Ontology Based Navigation Pattern Mining For Efficient Web Usage

Jothi Venkateswaran C. [1], Sudhamathy G. [2]

[1] Department of Computer Science, Presidency College (Autonomous), Chennai – 600 005, India
[2] Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore – 641 043, India
[1] cjpcmcahead@gmail.com
[2] sudhamathi10@hotmail.com

Abstract— Users of the web have their own areas of interest. Given the tremendous growth of the web, it is very difficult to redirect the users to their page of interest. Web usage mining techniques can be applied to study the users navigational behaviours based on their previous visit data. These user navigational patterns can be extracted and used for web personalization or web site reorganization recommendations. Web usage mining techniques do not use the semantic knowledge of the web site for such navigational pattern discovery. But, if ontology is applied along with web usage techniques, it can improve the quality of the detected patterns. This research work aims to design a framework that integrates semantic knowledge with web usage mining process that generates the refined website ontology that recommends personalization of web. As the web pages are seen as ontology individuals, the user navigational behaviours over a certain period are considered as the user expected ontology refinement. The user profiles and the web site ontology are compared and the variation between the two is proposed as the new refined web site ontology. The web site ontology has been semi-automatically built and evolves through the adaptation procedure. The result of implementation of this recommendation system indicates that integrating semantic information and page access patterns yield more accurate recommendations.

Keyword-Web Usage Mining, Semantic Web, Ontology, Web Page Recommendation

I. INTRODUCTION

World Wide Web is a large and dynamic information source, which is structurally complex and ever growing. Hence it is very difficult to fully exploit it. Semantic Web addresses this problem by giving information a well-defined meaning, better enabling computers and people to work in co-operation [BHL01]. The primary challenge of a web personalization system is to *learn the user interests* and based on that knowledge to provide users with the information that they need, without expecting from them to ask for it explicitly [MMTM06; SHY04].

In this paper, it has been explored, how it is possible to succinctly learn the user interests based on their navigational behaviour and to investigate the insights of this knowledge into web site customization. The present work addresses the adaptation of the web ontology, based on web usage data. A framework that employs web usage mining as well as web ontology methodologies has been presented. The proposed framework adapts the web in order to assist the users in their browsing tasks [LS06]. The *web usage mining techniques are used to find the user navigational paths and at the same time the web site ontology are constructed* using a tool and by consulting the web master.

The user profiles and the web site ontology are compared and the variation between the two is proposed as the new refined web site ontology. The web site ontology has been semi-automatically built and evolves through the adaptation procedure. For empirical studies, a real website access data, collected from the usage of the website of a university, has been used.

II. RELATED WORK

The web users browsing pattern depends on their interests, knowledge and needs. It has been proven that factors, such as visual experience and site attractiveness, logicality of navigation organization, placement of objects, colour schema and page loading time [BE01; BC00; LE99] also affect web surfing. Hence, in order to have an enhanced understanding of the needs and interests of the web users, the semantic information has to be integrated with the web usage data. Moreover, these web users needs and interests changes over a period of time and hence there is need for ability to constantly adapt with the changing needs and interests by using the ontology information of the web site.

The rapid expansion of the Internet has provided the possibility to explore users' navigational patterns and interaction with web-based systems. This allows computing of recommendations for users, either by simple pages, or constructing new web site topologies for user groups [MT05]. Usage patterns discovered through *web usage mining are effective in capturing page-to-page and user-to-user relationships* and similarities at the level of user sessions. Without the benefit of deeper domain knowledge, such patterns provide little insight into the

underlying reasons for which such items or users are grouped together. This can lead to a number of important shortcomings in personalization systems based on web usage mining.

The *previous approaches, however, are incapable of capturing more complex relationships at a deeper semantic level* based on different types of attributes associated with structured objects. Hence there is a need for creating a general framework for using domain ontology to automatically characterize usage profiles and to use this framework in the context of Web personalization by using the full semantic power of the underlying ontology. Nizar [NC09] proposes the integration of semantic information drawn from a web application's domain knowledge into all phases of the web usage mining process (pre-processing, pattern discovery, and recommendation/prediction). *The goal is to have an intelligent semantics-aware web usage mining framework.* The Internet consists of web sites that employ different kinds of structures as the backbone of their build-up. However, users are browsing the web according to its content, regardless of the structure. In the research work by Tarmo [TA07], they discuss the possibilities of applying ontology in exploring the web sites' structures and usage for producing viewing recommendations for the visitors.

A special log system for capturing access data is introduced as well as techniques applied for data mining. In this work, the ontology of user profiles is constructed by exploiting the user locality model. To improve the discovered patterns' quality, [AVMD07], presents a technique with the help of metadata about the content that they imagine is stored in domain ontology. This technique includes a dedicated pattern space constructed on top of the ontology, navigation primitives, mining procedure and recommendation methods. Providing users with assistance in their web navigation can help keep them in a web site, or even attract more visitors. This has always been a popular subject, especially in the e-commerce domain. Several systems have been developed towards this direction. *WebWatcher* [TDT97] has suggested links that may interest a user, based on the online behaviour of other users. Each user is asked, upon entering the site, what kind of information he is seeking. Before he departs, he is asked whether he has found what he was looking for. His navigation paths are used to deduce suggestions for future visitors that seek the same content. These suggestions are visualized by highlighting existing hyperlinks.

The *Avanti project* [FKN96] tries to predict the user's final objective as well as his next step. A model for the user is built, based partly on information the user provides about him. His interests are also extracted from his navigation paths. Visitors are provided with direct links to pages that are probably the ones they are looking for. In addition, hyperlinks that lead to pages of potential interest to each visitor are highlighted. A drawback of both the *WebWatcher* and the *Avanti system* is that they require the active participation of the users in the adaptation process, by asking them to provide information about themselves. On the other hand, the *Footprints* [WM99] system relies entirely on the navigation paths of the users. The system does not perform user identification. All navigation paths are recorded and the most frequent ones are presented to the visitor, in the form of maps or trails. Html pages also display next to each link the percentage of people who have followed it. Nevertheless, as in the *Web-Watcher* and the *Avanti systems*, no adaptation of the site's structure is performed.

Perkowitz [PE00], have presented a conceptual framework for adaptive web sites. They have focused on the semi-automatic creation of index pages, based on discovering clusters of pages. They assume that if a large number of visitors frequently visit a set of pages, this provides strong evidence that these pages are related. They have developed two cluster mining algorithms, *PageGather* and *IndexFinder*. The first one relies on a statistical approach to discover candidate link sets, while the second is a conceptual cluster mining algorithm, as it finds link sets that are conceptually coherent. They have also performed experiments on three web sites by placing the automatically generated pages online and observing the user response.

Bamshad [BRJ00], defined personalization as "*any action that tailors the web experience to a particular user, or a set of users*". On the other hand, according to Perkowitz [PE00], transformation is "*improving the site's structure based on interactions with all visitors*". The advantage of this approach is that it does not require user identification, which cannot be safely performed from usage data, unless the user contributes in an explicit or implicit way [MM03]. Nevertheless, most users are reluctant to give away personal information. Coenen [CSVW00], distinguishes between *tactical* and *strategic* adaptations in their framework for self-adaptive web sites. They call tactical the adaptations that can be performed in real time, without the webmaster's approval, since they do not affect the overall site structure. On the other hand, strategic adaptations are the ones that "*go against or conflict with the original beliefs of the site, and consequently have an important influence on the original site-structure*". Coenen [CSVW00], suggest that such modifications should be performed offline, with the approval of the webmaster.

The majority of the existing approaches in web adaptation lack in a crucial factor: they do not address the semantic aspect of the web. The ontological perspective is overlooked and the researchers' attention is drawn mainly by the site topology. Even though the improvement of the site topology is unquestionably significant, it is not possible to disregard the fact that users browse a site mainly for its content. Consequently, the content classification structure should also be adaptive through the evolution of the site ontology. The innovative concepts of the *Semantic Web* and *Web Usage Mining* are most suitable regions for applying such adaptation

methodologies, targeting to the direct benefit of the end users. The proposed framework defines the adaptation process as absolutely *transparent* to the user, requiring no active participation from them. In addition, the adaptations of the framework perform *web transformation*, instead of focusing on personalization tasks.

III. Architecture of the Innovative Approach

The *figure 1* presents an architecture implementing the theoretical principles of the proposed framework. The input for this process is the web server logs and it does not require the user involvement. The web logs are pre-processed and the web session database is created that consists of the user session records of the form, session id, list of web pages visited in their order. The session database is given as input to the web site ontology building tool *protégé* and the corresponding web site ontology is built. From the session database, the navigational paths that are frequently traversed by the various users of the web site are also constructed from the session database. The methodology for constructing the navigational paths from the user sessions are detailed in the section Navigational Path Construction and User Profiles Extraction.

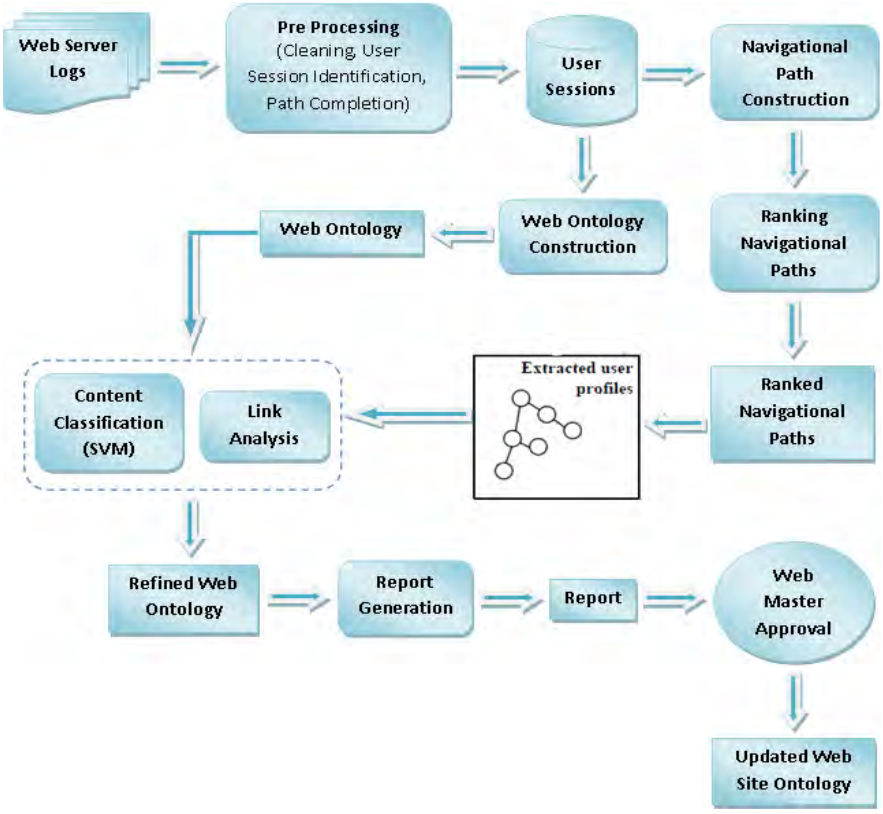

Fig. 1. Architecture for Refining Web Site Ontology

The page and user sessions are ranked using the inverse time weighing algorithm, which is defined as below. Based on the ranking of the user sessions and the web pages, the navigational paths are ranked. The resultant ranked navigational paths are used for user profiles extraction. The extracted user profiles are classified in relation to certain features of their pages. More specifically, two classification criteria have been used: *Link Analysis* and *Content Classification*.

Rank for a Page = p * sum (Hits for a page in a day / Number of days from current date) **(1)**

Where p takes the value between 0 and 1 and that is prefixed

Rank for a User Session = p * sum (Hits by a user session in a day / Number of days from current date) **(2)**

Where p takes the value between 0 and 1 and that is prefixed

Link Analysis refers to the connection that the pages of each user profile have according to the site structure. The *key factor* is whether the pages contained in a user profile are directly linked to each other or not. User profiles of unlinked pages might suggest the insertion of shortcut links between these pages, in order to achieve shorter navigational paths. From the user profiles of the linked pages changes in the appearance of existing links can be extracted. For example if an index page and some of its links comprise one or more user profiles, then by highlighting these links in the index page, first time visitors would be able to navigate the site easier.

The *second classification criterion* refers to the *content of the pages* contained in each user profile. The pages of the user profiles are classified using Support Vector Machines, in order to discover new associations between the concepts of the site ontology. More particularly, if a user profile includes pages belonging to concepts that were not previously linked; the ontology should then be modified to reflect the relevance these concepts have, according to the preferences of the users.

Based on the *link analysis and content classification*, a *report* containing proposals for the improvement of the web site is generated. This report contains proposals for the insertion of shortcut links from source pages to target pages that are frequently accessed together but are currently not linked. It also contains proposals for the change of appearance of popular hyperlink. Furthermore, the report contains proposals for the evolution of the web site ontology. After the proposed modifications have been revised by the webmaster, they can be applied to the web site. The site topology is then refined through the insertion of new shortcut links, as well as changes in the appearance of the existing ones. The web site ontology is also refined accordingly.

IV. NAVIGATIONAL PATH CONSTRUCTION AND USER PROFILES EXTRACTION

The user sessions are extracted one by one from the session database and they are processed as shown in the *figure 2*. The first step in this process is *path sequencing* for each user session. That is the pages accessed in a session are arranged sequentially to form a navigational path like structure. The next step is *path minimization* by removing the redundant pages. That is if the subsequent pages accessed in a user session are same, then count the page only once in the path sequence. If some of the user session paths just have single page accessed repeatedly or once, then remove such paths. Now it is required to calculate the sliding window size *w* for the selected web site. The sliding window size is same as that of the maximum menu depth of the web site. In the figure 2, it has been considered the sample *sliding window size as 3*, which is also found to be ideal sliding window size for any user preferred web site.

Next the *cleaned path sequence has been split* so that each subset of the path has just the same number of pages as that of the sliding window size. From these extracted sub sequences of the user navigational paths, it is required to remove the cyclic paths, which are the paths that starts and ends with the same page. These are the final set of navigational paths for the selected user session. These paths are now ranked based on the ranking found in the web pages and user sessions. Finally, the *ranked navigational paths are obtained*, from which the user profile is extracted. The above process has to be repeated for all the user sessions in the database. These *extracted user profiles are then compared with the existing web ontology* and the required modifications are done and that will help the users to browse the web site effectively without any delay and confusion.

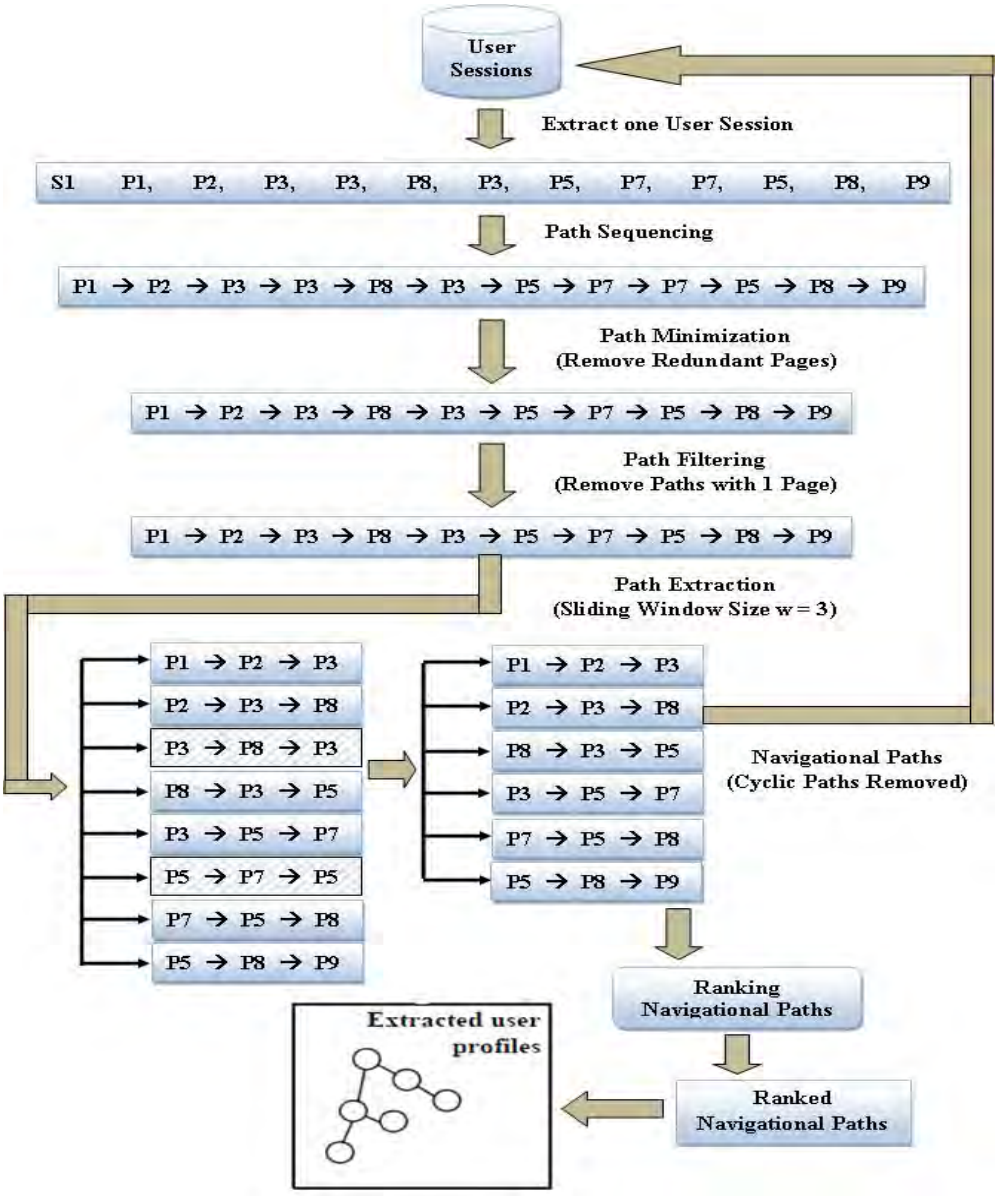

Fig. 2. Construction of Navigational Paths and User Profiles Extraction

V. RESULTS AND DISCUSSION

The proposed methodology has been automated using *C# .NET* and applied to the web logs of a university web site and the results have been observed. The ontology of the web site has been built by giving the user sessions as input to the *Protégé tool* and the pictorial representation of the same is given in the *figure 3*. The existing web site topology is represented pictorially in the *figure 4*. The topology of the web site has been refined through the insertion of the shortcut links between the pages that were not previously linked together. In addition the web site ontology has been modified in several ways based on the outcomes retrieved from the classified page sets.

The target pages have been found to be frequently visited after the source page. However the source page is not linked to the target pages, thus forcing the user to follow alternative paths in order to reach them. For instance '*Academic Calendar*' Page is accessed frequently after accessing the '*Time Table*' page under Examinations. With the existing web site topology this browsing took a longer path than that is proposed in the new topology. The web site ontology was modified in several ways, based on the outcomes retrieved from the classified pages. Based on these adaptations, the content organization of the site was altered to better satisfy the needs of its visitors. New associations have been discovered between concepts. These associations reflect the interests of the users as pages belonging to these concepts were frequently accessed together.

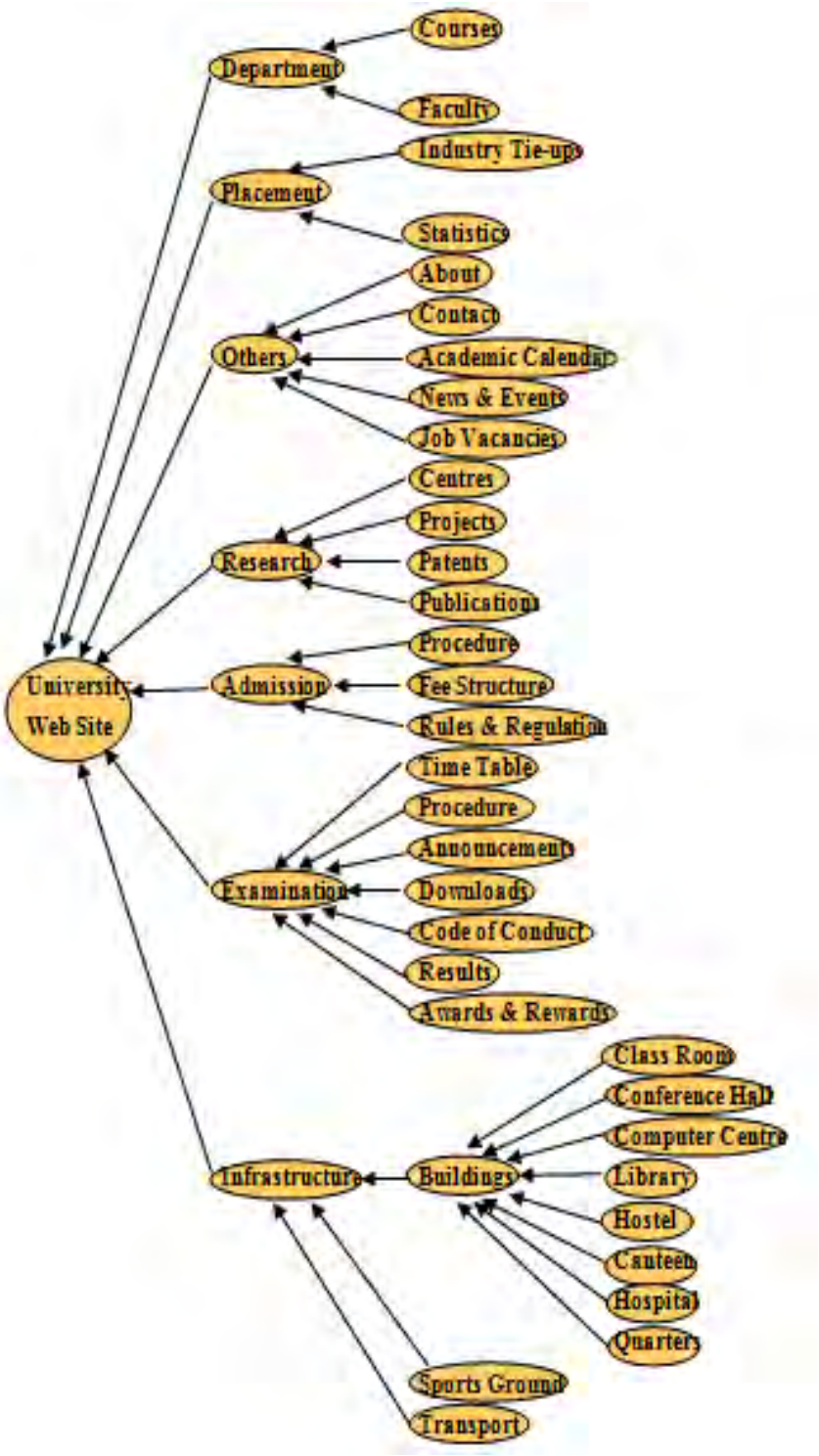

Fig. 3. Sample Web Site Ontology Before Applying the Method

The initial topology was based on the features existing in the web site irrespective of the user who will access the same. But in the revised topology the focus is on the user that is the staffs or the students. All the web pages related to students are grouped under *'Students'* and all the web pages that will be accessed by the staff are grouped under *'Staff'* as separate sections. Other general web pages are grouped under *'Home'*. The pages under *'Departments'* in the original topology are broken down as links under *'Staff'* and *'Students'* as per the frequent navigational paths discovered. Other pages under *'Academics'*, *'Admissions'*, *'Faculty'*, *'Placement'* and *'Research'* are taken under the broader grouping of *'Staff'* and *'Students'*.

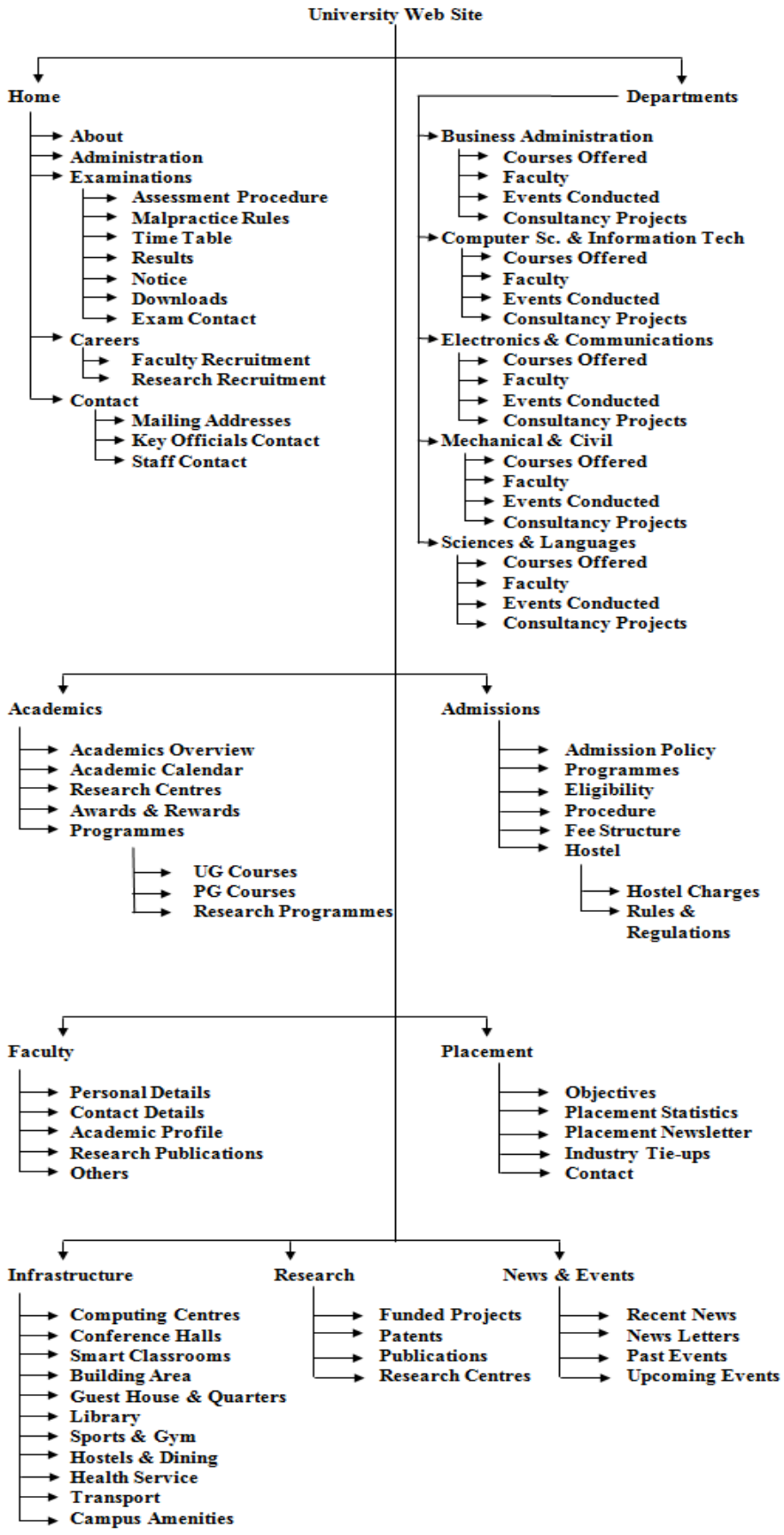

Fig. 4. Sample Web Site Topology Before Applying the Method

Thus reorganization of the concepts hierarchy was also performed. Further improvements included the creation of new categories such as *'Educational Buildings'* and *'Campus Amenities'* under *'Infrastructure'*, the removal of existing categories such as *'UG Courses'* and *'PG Courses'* under *'Academics'* as well as changes to the levels of hierarchy that the concept belongs to such as *'News'* and *'Events'* under *'Home'*. The high frequency with which this concept appeared in the page sets implies the significance that it has in the interests of the users. It would be thus appropriate to transfer this concept to the top level of the ontology.

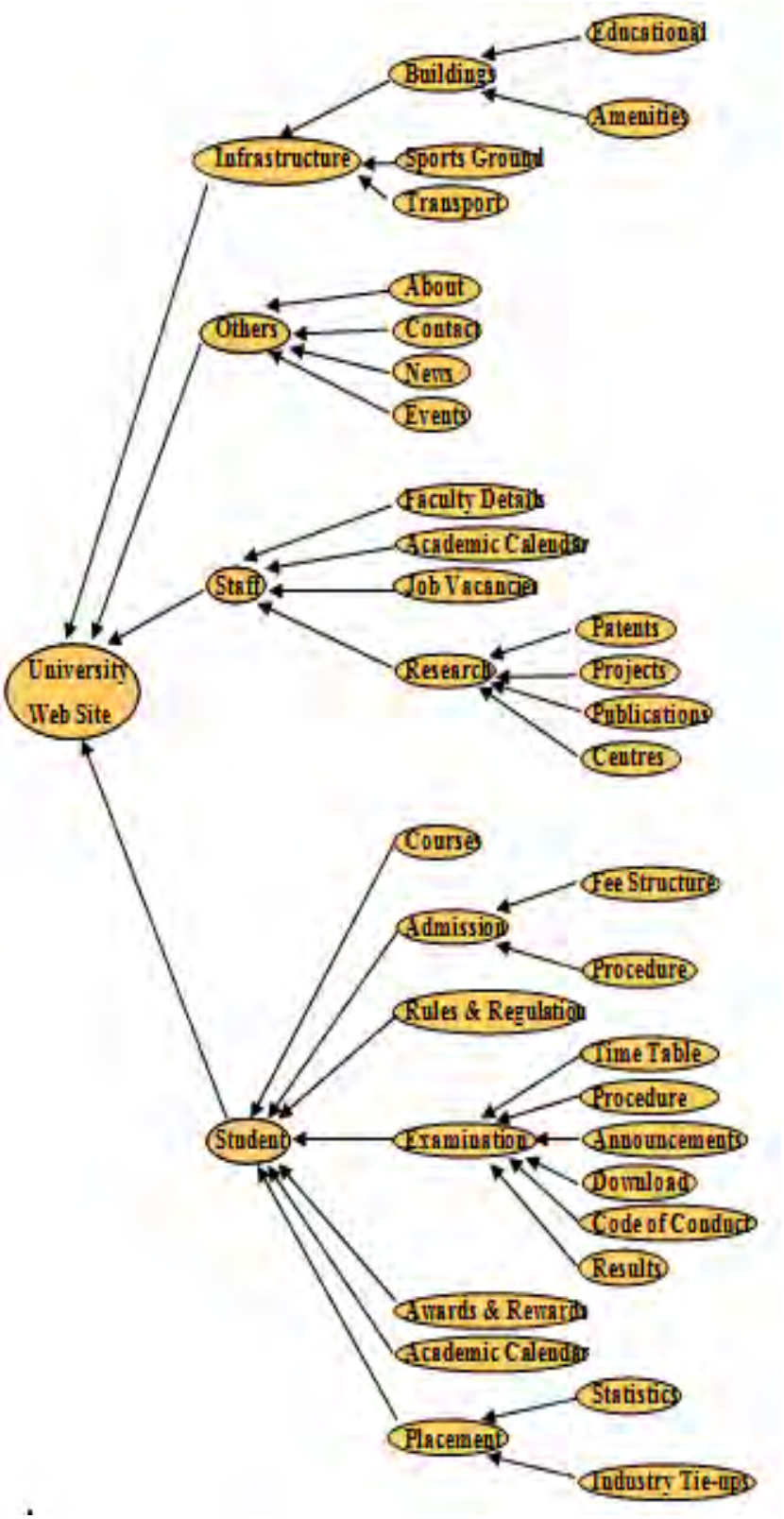

Fig. 5. Sample Web Site Ontology After Applying the Method

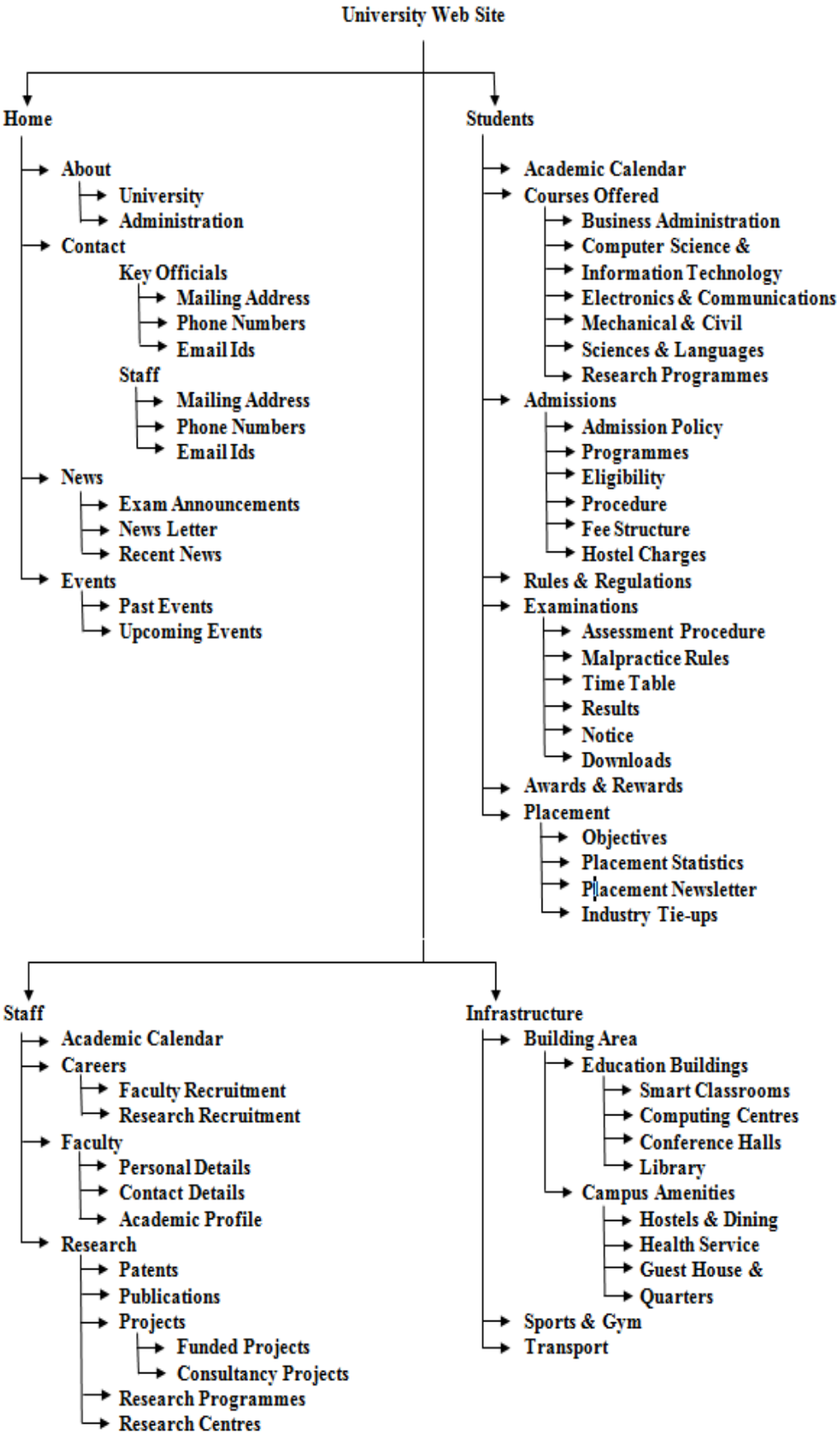

Fig. 6. Sample Web Site Topology After Applying the Method

The ontology of the site was extended to include multiple instances of concepts such as *'Academic Calendar'* under *'Students'* and *'Staff'* or multiple sub-concepts such as *'Funded Projects'* and *'Consultancy Projects'* under *'Research'*. The categorization of the web pages that was carried out, suggested that several pages belong to more than one concept. Moreover, in some cases, web pages and the corresponding concepts were categorized under different concepts than they previously were in the existing ontology. The site ontology should be therefore updated in order to reflect these facts. The resulting ontology and its corresponding topology have been shown in the *figure 5* and *figure 6*. Thus this proposed approach proves to be effective in providing web site modification recommendations that matches the user's preferences when domain ontology semantics is integrated with ranking of user navigational paths.

VI. SUMMARY

In this paper a web usage driven approach on the adaptation of the Semantic Web has been investigated. A framework has been introduced that enables adaptation of the web ontology to the needs and interests of web users. In addition, an architecture based on the principles of the framework has been presented. The proposed methodology uses the web logs of the users along with the semantic aspect of the web, in order to facilitate better web browsing. A University web site has been taken as the case study to analyse the impact of the proposed framework on the usability of the web. The ontology of the site has been refined in several ways. Apart from changes in specific web pages, enhancements of the whole formation of the site have been derived. Furthermore, useful knowledge has been acquired, regarding the overall usage of the site. The web pages that mostly interest the users were identified, leading to further improvements in their usability. Moreover, the regions of the web site that need more promotion have been revealed.

REFERENCES

[1] Robert, C., Bamshed, M., & Jaideep, S. (1997). "Web Mining: Information and Pattern Discovery on the World Wide Web", In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '97), (pp. 558-567).

[2] Robert, C., Bamshed, M., & Jaideep, S. (1999). "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems, (pp. 5–32).

[3] Zhang, H., & Liang, W. (2004). "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining", In Proceedings of the 5th World Congress on Intelligent Control and Automation.

[4] Yan, Li., Bo-Qin, F., & Qin-Jiao, M., (2008). "Research on path completion technique in web usage mining", In Proceedings of the International Symposium on Computer Science and Computational Technology, IEEE Xplore, Shanghai, (pp. 554-559).

[5] Renáta, I., & Sándor, J. (2007). "Analysis of Web User Identification Methods", World Academy of Science, Engineering and Technology, 34, (pp. 34-59).

[6] Spilipoulou, M., Mobasher, B., & Berendt, B. (2003). "A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", INFORMS Journal on Computing Spring, 15(2), (pp. 171-190).

[7] Jaideep, S., Robert, C., Mukund, D., Pang-Ning, T. (2000). "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations, 1(2), (pp. 12-23).

[8] Jose, B., & Mark, L. (2008). "Mining users' web navigation patterns and predicting their next step", NATO Secur. Sci. Ser. D-Inform. Commun. Secur., (pp. 45-55).

[9] Bamshad, M. (2004). "Web Usage Mining and Personalization", Practical Handbook of Internet Computing. Ed. Editor, CRC Press M.P. Singh., (pp. 1-37).

[10] Li, C. (2009). "Research on Web Session Clustering", Journal of Software, (pp. 460-468).

[11] Norwati, M., & Mehrdad, J. (2009). "Expectation maximization clustering algorithm for user modeling in web usage mining systems", Eur. J. Sci. Res., (pp. 467-476).

[12] Navin Kumar, T., Solanki, A.K., & Sanjay, T. (2010). "An Algorithmic Approach To Data Preprocessing in Web Usage Mining", International Journal of Information Technology and Knowledge Management, (pp. 279-283).

[13] Edmond, H.W., Michael, K.N., & Joshua, Z.H. (2004). "A Data warehousing and Data Mining Framework for Web Usage Management", Communications in Information and Systems, (pp. 301-324).

[14] Hallam-Baker, P., & Behlendorf, B. (1996). "Extended Log File Format, W3C Working Draft", WDlogfile-960323, URL: http://www.w3.org/TR/WD-logfile.html.

[15] Agosti, M., & Di Nunzio, G.M. (2007). "Web Log Mining: A Study of User Sessions", In Proceedings of 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, PersDL, Corfu, Greece, (pp. 70–74).

[16] Zaiane, O.R., Xin, M., & Han, J. (1998). "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", In Proceedings of Advances in Digital Libraries Conference, (ADL'98), Santa Barbara, CA, (pp. 19–29).

[17] Jansen, B.J., Spink, A., Blakely, C., & Koshman, S. (2007). "Defining a Session on Web Search Engines", Journal of the American Society for Information Science and Technology, 58(6), (pp. 862-871).

[18] Huaqiang, Z., Hongxia, G., & Han, X. (2010). "Research on Improving method of Preprocessing in web log mining", IEEE.

[19] Theint Theint, A. (2011). "Web log cleaning for mining of web usage patterns", IEEE.

[20] Sudheer Reddy, K., Partha Saradhi Varma, G., & Ramesh Babu, I. (2012). "Preprocessing the web server logs an illustrative approach for effective usage mining", ACM.

[21] Arvind Kumar, D., & Sunita, S. (2013). "A new approach for user identification in web usage mining Preprocessing", IOSR-JCE.

[22] Nithya, P., Sumathi, P. (2012). "Novel Pre-Processing Technique for web log mining by removing global noise and web robots." IEEE.

[23] Wasvand, C., Devale, P.R., Ravindra, M. (2014). "Survey on Data Preprocessing Method of Web Usage Mining", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), (pp. 3521-3524).