An Survey of Approaches for Mining Documents on Web based on User based Analysis

S.Senthilkumar[1]
[1] Research Scholar,
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 contain information that include image, video. The retrieved information are oriented towards the user's input for search. But the popular search engines like google uses algorithms for prioritizing and retrieving the results. However the information accessed also depends on the cookies as part of the local system. In real requirements based on the category , type and interest of users it is required to provide contents. This paper has a complete survey of different approaches/algorithms that are in existence for analyzing the web documents and providing to the user. The paper has a complete study of the different performance parameters that can be used for analyzing the results. The paper also draws conclusions for selecting the appropriate approach based on different scenarios and different user input criteria provided by users.

Keywords

Data Model, cluster, threshold, filtering, classification, Agent, learning

1. INTRODUCTION

Web users access data and update on a routine basis. The users stay away period on the web decides the importance and relevance of the web page for the users. The web browsers inherently maintain the recently accessed web page information in form of histories. However the history also becomes redundant over a period. The gender, age, topolographic information of the users also decides the importance of the web page for the users.The diagram below shows the model of a web architecture[26].

From Figure 1.1 , it can be understood that the user's request are filtered and the relevant information are presented to the users. Here there is a large amount of processing that happens during the retrieval of results. The web page engines like google.com has type word matching facility. With this ,as the user starts typing the word the first set of relevant words are displayed for the users to search against the web. The knowledge base that needs to be maintained for displaying the appropriate information is complex and dimensionally increases with the growing nature of the web.

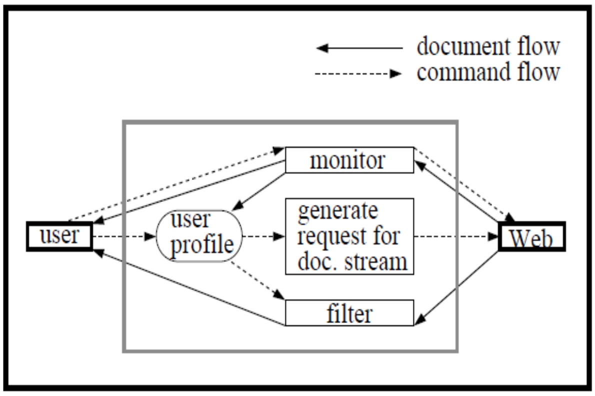

Figure 1.1 Architecture of the Web Model

2. RELATED WORK

Different web sites contain different information. The information could be organized contextually and presented. The nature of the web pages present large issues . The pages in e-commerce could be secured using false hit database algorithm and nearest neighbor algorithm[1]. Web usage mining can be a good solution to remove the drawbacks in data mining[2].

In Web mining, the temporal information is also very important. It decides how the cluster level changes. The type of data model could be snapshot graphs, tensors, segments. The clusters being designed are however incremental and grows with user types and search criteria[6].

Content based learning combined with user based learning can be useful for web page retrieval. It has more information relevant to the user behavior[7]. The results of the information retrieval system could be improved using a Concept lattice. This resolves the closure issues in building the set[9]. corpora could be built for the web documents for efficient retrieval. The corpora built could be trained for further classification. Boosting approaches can be a good solution [10].Rule based method and statistical methods can be used for classification of documents. User navigation also decides the mode of models can be used for predicting the type of users on the web. This could provide more inputs on the nature of the data the users can access[13]. Document level sentiment labels can be used in the context vectors used as the basis for measuring the distributional similarity between words. Test vectors can then be trained in a binary classifier for producing better results[14].

Ontology can also be a good solution for handling the historical documents. The historical documents can be mapped into a semantic knowledge base[15]. Full Bayesian Estimation can be used for understanding User information access behavior[16]. It could be also done by analyzing and simulating four kinds of kernel function and three ways of feature selection, polynomial kernel function and document frequency is chosen for the best way in SVM algorithm[17]. However this approach could be improved with feature reduction as the dimensionality of the information could be reduced. The web information need to stored, managed locally in a distributed environment. The users on web need access to information in a seamless manner. In such cases , there is needs for request to be routed across several nodes. The processes running in that nodes needs to interact with the requesting processes thereby sharing knowledge. It would be vital if the processes become self supporting providing clustering[12]. Maximum entropy and markov mixture decisions back to the users. In such cases agents can be a good solution.Agents could be user interface agents, coordinator agent, identify agent, CleanMiss agent, CleanNoisy agent, transformation agent and discretization agent [18][22]. More level of agents with higher functionality could also be developed. The agents could be hierarchically placed and designed for handling requested in any level to any order. Agents developed for web search engines should be scalable in nature[28]. The coordination among agents can be handled using holonic agents[29].

For a user search criteria ,the set of index pages can be created using content based, collaborative filtering, user based filtering. The search criteria could be improved with ontology based approaches. Machine learning with kernel classifiers can also be a good solution[19]. A framework could also be developed as a end solution for web users in the internet[21]. Rule Engine for the web could be developed using Enatailment Based OWL reasoning. Ontology mapping, inference, query support are the three major issues in EBOR paradigm[23]. Belief Desire Intention Models are good solutions for designing better agents. Each agent has one interpreter that will process the events from the agent's environment, and, in conjunction with the agent's plans and other mental states, determine the agent's behavior[24]. An automous web agent search model could be developed fully functional similar to CiteSeer[30].

Table-1, Survey table -1

Author/ Year	Approach/ Methodology	Merits	Limitations	Data Set	Parameters	Suggestions
Buche P./2013	ONtology-based Data INtEgration(ONDINE)	Core Ontology and Domain Ontology is being used. Fuzzy sets are used for analysis	Has not been experimented with Datatables generated from web.	XML Datatables	Score,Cosine Similarity ,	The model could be tested with documents that are timely changing like share market.
Guan Ziyu/2013	Hypergraph	co-occurrence structure is modeled using a hypergraph	The model does not accomdate for Optimizing the performance using multithreading, multicore, MapReduce or sampling techniques.	ClueWeb09 web collection	Running Time	-
Skabar Andrew/2013	Fuzzy Clustering Algorithm	Identifies overlapping clusters of semantically related sentences	Can be extended to Attribute Data	News Article Dataset	PageRank, Partition Entropy Coefficient, Purity, Entropy	-

Cui Hang/2003	Log Based Query Expansion Method	Query Expansion based on user interaction. User Query Sessions. Short and Long Queries. Gap between Query and Document space is reduced.	-	TREC data Encarta Website	Similarity, Interpolated 11-point average precision,	Can be tested for internal documents accessed in a university
Chen Ming-Syan/1998	Maximal Forward Algorithm	Backward References are removed. Number of database scans is reduced Available as part of SpeedTracer Tool.	-	Synthetic Data	Execution Time	History Information could also be used.
Chunyu Kit et al /2007	Best First Search Algorithm	no prior knowledge such as adhoc heuristics, no labelled data for training ,no similarity analysis of Web page structure	Rescuing weak keys Depressing false keys Extracting bitexts	Government Web index – http://www.info.gov.hk/chorgdex.htm.	Precision,Recall ,F-Score	Reliability could be improved.
Jean Paul/2012	DIVAlite agents	More visible agents for the screen Feature Based Model	Can be extended to SignLanguage Teaching	Characters	Complexity Number of lines of code	Can be integrated to the systems for guiding challenged people in railway stations.
Gabriel L. Somlo et al/2001	Incremental Clustering	Agent Feedback is provided using doubling incremental clustering. Use of Web information agents and Text filtering.	-	Web Data	Precision Recall LGF	Can be extended with web portals that has contents with different category of metadata.
Wolfgang Ketter et al/2008	Semantic Web Architecture	Advocate Agent Business Model Security support is provided for advocate agent. Support for collective Intelligence.	Optimization of user queries is not considered.	-	-	Optimization of results is not done.

3. RESEARCH FINDINGS

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

- Agents with ontology based approaches are appropriate solutions for web search engines.
- There is enough scope for applying different neural networks for search engines.
- The fuzziness of the data provides large scope for using Fuzzy based approaches.
- Semantic Search engines are of vital importance as it moves closer with the user's interest.
- There is possibility for building better feature vectors with more information about the user's ,user access patterns.
- A complete unsupervised system for providing results based on user interest could be developed.
- The agents could be built with autonomous nature considering the scope for developing new type of agents.
- The web search engines could be developed for a closed environment to understand the design challenges of the environment.
- There is enough scope for accommodating security aspects for the web retrieval.
- Feature Reduction Approaches could be used for modelling and presenting the search results in a efficient manner.

4. CONCLUSION

This paper has presented a detailed study on using data mining approaches with search engines from the view of the user interest. The study has been presented from the ends of mining, agents, learning approaches. The research findings has been presented that could be incorporated into the research.

5. REFERENCES

[1] Manjusha, R.,"Web mining framework for security in e-commerce" International Conference on Recent Trends in Information Technology (ICRTIT),PP.1043-1048,2011

[2] Sharma Kavita,Shrivastava,Gulshan Kumar, Vikas, "Web mining: Today and tomorrow" ,International Conference on Electronics Computer Technology (ICECT), Vol.1,PP.399 - 403,2011.

[3] Buche P. , Dibie-Barthelemy J. , Ibanescu L. , Soler L.," Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource",IEEE Transactions on Knowledge and Data Engineering, Volume: 25 ,PP.805 - 819,2013

[4] Guan Ziyu,Miao Gengxin,McLoughlin Russell,Yan Xifeng , Cai Deng,"Co-Occurrence-Based Diffusion for Expert Search on the Web"IEEE Transactions on Knowledge and Data Engineering, Vol.25,PP.1001 - 1014,2013.

[5] Skabar Andrew,Abdalgader,Khaled,"Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm" ,IEEE Transactions on Knowledge and Data Engineering,Vol.25 , PP.62-75,2013

[6] Giatsoglou M. , Vakali A.,"Capturing Social Data Evolution Using Graph Clustering" IEEE Internet Computing, Vol.17,PP.74 - 79,2013

[7] Lim Edward H Y , Tam Hillman W K , Wong Sandy W K , Liu James Nga-Kwok , Lee Raymond S T,"Collaborative content and user-based web ontology learning system", IEEE International Conference on Fuzzy Systems,PP.1050 - 1055,2009.

[8] Cui Hang,Wen Ji-Rong,Nie Jian-Yun Y.,Ma Wei-Ying Y.,"Query expansion by mining user logs",IEEE Transactions on Knowledge and Data Engineering, Vol.15,PP.829 - 839,2003

[9] Myat Nyeint Nyeint,Hla Khin Haymar Saw,"Organizing Web Documents Resulting from an Information Retrieval System Using Formal Concept Analysis",Proceedings of the sixth Asia-Pacific Symposium on Information and Telecommunication Technologies PP.198 - 203,2005.

[10] Huang Chien-Chung,Lin Kuan-Ming,Chien Lee-Feng,"Automatic training corpora acquisition through Web mining",IEEE/WIC/ACM International Conference on Web Intelligence, PP.193 - 199 ,2005.

[11] Chen Ming-Syan Syan,Park Jong Soo,Yu Phillip S.,"Efficient data mining for path traversal patterns",IEEE Transactions on Knowledge and Data Engineering, Vol.10,PP.209 - 221,1998.

[12] Chengzhi Zhang,Qingguo Zhang, "Topic Navigation Generation Using Topic Extraction and Clustering", International Symposium on Knowledge Acquisition and Modeling, PP.333 - 339 ,2008

[13] Manavoglu Eren,Pavlov Dmitry,Giles C. Lee, "Probabilistic user behavior models", Third IEEE International Conference on Data Mining,PP.203-210,2003

[14] Bollegala D. , Weir D. , Carroll J.,Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus,IEEE Transactions on Knowledge and Data Engineering,2012

[15] Witte R , Krestel R , Kappler T , Lockemann P,"Converting a Historical Encyclopedia of Architecture into a Semantic Knowledge Base", IEEE Intelligent Systems,2009.

[16] Chen Jian,Shtykh Roman Y.,Jin Qun,"Gradual Adaption Model for Estimation of User Information Access Behavior",3rd International Conference on Systems and Networks Communications,PP.378 - 383,2008

[17] Gang Xiao ,Jiancang Xie,"Performance Analysis of Chinese Webpage Categorizing Algorithm Based on Support Vector Machines (SVM)",Fifth International Conference on Information Assurance and Security, Vol.1,PP. 231 - 235,2009.

[18] Othman Zulaiha Hj Ali,Bakar Azuraliza Abu,Hamdan Abdul Razak,Omar Khairuddin Bin,Shuib Nor Liyana Mohd,"Agent based preprocessing",International Conference on Intelligent and Advanced Systems,PP.219-223,2007

[19] Magdalini Eirinaki,Michalis Vazirgiannis,"Web mining for web personalization",ACM Transactions on Internet Technology,Vol.3,PP.1-27,2003.

[20] Chunyu Kit, Jessica Yee Ha Ng,"An Intelligent Web Agent to Mine Bilingual Parallel Pages via Automatic Discovery of URL Pairing Patterns",Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops,PP.526-529,2007.

[21] Danielle Medeiros, Uirá Kulesza, André Mauricio Campos,"A framework for implementing web recommendation agents",Proceedings of the 18th Brazilian symposium on Multimedia and the web,2012.

[22] Leona F. Fass,"Some agent theory for the semantic web",SIGSOFT Software Engineering Notes,Vol.30,2005.

[23] Nick Bassiliades,"Agents and knowledge interoperability in the semantic web era", Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics,2012

[24] Ian Dickinson, Michael Wooldridge,"Towards practical reasoning agents for the semantic web",Proceedings of the second international joint conference on Autonomous agents and multiagent systems,July 2003.

[25] Jean Paul Sansonnet, Daniel Werner Correa, Patricia Jaques, Annelies Braffort, Cyril Verrecchia ,"Developing web fully-integrated conversational assistant agents",Proceedings of the 2012 ACM Research in Applied Computation Symposium,2012.

[26] Gabriel L. Somlo, Adele E. Howe,"Incremental clustering for profile maintenance in information gathering web agents",Proceedings of the fifth international conference on Autonomous agents,2001.

[27] Wolfgang Ketter, Arun Batchu, Gary Berosik, Dan McCreary,"A semantic web architecture for advocate agents to determine preferences and facilitate decision making", Proceedings of the 10th international conference on Electronic commerce(ICEC'08),2008.

[28] Andrea Paola Barraza, Angela Carrillo-Ramos ,"Basic requirements to keep in mind for an ideal agent-based web information retrieval system in ubiquitous environments", Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services(IIWAS'10),2010.

[29] Andreas Gerber, Matthias Klusch, Christian Ruß, Ingo Zinnikus,"Holonic agents for the coordination of supply webs",Proceedings of the fifth international conference on Autonomous agents(AGENTS'01),2001

[30] Kurt D. Bollacker, Steve Lawrence, C. Lee Giles,"CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications",Proceedings of the second international conference on Autonomous agents(AGENTS'98),1998