International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Enforcing Adaptive Personalization using Ontology

Sarika P. Aundhakar[1], N. B. Pokale[2]

[1]Computer Department, BSCOER, Narhe, Pune, India

[2]Professor, Computer Department, BSCOER, Narhe, Pune, India

Abstract: ***Due to increasing usage frequency of the search engines user becomes more and more familiar with searching techniques this leads to the fact that search engines becomes the inseparable part our life. As days are passing it is becoming tough to add all the favorite and important links of searched results in the browser to personalize the searched results. So a need of efficient personalization system is required which can store the searched results as long as user wants and automatically adapt the other searched URLs according to the users taste. Many systems are existed which barely store and retrieve the user's searched URLs upon users request but most of them are fail to adapt the user taste according to time. This paper put forwards an idea of user personalization using ontology for the searched news URLs in the search engine. Where every searched result are keep updating based on the user storage and time limit which makes way to store more new useful search results and discard older ones for the efficient and long lasting storage space management.***

Keywords: User Personalization, Ontology, OWL, Protégé,

1. Introduction

In today's world of web, lots of complex data is generated in every second, so performing searching operation on these data is a quite challenging task. When user search for any query on web, a lots of raw results are also get extracted. Web crawlers are the programs that that are to get meaningful data from the web pages by moving in iterative fashion. Fig 1 explains the architecture of the typical web crawler.

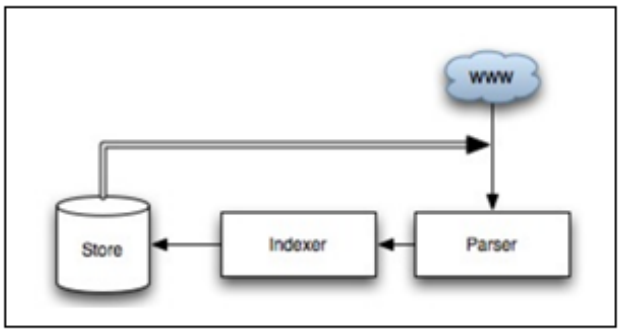

Figure 1: Architecture of web crawler

Web crawlers are widely classifies into the following categories.

- Breadth First Search Algorithm
- Depth First Search Algorithm
- Page Rank Algorithm
- Genetic Algorithm
- HITS Algorithm

Breadth First Search Algorithm: BFS algorithms are the most popular and widely used algorithms for the crawling. BFS is used to search across the neighbor nodes. It starts from the root and then goes parallel for the neighborhood. If it finds the desired entity then control will return else it goes to the next level and start searching in that level.

Depth First Search Algorithm: DFS starts its searching from the root and thus goes deeply to the child node until he gets the answer.

Page Rank Algorithm: These algorithms find out the status of the web pages by searching the backlinks attached to that page.

Following is the formula to find out page rank of a page
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

PR (A) ->Page Rank of a Website,
d ->damping factor
T1,….Tn ->links

Genetic Algorithm: Genetic algorithms make use of fittest function to accomplish the task.

HITS Algorithm: It finds out the relevance of the web page by finding the some set of scores.

Day by day users get smarter, so they don't want to waste much time on manual things like newspaper reading. Most of the users read the news from the web or mobile phones. So the main challenge faced by the news provider is to give the news to the user based on his/her interest or by observing the recent behavior of the user. So it is always been an advantage to store the important news link by the users, so personalization of web search interactively answers this question. Personalization web search has following key steps.

1. Web Data Processing.
2. User modeling in personalization.
3. Recommending Personalized Page Ranking Strategies.

The web personalization methods are widely classifies in three techniques.

- Content based personalization
- Link based personalization
- Function based personalization

In content based personalization, a set of words and its weight is considered for the personalization. Function based personalization uses the sorting algorithms to sort the context based on the user choice.

The remaining whole paper is structured as follows. Section 2 discusses related work and section 3 presents the design of our approach. The details of the results and some discussions we have conducted on this approach are presented in section 4 as Results and Discussions. A section 5 provides hints of some extension of our approach as future work and conclusion.

2. Literature Survey

This section represents all the related works of technologies used in our project. As discussed earlier more and more data get generated by daily activity of the user, it is very difficult to get desired data as along with operational data as unused data is also there. So to remove such unwanted data and thus increase the operational speed preprocessing techniques are used. Preprocessing is a technique of removing unwanted things that will not going to contribute in result by anyways. Below are the vital steps presents in the process of preprocessing.

- Data Cleaning
- Data Integration
- Data transformation
- Data reduction

Data Cleaning: Data Cleaning is a method where missing data is being filled to remove the noise from the data.

Data Integration: It's a process of developing the repository by gathering the data from the multiple resources

Data Transformation: Here the data is transformed to the more appropriate form so that the further processing will get easier.

Data Reduction: Here complex data's are minimized to the simpler form so that it gets easy to analyze the data.

Stop word removal, stemming, special symbol removal are some of the algorithms used for the preprocessing.

[1] Narrate the effects of various stemming and lemmatization approaches on the data retrieval methods. In this paper author discussed the different said techniques with their advantage and disadvantage over other methods. [2] State the context aware stemming algorithm which is the extension of the well-known port stemmer algorithm. CSA gives only meaningful output thus it reduces the error rate of port stemmer algorithm from 76% to 6.7 %. Because of the low error rate algorithm grabbed by the number of mining systems to improve their throughput.

[3] Surveyed the mentioned crawling methods in very deep manner. Along with the crawling some search algorithms are exploited with the advantage and disadvantage of each of these algorithms. At the last part of the paper author conclude that out of all these algorithms they find that the genetic algorithm is much better than other. The reason behind the conclusion is genetic algorithms goes in more iteration compared to other.

[4] Exploits the use different types of web crawlers used for the mobile systems. In this survey paper author put a one vital table that contains 19 web browsers used by the different operating systems of the mobile phones. They explain the origin of each of this web browser, also the platform in which they used. From this survey authors conclude that a web crawler has significant importance in mobile phones also. On later part of their work they extends there research to find out the importance of the crawlers in learning, commercial and social fields.

[5] Represents a BFS that applies on graphs for partitioning with more than 1 billion vertices. The main focus of this paper is to compare edge partitioning of the graph with the vertices partitioning. To show experimental evaluation of the algorithm they make use of BlueGene/L and Linux as an operating system and Poisson Random graph as a base of experiment. The author concludes that this method suits best for if the size of the graph to be traversed is large. Here they used only Poisson graph, the use of another graphs is kept as a future work subject for the authors.

[6] Discussed the problem of HITS algorithm which makes use of hyperlink as a base of search. To show the problem of HITS a tool called Link Viewer is developed. The main reason behind this tool is to show the step by step process of the extraction. In addition to this tool generation, two more methods i.e. projection method and base-set downsizing method are developed by the authors. Projection methods are used for projecting the Eigen vectors on the subspace of the root. So that the root elements will have some relation with the input query. In base-set downsizing method filtered all the pages without finding the links with the multiple pages.

[7] In order to find out the interested data from the large set of dataset association rule mining is used. In association rule mining, Apriori algorithm is a one which uses more frequently. In proposed method authors modifies the basic Apriori algorithm for more refinement in the answer. This modified algorithm has effective result over the basic Apriori algorithm.

To personalize the news [8] gives a novel approach where author analyses the user behavior, there likes, dislikes by observing the logs of the Google news clicks. So after analyzing the logs, Bayesian framework is used for predicting the interest of the user. To bring the whole idea into reality developer combines the content based recommendation with the collaborative filtering algorithm. So the hybrid recommender system is developed to personalize the news based on the click of the user. The experiments show that this hybrid method improves the quality of the result to the great extent and this increase the visitors to the site.

Most of the traditional personalization algorithms make use of TF-IDF methods to accomplish the work. [9] Elaborate new method which uses TF-IDF and domain ontologies to extract the answer. So the system makes use of ontology with TF-IDF it renamed as CF-IDF. The author used Hermes

research as a base for their contribution. At the last authors conclude that the proposed approach using ontology is much better than the traditional approach that uses TF-IDF only.

To personalize the current web search of the user [10] demonstrates the personalization method which have prior interaction of the user with web as a base. The information obtained by the system is realistic as it did more precisely which was not done in traditional methods. By using this information re ranking of web search is done. Here two factors are considered by the developer for personalization.

1. Past history of the user regarding his visits and movement.
2. Documents and emails read by the users. Since the vital behaviors are considered for the purpose of personalization the system gives effective throughput.

World Wide Web is used to give useful information to the user. When the same query is entered by the different users the extracted answers are same irrespective of the need of the user. But different user wants different information so the user interest matters a lot. [11] Demonstrates the useful approach that considers the need of the user for answer extraction purpose. Here the user profile is constructed with the help of collaborative filtering algorithms. This system considers the history of the user behavior for the individual day and not for the complete past of the user. [12] uses ontology for maintaining the user profile required for the web personalization. To do this spreading activation algorithm is used which gives more accuracy compare to the existing methodologies.

3. Proposed Methodology

Here in this section we described how exactly we implement our system with the help of steps shown in figure 2.

Step 1: In this step user login to his/her account by using authorized username and password given to him when he/she registered. Once they logged in, they entered the query for which personalization is needed to be done.

Step 2: As discussed earlier, preprocessing is a way to reduce the size of data by removing the unwanted things from the query. Here preprocessing is carried in three steps. First tokenization of entered query is done. Tokenization is nothing but a process of separating the sentences into the words so that it will get easy for the system to do further processing. Once tokenization is done, its output is given to the process of stemming as shown in figure. In stemming the derived words are converted to its base form without changing the meaning of the word. Stemming used for to enhance the system in terms of speed. In next step stop words are removed from the query like "a, is, the, an" etc.

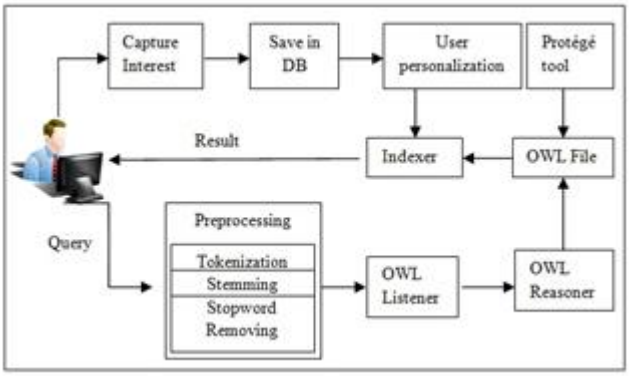

Figure 2: Overview of our Approach

Step 3: Extensible Markup Language (XML) is used to describe the data, not to display data unlike HTML. So here owl file generated by using protégé tool is accessed by using the xml expressions and tags of the xml. The main reason behind this ontology listener is to generate the object property of owl file.

Step 4: Here classification hierarchy is generated by using the protégé listener. Protégé allows different OWL Reasoner to be plugged in; the Reasoner sent with Protégé is called Fact++. The generated ontology files us passed to the Reasoner part to compute the hierarchy of classification. In addition to this, logical efficiency of the ontology is also checked. In Protégé 4 asserted hierarchy is the 'manually constructed' class hierarchy. There are some hierarchies which are automatically developed by the reasoned, these hierarchy's are known as inferred hierarchy. This step is one of the vital steps of proposed method as in this step only properties of owl file are extracted by the owl reasoner. To identify the child and parent of each entity the owl property object generated in previous step is taking as input by this step. Node representation is created by using hierarchy set in previous step. Once nodes are created, system rearranged those nodes in the form of tree object.

Step 5: With the help of protégé tool OWL-lite ontology is developed. Protégé is an ontology editor used which is openly available in the web.

Stanford University used for knowledge extraction. Till todays date protégé has more than 160,000 registered users. Protégé has built in support for two way personalization. Protégé editor allows the users to develop frame based ontologies as per the standards of OKBC (Open Knowledge Base Connectivity protocol). In this model, ontology can be created by a set of classes structured in a hierarchical type to represent a particular domain's main concepts. The protégé editor allows the user to develop ontology using the OWL language. To identify the child and parent classes, the OWL property object is used [14].

The ontology which system used is OWL-Lite. As the bane indicates OWL-Lite is a lite version of the ontology. Other than this OWL-Lite has much more limitations compare to the other version of the ontologies which are OWL DL and OWL Full languages. E.g. owl lite classes are used by using the named super classes and only few advantages of ontologies are used. The following features of OWL-Lite

which are correlated with rdf schemas are used in our application

- Class
- rdfs:subClassOf
- rdf:property
- rdfs:subPropertyOf
- rdfs:domain
- rdfs:range
- Individual

Step 6: Here term weight of the preprocessed user query is used to find out the OWL class with the help of tree object generated in previous step. After this, index for the hierarchy is used.
Step 7: Here indexed classes of the tree is used for fetching the URL which are stored in databases.
Step 8: This step is important for capturing the user evidence factors which are used for the personalization. Such as username, date, time, query, interested links, and session timings.
Step 9: The step refers to user personalization model where previous evidence of the same query entered by the user is checked. Checks for previous evidence for the query by the user.

If system finds the evidence then system fetches the evidence and then sends it to the indexer part. Indexer part is a part where evidence is properly displayed with the existing results. After this user personalization is done where user interests and process queries are used in timely manner. Below is the equation which represents the combined adaptive personalization.

$$P_m(X: Y) = \sum_{j=0}^{k} \cdot \sum_{i=0}^{N} Ul(1)$$

Where X= { i=1,…….N} is the set of URLs used for user personalization.
Where Y= { j=1,…….k} is the set of adaptive queries.
U_l is the User interest URL
Time based adaptation can be represented with the below equation.
$F_t = (t_c - t_l) > T:(Q-U_q)$ (2)

Where t_c is the current time, t_l is the lastly updated time, T is the adaptive time for the user, Q is the set of Queries and U_q is the user query to delete.
The above two equations can be summarized in the following algorithm.

Algorithm: Time based adaptive user personalization

// input: U_l is the interest user interest URL
// N is the number of user personalized URL
//K is the number of user adaptive queries
// t_c is the current time and t_l is the last query updated time
// T is the adaptive time for the user, Q is the set of Queries
// U_q is the user query need to delete.
//output: Q is the set of adaptive user personalized

Function : adaptive User Personalization (U_l,N,K,t_c,t_l,T,Q,U_q)

1. Get the N user interest URLs
2. Update all the URLs U_l into the database
3. **for all** $Q \in [1; k]$ do
4. Get the query Q, date, time and query count and update in database
5. **if** (t_c- t_l)>T then delete query U_q
6. Reset the set Q
7. **return set**

4. Results And Discussions

In our proposed system we evaluate the efficiency of the user personalization systems. Also we checked whether the techniques mentioned by us provide proper results to the user or not. Proposed system creates the ontology of more than 50 keywords of the general news categories like business, sports and health etc. with the respective URLs in the database. Then with the help of adaptive sequences user personalization is given to the user. Experiments shows that, our system gives more efficient results than the other as in our system sessions are increases on specific time interval.

We conducted a survey to know interest degree of the 10 users which captured on five different time intervals. So our survey shows always last interval always has more range over the first one. So this shows that system captures the more interest with the help of adaptive rules.

To show the experimental evaluation of the system we measures the associated user personalized URL and the adaptive interest of the user for the given query. For better understanding of retrieval effectiveness precision and recall parameters are used. And for this purpose again 10 users are considered. As the users personalized details are captured on fixed interval of time, it will get easy to find out the precision and recall.

For more clarity we assign
- A = Expected interest degree.
- B = the number of relevant interest not retrieved, and
- C = the number of irrelevant interest retrieved.

So, Precision = A/ (A+ C)
And Recall = A/ (A+ B)

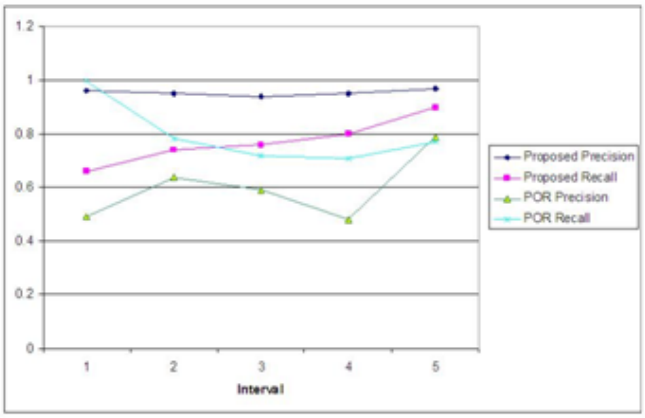

Figure 3: Precision and recall at different interval time for adaptive personalization process

By considering a regular interval of time precision and recall of the system is calculated as shown in figure 3. All the values of the user interests are normalized between 0 to 1.

Graph clearly indicates the difference of proposed model with the user personalization. In partial order manner [13] (POR) .By studying the algorithm we concludes that our system give much better result in terms of precision and recall as interval time is increasing. Whereas POR system has little lesser ratio of precision and recall compare to our system of adaptive personalization through ontology. This indicates our proposed system of adaptive user personalization literally opens many possibilities in enforcing the system of user personalization.

5. Conclusion and Future Scope

Proposed approach of user personalization successfully captures the user interested URLS on the click of the URL. And process this URL for the adaptive personalization theme where URLs are managed till given weight and time parameter. On exceeding these parameters system automatically adopt the new URLs which are searched by the designed ontology based focus search engine for the limited URLs of the news. Adaptive personalization can be enhancing to give more intelligence along with the coordination of the web browser to catch the user interest based on the surfing time on the specific web page.

References

[1] Michal Konkol and Miloslav Konopík , “Named Entity Recognition for Highly Inflectional Languages: Effects of Various Lemmatization and Stemming Approaches” Named Entity Recognition for Highly Inflectional Languages
[2] K.K. Agbele, A.O. Adesina, N.A. Azeez , A.P. Abidoye “Context-Aware Stemming Algorithm for Semantically Related Root Words” © 2012 Afr J Comp & ICT.
[3] Pavalam, S. M., et al. "A survey of Web crawler algorithms." IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011 ISSN (Online): 1694-0814
[4] Pavalam, S. M., M. Jawahar, and K. Akorli Felix. "Web Crawler in Mobile Systems." International Journal of Machine Learning and Computing 2.4 (2012): 531-534.
[5] Chow, Edmond, Keith Henderson, and Andy Yoo. "Distributed breadth-first search with 2-D partitioning." Lawrence Livermore Nat. Lab., Livermore, CA, Tech. Rep. UCRL-CONF-210829. 2005.
[6] Nomura, Saeko, et al. "Analysis and improvement of HITS algorithm for detecting Web communities." Systems and Computers in Japan 35.13 (2004): 32-42.
[7] Usharani, J., and Dr K. Iyakutti. "Mining association rules for web crawling using genetic algorithm." International Journal Of Engineering And Computer Science ISSN (2013): 2319-7242.
[8] Liu, Jiahui, Peter Dolan, and Elin Rønby Pedersen. "Personalized news recommendation based on click behavior." Proceedings of the 15th international conference on Intelligent user interfaces. ACM, 2010.
[9] Goossen, Frank, et al. "News personalization using the CF-IDF semantic recommender." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.
[10] Teevan, Jaime, Susan T. Dumais, and Eric Horvitz. "Personalizing search via automated analysis of interests and activities." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.
[11] Sugiyama, Kazunari, Kenji Hatano, and Masatoshi Yoshikawa. "Adaptive web search based on user profile constructed without any effort from users."Proceedings of the 13th international conference on World Wide Web. ACM, 2004.
[12] Sieg, Ahu, Bamshad Mobasher, and Robin Burke. "Web search personalization with ontological user profiles." Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007.
[13] Lau, Raymond YK, Dawei Song, Yuefeng Li, Terence CH Cheung, and Jin-Xing Hao. "Toward a fuzzy domain ontology extraction method for adaptive e-learning." *Knowledge and Data Engineering, IEEE Transactions on* 21, no. 6 (2009): 800-813.