Indian Journal of Science and Technology, *Vol 8(28), DOI: 10.17485/ijst/2015/v8i28/83984, October 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Frame Work for Semi-Supervised Clustering based on Color Constraints to Enhance Text Mining for Efficient Information Retrieval

S. Suguna[1]*, V. Sundaravadivelu[2] and B. Gomathi[1]

[1]Department of Computer Science, Sri Meenakshi Govt. Arts College for Women (A), Madurai – 625002, Tamil Nadu, India; kt.suguna@gmail.com; sweetgoms@gmail.com
[2]PG and Research Department of Computer Science, Thiru A. Govindasamy Govt. Arts College, Tindivanam – 604002, Tamil Nadu, India; vsundar.professor@gmail.com

Abstract

Background/Objectives: In this paper we have analyzed various issues with clustering and text mining. The collected documents are preprocessed and grouped using our proposed new algorithm based on position method. We proved our proposed color based constraint clustering algorithm out performs than K-Means and SOM algorithms in terms of time and reliability factors. **Methods/Statistical Analysis:** We identified the problem after analyzing the existing works with the help of articles from reputed journal papers and national and International level conferences. We proposed the new methodology for document grouping process, and color based constraint clustering process. Clustering can be considered as the most important semi-supervised learning problem which deals with finding a structure in a collection of unlabelled data. In this work the collected documents are preprocessed by stop word removal and stemming process and then the words are grouped according to their similarity using color code constraints. Performances of SOM and Kmeans, and color based constraint algorithms are analyzed for any kind of text document collections. **Findings:** In this work our proposed color based constraint (CBC) algorithm, SOM and K-Means algorithms performances are compared against time based frequency and reliability of retrieved documents. Here, the time needed to process the number of documents is analyzed. Reliability of retrieved documents can be made by using the number documents and the frequency measurement. We proved our proposed color based constraint clustering algorithm out performs than K-Means, and SOM algorithms in terms of time and reliability. **Application/Improvements:** Our work is useful for efficient information retrieval process. In future this work can be extended to maximize the grouping of words with minimum latency and one can also extend this work to develop an algorithm for maximize the grouping(clustering) of words in a document with color based constraints to increase the clustering performance for efficient text mining.

Keywords: Color Based Constraint, Clustering, Information Retrieval, Semi_Supervised Clustering Technique, Text Mining

1. Introduction

Clustering[1] is one of the traditional mining techniques. It is an unsupervised learning methodology. Clustering method try to inherent grouping of the text documents. Most current document clustering methods are based on Vector Space Model (VSM). Clustering divides data into meaningful groups (or clusters). Clustering capture the nature of data structure, it means that the characterization of the data should not change. Related documents are grouped or segmented (tokenization) for easy processing and browsing. The aim of the clustering[4] is to form a similar (or related) to one another within the group but dissimilar with the other group. The data or points in the group (or clusters) are homogeneous and belongs to the same data structure. The main aspects of the clustering is

*Author for correspondence

at any cause there should not be any overlapping because, it leads to complexity. Grouping of data (phrases, words) or clustering is characterization or frequency of data with reference to the search process. In some cases the value in different clusters are common to the some other clusters that is called Fuzzy clusters (Partially belong to various groups). Clustering text documents into different category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems.

Consider the survey of existing works. It is essential to determine an accurate similarity between documents[1]. Concept based similarity measures based on matching concepts at the sentence. The concept based similarity measures relay on three factors 1. Analyzed labeled terms are the concepts that capture the semantic structure of each sentence. 2. The frequency of concept used to measure the contribution to the meaning of the sentence as well as the main topic of the document. 3. The number of documents that contains the analyzed concepts is used to discriminate among documents in calculating the similarity. These three aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the sentence level.

Text mining process involves the Natural Language Processing and Data Mining to discover unknown[2-4] information or new pattern (knowledge). In content management tasks, the goal is to automatically categorize large amount of text documents in to groups or clusters. Constraints can be obtained from multiple auxiliary sources. Occurrence of two documents[5-7] in a dictionary can be used to infer a multi – link constraint between the documents, two documents in different categories of the dictionary can be considered as connected link. Using constraint from the auxiliary data sources, one can customize the clustering output for the particular task. The hierarchy of the document is close to the input dictionary structure in which the documents replaced.

Enhancing text clustering using concept based model concept is discussed in[8]. Most of text mining techniques are based on word and/or phrase analysis of the text. The statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. However, to achieve a more accurate analysis, the underlying mining technique should indicate terms that capture the semantics of the text from which the importance of a term in a sentence and in the document can be derived. Clustering with constraints is an important recent development in clustering of data (data mining). User have advantage by using constraints are it allow the users to incorporate the domain expertise in to clustering process by desirable properties in a clustering solution. Semi-supervised learning mechanism is recent trend or innovation in data mining. It is the midway (intersection) of supervised and unsupervised learning[9]. Two factors that involved in the process: 1. Addition of unlabeled data to supervised classifier learning and 2. Addition of some supervision into clustering.

Clustering with constraints allows the incorporation of back ground domain expertise with work in the form of instant knowledge in the form of instance level constraints. Constraints[10,11] are used in clustering algorithm to improve the clustering analysis. The constraints are used to bias the search for an appropriate clustering of the data. The constraints based approaches are 1. One is with strict enforcement, which finds the best feasible clustering 2. Once with partial enforcement, this finds the best clustering. Semi supervised clustering is to improve with the help of [12-15] user provided information. If constraints are selected improperly there would be degradation of clustering performance. Partial constraints required manually inspect the data points in a question which can be time consuming and costly. Clustering performance[16] can be improved by user by using semi-supervised clustering by specify the available prior knowledge about the data. Various studies point out that constraints plays an important role in improve the resulting data. But it is too important to choose constraints because in some time it might degrade the clustering performance. They used active query (constraint) selection mechanism, where the quires were selected using a min-max criterion. In[17], concept-based mining model, a labeled term either word or phrase is considered as concept.

Our proposed algorithm is explained in section 2. Section 3 deals with performance analysis of our proposed color constraint clustering algorithm, SOM and K-Means algorithms. Conclusion is explained in section 4.

2. Proposed Work

In this section we proposed the new methodology for document grouping process, and color based constraint clustering process. Clustering can be considered as the most important semi-supervised learning problem which deals with finding a structure in a collection of unlabelled

data. In Phase1 preprocessing process such as stop word removal and stemming process is discussed. In Phase 2 grouping process with the help of our proposed position based method is explained. In Phase 3 and 4 the SOM, K-Means, and our proposed color based constraint clustering algorithms are discussed. Finally performance evaluation based on time and reliability measurements are analyzed in phase 5.

2.1 Phase1: Preprocessing Process (Stop Word Removal and Stemming Process)

Step 1: Stop Word Removal: The text documents are collected for the mining process. A stop word is a commonly used word (such as "the", "a", "and") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. In order to save both space and time, these words are dropped at indexing time and then ignored at search time.

Step 2: Stemming Process: Stemmers used to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

2.2 Phase 2: Grouping Process

After completing the preprocessing steps such as stop word removal and stemming process the words are grouped according to their similarity. Our processed algorithm for grouping the documents is discussed below. The proposed methodology is explained with an example.

For example given sentence is,

"Clustering has been used in many application domains, including biology, medicine, marketing, communication, animal classification, image processing, pattern recognition, customer, administration and document retrieval and so on".

The above sentence is grouped in to two domains 1. Odd group (D1) 2. Even group (D2).

Position Representation of Characters

1	2	3	4	5	6	7	8	9	10
A /a	B/b	C/c	D/d	E/e	F/f	G/g	H/h	I/I	J/j
11	12	13	14	15	16	17	18	19	20
K/k	L/l	M/m	N/n	O/o	P/p	Q/q	R/r	S/s	T/t
21	22	23	24	25	26				
U/u	V/v	W/w	X/x	Y/y	Z/z				

Assume that words starts with the first character which are in odd position are belongs to odd set (group). The words starts with the first character which are in the even position are belongs to even set (group). The words in the sentence are grouped according to the first character. This process is extended up to the maximum combination of characters with respect to the position.

Words in the clusters (sub groups) are represents as points (values). Values in the sub groups are again grouped into sequences. Now the sub set (clusters) have the same group of values and in sequence (nearer) and pave easy way for search the required word from the document.

Step 1 (Categorization of Words)

Category 1 (D1) (Odd Position)	Category II (D2) (Even Position)
clustering, used	has, been
In, many, application	domains, biology
including, medicine	processing
marketing, communication	pattern , recognition
animal, classification	document, retrieval
image , customer	
administration, and	
and so, on	

Figure 1. Shows the process of segmentation of documents. Let D1 , D2 are domains and g1, g2, g3, g4, g5, g6 are clusters (sub groups). The document is categorized in to two categories and assumed to be D1 and D2. D1 have set of values or group of words with respect to the position. For example English character which comes in the respective position (odd) is grouped (clustered).Let us consider,

D1 = {$g_{(o)}1$, $g_{(o)}2$, $g_{(o)}3$, g(o)4, g(o)5, g(o)6,……$g_{(o)}n$} where g2, g3, g4, g5, g6 ……,gn are sub groups which contains group of words which belongs to a group in which the first character is same (similar). Again the words are arranged in to groups according the character sequence (first, second, ……,n character). Let "n" represents the last character in the word.

$g_{(o)}1$ = {w1 , w2,w3 ,w4,……,wn} , $g_{(o)}2${w1, w2,w3,w4,……..,wn),……$g_{(o)}n${w1,w2,w3,w4,……..,wn}

Step 2 Grouping of Words Starts with Same character

Group 1 (D1,g1)	**Group 1 (D2,g1)**
clustering, customer	has
communication, classification	**Group 2(D2,g2)**
Group 2(D1,g2)	been , biology
used	**Group 3(D2, g3)**
Group 3 (D1,g3)	many, medicine, marketing
in, including, image	**Group 4(D2, G4)**
Group 4 (D1,g4)	domain, document
application, animal	**Group 5 (D2, g5)**
administration ,and	processing, pattern
Group 5(D1,g5)	**Group 6 (D2,g6)**
on	recognition, retrieval
Group 6(D1,g6)	
so	

$W = \{w1_{(o)ch(1,2,3,4....n)}, w2_{(o)ch(1,2,3,4,...,n)}, w3_{(o)ch(1,2,3,4,...,n)}, w4_{(o)ch(1,2,3,4,...,n)},....., wn_{(o)ch(1,2,3,4,.....,n)}\}$. Let W be the total number of words in the particular cluster. Let $w1_{(o)}$, $w2_{(o)}$, $w3_{(o)}$, $w4_{(o)}$,……,$wn_{(0)}$ are group of words related to odd position category. $Ch(1,2,3...,n)$ represents various words with various combination are similar data to the group but nearer or for away with reference to the search data.

$W1_{(o)ch(1,2,3,4,.....,n)}$ have various combinations and they are consider as points in a cluster. The number of clusters depends on the maximum value and constraints. $W=max_{(ch(1,2,3,4,....,n),}$ Where $max_{(ch(1,2,3,4,....,n)}$ is the maximum number of clusters are depends on the combination of characters of similar property and w represents the total number of words which belongs to various sub groups or clusters under a domain D1.

In the first issue it is discussed for the words belongs to odd group (first character). The same can be applied for the second issue. Let us assume that D2 have set of values or group of words with respect to the position. For example English character which comes in the respective position (even) is grouped (clustered). In general let,

$D2 = \{g_{(e)}1, g_{(e)}2, g_{(e)}3, g_{(e)}4,g_{(e)}5,g_{(e)6,}$ ……,$g_{(e)}n\}$ where $g_{(e)}1, g_{(e)}2, g_{(e)}3, g_{(e)}4,g_{(e)}5,g_{(e)6,}$ ……$g_{(e)}$

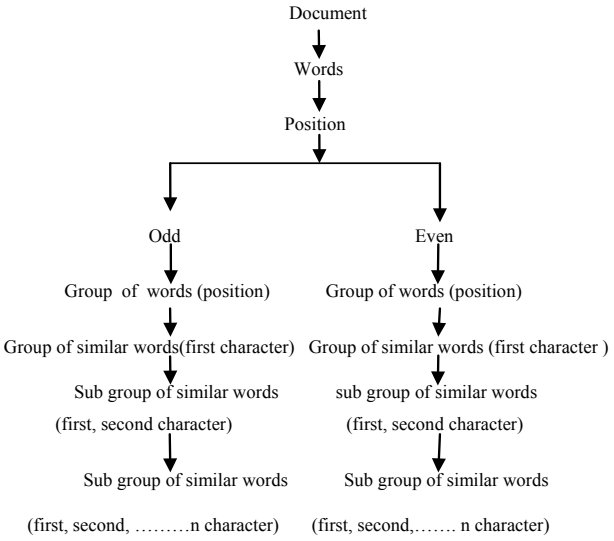

Figure 1. Segmentation of Documents.

n are sub groups which contains group of words which belongs to a group in which the first character is same (similar). Again the words are arranged in to groups according the character sequence (first, second,n character). Let "n" represents the last character in the word.

$g_{(e)}1 = \{w1 , w2,w3 ,w4, ……,wn\}$, $g_{(e)}2=\{w1,w2, w3,w4,……..,wn),……,g_{(e)}n\{w1,w2,w3,w4,……..,wn\}$

$W = \{w1_{(e)ch(1,2,3,4....n)}, w2_{(e)ch(1,2,3,4....n)}, w3_{(e)ch(1,2,3,4....n)}, w4_{(e)ch(1,2,3,4....n)},....., wn_{(e)ch(1,2,3,4....n)}\}$. Let W be the total number of words in the particular cluster.

Let $w1_{(o)}$, $w2_{(o)}$, $w3_{(o)}$, $w4_{(o)}$,……,$wn_{(0)}$ are group of words related to even position category. $Ch(1,2,3...n)$ represents various words with various combination are similar data to the group but nearer or for away with reference to the search data. $W1_{(o)ch(1,2,3,4,.....,n)}$ have various combinations and they are consider as clusters. The number of clusters depends on the maximum value. $W=max_{(ch(1,2,3,4,....,n),}$ Where $max_{(ch(1,2,3,4,....,n)}$ is the maximum number of clusters are depends on the combination of characters of similar property and w represents the total number of words which belongs to various sub groups or clusters under a domain D2.

Figure 2 Shows group of words. Figure 3 Shows the groups (clusters) based on the position and they are identified with the color "orange" for odd position and "even" position clusters are identified with "blue" color. Figure 4 Shows the sub groups (clusters) with respect to the position (first character of a word).

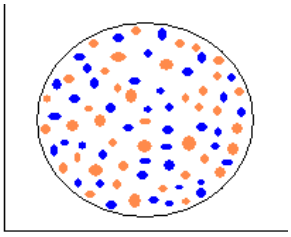

Figure 2. Group of words (Domain).

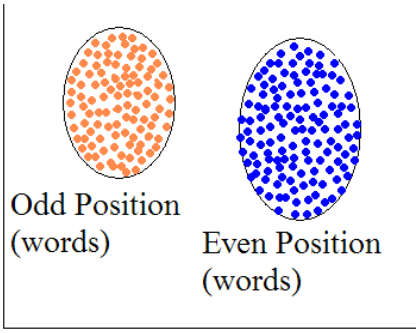

Figure 3. Clusters Based on Position.

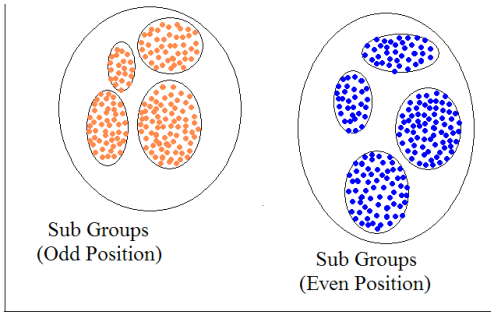

Figure 4. Clusters (Sub Group) Based on Position.

2.3 Phase 3: SOM

SOM is trained using unsupervised two-dimensional, discredited representation of the input space of the training samples. The aim is to learn a feature map from the spatially continuous input space, in which our input vectors live, to the low dimensional spatially discrete output space, which is formed by arranging the computational neurons into a grid. The stages of the SOM algorithm are initialization, sampling, matching, updating and continuation.

2.4 Phase 4: KMEANS

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into cells. The steps are collection, computation, separation, and grouping.

Algorithm - Step I (Positioning of Words)

```
Read w[] ; (words)
Declare
int i;
char a[] , b[];
char a,c,e,g,I,k,m,o,q,s,u,w,y = “odd”;
char b,d,f,h,j,l,n,p,r,t,v,x,z = “even”;
{
for ( i =0; i==\’o’; i++)
{
if (first_char_w[]==”odd”)
display a[]; (odd position words);
grouping of odd position words;
}
continue until last word;
{
if (first_char_w[]==”even”)
display b[]; (even position words);
grouping of even position words;
}
continue until last word;
}
```

Algorithm – Step II (Grouping of Words)

```
declare
 int i,j,k,n;
declare
  char a[100][100];
  Read n ; ( words )
    for(i=0;i<n;i++)
  {
    for(j=0;j<n;j++)
    {
  Read a([i][j]);
    } }
 Display (“Given word are :”);
  for(i=0;i<n;i++)
  {   for(j=0;j<n;j++)
    {  display a([i][j]);
    }
}
}
```

2.5 Phase 5: Clustering

Clustering usually groups keywords of documents into clusters and constructs them by selecting representative concepts from each cluster. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Our proposed Color based constraint clustering algorithm is discussed as below.

2.5.1 Cluster - Formation of Sub Sets

Clusters which are sub divided in to groups with similarities and dissimilarities. This can be explained with the following explanation. Let us assume that D is a Domain and it is segmented in to number of sub sets (groups). In each sub set numbers of points is grouped together with similarities and are dissimilar with other sub set. For example:

D = {D1, D2, D3, D4, ...Dn} where D1, D2, D3, D4 are the sub set of the domain. These are different groups with similar groups of data but dissimilar with each other are called clusters. The objects assigned to domains are nearer to them that is more frequently happened data.

D1 is a cluster which contains many points and is nearer to each other. This is represented by the distance between the points which are closely related to each other with similarities. The points which are consider for the process which are very closer to the mid value or centre value which is responsible for the iteration to achieve the goal. Until the condition satisfies the process continues (iterations) until the required is achieved.

D = { x1, x2 ,x3, x4, x5, x6,. . . . . ,xn}

p1 = {x1(a_g) , x2(a_g) , x3 (a_g) , ... , xn(a_g)}

p2 = {x1(ia_r) , x2(ia_r) , x3(ia_r) , ... , xn(a_r)} where a = active points; ia = inactive points; g = represents green color(active points); r = represents red color(inactive points) ; D = p1+p2. p1 = group of active elements. P2 = group of inactive elements. Inactive points are for away from the centroid and active elements are nearer to centroid and are very close to the search data.

Let i, j are two variables that are used to count the points or values which belong to similar group that is active or inactive. The color based constraints method is very easy to understand to identify the group or the nature of data. Here x1, x2, x3, x4, x5, x6, ……..,xn are domains (clusters) or a group (set of values). Points in the clusters are not considered for the process and these points are termed as inactive points. The points which are nearer or closely related to each point (neighborhood) and are closely to the centroid. These points are represented as active points. In different situations the various related points in the clusters are in the mode Active and Inactive and all depends on the situation of the process. The main purpose of the clustering is to optimize the number of iterations to achieve the goal.

2.5.2 Color Constraints

Color based constraint approach is easy to understand the grouping of data items (or clustering) and make the process in effective. The complexity may reduce though the constraints are similar. Color based constraints find the difference though they are nearer or linked. Depends on the query placed before for the processing and the set values are from the users pre-information. The color based active constraints reduce the complexity and pave the chance to select the more appropriate value or data for the process. In the same cluster i.e., the points in the clusters are near and distance. The distance among the points is measured with the centre point.

Figure 5 Represents a cluster which contains number of points (values). In the cluster $X1_{(a)}$,$X2_{(a)}$, $X3_{(a)}$, $Xn_{(a)}$, $X1_{(ia)}$,$X2_{(ia)}$, $X3_{(ia)}$, $Xn_{(ia)}$ represents points. Point "C" at the center with reference to other points. The diagram shows how the points are connected to each other with a distance. These points are assumed to be active and inactive depends on its use during the process. The active points are represented with "Green "(g) and the inactive points are represented with "Red" (r). The groups are identified with color and are easy to understand the process nature of the cluster values. "D" represents a domain and it is clustered and set into groups with number of sub clusters depends on the need.

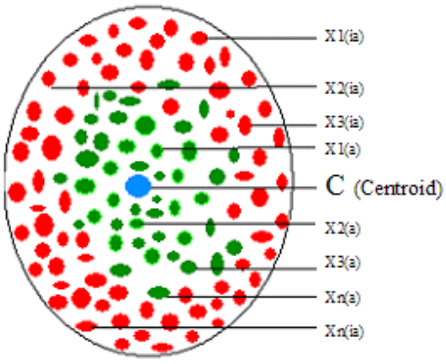

Figure 5. Color Constraint (Cluster) Representation.

Let "D" be the parent domain and D1, D2, D3 . . . ., Dn are siblings. Though the values in the clusters which are similar but dissimilar to other groups always have relations and the relations are used for the process when they are needed.

Let us assume that the point " C " is consider to be the centre point with respect to the entire cluster then the points for example $X1_{(a)}$,$X2_{(a)}$, $X3_{(a)}$, $Xn_{(a)}$ are nearer and $X1_{(ia)}$,$X2_{(ia)}$, $X3_{(ia)}$, $Xn_{(ia)}$ are away from the central point. Depending upon the situation , the distance points are consider to be inactive and represented with the color "Red " (r) and the nearer points are represented as active and represented with the color " Green"(g).

On the need the active and inactive points with reference to the reference value is grouped. With the color it is easy to understand the nature of the clusters and their characterization.

Let i , j are two variables which are used to segregate active and inactive point in a cluster to sub clusters with respect to "C" (centroid) until the last word in a given document Figure 6 Shows the sub clusters of active and inactive elements (word) of a given document.

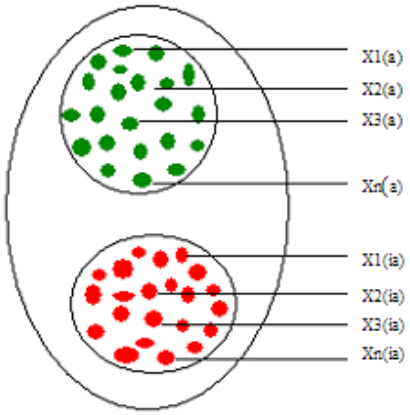

Figure 6. Color Constraints (Sub Clusters).

3. Performance Evaluation

The Core i5 processor is used. For system configuration, windows 7 operating system is used. The problem is done using the application server Visual Studio 2007. The front end C#.net and the Back end SQL Server is used. In this work our proposed Color Based Constraint (CBC) algorithm, SOM and K-Means algorithms performances are compared against time based frequency and reliability of retrieved documents. Here, the time needed to process the number of documents is analyzed and the results are shown in Figure 7. Reliability of retrieved documents can be made by using the number documents and the

Algorithm – Step III (Grouping of Words based on color constraint (active/inactive))

```
char x[100][100],aa[50][50],ia[50][50];
int i;
for (i=0;i<n;i++)
{
 read x[i];
}
read active;
read inactive;
for(i=0;i<n;i++)
{
 if (x==active)
 {
  aa[i] = active;
 }
 else
 {
  ia[i] = inactive;
 } }
display active elements;
for(i=0;i<n;i++)
{
 printf("active elements");
}
display inactive elements;
for(i=0;i<n;i++)
{ printf("inactive elements"); }
}
```

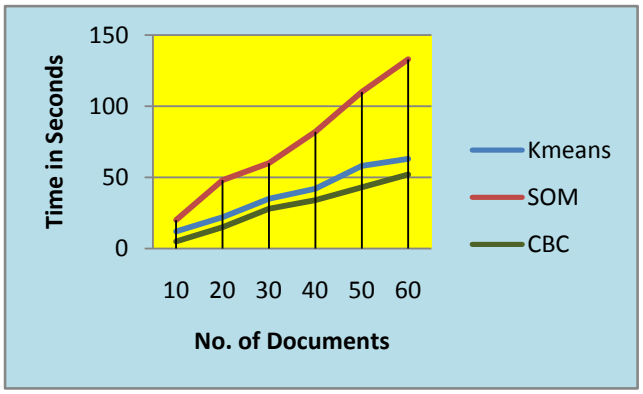

Figure 7. Time Based Frequency Analysis.

frequency measurement and the results are shown in Figure 8.

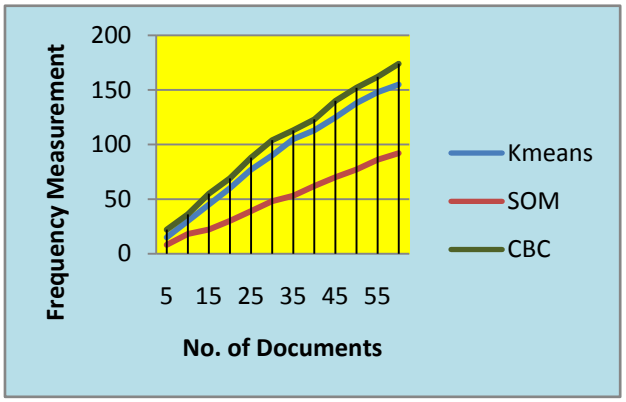

Figure 8. Reliable Documents Based Analysis.

4. Conclusion and Future Work

Thus a clustering based on color constraints enhances the text mining process. In this work the collected documents are preprocessed by stop word removal and stemming process. The preprocessed details are grouped based on position method then the documents are clustered using color based constraint clustering method, K-Means, and SOM algorithms. The algorithms performances are compared against time and reliability of retrieved documents. We proved our proposed color based constraint clustering algorithm out performs than K-Means, and SOM algorithms. In future this work can be extend to maximize the grouping of words with minimum latency and we plan to extend this work to develop an algorithm for maximize the grouping(clustering) of words in a document with color based constraints to increase the clustering performance for efficient text mining.

5. References

1. Shehata S, Karray F, Kamel M S. An efficient concept-based mining model for enhancing text clustering. IEEE Transactions On Knowledge and Data Engineering. 2010 Oct; 22(10):1360–71.
2. Feldman R, Dagan I. Knowledge Discovery in Textual Databases (KDT). Proceedings of First Int'l Conf. Knowledge Discovery and Data Mining. 1995 Aug; Montreal, Canada. p. 112–7.
3. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Proceedings of Knowledge Discovery and Data Mining (KDD) Workshop Text Mining. 2000 Aug 20-23; Boston, MA, USA. p. 525–6.
4. Nahm UY, Mooney RJ. A mutually beneficial integration of data mining and information extraction. Proceeding of 17th Nat'l Conf.Artificial Intelligence (AAAI '00). 2000; Austin,Texas. p. 627–32.
5. Davidson L Basu S. ACM transactions on knowledge discovery from data, A Survey of Clustering with Instance Level Constraints. 2007; W(x,z):1–41.
6. Prize KS, Fu, Jain A. An introduction to cluster analysis for data mining. 19th International Conference on Pattern Recognition ICPR 2008; 2008 Dec 8; Tampa, Florida.
7. Junker M, Sintek M, Rinck M. Learning for text categorization and information extraction with ILP. In: Cussens J, editor. Proceeding of First Workshop Learning Language in Logic; 2000; 1925:247–58.
8. Shehata S, Karray F, Kamel M. Enhancing text clustering using concept-based mining model. Proceeding of Sixth IEEE Int'l Conf Data Mining (ICDM). 2006; Hong Kong. p. 1043–8.
9. Hofmann T. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. Proceeding of 16th Int'l Joint Conf. Artificial Intelligence (IJCAI'99);. 1999 Jul 31-Aug 6; Stockholm, Sweden. p. 682–7.
10. Greene D, Cunningham P. Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering; In: Kok JN, Koronacji J, Mantaras RL, Matwin S, Mladenic D, Skowron A, editors. Proceeding of 18th European Conf. Machine Learning; 2007. Berlin, Heidelberg: Springer-Verlag; 2007. p. 140–151.
11. Bilenko MS, Basu S, Mooney R. Integrating constraints and metric learning in semi-supervised clustering. Proceedings of Int'l Conf. Machine Learning; 2004; Austin, TX. p. 11–8.
12. Xlong S, Aziml J, Fern XZ. Active learning of constraints for semi-supervised clustering. IEEE Transactions On knowledge And Engineering. 2014 Jan; 26(1): 43–54.
13. Basu S, Banerjee A, Mooney R. Active semi-supervision for pair wise constrained clustering. Proceedings of SIAM Int'l Conf Data Mining; 2004; Buena Vista,FL. p. 333–44.
14. Basu S, Davidson I, Wagstaff K. Constrained clustering: Advances in algorithms, theory, and applications. London: Chapman and Hall; 2008.
15. Ravi DD, Shamis L. A SAT-based framework for efficient constrained clustering. Proceeding of SIAM Int'l Conf. Data Mining; 2010; Columbus, Ohio, USA. p. 94–105.
16. Mallapragada PK, Jin R, Jain AK. Active query selection for semi-supervised clustering. Tampa, FL; 2008 Dec. p.14.
17. Khare A, Jadhav AN. An efficient concept based mining model for enhancing text clustering. International Journal of Advanced Engineering Technology. E-ISSN 0976-3945. 2010 Oct; 22(10): 1360–71.