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

S. Senthilkumar¹
¹Research Scholar,
Department of Computer Science and Engineering,
Manonmaniam Sundaranar University, India,
Tamilnadu, Tirunelveli - 627012

G. Tholkappia Arasu², Ph.D
²Principal,
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.

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 an end solution for web users in the internet[21]. Rule Engine for the web could be developed using Enataliment 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 autonomous web agent search model could be developed fully functional similar to CiteSeer[30].

Table-1, Survey table -1

<table>
<thead>
<tr>
<th>Author/ Year</th>
<th>Approach/ Methodology</th>
<th>Merits</th>
<th>Limitations</th>
<th>Data Set</th>
<th>Parameters</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buche P./2013</td>
<td>ONtology-based Data INtegration(ONDINE)</td>
<td>Core Ontology and Domain Ontology is being used. Fuzzy sets are used for analysis</td>
<td>Has not been experimented with Datatables generated from web.</td>
<td>XML Datatables</td>
<td>Score,Cosine Similarity ,</td>
<td>-</td>
</tr>
<tr>
<td>Guan Ziyu/2013</td>
<td>Hypergraph</td>
<td>co-occurrence structure is modeled using a hypergraph</td>
<td>The model does not accomodate for Optimizing the performance using multithreading, multicore, MapReduce or sampling techniques.</td>
<td>ClueWeb09 web collection</td>
<td>Running Time</td>
<td>-</td>
</tr>
<tr>
<td>Skabar Andrew/2013</td>
<td>Fuzzy Clustering Algorithm</td>
<td>Identifies overlapping clusters of semantically related sentences</td>
<td>Can be extended to Attribute Data</td>
<td>News Article Dataset</td>
<td>PageRank, Partition Entropy Coefficient, Purity, Entropy</td>
<td>-</td>
</tr>
<tr>
<td>Name</td>
<td>Method/Algorithm</td>
<td>Description</td>
<td>Dataset</td>
<td>Similarity</td>
<td>Conclusion</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Cui Hang/2003</td>
<td>Log Based Query Expansion Method</td>
<td>Query Expansion based on user interaction. User Query Sessions. Short and Long Queries. Gap between Query and Document space is reduced.</td>
<td>TREC data Encarta Website</td>
<td>Interpolated 11-point average precision, Similarity</td>
<td>Can be tested for internal documents accessed in a university</td>
<td></td>
</tr>
<tr>
<td>Chen Ming-Syan/1998</td>
<td>Maximal Forward Algorithm</td>
<td>Backward References are removed. Number of database scans is reduced Available as part of SpeedTracer Tool.</td>
<td>Synthetic Data</td>
<td>Execution Time</td>
<td>History Information could also be used.</td>
<td></td>
</tr>
<tr>
<td>Chunyu Kit et al /2007</td>
<td>Best First Search Algorithm</td>
<td>no prior knowledge such as adhoc heuristics, no labelled data for training , no similarity analysis of Web page structure</td>
<td>Government Web index – <a href="http://www.infor.gov.hk/chor">http://www.infor.gov.hk/chor</a> gdex.htm.</td>
<td>Precision,Recall F-Score</td>
<td>Reliability could be improved.</td>
<td></td>
</tr>
<tr>
<td>Jean Paul/2012</td>
<td>DIVAlite agents</td>
<td>More visible agents for the screen Feature Based Model</td>
<td>Can be extended to SignLanguage Teaching</td>
<td>Characters</td>
<td>Can be integrated to the systems for guiding challenged people in railway stations.</td>
<td></td>
</tr>
<tr>
<td>Gabriel L. Somlo et al/2001</td>
<td>Incremental Clustering</td>
<td>Agent Feedback is provided using doubling incremental clustering. Use of Web information agents and Text filtering.</td>
<td>Web Data</td>
<td>Precision Recall LGF</td>
<td>Can be extended with web portals that has contents with different category of metadata.</td>
<td></td>
</tr>
<tr>
<td>Wolfgang Ketter et al/2008</td>
<td>Semantic Web Architecture</td>
<td>Advocate Agent Business Model Security support is provided for advocate agent. Support for collective Intelligence.</td>
<td>Optimization of user queries is not considered.</td>
<td>-</td>
<td>Optimization of results is not done.</td>
<td></td>
</tr>
</tbody>
</table>
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 an 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, mining approaches. The research findings has been presented that could be incorporated into the research.

5. REFERENCES

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


