Bachelor's thesis (UAS)

Information Technology

Text Mining and Clustering

2013

Prabin Lama

CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K-MEANS ALGORITHM

– News headlines clustering

TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR´S THESIS | ABSTRACT
TURKU UNIVERSITY OF APPLIED SCIENCES

Information Technology

December 2013| 45 pages

Instructor: Patric Granholm

Prabin Lama

CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K-MEANS ALGORITHM

The increasing scope of the web and the large amount of electronic data piling up throughout the web has provoked the exploration of hidden information from their text content.

News articles published on different news portals throughout the web are the sources of the information.These can also be very good topics for the research on text mining. Clustering of similar news headlines and putting them under a single platform with the corresponding links to the news portal sites can be a very efficient option to the exploration of the same news article across multiple different news portals, which is, in fact, a tedious and time-consuming task.

This thesis presents the model which analyzes the news headlines across the different news portals, uses document pre-processing techniques and creates clusters of similar news headlines.

Data available on the web are structured, semi-structured or unstructured. Webpages are usually semi-structured because of the presence of html tags. The XML representation of semi-structured data facilitates the clustering of similar documents by the use of distance-based clustering techniques.

News headlines from different news portals are extracted and stored in an XML file. The XML file is then preprocessed using document preprocessing techniques. Techniques like tokenization, stop word removal, lemmatization and synonym expansion are used during the document preprocessing. The selected news headlines are then represented using the vector space modeling and term-frequency weighting scheme. Finally, the K-means clustering algorithm is applied to find similarities among the news headlines and create clusters of similar news headlines. A sample webpage is used to display the clusters of the news headlines with their corresponding links.

KEYWORDS:

Information extraction, document pre-processing, document clustering, K-means, news article headlines

CONTENTS

APPENDICES

FIGURES

TABLES

LIST OF ABBREVIATIONS (OR) SYMBOLS

FCM	Fuzzy C-Means
HTML	Hypertext Markup Language
IE	Information Extraction
LSI	Latent Semantic Indexing
NLP	Natural Language Processing
VSM	Vector Space Model
XML	Extensible Markup Language

1 INTRODUCTION

Along with the development of new smart technologies, the world is going digital. Because of this, the web is growing day by day and has become a huge source of the information. Looking up for the precise and relevant information and extracting it from the web has now become a time-consuming task. There are many techniques used for the information extraction from the web and text mining is one of them.

This thesis entitled "**Clustering System based on Text Mining using the K means algorithm,**" is mainly focused on the use of text mining techniques and the K means algorithm to create the clusters of similar news articles headlines. The project study is based on text mining with primary focus on data-mining and information extraction. The news headlines and the links to the different news portal are fetched via an XML file to the clustering system. The news headlines within the XML file are then preprocessed using document preprocessing techniques and finally grouped in the clusters based on their similarities. These clusters are displayed in a sample webpage with the corresponding links to the news portal sites.

2 LITERATURE REVIEW AND BACKGROUND

Existing works on web text mining and clustering are mainly focused on the different levels like: Web text clustering, Data text mining, Web page information extraction etc.

Web data clustering researchers Bouras and Tsogkas (2010) proposed an enhanced model based on the standard k-means algorithm using the external information extracted from WordNet hypernyms in a twofold manner: enriching the “bag of words” used prior to the clustering process and assisting the label generation procedure following it.

Murali Krishna and Durga Bhavani (2010) proposed the use of a renowned method, called Apriori algorithm, for mining the frequent item sets and devised an efficient approach for text clustering based on the frequent item sets. Maheshwari and Agrawal (2010) proposed centroid-based text clustering for preprocessed data, which is a supervised approach to classify a text into a set of predefined classes with relatively low computation and better accuracy.

Qiujun (2010) proposed a new approach to news content extraction using similarity measure based on edit distance to separate the news content from noisy information. Jaiswal (2007) performed the comparison of different clustering methods like K-Means, Vector Space Model (VSM), Latent Semantic Indexing (LSI), and Fuzzy C-Means (FCM) and selected FCM for web clustering.

2.1 Text Mining

Text mining, also known as text data mining or knowledge discovery process from the textual databases, generally, is the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. All the extracted information is linked together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

The classic approach of information retrieval based on keyword search from WWW makes it cumbersome for the users to look up for the exact and precise information from the search results. It is up to the user to go through each document to extract the relevant and necessary information from those search results. This is an impractical and tedious task .Text mining can be a better solution for this as it links all the extracted information together, pushes all the irrelevant information aside, and keeps the relevant ones based on the question of interest.

2.2 Document Pre-processing

Document pre-processing is the process of introducing a new document to the information retrieval system in which each document introduced is represented by a set of index terms. The goal of document pre-processing is to represent the documents in such a way that their storage in the system and retrieval from the system are very efficient. Document pre-processing includes the following stages:

2.2.1 Tokenization

Tokenization in text mining is the process of chopping up a given stream of text or character sequence into words, phrases, symbols, or other meaningful elements called tokens which are grouped together as a semantic unit and used as input for further processing such as parsing or text mining.

Tokenization is a useful process in the fields of both Natural language processing and data security. It is used as a form of text segmentation in Natural Language processing and as a unique symbol representation for the sensitive data in the data security without compromising its security importance.

Usually, tokenization occurs in a word level but the definition of the “word” varies accordingly to the context. So, the series of experimentation based on following basic consideration is carried for more accurate output:

- All alphabetic characters in the strings in close proximity are part of one token; likewise with numbers.
- Whitespace characters like space or line break or punctuation characters separate the tokens.
- The resulting list of tokens may or may not contain punctuation and whitespace

In languages such as English (and most programming languages) where words are delimited by whitespace, this approach is straightforward. However, tokenization is more difficult for languages such as Chinese which have no word boundaries. Simple whitespace-delimited tokenization also presents difficulties when word collocations such as New York should be treated as one token. Some ways to address this problem are by developing more complex heuristics, querying a table of common collocations, or fitting the tokens to a language model that identifies collocations in a later processing step.

For example:

Input: “Friends, Romans and Countrymen”

Output: Tokens

- Friends
- Romans
- Countrymen

2.2.2 Stop Word Removal

Sometimes a very common word, which would appear to be of little significance in helping to select documents matching user’s need, is completely excluded from the vocabulary. These words are called “stop words” and the technique is called “stop word removal”.

The general strategy for determining a “stop list” is to sort the terms by collection frequency and then to make the most frequently used terms, as a stop list, the members of which are discarded during indexing.

Some of the examples of stop-word are: a, an, the, and, are, as, at, be, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with etc.

2.2.3 Lemmatization

Lemmatisation (or lemmatization) in linguistics, is the process of reducing the inflected forms or sometimes the derived forms of a word to its base form so that they can be analysed as a single term.

In computational linguistic, lemmatisation is the algorithmic process of getting the normalized or base form of a word, which is called lemma, using vocabulary and morphological analysis of a given word. It is a difficult task to implement a lemmatizer for a new language as the process involves complex tasks such as full morphological analysis of the word, that is, understanding the context and determining the role of a word in a sentence (requiring, for example, the grammatical use of the word).

According to Plisson et al. (2005), lemmatization plays an important role during the document pre-processing step in many applications of text mining. Beside its use in the field of natural language processing and linguistics, it is also used to generate generic keywords for search engines or labels for concept maps.

Lemmatisation and stemming are closely related to each other as the goal of both processes is to reduce the inflectional forms or derivationally related forms of a word to its base form. However, stemming is a heuristic process in which the end of the words or the affixes of the derivational words are chopped off to receive the base form of the word. Lemmatisation goes through the whole morphological analysis of the word and uses the vocabulary to return the dictionary or base form of the word which is called lemma.

For instance: If operated on a token “saw”, stemming may return just “s”. While lemmatisation will go through the morphological analysis and return see or saw depending upon the use of word “saw” as a verb or noun in the sentence.

2.2.4 Synonym Expansion

According to McCarthy et al. (2007), synonym expansion, also known as lexical substitution, is the task of replacing a certain word in a given context with another suitable word similar in meaning.

When a word has multiple meanings, synonym expansion tries to find the correct meaning of the word used in a sentence by identifying its synonyms (or substitutes) in a given context. Given a sentence, for example “she was a bright girl”, the task is to find a synonym that could replace the word bright without changing the meaning of the sentence. Let us assume that we take “bright” as the target word, then the word “brilliant” could be the suitable substitution for the selected word, which would both maintain the meaning of the target word and at the same time fit the context of the sentence.

2.3 Document Representation

Document representation is a key process in the document processing and information retrieval systems. To extract the relevant documents from the large collection of the documents, it is very important to transform the full text version of the documents to vector form. A such transformed document describes the contents of the original documents based on the constituent terms called index terms. These terms are used in indexing, the relevant ranking of the keywords for optimized search results, information filtering and information retrieval. The vector space model, also called vector model, is the popular algebraic model to represent textual documents as vectors. Using the vector space model, documents are represented using the term frequency (tf), inverse document frequency (idf) or tf- idf weighting scheme.

2.4 Information Extraction

Information Extraction (IE) is an important process in the field of Natural Language Processing (NLP) in which factual structured data is obtained from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the further extraction process. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include:

Term analysis: This identifies one or more words called terms, appearing in the documents. This can be helpful in extracting information from the large documents like research papers which contain complex multi –word terms.

Named-entity recognition: This identifies the textual information in a document relating the names of people, places, organizations, products and so on.

Fact extraction: This identifies and extracts complex facts from documents .Such facts could be relationships between entities or events.

2.5 Clustering / organization of documents

According to Fung, (2001), clustering is the process of grouping contents based on the fuzzy information, such as words or word phrases in a set of documents. In simple words, clustering is the process of grouping the set of physical or abstract objects into classes of similar objects. An object belonging to one cluster is dissimilar to the object belonging to another cluster.

The importance of document clustering is widely acknowledged by researchers Qiujun (2010), Jaiswal (2007) and Shah and Elbahesh (2004) for the better management, smart navigation, efficient filtering and concise summarization of the large collection of documents such as World Wide Web (WWW).

2.5.1 K-means Clustering

There are various methods of clustering. K-means is one of the most efficient methods for clustering.

From the given set of n data, k different clusters; each cluster characterized with a unique centroid (mean) is partitioned using the K-means algorithm. The elements belonging to one cluster are close to the centroid of that particular cluster and dissimilar to the elements belonging to the other cluster.

2.5.2 How does K-means Clustering Algorithm work?

The letter “k” in the K-means algorithm refers to the number of groups we want to assign in the given dataset. If “n” objects have to be grouped into “k” clusters, k clusters centers have to be initialized. Each object is then assigned to its closest cluster center and the center of the cluster is updated until the state of no change in each cluster center is reached.

From these centers, we can define a clustering by grouping objects according to which center each object is assigned to. The detailed algorithm of k-means is presented in Section 5.2.3.

2.6 Motivation

The web pages contain a huge amount of information in unstructured or semi-structured format which can be extracted and transformed to valuable information as per our requirement. Different techniques are available for this purpose. Text mining is one of them and is an important step in knowledge discovery process as already explained above. News articles are one of the important sources of information which keep people up-to-date with the current happenings in the world. Sometimes, for the similar headlines of the news articles, the content may differ across different news portal. The availability of the similar headlines and their corresponding links to the news portals under a single roof would be a very efficient and sophisticated option to explore information on a similar topic.

This project provides a better, deeper, and a clearer understanding of the overall process involved in transforming the headlines extracted as XML form from the web pages, grouping them into clusters based on the similarity of the headlines and displaying them under a single platform with the corresponding links to the original webpages for better exploration of information on a similar topic across different news portals

2.6.1 Problem Statement

A person who is reading particular news in one news portal may also be interested in reading similar news coverage on other news portal to find more about the topic. The problem here is finding the news portal that covers the similar news. The usual approach is to visit each likely news portal and then manually look for the similar news in them to find whether the news the person is looking for is present or not. This is a problematic, time-consuming and tedious task. This even reduces the user’s interest in reading that particular news as well as his enthusiasm to acquire more information on that topic. On the other hand, people are so much into smart technologies these days that they always look for the technology that can satisfy their interest without having them to put in much effort and time.

The solution purposed here is based on the idea of text mining and clustering. The basic idea is to create clusters of similar news headlines and build news content warehouse. This clustered information will then be displayed using a GUI created using asp.net where a reader can find all similar news with the corresponding links in a single webpage which solves the issues of browsing each and every news site and looking for the required news in them.

2.6.2 Research Question

The main objective of the project is to answer the question “how it is possible to view and compare similar news which is published in different news portals from one single platform with the use of text mining?”

To answer this question, different text mining and clustering techniques are used to find the similarities between the news headlines and the GUI is created to display the clusters of similar news headlines with their corresponding links to the original sites.

2.6.3 Objectives

The objectives of this research work are:

- To provide a single platform to place the clusters of similar news headlines and their corresponding links to the original sites.
- To reduce the complexity of visiting each site manually to read the similar news.

2.6.4 . Scope and Limitation

The scope of this project is the headlines of news articles published in different news portals. The system is limited in finding the similarities between the news headlines fetched manually via XML file. Similarity in news content across different news portals for particular headlines is not covered while making clusters of the similar headlines.

3 REQUIREMENTS ANALYSIS

3.1 Data Collection

In order to carry out the study, the headlines of news articles are collected from the random online news portals (online) or some random headlines can be fetched (offline). Those collected data are stored in an XML file for further document preprocessing. The XML file is then fed to the system by referencing its location in the local machine.

3.2 Functional requirements

A functional requirement is something the system must do and includes various functions performed by specific screen, outlines of work-flows performed by the system and other business or compliance requirements the system has to meet.

3.2.1 Use Case Diagram

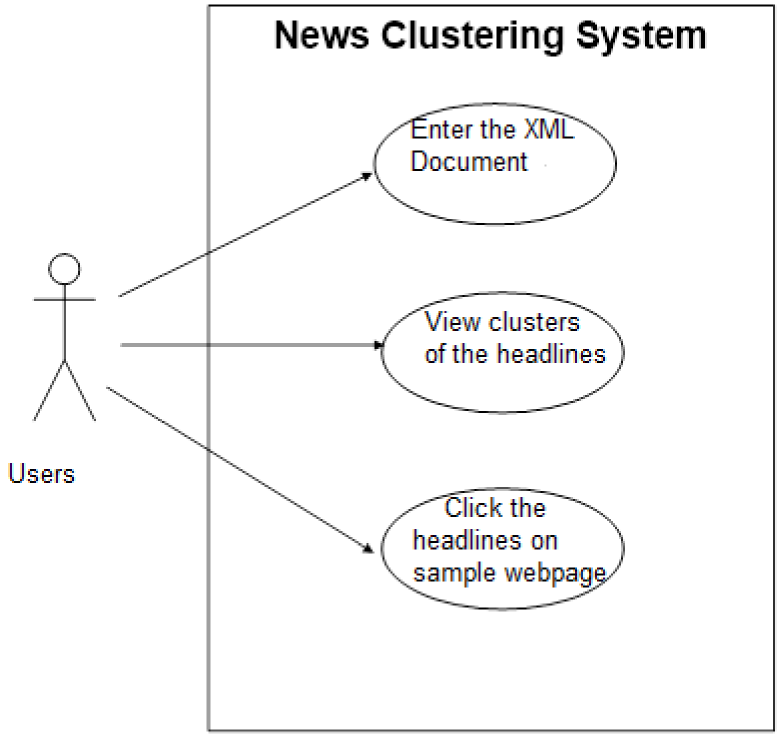

News Clustering System
Enter the XML Document
View clusters of the headlines
Click the headlines on sample webpage
Users

Figure 1. Use case diagram of the system.

3.3 Non-functional Requirements

The non-functional requirements represent requirements that should work to assist the project to accomplish its goal. The non-functional requirements for the current system are:

3.3.1 Interface

The project constructed is console-based. The output is displayed on a console screen and a sample webpage is created to display the clusters of the news headlines.

3.3.2 Performance

The developed system must be able to group the given headlines into clusters based on the K means algorithm.

3.3.3 Scalability

The system must provide as many options like changing headlines in the XML file and then the changes occur in the clusters as well.

3.4 Resource Requirements

- Microsoft Visual Studio for development.
- Window XP or greater.

4 FEASIBILITY ANALYSIS AND SYSTEM PLANNING

4.1 Feasibility Analysis

The feasibility study is an important issue while developing a system. It deals with all the specifications and requirements regarding the project and gives the complete report for project sustainability. The feasibility studies necessary for the system development are mentioned below:

4.1.1 Economic feasibility

It is the study to determine whether a system is economically acceptable. This development looks at the financial aspects of the project. It determines whether the project is economically feasible or not. The system designed is a console application which needs Visual Studio 2010 and all other hardware requirements for it so that the start-up investment is not a big issue.

4.1.2 Technical feasibility

The system is developed for general purpose use. Technical feasibility is concerned with determining how feasible a system is from a technical perspective. Technical feasibility ensures that the system will be able to work in the existing infrastructure. In order to run the application, the user only needs to have Visual Studio 2010 installed and to be able to edit the XML file. These all requirements can be easily fulfilled.

4.1.3 Operational feasibility

Operation feasibility is concerned with how easy the system is to operate. As it is a console-based application, it is quite easy to handle with normal Visual Studio 2010 skills. For the efficient operation of the system, the user needs a general computer. The GUI is a sample webpage, so it does not require any special skill to view and click the links. The proposed system is operationally feasible.

4.2 System Planning

4.2.1 Activity Diagram

The activity diagram is the graphical presentation of stepwise computational and organizational actions (workflows) of a system. The activity diagram for the system is shown in Figure 2.

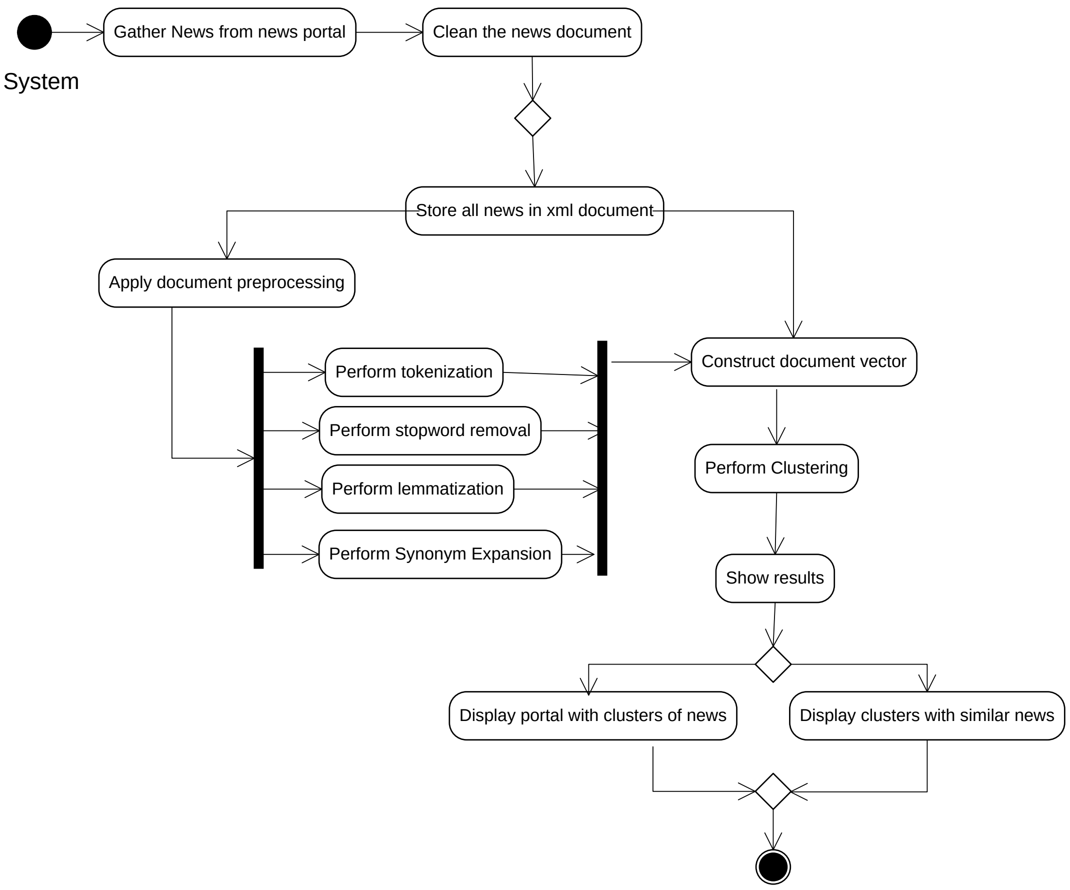

System
Gather News from news portal
Clean the news document
Store all news in xml document
Apply document preprocessing
Perform tokenization
Perform stopword removal
Perform lemmatization
Perform Synonym Expansion
Construct document vector
Perform Clustering
Show results
Display portal with clusters of news
Display clusters with similar news

Figure 2. Activity diagram for the system.

5 RESEARCH METHODOLOGY

According to Kothari (2004), research comprises defining and redefining problems, formulating hypothesis or suggested solutions, collecting, organizing and evaluating data, making deductions and reaching conclusion and further testing the conclusion whether they fit into formulating hypothesis. Research methodology may be broadly classified as Exploratory and Conclusive. This chapter discusses the research methodology and the design strategy implemented to answer the research question of this project.

5.1 Research Approach and Design Strategy

Since all the document pre-processing techniques and the algorithm planned to use during this project are existing ones, the research approach is more of exploratory than conclusive.

The use of existing the K-means algorithm for creating clusters of the news headlines from the corpus of dataset achieved through various document preprocessing and text mining techniques and generation of an idea which can be further explored and carried on to develop a new system related to the web data mining explains that the research methodology used during this project is Exploratory.

Data mining is a good research field that uses data either originated from the researcher for addressing specific problem at hand or secondary data which have already been collected for other purposes like generating new hypothesis or new idea based on the existing studies and researches. Secondary data can be classified as internal and external. The research strategy for this project is the analysis of secondary data which consists of news articles headlines gathered from different news portals and stored in an XML file. The research approach and design strategy of this study is illustrated in Figure 3.

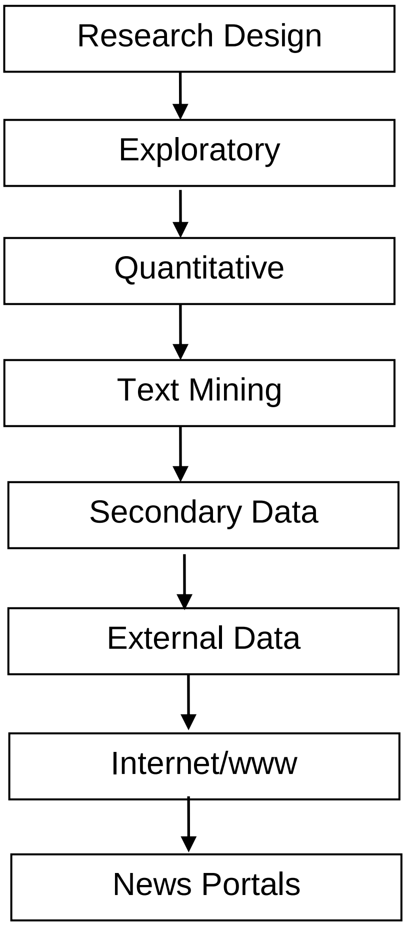


Figure 3. Research approach and design strategy of the study.

5.2 The Overall Research Process

In this section, the overall process used for the project study is described. The process includes step-by-step actions performed for the clustering of the news headlines from different news portals. The overall research process is illustrated in Figure 4.

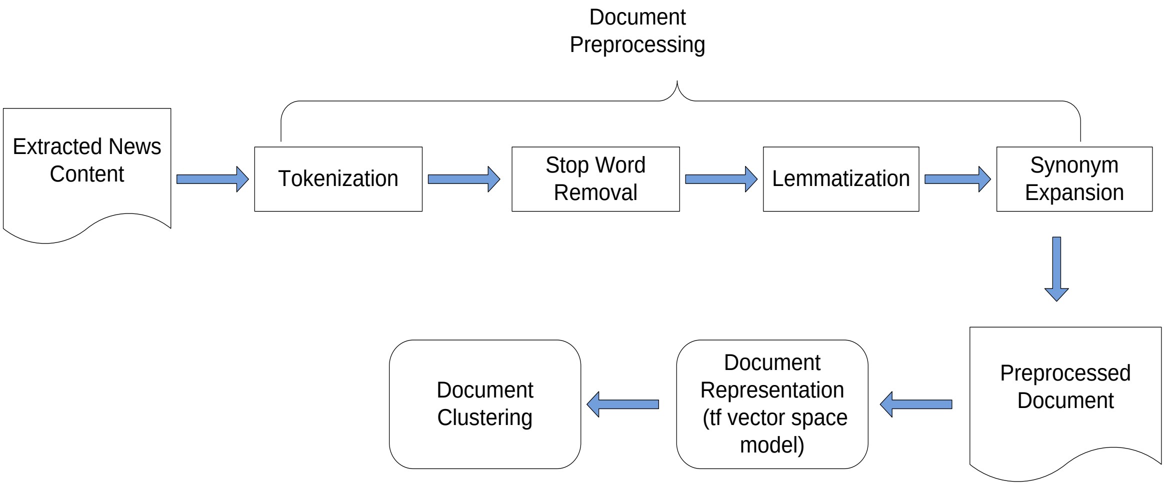


Figure 4. The overall research process.

5.2.1 Document pre-processing

When the news content along with their links and news heading is extracted in an XML document, it is further processed for text mining using document pre-processing techniques. The document pre-processing process includes the following steps:

5.2.1.1 Tokenization

The detail concept and explanation about Tokenization have been mentioned in Section 2.2.1.The text used in this project is English. So, instead of using the complex methods, the tokenization process is accomplished by using space to split the sequence of characters to token.

5.2.1.2 Stop Word Removal

Sometimes extremely common words which would appear to be of little value in helping select documents matching a user's need are excluded from the vocabulary entirely. These words are stop words and the process is called stop word removal. For the purpose of stop word removal, we create a list of stop words such as a, an, the, and prepositions. Hence the tokens contained in the stop word list are discarded.

5.2.1.3 Lemmatization

The detail concept and explanation about lemmatization have been mentioned in Section 2.2.3. After the text is tokenized, each inflected term is reduced to its base form so that it can be analysed as a single term. Lemmatization goes through the morphological analysis of the word in the text, so it is a preferred method to stemming.

5.2.1.4 Synonym Expansion

The detail concept and explanation about Synonym Expansion have already been mentioned in Section 2.2.4. Synonym expansion is carried out by searching each token in the dictionary and transforming each word to the base words. The dictionary consists of a list of words and all of their synonyms.

5.2.2 Document Representation

Vector space model is the one of the efficient methods of representing documents as vectors using the term frequency weighting scheme as mentioned in Section 2.3. The entire collection of dataset from the XML file is represented as vectors using the Vector space model.

5.2.3 Clustering Using K-means Algorithm

After the construction of the document vector, the process of clustering is carried out. The K-means clustering algorithm is used to meet the purpose of this project.

The basic algorithm of K-means used for the project is as following:

K-means Algorithm

Use:

For partitioning where each cluster’s center is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

Output:

A set of *k* clusters.

Method:

```
Step 1: Choose $k$ numbers of clusters to be determined.
Step 2: Choose $C_k$ centroids randomly as the initial centers of the clusters.
Step 3: Repeat
        3.1: Assign each object to their closest cluster center using Euclidean distance.
        3.2: Compute new cluster center by calculating mean points.
Step 4: Until
        4.1: No change in cluster center OR
        4.2: No object changes its clusters.
```

5.2.4 Presentation of the results

The results obtained from the entire process, which, in fact are the clusters of similar news headlines, are presented in the console and also on a sample web page. The sample web page shows clusters of news headlines with underlying links to the news portals containing the corresponding news.

6 EXPERIMENTAL SETUP

The confidence in the overall system performance is yet to be justified by the experimental results. This chapter discusses the experimentation carried to evaluate the clustering system. The requirements for the experimentation are as following:

6.1 Corpus

The corpus is a large collection of documents. The corpus for this project is the headlines of the news articles extracted in an XML file along with the corresponding site names. The headlines are stored as structured text.

6.2 Dataset

The dataset is constructed by retrieving the structured text from the XML file which consists of the structured news headlines. Then the dataset is fetched to the document pre-processing model as described in Section 5.2.1.

6.3 Approach

For the experiment, the K-means algorithm is applied on the dataset. The clusters are created based on the minimum distance between the documents and the cluster centroids. For each dataset of n clustering examples, n experiments are run, where each clustering is taken in turn as the single example “test set” with the n−1 remaining clustering as the “training set”.

The experiment uses the following logical steps:

1) Read n datasets from the XML file.

2) Apply document preprocessing techniques as discussed in Section 5.2.1 to every dataset.

3) Create the n-dimensional document vector for the dataset.

4) Pass the document vector to the K-means clustering algorithm.

5) Observe the clusters obtained and measure the performance of the model based the cohesive measures of the clusters.

6) Carry out the experiment for different dataset containing news headlines of another day.

7 IMPLEMENTATION

Implementation is the process of executing a plan or design to achieve some output. In this thesis, the implementation encompasses the extraction of news headlines in the XML document, fetching them in the system to go through the document pre-processing techniques, forwarding the pre-processed documents to the clustering system and obtaining clusters of similar news headlines as a final output.

7.1 Implementation model

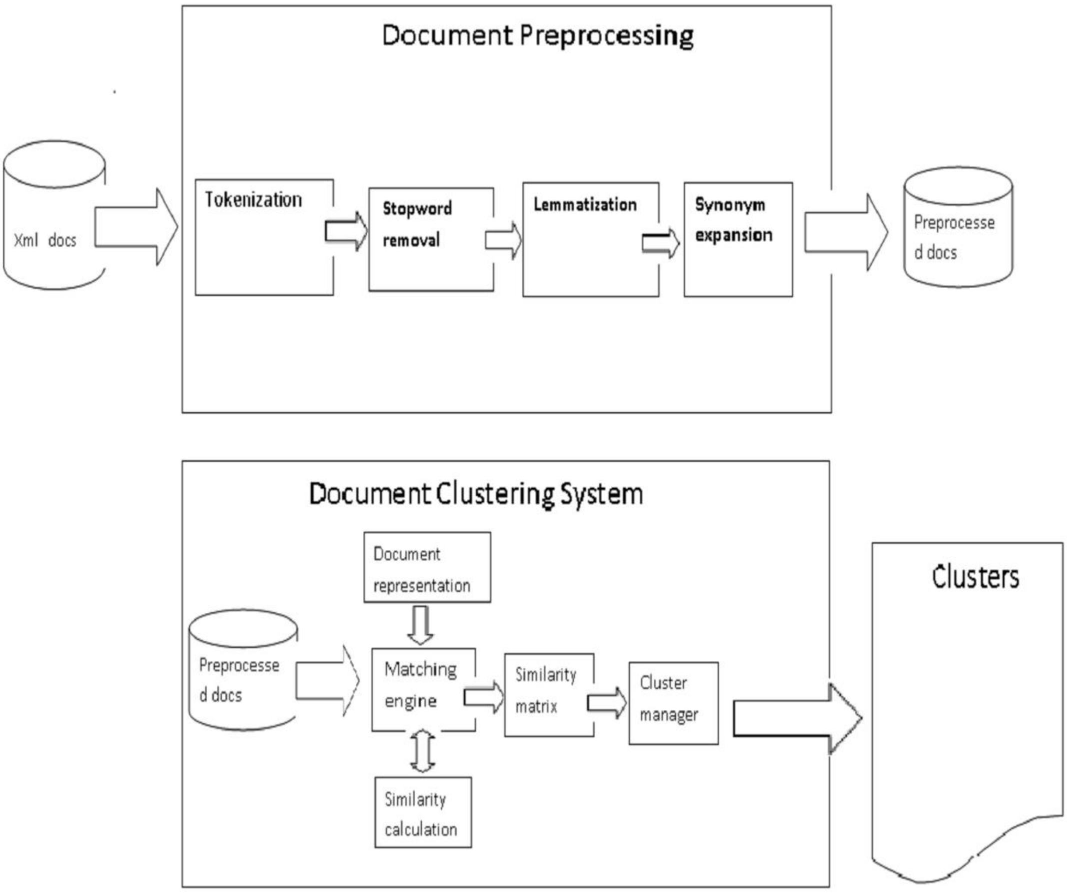


Figure 5. Implementation model of the project.

7.2 Implementation phases

The processes involved during the implementation phases are described in detail in Section 5.2. The following flowcharts are the diagrammatic representations of those processes.

7.2.1 Flowchart for Document Pre-processing

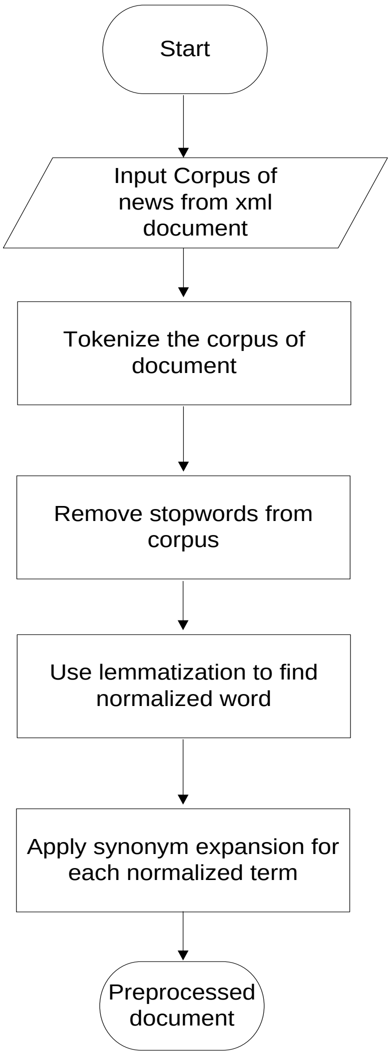

Start
Input Corpus of news from xml document
Tokenize the corpus of document
Remove stopwords from corpus
Use lemmatization to find normalized word
Apply synonym expansion for each normalized term
Preprocessed document

Figure 6. Flowchart for document preprocessing.

7.2.2 Flowchart for Clustering Process

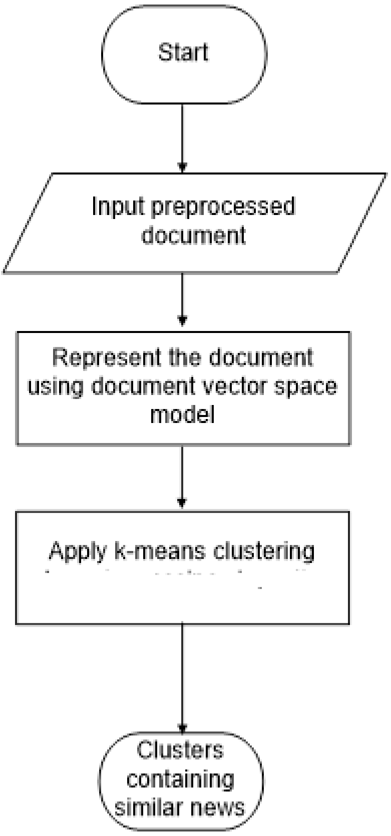

Start
Input preprocessed document
Represent the document using document vector space model
Apply k-means clustering
Clusters containing similar news

Figure 7. Flowchart for clustering component.

8 TESTING AND ANALYSIS

8.1 Testing and Analysis with custom data sample

K-means is a heuristic method of constructing clusters of the documents. There is no guaranteed output but the execution time is really fast. The K-means algorithm constructs the clusters of the documents based on their minimal distances to the centroid of the cluster. The output depends on optimal calculation of centroids of the clusters. To explain this, the following custom sample data is used.

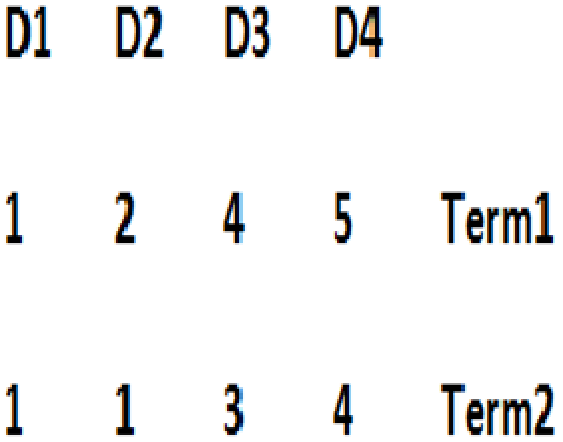
8.1.1 Document Input

The sample data with four documents (D1, D2, D3 and D4) consisting distinct terms Term 1 indicated as X and Term 2 indicated as Y in the following table is taken as an input document for the experimental analysis.

Documents	Attribute 1(x) Term 1	Attribute 2(Y) Term 2
D1	1	1
D2	2	1
D3	4	3
D4	5	4

Table 1. Custom data sample with the term frequency.

The matrix representation of the above data is given below:



D1	D2	D3	D4	
1	2	4	5	Term1
1	1	3	4	Term2

Figure 8. Matrix representation of custom data sample.

The graphical representation of the given custom data sample is given below:

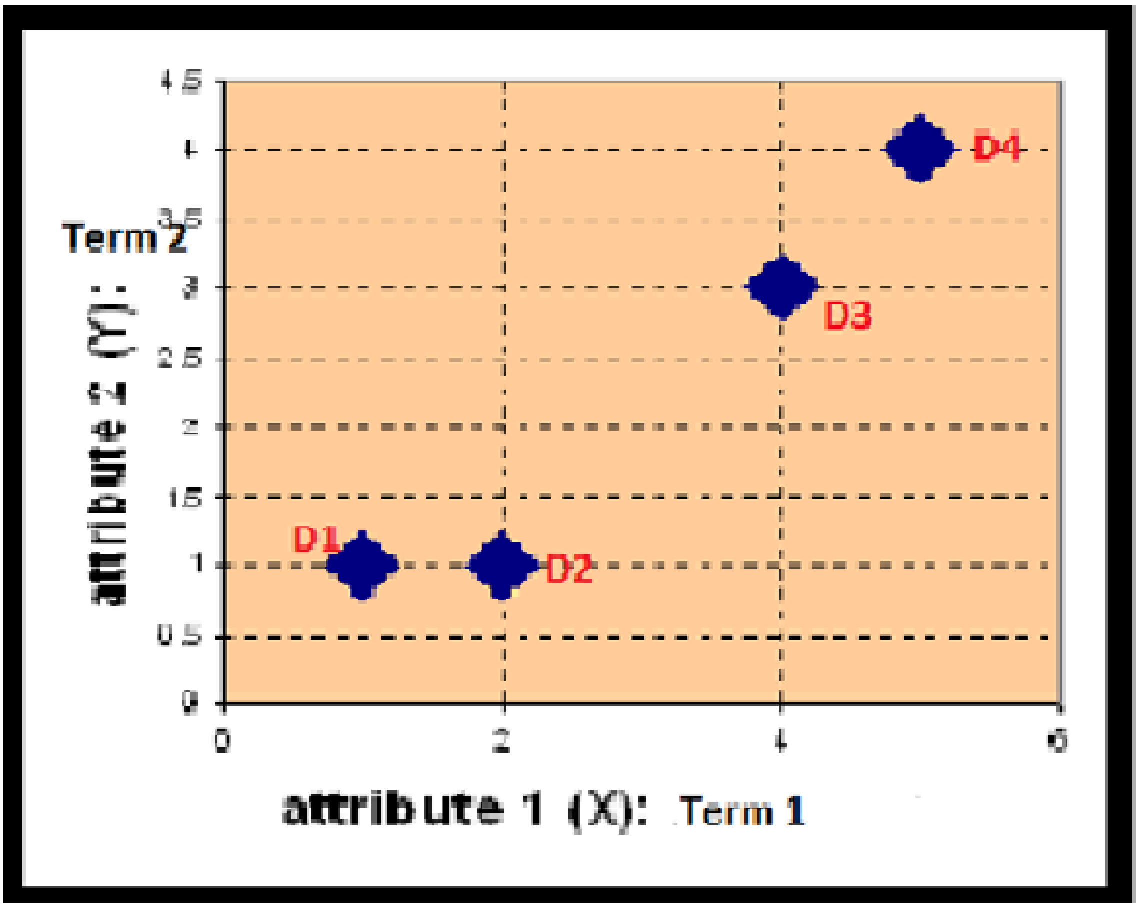


Figure 9. Graphical representation of custom data.

The input document is now processed for clustering with the use of the K-means algorithm. The steps involved in K-means algorithm have already been explained in Section 5.2.3.

Step 1:-

Here, K=2 (Number of clusters to be created)

Documents D1 and D2 are assumed as the first cluster centroids.

Let C_1 and C_2 be the coordinates for D1 and D2 respectively.

So, C_1= (1, 1) and C_2= (2, 1)

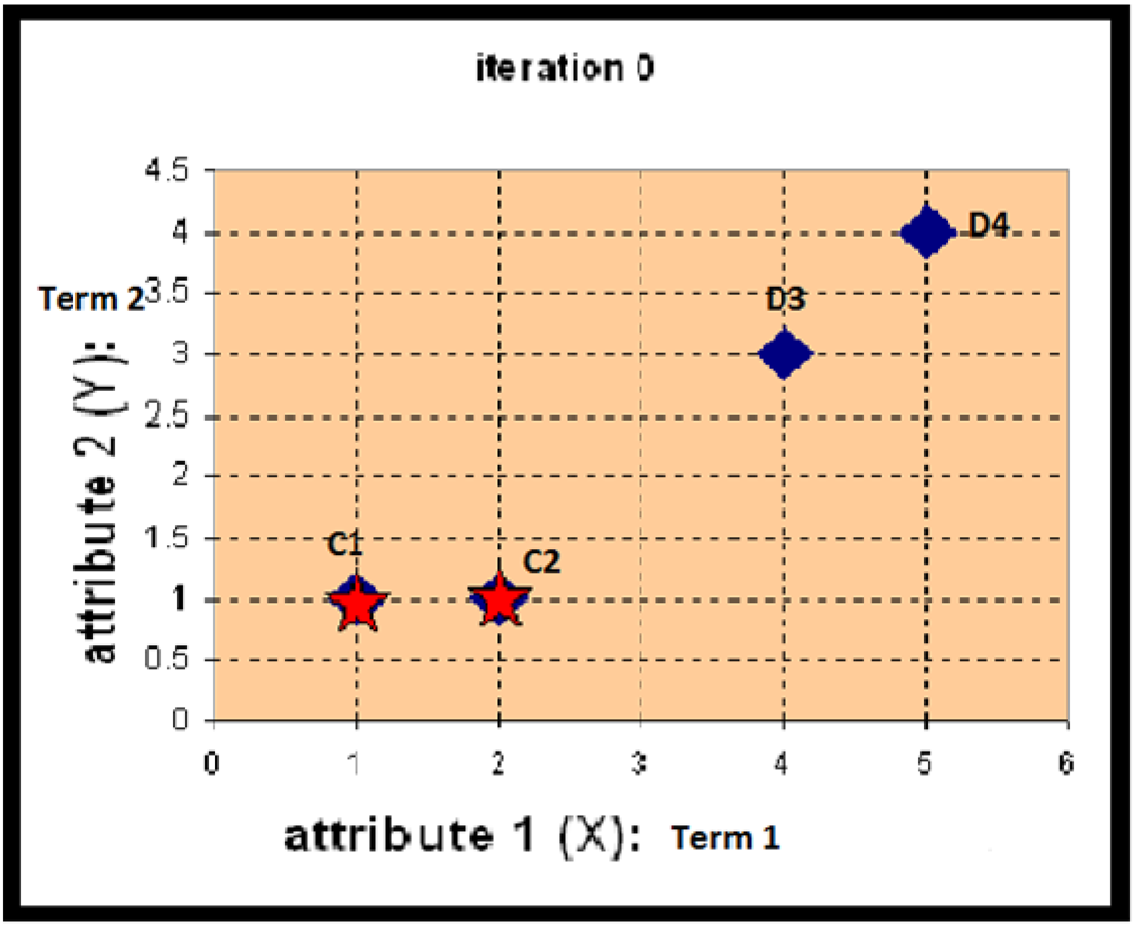


Figure 10. Graphical representation of Iteration 0 of the K-means algorithm.

Step 2:-

Distances between each document (D3 and D4) and centroids C_1 and C_2 are calculated using the Euclidean distance formula.

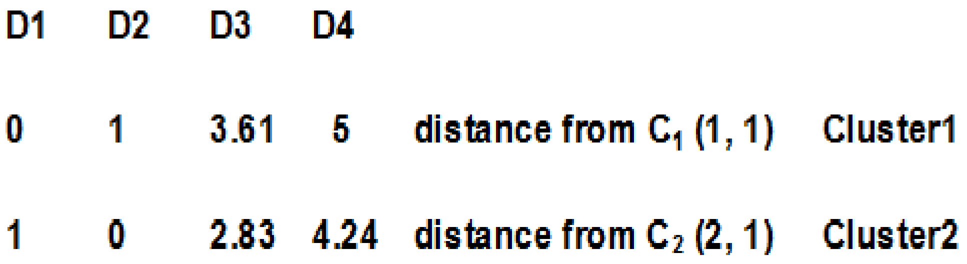
For example: Distance from document D3 (4, 3) to first centroid C_1 (1, 1) is

Square root of [square of (4-1) + square of (3-1)]

=3.61

and the distance of D3 to C_2 = 2.83.

The same method is carried out to calculate distance from D1, D2 and D4 to the centroids C_1 and C_2 and the final result is displayed in matrix form.

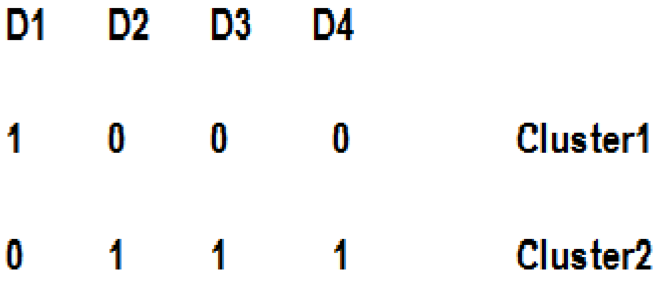


D1	D2	D3	D4		
0	1	3.61	5	distance from C_1 (1, 1)	Cluster1
1	0	2.83	4.24	distance from C_2 (2, 1)	Cluster2

Figure 11. Matrix representation of distance between documents and centroids.

Based on the minimum distance from the centroid, the documents are assigned to the clusters.

D1 is assigned to cluster 1 as its distance is minimal to C_1 compared to $C_{2.}$ D2 is assigned to cluster 2 as its distance is minimal to C_2 compared to C_1 and so on for the other documents.



D1	D2	D3	D4	
1	0	0	0	Cluster1
0	1	1	1	Cluster2

Figure 12. Matrix representation of the documents in the cluster.

In the above matrix representation (Figure 12), value 1 in the row indicates that the document belongs to the corresponding cluster and value 0 represents that the document is not a member of the cluster.

Thus, Cluster 1 has D1 as its member and Cluster2 has D2, D3 and D4 as its members.

Step 3:-

In this step, Iteration 1 of the algorithm runs, where the new centroid of each cluster based on the new membership is calculated.

Cluster 1 has only one member which is D1. So, the centroid remains the same (C_1). Cluster 2 has three members, thus the new centroid is the average coordinates of the three members (D2, D3 and D4).

C_2= ((2+4+5)/3, (1+3+4)/3) = (11/3, 8/3)

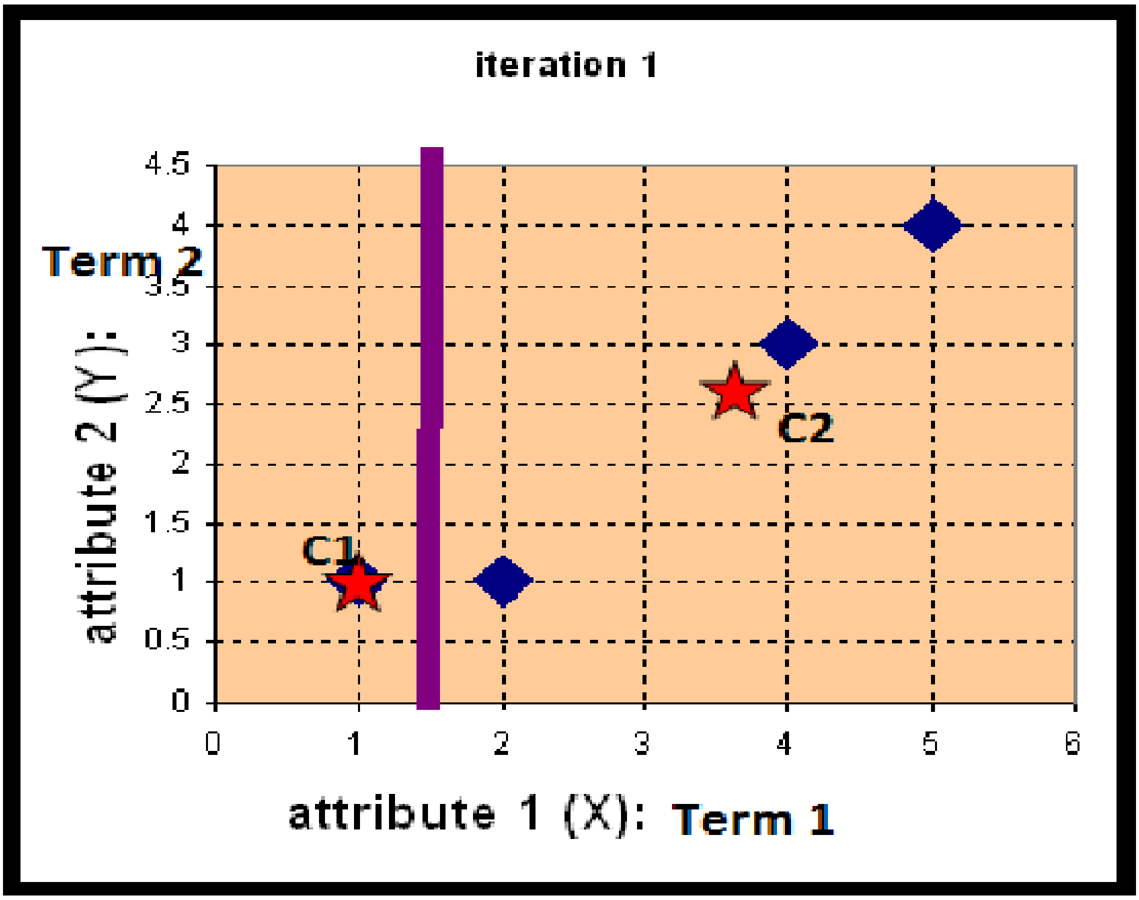
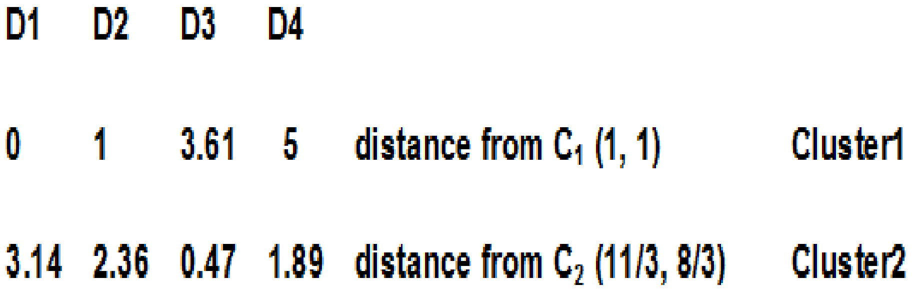


Figure 13. Graphical representation of Iteration 1 of the K-means algorithm.

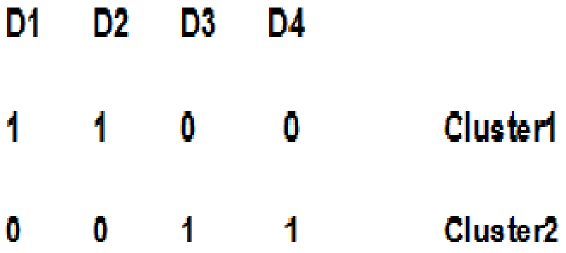
After the new centroids are computed, the distances of all the documents to them are computed as in Step 2.



D1	D2	D3	D4		
0	1	3.61	5	distance from C_1 (1, 1)	Cluster1
3.14	2.36	0.47	1.89	distance from C_2 (11/3, 8/3)	Cluster2

Figure 14. Representation of distance between documents and new centroids.

Figure 14 shows the documents and their distances from the new centroids. Based on the minimal distance to each centroid, the documents are redistributed to their nearest clusters.



D1	D2	D3	D4	
1	1	0	0	Cluster1
0	0	1	1	Cluster2

Figure 15. Matrix representation of the documents in the cluster.

Based on the above document matrix (Figure 15), cluster 1 has two members D1 and D2 and cluster 2 also has two members D3 and D4.

Step 4:-

Cluster 1 has two members D1 and D2 and Cluster 2 also has two members D3 and D4. Thus, the new centroid for each cluster is calculated.

C_1 = Average of coordinates of D1 and D2 belonging to cluster 1.

C_2 = Average of coordinates of D3 and D4 belonging to cluster 2.

C_1 = ((1+2)/2, (1+1)/2) = (3/2, 1)

C_2 = ((4+5)/2, (3+4)/2) = (9/2, 7/2)

All these processes take place in iteration 2.

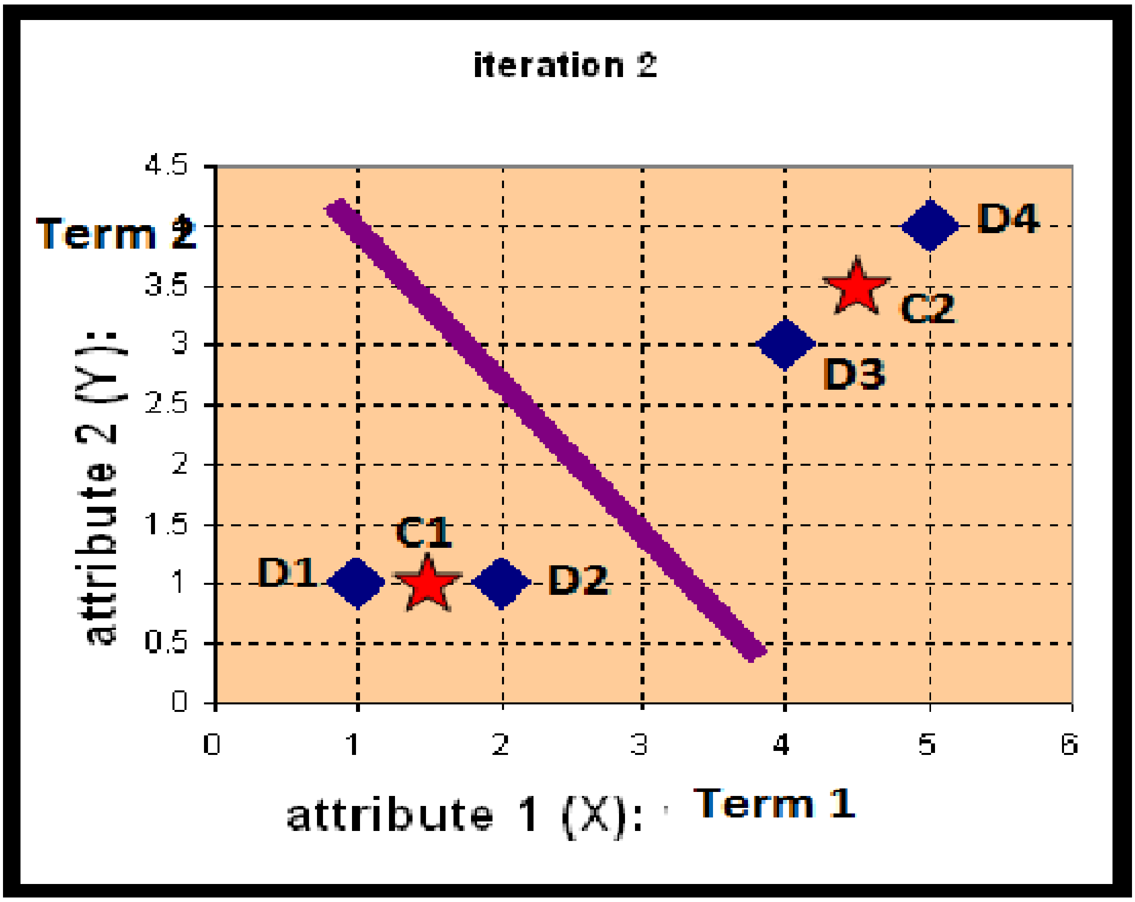
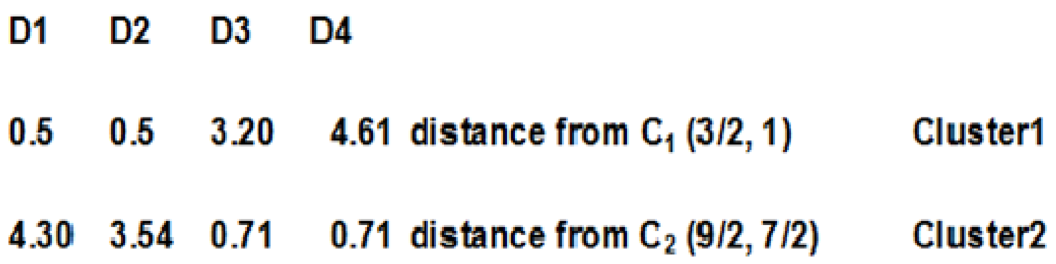


Figure 16. Graphical representation of Iteration 2 of the K-means algorithm.

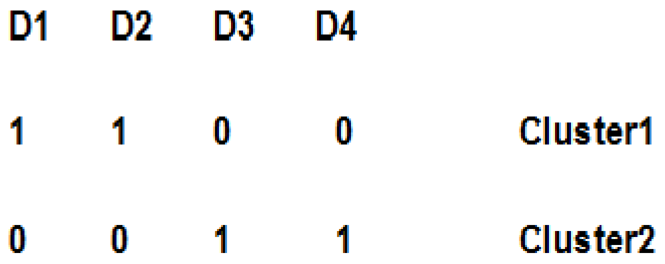
The distance from each documents to the new centroids are calculated as described in Step 2.



D1	D2	D3	D4		
0.5	0.5	3.20	4.61	distance from C_1 (3/2, 1)	Cluster1
4.30	3.54	0.71	0.71	distance from C_2 (9/2, 7/2)	Cluster2

Figure 17. Representation of distance between documents and new centroids.

Based on the distances calculated above, the documents are redistributed to their nearest clusters.



D1	D2	D3	D4	
1	1	0	0	Cluster1
0	0	1	1	Cluster2

Figure 18. Matrix representation of the documents in the cluster.

Based on the above document matrix (Figure 18), Cluster 1 has two members D1 and D2 and Cluster 2 also has two members D3 and D4.

The result obtained from Step 3, represented in Figure 15 and the result obtained from Step 4, represented in Figure 18, are same. The grouping of documents from the iteration 1 and the iteration 2 shows that the documents do not move to new cluster anymore. Thus, the computation of the K- means algorithm stops. The following result is obtained.

Documents	Term1	Term2	Result(Cluster)
D1	1	1	1
D2	2	1	1
D3	4	3	2
D4	5	4	2

Table 2. Final result of the custom data after clustering.

8.1.2 Analysis

From the final result obtained from the test, the conclusion is drawn that the optimal calculation of centroids is required for the better and more accurate clustering of the documents. The iteration keeps on going until the stability of the clusters is obtained.

The real data is the XML documents with a collection of a large number of news headlines. The clustering process of news headlines involves all the processes described in Section 8.1.1. The manual calculation of the centroids of all the documents taken as input for the project and displaying the final results is a tedious and time-consuming task. So, based on the testing and analysis of the final output of the custom data sample, the results of the clusters displayed by the clustering system are analyzed and the conclusion of the project is drawn.

9 CONCLUSION

The first goal of the current study was to use text mining techniques on web news articles headlines. The project study involved the great deal of work on various areas of information retrieval and text mining and focused on the various methods for document pre-processing and document clustering.

Text mining and clustering techniques are really powerful. This project study was completely based on these techniques. The system was created for finding the similarities among the news articles headlines. Various techniques were applied for preparing the corpus of a pre-processed document. Lastly, the k-means clustering algorithm was used for creating the clusters of similar news articles headlines.

The similar news headlines were grouped into a single cluster. The real world application of the project study would help people to find the similar news headlines on different news portal from a single platform. This would not have been possible without the use of text mining and clustering techniques. In general, it is not feasible to manually look for similar news in each of the portals and then compare each of them to find similarities between them.

10 LIMITATIONS AND FUTURE WORK

10.1 Limitations

The proposed method only clusters news articles headlines on the basis of their similarities. The news headlines are not classified into categories such as national, sports, politics, entertainment, international, etc. We applied K-means Clustering to construct the clusters containing similar news. The result produced would have been more accurate if the combined method of clustering such as hybrid clustering based on partition clustering and hierarchical clustering methods had been used.

10.2 Future Work

As mentioned in limitations, the clustered news headlines have not been classified into categories. In future, categorization methods can be added to render results and user interface more manageable and user friendly. After successful appreciation of the model, enhancement can be done to cover the news articles over large domains. Web crawlers and parsers can be developed to extract information from the whole news portal sites. Text mining and clustering techniques can be used to create the clusters on the basis of contents and those contents present on the different news portals can be displayed under the single platform. That single platform can be any webpage, web application or any mobile application. In addition, this project study can be extended to apply the methods proposed for web documents other than articles. The developed model can be enhanced to use in web content mining.

REFERENCES

Ali Shah, N. & M. ElBahesh, E. “Topic-Based Clustering of News Articles”, University of Alabama at Birmingham. Retrieved September 23, 2013 from:

http://pdf.aminer.org/000/006/766/topic_based_clustering_of_news_articles.pdf

Bouras, C. & Tsogkas, V. 2010. "W-kmeans: Clustering News Articles Using WordNet". Retrieved September 5, 2013 from:

http://ru6.cti.gr/ru6/publications/116262780379.pdf

Buscaldi, D., Rosso, P. & Arnal Sanchis, E. 2005. “A WordNet-based Query Expansion method for Geographical Information Retrieval”. Retrieved September 2, 2013 from:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.8031&rep=rep1&type=pdf

D. Manning, C., Raghavan, P. & Schütze, H. 2008. “Introduction to Information Retrieval”, Cambridge, England: Cambridge University Press. Retrieved September 4, 2013 from:

http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Falinouss, P. 2007. “Stock Trend Prediction Using News Articles.” Master’s Thesis, Lulea University of Technology, Sweden. Retrieved September 20, 2013 from:

http://epubl.ltu.se/1653-0187/2007/071/LTU-PB-EX-07071-SE.pdf

Fung, G. 2001. “A Comprehensive Overview of Basic Clustering Algorithms”. Retrieved September 6, 2013 from:

http://pages.cs.wisc.edu/~gfung/clustering.pdf

Huang, C., Simon, P., Hsieh, S. & Prevot, L. 2007. “Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Word break Identification”. Retrieved September 2, 2013 from:

http://delivery.acm.org/10.1145/1560000/1557791/p69-huang.pdf?ip=86.50.66.20&id=1557791&acc=OPEN&key=BF13D071DEA4D3F3B0AA4BA89B4BCA5B&CFID=389226789&CFTOKEN=39573449&__acm__=1387138322_f651d65355dcc035ef1e98e656194624

Jaiswal, P. "Clustering Blog Information" 2007. Master's Thesis Projects, Paper 36, San Jose State University. Retrieved September 18, 2013 from:

http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1035&context=etd_projects

Kobayashi, M. & Takeda, K. 2000. "Information retrieval on the web". Retrieved September 20, 2013 from:

https://www.ischool.utexas.edu/~i385d/readings/Kobayashi.pdf

Kothari C.R. 2004. Research Methodology: Methods and Techniques. Retrieved October 8, 2013 from:

http://books.google.com/books?id=8c6gkbKi-F4C&printsec=frontcover&source=gbs_atb&redir_esc=y#v=onepage&q&f=false

Plisson J., Lavrac N. & Mladenic D. 2004. “A Rule based Approach to Word Lemmatization”. Retrieved September 25, 2013 from:

http://eprints.pascal-network.org/archive/00000715/01/Pillson-Lematization.pdf

Maheshwari, P. & Agrawal, J. 2010. "Centroid Based Text Clustering". Retrieved October 3, 2013 from:

http://www.ijest.info/docs/IJEST10-02-09-36.pdf

Qiujun, L. 2010. "Extraction of News Content for Text Mining Based on Edit Distance". Retrieved October 3, 2013 from:

http://www.jofcis.com/publishedpapers/2010_6_11_3761_3777.pdf

Murali Krishna, S. & Durga Bhavani, S. 2010. "An Efficient Approach for Text Clustering Based on Frequent Itemsets". Retrieved October 2, 2013 from:

http://www.textmining.xpg.com.br/ejsr_42_3_04.pdf

Vectomova, O. & Wang, Y. 2006. "A study of the effect of term proximity on query expansion" (Abstract). Retrieved October 2, 2013 from:

http://jis.sagepub.com/content/32/4/324.abstract

McCarthy, D. & Navigli, R. 2009. “English Lexical Substitution”. Retrieved October 4, 2013 from:

http://www.dianamccarthy.co.uk/task10index.html

Appendices

Appendix 1.0

Sample code for overall clustering and text mining (NewsClustering.cs).

```
class NewsClustering
   {
      //create the list of objects of class headInfo to hold heading detail
      public static List<NewsLibrary.NewsInfo> NewsList = new
List<NewsLibrary.NewsInfo>();
      public static List<List<NewsLibrary.NewsInfo>> clusterList = new
List<List<NewsLibrary.NewsInfo>>();


      static void Main()
      {
         //pass xml file of news and obtain list of object of news
         NewsLibrary.PopulateNews news = new NewsLibrary.PopulateNews();


         NewsList = news.populate(@"..\\..\\..\\title.xml");


         //pass list of news objects to tokenization and obtain token of headline
         NewsLibrary.Tokenization tokenization = new
NewsLibrary.Tokenization();
         NewsList = tokenization.Tokenize(NewsList);


         //pass list of news objects to stopword and obtain stopword free
headline
         NewsLibrary.StopWord stopword = new NewsLibrary.StopWord();
         NewsList = stopword.remove(NewsList);


         //pass list of news objects to lemmatization and lemmmatize tokens in
headline
          NewsLibrary.Lemmatization lematization = new
NewsLibrary.Lemmatization();
          NewsList = lematization.Lematize(NewsList);


         //past list of news objects to synononym  expansion
         NewsLibrary.SynonymExp Syn = new NewsLibrary.SynonymExp();
         NewsList = Syn.FindSynonym(NewsList);

        //pass list  news objects to KMeansClustering
         NewsLibrary.KMean KMean = new NewsLibrary.KMean();
         clusterList=KMean.Cluster(NewsList, 15);

         //Generate cluster XML file
        NewsLibrary.GenerateXML XMLFile = new
NewsLibrary.GenerateXML();
        XMLFile.Create(clusterList);
    }
```