ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 6, August 2012

Optimize Space Search Using FCC_STF Algorithm in Fuzzy Co-Clustering through Search Engine

Monika Rani, Satendra Kumar, Vinod Kumar Yadav

***Abstract*— Fuzzy co-clustering can be improved if we handle two main problem first is outlier and second curse of dimensionality .outlier problem can be reduce by implementing page replacement algorithm like FIFO, LRU or priority algorithm in a set of frame of web pages efficiently through a search engine. The web page which has zero priority (outlier) can be represented in separate slot of frame. Whereas curse of dimensionality problem can be improved by implementing FCC_STF algorithm for web pages obtain by search engine that reduce the outlier problem first. The algorithm FCCM and FUZZY CO-DOK are compared with FCC_STF algorithm with merit and demerits on the bases of different fuzzifier used. FCC_STF algorithm in which fuzzifier fused into one entity who have shown high performance by experiment result of values (A1,B1,Vcj,A2,B2) seem to less sensitive to local maxima and obtain optimization search space in 2-D for web pages by plotting graph between J(fcc_stf) and Vcj.**

***Index Terms*— Co-Clustering, Curse of Dimensionality, Fuzzy Clustering, Search Engine, Web Documents,.**

I. INTRODUCTION

Fuzzy clustering method is offered to construct clusters with uncertain to boundaries and allow that one object to belong to one or more cluster with some membership degree. Fuzzy co-clustering (bi-clustering) is a technique to simultaneously cluster data (web document) and feature(words inside the document).Fuzzy co-clustering have some advantages like dimensionality reduction, interpretable document cluster, improvement in accuracy due to local modal of clustering.

In this thesis solution for outlier and curse of dimensionality is provided. To reduce outlier fuzzy membership of each web document is found, so each document belong to particular cluster base on the priority of page assign in frame of five with additional to sixth frame of zero web page priority. And second problem is cure of dimensionality- fuzzy co-clustering with Ruspini's condition (FCR) and later FCCM [4] & fuzzy codok are added to this family with advance FCC_STF. In FCC_STF(Fuzzy co-clustering with single term fuzzifier) instead of two entities (like in fuzzy codok), here fuzzifier is fused into one entity.

***Monika Rani**, Department of Computer Science & Engineering ,YMCA University of science and Technology, Faridabad, India, 08860130090,*

***Satendra Kumar**, Department of Computer Science & Engineering, YMCA University of science and Technology, Faridabad, India, 08929511732,*

***Vinod Kumar Yadav**, Department of Computer Science & Engineering, Kamla Nehru Institute of Technology, Sultanpur, India, 09451611568 ,*

II. OUTLIER PROBLEMS

Data (web document) object that do not comply with the general behavior or model (cluster) of the data. These data object are outlier. Most data mining method discard outliers as noise or exceptions. However, in some application such as fraud detection, the rare event can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Outliers may be detected using statistical test that assume a distribution or probability model for the data or using distance measure where object that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation-based methods identify method identify outlier by examining differences in the main characteristics of object in a group.

To reduce outlier fuzzy membership of each web document is found, so each document belong to particular cluster. Thus searching of web document becomes efficient.

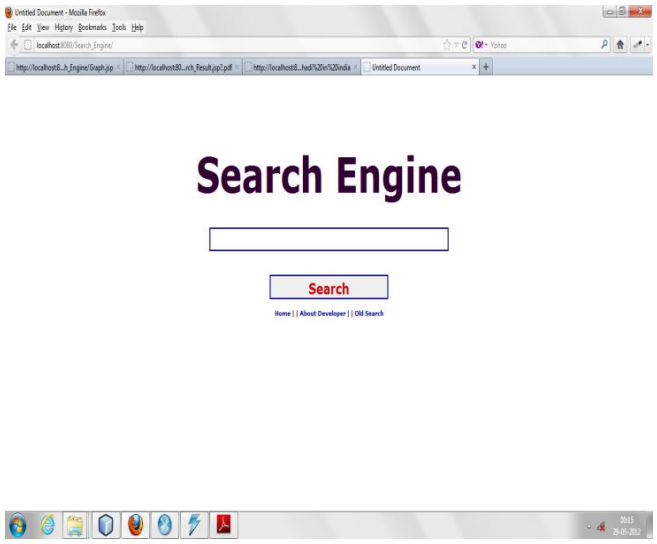
Here five frames are considered where we can use page replacement algorithm like.

A. FIFO (Fist in First Out)

B. LRU (Least Recently Use)

But as per preference here, PRIORITY (number of visit) basis is chooses mean on the basis of the priority page is set in the form of size 5 which the search engine will give the output for the high priority pages as given in the description (description name) by the description link.

ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 6, August 2012



. Fig.1 Search Engine without query

A. Search Engine

This search engine works in this manner as:

- Link Name: The source name is describe here from which data has to be taken means source for web pages .These web pages refer when the link description define the keywords and these keywords are search in source (data base) for particular web page.
- Link Description: Here we mention the statement (set of keyword) for which specific data is represented in web pages.
- Keyword: Keyword mention help in finding required web page and the key word separate by comma define particular key word for particular web page. Basically following step take place :

1-obtaine set of keyword when enter in search engine.

2-tokenize the text in the search engine and for particular keyword find appropriate web page from database.

3- break the keyword (remove stop word) , means if there are three key word then in search engine then , loop for k=3 , k=2 , k=1 is search bases for webpage in data base.

4- Do linguistic pre-processing a list of normalized tokens, which are present in web page by using key word as search bases of web pages.

5 –finally obtain the link for the web page in the five frames with high priority and one additional to these five frames6th with zero priority web pages.

B. Example

Link Name:- www.education.ac.in, education site is used in the link name with limit of varchar(45) to enter the character.

Link descriptor:- define the content show on the link or the matter present on the link its limit is define as varchar(450) to enter the character, here “we deal in all type of education” is given as link descriptor.
Keyword:– define the key word on which particular web page is found , the word limit of key word varchar (400) , by the key word match web page is brought from web data base.

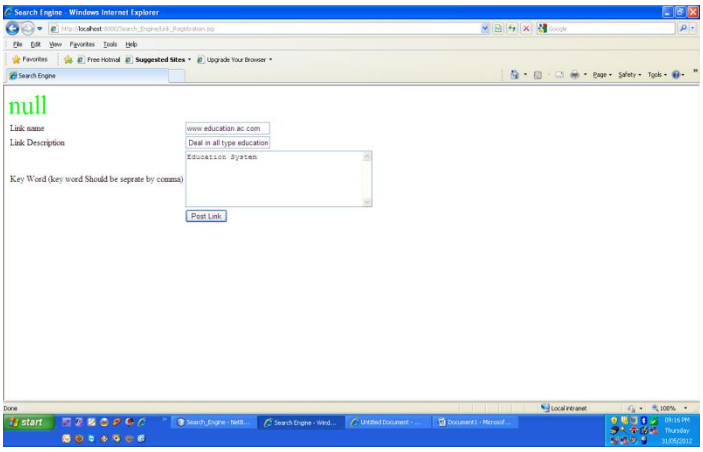


Fig.2 Link Registration

Once link name, descriptor and key word is mention then it is post. Below is view after post link.

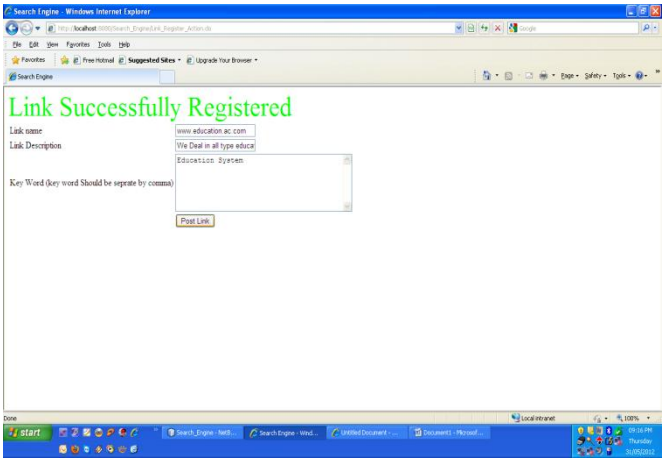


Fig.3 Link Successfully Registered

III. EXPERIMENTAL RESULT

The search engine made to search on the bases of keyword, example –“education system parameter”.

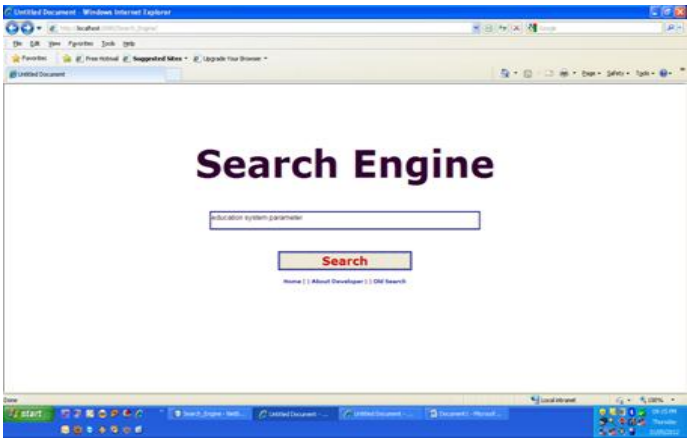


Fig.4 Search Engine with query

The main problem of fuzzy co-cluster is that the outlier are more in number , means there are web pages which lie outside the clusters.

Thus , this search engine reduces the number of outlier web pages by applying priority bases algorithm in five frame and soon on for next five frame with high priority and addition to this here 6th is use to represent the web page whose priority is "ZERO" by this we get 5 pages of high priority along with zero priority page , total six frame and in this approach the "ZERO" priority pages also found which was previously an outlier only .Thus this approach give chance to find out web page with zero priority .

- "ZERO" priority page is also important as its link descriptor might not define correctly and that page contain information which we require, just because of bad approach of searching we can loss that web page as outlier, thus this approach help to find the appropriate information as per requirement .Also improve the efficiency to search web document. Thus we obtain appropriate information as per requirement.
- After executing the query search engine show the total number of row is 18 and six frame in which five frame denote the web page of high priority and additional one show the frame for "zero" priority.

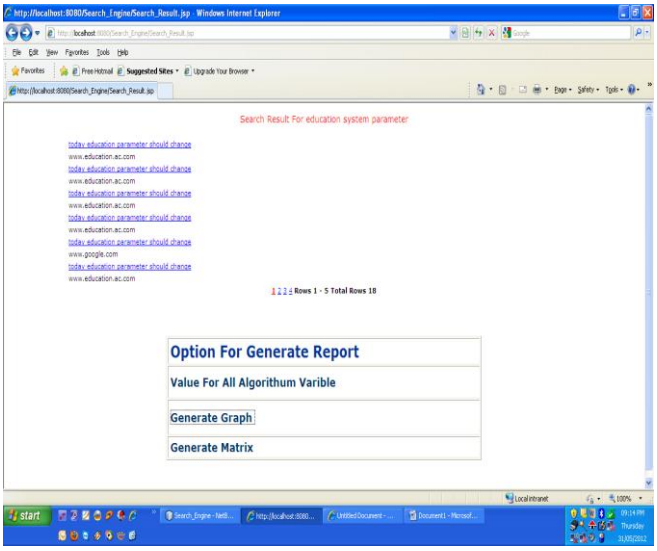


Fig.5 Frame for search engine on priority bases of web page

IV. CURSE OF DIMENSIONALITY PROBLEM

The curse of dimensionality :refer so various phenomena that arise when analyzing and organizing high dimensional space (100-1000) that do not occur in low dimensional setting such as the physical space commonly modeled with just 3-D when considering problem in dynamic optimization .To solve curse of dimensionality Fcc_Stf algorithm[5] is used.

FCC-STF Algorithm

1. Set parameters C, Tu, Tv and E;
2. Randomly initialize Uci;
3. REPEAT
4. Update Vcj using (equation);
5. If any Vcj is negative
6. Perform clipping and renormalization on Vci;
7. Update Uci using (equation);
8. If any Uci is negative
9. Perform clipping an renormalizing on Uci;
10. Until max (|Uci(t+1)-Uci(t)|)<E.

Where C, N, K denotes the number of co-clusters, documents and words.

- Uci and Vcj denote membership of document and word respectively
- dij denote the correlation degree between document and words
- Tu and Tv are the degree of fuzziness parameters to adjust the level of fuzziness in the document and word clusters respectively.

The FCC_STF new algorithm is proposed as compare to FCCM and fuzzy codok, which discovering relevant information effectively. Here values of variable like A1, B1, Vcj, A2, and B2.

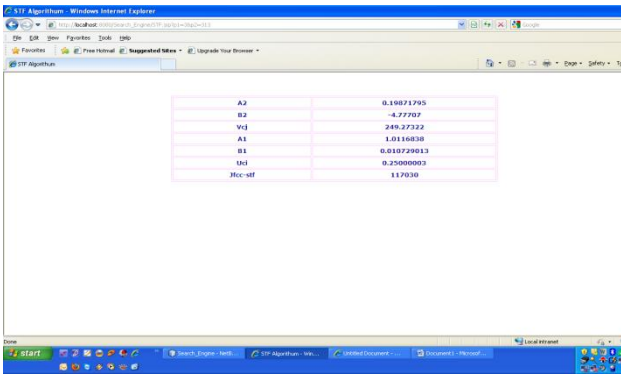
$$u_{ci} = \frac{2T\left(\sum_{j=1}^{K} v_{cj}^{\,2} - 1\right) - \sum_{j=1}^{K} v_{cj} d_{ij} + (A_1 / B_1)}{2T\left(\sum_{j=1}^{K} v_{cj}^{\,2} - 1\right)}$$

$$v_{cj} = \frac{2T\left(\sum_{i=1}^{N} u_{ci}^{\,2} - \sum_{i=1}^{N} u_{ci}\right) - \sum_{i=1}^{N} u_{ci} d_{ij} + (A_2 / B_2)}{2T\left(1 - 2\sum_{i=1}^{N} u_{ci} + \sum_{i=1}^{N} u_{ci}^{\,2}\right)}$$

where:

- $A_1 = 1 - \sum_{d=1}^{C} \left(\frac{2T\left(\sum_{j=1}^{K} v_{cj}^{\,2} - 1\right) - \sum_{j=1}^{K} v_{cj} d_{ij}}{2T\left(\sum_{j=1}^{K} v_{cj}^{\,2} - 1\right)} \right)$
- $B_1 = \sum_{d=1}^{C} \left[1 / \left\{ 2T\left(\sum_{j=1}^{K} v_{cj}^{\,2} - 1 \right) \right\} \right]$
- $A_2 = 1 - \sum_{q=1}^{K} \left(\frac{2T\left(\sum_{i=1}^{N} u_{ci}^{\,2} - \sum_{i=1}^{N} u_{ci}\right) - \sum_{i=1}^{N} u_{ci} d_{iq}}{2T\left(1 - 2\sum_{i=1}^{N} u_{ci} + \sum_{i=1}^{N} u_{ci}^{\,2}\right)} \right)$
- $B_2 = K / \left\{ 2T\left(1 - 2\sum_{i=1}^{N} u_{ci} + \sum_{i=1}^{N} u_{ci}^{\,2}\right) \right\}$

Table.1 Value of variable in Fcc_Stf algorithm



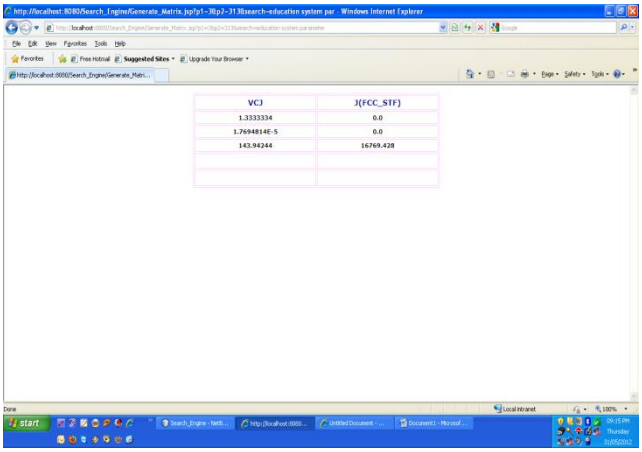
Optimize function J(fcc_stf) is defined for Fcc_Stf algorithm:

$$J_{FCC-STF} = \sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ci} v_{cj} d_{ij}$$

$$+T\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{K}\left[\left(u_{ci}+v_{cj}\right)-u_{ci}v_{cj}\right]^2$$

$$+\sum_{i=1}^{N}\lambda_i\left(\sum_{c=1}^{C}u_{ci}-1\right)+\sum_{c=1}^{C}\gamma_c\left(\sum_{j=1}^{K}v_{cj}-1\right)$$

Table 2 Table of Vcj and J(fcc_stf) value



Optimize Search Space**:** The value of vcj and J(fcc_stf) is use find optimize space search graph in 2-D :-

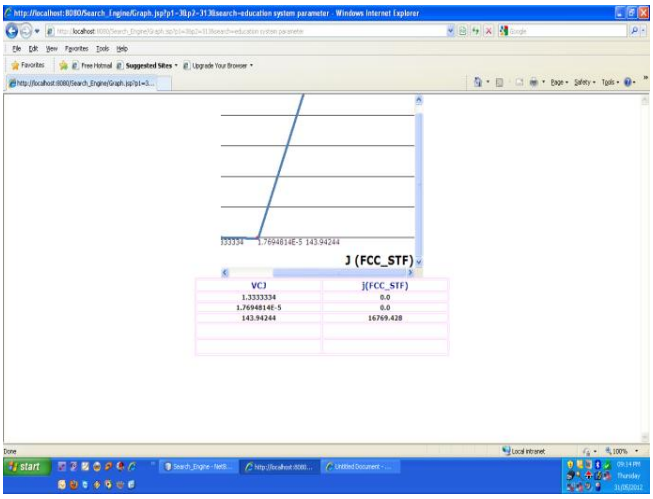


Fig.6 Graph between Vcj and J(Fcc_Stf)

V. CONCLUSION

1. The document get membership Fuzzy co-clustering provides efficient result in situation of vague and uncertainty. Fuzzy co-clustering could in many ways outperform standard fuzzy clustering when applied on a highly overlapping dataset.

2. To avoid the problem of always getting one fuzzy co-cluster containing all documents and words where imposing Ruspini's condition[6] on both document and word membership (in fuzzy co-clustering with Ruspini's condition) make it for word membership in co-cluster must be equal to 1 (in FCCM ,fuzzy codok, fcc_stf).

3. FCCM algorithm have of overflow problem which is mitigating by fuzzy codok but fuzzy codok allowed negative value therefore clipping and renormalization is require in the optimization.

4. In FCCM fuzzifier fuzzy entropy, in fuzzy codok fuzzifier fuzzy gini index , in FCC_STF we replace fuzzy codok's fuzzifirer with a new single term fuzzifier.

5. To reduce outlier fuzzy membership of each web document is found on the priority basic for search engine showing five frame of high priority along with one additional for "zero" priority web page(for outlier).

6. To reduce the problem of curse of dimensionality[5] use FCC_STF, show high performance by experimental result of values (A1,B1,Vcj,A2,B2) seem to less sensitive to local maxima this suggest that hyper dimensionality surface of FCC_STF.

7. The optimization search space is obtained by simplified 2-D plot of J(FCC_STF) and Vcj.

VI. FUTURE SCOPE

1. Search engine efficiency can be improve by using algorithm FIFO (first in first out), LRU (least recently use) , or the no. of visit for particular web page and frame size can be set accordingly .

2. For search engine we can use standard like XML format which can retrieve page in any language like English, French, italic, Germany etc., this will end communication barrier to retrieve web page.

3. In FCC_STF by using different fuzzifier parameter can improve efficiency of algorithm and also value of objective function along with optimize space for search in 2-D when graph is plotted between Vcj and J(fcc_stf).
4. Even by using web usage mining approach, client visit can be recorded in cookies and on the bases of test case can be made and fuzzy co-clustering can use it to improve retrieval efficiency.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques Morgar Kauffmann, pp, 21-25, 2000.Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

[2] Raymond Kosala and Hendrik Blockeel ACM SIGKDD, July 2000 Presented by Shan Huang, 4/24/2007Revised and presented by Fan Min, 4/22/2009.

[3] Web Usage Mining: By BamshadMobasher Dept. of Computer Science, DePaul University, Chicago.

[4] Fuzzy co-clustering of document and key word :Krishna Kummamuru, Ajay Dhawale, and Raghu Krishnapuram ,IBM India Research Lab ,Block 1, IIT, HauzKhas.

[5] Fuzzy Co-clustering of Web Document William-Chandra Tjhi and Lihui Chen Nanyang Technological University, Singapore School of Electrical & Electronic Engineering, Division of Information Engineering.

[6] A partitioning based algorithm to fuzzy co-cluster documents and words,William-Chandra Tjhi, Lihui Chen
[7] The web is the killer application for KDDM (R.kohavi kohavi-2001)

[8] From Wikipedia, the free encyclopedia.

[9] An introduction to information retrieval (e –book)Robert Cooley, Jaideep Srivastava Dept. of Computer Science, University of Minnesota, Minneapolis, MN cooley@cs.umn.edu, srivasta@cs.umn.edu.

[10] Robust weighted fuzzy C-means clustering, A.H.Hadjahmadi, M.M.Homayunpour and S.M. Ahadi.

AUTHORS PROFILE

Monika Rani born on 10 april 1988 at meerut, She received her B.Tech Degree in Information Technology from shobhit institute of engineering and technology, meerut(UPTU Lucknow),India in 2010, she has completed M.Tech in Information Tecnology from YMCA University of Science and Technology Faridabad, India. Her areas of interest in research are datamining, web mining, infromation retrieval, software testing.

Satendra Kumar born on 20 Feb 1988 at moradabad, He did his B.Tech in Computer Science and Information Technology from IET,MJP Rohilkhand University, Bareilly(U.P) in 2008, He has completed his M.Tech in Computer Engg. from YMCA University of Science and Technology Faridabad, India. Currently he is an assistant Professor in SITM Rewari at CSE department. His areas of interest in research are Database, Data Mining and Soft Computing.

Vinod Kumar Yadav born in Jaunpur, UP, India. He received the B.Tech. Degree in Computer Science and Information Technology in 2008 from I.E.T., M.J.P. Rohilkhand University Bareilly, India. He is currently pursuing M.Tech in Computer Science and Engineering from Kamla Nehru Institute of Technology, Sultanpur, UP, India. His areas of interest in research are Cryptography, Database and Network Security.