Certain Investigations for Retrieving Web Documents using Soft Computing Techniques

B.Sundaramurthy[1]
[1] Research Scholars,
Department of Computer Science and Engineering,
Manonmaniam Sundaranar University, India,
Tamilnadu, Tirunelveli - 627012

G.Tholkappia Arasu [2], Ph.D
[2] Principal,
AVS Engineering College,
India, Tamilnadu,
Ammapet,Salem – 636003

ABSTRACT
Web Documents are of vital importance as they provide information from basic to the core requirements of users. The web documents have similarity based on the content and information represented. The key information in the web documents can be analyzed and can be represented as features. This leads to the importance of using similarity measures for categorizing and representing the web documents back to the users. This is a vital challenge as the documents could be clustered and represented in multiple dimensions for retrieval. This paper performs a complete survey of different techniques that are available for retrieving the web documents. The paper also presents investigations of the different performance parameters available for measuring the performance of results. The paper also presents research directions for using soft computing techniques for web document retrieval.

Keywords
clustering, feature vector, neural network, index, dimension, Filters, probability.

1. INTRODUCTION
Today's web has documents represented in different forms. The contents include images, text and also contains hyperlinks to several levels. This is of major importance as the data provides better views to required users. The information presented needs to be analyzed more in detail , as the user's interest varies. There is a growing demand in terms of providing information to users. However in parallel user's view of looking for appropriate information is also growing in several dimensions. This needs to be analyzed and presented in detail . The model presented below analyzes the basic model for routing requests in the web[1].

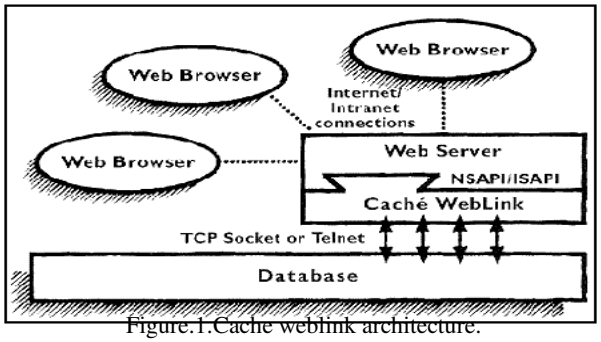

Figure.1.Cache weblink architecture.

In Internet , the web pages are maintained across several servers. The copies of data is maintained continuously and replica management is also handled properly. URL rewriting could be used for redirecting the requests of the users for providing the appropriate pages. It could be a entire replication or object replication [2]. The data being clustered needs to be managed and stored efficiently. The topography of the databases is also important for representing and managing the web pages. Web on the standard level is growing from Web 1.0 to Web 2.0 and so on.

2. RELATED WORK
The clustering approach can use the Web log files, co-clustering Web objects and investigating social networks from Web data. This could be of vital importance as it provides the dynamic information recorded from time to time[2]. The information could be scaled and mined for efficiency. A hierarchical environment could also be developed taking into account the site content and structure, the HTML document structure and the term importance. A partitional clustering algorithm is a best solution for such scenario[4]. The user accessing the web page can stay on the web page and can switch between web pages. This provides inputs on the user's interest for a web page. An evolutionary two-layer clustering algorithm could also be used for web clustering. The learning vector quantization (LVQ) along with the weighted fuzzy c-means approach is handled with the results of the first layer[6]

The data is inherently fuzzy and provides more learning information during web page retrieval. Cluster scoring function could be used for deciding the clustering mechanism[7].Duplicated page deletion along with matching degree across web pages is also vital in deciding and providing the appropriate page[8]. This can be important as it removes the redundancy of web pages in terms of the content across the web pages. A topic model can be used to associate annotated document with a distribution of topics, constructs a graph including tags, document and topics by performing a Random Walks for clustering[9].Phrase based analysis with incremental clustering can improve the results efficiently in comparison to word based analysis[10].

Clustering of web pages could be done based on similarity between hyperlinks[13]. The inherent data access mechanism in web users suggest that the rough set approach can be used for providing relevant more web pages[15]. It is proved that

the similarity measure gives better results during rough set based approaches. Ontology can also provide better results if combined with fuzzy based approaches[16]. Time locality of the navigational acts could also be of vital importance for clustering as it provides details regarding the time spent by the users for a web page[19]. Each of the cluster instances return results that needs to be managed and presented for user's view. In this case the features are also of multi type in nature. MultiView Information Bottleneck could be a good solution in such cases[23].Temporal cluster migration matrices can be applied for presenting web pages in demand that change with the temporal aspects of the user access[24]. Users browsing patterns could be analyzed for interpretration. Learning could be done for a set of webpages using Kohonen's self organizing maps (SOM) and interpreting those results using Unified distance Matrix (U-matrix)[26].This could be extended to research articles accessed by users in a domain.

In real time requirements the datasets include movie ,stock market, climate datasets that are recorded over a period of time. The datasets are all dense in nature and are time varying. coverage and Precision could be used for evaluating the datasets. Similarity measures combined with noise estimation can be used for clustering the webpages[29].

Survey table -1

Author/ Year	Approach/ Methodology	Merits	Limitations	Data Set	Parameters	Suggestions
Lu Caimei et al /2011	Tripartite Clustering k-means and Link-K-Means is also applied for social tagging	(i). Social tagging is proved be to be providing vital information (ii). Cluster Structure of tag nodes is analyzed.	Probability based association of topics to clusters is not possible	(i). Social Bookmarking data (ii). Social tagging Data	(i). Centroid for cluster (ii). FScore (iii). Purity (iv). NMI	Multiclass Posterior Probability Support Vector Machines could be used.
Di Giacomo E. et al /2007	Topology Based Approach Graph for indicating Semantic relationships	Graph of categories is used for indexing	Speed of Programming environment needs to be improved.	First 200 snippets returned by Google	(i). Relevance Score (ii). Hitratio (iii). Average Number of correct pages found (iv). T-test on hit ratio	Language based analysis for web page indexing could be applied.
Yanagimoto Hidekazu/2010	Eigen Based Approach	Noise is also handled at web page level	Eigen value Decay is not considered	Real social bookmarking data	Eigen Similarity measures	Fuzzy based approach could be provided with boosting approaches
Yang Christopher C. et al /2011	Scalable Distance based clustering	Less number of clusters are grouped even when they are densely reachable.	Configuring the density of clusters is not reachable	Dataset of political Subjects	(i).TFIDF (ii).Cosine Similarity (iii).Micro accuracy (iv).Macro accuracy (v).Eps-neighbourhood distance	Social Networking forum topics could be tagged for providing better results.
Loia Vincenzo et al/2007	Collaborative proximity-based fuzzy clustering	Semantic MetaData based search is applied for web clustering Structure Discovery is handled efficiently.	Feature Reduction is not complete.	Dataset built from Swoogle search engine	(i).Proximity, Performance Index,precision, (ii).Recall,Lack of Classification, (iii).Classification Error	Boosting based approaches could be used.

Author/ Year	Approach/ Methodology	Merits	Limitations	Data Set	Parameters	Suggestions
Schroeder Trevor/2000	L7 Server is being suggested for Clustering	A detailed Tabulation of L4/2,L4/3,L7 is being presented.	Lacks detail on how the queuing request is being managed.	-	Chip density, Speed	More input could be presented on the datasize and user management types.
Huang Sheng et al/2006	Multitype Features Coselection for Clustering	(i).MFCC is compared with Colearning (ii).Feature coselection is evaluated using two WebKB benchmarks and one ODP data set	Fuzzy based nature of datasets could be analyzed.	CT dataset	(i). voting, average value, (ii). max value, average rank, and max rank, error rate, F1-Measure, andentropy, feature selection percentage, the number of clusters, selection of centroids.	(i). More Datasets could be built and tested (ii). Different clustering algorithms could be applied and tested.
Papadakis Nikolaos K., et al/2005	Statistical based Web retrieval	Statistics origin is used for web page retrieval A detailed tabulation on wrapping of datasources is presented. Performance has been compared with OMRI and MDR	Web pages that Contain frames is not possible to Be handled under this Model.	63000 HTML pages from 50 different Datasources.	Variance,cluster compactness, cluster separation, correlation Coffecient, Covariance, Precision, recall	Could be tested and extended for a cloud Environment
Krishnapuram Raghuram et al /2001	Fuzzy-C-medoids algorithm Robust fuzzy – medoids algorithm	(i) Feature Vector is used for representation of web page information. (ii) Stop Word and Elimination algorithm.	Web content does not analyze the summarizing phrases.	One data set of 1042 abstracts Second Dataset of 59 HTML pages	IDF, Disimilarity, Cosine Measure	Could be extended to stories on the internet and summaries could be presented.
Lu Caimei/2008	Link Based Clustering Algorithm	Information from Sibling pages is used in clustering	Temporal information is not considered	WebKB4 Dataset citation dataset Cora7	Precision	Tree/Graph based approach could be used for understanding the depth of the web traversal.
Rangarajan Santosh K./2004	ART1 algorithm	Web access patterns is used with ART network	Log file does not have information regarding the security .	Data Managed from NASA Web server	Average Intercluster distance, Average IntraCluster distance, Accuracy	Could be tested with Pattern Recognition based approaches.

3.OUTCOMES OF THE SURVEY

The survey on different approaches of web clustering has led to following findings:

- There is enough scope for analysis of web documents considering the semantic information of the documents.
- The approaches existing could be extended with more better learning approaches with soft computing based models.
- Real time data could be built and tested as there is large scope in fuzzy nature of the dataset.
- Mining approaches could also be appropriately suggested and used for investigating the nature of the datasets and for improving the results.
- Linguistic issues in web pages could also be analyzed for evaluating the results.
- Webserver design issues are also important as it can affect the performance of the clustering results.
- Cluster formation could be done better improving the probability of relativity of web page content.
- Queing approaches could be modelled to understand the nature of traffic in the web .
- The type of web content is of vital importance as it leads to data specific issues on the web.

4. CONCLUSION

This paper has performed a complete survey of the various techniques that are available for clustering the web documents. A detailed tabulation is being presented with the findings . This could be extended by selecting a appropriate problem that exist as part of the finding and can be analyzed for the feasibility issues. More learning to web searches can be completed with tools like orange. A complete framework could also be proposed for web page retrieval that is robust, scalable and learns with real time user access patterns.

5. REFERENCES

[1] http://docs.intersystems.com/cache41/wlk/wl_intro.html

[2] Norihito Fujita, Yuichi Ishikawa, Atsushi Iwata, Rauf Izmailov,"Coarse-grain replica management strategies for dynamic replication of web contents", The International Journal of Computer and Telecommunications Networking,Vol.45, pp.19-34, 2004,

[3] Zhang, Yanchun ; Xu, Guan-Dong," Using Web Clustering for Web Communities Mining and Analysis", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, ,pp.20-31, 2008

[4] Agavriloaei Ioan, Alexandrescu Adrian , Craus Miticä, "Improving web clustering through a new modeling for web documents" International Conference on System Theory, Control, and Computing (ICSTCC), PP.1 – 6, 2011

[5] Lu Caimei,Hu Xiaohua Tony,Park Jung-ran,"Exploiting the Social Tagging Network for Web Clustering" , IEEE Transactions on Systems, Man and Cybernetics, Volume 41 , PP.840 – 852,2011

[6] Wu Rui,"Clustering Web Access Patterns Based on Hybrid Approach"Fifth International Conference on Fuzzy Systems and Knowledge Discovery,Vol.1, PP.52 – 56,2008.

[7] Crabtree Daniel , Gao Xiaoying , Andreae Peter," Improving Web clustering by cluster selection"IEEE/WIC/ACM International Conference on Web Intelligence, PP. 172 – 178,2005

[8] Li Tao-Ying , Chen Yan,Web Page Clustering Based on Searching Keywords ,"International Conference on Intelligent Computation Technology and Automation (ICICTA)",Vol.3, PP.1163 – 1166,2010

[9] Sun Jiashen , Wang Xiao-jie , Yuan Caixia , Fang Guannan, Annotation-aware web clustering based on topic model and random walks,IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS),,PP.12-16,2011.

[10] Supreethi K. P. , Prasad E. V.," Web Document Clustering Technique Using Case Grammar Structure" International Conference on Conference on Computational Intelligence and Multimedia Applications, 2007. Vol.2,, PP.98 – 102,2007

[11] Di Giacomo E. , Didimo Walter , Grilli Luca , Liotta Giuseppe," Graph Visualization Techniques for Web Clustering Engines"IEEE Transactions on Visualization and Computer Graphics,Vol.13 , PP. 294 – 304,2007

[12] Yanagimoto Hidekazu , Yoshioka Michifumi , Omatu Shigeru,"Web Clustering Using Social Bookmarking Data with Dimension Reduction Regarding Similarity",International Conference on Advances in Social Networks Analysis and Mining, pp.386-390,2010

[13] Takahashi Kou , Miura Takao , Shioya Isamu, "Clustering Web Documents Based on Correlation of Hyperlinks" International Conference on Data Engineering , PP.1225,2005

[14] Yang Christopher C. , Ng Tobun Dorbin,"Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering"IEEE Transactions on Systems and Humans,Systems,Man and Cybernetics, Vol.41 ,PP.1144 - 1155,2011.

[15] Mishra R.,Kumar P.,Bhasker B.,"Rough set based clustering in dense web domain" International Conference on Intelligent Systems Design and Applications (ISDA), PP.521 - 526 ,2012 .

[16] Gholamzadeh Nayereh , Taghiyareh Fattaneh," Ontology-based fuzzy web services clustering" International Symposium on Telecommunications,PP.721 - 725 ,2010.

[17] Loia Vincenzo,Pedrycz Witold ,Senatore Sabrina,"Semantic Web Content Analysis: A Study in Proximity-Based Collaborative Clustering"IEEE Transactions on Fuzzy Systems,Vol.15,PP.1294 - 1312,2007 .

[18] Schroeder Trevor,Goddard Steve,Ramamurthy By rav,"Scalable Web server clustering technologies" IEEE Network,Vol.14 ,PP.38 - 45,2000.

[19] Petridou, Sophia G.,"Time-Aware Web Users' Clustering" ,IEEE Transactions on Knowledge and Data Engineering,Vol.20,PP. 653 – 667, 2008.

[20] Huang Sheng,Chen Zheng,Yu Yong,Ma Wei-Ying Y.,"Multitype features coselection for Web document clustering" IEEE Transactions on Knowledge and Data Engineering,Vol.18,PP.448 - 459,2006.

[21] Papadakis Nikolaos K.,Skoutas Dimitrios N. , Raftopoulos Konstantinos , Varvarigou Theodora A. "STAVIES: a system for information extraction from unknown Web data sources through automatic Web wrapper generation using clustering techniques" IEEE Transactions on Knowledge and Data Engineering, Vol.17,PP.1638 - 1652,2005.

[22] Chen Yan , Qiu Lilli , Chen Wei-Yu , Nguyen Luan , Katz Randy H.," Efficient and adaptive Web replication using content clustering" IEEE Journal on Selected Areas in Communications, Vol.21,PP.979 - 994 ,2003

[23] Gao Yan , Gu Shiwen , Xia Liming , Fei Yaoping, "Web Document Clustering with Multi-view Information Bottleneck" International Conference on Intelligent Agents, Web Technologies and Internet CommerceComputational Intelligence for Modelling, Control and Automation, PP.148,2006.

[24] Lingras Pawan,Hogo Mofreh A.,Snorek, Miroslav,"Temporal Cluster Migration Matrices for Web Usage Mining"IEEE/WIC/ACM International Conference onWeb Intelligence, PP.441 - 444,2004.

[25] Krishnapuram Raghuram,Joshi Aunupam,Nasraoui Olfa ,Yi Liyu, "Low-complexity fuzzy relational clustering algorithms for Web mining" ,IEEE Transactions on Fuzzy Systems,Volume: 9 , PP.595 - 607,2001.

[26] Menon Kartik , Dagli Cihan H.,"Web personalization using neuro-fuzzy clustering algorithms" International Conference of the North American Fuzzy Information Processing Society, 2003.PP.525 - 529,2003.

[27] Lu Caimei,PA Zhang Xiaodan, Park Jung-ran,Hu Xiaohua, He Tingting, "Web clustering based on the information of sibling pages" ,IEEE International Conference on Granular Computing, 2008.

[28] Sridharan K. , Chitra M.,"A fuzzy category based aggregation technique: A mutual approach for clustering and query processing and its application to web mining", Third International Conference on Computing Communication & Networking Technologies (ICCCNT), PP.1-7,2012

[29] Rongfei Jia , Maozhong Jin , Xiaobo Wang,"Web Objects Clustering Using Transaction Log" Third International Conference on Knowledge Discovery and Data Mining,PP.182 - 186,2010.

[30] Rangarajan Santosh K. , Phoha, Vir V. , Balagani, Kiran S. , Selmic, Rastko R. ; Iyengar, S Sitharama Sitharama,"Adaptive neural network clustering of Web users",IEEE computer society,Vol. 37 , PP. 34 - 40,2004