International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.195-202
http://dx.doi.org/10.14257/ijdta.2016.9.12.18

A Survey on Ontology based Web Usage Mining

Vandana M. Patil[1], J. B. Patil[2]

[1]*Research Scholar, Department of Information Technology, R. C. Patel Institute of Technology, Shirpur (MS), India*
[2]*Professor, Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur (MS), India*
[1]*shraddhasam@rediffmail.com*, [2]*jbpatil@hotmail.com*

Abstract

The exponential increase in information, users and number of Websites on WWW has given rise to number of challenges. The most important challenge is the effective and systematic management of this massive Web data. For Web users, it is very difficult to access relevant information quickly and efficiently. And for Web site owners, it is very difficult to satisfy their users' information needs effectively. Web Usage Mining has been used to deal with aforesaid issues. The Web Usage Mining techniques are solely based on knowledge acquired through the analysis of the users' navigational behavior. Hence, quality of discovered patterns is low. Recent studies show that, semantically enriched Web Usage Mining enhances the quality of discovered patterns. The semantically enriched Web is called as Semantic Web, and this new form of Web Usage mining is called as Semantic Web Usage Mining. It is also called as Ontology based Web Usage Mining, as Ontologies act as backbone for conceptual description of semantic knowledge in Semantic Web. In this paper, we have presented brief overview of conventional Web usage mining and performed an extensive survey of research work done in ontology based web usage mining.

Keywords*: Web Mining; Web Usage Mining; Ontology based Web Usage Mining; Semantic Web Mining.*

1. Introduction

The explosive growth of Web data makes it difficult to manage Web data efficiently and systematically. Web mining has been used over the last few years to deal with the unusual nature of Web data and to improve the accessibility of Web data. But due to lack of background semantic knowledge, conventional Web usage mining techniques cannot penetrate into more complex relations and properties those reside in the Web pages.

1.1. Web Mining

The concept of Web mining is firstly proposed by Oren Etzioni in 1996. Web mining is the branch of data mining deployed to Web data for automatic discovery and extraction of information from Web data [1]. The Web data suffers from some peculiar characteristics such as dynamic, chaotic, huge volume, semi- structured or unstructured, diversity in meaning, heterogeneity etc. These characteristics have given rise to several information overload problems directly or indirectly, such as finding relevant information, creating new knowledge out of information available on the Web, personalization of information, learning about customers or individual users, etc. Over the last few years, Web mining is used to address these problems and to improve the accessibility of Web data [2].

ISSN: 2005-4270 IJDTA

1.2. Web Mining Categorization

The Web data is broadly categorized as content data, structure data and usage data [3]. Web content and structure data are the real or primary data on the Web while Web usage data is the secondary data i.e. data derived from user interactions during Web navigation [2]. Based upon type of Web data used as input, Web mining techniques are classified into three categories: Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). All three approaches attempt to extract knowledge from the Web, produce some useful results from the knowledge extracted, and apply the results to certain real-world problems [4]. Major application areas [4, 5] of Web mining include Web search, E-commerce, advertising, fraud detection, improving Web site design and performance etc.

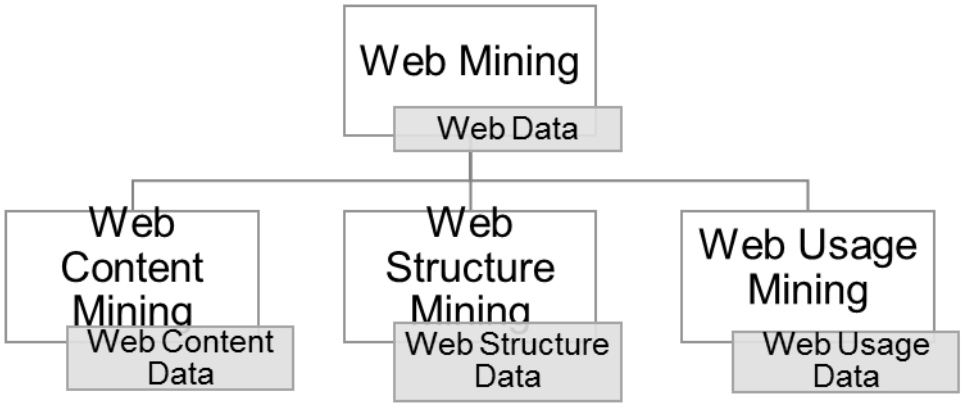

Figure 1. Web Mining Categorization

Among these, Web Usage Mining has become a very popular area of research. WUM basically focuses on how the users of Web sites interact with web sites, list of Web pages visited, order of visit, timestamp of visits, duration of them, etc.

1.3. Web Usage Mining

Web Usage Mining is a process of extracting interesting patterns from Web log file. Hence, it is also called as Web log mining. During the navigation of a Web site, a user's each visit to the site is recorded as a navigation trace in the Web log file. These navigation traces form patterns, called as Web navigational patterns. These patterns reflect the user's behavior during Web navigation. A Web navigational pattern can be a sequence of pages that allows the user to reach a specific set of content to satisfy their requirements. A Web log file is located in three different places: i) Web servers ii) Web proxy servers and iii) Client browsers [6]. Web log file is a rich source of users' navigation information. The information preserved by Web log file includes user's domain, subdomain and hostname, resources requested by the user, date and time of request, any errors returned by the server etc.

Web Usage Mining process consists of five steps:

i) **Usage data gathering**: This step gathers the log data from one or more sources. The main sources of usage data are Web server logs, proxy server logs, registration data, clickstream data, etc. [4, 6].

ii) **Data preprocessing**: The input usage data tends to be noisy, incomplete and inconsistent. Hence to improve the quality of data, Data preprocessing step converts this raw input usage data into rich usage data. Data preprocessing tasks include data

acquisition, Data cleaning, data integration, data transformation, data reduction, path completion, user identification, session identification, etc. [6].

iii) **Pattern discovery**: The pattern discovery phase deals with actual deployment of data mining techniques to the preprocessed data for finding frequent access patterns from sessions discovered in the previous phase. This phase consists of various techniques such as statistical analysis, association rule mining, sequential pattern mining, clustering, classification, path analysis, etc. [7].

iv) **Pattern analysis**: The output of mining algorithms is often not in a form suitable for direct human consumption. Hence, this phase aims at the analysis of discovered patterns such that they could be useful for further applications. The analysis techniques include statistical analysis, visualization, On-Line Analytical Processing (OLAP), database querying, etc.

v) **Pattern applications**: The navigation patterns discovered can be applied to various application areas such as: i) Improving the web page /site design, ii) Recommender systems, iii) Web personalization, and iv)Learning user or customer behavior.

1.4. Ontology based Web Usage Mining

Although conventional Web Usage Mining techniques have shown promising performance in efficient management of web data, their deployment to live web data is limited due to lack of background semantic knowledge. Specifically, the pattern analysis phase suffers from two key issues such as pattern interpretation and pattern retrieval [8].

1.4.1. Pattern Interpretation: Pattern Interpretation of mined data is difficult due to the syntactic nature of web data. It has to deal with the semantic gap between URLs and events performed by users, in order to understand what usage patterns reveal in terms of site events. To reduce this semantic gap, some researchers have proposed the concept of integrating background semantic knowledge in WUM process. This new form of WUM is called as semantic web mining. It is also called as ontology based Web Usage Mining since ontologies play an important role in the semantic enrichment of web data [9].

1.4.2. Pattern Retrieval: The mining techniques such as association rule mining and sequential pattern mining yield a huge number of patterns where most of them are useless, incomprehensible or uninteresting to users. Pattern analysts have difficulty in identifying new and interesting patterns for the application domains. Pattern retrieval deals with the difficulties involved in managing a huge set of patterns, to focus on a subset of them for further analysis. The two approaches to achieve this are, clustering and filtering. Clustering refers to group a set of related patterns according to given similarity criteria; whereas, filtering refers to selection of patterns those have specific properties [10].

2. Review on Ontology Based Web Usage Mining

Conventional Web Usage Mining alone can be problematic in some cases such as there is not enough usage data, in order to extract patterns related to certain categories or when the site content changes as new pages are added to Website but not included in the Web logs [11]. Hence, many researchers had proposed the Ontology based WUM and proved that it enhances the quality of generated usage patterns through experimental results. The literature survey presented here concentrates on integration of semantics in one or more phases of Web usage mining process.

Vanzin *et al* have discussed the two key issues such as pattern interpretation and pattern retrieval incurred in the pattern analysis phase of Web usage mining. They have proposed the use of ontologies to support the interpretation of Web usage sequential patterns [8]. They have also focused on filtering functionality [9] and rummaging

functionality which deal with the pattern interpretation and pattern retrieval problems, respectively [10].

Bredent has described two tools, Web Usage Miner (WUM) and STRATDYN. The WUM tool discovers frequent sequences and also allows inspection of the different paths through the Website. The STRATDYN is developed as an add-on module to WUM tool that extends the capability of WUM tool. It tests differences between navigation patterns for statistical significance. The paper emphasizes the usefulness of integrating the site's semantics in the classification of navigation behavior and in the visualization of results. The site's semantics denote the formal description of the meaning of a site's different Web pages, in the form of ontologies. The semantic integration allows more insights into the process of navigation and helps Website designers in Website adaptation as per users' needs [12].

Eirinaki *et al* have presented architecture of a Web personalization system, called Semantic Web Personalization (SEWeP). The architecture integrates the Web uses mining process with site semantics in order to enrich the set of recommendations provide to the end user. The innovative feature of this architecture is the introduction of C-logs (Concept Logs). The C-logs is an extended form of Web usage logs that encapsulate the knowledge derived from the link semantics or content semantics. These C-logs are then used as input to Web usage mining process. Another innovative feature of the proposed architecture is that the semantic annotation of content is performed using conceptual hierarchy. The architecture uses association rule mining for extracting navigational pattern and recommendations [13].

Eirinaki *et al* have addressed the problem of multilingualism which arises when the content of a web site appears in more than one language. The paper proposes an automatic method for uniformly characterizing a Web site's documents using a common vocabulary. The paper also introduced a novel recommendation method which integrates Web usage data with Web content semantics. Both methods emphasize that the integration of semantics in WUM process gives enhanced results [14].

Zhong *et al* have presented a theoretical framework for two fundamental issues, mismatch and overload. Mismatch means some interesting and useful data /patterns has not found or missed out. Overload means some gathered data is not as per users' expectations. These issues affect the performance of WUM. Hence, the paper proposes the ontology based web mining model which uses ontologies to represent discovered patterns. It consists of ontology extraction, reasoning on the ontologies and capturing evolving patterns. The innovative feature is that the approach deals with the pattern evolution i.e. the system can update its ontology by adding new positive patterns, by removing inadequate patterns or by updating some existing positive patterns if they cause incorrect decision [15].

Bose *et al* have proposed a novel technique to incorporate the conceptual characteristics of Web sites into a usage based recommendation model. The conceptual characteristics of a Web site are described through the concept of hierarchy of Web site. The technique uses a framework based on biological sequence alignment for matching two sequences. The value that estimates the quality of matching is termed as similarity score. The author has introduced a scoring system that generates a similarity score from Web site's concept hierarchy. The author quotes that the model is flexible enough to be extended to incorporate other kinds of domain information such as Web site topology and semantic classification of Web documents [16].

Khasawneb *et al* have proposed an ontology based approach for Web log data preprocessing phase. A fast active user based user identification algorithm is proposed for user identification. For session identification, author had proposed an ontology based method based on Website structure [17].

Rokia *et al* have proposed use of metadata about the content to enhance the discovered patterns' quality. The metadata is stored in domain ontology. The proposed approach is based on association based recommendation and uses click stream analysis powered by an explicit representation of domain knowledge in domain ontology [18].

Nizar *et al* have introduced a comprehensive generic framework termed as an intelligent semantics-aware WUM framework- SemAware. The framework integrates semantic information in the form of domain ontology into the pattern discovery and the pattern mining phase of WUM. In pattern discovery phase, a semantic distance matrix is used in the sequential pattern mining algorithm to prune the search space and partially relieve the algorithm from support counting. The paper introduces a novel method for enriching the Markov transition probability matrix with semantic information to solve the problem of the tradeoff between accuracy and complexity in Markov models used for prediction as well as the problem of ambiguous predictions [19].

Adda *et al* have presented a framework for mining meaningful usage patterns within a semantically enhanced Web portal. The framework concentrates on use of semantic relations between domain concepts. Author further quotes that reflecting domain relations in the navigation patterns results in a new pattern structure and requires a dedicated mining strategy that combines elements from sequential, generalized and graph pattern mining. Author had also described the dedicated level-wise mining method XPMiner for this new pattern structure [20].

Yilmaz *et al* have presented an approach for recommendation system using ontology based WUM and object clustering. The approach is based on ontological data where web pages are seen as objects and sessions as a sequence of objects. User sessions are clustered on a semantic level to capture different behavioral groups [21].

Vijaykumar *et al* have proposed a novel architecture for online recommendation system that uses a novel user navigation classification approach and website ontology concept scoring algorithm for identifying the user intent and predict future browsing pattern of online user using novel Web usage mining method. The author had proposed the Greatest Common Subsequence Detection method to classify user and capture the imminent browsing pattern of user [22].

Kilic *et al* have proposed an approach for clustering Web navigation patterns based on semantic similarity between patterns. Web navigation information of users is integrated with the set of concepts defining Web pages. Each session is a sequence of Web pages and each Web page is represented with a set of concepts from the defined ontology. Sessions are clustered to find meaningful partition to maximize intra-cluster similarity while minimizing inter-cluster similarity. The proposed approach combines the concept based clustering with time spent information [23].

Pinar *et al* have developed a framework to investigate the effect semantic integration on Web navigational pattern generation process. The ontologies are used in pattern discovery phase of WUM. The framework uses sequential association rule mining technique and SPADE algorithm for sequential association rule mining [24].

Hoxha *et al* have presented an approach for semantic formalization of usage logs which lays the basis for effective techniques of querying expressive usage patterns as well as intelligent mining and recommendation methods. It also presents a query answering approach to find expressive patterns of usage behavior. The logs are semantically formalized using domain ontologies and RDF representation of accessed Web resources. The main distinguishing feature of the approach is that it can discover patterns of cross site user behavior with semantic and temporal based constraints [25].

Shirgave *et al* have proposed a semantically enriched WUM method which is an extension to WebPUM approach described in [26]. The approach enriches the Web log data with rich semantic data characterizing content of Web pages and Web site structure characterizing Web site topology i.e. semantic relationship in Web pages [27].

Vijaykumar *et al* have incorporated the website knowledge into WUM process along with server access log file. Website knowledge is represented via concept based website graph. It is a combination of website graph and concept hierarchy of concerned website. The website graph consists of vertices representing the web pages and edges representing the hyperlinks between web pages. Concept hierarchy represents the organization of website content in terms of Website ontology. It is a collection of domain concepts organized using IS-A and HAS-A relationship [28].

Tarrannum *et al* have presented a framework for recommending better Web pages based on the queries fired by users and thus provides a better search utility over Google search engine using ontology and WUM. Authors had built the domain ontology of Web pages of a given Website to represent the domain concepts, the relationships between the concepts with constraints, the instances of concepts, Web pages and the links between Web pages and specific domain terms [29].

Hoppe *et al* have explored the use of ontology to develop a profiling application based on the available online navigational data [30].

3. Key Principles

After this extensive survey of the related research work done in Ontology based Web Usage Mining, we identify a set of underlying principles:

1) Web Usage Mining (WUM) aims to extract navigational patterns from Web server logs. The Web Usage Mining process consists of five steps: usage data gathering, data preprocessing, pattern discovery, pattern analysis, and pattern applications.
2) Although, conventional Web Usage Mining techniques have shown promising performance in efficient management of web data, they suffer from two key issues: 1) their deployment to live web data is limited due to lack of background semantic knowledge, 2) mining algorithms used during pattern discovery phase yield a very huge number of usage patterns.
3) These issues make it difficult for analysts to retrieve the new and interesting patterns and interpret what they reveal about the domain. Hence, to deal with these issues, our research is targeted at the pattern analysis phase.
4) The goal of the pattern analysis phase is to eliminate irrelevant patterns and to extract the interesting ones that constitute knowledge. The pattern analysis phase deals with the aforementioned key issues as: pattern interpretation and pattern retrieval respectively. ***Pattern Interpretation*** deals with the semantic gap between URLs and events performed by users. It is targeted at assessing the meaning and relevance of patterns with regard to the domain. It helps to understand what usage patterns reveal in terms of site events. To reduce this semantic gap, we will use Website ontology for integrating background semantic knowledge in WUM process. ***Pattern retrieval*** deals with the difficulties involved in managing a huge set of patterns, to focus on subset of them for further analysis. The two approaches to achieve this are, clustering and filtering. Clustering refers to group a set of related patterns according to given similarity criteria; whereas, filtering refers to selection of patterns those have specific properties.
5) The key objectives of our research will be:
   i. Develop a Website ontology.
   ii. Gathering usage data from the Web log file.
   iii. Preprocess raw usage data and convert it into rich usage data useful for further phases.
   iv. Discovering frequent access patterns through actual deployment of data mining techniques to the preprocessed data.
   v. During pattern analysis phase,

a. For pattern retrieval, use filtering and/or clustering mechanisms to retrieve subsets of patterns with specific characteristics, in order to deal with the large volume of patterns.
b. For pattern interpretation, the physical patterns are mapped into conceptual patterns. The physical patterns are sequences of URLs. The conceptual patterns are sequences of corresponding concepts described in Website ontology.

vi. Develop a framework for evaluation of proposed approach.

4. Conclusions

In this paper, we have discussed the limitations of conventional Web usage mining techniques incurred due to lack of background semantic knowledge. We have also presented the related work done in Ontology based Web Usage Mining. Finally, we can conclude from the empirical results presented by various researchers that the semantic integration of background knowledge in the form of ontology enhances the quality of patterns generated by the Web usage mining process. These patterns are useful in variety of applications such as Web personalization, recommender systems, Web site design improvement, Web search, E-commerce etc.

References

[1] E. Oren, "The World-Wide Web: quagmire or gold mine", Communications of the ACM, vol. 39, no. 11 **(1996)**, pp. 65-68.
[2] K. Raymond and H. Blockeel, "Web mining research: A surve", ACM Sigkdd Explorations Newsletter, vol. 2, no. 1, **(2000)**, pp. 1-15.
[3] E. Magdalini, and M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology (TOIT), vol. 3, no. 1, **(2003)**, pp. 1-27.
[4] H. Chen, X. Zong, C. W. Lee and J. H. Yeh, "World Wide Web usage mining systems and technologies", Journal of Systemics, Cybernetics and Informatics, vol. 1, no. 4, **(2003)**, pp. 53-59.
[5] S. Jaideep, R. Cooley, M. Deshpande and P. N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data", ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, **(2000)**, pp. 12-23.
[6] F. F. Michele and P. L. Lanzi, "Recent developments in web usage mining research", In Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg, **(2003)**, pp. 140-150.
[7] E. Magdalini, D. Mavroeidis, G. Tsatsaronis and M. Vazirgiannis, "Introducing semantics in web personalization: The role of ontologies", In Semantics, web and mining, Springer Berlin Heidelberg, **(2006)**, pp. 147-162.
[8] W. Yan, "Web mining and knowledge discovery of usage patterns", Cs 748T Project, **(2000)**, pp. 1-25.
[9] V. Mariângela and K. Becker, "Exploiting knowledge representation for pattern interpretation", In Proc. Workshop on Knowledge Discovery and Ontologies, **(2004)**, pp. 61-72.
[10] V. Mariângela and K. Becker, "Ontology-based rummaging mechanisms for the interpretation of Web usage patterns", In Semantics, Web and Mining, Springer Berlin Heidelberg, **(2006)**, pp. 180-195.
[11] V. Mariângela, K. Becker and D. D. A. Ruiz, "Ontology-based filtering mechanisms for web usage patterns retrieval", In E-Commerce and Web Technologies, Springer Berlin Heidelberg, **(2005)**, pp. 267-277.
[12] B. Bettina, "Using site semantics to analyze, visualize, and support navigation", Data Mining and Knowledge Discovery, vol. 6, no. 1, **(2002)**, pp. 37-59.
[13] E. Magdalini, M. Vazirgiannis and I. Varlamis, "SEWeP: using site semantics and a taxonomy to enhance the Web personalization process", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, **(2003)**, pp. 99-108.
[14] E. Magdalini, C. Lampos, S. Paulakis and M. Vazirgiannis, "Web personalization integrating content semantics and navigational patterns", In Proceedings of the 6th annual ACM international workshop on Web information and data management, **(2004)**, pp. 72-79.
[15] L. Yuefeng and N. Zhong, "Capturing evolving patterns for ontology-based web mining", In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, **(2004)**, pp. 256-263.
[16] B. Amit, K. Beemanapalli, J. Srivastava and S. Sahar, "Incorporating concept hierarchies into usage mining based recommendations", In Advances in Web Mining and Web Usage Analysis, Springer Berlin Heidelberg, **(2006)**, pp. 110-126.

[17] K. Natheer and C. C. Chan, "Active user-based and ontology-based web log data preprocessing for web usage mining", In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, **(2006)**, pp. 325-328.
[18] A. Mehdi, P. Valtchev, R. Missaoui and C. Djeraba, "Toward recommendation based on ontology-powered web-usage mining", Internet Computing, IEEE, vol. 11, no. 4, **(2007)**, pp. 45-52.
[19] M. R. Nizar and C. I. Ezeife, "Using domain ontology for semantic web usage mining and next page prediction", In Proceedings of the 18th ACM conference on Information and knowledge management, ACM, **(2009)**, pp. 1677-1680.
[20] A. Mehdi, P. Valtchev, R. Missaoui, and C. Djeraba, "A framework for mining meaningful usage patterns within a semantically enhanced web portal", In Proceedings of the Third C* Conference on Computer Science and Software Engineering, ACM, **(2010)**, pp. 138-147.
[21] Y. Hakan and P. Senkul, "Using ontology and sequence information for extracting behavior patterns from web navigation logs", In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on IEEE, **(2010)**, pp. 549-556.
[22] M. Vijayakumar and C. H. Muthaial, "An ontology based approach to implement the online recommendation system", In Journal of Computer Science, vol. 7, no. 4, **(2011)**, pp. 573-581.
[23] K. Sefa, P. Senkul and I. H.Toroslu, "Clustering Frequent Navigation Patterns from Website Logs by Using Ontology and Temporal Information", In Computer and Information Sciences III, Springer London, **(2013)**, pp. 363-370.
[24] S. Pinar and S. Salin, "Improving pattern quality in web usage mining by using semantic information", Knowledge and information systems, vol. 30, no. 3, **(2012)**, pp. 527-541.
[25] H. Julia, M. Junghans, and S. Agarwal, "Enabling semantic analysis of user browsing patterns in the web of data", arXiv preprint arXiv: 1204.2713, **(2012)**.
[26] M. Jalali, N. Mustapha, N. Sulaiman and A. Mamat, "WebPUM: A Web-based recommendation system to predict user future movements", Expert Systems with Applications, vol. 37, no. 9, **(2010)**, pp. 6201–6212.
[27] "Expert Systems with Applications", vol. 37, no. 10, **(2010)**, pp. 7295.
[28] S. Suresh and P. Kulkarni, "Semantically enriched web usage mining for predicting user future movements", International Journal of Web & Semantic Technology, vol. 4, no. 4, **(2013)**, pp. 59.
[29] T. Kumar, V. H. S. Guruprasad, K. M. B. Kumar and I. Baig, "A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph", International Journal of Computer and Electrical Engineering, vol. 6, no. 1, **(2014)**, pp. 67.
[30] S. H. R. N. Tarannum and R. R. Keole, "A Preliminary Review of Web-Page Recommendation in Information Retrieval Using Domain Knowledge and Web Usage Mining", **(2015)**.
[31] H. Anett, A. Roxin and C. Nicolle, "Ontology-based Integration of Web Navigation for Dynamic User Profiling", Informatica Economica, vol. 19, no. 1, **(2015)**, pp. 10.

