USAGE OF SIMILARITY METHODS AND CLUSTERING METHODS FOR WEB PAGES CLUSTERING

San San Tint

Department of Research and Development II, University of Computer Studies, Mandalay, Myanmar

Abstract

We have been studying different kinds of papers about web pages clustering techniques with many algorithms. Almost papers published in the Web are used popular algorithms. Among them similarity methods and clustering methods are very popular algorithms for extraction data in the Web. We use some methods to summarize by means of well known those methods. Some papers are applied to different methods. It is motivated to collect and summarize what are differences between them. This paper presents a comprehensive analysis of different methods and intends to be more recognized those methods. Most of sections are well composed for detail explanations about papers studied as a list of numbers and a table. Today web pages are saturated on the Internet. Content of web pages are extracted from cluster of web page using related methods. Clustering systems are useful to extract knowledge. The results obtained show that the status of using similarity and clustering method provides the best results for people who study about web.

Keywords

Clustering, Common Approach, Distance, Similarity, Web Pages

1. Introduction

Nowadays, the Internet is increasing in size at the rapid rate so locating information is becoming more difficult. Information source containing the data was extracted to perform a task. We can access data when the information source has been found in the web [1] [2] [3]. The metadata such as a document's keywords, description [4] assists web clustering and its proving plays an important role in the improvement of web applications. A common approach to capture this kind of metadata is called web- page categorization [5]. A set of classes representing a topic area that we aim at developing scope containing the page assigned in our case study.

Each keyword is searched for its hypernym tree in the WordNet. These keywords are replaced with the common hypernym within the specified threshold, the replacement term are reconstructed to use in the self-organizing maps (SOM) by input as a vector representation of the keyword. The web pages are clustered by SOM algorithm how these are related each other. The important technique for analyzing data is Clustering. Clustering is similar to the classification but there has some differences. Classification when classes are not known [6] [7] [8]. Both classification and clustering have benefit from the integration of prior external class knowledge, which reflects specific classification concept or organization. In the studying of Classification in which domain knowledge, document clustering is classified by many clustering algorithms as well as similarity methods like as [9] [10] [11] [12] [13]. We leave other methods such as Hamming distance, Levenstein distance, and K nearest Neighbor because of getting papers to

study from internet randomly not including those papers related them. This paper attends to know what methods are most widely used in the web page clustering.

2. Literature Review

In text clustering methods, three famous methods are an ascending hierarchical clustering method, a SOM-based clustering method and an ant-based clustering method. Most of papers based on WordNet use Self- Organization Maps. They who implement clustering usually apply the methods to examine in several experiments using 2 similarity measurements: the cosine distance, and the Euclidean distance. For evaluation, they commonly use the F-measure. The results obtained from the SOM-based clustering method using the cosine distance provide the best results [14] [15] [16] [17] [18] [19] [20].

A semantic similarity measure based on documents in topic maps is developed by [20]. In later search and extraction, topic maps become an industrial standard for knowledge related with mining. After transforming documents into a topic map, the similarity between a pair of documents based coded knowledge represents as a correlation between the sub-trees (common patterns). The study on the text mining datasets can be very revealing about of new similarity measure that is somewhat effective than as compared to commonly used old ones in that study like document clustering [21].

In [22], the method proposed in the result section was evaluated with metric values of number of plants, computation - time and memory usage. The knowledge engineering approach achieving results presented that not to waste space of data and to be faster knowledge retrieval when we had unknown variables.

SOM compromises a powerful model for Web mining and defines a visual overview of a set of Web documents. A document SOM can order spread of the documents in the set semantically. In document collections, SOM-based visualization system is a powerful information retrieval tool as well as a useful browser for a set of Web documents. A user needs a useful clustering algorithm related the content of the text collection when the user has lack of knowledge. Most of framework can apply some approaches that are based on self-organizing maps and the unsupervised top-down neural network based approach with different domains and languages [23].

3. Document Clustering

Document clustering is text processing that group's document with similar concepts. An unsupervised learning approach has been applied rather than supervise learning because of no need to guide the training process when the topical information is unavailable.

The supervised learning approach is useful one when the topical information is available for the training process in preclassified information guides. Although the differences between two approaches, both classification and clustering techniques can catch advantage of different weights of words' relationships even the corpus providing topical information.

In the system [23] [24], the users not only exploit the domain knowledge but also reduce the gap between concepts. And if they derive data clustering decision using different weights. The self organization map is a network for guided or unguided clustering. SOM connects among things that are nonlinear projection, vector quantization, and data - clustering functions.

Self- Organization approach is motivated in a similar manner but further integrate topical and semantic information from WordNet. Because a document training set with preclassified

information applies relationships between a word and its preference class, a novel document vector representation approach is used to extract these relationships for document clustering. Furthermore, this system includes merging statistical methods, competitive neural models and semantic relationships from symbolic WordNet [25].

4. Self- Organization Maps

In [20] and [23], it has created the self-organizing maps (SOM) as a particular kind of neural networks. SOM has multiple views on the different definitions that are the followings: a model of specific aspects of biological neural nets, a model of unsupervised machine learning and an adaptive knowledge representation scheme. SOM is a tool for statistical analysis and $i = c$. At the beginning of the learning process, the neighborhood must be a wider area. During learning process, neighborhood area's width and height must decrease.

At the same time, the updating steps imposed the forms of a globally ordered map. In a hexagonal map topology, a map unit has six immediate neighbors, and those prefer the common topologies. Only a hexagonal type of the two-dimensional array like grid of neurons with the SOM map continues to be a planar rectangle so that the hexagonal neighborhood topology effect is gained by shifting rows number of the rectangular map to the right and keeping rows untouched.

A rectangular topology, a map unit has only four immediate neighbors while the number of neighbor units affected during the learning compared to six in the hexagonal topology [24] [26] [27] [28].

After the training and preparing a map, its corresponding vectors have connected with stationary values. A topology-preserved map has been come out as a result. Similar reference vectors close to each other while dissimilar ones are far from each other on the map. Two inputs related each other in the input data space are mapped onto the same or nearly the same neurons on the map. Each neuron owns reference vector representing similar data items of the input space, and neighboring neurons with similar vectors made a cluster [29] [30].

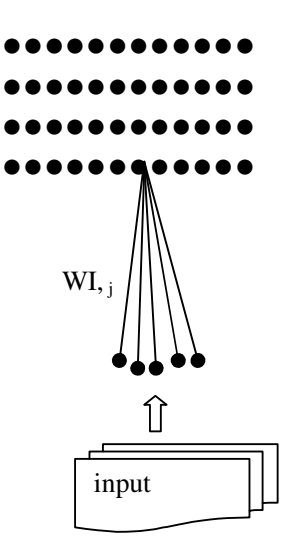

Figure1. Two dimensional Maps

5. Methods of Distance /Similarity Measure

Following subsections show briefly about distance and similarity methods [18 - 33]. We describe some equations which are the most well known and commonly used in the fields related with similarity measure. In this paper, we summarize formula in term of equations to collect to choose for applications in the interest areas. There are thirteen equations in this paper.

5.1. Euclidean distance

$$d(i,j)=\sqrt{\sum_{1}^{n}(x_{i1}-x_{j1})^2} \tag{1}$$

5.2. Cosine Distance

$$Cos(d_i,d_j)=\frac{\sum_{t_k}[TF\times IDF(t_k,d_i)]\cdot[TF\times IDF(t_k,d_j)]}{\|{d_i}^2\|\cdot\|{d_j}^2\|} \tag{2}$$

5.3. Cosine Similarity

$$\mathrm{SIM_c}\ (\overrightarrow{t_a},\overrightarrow{t_b})=\frac{\overrightarrow{t_a}\cdot\overrightarrow{t_b}}{|\overrightarrow{t_a}|\times|\overrightarrow{t_b}|} \tag{3}$$

5.4. Manhattan

$$\mathrm{Manhanttan}(d_i,d_j)=\sum_{1}^{n}\left|w_{k_i}-w_{k_j}\right| \tag{4}$$

5.5. Pearson correlation

$$\mathrm{SIM_c}\ (\overrightarrow{t_a},\overrightarrow{t_b})=\frac{m\sum_{t=1}^{m}\omega_{t,a}\times\omega_{t,b}-TF_a\times TF_b}{\sqrt{[m\sum_{t=1}^{m}\omega_{t,a}^2-TF_a^2][m\sum_{t=1}^{m}\omega_{t,b}^2-TF_b^2]}} \tag{5}$$

5.6. Jaccard Coefficient

$$\mathrm{SIM_J}\ (\overrightarrow{t_a},\overrightarrow{t_b})=\frac{\overrightarrow{t_a}\cdot\overrightarrow{t_b}}{|\overrightarrow{t_a}|^2+|\overrightarrow{t_b}|^2-\overrightarrow{t_a}\cdot\overrightarrow{t_b}} \tag{6}$$

5.7. Extended Jaccard Measure

$$s(x_i,x_j)=\frac{x_i^T\cdot x_j}{\|x_i\|^2+\|x_j\|^2-x_i^T\cdot x_j} \tag{7}$$

5.8. Normalized Information Distance

$$d(x,y)\le f(x,y)+O(1/K) \tag{8}$$

5.9. N-Gram Similarity and Distance

$$s_n(\Gamma_{k,l}) = max\left(s_n(\Gamma_{k-1,l}), s_n(\Gamma_{k,l-1}), s_n(\Gamma_{k-1,l-1}), +s_n(\Gamma^{n}{}_{k-n,l-n})\right) \quad (9)$$

5.10. Silhouette Coefficient

$$SC = \frac{\sum_{\rho \epsilon DS} S(\rho, D_M)}{|DS|} \quad (10)$$

5.11. Averaged Kullback-Leibler Divergence

$$D_{KL}(P \parallel Q) = Plog\left(\frac{P}{Q}\right) \quad (11)$$

5.12. Normalized Google Distance

$$s(t_x, t_y) = 1 - NGD(t_x, t_y)\alpha \quad (12)$$

5.13. Dice Coefficient Measure

$$s(x_i, x_j) = \frac{2x_i^T \cdot x_j}{\|x_i\|^2 + \|x_j\|^2} \quad (13)$$

6. Common Clustering Technique

The list presents summarization of clustering techniques used in the papers mostly [31] [32] [34-41].

1. Hierarchical Methods
   - Agglomerative Algorithms
   - Divisive Algorithms
2. Partitioning Methods
   - Relocation Algorithms
   - Probabilistic Clustering
   - K-medoids Methods
   - K-means Methods
   - Bisecting K-means
   - Fuzzy c-means
3. Density-Based Algorithms
   - Density-Based Connectivity Clustering
   - Density Functions Clustering
4. Grid-Based Methods
   - Basic Grid-based Algorithm
5. Model-Based Clustering
6. Methods Based on Co-Occurrence of Categorical Data
   - Robust Clustering using links (ROCK)
   - Sieving Through Iterated Relational Reinforcement(STIRR)
   - Clustering Categorical Data Using Summaries algorithm (CACTUS)
7. Constraint-Based Clustering

8. Clustering Algorithms Used in Machine Learning
9. Gradient Descent and Artificial Neural Networks
10. Evolutionary Methods
11. Scalable Clustering Algorithms
12. High Dimensional Data
    - Subspace Clustering
    - Projection Techniques
    - High Dimensional Data Clustering
    - Explicit Clustering
    - Implicit Clustering

After calculation the similarity, we need to know about clustering methods to construct similar groups. Thus we present well known equations: K-mean, bisecting k-means (Algorithm) and Fuzzy c-mean.

6.1. K-means

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q \tag{14}$$

6.2. Bisecting k-means

1. Choose the largest cluster to split.
2. Use k-means to split this cluster into two subclusters.
   (Bisecting step)
3. Repeat step 2 for some iterations (in our case 10 times) and choose the split with the highest clustering overall similarity.
4. Step 4: Go to step1 again until the desired k clusters are obtained.

6.3. Fuzzy c-means

$$E(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^{m} \parallel \bar{x}_j - \bar{v}_i \parallel^2 \tag{15}$$

$$E(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^{m}\left(1-K(\bar{x}_j,\bar{v}_i)\right) + \frac{\alpha}{|N_j|}\sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^{m}\sum_{\bar{x}_r\in N_j}(1-u_{ir})^{m} \tag{16}$$

7. Overall System Implementation for Web Pages Clustering

Reviews of most of papers related on web page clustering are known over all block diagrams in which using two main functions: distance and similarity methods and clustering methods in fig. 2. After acquisitions web pages on the Internet, we need to determine methods which are suitable for a standard dataset generally.

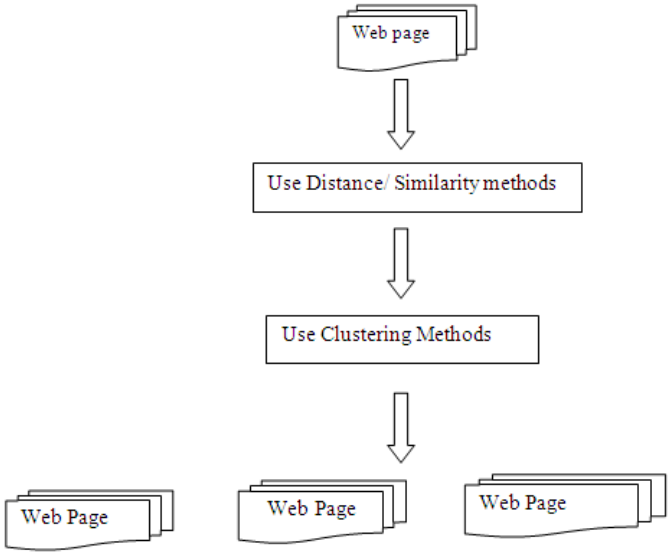

Figure 2. Representation of General Approach for Clustering

8. Overall Result

This section presents the results and clarifies our survey of usages of methods for papers studied in this paper. Here, we show usage times in survey papers. Most of papers use common methods such as Euclidean distance, Cosine Distance, Cosine Similarity, Manhattan, Pearson correlation, Jaccard Coefficient, Extended Jaccard Measure and N-Gram Similarity and Distance. Some methods not shown in fig. 3 are rare to use in clustering web pages. In fig. 4, it shows different usages of clustering methods in our survey. The figure 5 shows the number of sample usages of Partitioning Clustering Methods of survey papers. Comparisons in figures can be concluded what methods are widely useful today.

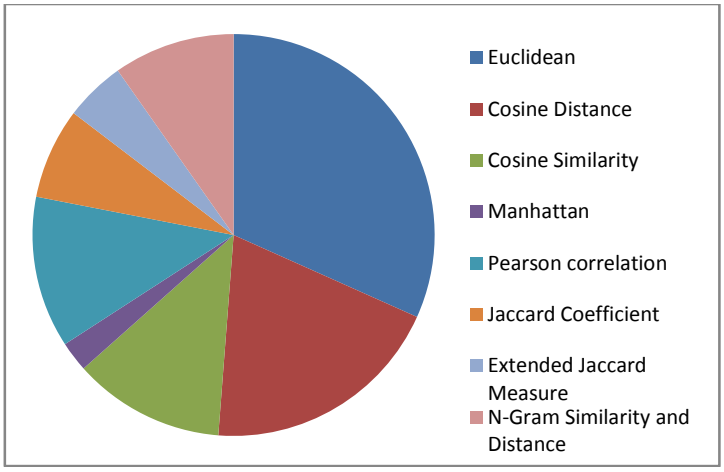

Figure 3. Usage of Distance/ Similarity Methods

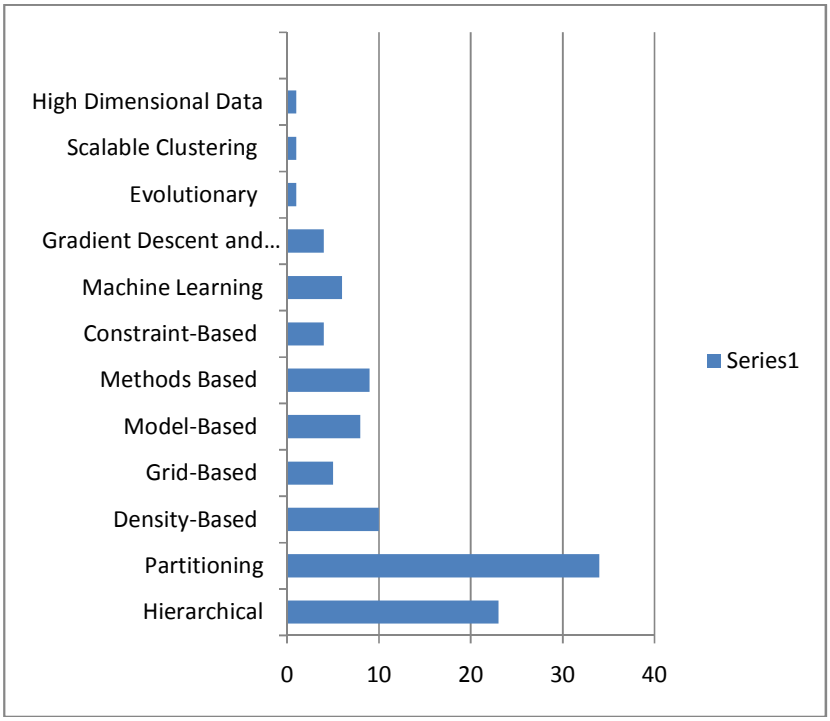

Figure 4. Usage of Clustering Methods

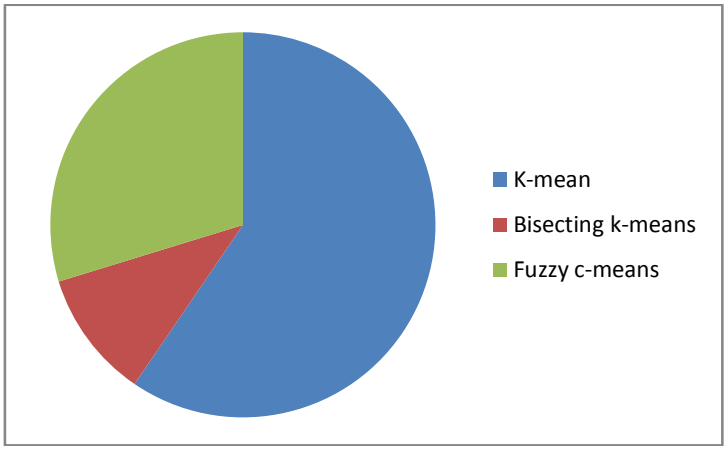

Figure 5. Usage of Partitioning Clustering Methods

9. Conclusion

This paper concludes summarization of similarity equations and clustering methods of usage which aims at describing equations with reasonable amount of papers. We have developed a result for usage of document similarity measure and cluster from human judgment. Average human predicts similarity more consistently with on a standard dataset. Our results provide

somewhat support for people general knowledge of web page clustering. This paper demonstrates effective methods that must be analyzed and organized according to the user's approaches.

Acknowledgements

Our heartfelt thanks go to all people, who support us at the University of Computer Studies, Mandalay, Myanmar. This paper is dedicated to our parents. Our special thanks go to all respectable persons who support for valuable suggestion in this paper. This paper is dedicated to our parents and my younger sister. Our acknowledgments especially go to our families for their financial and moral support to the publication of this paper.

References

[1] D.Brezeale, "The Organization of the Internet Web Pages using WordNet and Self Organization Maps", Master thesis, University of Taxes at Arlington, August 1999.

[2] X. He, H. Zha , C.H.Q. Ding, and H. D. Simon, "Web document clustering using hyperlink structures", Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, US, NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA.

[3] X. Huang, and W. Lai, "Web Graph Clustering For Displays And Navigation Of Cyberspace", The University of Southern Queensland, Australia, Swinburne University of Technology, Australia.

[4] B. Liu, "Web Data Mining", Department of Computer Science, University of Illinois at Chicago, USA.

[5] Z. Markov and D. T. Larose, "Data Mining The Web", Central Connecticut State University New Britain, CT.

[6] C. Kaur , and R. R. Aggarwal, "Web Mining Tasks And Types: A Survey", Department of Computer Science & Engineering, Thapar University, Patiala, Volume 2, Issue 2, February 2012 ,.

[7] D. Gibson, R. Kumar, and A. Tomkins, "Discovering Large Dense Subgraph in Massive Graphs", VLDB Conference, 2005.

[8] C. C. Aggarwal, "Managing and Mining Graph Data", IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA,

[9] D. F Kibler and S. Hampson, "Learning Weight Matrices for Identifying Regulatory Elements", Information Computer Science Department, University of California, Irvine, USA.

[10] N.P. Vander, H.G. T. Morsche, and R.R.M. Mattheij, "COMPUTATION OF EIGENVALUE AND EIGENVECTOR DERIVATIVES FOR A GENERAL COMPLEX-VALUED EIGENSYSTEM", Electronic Journal of Linear Algebra, A publication of the International Linear Algebra Society Volume 16, pp. 300-314, October 2007.

[11] S. KOUACHI, '"EIGENVALUES AND EIGENVECTORS OF TRIDIAGONAL MATRICES", Electronic Journal of Linear Algebra ISSN 1081-3810, A publication of the International Linear Algebra Society, Volume 15, pp. 115-133, April 2006.

[12] A. Farkas, "The Analysis of the Principal Eigenvector of Pairwise Comparison Matrices", Budapest Tech, Faculty of Economics, 1084 Budapest, TavaszmezÅ út 17, Hungary.

[13] X. Chen, L. Guo, and Z. Fan, "LEARNING POSITION WEIGHT MATRICES FROM SEQUENCE AND EXPRESSION DATA", 10:6 WSPC/Trim Size: 11in x 8.5in for Proceedings, May 14, 2007.

[14] A. Amine, Z. Elberrichi, and M. Simonet, "Evaluation of Text Clustering Methods Using WordNet", The International Arab Journal of Information Technology, Vol. 7, No. 4, October 2010.

[15] M. Rafi, and M. S. Shaikh, "An improved semantic similarity measure for document clustering based on topic maps", Computer Science Department, NU-FAST, Karachi Campus, Pakistan, muhammad.rafi@nu.edu.pk.

[16] A. Meenakshi and V. Mohan, "KNOWLEDGE MANAGEMENT IN EDAPHOLOGY USING SELF ORGANIZING MAP (SOM)", International Journal of Database Management Systems (IJDMS) Vol.4, No.5, October 2012.

[17] E. Ş. Chifu and I. A. Leţia, "Self-organizing Maps in Web Mining and Semantic Web", Technical University of Cluj-Napoca, Romania.

[18] L. Khan and F. Luo, "Ontology Construction for Information Selection", Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688.
[19] W. Wong, W. Liu and M. Bennamoun, "Featureless similarities for terms clustering using tree-traversing ants", School of Computer Science and Software Engineering, University of Western, Australia, Crawley WA 6009.
[20] T. F. Gharib, M. M. Fouad, A. Mashat, and I. Bidawi, "Self Organizing Map -based Document Clustering Using WordNet Ontologies", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814, www.IJCSI.org.
[21] L. Khan and F. Luo, "Ontology Construction for Information Selection", Department of Computer Science, University of Texas at Dallas, Richardson.
[22] B. Russell, H. Yin, and N. M. Allinson, "Document Clustering Using the 1 + 1 Dimensional Self-Organising Map", University of Manchester Institute of Science and Technology (UMIST, Department of Electrical Engineering and Electronics, Manchester M60 1QD, United Kingdom.
[23] A. Tahamtan, A. Anjomshoaa, E. Weippl, and A. M. Tjoa, " SOM-Based Technique for a User-Centric Content Extraction and Classification of Web 2.0, with a Special Consideration of Security Aspects", Vienna University of Technology, Dept. of Software Technology & Interactive Systems, and Information & Software Engineering Group, Secure Business Austria, Favoritenstrae 16 - 2nd floor, 1040 Vienna, Austria.
[24] K. M. Fouad and M. O. Hassan, " Agent for Documents Clustering using Semantic-based Model and Fuzzy", Computer Science Dept., College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia (KSA) and Computer Engineering Dep., Faculty of Engineering Cairo University, Egypt, International Journal of Computer Applications (0975 – 8887) Volume 62– No.3, January 2013.
[25] L. Rokach, and O. Maimon, "CLUSTERING METHODS", Department of Industrial Engineering, Tel-Aviv University.
[26] P. Berkhin, and S. Jose, "Survey of Clustering Data Mining Techniques", Accrue Software, 1045 and Forest Knoll, CA, 95129.
[27] J. Hou, and Y. Zhang, "Utilizing Hyperlink Transitivity to ImproveWeb Page Clustering", Department of Mathematics and Computing, University of Southern Queensland, Toowoomba, Qld 4350, Australia.
[28] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi," The Similarity Metric", IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 50, NO. 12, DECEMBER 2004.
[29] G. Kondrak, "N-Gram Similarity and Distance", Department of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8, Canada.
[30] H. Liu and M. Schneider, " Similarity Measurement of Moving Object Trajectories", Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32611, USA.
[31] D. E. Herwindiati and S. M Isa, "The Robust Distance for Similarity Measure of Content Based Image Retrieval", Proceedings of the World Congress on Engineering 2009, Vol II, WCE 2009, July 1 - 3, 2009, London, U.K.
[32] R. C. Veltkamp and L. J. Latecki, "Properties and Performance of Shape Similarity Measures", Dept. Computing Science, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands, Dept. of Computer and information Sciences, Temple University, Philadelphia, PA 19094, USA.
[33] A. Huang, "Similarity Measures for Text Document Clustering", Department of Computer Science, The University of Waikato, Hamilton, New Zealand.
[34] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data", Department of Computer Science, Trinity College, Dublin 2, Ireland.
[35] L. H. Ungar and D. P. Foster, " Clustering Methodsfo r Collaborative Filtering", CIS Dept. and Dept. of Statistics, University of Pennsylvania Philadelphia, PA 19104.
[36] M. Steinbach, G. Karypis and V. Kumar, "A Comparison of Document Clustering Techniques", Department of Computer Science and Engineering, University of Minnesota.
[37] T. Soni Madhulatha, "AN OVERVIEW ON CLUSTERING METHODS", Associate Professor, Alluri Institute of Management Sciences, Warangal. IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.
[38] M. Youssef, and A. Agrawala, "LOCATION-CLUSTERING TECHNIQUES FOR WLAN LOCATION DETERMINATION SYSTEMS", Department of Computer Science University of Maryland at College Park.
[39] P. Andritsos, "Data Clustering Techniques", Qualifying Oral Examination Paper, Department of Computer Science, University of Toronto, March 11, 2002.

[40] F. M... Alvarez, A. Troncoso, J.C. Riquelme, and J.M. Riquelme, " Partitioning-Clustering Techniques Applied to the Electricity Price Time Series", Area of Computer Science, Pablo de Olavide University, Spain, Department of Computer Science, University of Seville, Spain, and Department of Electrical Engineering. University of Seville, Spain, IDEAL 2007, LNCS 4881, pp. 990–999, 2007.

[41] P. Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc, 1045 Forest Knoll, CA.

Author

She is Associate Professor, Head of Department of Research and Development II in University of Computer Studies, Mandalay, Myanmar. Her research areas include Information Retrieval, Cryptography and Network Security, Web Mining and Networking with T CP/IP. She received her B.Sc. (Physics), M.Sc. (Physics) from Yangon University, Myanmar and M.A.Sc. (Computer Engineering) and Ph.D. (Information Technology) from University of Computer Studies, Yangon, Myanmar.