INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY (IJARET)

ISSN 0976 - 6480 (Print)
ISSN 0976 - 6499 (Online)
Volume 5, Issue 1, January (2014), pp. 07-17
© IAEME: www.iaeme.com/ijaret.asp
Journal Impact Factor (2013): 5.8376 (Calculated by GISI)
www.jifactor.com

IJARET
©IAEME

WEB PAGE CLUSTERING USING CEMETERY ORGANIZATION BEHAVIOR OF ANTS

Priyank Thakkar[1], **Samir Kariya**[2], **K Kotecha**[3]

[1]Assistant Professor, CSE Department, Institute of Technology, Nirma University,
Ahmedabad - 382 481, Gujarat, India
[2]Assistant Professor, IT Department, B. H. Gardi College of Engineering & Technology,
Rajkot - 361 162, Gujarat, India
[3]Director, Institute of Technology, Nirma University, Ahmedabad - 382 481, Gujarat, India

ABSTRACT

Clustering is the unsupervised classification of patterns (data items, observations or feature vectors) into groups (clusters). Clustering problem has been addressed by the researchers of many disciplines in different contexts. Due to the escalating amount of data available online, the World Wide Web has become one of the most precious resource for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. In this paper, we focus on web page clustering based on their content. A web page clustering system is valuable in web search for grouping search results into strongly related sets of documents. It can improve similarity search by focusing on sets of pertinent documents. At the same time, as the large variety of noisy information is embedded in web pages, web page clustering is much more intricate than pure-text clustering. This paper addresses web page clustering problem through the technique inspired by cemetery organization behavior of ants. Technique proposed by us begins by reducing the dimensionality of index of web pages with the application of Latent Semantic Indexing (LSI). Web pages are then transformed to two dimensional grid space using cemetery organization behavior of ants. Web pages represented in this two dimensional grid space are finally clustered using k-means algorithm. Paper also demonstrates impact of dimensionality reduction by means of LSI and distance measure on web page clustering results is also demonstrated.

Keywords: Web Page Clustering, Latent Semantic Indexing, Cemetery Organization Behavior of Ants.

1. INTRODUCTION

A web page clustering system can be precious in web search for grouping search results into strongly associated sets of documents. It can improve similarity search by centering on sets of pertinent documents [9]. At the same time, as the large diversity of noisy information is embedded in web pages, web page clustering is much more intricate than pure-text clustering.

Clustering also serves asvaluable technique for analyzing the Web. Duplications, patterns and other interesting structure on the web can be exposed by matching the content-based clustering and the hyperlink structure. There exist a range of types of clustering, depending on the way that clusters are characterized, the types of algorithms and the cluster properties employed for clustering [9].

2. WEB PAGE CLUSTERING

In web mining, web page clustering is one of the foremost preprocessing steps [9].Web page clustering puts together web pages in groups based on similarity or other relationship measure.

2.1 Representation of Web Pages

As long as an appropriate representation of objects exists, clustering can be applied to anyset of objects. Most general representation is the attribute–value or feature-value representation. In this representation a number of attributes (features) are recognized for the entire population, and each object is represented by a set of attribute-value pairs. Instead, if one fixes the order of the features, a vector of values (data points) can be used in its place. The document vector space model is precisely the identical type of representation, where the features are terms [9].

2.1.1 Vector Space Model

In vector space model, web pages are defined as vectors (or points) in a multidimensional Euclidean space where the terms represent axes (dimensions). Depending on the type of vector components (coordinates), there are three essential versions of this representation: Boolean, term frequency (TF), and term frequency-inverse document frequency (TFIDF) [9].

The easiest way to use a term as a feature in document representation is by inspecting whether the term is present in the document or not. That is why the term is considered as a Boolean attribute, and the representation is called Boolean.

In term frequency (TF) approach, function of the term counts, usually normalized with the document length is used as the coordinates of the web document $\overrightarrow{\mathrm{d_j}}$. For each term t_i and document $\mathrm{d_j}$ in document collection (D) (where each document is represented using m terms), term frequency is defined as

$$\mathrm{TF}(\mathrm{t_i}, \mathrm{d_j}) = \begin{cases} 0 & \text{if } \mathrm{n_{ij}} = 0 \\ \dfrac{\mathrm{n_{ij}}}{\sum_{k=1}^{m} \mathrm{n_{kj}}} & \text{if } \mathrm{n_{ij}} > 0 \end{cases} \qquad (1)$$

In the TFIDF representation, a product of TF and IDF components is used to calculate each component of the document vector. It is calculated as

$$\mathrm{d_i^j} = \mathrm{TF}(\mathrm{t_i}, \mathrm{d_j})\mathrm{IDF}(\mathrm{t_i}) \qquad (2)$$

where inverse document frequency of term t_i is defined as under:

$$\mathrm{IDF}(\mathrm{t_i}) = \log \frac{1 + |\mathrm{D}|}{|\mathrm{D_{t_i}}|} \qquad (3)$$

3. RELATED WORK

A new web page clustering algorithm QDC was proposed in [12].In this algorithm, user's query is used as a part of a reliable measure of cluster quality. The five key novelties proposed by this paper are: use of association between clusters and the query in query directed cluster quality guide, use of cluster depiction similarity in addition to cluster overlap to generate semantically coherent clusters, use of a new cluster splitting method to fix the cluster drifting or cluster chaining problem, use of query directed cluster quality guide to improve heuristic for cluster selection, use of ranking of the pages by relevance to the cluster to improve clustering results.

Methods such as hierarchical (divisive and agglomerate) clustering, partitioning (k-medoids, k-means, probabilistic) approaches, density-based clustering, grid-based clustering, fuzzyc-means clustering, Kohonen self-organizing maps and many more [6, 7, 13, 15] have been used for web page clustering by researchers. Many algorithms such as partitioning and hierarchical algorithms use data similarity measures to build clusters; when applied directly to web page data; the similarity based methods are not effective at producing semantically meaningful clusters. A method based on Harmony Search (HS) optimization is proposed in [8] to deal with web page clustering. By modeling clustering as an optimization problem, they recommend a pure HS based clustering algorithm that finds near global optimal clusters within a reasonable time. They have also hybridized K-means and harmony clustering to achieve better clustering results.

4. PROPOSED APPROACH

In our proposed approach, we follow the three steps as discussed under.

4.1 Generating Term-Document Matrix

First all the web pages are saved as the text documents and then the under mentioned steps are followed.

- Tokenized documents are produced by removing all punctuation marks and character strings without spaces.
- All the characters existing in the document are transformed to lowercase to carry out keyword matching in the document.
- All the stop words are removed and then resulting documents are used to generate term-document matrix. We have used TFIDF representation of the documents.

4.2 Latent Semantic Indexing

LSI proposed by Deerwester [20], uses a statistical technique, called Singular Value Decomposition (SVD) [21]. LSI begins with $m \times n$ term-document matrix A. SVD factors matrix A into product of three matrices, i.e.

$$\mathrm{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} \qquad (4)$$

where U and V are matrices of the left and right singular vectors respectively. Σ is the diagonal matrix of singular values. A key feature of SVD is that we can delete some insignificant dimensions in the transformed space to optimally approximate matrix A. The importance of the dimensions is indicated by the magnitudes of the singular values in Σ. LSI approximates A with a rank k matrix

$$\mathrm{A_k = U_k \Sigma_k V_k^T} \qquad (5)$$

where $\mathrm{U_k}$ includes the first *k* columns of the matrix U and $\mathrm{V_k^T}$ includes first *k* rows of matrix V. $\Sigma_\mathrm{k} = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_\mathrm{k})$ is the first k factors. In our approach we consider document *d* as $m \times 1$ matrix and then to transform it in the reduced dimension space we use

$$\hat{\mathrm{d}} = \mathrm{d^T U_k \Sigma_k} \qquad (6)$$

where k is the rank of the matrix and it allows us to control the dimensionality of the document in the transformed space.

After reducing the dimensionality of the index of the web pages as discussed above, we have applied clustering algorithm inspired by the cemetery organization behavior of ants. This algorithm basically transforms web pages to the two dimensional grid space. This algorithm is discussed in the next sub-section.

4.3 Clustering Based on Cemetery Organization Behavior of Ants

The general idea is that isolated items (web pages in our case) should be picked up and dropped at some other location where more items of that type are present [18]. The Algorithm pioneered by Lumerand Faieta [22] states to project the space of attributes onto some lower dimensional space, typically of dimension $\mathrm{z} = 2$.LF algorithm works as follows. Instead of embedding the set of web pages into R^2, this embedding is approximated by bearing in mind a grid, that is, a subspace of Z^2. Ants that are moving in this discrete space can straightforwardly perceive a surrounding region of area s^2(a square $\mathrm{Neigh}_{(\mathrm{s} \times \mathrm{s})}$of $\mathrm{s} \times \mathrm{s}$ sites surrounding site *r*). Let $\mathrm{d}\left(\mathrm{o}_{\mathrm{i},}\mathrm{o}_\mathrm{j}\right)$ be the distance between two web pages $o_i$and o_j in the space of attributes. Letusalsoimaginethatanantissituatedatsite*r*at time *t*, and finds a web page o_i at that site. The “local density”f $(\mathrm{o_i})$ with respect to web page o_i at site *r* is given by

$$\mathrm{f(o_i)} = \begin{cases} \dfrac{1}{\mathrm{s}^2} \sum_{\mathrm{o_j} \in \mathrm{Neigh_{s \times s}}(\mathrm{r})} \left[1 - \dfrac{\mathrm{d(o_i, o_j)}}{\alpha}\right] & \text{if f} > 0, \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

$\mathrm{f(o_i}$)is a measure of the average similarity of web page o_i with the other web pages $\mathrm{o_j}$ present in the neighborhood of o_i. Scale for dissimilarity is defined by α. It is important as it determines when two items should or should not be located next to each other.

Lumer and Faieta has defined picking up and dropping probabilities as follows:

$$\mathrm{p_p(o_i)} = \left(\frac{\mathrm{k_1}}{\mathrm{k_1 + f(o_i)}}\right)^2 \qquad (8)$$

$$\mathrm{p_d(o_i)} = \begin{cases} \mathrm{2f(o_i)}, & \text{if } \mathrm{f(o_i) < k_2} \\ 1 & \text{if } \mathrm{f(o_i) \geq k_2 s} \end{cases} \qquad (9)$$

where$\mathrm{k_1}$ and$\mathrm{k_2}$are two constants.

High-Level Description of the Lumer-Feita Algorithm [18][22]

```
  /*Initialization*/
  Foreveryweb page$o_i$do
    Place$o_i$randomlyongrid
  EndFor
  For allantsdo
    Placeantatrandomlyselectedsite
  EndFor
  /*Mainloop*/
  For t=1to$t_{max}$ do
  For allantsdo
     If ((ant unladen) and (site occupied by web page $o_i$)) then
         Compute $f(o_i)$ and $p_p(o_i)$
         Draw Random Real number R between 0 and 1
         If ($R \leq p_p(o_i)$)then
             Pick up web page $o_i$
         End If
     Else If ((ant carrying web page $o_i$) and (site empty)) then
         Compute $f(o_i)$ and $p_d(o_i)$
         Draw Random Real number R between 0 and 1
         If ($R \leq p_d(o_i)$)then
             Pick up web page $o_i$
         End If
     End If
     Move to randomly selected neighboring site not occupied
      by other ant
 End For
 End For
 Print location of web pages
```

4.4 Modifications to Lumer-Faieta Algorithm [18][22]

Lumer and Faieta algorithm depicted above tend to produce more number of clusters than desired. This was also the case in our initial implementation. They have suggested three features to correct this behavior.

- Ants with different moving speeds: Let v be the speed of an ant (v is the number of grid units walked per time unit by an ant along a given grid axis); v is distributed uniformly in $[1, v_{max}]$. We use $v_{max} = 6$ in our simulations. v also influences, through the function f, the tendency of an ant to either pick up or drop a web page:

$$f(o_i) = \begin{cases} \frac{1}{s^2}\sum_{o_j \in \mathrm{Neigh}_{sXs}(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha\left(1 + \frac{v-1}{v_{max}}\right)}\right] & \text{if } f > 0, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Therefore, slow moving ants are more careful than fast ants in their judgment of the average similarity of a web page to its neighbors. Clusters over diverse scales are developed simultaneously due to the miscellany of ants.

- A short-term memory: Last *m* web pages that ants have dropped can be memorized by them along with their locations. The ant compares the properties of the web page with those of the *m* memorized web pages and goes toward the location of the most similar instead of moving randomly each time it picks up a web page.This behavior pilots to a reduction in the number of equivalent clusters, since similar web pages have a low probability of instigating independent clusters.
- Behavioral switches: Web pages are less and less likely to be manipulated as clusters of similar objects form. Therefore it can be said that the system demonstrates some sort of self-annealing. To allows a "heating up" of the system to escape local non-optimal configurations Lumer and Faieta have added the opportunity for ants to start destroying clusters if they haven't performed an action for a given number of time steps.[18][22].

We have integrated first two features in our implementation. Incorporation of the third feature deteriorated the quality of clusters formed in our simulations.

5. RESULTS AND DISCUSSION

5.1 DataSets

To verify the effectiveness of the web page clustering system proposed in this paper, we have conducted experiments on three different datasets as shown in Table 1. The aim behind using different datasets for experimentation is to prove the consistency of the proposed system. First two datasets (Bank Search [25] and Syskill & Webert [23][24]) are publically available and the third dataset is created by us by downloading 364 pages of eight different categories (we name this dataset “All Text Combine” Dataset). From the total available documents in Bank Search dataset [25], we have selected 300 documents each from two different categories and used them in our simulations.

Table 1. Data Set Statistics

	Number of	Number of Attributes	Number of Clusters
Bank Search	600	22513	2
Syskill & Webert	331	21231	4
All Text Combine	364	25927	8

5.2 Evaluation Measures

Assume a confusion matrix with m classes (number of rows) and k clusters (number of columns) as shown in Table 2. For the number of web pages from cluster *j* that belong to class *i*, precision and recall are defined as follows.

$$\text{Precision } P(i,j) = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}} \qquad (11)$$

$$\text{Recall } R(i,j) = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}} \qquad (12)$$

Table 2. Confusion matrix for *m* classes and *k* clusters

Classes	Clusters				
	1	…	j	…	k
1	n_{11}	…	n_{1j}	…	n_{1k}
…					
I	n_{i1}	…	n_{ij}	…	n_{ik}
…					
M	n_{m1}	…	n_{mj}	…	n_{mk}

We have used F-measure as the evaluation measure for our system.More exact account for the error is provided by F-measure than the overall accuracy. F-measure is actually the harmonic mean of precision and recall.

$$\mathrm{F(i,j)} = \frac{2 \cdot \ \mathrm{P(i,j)} \cdot \mathrm{R(i,j)}}{\mathrm{P(i,j)} + \mathrm{R(i,j)}} \qquad (13)$$

Maximum of F(i, j) over all clusters is taken and then sum across classes. As classes normally include different numbers of documents, their contribution to the sum is weighted with the fraction of documents in each. Thus, we obtain the F-measure for the entire clustering.

$$\mathrm{F} = \sum_{\mathrm{i=1}}^{\mathrm{m}} \frac{\mathrm{n_i}}{\mathrm{n}} \max_{j=1,\dots,k} F(i,j) \qquad (14)$$

where $\mathrm{n_i} = \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{n_{ij}}$(the number of web pages belonging to class i, or total of row i) and $\mathrm{n} = \sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{n_{ij}}$ (the total number of web pages in the sample)

To reduce the stochastic noise in evaluation, we have simulated our algorithm 10 times and all the results which are presented in this paper are averaged over these 10 runs. Fig. 1, 2 and 3 depict the results on Bank Search, Syskill & Webert and All Text Combine datasets respectively. All the figures show web pages mapped on two dimensional grid space. In our experiments best results are achieved while using 1-cosine similarity as the distance measure in Bank Search dataset. In case of other two datasets, best results are produced when Euclidean distance is used as the distance measure.

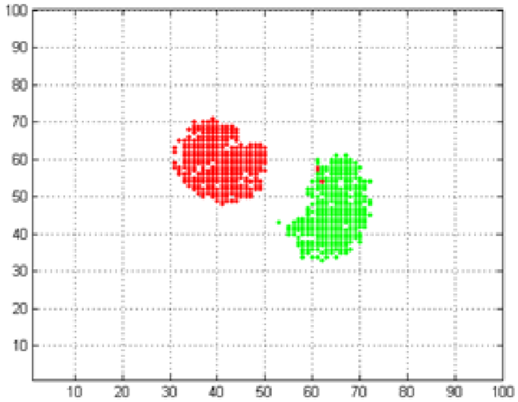

Fig. 1: Results on Bank Search Data Set

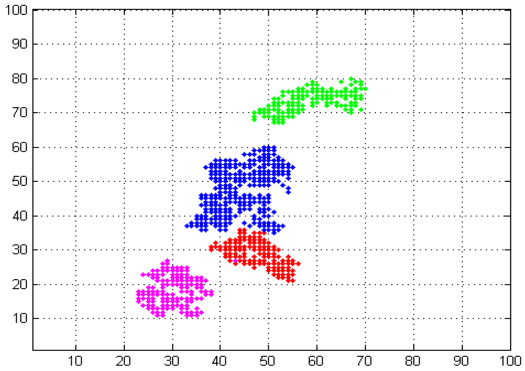

Fig. 2: Syskill and Webert Data Set

Once web pages are mapped onto two dimensional space, k-means algorithm is used to produce the final clustering.

Table 3, 4 and 5 illustrate final clustering results in terms of F-measure for Bank Search, Syskill and Webert and All Text Combine data sets respectively. We have evaluated our proposed system using 1-cosine and Euclidean distance as the distance measure for all three data sets. Results also divulge impact of dimensionality reduction on clustering results as we have varied number of dimensions from 300 to 700. We have also simulated our algorithm without reducing the dimensionality of the web page corpus. However results are far superior when we have reduced the dimensionality appropriately.

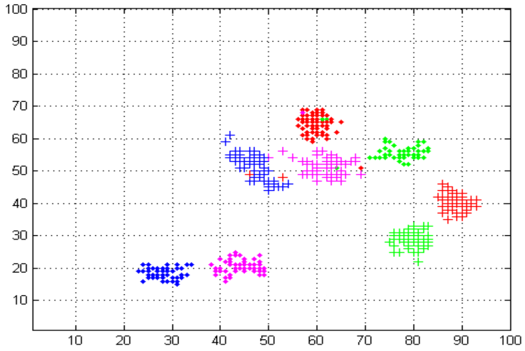

Fig. 3: All Text Combine Data Set

Table 3. F-Measure for Bank Search Data Set

No. of Dimensions	Distance Measure	
	Euclidean	**1 - Cosine Similarity**
300	0.6366	0.8340
400	0.6566	0.8747
500	0.6634	**0.8859**
600	0.6687	0.8800
700	0.6313	0.8523
All (22513)	0.5307	0.6185

As revealed from Table 3, in Bank Search dataset, best results of 0.8859 is obtained when number of dimensions is 500 and 1-cosine is used as the distance measure.

Table 4. F-Measure for Syskill and Webert Data Set

No. of Dimensions	**Distance Measure**	
	Euclidean	**1 - Cosine Similarity**
300	0.8882	0.2910
400	0.9049	0.4012
500	**0.9155**	0.5084
600	0.9008	0.5113
700	0.8904	0.4788
All (21231)	0.6658	0.3196

As shown in Table 4, in Syskill & Webert dataset, best results of 0.9155 is attained when number of dimensions is 500 and Euclidean is used as the distance measure.

In All Text Combine dataset, as indicated in Table 5, best results of 0.8964 is achieved when number of dimensions is 500 and Euclidean is used as the distance measure.

Fig. 4 depicts F-measure over 10 different runs for each of the dataset. In each of the run we have used 500 dimensions foreach of the dataset.

Table 5. F-Measure for All Text Combine Data Set

No. of Dimensions	**Distance Measure**	
	Euclidean	**1 - Cosine Similarity**
300	0.8759	0.6527
400	0.8928	0.6859
500	**0.8964**	0.6911
600	0.8939	0.6966
700	0.8801	0.6759
All (25927)	0.6427	0.4745

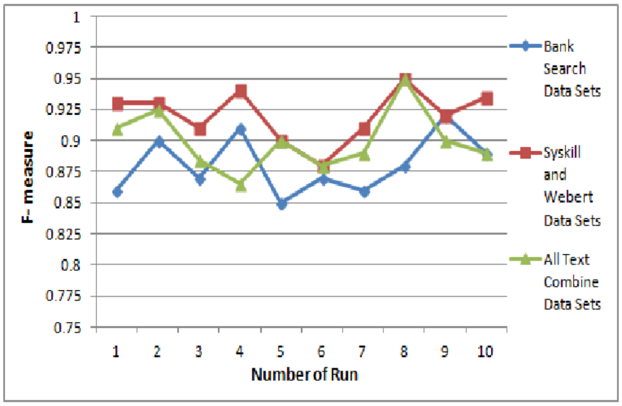

Fig. 4: F-measure over 10 different runs for each data sets

6. CONCLUSION AND FUTURE WORK

This paper proposes technique that combines algorithm inspired by cemetery organization behavior of ants and k-means for clustering the web pages. Latent Semantic Indexing (LSI) is first used to represent the documents in lower dimension space and then algorithm based on cemetery organization behavior of ants is used to transform the web pages on two dimensional grid space. Once the web pages are represented in two dimensional grid space, k-means is used to cluster them. Implementation results are promising and show the effectiveness of the proposed framework. Impact of dimensionality reduction is also demonstrated. It can be seen that selecting right number of dimensions to represent web pages improves the result of clustering. Clustering results are best when an appropriate number of dimensions are used to represent web pages. In future, we would like to incorporate link information among web pages in the representation and evaluate the impact of this on the results.

REFERENCES

1. K.-C. H. Chun-Wei Tsai and M-C. Chiang, “Ant colony optimization with dual pheromone tables for clustering”, IEEE International Conference on Fuzzy Systems, June 2011.
2. H. K. and A. K., “Clustering Algorithm Employ in Web Usage Mining: An Overview”, Bharati Vidyapeeths Institute of Computer Applications and Management, New Delhi, March 2011.
3. W. Xiong and C. Wang, “A novel hybrid clustering based on adaptive ACO and PSO”, IEEE, 2011.
4. M. V. S. G. Mr. Pankaj K. Bharne and M. S. K. Yewale, “Data clustering algorithms based on swarm intelligence”, IEEE, 2011.
5. O. M. Jafar and R. Sivakumar, “Ant-based clustering algorithms: A brief survey”, International Journal of Computer Theory and Engineering, October 2010.
6. K. Gupta and M. Shrivastava, “Web usage mining clustering using hybrid FCM with GA”, International Journal of Advanced Computer Research, June 2010.
7. V. B. Praveen, “Influence of various clustering algorithms on web personalization”, Proceeding of the International Workshop on Machine Intelligence Research, 2009.
8. Rana Forsati, Mehrdad Mahdavi, Mohammadreza Kangavari and Ba- nafsheh Safarkhani, “WEB PAGE CLUSTERING USING HARMONY SEARCH OPTIMIZATION”, Department of Computer Engineering, Tehran Azad University, Tehran, Iran, IEEE, 2008.
9. Z. Markov and D. T. Larose, “Data Mining The Web: Uncovering Patterns in Web Content, Structure, and Usage”, John Wiley & Sons, 2007.
10. B. Liu, “Web Data-Mining: Exploring Hyperlinks, Content, and Usage Data”, Springer, 2007.
11. J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, ElsevierInc. 2006.
12. Daniel Crabtree, Peter Andreae and Xiaoying Gao, “Query Directed Web Page Clustering”, Victoria University of Wellington New Zealand, IEEE, 2006.
13. R. Xu and D. W. II, “Survey of Clustering Algorithms”, IEEE Trans. On Neural Networks, May 2005.
14. L. Wanner, “Introduction to Clustering Techniques”, July 2004
15. Kate A. Smith and Alan Ng, “Web page clustering using a self- organizing map of user navigation patterns”, Monash University, P.O. Box 63B, Victoria 3800, Australia, Elsevier Science, 2003.

16. Jerome Moore and Eui-Hong, “Web Page Categorizing and feature selection using Association Rule and Principal Component Clustering”, University of Minnesota, IEEE, 2000.
17. A.K. Jain, M.N.Murty, and P.J.Flynn, “Data clustering: A review”, ACM Computing Surveys, September 1999.
18. E. Bonabeau, M. Dorigo, and G. Theraulaz, “Swarm Intelligence: From Natural to Artificial Systems”, Sante Fe Institute Studies in the Sciences of Complexity, Oxford University, 1999.
19. M. V. Shrivastava and M. N. Gupta, “Performance improvement of web usage mining by using learning based k-mean clustering”, International Journal of Computer Science and its Applications.
20. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, “Indexing by Latent Semantic Analysis”, Journal of the American Society for Information Science, 41, pp. 391–407, 1990.
21. G. H. Golub, and C. F. Van Loan,“Matrix Computations”, The Johns Hopkins University Press, 1983.
22. Lumer, E., and B. Faieta, "Diversity and Adaptation in Populations of Clustering Ants", In Proceedings of third International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3, 499-508. Cambridge, MA: MIT Press, 1994.
23. Michael Pazzani, Jack Muramatsu and Daniel Billsus. “Syskill Webert: Identifying Interesting Web Sites”, In AAAI, Vol. 1(1996), pp. 54-61.
24. Bache, K. & Lichman, M., UCI Machine Learning Repository [http://archieve.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.
25. Mark P. Sinka, David W. Corne. “The BankSearch Web Document Dataset: Investigating Unsupervised Clustering and Category Similarity”, Journal of Network and Computer Applications 28 (2004), 129-146, Science Direct.
26. Alamelu Mangai J, Santhosh Kumar V and Sugumaran V, “Recent Research in Web Page Classification – A Review”, International Journal of Computer Engineering & Technology (IJCET), Volume 1, Issue 1, 2010, pp. 112 - 122, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
27. Sudip Kumar Sahana, Dr. Aruna Jain and Abijit Mustafi, “A Comparative Study on Multicast Routing using Dijkstra’s, Prims and Ant Colony Systems”, International Journal of Computer Engineering & Technology (IJCET), Volume 1, Issue 2, 2010, pp. 16 - 25, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
28. R. Manickam, D. Boominath and V. Bhuvaneswari,, “An Analysis of Data Mining: Past, Present and Future”, International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 1, 2012, pp. 1 - 9, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.