Personalization in the semantic web era: a glance ahead

P. Markellou[1,2], M. Rigou[1,2], S. Sirmakessis[2,3] & A. Tsakalidis[1,2]
[1]*Department of Computer Engineering and Informatics, University of Patras, Greece*
[2]*Research Academic Computer Technology Institute, Patras, Greece*
[3]*Department of Applied Informatics in Administration and Economics, Technological Educational Institution of Messolongi, Greece*

Abstract

The problem of information overload when browsing and searching the web becomes more and more crucial as the web keeps growing exponentially and personalization features as the most popular remedy for it. The evolution of personalization as we experience it in recent years has been dramatically influenced by web mining, a research area developing around three main axes: web content, web usage and web structure mining. Lately most research efforts have moved towards combining techniques from more than one domain to achieve extreme personalization in large and complicated web structures. The introduction of the *semantic web* brings into the picture an enticing dimension: the ability to semantically link various resources (documents, images, people, concepts, etc). Without semantic knowledge about the specific application domain, the identification of underlying properties, attributes and interconnections has often turned out to be (if not impossible) a time and cost demanding task. This paper explores the potential introduced for personalization by combining web content, usage and structure mining with the semantic web infrastructure and investigates how old problems can now be resolved or alleviated. Moreover, fields foreseen to produce new scientific results are identified.
Keywords: personalization, web content, structure and usage mining, semantic web, ontologies.


Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

1 Introduction

The Web has become a huge repository of information and keeps growing exponentially under no editorial control, while the human capability to find, read and understand content remains constant. Providing people with access to information is not the problem; the problem is that people with varying needs and preferences navigate through large Web structures, missing the goal of their inquiry [18]. Web personalization is one of the most promising approaches for alleviating this information overload, providing tailored Web experiences, as "*its objective is to provide users with what they want or need, without having to ask (or search) for it explicitly"* [21].

Initial attempts of implementing personalization were limited to *check-box personalization,* in which portals allowed the users to select the links they would like on their "personal" pages, but this has proved of limited use since it depends on the users knowing in advance the content of their interest. Moving towards more intelligent (or AI) approaches, *collaborative filtering* was deployed for implementing personalization based on knowledge about likes and dislikes of past users that are considered similar to the current one (using a certain similarity measure). Such techniques required users to input personal information about their interests, needs and/or preferences, but web users are not usually cooperative in revealing these types of data and soon researchers resorted to *observational personalization.* Observational personalization is based on the assumption that we can find clues about how to personalize information, services or products in records of user previous navigational behaviour [21]. The evolution of personalization as we experience it the recent years has been dramatically influenced by web mining, a scientific area defined as *"the use of data mining techniques for discovering and extracting information from Web documents and services"*. Eirinaki and Vazirgiannis provide a mining-oriented definition of web personalization in [7], according to which it is "*any action that adapts the information or services provided by a web site to the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the web site"*.

Web mining is distinguished as web content, structure or usage mining, depending on which part of the Web is mined [15]. It should be emphasized thought that the distinctions between the three main categories of web mining are not clear-cut [17]. Web content mining might utilize text and links and even the profiles that are either inferred or inputted by the users. User profiles are mostly used for the user modeling applications or personal assistants. The same is true for web structure mining; in addition to the link structures it may use information regarding links. Or, traversed links can be inferred from the documents that were requested during user sessions as recorded in the logs generated by the server. In the majority of cases, web applications base personalization on web usage mining, which undertakes the task of gathering and extracting all data required for constructing and maintaining user profiles based on the recorded behavior of each user. In practice, the three web mining categories can be used in isolation or combined in an application; especially in the case of web content and structure

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

mining since links may be considered as part of the content of a web document. In fact, such hybrid approaches deploying web usage mining as well as structure and/or content mining have been proposed for delivering more effective personalization. For some examples of combined applications see [3],[14]. This combination implies the advancing of web mining to a more abstract level. To achieve this abstraction, web data (usage, content, structure) are represented using another emerging model of representation, ontologies. This representation, closes the gap between semantic web and web mining areas, to create a fast-emerging research area, that of semantic web mining.

This paper explores the potential introduced for personalization by combining web content, usage and structure mining with the semantic web infrastructure and investigates how old problems can now be resolved or alleviated. More specifically, section 2 provides a brief introduction to the notion the semantic web and its enabling technologies and investigates the way web mining has been affected by the semantic web infrastructure. Section 3 reviews representative personalization applications that exploit the semantic web infrastructure when mining the traces web users leave behind them while navigating. Finally, section 4 concludes with some foreseen future directions for new research in the area.

2 Mining the semantic web

The semantic web is an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in better cooperation. The W3C Semantic Web Activity [28] in collaboration with a large number of researchers and industrial partners, is tasked with defining standards and technologies that allow data on the Web to be defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across applications. The notion of being able to semantically link various resources (documents, images, people, concepts, etc) is of primary importance. With this we can begin to move from the current web of simple hyperlinks to a more expressive semantically rich web, where we can incrementally add meaning and express a whole new set of relationships among resources (hasLocation, worksFor, isAthorOf, hasSubjectOf, dependsOn, etc), making explicit the particular contextual relationships that are implicit in the current web. This opens new doors for effective information integration, management and automated services.

Semantic Web Mining aims to combine two fast-developing research areas; the Semantic Web and Web Mining. Berendt et al. in [1] give an overview of where the two areas meet today, and sketch ways of how a closer integration could be profitable. The idea behind using the semantic web for generating personalized web experiences is to improve web mining by exploiting the new semantic structures. Two significant technologies towards that direction are the eXtensible Markup Language (XML) [9] and the Resource Description Framework (RDF) [25]. And while XML is simple, easy to use, and gives unrestricted freedom to add structure to web documents, it does not impose any underlying rules for expressing the content in a universally understandable way.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

RDF on the other hand, is based on strict rules for describing content, and therefore is easier to be "understood" by machines but imposes difficulty to users for annotating documents. Just the combination of XML and RDF proves insufficient since, even if there exists a "grammar" for expressing content using RDF, the same concept may be expressed using different identifiers, making interoperability across different systems difficult. Ontologies are the remedy to this problem. An ontology is defined as "a formal explicit specification of a shared conceptualization" [G93]. Tim Berners-Lee considers ontologies to be a critical part of the Semantic Web, since they provide a common vocabulary for solving such terminology problems.

2.1 Web content and structure mining

In the semantic web, content and structure are strongly intertwined and the distinction between content and structure mining vanishes. However, the distribution of the semantic annotations may provide additional implicit knowledge. Web structure mining benefits by incorporating into the proposed algorithms and methods the meta-information included in the hyperlinks and the text surrounding them. What is more, Web content mining is enhanced if Web pages are characterized using an abstract representation based on ontologies.

In the web personalization domain, a trivial task is the identification of similar web documents (product pages, teaching topics, similar search results, etc.) in order to be recommended to a user or a cluster of like-minded users. Web documents are mainly characterized by extracted keywords and by a rank that takes into account link structures [2] and similarity calculations between documents are based on binary matching. By using instead of keywords concepts and moreover concepts in an ontology, we can achieve more flexible document matching, handling both term specializations and generalizations. There exist several similarity measures for handling the more simple case of calculating the similarity between two given terms of the ontology. Richardson et al. [26] and Resnik ([24],[23]) propose different measure inside a taxonomy such as WordNet, and Lin [16] proposes a comparison between these measures and others, such as Wu and Palmer [29], Miller and Charles, and a novel similarity measure. [5][11] and [8] also propose the use of Wu and Palmer measure in the ontology context. Moreover, since documents in the semantic web are characterized using ontology terms, this knowledge can be used in order to form semantically coherent clusters based on a similarity measure between sets of weighted words. Traditional such techniques rely mostly on exact keyword matching, and do not take into account the fact that there may be some semantic proximity among keywords. Halkidi et al. in [11] propose a clustering scheme, based on a novel similarity measure that can be applied to sets of terms that are hierarchically related. A challenging task once the clusters have been found is their adequate labeling.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

2.2 Web usage mining

At the user session level the usage patterns discovered by web usage mining are effective in capturing item-to-item and user-to-user correlations. However, 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. It is possible to capture some of the site semantics by integrating keyword-based content-filtering approaches with collaborative filtering and usage mining techniques. These approaches, however, are incapable of capturing more complex relationships at a deeper semantic level based on the attributes associated with structured objects [20]. Incorporating semantics into usage mining and associating them with an ontology may provide all the required insight to justify the mined findings. Semantic web usage mining can for instance be performed on log files that record user behavior in terms of an ontology ([12], [27]). Semantic log files can then be mined to cluster for instance users with similar interests in order to provide personalized views on the ontology. The typical Web personalization process, which is based on the Web-site's logs, can be enhanced by taking into account the semantic proximity of the content [8]. This way, the produced recommendations can be enriched with content bearing similar semantics (that is assigned to the same semantic cluster). Moreover, since the logs incorporate semantic information, category-based rules can be extracted, i.e. association rules based on category terms instead of URIs.

3 Personalization on the semantic web

As have been mentioned before, the challenge for the next generation personalization and recommendation systems is the integration of semantic and ontological knowledge into the various parts of the web mining process. Indeed, when there is not enough usage data in order to extract useful patterns related to certain categories, or when the web site content changes and new pages are added but are not yet included in the web log then the usage-based personalization can be insufficient. The incorporation of information related to the content and/or the structure of the web site provides a way of overcoming such problems, thus improving the whole personalization process. An increasing number of researchers focus on the use of structured semantics and ontologies in the web site design and implementation. In the next paragraphs, we present research efforts that combine web mining techniques and semantic/ontological knowledge.

In [8] C-logs (concept logs) are introduced, which comprise a conceptual abstraction of the original web usage logs based to the web site's semantics. C-logs are used as input to the web usage mining process, resulting in a broader yet semantically focused set of recommendations. Specifically, web content is semantically characterized using a domain-specific taxonomy and a combination of IR techniques. The keywords that are extracted using these techniques are mapped to the categories of the taxonomy, resulting in a uniform and consistent vocabulary. After this process, every document falls under one or more

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

taxonomy categories. C-logs are created when each record in the processed web server log is updated to include the related set of concepts (categories).

Another work that uses the idea of semantic enrichment of web log files with ontology concepts is presented in [22]. The approach is based on an ontology underlying the web site. Using the ontology's taxonomy, it describes user actions at different levels of abstractions while using the ontology's concepts and relations, it captures the multitude of user interests expressed by a visit to one page. However, it doesn't cover the majority of the web sites that do not support semantics.

Other parts of the web mining process that can be incorporate semantic knowledge are the internal representation models. These models are used for constructing individual and aggregate (when working with groups of users) profiles and can be based on domain ontologies. Specifically, the work of Mobasher & Dai in [4] aims to automatically characterize usage and content profiles (defined in [19]) containing a set of structured web objects. Firstly, a usage profile representation through web usage mining process is created as a set of structured objects embedded in visited pages. Then a domain-level aggregate profile that characterizes a collection of similar users based on the common properties of objects in the domain ontology that were accessed by these users. The personalization decisions are based on these profiles and succeed to use the full semantic power of the underlying ontology.

Data mining algorithms and techniques may then be performed on these semantic web logs to extract knowledge about groups of users, users' preferences, rules, etc. This implies that they should be able to deal with complex semantic objects. In most cases they comprise extensions of well-known data mining algorithms such as clustering, classification, regression, association rules mining in order to take into consideration semantic and ontological knowledge [20]. These techniques are known as Relational Data Mining [1] or formally as Inductive Logic Programming – ILP [6] and they search for patterns that involve multiple relations in a relational database. In [22] the data analysis process is performed on a knowledge portal (SEAL – SEmantic portAL), exploiting its inherent RDF annotations.

A joint probabilistic latent semantic analysis (PLSA) framework to develop a unified model of web user behavior based on the usage and content data associated with the site is described in [13]. This model provides great flexibility and can be used for a variety of web mining and analysis tasks. It discovers and characterizes web user segments and produces dynamic and personalized recommendations based on these segments. The experiments show a more accurate representation of user behavior, resulting in turn in higher quality patterns that can be used effectively in web recommendations.

The output of personalization process should also incorporate semantic knowledge in order to produce intelligent adaptations in the content, the structure and the presentation of the web site. In [8] the recommendations that are produced are enriched with documents that fall under the same cluster as the ones that would otherwise be presented to the end user. In this work (SEWeP – Semantic Enhancement for Web Personalization), the notion of category-based

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

recommendations is also introduced, i.e. recommendations on general categories (terms of the taxonomy). The presented recommender system in [10] succeeds to recommend other items in the same class of products that match the user model by mapping products to an abstract layer of semantic features. Moreover, its main advantages are that it understands the customer tastes and recommends items across categories, explains the recommendations in terms of qualitative features, enhances the user experience and builds user's confidence to it.

4 Conclusion

In this work, we revise the new opportunities for personalization introduced by the incorporation of semantic web mining. Firstly, we described the semantic web notion and the underlying technologies. Then, we investigated the way web mining has been affected by the semantic web infrastructure and reviewed research efforts in the area. Since, the distinctions between the three axes of web mining are in many cases ambiguous the current trend is to combine techniques from one, two or all axes with ontologies creating a fast-emerging research area, Semantic Web Mining. Research in the domain of web personalization based on semantic web mining is promising and offers prospects for more effective personalization and recommendation systems. It is certain that the web will keep growing, even in a somewhat different form than how we know it today and thus the need for new methods for handling the large volume of information available in this universal framework will remain prevailing.

Acknowledgements

This work is partially supported by the EU/IST NEMIS project (Network of Excellence in Text Mining and its Applications in Official Statistics) IST-2001-37574.

References

[1] Berendt B., Hotho A. & Stumme G., Towards Semantic Web Mining. *Proc. of the ISWC 2002*, eds. I. Horrocks & J. Hendler, Springer-Verlag Berlin, LNCS 2342, pp. 264-278, 2002.

[2] Brin, S. & Page, L., The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, **30(1-7),** *Proc. of the 7th International World Wide Web Conference (WWW7)*, pp. 107-117, 1998.

[3] Chakrabarti S., Dom B., Gibson D., Kleinberg J., Kumar S., Raghavan P., Rajagopalan S. & Tomkins A., Mining the link structure of the world wide web. *IEEE Computer*, **32(8)**, pp. 60-67, 1999.

[4] Dai, H. & Mobasher, B., Using Ontologies to Discover Domain-Level Web Usage Profiles. *Proceedings of the 2nd Workshop on Semantic Web Mining, 6th European Conference on Principle and Practice of Knowledge*

Discovery in Databases (ECML-PKDD'02), Helsinki, Finland, pp. 61-82, 2002.

[5] Desmontils, E. & Jacquin, C., Indexing a Web Site with a Terminology Oriented Ontology. *The Emerging Semantic Web*, eds. I.F. Cruz, S. Decker, J. Euzenat & D.L. McGuinness, IOS Press, pp. 181-198, 2002.

[6] Dzeroski, S. & Lavrac, N., (eds). *Relational Data Mining*. Springer, 2001.

[7] Eirinaki, M. & Vazirgiannis, M., Web mining for Web personalization. *ACM Transactions on Internet Technology (TOIT)*, **3(1)**, pp. 1-27, 2003.

[8] Eirinaki, M., Vazirgiannis, M., & Varlamis, I., SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process. *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'03)*, Washington DC, pp. 99-108, 2003.

[9] Extensible Markup Language, http://www.w3.org/XML/

[10] Ghani, R. & Fano, A., Building Recommender Systems Using a Knowledge Base of Product Semantics. *Proceedings of the Workshop on Recommendation and Personalization in E-Commerce*, at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, Spain, 2002.

[11] Halkidi M., Nguyen, B., Varlamis, I. & Vazirgiannis M., THESUS: Organizing Web Document Collections Based on Link Semantics. *The VLDB Journal*, special issue on Semantic Web, **12(4)**, pp. 320-332, 2003.

[12] Hotho A., Maedche A., Staab S. & Studer R., SEAL-II - the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science (J.UCS)*, **7(7)**, pp. 566-590, 2001.

[13] Jin, X., Zhou, Y. & Mobasher, B., A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content. *Proceedings of AAAI Workshop on Semantic Web Personalization*, held in conjunction with the 19th National Conference on Artificial Intelligence - AAAI 2004, San Jose, California, 2004.

[14] Joachims T., Freitag D. & Mitchell T., Webwatcher: A tour guide for the world wide web. *Proc. of the Int. Joint Conference on Artificial Intelligence IJCAI-97*, pp. 770-777, 1997.

[15] Kosala R. & Blockeel, H., Web Mining Research: A Survey. *SIGKDD Explorations*, **2(1)**, pp. 1-15, 2000.

[16] Lin D., An Information-Theoretic Definition of Similarity. *Proc. of 15th Int. Conference on Machine Learning (ICML)*, Morgan Kaufmann, San Francisco, pp. 296-304, 1998.

[17] Markellos, K., Markellou, P., Rigou, M., & Sirmakessis, S., Web mining: Past, present and future. *Text Mining and its Applications, Results of the NEMIS Lanch Conference,* ed. S. Sirmakessis, Springer-Verlag Berlin Heidelberg, Studies in Fuzziness and Soft Computing, pp. 25-35, 2004.

[18] Markellou, P., Rigou, M. & Sirmakessis, S., Mining for Web Personalization (Chapter 2). *Web Mining: Applications and Techniques*, ed. A. Scime, Idea Group Publishing Inc, pp. 27-48, (in press) 2005.

[19] Mobasher, B., Dai, H., Luo, T., Sung, Y. & Zhu, J., Integrating Web Usage and Content Mining for More Effective Personalization. *Proc. of the International Conference on E-Commerce and Web Technologies (ECWeb2000),* Greenwich, UK, 2000.
[20] Mobasher, B., Web Usage Mining and Personalization. *Practical Handbook of Internet Computing*, ed. M. P. Singh, CRC Press, (in press) 2004.
[21] Mulvenna, M., Anand, S. & Bchner, A., Personalization on the Net using Web mining. *Communications of the ACM,* **43(8)**, pp. 122-125, 2000.
[22] Oberle, D., Berendt, B., Hotho, A. & Gonzalez, J., Conceptual User Tracking. *Proceedings of the 1st International Atlantic Web Intelligence Conference (AWIC)*, Madrid, Spain, pp. 150-164, 2003.
[23] Resnik P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research,* 11, pp. 95-130, 1999.
[24] Resnik P., Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI-95*, pp. 448-453, 1995.
[25] Resource Description Framework (RDF), http://www.w3.org/RDF/.
[26] Richardson, R., Smeaton, A. & Murphy, J., Using wordnet as a knowledge base for measuring semantic similarity between words. *Proc. of the AICS Conference,* Trinity College, Dublin, 1994.
[27] Spyns, P., Oberle, D., Volz, R., Zheng, J., Jarrar, M., Sure, Y., Studer, R. & Meersman, R., OntoWeb - A Semantic Web Community Portal. *Proc. of the Fourth International Conference on Practical Aspects of Knowledge Management (PAKM02)*, Springer-Verlag, pp. 189-200, 2002.
[28] W3C Semantic Web Activity, http://www.w3.org/2001/sw/.
[29] Wu Z. & Palmer M., Verb Semantics and Lexical Selection. *Proc. of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pp. 133-138, 1994.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)