ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

A Novel Approach for Web Intelligence

M. Ambika, Dr. K. Latha

Dept of CSE, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu, India.

Dept of CSE, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu, India.

Abstract— Web Intelligence is a fast-growing area of research that combines multiple disciplines including artificial intelligence, machine learning, data mining, natural language processing. Making the web intelligent is the art of customizing items in response to the needs of the users. Predicting users' behaviors will expedite and enhance browsing experience, which could be achieved through personalization. Web Intelligence provides a platform which empowers internet users to find the most appropriate and best information for their interest. This paper proposes a novel approach for making the web adroit.

Keywords— web mining, web usage mining, web personalization.

I.INTRODUCTION

In the today's era of information technology, the Web provides every Internet citizen with access to the abundance of information. We can say that "We are drowning in data, but starving for knowledge." Thus considering the impressive variety of the web, retrieving interesting, relevant and required information has become a very difficult task. A popular and successful technique which has shown many promises is web mining.

Web mining has become a hot research topic, which combines two of the prominent research areas comprising the data mining and the World Wide Web. "Web mining is a technique to explore, collate and extract patterns in the data content of web sites by using traditional data mining techniques". More over web mining research is a multidisciplinary field from several research communities, such as database, information retrieval, and artificial intelligence research communities, especially from machine learning and natural language processing.One main issue to be considered in web mining is "Reach the Right Person with the Right Message at the Right Time". The Web should have the ability to sense and adapt to the needs and preference of the user. One effective approach to this is to make the web intelligent. Here each and every individual visitor will be treated as individuals with their own goal and target when searching for information on the web. In the paper we proposed a novel approach to perform web personalization with intelligence.This paper is structured as follows: Section II, give an overview about web mining, Section III, shows deep and intense study of web intelligence. Section IV and V elaborates the proposed methodology and evaluation techniques. Finally, we concluded the paper in section VI and outlined some promising area of future research.

II. WEB MINING

According to Oren Etzioni (1996), web mining is the application of data mining techniques to automatically discover and extract information from World Wide Web documents. The information gathered through web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns. Web mining can also be used to understand customer behaviour, evaluate the effectiveness of a particular web site. Some of the issues related to the web [1] are information retrieval, information extraction, personalization of the web information, learning the consumers and individual user's behaviour which can be used for user modelling and profiling.

Web mining deals with the discovery and analysis of useful information from the www. As suggested by Kosala and Blockeel (2000) and Qingyu Zhang et al. (2008), Fig. 1 depicts the Web mining sub tasks. The

relevant data are retrieved from the web and then selection and pre-processing are done. Machine language or data mining techniques can be used for discovering the patterns, and finally, it is validated and visualization [1] [13].

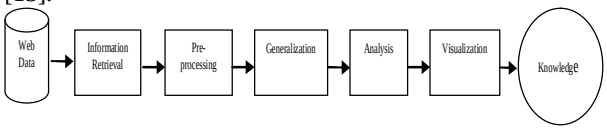

Fig. 1 Web mining task

Web mining process is of three different types based on the nature of the data such as [1][3][5],

Web Content mining: extract useful information or knowledge from the web page. It deals with text, image, records, etc.

Web Structure Mining: is the specialization of web content mining except it is tag content, mainly deals with the hyperlink structure of the web sites.

Web Usage mining: deals with the data mining techniques applied to web usage information such as web server logs, proxy logs, http logs, etc. It can be used to predict user behavior while the user interacts with the web.

III. WEB INTELLIGENCE

The Web has been considered as the largest repository of information. In web based learning gathering accurate, effective and useful information or knowledge has become a challenging issue. Each and every user has their goal when searching for information on the Web [19] [20]. Some of the issues in the current web search engines are,

- Lack of user adaption
- Retrieve results based on web popularity rather than user's interests
- Relevant results beyond first few pages have a much lower chance of being visited by user
- Provides the same result for the same query.
- Relevance estimate is system centered approach.

One solution to above issues is to make the web, intelligent with the ability to sense and adapt to the users need. Every individual visitor is treated as individuals, with targeted content and offers that appeal to their implicit or explicit needs [9]. As a result it can achieve

- ✓ tailoring search results to individuals based on knowledge of their interests
- ✓ identify relevant documents and put them on top of the result list
- ✓ filter irrelevant search results
- ✓ relevance is crucial
- ✓ making visitor's time more productive and engaging.

Thus Web intelligence is an art, which can make the web as Wisdom web. Ning Zhong et.al (2000) coined the term Web Intelligence. It is the area of study and research of the application of artificial intelligence and information technology on the web in order to create the next generation of web empowered system [22]. It can be defined as a process of helping users by providing customized or relevant information on the basis of web experience to a particular user or set of users. It can also be defined as a recommendation system that performs information filtering with intelligence. The general phases of web intelligence process are (Fig. 2),

- Learning phase,
- Pattern extraction phase
- Information retrieval or Recommendation.

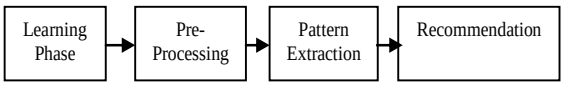

Fig. 2 The phases of web intelligence

A. Learning Phase

Data collection phase is also known as learning phase. This phase is used to develop a user profile definition and user segments. A user profile is a collection of attributes that user will need to either maintain or derive in order to support personalization [17]. It is very important phase. Other phase depend on this phase. The user profile can be collected by two ways.

- implicit profile
- explicit profile

Implicit profile: The attributes can be derived from browsing patterns, cookies, and other sources. No feedback from the user is collected. It can also be derived from Geo Locations, Behavioural Learning, Contextual Learning, Social Collaborative Learning, and Simulated Feedback Learning. It is based on the characteristic of an individual's such as Interest, Social Category, Context, role, functional area. It can also make use of current session activities, past session activities.

Explicit profile: The attributes come from online questionnaires, registration forms, integrated CRM or sales force automation tools, and legacy or existing databases. External feedback is collected from the user which provides rating or preferences. User personal information like age, occupation, religion, economic, environment etc can be used to generate user profile.

Pre-processing consists of elaborating the raw web access logs to produce data in a format usable by the Pattern Discovery phase.

B. Pattern Discovery

Pattern discovery aims to detect interesting patterns in the pre-processed Web usage data by deploying various data mining techniques. These methods usually consist of (Eirinaki & Vazirgiannis, 2003):

- *Association rule mining:* A technique used for finding frequent patterns, associations and correlations between sets of items. In the Web personalization domain, this method may indicate correlations between pages not directly connected and reveal previously unknown associations between groups of users with same interests.
- *Clustering:* a method used for grouping together items that have similar characteristics. In our case items may either be users (that demonstrate similar online behavior) or pages (that are similarity utilized by users).

- *Classification:* A process that assigns data items to one of several predefined classes. Classes usually represent different user profiles.
- *Sequential pattern discovery:* An extension to the association rule mining technique, used for revealing patterns of co-occurrence, thus incorporating the notion of time sequence. A pattern in this case may be a Web page or a set of pages accessed immediately after another set of pages.

C. Recommendation Phase

The aim of a recommender system is to determine which Web pages are more likely to be accessed by the user in the future [12]. In this phase active user's navigation history is compared with the discovered Navigation patterns in order to recommend a new page or pages to the user in real time. Generally not all the items in the active session path are taken into account while making a recommendation. A very earlier page that the user visited is less likely to affect the next page since users generally make the decision about what to click by the most recent pages. Therefore the concept of window count is introduced. Window count parameter '*n*' defines the maximum number of previous page visits to be used while recommending a new page.

IV. PROPOSED METHODOLOGY

This paper proposes a new novel approach to perform web intelligence. The entire process is parted into two phase such as online phase and offline phase. The subsequent fig. 3 gives the general framework of the proposed method.

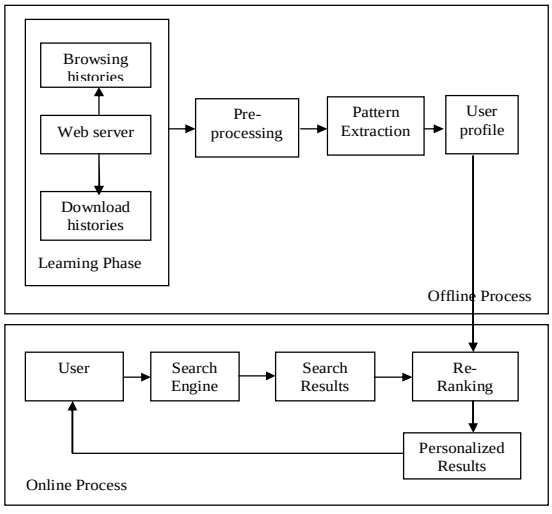

Fig. 3 The framework of web personalization

The user profile is generated by collecting data from various sources and by performing necessary transformation and pre-processing. Personalized and useful patterns are extracted in the off line process. During online phase, the current active session of the user are taken into account, and based on the patterns generated from the uses previous experience, the personalized information will be provided to the user.

A. Learning Phase or profile generation

In our investigation the effectiveness of personalized search is based upon the generation of user profile. The user profile has information that can distinguish one user from a multitude of other users. Profiles normally include topics of interests but may also include topics of disinterest by taking into account relevant and non-relevant documents [26].

1) Creation of user profile: There are two different ways to collect users' interest.

- Implicit feedback
- Explicit feedback

Our system uses the implicit feedback approach to decipher the interests of a user. The information sources currently used are:

1. The documents on the user's machine are taken as being of interest to him/her. The assumption is that if a user keeps a document on his/her machine there is a strong possibility that the user is interested in those documents.
2. The browsing history is another source of information. Although not all web pages browsed are of interest to the user but pages that are frequently visited and those on which more time is spent can be taken as useful documents for the user's interest.
3. The bookmarked pages are definitely interesting for a user.
4. The download history can be used as a starting point for deciding favorite directories for download. This would be more useful in case of those users who organize files on their machines properly.

We believe that this approach has several advantages over previous work in which user profiles were built from user browsing hierarchies [18].

2) User profile representation: A profile constitutes the user's long-term information desires. There are two main distinct types of user profiles [27]:

a) Content-based profile: the user profile is depicted similar to a query, as a vector of terms.

b) Collaborative profile: this approach is based on the rating patterns of similar users. It is assumed that individuals with similar rating patterns seem to like the same kind of information - "like minded people". Hence, a collaborative profile may be expressed as a list of similar users.

3) Vector Space Model:

In our approach content based profile generation is used. In this model the documents are represented as feature vectors where features are the keywords extracted from a given set of the documents. If a term is more frequent in a document it is expected to be given more weight than a less frequent word as it represents the document better. In the term frequency model the value of a feature is term frequency divided by the total number of words in the document [29]. The term frequency tf_{ij} of a term t_i in a document d_j is defined as:

$$tf_{ij} = \frac{number\ of\ occurences\ of\ t_i\ in\ d_j}{sum\ of\ number\ of\ occurences\ of\ all\ terms\ in\ d_j} \quad (1)$$

The inverse document frequency of a term ti in a set of documents D is defined as:

$$IDF_i = \log\left(\frac{number\ of\ documents\ in\ D}{number\ of\ documents\ in\ D\ containing\ t_i}\right) \quad (2)$$

A TF-IDF score w_{ij} is the product of tf_{ij} and idf_i. The weight is calculated using the equation (3):

$$w_{ij} = tf_{ij} \times IDF_i \quad (3)$$

B. Pre - Processing

Pre-processing involves interpreting raw user data and selecting which data to use so that is makes better sense for the rest of the system. The input data can be of different types (pdf, doc, ppt, html etc). They are first converted to text and then preprocessed. The preprocessing step involves stop-word removal and stemming. These are then converted to feature vectors where the features are the terms in the documents after the preprocessing step.

Filtered user data from this process is then the input for initial pattern extraction for new users. If the user is not new this filtered data can be combined with existing user data for discovery of new user patterns.

C. Pattern Extraction

The unique behavior of web user has to be analyzed. Clustering technique is used to find common patterns, group similar objects, or to organize them in hierarchies. The grouping of similar objects should be such that members within a group are closer to each other than to members of a different group.

Evolutionary particle swarm optimization based clustering EPSO-Clustering [28] is based on the idea of the generation based evolution of the swarm. The swarm evolves through different intermediate generations to reach a final generation. Particles are initialized in the first generation and after each generation the swarm evolves to a stronger swarm by consuming the weaker particles of that generation by the stronger ones. The stronger the particle is, the greater its chance of survival to the next generation. Stronger particles make mature and stable generations. The strongest generation reached with an optimal number of clusters and lowest intra-cluster distance represent an optimal solution of the problem. The pseudo code of the process is given in Algorithm 1.

Algorithm 1: Evolutionary Particle Swarm Clustering

1. Initialization of the particles
   - i. Initialize $V_i(t)$, $X_i(t)$, V_{max}, φ, q_1 and q_2
   - ii. Initialize swarm size, generation
   - iii. Initialize particles to input data
2. Iterate generation
   - i. Iterate swarm
     - a. Find won data vectors
     - b. Update velocity and position
   - ii. Evaluate the strength of the swarm
     - a. Iterate generation
     - b. Consume the weaker particles
     - c. Recalculated positions
3. Exit on number of generation exceed or stopping criteria fulfill

Where, $V_i(t)$ is the current velocity, $V_i(t+1)$ is the new velocity, w is the inertia weight, q_1 and q_2 are the vectors which weigh the cognitive and social components and $r1$ and $r2$ are two randomly generated numbers ranging from 0 to 1. The range for the velocities of the particles is from $-V_{max}$ to V_{max}.

After each iteration the swarm adjusts the positions of all particles by associating the nearest data vectors calculated using Euclidian distance equation 4, recalculates the attributes of the particles, organizing itself according to the new data vectors won by each particle.

$$d(x_n, z_i) = \|x_n - z_i\| \quad (4)$$

D. Recommendation Phase

The user query has been processes based on the user profile and the personalized results have been suggested to the use. It is also known as retrieval process. During web search the results by the search engine are converted to feature vectors using the same pre-processing techniques that were used for the user profile. Thus each search result is represented by a feature vector. These feature vectors are then passed to Similarity Scorer which assigns them scores based on their similarity to interest vectors. Each result is assigned a score equal to the maximum of the similarity scores with each interest vector. The results along with their scores are passed to Re-Ranker which sorts the results based on the scores assigned and the modifies the ordering that is ultimately presented to the user.

V. RESULTS AND EVALUATION

To assess the performance of the approach we tested the algorithm on the NASA web log file and analyzed the logs containing one day of HTTP requests after passing the log through all the preprocessing steps. For our experiment we have considered 100 data vectors, which have to be categories under 3 clusters. The results are depicted in the following table1.

TABLE I

sEPSO AND K-MEANS – CLUSTERING COMPARISON

EPSO			K- Means		
Clusters nos.	**No of Data vectors**	**Avg. Distance from centroid**	**Clusters nos.**	**No of Data vectors**	**Avg. Distance from centroid**
1	89	33.1459	1	89	33.15978
2	9	87.1016	2	9	88.73784
3	2	123..5667	3	2	123.5658
Fitness (sum of distance)		243.8142			245.46342
Mean intra-cluster distance		81.2712			81.82114

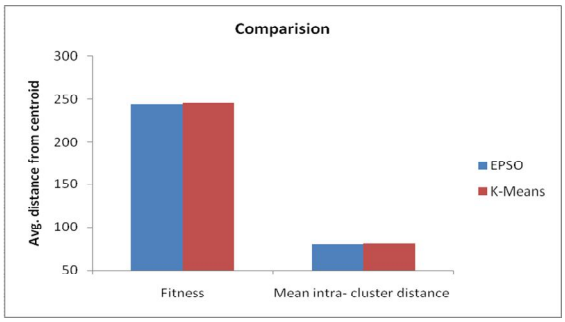

Fig. 4 Comparitive graph for EPSO and K-Means clustering

For comparison purpose the intra distance within the cluster was taken. EPSO algorithm has been compared with the K-Mean cluster algorithm. Compared to the K-means the average distance has been reduced such that members within a group are closer to each other than to members of a different group. Thus the swarm was uniformly distributed along the input data space. Our proposed method can also be evaluated based on the following factors.

Precision: It is the ratio of the number of recommended pages actually viewed by the user to the total number of recommended pages.

$$precision = \frac{No\,of\ recommended\ pages\ viewed}{Total\ no\ of\ recommended\ pages} \quad (5)$$

Coverage: It is the ratio of the number of recommended pages actually viewed by the user to the number of remaining pages in the user transaction leaving the active session pages.

$$coverage = \frac{No\,of\ recommended\ pages viewed}{No\,of\ remaining\ pages} \quad (6)$$

F1- measure: It is the combination of both precision and recall.

$$F1-measure = \frac{(2 * precision * recall)}{precision + recall} \quad (7)$$

VI. CONCLUSION

Web is growing rapidly, but on the other hand the user's capability to access web content remains constant. Currently, Web personalization is the most promising approach to alleviate this problem and to provide users with tailored experiences. We built a system that creates user profiles based on implicitly collected information: User browsing histories and download histories thereby improving their performance by addressing the individual needs and preferences of each user, increasing satisfaction of user. Thus achieving 100% personalization is not possible. But we can reduce the searching time to mine massive amount of data by shifting internet from search based to find based.

REFERENCES

[1] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD: Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, Vol. 2, Issue 1 – page 1, July 2000.

[2] Sanghamitra Bandyopadhyay . Sankar K.Pal, Classification and Learning with Genetic algorithms – Application in Bioinformatics and Web Intelligence, ISSN – 1619-7127, Springer Berlin Heidelberg – 2007 (pg: 242 – 256).

[3] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[4] P. Ravi Kumar and Ashutosh Kumar Singh, Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of Applied Sciences 7 (6): 840-845, 2010, ISSN 1546-9239.

[5] Srivastava J, Desikan P and V Kumar, Web Mining - Concepts, Applications & Research Direction, in 2002 Conference.

[6] Rochester Institute of Technology, Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data, MS Project Report, Rochester Institute of Technology, 2008.

[7] Han, J., Kamber, M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

[8] Manoj Swami, Manasi Kulkarni, Understanding Web Personlization with Web Usage Mining and its Application : Recommender System, International Journal of Emerging Techonolgy and Advanced Engineering, vol. 3,Issue 5, May 2013, pp.726-730.

[9] P. Markellou, Maria Rigou, Spiros S., Mining for Web Personalization,

[10] Honghua Dai, Bamshad Mobasher, Integrating Semantic Knowledge with Web Usage Mining for Personalization.

[11] C. Ramesh, Dr. K. V. Chalapati Rao, Dr. A. Goverdhan, A Semantically Enriched Web Usage Based Recommendation Model. International Journal of Computer Science and Information Technology (IJCSIT) Vol 3, No 5, Oct 2011.

[12] Daniar Asanov, Algorithms and Methods in Recommender Systems.

[13] Emmanouil Vozalis, K.G. Margaritis, Analysis of Recommender Systems' Algorithms. Parallel and Distributed Processing Laboratory.

[14] A.C.M. Fong, B. Zhou, Jie Tang, Guan Y. Hong, Generation of Personalized Ontology Based on Consumer Emotion and Behavior Analysis. IEEE Transactions on Affective Computing, Vol 3, No 2, April-June 2012.

[15] Nizar R. Mabroukeh, C. I. Ezeife, Ontology-based Web Recommendation from Tags. ICDE Workshop 2011. 2011 IEEE.

[16] A.C.M. Fong, B. Zhou, Jie Tang, Guan Y. Hong, Web Content Recommender System based on Consumer Behavior Modeling. IEEE Transactions on Consumer Electronics, Vol. 57, No. 2, May 2011

[17] Micro Speretta, Susan Gauch, Personalized Search Based on User Serach Histories, Proceesings of IEEE/WIC/ACM International Conference on Web Intelligence, 2005.

[18] J. Chaffee, S.Gauch, Personal Ontologies for Web Navigation. In proceedings of the 9th International Conference on Information and Knowledge Management, pp 227-234, 2000.

[19] Honghua Dai, Bamshad Mobasher, A road map to more web personalization: Integrating Domain Knowledge with web usage mining.

[20] Gabriella Pasi, Issue in personalizing information retrieval, IEEE Intelligent Informatics Bulletin, Vol. 11, No.1, 2010.

[21] Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul de Bra, Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems, Computers & Education, Elsevier, 53 (2009) 828–840.

[22] Zhong, Ning, Liu Yao, Jiming, Yao, Y.Y. Ohsuga, S. (2000), Web Intelligence (WI), Web Intelligence, Computer Software and Applications Conference, 2000. COMPSAC 2000. The 24th Annual International, p. 469, doi:10.1109/CMPSAC.2000.884768, ISBN 0-7695-0792-1

[23] Dingqi Yang, Daqing Zhang, Zhiyong Yu, Zhu Wang, A Sentiment-Enhanced Personalized Location Recommendation System, 24th ACM Conference on Hypertext and Social Media, May 2013.

[24] Tricia Rambharose, Alexander Nikov, Computational intelligence-based personalization of interactive web systems, WSEAS Transactions on Information Science and Applications, ISSN: 1790-0832, Issue 4, Volume 7, April 2010.

[25] Supiya Ujjin and Peter J. Bentley, Particle Swarm Optimization Recommender System.

[26] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. pages 54-89, 2007

[27] Tsvi Kuflik and Peretz Shoval, Generation of User Profiles for Information Filtering – Research Agenda, SIGIR, ACM, 2007
[28] Shafiq Alam, Gillian Dobbie, Patricia Riddle, An Evolutionary Particle Swarm Optimization Algorithm for Data Clustering, IEEE Swarm Intelligence Symposium, 2008.
[29] B. Everitt, "Cluster Analysis," 2nd Edition, Halsted Press, New York, 1980.
[30] Sarabjot Singh Anand and Bamshad Mobasher, Intelligent Techniques for Web Personalization, LNAI 3169, pp. 1–36, Springer-Verlag Berlin Heidelberg 2005