ISSN:0975-9646

Website Personalization based on Link Analysis, Navigational Patterns and Web Content

Atul N. Pawar[1], Dr. Suresh K. Shirgave[2]

[1]*Department of Computer Science and Engineering,*
Rajarambapu Institute of Technology, Sakharale, Maharashtra 415414 India
[2]*Department of Information Technology,*
DKTE's Society's Textile and Engineering Institute, Ichalkaranji, Maharashtra 416 115 India

***Abstract*— The recommendation systems are used in variety of the application for rating of products or identifying user preferences. In this paper we present a hybrid recommendation system using different approaches such as web content mining, web structure mining and web usage mining. An important application of the recommendation system is Website Personalization in which information is gathered, processed and analyzed to deliver correct information to each visitor at right time. The techniques used in web mining have certain drawbacks if used independently. An integrated approach will overcome the drawbacks existing in these techniques.**

***Keywords*— Recommendation System, Usage base Page Ranking, Markov Models, Ontology and Semantic Similarity.**

I. INTRODUCTION

World Wide Web is massive, explosive and unstructured data. In earlier WWW was used as source of information. As use of WWW increases, it also plays role in banking, marketing and trending. Many of the businesses are running on the WWW. Web mining is component of data mining, focusing on the World Wide Web as a main resource or data for the mining. As data mining process is used extract the knowledge from the information, web mining is also used for extracting the knowledge from web and can be used for different purposes. Web mining uses web site structure, contents of individual page, server logs and services as its primary data resource. This knowledge can be used increase benefits of commercial web sites, approaching new trends in market and mainly help users.

A very important use of web mining is website personalization. The system used for web site personalization is called as recommendation system uses many web mining techniques based on Big Data Analytics, Artificial Intelligence and Information Retrieval. Website personalization can be defined as adopting the requirements or preferences of users for managing the contents of the website. This dynamic website content generation can be achieved with two methodologies. One easy approach is to present a menu for the user, so he/she can manually select the preference and change the contents. In web site personalization, the part of collecting and analysing information about web and user is called as web mining which deals with web structure mining, web content mining and web usage mining.

Web structure mining is used to discover useful knowledge from the structure of the web site. With this method relationship between the WebPages are identified with the help of information based on how they are linked with each other. This hyperlinked information enables to cluster WebPages which has similar structure.

Web content mining is used to extract useful information from web page content. Content mining deals scanning and mining of the text present in Webpage. Web pages designed with HTML also contained pictures and graphs which are also relevant to content mining. In the scanning process, contents on the web pages are scanned and clustered with similar keywords or texts. When user fires query with the search engine, a cluster with keyword existing in query is triggered and results are displayed to user. In web personalization process, pages visited by user are scanned for keywords existing on web contents for identification of user interest. Content mining has provided huge area in semantic web. Semantic webs are intuitively designed for providing more information on web contents with metadata and properties, which helps to classify information accurately. In semantic web domain ontology provides easy way for classification of information. Considering the amount of information available on web semantic web content mining provides more benefits in web personalization as compared with web content based on text mining.

Web usage mining extracts user access patterns from web log data or server logs. Website Personalization is based on the Web usage mining i.e. extracting useful information from the server logs (user history) termed as Navigational Patterns. With this one can easily find out user interest and provide useful information to the user. Most of the related algorithms uses the same techniques for the recommendations and ignores structure of the website, underestimating rank of pages. Some another approaches used for the web mining are clustering, classification and association rules. In clustering, data set are structured into set of groups using similarity measures such as Hamming distance or Euclidean distance. In runtime the current session is measured to the predefined groups to set the group of the session. While in classification approach, a known structure is applied to the new data. Finally Association rules try to correlated user behaviour with previous stored behaviour. All these approaches are resulting in common output, recommendations.

II. RELATED WORK

With more and more information available on the internet, the task of making personalized recommendations to assist the user's navigation has become increasingly important. Web usage mining requires to model user web navigation behaviour. Markov models are very useful to model this scenario. Some of the algorithms proposed uses clustering approach. A dynamic clustering based method can be used to increase accuracy of Markov Models [2]. In clustering based methods, classic distance based clustering evaluation method determines distance between an object and cluster centroid are not suitable in the model based clustering domain [3]. Yang Liu, Xiangji [2007], proposed a model to capture user access sequences as stochastic process, using mixture of Markov models for defining relationship in user accesses. The prediction accuracy of these Markov models can be increased with higher order Markov models. Current frameworks are not suitable hence fails [4]. Apart from these Markov based models; some new models are exists for prediction of pages such as WebPUM [5]. In this approach navigation patterns are clustered for online prediction.

Form another approaches, content Mining is also very useful in web site personalization. Web content mining describes the discovery of useful information from the web content/data/documents. Content mining deals with Unstructured Text data mining, semi structured and structured data mining [6]. Unstructured data mining extracts information form unstructured test files, termed as Knowledge Discovery in the Text (KDT). Semi-structured and structured data mining uses HTML or XML files for information extraction. During content mining, pages are categorized into collection conforming to common schema and common template [7].

In early, k-means algorithm was used to identify the recommendation set. A Markov model based approach is proposed by [8] which is applied for the learning extraction models. For semi-structured and structured documents various approaches can be used. In Multilevel Database approach hypertext documents are used as data repositories which contain lower level information in databases. At higher levels Meta data or generalization are extracted from lower levels [9] As lots of the information is available on web, management of meta data it becomes critical. Domains re used to define schema for these Meta data and can be used globally [10]. An incremental integration of a portion of the schema was done from each information source rather than relying on a global heterogeneous schema [11]. One more approach is also used Web Query System. Web based query systems or languages such SQL are used for this. W3QL combines structure queries combine's structure queries and content queries base on the information retrieval techniques [12]. WebLog – a logic based query language for restricting extracts information from web information sources was designed to overcome drawbacks in heterogeneous environment [13]. Whereas ontologies are content theories about the classes of individuals, properties of individuals, and relations between individuals that are possible in a specified domain of knowledge [14].

Ontology generation is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts from a corpus of natural language text, and encoding them with an ontology language for easy retrieval[15]. Automatic Generation of Ontology Based on Database discussed about rules for generation of the ontology elements based on relational database [16].

Automatic Ontology Creation from Text for Nat With more and more information available on the internet, the task of making personalized recommendations to assist the user's navigation has become increasingly important. Web usage mining requires to model user web navigation behaviour. Markov models are very useful to model this scenario. Some of the algorithms proposed uses clustering approach [17].

By viewing the Web user's navigation in a Web site as a Markov chain, Markov model can be build a for link prediction based on past users' visit behaviour recorded in the Web log file. Assume that the pages to be visited by a user in the future are determined by his/her current position and/or visiting history in the Web site. Construction of a link graph from the Web log file, which consists of nodes representing Web pages, links representing hyperlinks, and weights on the links representing the numbers of traversals on the hyperlinks. By viewing the weights on the links as past users' implicit feedback of their preferences in the hyperlinks, we can use the link graph to calculate a transition probability matrix containing one-step transition probabilities in the Markov model.

The Markov model is further used for link prediction by calculating the conditional probabilities of visiting other pages in the future given the user's current position and/or previously visited pages. An algorithm for transition probability matrix compression is used to cluster Web pages with similar transition behaviours together to get a compact transition matrix. The compressed transition matrix makes link prediction more efficient [18].

III. EXISTING SYSTEM

The existing system uses Usage based Page rank, Localized Usage based Page rank and Hybrid Probabilistic Predictive algorithms for generation of the recommendation using Markov models. Both of these algorithms use Navigational Graph which captures previous user sessions.

A. Navigational Graph (NG)

Navigational Graph is weighted directed graph. All paths followed in session are started with one special node called as 'Root' and ends with special node called as 'End'. An edge in graph represents visit of user from one page to second page. And weight assigned to the edge is number of times user has visited that page from previous page. For a single session, there exists a path from start node to end node and intermediate nodes in the path are nodes representing pages visited by user. Navigational Graph represents number of times a page is visited as page weight and number of times a link is followed as weight of the edge.

B. Usage based Page Rank(UPR)

Usage based Page Rank algorithm computes the rank of the page as n[th] iteration of the following formula.

$$UPR_i^n = \varepsilon \sum_{x_j \epsilon\, In(x_i)} \left(UPR_i^{n-1} \times \frac{w_{j\to i}}{\sum_{x_k \epsilon Out(x_j)} w_{j\to k}} \right) + (1-\varepsilon) \frac{w_i}{\sum_{x_j \epsilon WS} w_j}$$

Where, *UPR* is normalized matrix whose column sum to 1, $In(x_i)$ indicates set of in-links of pagex_i, indicates set of out-links of page$Out(x_j)$, $w_{j\to i}$indicates weight of edge in NG, w_i indicates weight of the page in NG, $\sum_{x_j \epsilon WS} w_j$indicates sum of weights of all pages in web site and $(1-\varepsilon)$ is the dumping factor set to 0.15. The ranks generated by UPR are used as transition probabilities of the pages and matrix of the same is called as transition matrix (*TP*).

C. Localized Usage based Page Rank(l-UPR)

Ranks generated by UPR are applied to small subset of NG called as personalized NG (prNG). The fraction of NG is expanded with a specified depth *d* in *l*-UPR. Expansion process include removal previously nodes from NG in the current path.

D. Hybrid Probabilistic Predictive Model(hPPM)

hPPM defines transition probabilities between pages using path prediction and selecting most probable path among candidate path depending upon the user's current path followed. This algorithm uses Navigational Graph to identify path probabilities and extends Morkov Models for the prediction. It utilizes sequential dependency of navigational behaviour and computes transition probabilities using chain rule. For m[th] order Markov Model, the path probability of following the path $x_1 \to x_2 \to \ldots \to x_k$ equals to,

$$P(x_1 \to \ldots \to x_k) = P(x_1) * \prod_{i=2}^{k} P(x_i | x_{i-m} \ldots . x_{i-1})$$

IV. PROPOSED SYSTEM

A new recommendation system integrates transition probabilities generated by UPR, *l*-UPR and hPPM algorithms and semantic similarities generated by semantic content mining. Recommendations, using semantic content mining are generated by creating ontology for the domain, calculating taxonomical distances between concepts and calculating the semantic similarity between web pages. Thus achieving integration of three web mining approaches; web structure mining, web usage mining and web content mining.

A. Creating Ontology

A concept is one node in RDF triple. Ontology i.e. concept hierarchy is built by identifying concepts those are defined on web pages and specifying relationship between the concepts. Two concepts are directly related to each other if they are part of same RDF triple.

B. Semantic Distance and Similarity Measurement

Distance between two concepts is taxonomical distance in concept hierarchy and is computed using Shortest Distance Path algorithm. These distances between two concepts are normalised using following formula,

$$normalized\ distance(n1, n2) = \frac{distance(n1\ and\ n2)}{largest\ distance}$$

Similarity is measured as inverse of the normalized distance. As distance between two nodes is less, maximum is the similarity and distance is maximum, similarity is less.

$$similarity\ (node1, node2) = 1 - normalized\ distance$$

Finally similarity between two pages is defined as the average similarity that exists between all nodes defined on those pages. Consider, the number concepts defined on page x are n1 and number of concepts defined on page y are n2. Then similarity between page x and page y is defined as below.

$$page\ similarity\ (x, y) = \frac{\sum_{i\epsilon\ n1\ and\ j\epsilon n2}\ similarity\ (i, j)}{n1 + n2}$$

Thus, similarity matrix *DM* represents all the semantic distance between all pages and is of size

C. Integrating recommendations

UPR, *l*-UPR and hPPM results in set of recommendations based on the structure and behavioural patterns. These results are updated with similarity indices calculated based on the semantic web content mining. Thus probability transition matrix *TP* is integrated with similarity matrix *DM*. The size of these matrices is same-$N \times N$. Integration of the results is done as follows:

$$R[O_i, O_j] = TP[O_i, O_j] + \begin{cases} 1 - \frac{DM[O_i, O_j]}{\sum DM[O_i, O_k]}, & DM[O_i, O_j] > 0 \\ 0 & otherwise \end{cases}$$

Where, matrix *R* is the resultant matrix obtained after integration. O_i and O_j indicate two pages represented with index numbers assigned to them in both matrices. Transition probabilities are updated to reflect semantic similarities.

V. EXPERIMENTAL SETUP AND EVALUATION

The recommendation system discussed earlier works offline. It uses some sessions as input to the system and some sessions are used for the testing.

A. Data Sets Used

System uses server logs of two different web sites http://data.semanticweb.org and http://dbpedia.org. These logs are used for creating sessions and are referred as SWDF data set and DBPedia data set respectively. For SWDF data set logs of three months are used for creating sessions, while for DBPedia, logs of one month are used. Table I shows details of data sets and sessions.

Data Set	No. of Input Session	No. of Test Sessions	No. of Pages
SWDF	630	65	2225
DBPedia	168	5	1981

Table I- Training and Testing Sessions

B. Evaluation of System

The system is evaluated on the basis of correctness in the prediction and improvement in correctness after integrations of results. From any test session a path is given as input to the system and it is checked that the next page in the test session, if it exists in the predictions done by system. The proposed system is recommendation system, and hence correctness of the system is directly measured on the basis of correct predictions done by the system. The prediction probabilities of all the framework is determined with correct and incorrect predictions generated by framework. For given input sets, if number of correct predictions done by framework are *Pc* and number of incorrect i.e. false predictions are *Pf*, then prediction probability *P* of the framework is given by,

$$Prediction\ Probabilty\ (P) = \frac{Pc}{Pc + Pf}$$

C. Results obtained before integration

Table II, shows the prediction probabilities of the framework obtained before integration of the results.

In Table II, '1 MM 1 depth' indicates, first order Morkov Model used in Usage based Page Rank with depth one for the path. '1 MM 2 Depth' indicates Markov Model used with order one with depth two. '2 MM 2 Depth', '2 MM 2 Depth' and '4 MM 2 Depth' indicates higher order Morkov Model used in l-Usage based page Rank with depth of path two and order two, three and four respectively.

In case of SWDF dataset, for first order Morkov Model with different depth, results remains same because of the framework considers only next immediate recommendations. Removing of the pages which are already exists in the prNG does not make any changes in the results. But in case of DBPedia with depth 2, removing already visited page makes children of that page as first recommendations with a significant positive effect, increasing the prediction probability.

Data Set	1 MM 1 Depth	1 MM 2 Depth	2 MM 2 Depth	3 MM 2 Depth	4 MM 2 Depth
SWDF	0.283784	0.283784	0.012048	0.021622	0.064935
DBPedia	0.061228	0.20482	1.0	0	0

Table II- Prediction Probabilities before integration of transition and similarity matrix

as order goes on increasing in SWDF also prediction probability increases. This does not exist with DBPedia. In case of DBPedia second order Morkov Model with depth 2 has probability 1, while it is 0 for higher orders. In DBPedia, sessions used for test are only 2% of the input session and SWDF that is 10.31%. The numbers of input sessions of DBPedia are 27% of the SWDF.

All sessions of SWDF and DBPedia that are given as input to the framework, they uses 2225 and 1981 pages respectively. If it is assumed that all the sessions contain distinct page visits, then SWDF data set has 4 distinct pages are in one session, while DBPedia has 12 distinct pages per session. As session length used in the framework varies from 5 to 30, occurrence of pages in the sessions of SWDF is more as compared to DBPedia, this result in more numbers of page visits and path frequencies. The prediction capability of the framework is also dependent on which data set is used. Prediction probability of the framework can be increased by increasing the number of sessions used as input to the framework.

D. Results obtained after integration

The results obtained integration of transition matrix and similarity matrix is shown in Table III.

Data Set	1 MM 1 Depth	1 MM 2 Depth	2 MM 2 Depth	3 MM 2 Depth	4 MM 2 Depth
SWDF	0.29632	0.29632	0.012048	0.021622	0.064935
DBPedia	0.063228	0.20673	1.0	0	0

Table III- Prediction Probabilities after integration of transition and

Comparing Table II and Table III, the results obtained after integration of transition matrix and similarity matrix shows small improvement in the prediction probability of the framework. These improvements with SWDF and DBPedia data sets are shown in Figure 1and Figure 2 respectively.

The similarity matrix used in the integration was obtained from the concept hierarchy build built and is based on the taxonomical distances between the concepts. The graph created is strongly connected. All concepts are connected with each others with small distance in nodes, so distances between pages are also very low. When transition matrix and similarity matrix, both are integrated, there is linear change in the all the values of resultant matrix. The linear change in all values of matrix preserves the differences that exist in earlier matrix.

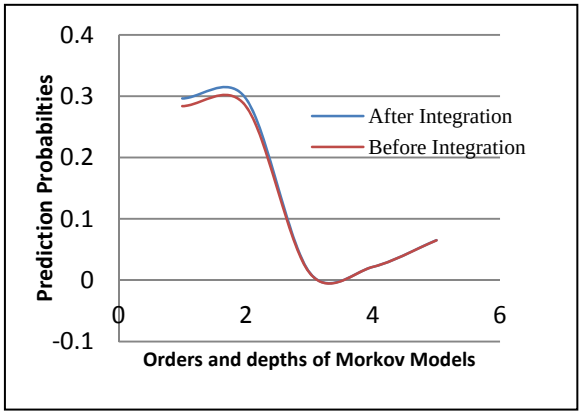

Figure 1- Variations in the results of SWDF Data set

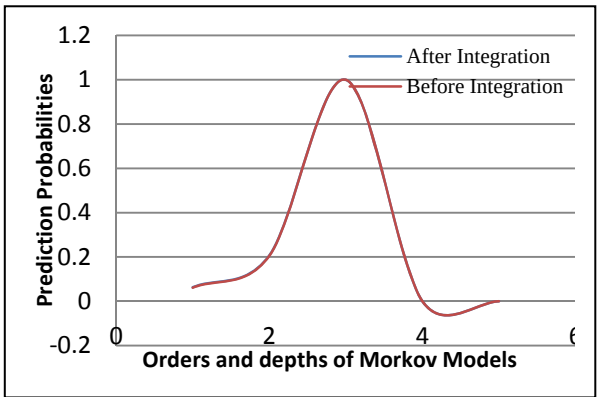

Figure 2 Variations in the results of DBPedia Data set

VI. CONCLUSIONS

A variety of recommendation systems with different approaches are available for web site personalization based on learning of the usage data used to identify behaviour of past user. Many of them are exclusively based on the usage data and not considering the web site structure as well as contents on the page. Exclusive use of different approaches leads to be present drawbacks in terms of missing the use proper information available. By instinct, data used by the different approaches try to overcome the drawbacks of single approach to be used. The proposed framework uses a hybrid approach to incorporate the web usage mining, web structure mining and web content mining.

The algorithm UPR, which uses usage patterns and link structure, is integrated with semantic similarities between pages. As earlier UPR is applied in two contexts; one is personalized subgraph of NG and second is probabilistic predictive models. The results obtained in earlier shows that results obtained with hPPM are not promising with higher order Markov models. Results generated by the framework shows small improvement in the recommendations after the integration. These improvement subsists in the lower order Markov models, while results of higher order Markov models remains unchanged.

The approach used in the framework, for creation of concept hierarchy is based on the relationship of concepts and has ignored the relationship that exists in the different types properties of the concepts. There exists a path from every node to other nodes in the graph- representing the concept hierarchy. These concepts, defined on web page as part of it, reflect very small difference in similarity of pages. When similarities are integrated, changes in values of transition matrix are continuous. Deviation in the similarities of pages can be improved with number of concepts on web pages and relationship between the properties of concepts such as object properties, data type properties or annotation properties. The semantic web pages used for content mining are required to follow the standards used for designing the semantic web page, so they can be used effectively.

REFERENCES

[1] Kosalaand Bockeel, “A Review of Key Data Mining Technologies/ Applications”, IRMA Cooley et al., (2000)

[2] Borges, José, and Mark Levene. "A dynamic clustering-based Markov model for web usage mining." arXiv preprint cs/0406032 (2004).

[3] Thuraisingham, Bhavani M., Latifur Khan, Murat Kantarcioglu, Sonia Chib, Jiawei Han, and Sang Hyuk Son. "Real-Time Knowledge Discovery and Dissemination for Intelligence Analysis." In hicss, pp. 1-12. 2009.

[4] Eirinaki, Magdalini, and Michalis Vazirgiannis. "Web mining for web personalization." ACM Transactions on Internet Technology (TOIT) 3, no. 1 (2003): 1-27.

[5] J alali, Mehrdad, Norwati Mustapha, Md Nasir Sulaiman, and Ali Mamat. "WebPUM: A Web-based recommendation system to predict user future movements." Expert Systems with Applications 37, no. 9 (2010): 6201-6212.

[6] Azmy, Michael. "Web content mining research: A survey." ACM SIGMOD Explorations 1, no. 01 (2005): 203-212.

[7] Arasu, Arvind, and Hector Garcia-Molina. "Extracting structured data from web pages." In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 337-348. ACM, 2003.

[8] McCallum, Andrew, Dayne Freitag, and Fernando CN Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." In Icml, vol. 17, pp. 591-598. 2000.

[9] Wang, Jiawei Han Yongjian Fu Wei, Jenny Chiang Wan Gong Krzysztof Koperski, Deyi Li, Yijun Lu Amynmohamed Rajan Nebojsa Stefanovic, and Betty Xia Osmar R. Zaiane. "DBMiner: A system for mining knowledge in large relational databases." In Proc. Intl. Conf. on Data Mining and Knowledge Discovery (KDD’96), pp. 250-255. 1996.

[10] Khosla, I., B. Kuhn, and N. Soparkar. "Database search using informatiuon mining." In Proc. of. 1996.

[11] King, Roger, and Michael Novak. "Supporting information infrastructure for distributed, heterogeneous knowledge discovery." In Proc. SIGMOD, vol. 96. 1996.

[12] Konopnicki, David, and Oded Shmueli. "W3qs: A query system for the world-wide web." In VLDB, vol. 95, pp. 54-65. 1995.

[13] Quass, Dallan, Anand Rajaraman, Jeffrey Ullman, Jennifer Widom, and Yehoshua Sagiv. "Querying semistructured heterogeneous information." Journal of Systems Integration 7, no. 3-4 (1997): 381-407.

[14] Touzi, Amel Grissa, Hela Ben Massoud, and Alaya Ayadi. "Automatic ontology generation for Data mining using FCA and clustering." arXiv preprint arXiv:1311.1764 (2013).

[15] Navigli, Roberto, and Paola Velardi. "Learning domain ontologies from document warehouses and dedicated web sites." Computational Linguistics 30, no. 2 (2004): 151-179.

[16] Zhang, Lei, and Jing Li. "Automatic generation of ontology based on database." Journal of Computational Information Systems 7, no. 4 (2011): 1148-1154.

[17] Balakrishna, Mithun, and Munirathnam Srikanth. "Automatic ontology creation from text for national intelligence priorities framework (NIPF)." In Proceedings of 3rd International Ontology for the Intelligence Community (OIC) Conference, pp. 8-12. 2008.

[18] Eirinaki, Magdalini, and Michalis Vazirgiannis. "Web site personalization based on link analysis and navigational patterns." ACM Transactions on Internet Technology (TOIT) 7, no. 4 (2007): 21