Knowledge Oriented Personalized Search Engine: A Step towards Wisdom Web

Aarti Singh
Assoc. Prof.,MMICT & BM
Maharishi Markendeshwar University,
Mullana,Haryana, India

Basim Alhadidi
Assoc.Prof., Deptt. of Comp. Sc.
AlBalqa' Applied University,
Salt, Jordan

ABSTRACT

WWW has undergone three generations from information web to social web to semantic web. It has started its journey towards the fourth generation which expects wisdom from the web and so termed as the wisdom web. In present era, where computers and internet has become inseparable parts of our life, user wants the Web to sense their requirements and interests and serve the contents accordingly. Search engines play major role in information extraction and delivery and present models of search engines are still struggling for providing personalized information to the users. This work is an extension of authors earlier published article on next generation of WWW where the idea for change in search engine model was coined. In this work author presents knowledge oriented personalized search engine framework which can provide personalized contents to its users. This framework provides a direction for the next generation of WWW and contributes towards wisdom web.

General Terms

Web Personalization, Web Mining, Agent Technology, Search Engine.

Keywords

Wisdom Web, Agent technology, Web Personalization, Web Semantics, Web Mining

1. INTRODUCTION

In 2001 Tim Berner Lee coined the idea of Semantic Web [9] where machines could understand the meaning of information displayed on the web pages and return context based information to the user. His vision gave a new direction to WWW [6], and helped converting unrelated, scattered data sources into systematic pool of information. From last one decade researchers had been working in different directions of web technologies for the implementation of semantic web. Now, that we have achieved semantic web, the users and research community is striving for personalized information delivery. WWW is on a journey of wisdom and maturity, where it can sense what user wants and serve relevant contents on its own. Table 1 given below throws light on the journey of WWW since its inception.

TABLE I. ROADMAP OF THE JOURNEY OF WWW

Generation	Time Span	Feature	Purpose
Ist Generation	1990-2002	Business oriented	Purpose was information exchange among users. End user could only consume the information provided but could not contribute in it. Web site & Web browsers were the major components.
IInd Generation	2003	User Oriented	WWW became integral part of common man's life. Social networking, blogs and Wiki's gave ability to the user to share their interests, ideas and participate in information generation, faster than ever before. Search engines came in existence and made information searching and dissemination easier. Web became more like an online community.
IIIrd Generation	Idea was coined in 2001, under development and use.	Context based Information Oriented	With information added on the web exponentially, users suffered from Information overload problem and became interested in context based information retrieval, rather than keyword based information. Search engines are becoming more sophisticated. Semantics of information is paid much attention while developing web

			contents.
IVth Generation	Expected to be achieved in next decade	Knowledge Oriented	After context based information retrieval, user is interested in customized contents specially meant to cater one's needs and interests. Researchers and Developers are focusing on web personalization techniques to fulfill this desire.

This work is an extension of an earlier work [20] by the author in which it was highlighted that the emergence of wisdom web is contributed by research efforts in many directions such as web technologies, agent technology, web mining and search engines which facilitate information extraction.

Figure 1 given below illustrates the process of emergence of Wisdom Web due to fusion of various technologies as suggested in [20].

The above figure illustrates the role of search engines, intelligent agents, agent based crawlers, application of web mining techniques on web server logs and emergence of knowledge through this process which is further used in making useful recommendations for the users.

We currently experience and work with this kind of web, although it is still far from personalized web, since the techniques applied in mining web data and results produced are general and are not meant for a specific individual, thus the results may or may not be fruitful. Transformation of WWW to Wisdom Web requires change in the way of web access and therefore requires new architecture. This work provides a new model of web personalization search engine, which is capable of providing knowledge, based recommendation to an individual and thus can contribute towards Wisdom Web.

Rest of the paper is organized as: Section 2 contains review of relevant literature highlighting contributions of various tools and techniques in present web specifically focusing on role of agent technology and advances in web mining domain. Section 3 provides details of the proposed framework. Section 4 concludes the work and highlights future scope.

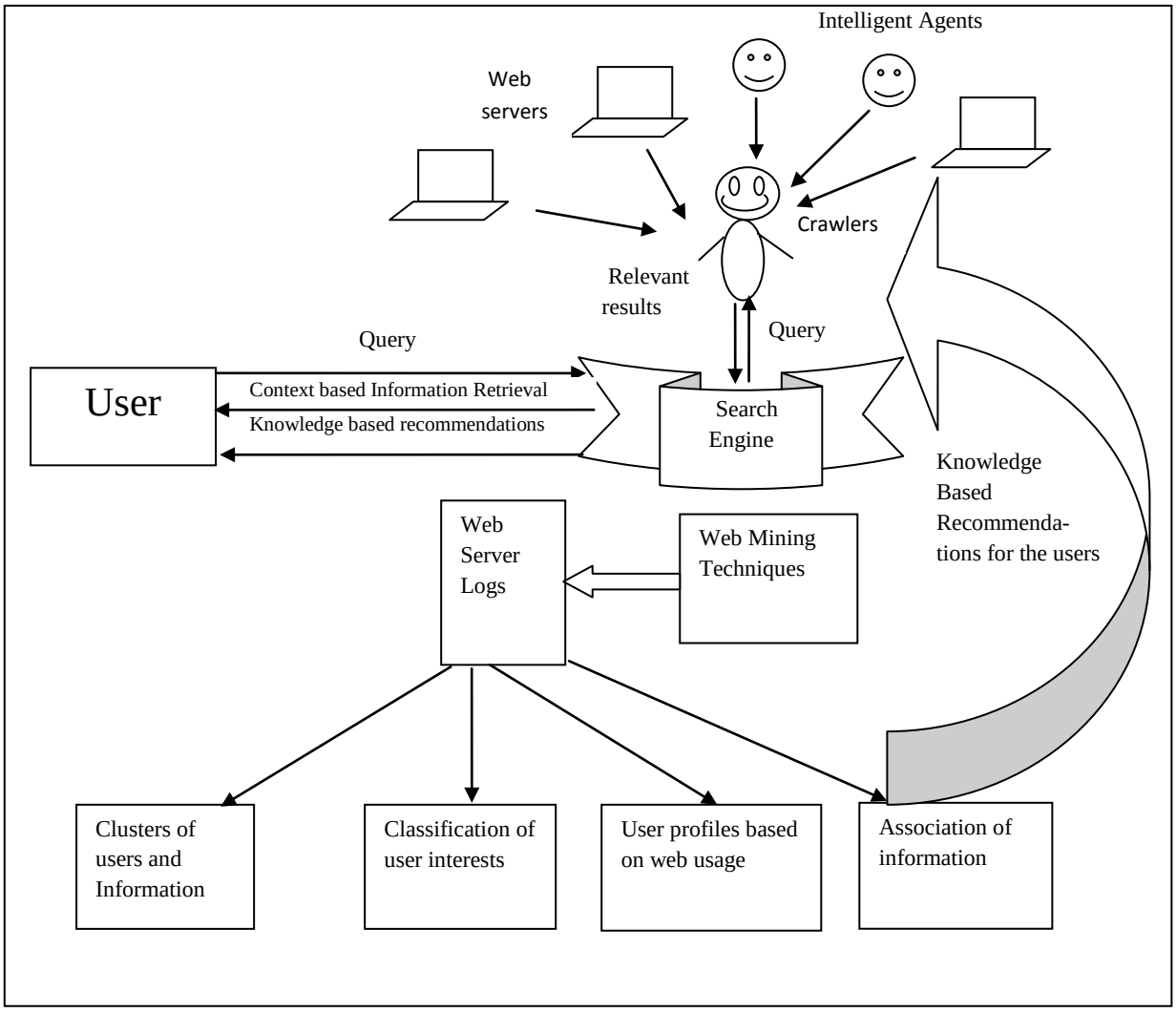

Figure 1: High Level View of Emerging Wisdom Web [20]

2. LITERATURE REVIEW

Zhong et. al in [23] elaborated the outline of web intelligence along with various technologies contributing to it. They highlighted that intelligent web requires extension in the knowledge and intelligence of artificial assistants i.e. the software agents. Zhan et. al in [22] introduced a multi-agent cooperation module for web mining which act as knowledge crawler, extraction machine, generalization machine and analysis machine for improving efficiency of web mining process. This framework provides method to choose best site to hop by using weights in ant theory, however formulation of integrity of different steps in mining process is left as future work. Li et.al in [10] proposed *embedding web mining ability in the crawler itself*. However the process of mining of downloaded web pages and RSS feeds is not paid much attention and not elaborated in detail. Fu and shih in [5] have proposed to mine web usage data on client side instead of traditional server side web usage mining. Their framework involves *use of intelligent agent for making personalized web page recommendations*. However implementation of prototype system is still under progress.

Shah et.al in [17] highlighted that existing search engines doesn't cater to the interests, knowledge level and literacy rate of the person seeking learning material from the web. Thus user spends more time on searching and filtering of contents suitable for the need, as compared to the time devoted in learning the concepts. Their work proposed an agent based intelligent e-learning model which aims to provide personalized search engine result page (SERP). However implementation and analysis of the framework is still left for future. Mobasher et. al in [11] elaborated the role of web usage mining in providing personalized contents to users by creating user profiles. Abraham et. al in [1] proposed an ant clustering algorithm to discover web usage patterns and linear genetic programming approach to analyze the visitor trends.

Singh A. in [19] proposed *an agent based framework for mining semantic web contents employing clustering techniques*. The work aimed at providing context based knowledge oriented results to the users. Joshi et. al in [7] have elaborated robust and possibilistic clustering and implemented an intelligent middleware to enable adaptive web access in bandwidth and resource constrained scenarios such as mobile systems. However they emphasized that web mining demands the development of new techniques for robust fuzzy clustering. Berkhin P. in [2] surveyed various clustering techniques available for data mining. Pierrakos et. al in [14] provided a survey of web usage mining for the web personalization perspective. Their work highlighted that web users are more interested in shopping from a site that offers personalization and that web personalization could be expanded by automating the adaptation of web based services for the users. Kumar et. al in [15] proposed a context-aware focused web search systems that considers various context features and returns relevant search results to the user. Authors remain silent about implementation of the framework.

From the literature review it is clear that users are demanding for personalized web services and contents. Research community is also making efforts for realizing the vision of personalized web [4,14]. Also it is evident that semantic web, context based searching in crawlers and web mining techniques specifically usage mining are playing important role in overall web personalization research. Literature survey [3,17,19,21,22] highlights that research fraternity have realized importance of agent technology in web based scenario and employed them to provide solutions at various phases. Analysis of various aspects of the reviewed literature provided motivation for proposing new model of search engine usage which will provide knowledge based personalized contents to the user. The details of the proposed framework are discussed in the upcoming section.

3. PROPOSED WORK

Present work focuses on the desire to have a knowledge oriented web where the user is provided with the information which suits individual's interests. Although present research in web usage mining allows us to create user's profiles [11] based on the web contents accessed, time of accessing a particular web page, hyperlinks visited etc. but these observation lead to profile creation of an unknown individual , thus don't lead to actual personalization desired. Most of the time users access internet, they make use of search engines to retrieve useful information, to access website of an organization whose address is unknown to them, to search for new movies, songs, videos etc. Every time the user accesses search engine, that individual is treated as a new user. The web log data analyzed by various web mining tools and techniques is used to provide recommendation of web contents for all the users searching for a particular kind of data. For example, if an internet user accesses websites of international journals in computer science and it is observed earlier that an individual accessing journals is also interested in information related to Conferences and seminars and in publishing research papers, then links related to conference and seminars, and for publishing papers in computer science are also presented to the user, believing that user might feel interested in that information. However it might not be true all the time and might irritate the user getting unnecessary links. Comparatively, information provided to the user in his/her mail account (in the form of advertisements on the side bars) is more relevant since there it is provided based on the web contents accessed by an individual. This difference provided the motivation to propose a new architecture of *Knowledge Oriented Personalized Search Engine (KOPSE)*, which can analyze and record information accessed by an individual and will be able to provide relevant contents based on the knowledge of past usage behavior. Figure 2 given below provides the high level view of *KOPSE. Through this work the aim is not to redefine the basic working of a search engine, but to add new module responsible for providing personalized and knowledge oriented web access. Therefore there is no need of elaborating the working of a search engine (refer to [13] for detailed working of crawlers in search engine).*

KOPSE makes use of intelligent agents for embedding personalization component in existing search engines. It mainly comprises of Profile Manager Agent, Usage Mining Agent and Query Responder Agent. Every user interested in personalized web access, will have to create an account in KOPSE through normal sign up procedure which will create a username and password for an individual, to be used in all future interactions with the search engine. The ecology of the agents involved in KOPSE is as follows:

- ***Profile Manager Agent (PMA)****:* is responsible for managing the list of registered users and their login details. Additionally it keeps a priority based list of user interests. Initially the user will be asked to provide priority based list of interests at the time of sign up and later this list will be revised based on actual usage pattern of an individual.

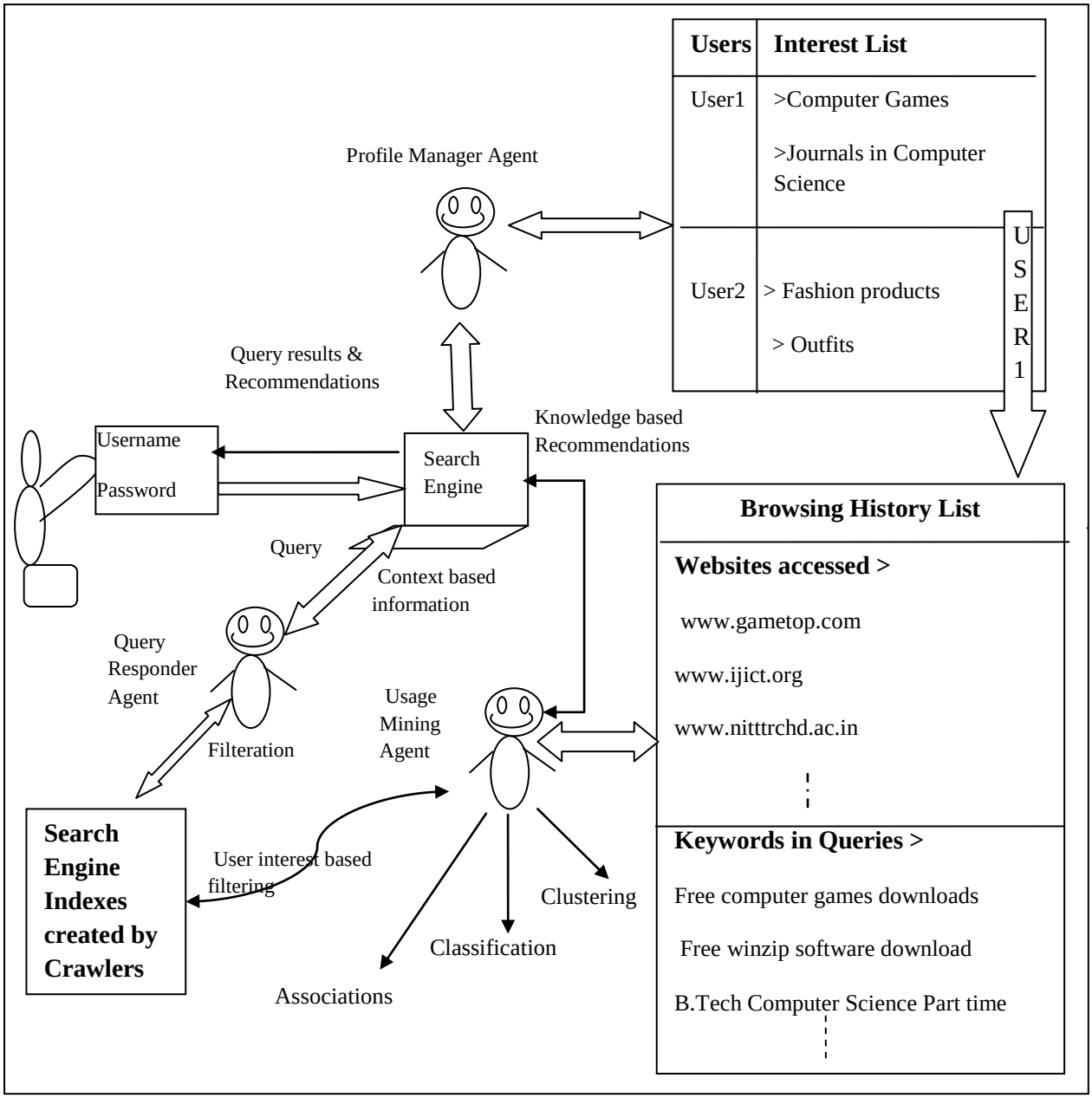

Figure 2:. Framework for Knowledge Oriented Personalized Search Engine (KOPSE)

For example if a user initially provides his interests as fashion products and clothing with priority value 1 & 2 but later on searches the web for cooking recipes or latest movies then the interest list will be extended with categories cooking recipes & new movies at priority value 3 & 4. If user has given priority 1 to fashion products but later accesses clothing websites most of the time, priority of clothing will be increased i.e. clothing will be assigned priority 1 and fashion product as 2. This applies to all categories. Whenever the same category is visited continuously for thresholds value (say 5 times in this case) its priority is increased to next higher value. Same applies to categories not visited for a threshold number of times, its priority is decreased to next lower level. Priority of interest domains varies from 1 to 10 and there may be any number of items at same priority. Priority simply means that the user is interested in getting updates in domains of interest in that sequence. This way there will be constant monitoring of the interests of the user by PMA and any change in user interest will be recorded and taken care of while providing contents.

- ***Usage Mining Agent (UMA):*** is responsible for mining the browsing history list of an individual and extracting useful knowledge about user interests. One usage mining agent is dedicated to every registered user and draws knowledge about that individual's interests and preferences. For this purpose a browsing history list is maintained for every individual recording access history for predefined number or days or up to predefined storage level. UMA applies various mining techniques on the browsing history list periodically and generates associations, classifications and clusters of web usage.

- ***Query Responder Agent (QRA)****:* is responsible for accepting the query entered by the user and perform context based search from the indexes created by the crawlers, in the traditional form. It receives the user query and creates a result list based on keywords and context of query. While QRA responds query in hand, UMA also works simultaneously, and creates a list of items from the index which might be of interest to the user; this list is based on the browsing history list maintained by it and knowledge drawn from association, classification and clustering techniques applied over this list. It then displays the updates on the already visited contents and websites and makes recommendation for the new contents which might be of interest to the user.

Thus using KOPSE whenever a user logs in, whether that individual searches something or not, KOPSE will suggest the contents of interest. This way user will not be required to search for updates in areas of his/her interests rather it will be provided to him just by logging in KOPSE.

3.1 Work Flow

The flow of work in KOPSE is illustrated in figure 3 given below. Step by step explanation is as follows:

1. User signs up in KOPSE to create a username and password. This information is entered in the list maintained by Profile Manager Agent. User interest domains are asked and maintained in the user interest list for future usage.

2. On successive logins user provides his username and password which is passed to PMA for verification.

3. PMA accesses the user profile and interest list and verifies the user, if verified user session starts.

4. User enters a query in KOPSE for accessing information.

5. Every query received by KOPSE interface is passed to QRA for extracting relevant information from the indexes. At the same time every query is scanned by PMA to analyze whether there is some new term apart from those already mentioned in the interest list, if so, new term are appended in the end of interest list. PMA keeps count of the times user searches/accesses terms specified in the interest list to maintain the priority of the list.

6. QRA finds the relevant results corresponding to the query through the traditional methods of information extraction and provides the results to the user.

7. All information accessed through KOPSE i.e. all the terms searched for or websites visited are maintained in the browsing history list dedicated for the user. UMA keeps an eye on every query provided by the user and the information accessed from the response generated by QRA. UMA records this information in browsing history list maintained for the user. and later mines it to draw useful knowledge about user interests.

8. Once the user session is over, UMA mines the browsing history list, classifies user interests, forms associations and applies clustering techniques to draw useful knowledge about the user interests. When user accesses the web using KOPSE for the first time, then browsing history list is populated and from the next session onwards UMA starts making personalized recommendation on interesting web contents.

9. When user accesses KOPSE for the second time, on user verification PMA passes the priority based interest list for that user to UMA which picks interest fields one by one, filters the index to find relevant and updated information on that field and provides it to the user as part of knowledge based recommendation. If user feels interested, he may access those contents, if not user can proceed with normal query based accessing method. Then the recommendations on his interests will be displayed in the side bars. For the entire session, UMA will go on providing recommendations on the contents related to his interest fields. One set of recommendations will be displayed for a fixed time (say 5 minutes), if the user doesn't access anything from the set, next set will be displayed. However if user accesses something from a set if will be displayed for one more time slot.

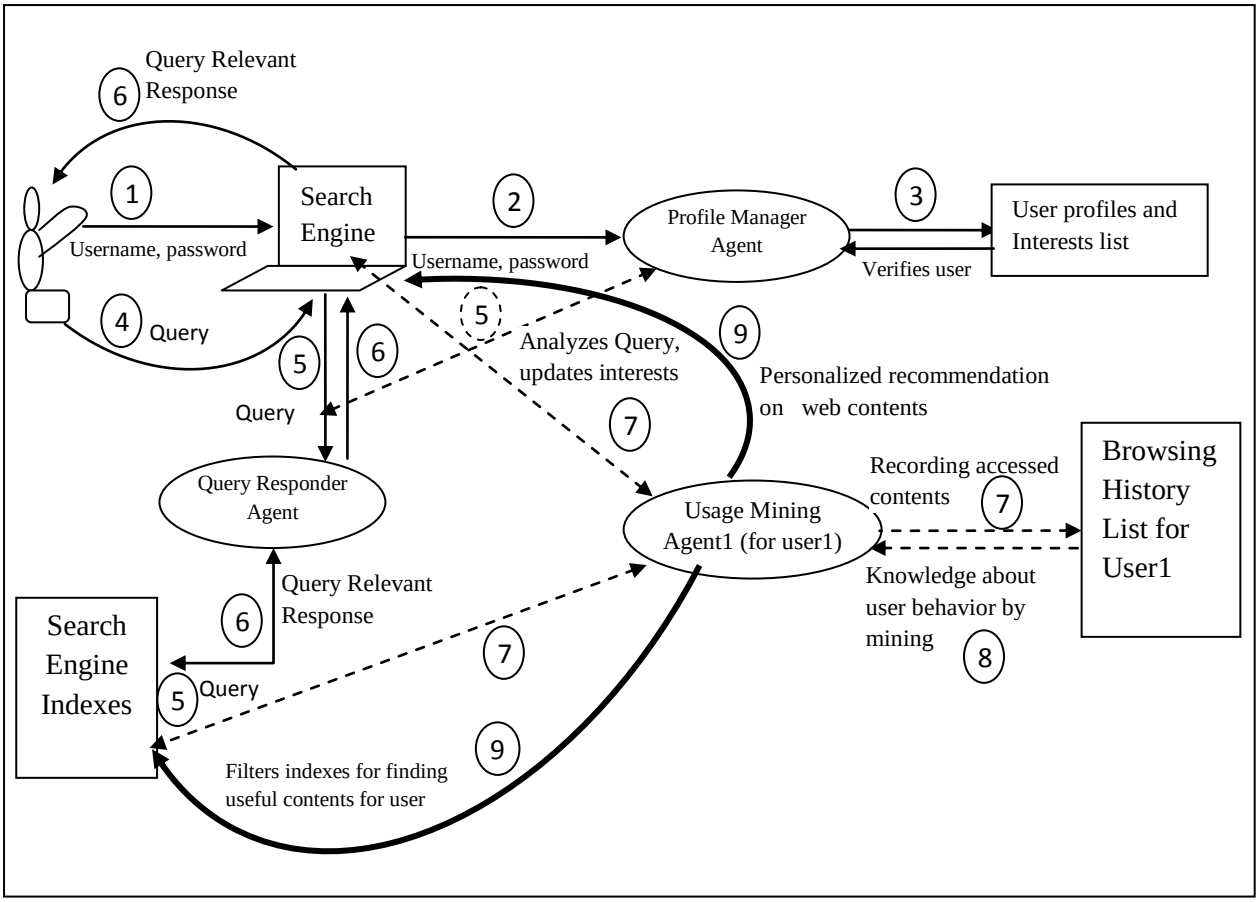

Figure 3:. Flow Diagram of KOPSE

Through this process the user may know the updates in his/her interest fields just by visiting KOPSE, without providing any query. Since UMA will filter web contents based on his/her priority and will serve contents.

Algorithms for various agents involved in KOPSE are provided in the upcoming section

.

3.2 Algorithms

Figure 4(a)-4(c) given below provide the algorithm for PMA(), QRA( ) and UMA( ) respectively.

```
Profile Manager Agent ( )

Input: user_profile;

Output: user_verification_result, interest_list;

 { On(new user profile) activate;

   { Create new user profile and interest list;

   Accept user interests and put them in list;  }

  On (existing user profile)

   {Match profile information from the list;

   If (verified) send priority list of user interests to UMA; }

 For every query or web page accessed observe key terms   and maintain count;

If ((item_priority ≠ 1) && (no of accesses)> 5)

  Increase priority to next higher level;

   If item not accessed in five consequent sessions reduce priority by one;

  }
```

Figure 4(a): Algorithm for Profile Manager Agent

```
Query Responder Agent ( )

Input: user query;

Output: list of relevant results;

 On (new query) activate;

 {Shortlist the index based on key terms in query and the context based knowledge;

   Provide relevant results back to the user;

}
```

Figure 4(b): Algorithm for Query Responder Agent

```
Usage Mining Agent ( )
Input: user profile, user query;
Output: list of recommended pages;
{     On (user profile)
   {  Receive list of interests for the user;
   Start filtering index based on user interests.
   Provide knowledge based recommendations of web contents to the user ;
   }
On (query)
 { Record every query in the browsing history list;
   Record all the web page accessed in a session in the history list;
}
  On (end_of_session)
 { Apply mining techniques on browsing history list;
    Draw useful knowledge about user interests and behavior;
  }
}
```

Figure 4(c): Algorithm for Usage Mining Agent

4. CONCLUSIONS AND FUTURE WORK

This work has explored various technologies contributing towards emergence of Wisdom Web. Literature review highlighted the fact that users are suffering from information overload problem and are striving for personalized web contents as per their requirements and interests. Keeping this fact in view this work proposed a new module for search engines, which can be embedded in the existing search engine architectures and can provide knowledge oriented personalized contents for the users. This work is in implementation phase, author is working with MobileC and Ch editors for implementing the suggested framework. Implementation and evaluation of this framework is aim of the future work. However, once implemented this module can meet much awaited dreams of the users.

5. REFERENCES

[1] Abraham A. and Ramos V., 'Web Usage Mining using Artificial Ant Colony Clustering and Genetic Programming'. Published in Proceedings of IEEE Congress on Evolutionary Computation, Australia, 1384-1391.

[2] Berkhin P., 'A Survey of Clustering Data Mining Techniques'. Technical report, Accrue Software, San Jose, CA, 2002.

[3] Chauhan N., Sharma A.K., 'Design of an Agent Based Context Driven Focused Crawler'. Published in BVICAM'S International Journal of Information Technology, 2008, pp. 61-68.

[4] Eirinaki M. & Vazirgiannis M., 'Web Mining for Web Personalization'. Published in ACM Transactions on Internet Technology, Vol. 3, No. 1, February 2003, pp. 1-27.

[5] Fu Y., Shih M.Y., 'A Framework for Personal Web Usage Mining'. International Conference on Internet Computing, Las Vegas, NV, pp. 595-600, 2002.

[6] 'History of WWW', Source: http://www.w3.org/People/Berners-Lee/ShortHistory.html

[7] Joshi A. and Krishnapuram R., 'Robust Fuzzy Clustering Methods to Support Web Mining'. Published in Proceedings of Workshop in Data Mining and Knowledge Discovery, SIGMOD, 1998, pp. 15-1–15-8.

[8] Kosala R. & Blockeel H., 'Web Mining Research: A Survey'. Published in ACM SIGKDD, Vol. 2, Issue 1, July 2000.

[9] Lee T Berners, Hendler J and Lassila C, 'The Semantic Web'. Published in The Scientific Americans, Vol. 5, No. 1, p. 36, 2001.

[10] Li Q.Cheng, Lin S. and Dong Z.H., 'Research of Web Information Mining by using Crawler Techniques'. Published in Proceedings of the 2008 IEEE International Conference on Information and Automation, June 20-23, 2008, Zhangjiajie, China.

[11] Mobasher B., Cooley R. and Srivastava J., 'Automatic Personalization Based on Web Usage Mining'. Published in Communications of the ACM, Vol. 43, No. 8, August 2000, pp. 142-151.

[12] Pant G., Srinivasan P. and Menczer F., 'Topical Web Crawlers: Evaluating Adaptive Algorithms'. Published in ACM Transactions on Internet Technology, Vol. 4, Issue 4, pp. 378-419, November 2004.

[13] Pant, G., Srinivasan, P., and Menczer, F.. 'Crawling the Web'. Published in Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Edited by M. Levene and A. Poulovassilis, pages 153-178. Springer-Verlag, 2004.

[14] Pierrakos D., Paliouras G., Papatheodorou C. and Spyropoulos D. C., 'Web Usage Mining as a Tool for Personalization: A Survey'. Published in User Modeling and User Adapted Interaction, Vol. 13, No. 4, pp. 311-372, 2003.

[15] S. Kumar and N. Chauhan, 'A Context Model for Focused Web Search'. Published in International Journal of Computers and Technology, Vol. 2, No. 3, June 2012.

[16] Shadbolt N, Hall W and Lee T Berners, 'The Semantic Web Revisited'. Published in IEEE Intelligent Systems, Vol. 21, No. 3, pp. 96-101, 2006.

[17] Shah A. and Jain S., 'An Agent Based Personalized Intelligent e-learning'. Published in International Journal of Computer Applications, Vol. 20, No. 3, April 2011, pp.40-45.

[18] Sharma K., Shrivastava G. & Kumar V., 'Web Mining: Today and Tommorrow'. In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[19] Singh A., 'Agent Based Framework for Semantic Web Content Mining'. Published in International Journal of Advancements in Technology, Vol. 3, No. 2, April 2012, pp. 108-113.

[20] Singh A., 'Wisdom Web: The WWW Generation Next'. Published in International Journal of Advancements in Technology, Vol.3, No. 3, July 2012, pp. 123-126.

[21] Singh A., Juneja D. and Sharma A.K., 'An Extensive Analysis of Implementation Issues in Semantic Web'. Published in ICFAI University Journal of Information Technology, Vol. V, No. 4, December 2009, pp. 67-74.

[22] Zhan L. and Zhijing L., 'Web Mining Based on Multi-Agents'. Published in Proceedings of the fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), 2003.

[23] Zhong N., Liu J. and Yao Y., 'In search of the Wisdom Web'. Published in IEEE journal of Computer, Vol. 35, issue 11, November 2002, pp. 27-31.