Australian Journal of Basic and Applied Sciences, 8(9) June 2014, Pages: 22-30

AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Comparative Analysis for Alignment Based Document Clustering

[1]T.Veeramani and [2]Dr. R. Nedunchelian

[1]*Research Scholar, Department of Computer Science and Engineering, Saveetha School ofEngineering , Saveetha University,Chennai, Tamil Nadu, India.*
[2]*Professor, Department of Computer Science and Engineering, Saveetha School ofEngineering , Saveetha University, Chennai, Tamil Nadu, India.*

ARTICLE INFO

Article history:
Received 2 March   2014
Received in revised form
13 May 2014
Accepted 28 May 2014
Available online 23 June 2014

Keywords:
Text Mining, Clustering, K-means Clustering, Enhanced K-means, Weighted K-means

ABSTRACT

Background: Document Clustering is a technique that organizes a large quantity of unordered text Document into small number of meaning full and coherent cluster. Clustering approach facilitates the presentation of search result in more compact form and enables thematic browsing result set. **Objective:** The main problem of existing web search result based on poor vector representation of snippets. The Data units returned from the underlying database are normally encoded into the result page dynamically for human browsing which essential for many application such as internet comparison, shopping, and also be extracted out and assigned meaningful labels. **Result:** We present a clustering approach such K-Means, Weighted K-Means and Enhanced K-Means Algorithm. This method is capable of handling a variety of clustering approach based on Alignment Algorithm. **Conclusion:** Our Experimental result shows that the precision and result are achieved to improve the performance of clustering system is highly effective.

ToCite ThisArticle:T.Veeramani, Dr. R. Nedunchelian., Comparative Analysis for Alignment Based Document Clustering. ***Aust. J. Basic & Appl. Sci.,*** *8(9): 22-30, 2014*

INTRODUCTION

Data miningis an extraction of hidden predictive information from large databases. Data mining is the procedure of investigating data from alternate points of view and outlining it into convenient data. Data mining includes the utilization of refined data analysis tools to uncover previously obscure, quality samples and connections in huge data sets. These devices can incorporate factual models, numserical algorithms, and machine learning strategies. Therefore, Data mining comprises of more than gathering and overseeing data; it additionally incorporates analysis and forecast. With the enormous measure of data accumulated in records, databases, and different repositories, it is progressively imperative to create effective methods for analysis and maybe elucidation of such data and for the extraction of fascinating information that could help in decision making. Constraints of Data Mining are principally data or personnel related as opposed to technology related.
Data mining comprises of five major components:

- Extract, convert, and burden transaction data onto the data warehouse framework.
- Store and deal with the data in a multidimensional database framework.
- Provide data access to business experts and data innovation experts.
- Analyze the data by application programming.
- Present the data in a handy format, for example, a diagram or a table.

Text Mining:

Miller illustratesText mining as "the mechanized or partially computerized transforming of text". Characterizes text mining with a methodology model demonstrating segments and interacting steps particular to texts. Fig 1 shows the process of text mining, Fig 2 shows the sample search result.At the starting there is the crude text input indicated as text corpus indicating an accumulation of text records, in the same way as memos, reports, or productions. (Krushmerick *et al*., 1997) Grammatical parsing and preprocessing steps convert the unstructured text corpus into a semi-organized configuration meant as a text database. Depending on the input material this procedure stop may be profoundly intricate including record group transformations or modern meta data handling, or a basic task and as simply perusing in the texts.

Corresponding Author: T.Veeramani, [1]Research Scholar, Department of Computer Science and Engineering, Saveetha School ofEngineering , Saveetha University,Chennai

Thusly an organized representation is made by processing a term-report matrix from either the text corpus or the text database. (Su*et al*, 2009) the term-record matrix is a bag-of-words component holding term frequencies for all reports in the corpus. This regular data structure forms the premise for further text mining analysis, like

- Text classification, i.e., assign a priori known labels to text documents,
- Syntax analysis, i.e., analyzing the syntactic structure of texts,
- Relationship identification, i.e., finding connections and similarities between distinct subsets of documents in the corpus,
- Information extraction and retrieval, and
- Document summarization, i.e., extracting relevant and representative keywords, phrases, and sentences from texts.

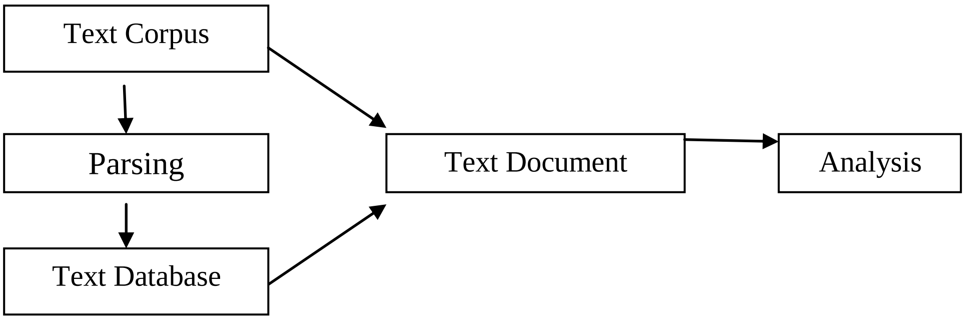


Fig. 1: Model for Text Mining Process.

Clustering:

Clustering plays an important role in many engineering fields such as pattern recognition, system modeling, image processing, communication, data mining, etc.Clustering is the process of grouping data into clusters, where objects within each cluster have high similarity, but are dissimilar to the objects in other clusters.

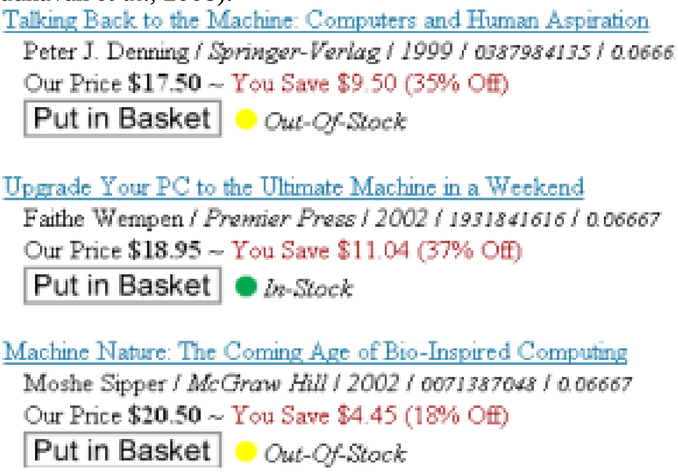
Document Clustering:

Document clustering is a kind of text data mining and organization technique that automatically groups related documents into clusters. Traditionally single words occurring in the document are identified to determine the similarities among documents.

There are several ways to model a text document. For example, it can be represented as a bag of words, where words are assumed to appear independently and the order is immaterial. The bag of word model is widely used in information retrieval and text mining. Words are counted in the bag, (Krushmerick., 1997) which differs from the mathematical definition of set. Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension. Here we use the frequency of each term as its weight, which means terms that appear more frequently are more important and descriptive for the document.

Related works:

Information extraction and annotation has been an active research area. In wrapper induction systems (Krushmerick *et al*., 1997), (Lee, 1997) theyrely on human users to mark and label the desired information. They induce a series of rules called wrapper toextract the same set of information on result pages from the same web database. Hence, the system achieves highextraction accuracy through supervised training and learning process they suffer from poor scalability and notsuitable for online applications.Conceptual-model-based data extraction (L. Kaufman and P. Rousseeuw, 1990) uses ontologies with heuristics to extract informationautomatically from the result pages and label them. Ontologies are defined as structural framework for organizinginformation. Ontologies for various domains are constructed manually.Several works in (Elmeleegy *et al*., 2009), (Lu *et al*., 2007) automatically assigns meaningful labels to the data units of SRRs. In dataextraction from large websites (Elmeleegy *et al*., 2009) annotates data units with their closest labels on the result page. This method haslimited applicability since they do not encode data units with labels on result pages. In ODE (Lu *et al*., 2007), first ontologies areconstructed using query interface and result pages from the same web database. Domain ontologies are used to labeleach data unit and with the same label they are aligned. This method is sensitive to quality and completenessattributes. Previous approaches of automatic data alignment techniques are based on few features: HTML tag paths (Madhavan *et al*., 2008), Visual feature (Liu *et al*., 2001), splitting of SRR into text segments (Embley *et al*,1999).(Zhao *et al*., 2005) describes an Automatic Wrapper Generation System calledViNt for extraction of SRRs from web pages.(Lu *et al*., 2007) describes the multi-annotator approach for constructingannotation wrapper. Our approach in this paper is similar to the approach referred here. It uses six different types ofannotation and constructs annotation wrapper. Annotation here refers to the assignment of meaningful labels. .(Lu *et al*., 2007) is the extension of the multi-annotator approach followed in (Madhavan *et al*., 2008).



Talking Back to the Machine: Computers and Human Aspiration
Peter J. Denning / *Springer-Verlag / 1999 / 0387984135 / 0.0666*
Our Price **$17.50** ~ You Save $9.50 (35% Off)
Put in Basket ● *Out-Of-Stock*

Upgrade Your PC to the Ultimate Machine in a Weekend
Faithe Wempen / *Premier Press / 2002 / 1931841616 / 0.06667*
Our Price **$18.95** ~ You Save $11.04 (37% Off)
Put in Basket ● *In-Stock*

Machine Nature: The Coming Age of Bio-Inspired Computing
Moshe Sipper / *McGraw Hill / 2002 / 0071387048 / 0.06667*
Our Price **$20.50** ~ You Save $4.45 (18% Off)
Put in Basket ● *Out-Of-Stock*

Fig. 2: Sample Search Result.

It explains the relationship between the text nodes and the data units. It also enhances the alignment and the cluster-shifting algorithm to improve the efficiency. (Arlotta *et al*., 2003) basically annotate data units with the closest labels onresult pages. However, this method has very limited applicability because many Web sites do not encode data units withtheir labels on result pages. DeLa (J. Wang and F.H. Lochovsky, 2003) uses several heuristics to assign labels to the extracted SRRs. Also DeLa uses thelocal Interface Schema whereas our approach differs from it by using the Integrated Interface Schema. (H. He*et al*., 2005) describes the WISE Integrator approach for automatic integration ofsearch queries from different websites.

Preprocessing:

It is widely known that preprocessing text data before feeding it into clustering algorithm is essentials and can have great impact on algorithm performance. Here Document preprocessing there are several ways to model a text document. For example, it can be represented as a bag of words, where words are assumed to appear independently and the order is immaterial. The bag of word model is widely used in information retrieval and text mining. Words are counted in the bag, which differs from the mathematical definitionof set. Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension.

Let $\mathbf{D= \{d_1,d_2,\ldots,d_n\}}$ be a set of documents and $\mathbf{T= \{t_1,t_2,\ldots,t_m\}}$ the set of distinct *terms* occurring in D. We discuss more precisely what we mean by’term’ below for the moment just assume they are words. A document is then represented as a m-dimensional vector td. Let tf(d,t)denote the frequency of term d ε D. Then the vector representation of document d is

$T_d=(tf(d,t_1),\ldots,tf(d,t_m))$:

Although more frequent words are assumed to be more important asmentioned above, this is not usually the case in practice.

The following standard preprocessing steps are performed.

(i) Stop-word Elimination
(ii) Text Stemming

Text Stop-words Elimination:

A stop word itself doesn’t bring any semantic meaning but in connection with other words can form meaningful phrases. Therefore, term that occur in stop-word list are specially make to be ignored from document index term, but not removed.

Text Stemming:

A version of Porter’s Stemming is used in this step to remove *prefixed* and *suffixes*, normalizing terms to its root form. This process can reduce vocabulary of the collection. These stemmed terms are linked to its original form, which are preserved to be used in subsequent phases. For example **production, produce, produces and product** will be mapped to the stem produce. The underlying assumption is that different morphological variations of words with the same root/stem are thematically similar and should be treated as a single word.

Term weighting Techniques:

Term-weighting uses statistical regularities in the documents to estimation the significance weights for the terms. Term-weighting function can measure how specific terms are to a topic by exploiting the statistic variations in the distribution of the terms within the relevant documents and with a complete document collection (L. Kaufman and P. Rousseeuw, 1990). The following methods are some of the most commonly used weighting factors.

Term Frequency (TF):

Term that repeat multiple times in a document are considered important. Terms that appear in many documents are considered common and are not indicate of document content. Based on this idea, the number of documents in a document collection defined as N. The frequency of a term t_i in a document can be used for document specific weighting and denoted $tf(t_i)$.It is only a measure of a term's significance within a document. It assigns high weights to terms that appear more frequently within a document.

Inverse Document Frequency (IDF):

This was proposed by sparck Jones in 1972. The IDF is used to measure the specificity of terms in a set of documents. The intuition was that a query term that occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents. The measure was a heuristic implementation of this intuition. Document Frequency $df(t_i)$ is the number of documents that contain the term $t_{i.}$ (Wu *et al*., 2003).

The formula of the *idf proposed*by Sparck Jones can be expressed by the following.

$$\mathbf{Idf(t_i)=log(N/df(t_i))}$$

Here,

- N is Total Number of Documents.
- Df(t_i) is t Number of times term occurance.

Term Frequency Inverse Document Frequency (tf-idf):

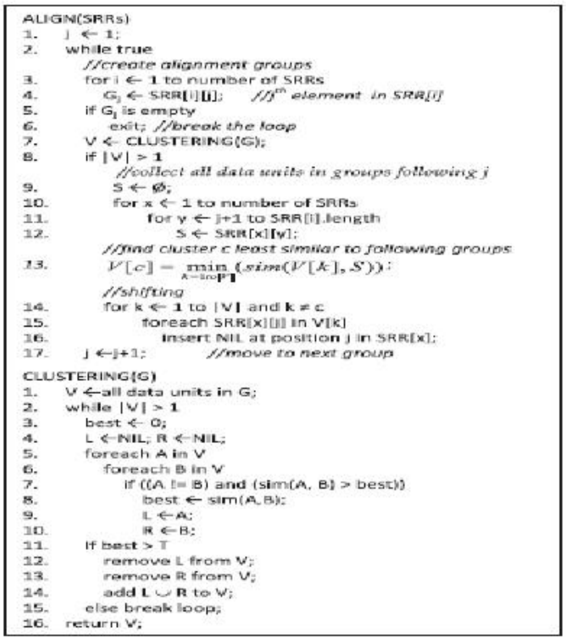
The extensive empirical studies of combination of weighting factors have been conducted by many research hers. A term-weighting method that involves multiplying the IDF measure (possibly one of a number of variants) by a TF measure (again possibly one of a number of variants, not just the raw count) (Lu*et al*.,2007)is a very popular term-weighting method developed (J. Lee, 1997). The weight of a term t_i represented by the *tf-idf*value is the combination of the exhaustively statistic*(tf)* to a term t_i. It can be expressed by the formula:

$$\mathbf{W(t_i)=tf(t_i) \times idf(t_i)}$$

Data Alignment:

The data units are not aligned whenever the search result records are extracted from the web page. The main purpose of data alignment is to group the data units from different records into the semantically same group. This alignment of data records facilitates easier annotation of data. It is based on the assumption that the data units in different SRRs of the same semantic usually have the fixed layout and presentation. Based on this assumption a record expression(REXP)[9] is constructed for each result record. An REXP is a string comprising of sequence of symbols that represents either the presentation style of the node or the separator/ delimiter. For example, in our current implementation of REXP,\S" denotes a pure text node with bold style, \s" denotes a pure text node without bold style, \L" denotes a link node with bold style, \l" denotes a link without bold style, \^" denotes starting a new line, etc. Separators are nodes that contain onlynon-letter and non-digit characters appearing in HTML text.

The REXP for each SRR can be constructed easily. Example 1: The REXP of the first record in Figure 1 is \l^ss/s/s/s^sS»s^s", where \=" and \»" are the separators appearing in HTML text of the record. Note that as shown the text \Peter J. Denning" and a \=" are en-coded together, so the first \s" in the REXP represents \Peter J. Denning=".\Put in Basket" is not included because buttons, icons and imagesare currently ignored. The last \s" represents \Out-Of-Stock". With this Record expression a suffix tree is constructed. Then the most common longest string (MCLS) is selected and its corresponding components are aligned to form groups. The data Alignment also concentrates on the Data Unit Similarity, Data Content Similarity, Presentation Style Similarity, Data Type Similarity, Tag Path Similarity as described in (O. Zamir and O. Etzioni, 1998). Also for improving the efficiency of data grouping and aligning, Alignment and cluster-shifting technique (O. Zamir and O. Etzioni, 1998) is also used. The algorithm concentrates on four steps namely Merging text nodes, Aligning text nodes, Splitting Composite Text nodes and Align Data Units. The alignment algorithm is shown in Fig 3.



```
ALIGN(SRRs)
    j ← 1;
    while true
        //create alignment groups
        for i ← 1 to number of SRRs
            G_i ← SRR[i][j];     //j^th element in SRR[i]
        if G_j is empty
            exit; //break the loop
        V ← CLUSTERING(G);
        if |V| > 1
            //collect all data units in groups following j
            S ← Ø;
            for x ← 1 to number of SRRs
                for y ← j+1 to SRR[i].length
                    S ← SRR[x][y];
            //find cluster c least similar to following groups
            $V[c] = \min_{k=1 to |V|} (sim(V[k], S))$;
            //shifting
            for k ← 1 to |V| and k ≠ c
                foreach SRR[x][j] in V[k]
                    insert NIL at position j in SRR[x];
        j ←j+1;          //move to next group

CLUSTERING(G)
    V ←all data units in G;
    while |V| > 1
        best ← 0;
        L ←NIL; R ←NIL;
        foreach A in V
            foreach B in V
                if ((A != B) and (sim(A, B) > best))
                    best ← sim(A,B);
                    L ←A;
                    R ←B;
        if best > T
            remove L from V;
            remove R from V;
            add L ∪ R to V;
        else break loop;
    return V;
```

Fig. 3: Alignment and Clustering Algorithm.

Clustering Algorithm:

Clustering is the methodology of arranging data objects into a set of disjoint classes called clusters.One of the basic problems that arise in a variety of fields, including pattern recognition, machine learning and statistics, is clustering. The fundamental data clustering problem may be defined as discovering groups in data or grouping similar objects together. Each of these groups is called a cluster.

K-Means Clustering Algorithm:
K-Means Algorithms steps:

(i) Choose random k points and set as cluster centers.
(ii) Assign each objects to the closest centroid's cluster.
(iii) When all objects have been assigned, recalculate the
Positions of the centroids.
(iv) Go back to Steps 2 unless the centroids are not changing

Fig. 4:K-means Algorithm Steps.

Enhanced K-Means Clustering Algorithm with Improved Initial Centeroids:

In this section, presents anupgraded strategy for upgrading the performance of K-Means clustering algorithm.An improved system is proposed to enhance the ability of the K-Means clustering algorithm. But in the K-Means technique the initial centroids are chosen arbitrarily. Fig 4 shows the K-means algorithm steps, Fig 5 shows the weighted k-means clustering algorithm.So this strategy is exceptionally delicate to the beginning stages and it doesn't guarantee to prepare the special clustering effects. An upgraded algorithm is utilized to enhance the precision and effectiveness of the K-Means clustering algorithm.

In this algorithm two systems are utilized, one technique for discovering the better initial centroids. An alternate strategy for a proficient way for allocating data points to right clusters. The technique utilized for discovering the initial centroids computationally exorbitant. In this work, concentrate on another approach for discovering the better initial centroids with decreased time intricacy.

In the Enhanced K-Means algorithm first checking, whether the given data set hold the negative worth characteristics or not. Assuming that the data set holds the negative quality characteristics then changing the all data points in the data set to the positive space by subtracting the every data point characteristic with the base property estimation in the given data set. Assuming that data set holds the all positive worth qualities then the change is not needed.

In the following step, for every data point, ascertain the distance from origin. At that point, the first data points are sorted in agreement with the sorted distances. In the wake of sorting partition the sorted data points into K equivalent sets. In each one set take the mean values the initial centroids. These initial centroids lead to the better exceptional clustering outcomes. Next, for every data point the distance computed from all the initial centroids. The following stage is an iterative methodology which makes utilization of a heuristic methodology to decrease the obliged computational time. The data points are allocated to the clusters having the closest centroids in the following step. Clusterid of a data point means the cluster to which it has a place. Nearestdist of a data point indicates the display closest distance from closest centroid.

Next, for each one cluster the new centroids are computed by taking the mean of its data points. At that point for every data point the distance ascertained from the new centroid of its available closest cluster. Assuming that this distance is less than or equivalent to the past closest distance, then the data point stays in the same cluster, overall for every data point, need to ascertain the distance from all centroids. After calculated the distances, the data points are assigned to the appropriate clusters and the new ClusterId's are given and new NearestDist values are updated. This reassigning process is repeated until the convergence criterion is met.

Algorithm: The Enhanced Method:

```
Require: D = {$d_1$, $d_2$, $d_{3,}$..., $d_i$..., $d_n$ } // Set of n data Points.
$d_i$ = { $x_1$, $x_2$, $x_3$,..., $x_i$,..., $x_m$ } // Set of attributes of one data point.
K// Number of desired clusters.
Ensure: A set of K clusters.
Steps:
1: In the given data set D, if the data points contains the both positive and
Negative attribute values then go to step 2, otherwise go to step 4.
2: Find the minimum attribute value in the given data set D.
3: For each data point attribute, subtract with the minimum attribute value.
4: For each data point calculate the distance from origin.
5: Sort the distances obtained in step 4.
Sort the data points accordance with the distances.
6: Partition the sorted data points into K equal sets.
7: In each set, take the mean value as the initial centroid.
8: Compute the distance between each data point di ($1 \leq i \leq n$) to all the initial
    Centroids$c_j$ ($1 \leq j \leq k$).
9: Repeat
10: For each data point $d_i$, find the closest centroid $c_j$ and assign$d_i$ to cluster j.
11: Set ClusterId[i]=j. // j: Id of the closest cluster.
12: Set NearestDist[i]= d ($d_i$, $c_j$).
13: For each cluster j ($1 \leq j \leq k$), recalculate the centroids.
14: For each data point $d_i$,
14.1 Compute its distance from the centroid of the present nearest cluster.
14.2 If this distance is less than or equal to the present nearest distance, the data
    Point stays in the same cluster.
Else
14.2.1 For every centroid $c_j$ ($1 \leq j \leq k$) compute the distance d ($d_i$, $c_j$).
End for;
Until the convergence criteria is met.
```

Weighted K-Means Clustering Algorithm:

```
    Initialize the number of clusters K.
1.  Randomly selecting the centroids $(\mathbf{c_1 c_2}, \dots, \mathbf{c_k})$ in the data set.
2.  Calculating the weights$\mathbf{w_{(i)}}$of the corresponding centroids$(\mathbf{c_1}, \mathbf{c_2}, \dots, \mathbf{c_k})$,
Calculate Sum $\mathbf{s_{(i)}} = \sum_{j=1}^{n} \mathbf{c_{ij}}$ , $\mathbf{i} = \mathbf{1}, \mathbf{2}, \dots, \mathbf{k}$
      And $\mathbf{w_{(i)}} = (\mathbf{s_{(i)}} - \mathbf{c_{ij}})/\mathbf{s_{(i)}}$ where j=1,2,…,n
      Where$\mathbf{w_{(i)}}$ is corresponding weight vector to the$\mathbf{c_{(i)}}$.
3.  Find the distance between the centroids using the Euclidean Distance equation.
                 $\mathbf{d_{ij}} = \left\| \mathbf{w_{(i)}} * (\mathbf{x_{i-}c_k}) \right\|^2$
4.  Update the centroids using this equation.
5.  Stop, when the new centroids is nearer to old one; otherwise, go to step-4.
```

Fig . 5: Weighted K-means Clustering Algorithm.

Experimental Analysis and Results:

Comparative analysis was performed using K-Means clustering, Enhanced K-Means algorithm with refer library data sets are implemented in MATLAB**.** The Xie - Beni index was used as validation measure for comparative analysis. The dataset is described hereunder.

Experimental Data :

1200 abstracts were collected from journals belonging to ten different areas. For each area, 400 abstracts have been selected. Table 1 lists the areas and the names of the journals. This data set was divided evenly into ten subsets and each subset contained 100 abstracts.

Table 1: Journal Abstract Datasets.

S.NO	SUBJECT	NAME OF THE JOURNAL
1	Zoology	Journal of Zoology
2	Computer Science	Journal of Computer Science
3	Economics	Journal of Economics
4	Education	Journal of Education
5	Geology	Journal of Geology

The abstracts were cut into sentences. Then the tokens were identified from each sentence as terms. All the stop words and stemming words were filtered by using porter Stemmer method. Finally, a document was converted into a list of Terms. The terms were used to construct the Weight feature value for each term.

Experimental Results and Analysis:

The basic clustering algorithms, K-Means,Enhanced K-Means, Simple Vector space model , were selected for comparison. In this experiment, the Enhanced K-Means methods are executed to alleviate the effect of a random factor. All the results are listed in Table 2.

Table 2 Experimental Results and Analysis

Algorithms	No.of Instance	Data Set 1		Data Set 2		Data Set 3	
		Xie-Beni	Time(sec)	Xie-Beni	Time(sec)	Xie-Beni	Time(sec)
K-Means	400	0.081	0.62	0.071	0.59	0.063	0.54
Weighted K-Means	400	0.077	0.51	0.058	0.47	0.057	0.47
Enhanced K-Means	400	0.044	0.36	0.045	0.35	0.045	0.37

Table 2 shows the xie-beni index measures is low for proposed method compare to existing method. From this, the proposed method Enhanced K-Means is the best method to cluster the documents.

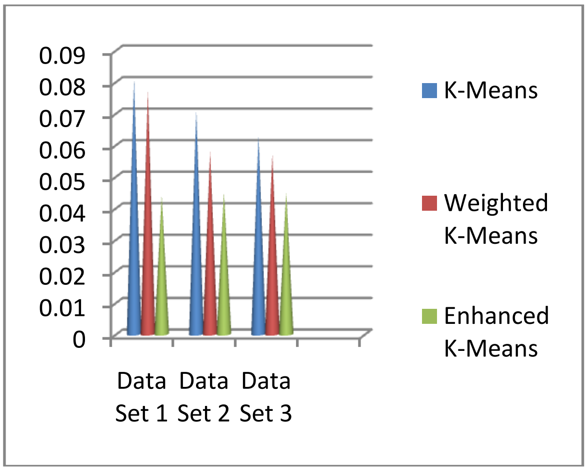


Fig.6: Comparison of K-Means Clustering Algorithm with Proposed Algorithm.

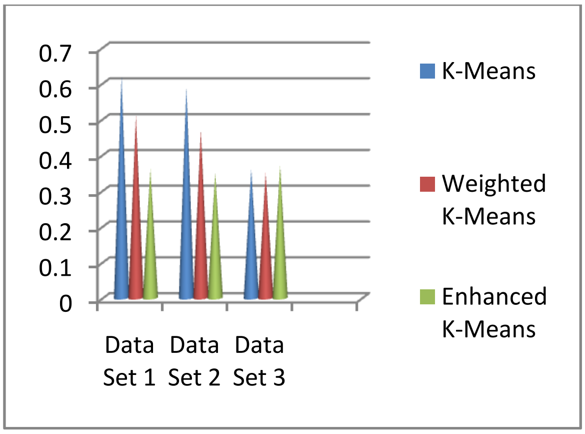


Fig.7: Comparison with respect to CPU Time between K-Means and Proposed Algorithm.

From the Fig 6 and Fig 7 one can know the time takes to found the document cluster using proposed method is improved compared to method K- Means, Weighted K-Means Algorithm.

Conclusion

In this study, a new approach to deals with Enhanced K-Means Algorithm used for Document Clustering. Though it presents a remarkable advantage in reducing the time due to Enhanced K-Means. To reduce the computational complexity and improve running efficiency, it has shown to behave well when compare to more statistical measures. Experiments results shows that clustering perform much better than original K-Means method. Many other Clustering methods used text categorization, but none of them consider the Term Frequency factor. In this method Alignment Algorithm based Cluster is future research we will investigate to use Term Frequency and Document Clustering. Our Experimental result shows that the precision and result are achieved to improve the performance of clustering system is highly effective.

REFERENCES

Arlotta,L., V. Crescenzi, G. Mecca and P. Merialdo, 2003. “Automatic Annotation of Data Extracted from Large Web Sites,” Proc. Sixth Int’l Workshop the Web and Databases(WebDB).

Elmeleegy,H., J. Madhavan and A. Halevy, 2009. “Harvesting Relational Tables from Lists on the Web,” Proc. Very Large Databases (VLDB) Conf.

Embley,D., D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng and R. Smith, 1999. “Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages,” Data and Knowledge Eng., 31(3): 227-251.

He,H., W. Meng, C. Yu and Z. Wu 2004. ‘Automatic Integration of Web Search Interfaces with WISE-Integrator,’ VLDB J., 13-3.

He,H., W. Meng, C. Yu and Z. Wu, 2005. ‘Constructing Interface Schemas for Search Interfaces of Web Databases’ Proc. Web Information Systems Eng. (WISE)Conf.

Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Krushmerick, N., D. Weld and R. Doorenbos, 1997. “Wrapper Induction for Information Extraction,” Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI).

Lee,J., 1997. “Analyses of Multiple Evidence Combination,” Proc. 20th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval.

Liu,L., C. Pu and W. Han, 2001. “XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources,” Proc. IEEE 16th Int’l Conf. Data Eng. (ICDE).

Lu,Y., H. He, H. Zhao, W. Meng and C. Yu, 2007. “Annotating Structured Data of the Deep Web,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE).

Madhavan,J., D. Ko, L. Lot, V. Ganapathy, A. Rasmussen and A.Y. Halevy, 2008. Google’s Deep Web Crawl,” Proc. VLDB Endow- ment, 1(2): 1241-1252.

Meng,W., C. Yu and K. Liu, 2002. “Building Efficient and Effective Metasearch Engines,”ACM Computing Surveys, 34(1): 48-89. [25] S. Mukherjee, I.V.

Mukherjee, S.,I.V. Ramakrishnan and A. Singh, 2005. “Bootstrapping Semantic Annotation for Content-Rich HTML Documents,” Proc. IEEE Int’l Conf. Data Eng. (ICDE).

Ramakrishnan and A. Singh, 2005. “Bootstrapping Semantic Annotation for Content-Rich HTML Documents,” Proc. IEEE Int’l Conf. Data Eng. (ICDE).

Su,W., J. Wang and F.H. Lochovsky, 2009. “ODE: Ontology-Assisted Data Extraction,” ACM Trans. Database Systems, 34-2 article.

Thomas, W. Miller, 2005. Data and Text Mining: a Business Applications Approach , Pearson Edition.

Wang, J. and F.H. Lochovsky, 2003. “Data Extraction and Label Assignment for Web Databases,” Proc. international conference on World Wide Web (WWW-12), 187-196.

Wu, Z.*et al.*, 2003. “Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine,” Proc. IEEE/WIC Int’l Conf. Web Intelligence (WI ’03).

Zamir,O. and O. Etzioni, 1998. “Web Document Clustering: A Feasibility Demonstration,” Proc. ACM 21st Int’l SIGIR Conf. Research Information Retrieval.

Zhai, Y. and B. Liu, 2005. “Web Data Extraction Based on Partial Tree Alignment,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05).

Zhao, H., W. Meng, Z. Wu, V. Raghavan and C. Yu, 2005. Fully automatic wrapper generation for search engines. In WWW ’05, 66-75.