection approach is determined by the number of false positives, i.e., the number of news articles that were mistakingly discarded in each set. The number of false positives is then converted into a percentage of accuracy. We also provide an analysis of the importance of the parameter α, which etermines the number of clusters generated during the process of selecting less-informative news articles.

5.1 The Significance of α-Cut Values

The value of α plays a key role, along with the ranking function, to select which news articles should be discarded. The value of α determines the appropriate number of clusters to be generated, in addition to the number of clusters in which an article is contained. In Figure 5.1, we observe that the number of clusters (of size ≥ 2) increases as the value of α decreases. Also, as the number of clusters increases, the

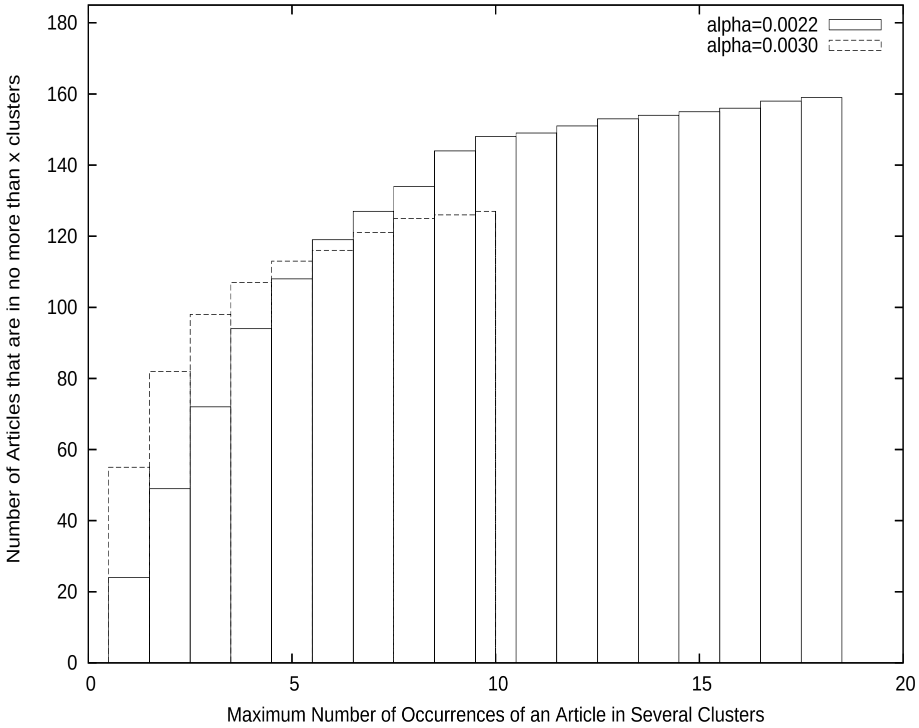


Figure 5.1: Accumulated Histogram of the frequency of cluster appearance of news articles. The box values indicate the maximu number of clusters that an article may belong to.

number of clusters to where an article belongs also increases, which occurs when the number of clusters is greater than half of the total number of news articles.

Figure 5.1 shows the accumulated histogram of the frequency of cluster appearance of news articles for a test set that contains 162 news articles from four different RSS news feeds: `http://www.cbsnews.com/track/rss/sections/world/main202.shtml??source=RSS`, `http://www.iht.com/pages/africa/index.ph`, `http://www.seattletimes.com`, and `http://www.usatoday.com/news/world/digest.htm`. When $\alpha = 0.00013$, the number of clusters is larger than the number of articles, and therefore most articles appear in at least five different clusters. It turns out that if α is smaller, the average size of the clusters is bigger. On the other hand, for a bigger value of alpha ($\alpha = 0.003$), The number of clusters are reduced from 362 to 108 and most of the news articles appear in only one cluster.

Since a smaller α value generates more clusters than the number of news articles in a collection, we decrease the probability of eliminating a cluster entirely, during the process of discarding redundant or less-informative news articles, However, the cluster generated may be loosely related if α is too small.. We often favor larger α values, as it turns out a larger α value generates less and smaller clusters in which articles are closely related.

In our detection approach, clusters form groups of news articles that are α-related. We assume that each cluster of news is created independent of the cluster of other news, e.g., a cluster may contain news articles related to the war in Iraq, and another cluster consist of news articles on the celebration of the new year. Disallowing all the news articles to be removed from any cluster, we can eliminate redundant and less-informative from each cluster and still retain at least one news article that represents the main "topic" in the cluster. news regarding to the main "topic" of the news in each cluster. We used a subset of 34 articles from the RSS news feeds mentioned earlier to show the differences in the clusters generated for 2 different values of α . Table 5.2 shows the clusters generated when $\alpha = 0.00011$, and half of the total number of articles were discarded. Table 5.1 shows the clusters for $\alpha = 0.0022$. As it turns out, both values of α yield a good result, however, for the smaller α article number 18 is a false positive, and it should have been kept since it is quite different from the other articles.

5.2 Detection of Redundant RSS New Articles

During the process of gathering different test sets, we observe that many news portals release RSS files which contain repeated news articles, or articles that provide updated information of a recent event. Hence, it is common that the text contained in the title

Articles in Original Cluster	After Elimination	Articles in Original Cluster	After Elimination
0 1 2 4	0 1	3 5 6	5 6
0 6	0 6	6 7	6 7
0 2 4 8	0 8	7 8 25	7 8
0 9	0 9	15 16	15 16
12 19	12 19	20 23	23
2 4 18	18	3 9	9
4 12	12	6 14 21	6
4 15	15	1 17	1 17
0 1 25	0 1	24 25 26 27 28 29 30 31 32 33	28
8 11 22	8	8 25 26	8
14 28	28		

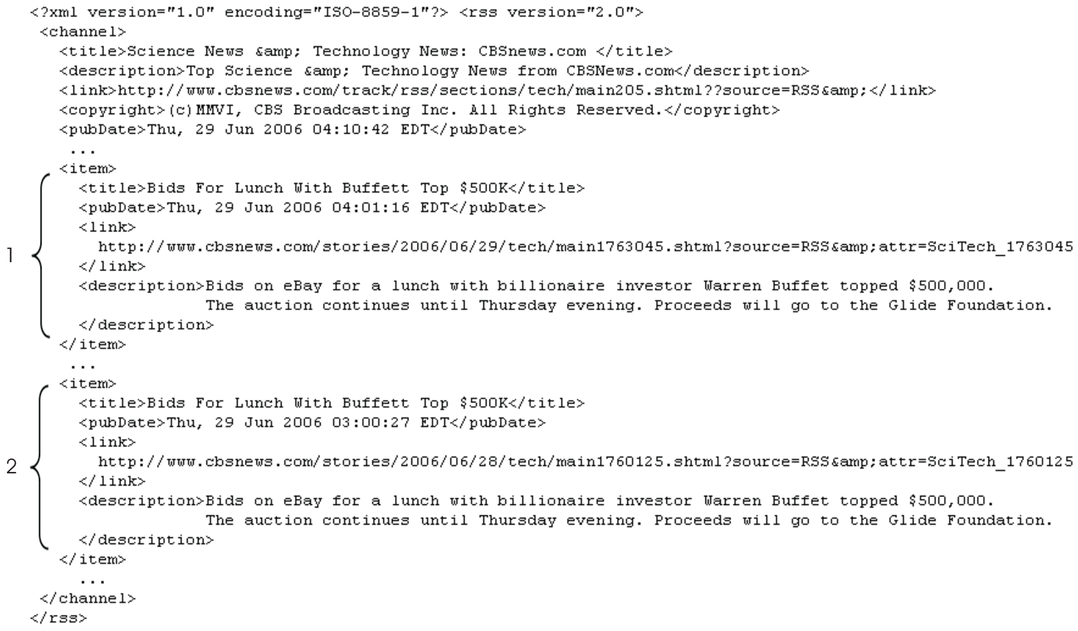
Table 5.1: Clusters before and after discarding 50% of news articles, $\alpha = 0.0022$.

and description tags of two news articles are identical or very similar, even though the URL links are different (see news articles 1 and 2 in Figure 5.2). As a result, the content of the title and description tags of an RSS news article could be used as a unique identifier during the process of detecting redundant news articles.

It's not uncommon to find duplicated news articles either within a single RSS feed, or an identical news article in different RSS news feeds. During the process of calculating the similarity among the documents, if the system detects a duplicated news article, i.e., if $Sim(i,j) = 1 = Sim(j,i)$, or if $Sim(i,j) = 1$ and $Sim(j,i) < 1$, the news article i is redundant with respect of news article j and it is discarded immediately; prior to the clustering process. In Table 5.3 we have two pairs of duplicated (redundant) articles {120,135} and {1,72}. Also, article #162 is subsumed in article #154, i.e., $Sim(162, 154) = 1$ and is discarded too. Articles #4 and #99, are similar, but article #4 which is less-informative than article #99, is discarded and article #99 is kept. Article #96 is subsumed in article #97 and is discarded before

Articles in Original Cluster	After Elimination	Articles in Original Cluster	After Elimination
0 8 22 24 26	0	0 8 24 25 26 31	0
0 11 25 29 30 31	0 11	0 4 6 8 24 31	0 4
0 4 6 9 15 21	0 4 9	0 1 2 3 6 7 8 11 25 31	0 1 3 7 11
0 6 9 11 19 21 25	0 9 11 19	0 1 2 17 25	0 1 17
0 1 2 3 4 5 6 7 8 31	0 1 3 4 5 7	0 2 3 4 6 8 9 11	0 3 4 9 11
0 1 2 4 17 18	0 1 4 17	0 3 6 7 11 19 25	0 3 7 11 19
0 6 9 15 19 21	0 9 19	0 4 17 20	0 4 17 20
0 7 11 19 20	0 7 11 19 20	0 2 4 8 9 11 18 22	0 4 9 11
0 4 7 8 11 20	0 4 7 11 20	0 1 2 3 4 6 7 8 11 31	0 1 3 4 7 11
2 3 8 25 26 27 31	3	3 5 6 7 28 31	3 5 7
3 5 7 28 30 31	3 5 7	3 10 11 27 30	3 10 11
2 3 4 8 9 10 11 27	3 4 9 10 11	2 4 8 16	4 16
4 15 16	4 16	4 8 16 24	4 16
2 4 17 18 27	4 17	4 6 13 24	4 13
4 9 12 15 21	4 9 12	2 4 6 8 11 31 32	4 11 32
4 13 20 24	4 13 20	2 4 7 8 11 12	4 7 11 12
2 4 8 9 10 11 12 18 22 27	4 9 10 11 12	4 8 10 11 20 22 32	4 10 11 20 32
6 7 11 14 19 21 25 28	7 11 19	7 11 23 28	7 11 23
7 8 11 20 23	7 11 20 23	7 11 19 20 23	7 11 19 20 23
6 9 11 19 21 25 28	9 11 19	8 9 10 11 22 23	9 10 11 23
9 11 12 19 21	9 11 12 19	11 12 27 29	11 12
23 28 32	23 32	24 25 26 27 28 29 30 31 32 33	32

Table 5.2: Generated Clusters before and after discarding 50% of news articles. $\alpha = 0.00011$



```
<?xml version="1.0" encoding="ISO-8859-1"?> <rss version="2.0">
 <channel>
   <title>Science News &amp; Technology News: CBSnews.com </title>
   <description>Top Science &amp; Technology News from CBSNews.com</description>
   <link>http://www.cbsnews.com/track/rss/sections/tech/main205.shtml??source=RSS&amp;</link>
   <copyright>(c)MMVI, CBS Broadcasting Inc. All Rights Reserved.</copyright>
   <pubDate>Thu, 29 Jun 2006 04:10:42 EDT</pubDate>
    ...
   <item>
     <title>Bids For Lunch With Buffett Top $500K</title>
     <pubDate>Thu, 29 Jun 2006 04:01:16 EDT</pubDate>
     <link>
       http://www.cbsnews.com/stories/2006/06/29/tech/main1763045.shtml?source=RSS&amp;attr=SciTech_1763045
     </link>
     <description>Bids on eBay for a lunch with billionaire investor Warren Buffet topped $500,000.
                  The auction continues until Thursday evening. Proceeds will go to the Glide Foundation.
     </description>
   </item>
    ...
   <item>
     <title>Bids For Lunch With Buffett Top $500K</title>
     <pubDate>Thu, 29 Jun 2006 03:00:27 EDT</pubDate>
     <link>
       http://www.cbsnews.com/stories/2006/06/28/tech/main1760125.shtml?source=RSS&amp;attr=SciTech_1760125
     </link>
     <description>Bids on eBay for a lunch with billionaire investor Warren Buffet topped $500,000.
                  The auction continues until Thursday evening. Proceeds will go to the Glide Foundation.
     </description>
   </item>
     ...
 </channel>
</rss>
```

Figure 5.2: Redundant articles in an RSS News Feed. Identical titles and descriptions, different links

the clustering process. Article #86 is considered in the clustering process, however, it is less-informative than article #97 and thus article #86 is eliminated.

5.3 The Degree of Accuracy of Our Detection Approach

We have considered 25 different test sets, each of which contains articles from one or several RSS news feeds. We manually evaluated the articles chosen to be discarded and counted the number of False Positives. We calculated the percentage of accuracy using the formula:

$$1 - \frac{False\ Positives}{Discarded}$$

The twenty-five test sets provide a high degree of accuracy, even when the percentage of articles discarded is high. The percentages of articles to be discarded are considered are 10%, 20%, 30%, 40% and 50%. Figure 5.3 shows the results when 50% of the articles in the 25 test sets are discarded. Table 5.4 shows the number of false positives for all the different percentages of discarded articles. Table 5.5 shows the false positive values converted in accuracy percentages.

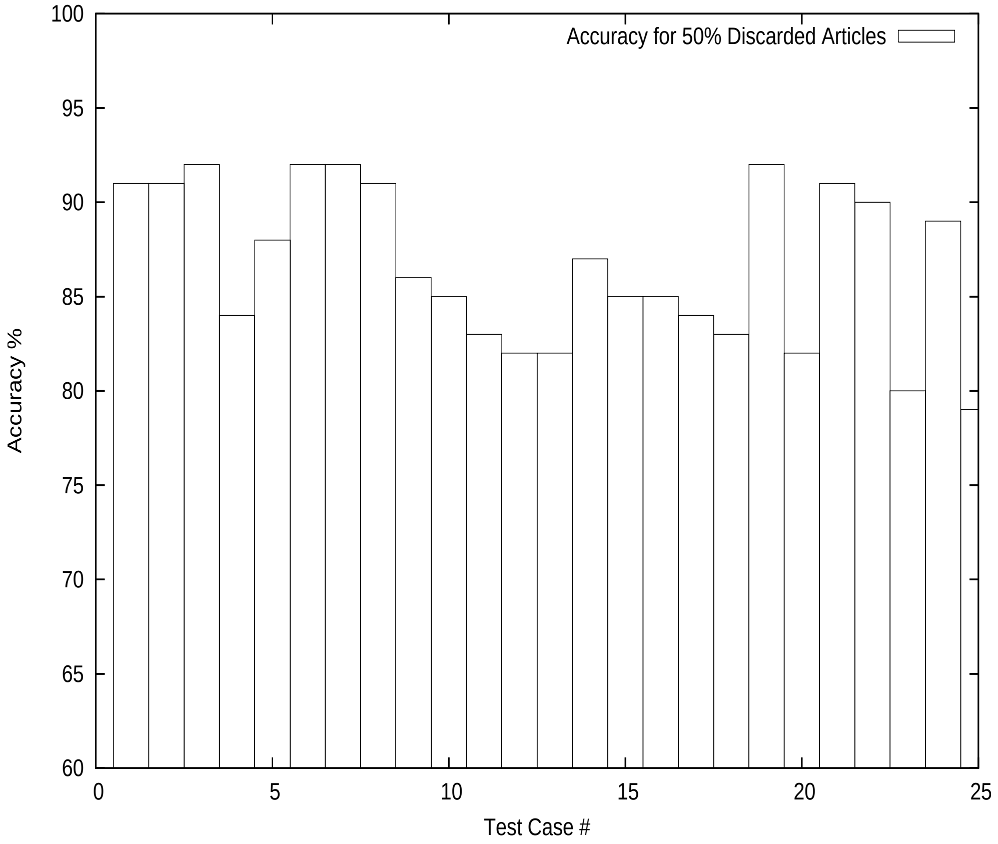


Figure 5.3: Accuracy percentages for 25 test sets

Id	Title and Description
120	Official: Missing Indonesian Jet Not Found The wreckage of an Indonesia jet that went missing has not been found despite
125	Official: Missing Indonesian Jet Not Found The wreckage of an Indonesia jet that went missing has not been found despite
1	Thailand bombings Deadly New Year's bombings failed to shatter Bangkok's charms for many visitors . . .
72	Thailand bombings Deadly New Year's bombings failed to shatter Bangkok's charms for many visitors . . .
154	Survivors Found Days After Ferry Sinks Survivors found days after sinking of ferry in Indonesia; 400 people still missing
162	Survivors Found Days After Ferry Sinks Survivors found days after sinking of ferry in Indonesia.
4	AFP photographer latest journalist abducted in Gaza AFP photographer was abducted when ...
99	Peruvian journalist is seized in Gaza Palestinian gunmen kidnapped a photographer working for Agence France-Presse in Gaza, while several militants were seized in separate abductions that ignited . . .
86	U.S. deaths in Iraq reach 3,000 The rising toll of U.S. troops killed in Iraq hit another grim milestonE
92	U.S. military fatalities in Iraq reach 3,000 The number reflects how much more dangerous a soldier's job in Iraq has become in the face of a . . .
97	U.S. military fatalities in Iraq reach 3,000 The number reflects how much more dangerous and muddled a soldier's job in Iraq has become in the face of a growing and increasingly sophisticated insurgency.

Table 5.3: A number of news articles which include two pairs of identical news, pairs {1,72} and {120,125}. Articles #162 and #92 are subsumed in articles #154 and #97 respectively. Articles #4 and #86 are less-informative than articles #99 and #97 respectively.

Topic	source	Art	10%	20%	30%	40%	50%
World	3	65	0	1	3	3	3
Techno	2	64	0	0	2	3	3
USA	3	53	1	1	2	2	2
World	2	51	0	0	2	3	4
USA	2	50	2	2	2	2	3
Sports	2	50	0	1	1	2	2
USA	2	49	0	0	1	1	2
World	2	46	1	1	1	2	2
Sports	2	44	1	1	2	2	3
World	2	40	0	1	2	3	3
USA	3	36	1	1	1	2	3
World	2	34	1	1	1	2	3
Entert	3	33	0	1	1	2	3
World	1	30	0	0	1	2	2
Entert	2	27	0	0	1	1	2
World	3	26	0	1	1	2	2
USA	1	25	1	1	1	1	2
World	1	24	0	1	2	2	2
Sports	1	24	0	0	0	1	1
USA	1	22	0	0	1	2	2
Entert	2	22	0	0	1	1	1
USA	1	20	0	0	0	1	1
World	1	20	2	2	2	2	2
USA	1	19	0	0	1	1	1
World	1	19	0	1	1	1	2

Table 5.4: False Positives for 25 test sets

Test Set	Topic	Sources	# of Articles	10%	20%	30%	40%	50%
1	World	3	65	100%	92%	84%	88%	90%
2	Techn	2	64	100%	100%	89%	88%	90%
3	USA	3	53	81%	90%	87%	90%	92%
4	World	2	51	100%	100%	86%	85%	84%
5	USA	2	50	60%	80%	86%	90%	88%
6	Sports	2	50	100%	90%	93%	95%	96%
7	USA	2	49	100%	100%	93%	94%	91%
8	World	2	46	78%	89%	85%	89%	86%
9	Sports	2	44	77%	88%	92%	88%	90%
10	World	2	40	100%	87%	83%	81%	85%
11	USA	3	36	72%	86%	90%	93%	94%
12	World	2	34	70%	85%	90%	85%	88%
13	Entert	3	33	100%	84%	89%	84%	81%
14	World	1	30	100%	100%	88%	83%	86%
15	Entert	2	27	100%	100%	87%	90%	85%
16	World	3	26	100%	80%	87%	80%	84%
17	USA	1	25	60%	80%	86%	90%	84%
18	Sports	1	24	100%	100%	100%	89%	91%
19	World	1	24	100%	79%	72%	79%	83%
20	Entert	2	22	100%	100%	100%	88%	90%
21	USA	1	22	100%	100%	84%	77%	81%
22	USA	1	20	100%	100%	100%	87%	90%
23	World	1	20	0%	50%	66%	75%	80%
24	USA	1	19	100%	100%	100%	100%	89%
25	World	1	19	100%	94%	98%	98%	78%

Table 5.5: Accuracy Percentage for 25 test cases.

Chapter 6

Conclusions

In this thesis, we have proposed a solution to the RSS information overflow problem by introducing a filtering approach which selectively eliminates redundant or less-informative RSS feeds entries. Our filtering method has been proved to efficiently detect redundant or less informative news articles using a heterogeneous set of test cases. The average accuracy rate in detecting redundant and less-informative news articles is close to 90%. Thus, we conclude that our method provides a framework for selectively filtering undesirable and exceeding amount of news articles, which is accomplished by ordering news articles according to their degrees of similarity with other news articles. News articles that are highly similar or embedded in other news articles are discarded.

The proposed filtering approach, which adopts the Fuzzy-Set information retrieval (IR) model, a distance matrix, and a fuzzy equivalence relation with *max-prod* transitivity, has a solid theoretical and mathematical foundation, since it is developed using the fuzzy-set theory, equivalence relations, and other uncertainty measures such as the Dempster-Shafer Theory of Evidence. Furthermore, the proposed method is flexible in terms terms of the number of news articles to be eliminated regardless the number of RSS news feeds and the total number of news articles involved in the filtering process. Our filtering approach can easily be adopted for load shedding of data streams in sensor networks.

As the number of RSS news feeds continue to increase over the Internet, which translate into significant more news articles to become available to the Web users, our proposed method in filtering redundant and less-informative news articles would only become more significant in minimizing the workload of the end user who must scan through huge amount of news articles to retrieve unique news otherwise.

Bibliography

[BYR 99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[BDG 95] S. Brin, J. Davis, and H. Garcia-Molina, *Copy Detection Mechanisms for Digital Documents.* In Proceedings of ACM SIGMOD, pp. 398-409, 1995.

[CHI 05] S. Chien and N. Immorlica. *Semantic Similarity Between Search Engine Queries Using Temporal Correlation.* In Proceedings of the WWW Conference, pp. 2-11, 2005.

[KSY 04] G.K. Klir, U. St. Clair, and B. Yuan. *Fuzzy Set Theory, Foundations and Applications*, 1997.

[GDH 04] E. Gabrilovich, S. Dumais, and E. Horvitz. *Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty.* In Proceedings of World Wide Web, pp. 482-490, 2004.

[HKK 04] K.M. Hammouda and M.S. Kamel. *Efficient Phrase-Based Document Indexing for Web Document clustering.* IEEE TKDE, 16(10), pp. 279-1296, 2004.

[LUG 97] G.K. Luger. *Artificial Intelligence, Structures and Strategies for Complex Problem Solving.* 1997.

[MAN 94] U. Manber. *Finding Similar Files in Large File System.* In USENIX Winter Technical Conference, 1994.

[NH 96] H. Nevin. *Scalable Document Fingerprinting.* In Proceedings of the 2^{nd} USENIX Workshop on Electronic Commerce, pp. 191-200, 1996.

[OMK 91] Y. Ogawa, T. Morita, and K. Kobayashi. *A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method.* Fuzzy Sets and Systems, Vol. 39, pp. 163-179, (1991).

[Porter 80] M. Porter. An Algorithm for Suffix Stripping. Program, 14(3): 130-137 (1980).

[SGM 95] N. Shivakumar and H. Garcia-Molina. *The SCAM Approach to Copy Detection in Digital Libraries.* Diglib Magazine, November 1995.

[WIKI 05] http://wikipedia.org/.

[YNG 05] R. Yerra and Y-K. Ng. *Detecting Similar HTML Documents Using a Fuzzy Set Information Retrieval Approach.* In Proceedings of IEEE International Conference on Granular Computing (GrC'05), pp. 693-699, 2005.

[ZK 05] Y. Zhao and G. Karypis. *Topic-driven Clustering for Document Datasets.* In Proceedings of SIAM International Conference on Data Mining, pp. 358-369, 2005.

[ZAD 70] L.A. Zadeh. *Similarity relations and fuzzy orderings.* Information Sciences, Vol. 3, pp. 177-200, 1970.