

Ontology Based Extraction of Web Site Key Object

Vishnu Dutt Tripathi¹, Prateek Gupta²

¹ *MTech Scholar, Department of Computer science and Engineering, SRIST, Jabalpur, vishnu.dutt.0101@gmail.com, India;*

² *HOD, Department of Computer science and Engineering, SRIST, Jabalpur, pguptace@yahoo.com, India;*

Abstract - Network mining has been traditionally used in different application domains in order to enhance the Content that Web users are accessing. Likewise, Website administrators are interested in finding new approaches to improve their Web site content according to their users' preferences. Furthermore in Semantic Web has been considered as an alternative to represent Web content in a way which can be used by intelligent techniques to provide the organization, meaning, and definition of Web content. In This work, we define the Website Key Object Extraction problem, whose solution is based on a Semantic Network mining approach to extract from a given Website core ontology, new relations between object according to their Web user interests. This methodology was applied to areal Website, whose results showed that the automatic extraction of Key Objects is highly competitive against traditional surveys applied to Web users.

I. Introduction

The rapid growth of the World Wide Web, the assembly of large- scale volumes of Web data, and ever exponentially add to applications have led to the development of ever smarter approaches to extract patterns and build knowledge with the aid of artificial intelligence techniques. These techniques have been used, together with information's technology, in a broad range of applications.

This is where semantics, social network analysis, Web structure, content, usage, and more aspects have already been and will increasingly be included in many application domains. One of such domains is applicable to how Web users browse the Web pages and Websites looking for information. Frequently, they require and there is a better possibility of them staying or returning to the Website if they find the content they are searching for in a Website. For this, Website administrators determine to reach the greatest user base they can, therefore it is within their interest to provide accurate and correct content. However, different difficulties are current in this application domain. On the one hand, Web users' interests frequently change and it is frequently unclear to assume at first sight what the users' interests are. On the other hand, whether the content has been rightly presented is a relevant question for any Website administrator. Furthermore the content may be presented in certain formats, which could stem from free text to images or videos. In this sense, not only it is unclear what is the content that users are looking

for, but also their preferences in agreements of the format that should be considered.

A typical Website is controlled primarily of a free text being formatted within the limitations imposed by the HTML standard. However, they also consist of almost data formats such as images, videos, etc. An example of this is the almost successful sites of the so-called Web 2.0 such as You Tube where the main convene of interest lies in Web videos. A drawback to this trend is that the formats shown above do not provide information as regards their content which can easily be retrieved by a computer and, therefore, a short degree of content analysis can be carried out in relation to them. The foremost improvement of this work is a methodology that enables the extraction of significant Website Key Objects, following a Semantic Web mining approach. In this case, Semantic Web mining will be heeded as using data mining algorithms in order to deduce relevant information from the Semantic Web representation of a given Website. Specifically, the idea is to deduces a new relation between structured components from a Website (WebObjects), represented by a simple core ontology. This relation is deduces from the Web user's perspective, represented by their collected drill, from which patterns are extracted.

Background: - The rapid growth of the World Wide Web, the assembly of large- scale volumes of Web data, and ever exponentially increasing applications have led to the development of ever smarter approaches to deduce patterns and build knowledge

with the aid of artificial intelligence techniques. These techniques have been used, together with information technology, in a broad range of applications. This is where semantics, social network analysis, Web structure, willing, usage, and other aspects have already been and will increasingly be included in various application domains.

Significant information extraction using Web mining :- significant information extraction from Web content has been a major focus for many researchers, where different degrees of information, such as words, text passages, or WebObjects, have been taken into account. Furthermore, various methodologies have been proposed, and some of the most relevant approaches will be discussed in the following.

The methodology for finding Website keywords forms the basis of this work, in which information retrieval and Web usage mining techniques were used within the Knowledge Discovery in Databases (KDD) framework, to find the keywords that define the search process for a group of users. The process described is based on five basic steps. The first one is associated with the Vector Space Pattern definition from a given Website and the processing of Weblogs, in order to include the end-user information. Afterwards, this methodology focuses on finding the relationship between the page interest and time spent, as well as selecting the most important pages from the extracted user sessions. Finally, by using different clustering techniques (particularly k-Means and Kohonen Self-Organizing Article Maps), the process to discover Keywords in clusters takes place. The attempts for attracting users to Websites have been created since the Web became a massive source of information, and the study of usability in Websites has been one of the most widespread research domains. One of the first approaches to create Website usability patterns is the Common User Access (CUA) proposed. Another approach focuses in how to present the content in terms of typography, design, presentation of elements and other end-user properties associated with their interaction with visualization components. These patterns have shown to be effective, but they lack information about the direct feedback for the users of the site.

Key phrases, which can be single keywords or multiword key terms, are linguistic descriptors of documents. They are often sufficiently informative to allow human readers get a feel for the essential topics and main content included in the source documents. Key phrases have also been used as features in many text-related applications such as

text clustering, document similarity analysis, and document summarization. Manually extracting key phrases from a number of documents is quite expensive. Automatic key phrase extraction is a maturing technology that can serve as an efficient and practical alternative. In this paper, we present an ontology-based approach to building a Vietnamese key phrase extraction system for Vietnamese text. The rest of the paper is organized as follows: Section 2 states the problem as well as describes its scope, Section 3 introduces resources of in Wikipedia contains a rich body of lexical semantic information, the aspects of which are comprehensively described in (Zesch et al., 2007). Additionally, the redirect system of Wikipedia articles can be used as a dictionary for synonyms, spelling variations and abbreviations.

A page: - A basic entry in Wikipedia is a page that represents either a normal Wikipedia article, a redirect to an article, or a disambiguation page. Each page object provides access to the article text (with markup information or as plain text), the assigned categories, the ingoing and outgoing article links as well as all redirects that link to the article. Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the process of extracting useful information from server logs i.e. user's history. Web usage mining is the process of finding out what users are looking for on Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web usage mining is a process of data mining whose users use who's contents .Example..One user government jobs contents and another user use private job contents.

II. Method

The proposed approach to extract the Website Key Objects is presented. Firstly, the problem definition and the general notation of terms are introduced. Secondly, Web content and Web usage mining terms and methodologies used to achieve the Website Key Object definition are presented. Finally, the core methodology and main contribution of this work are detailed.

Proposed system: - In our approach, the usage of metadata to describe Web Objects will be considered as the basis to constitute the information source, in order to build a vectorial representation of its content. However, the end-user's point of view will be considered as the principal research topic of this approach. Therefore, the contents of

the Website and Weblogs are combined for processing.

In order to achieve an accurate WSKO extraction, the Website can be represented as a core-ontology from which concepts and its relations will be used as input to the WSKO extraction process.

WebObjects: - WebObjects and Website Key Objects terms are presented, as well as the proposed ontological representation and mathematical notation. Likewise, the Website Key Object (WSKO) extraction problem is introduced.

In our approach, the usage of metadata to describe WebObjects will be considered as the basis to constitute the information source, in order to build a vectorial representation of its content. However, the end-user's point of view will be considered as the principal research topic of this approach. Therefore, the contents of the Website and Weblogs are combined for processing.

Website Key Objects: - WebObjects or groups of WebObjects that attract the Web user's attention. Key Objects can be considered as elements on a given Website that provide knowledge of both content and format that appear interesting to end-users. Enhancements can be made in presentation as well as in content when Key Objects are identified, and used to improve the structure of a Website.

Website Key Object Extraction: - In order to achieve an accurate WSKO extraction, the Website can be represented as a core-ontology (Stumme et al., 2006), from which concepts and its relations will be used as input to the WSKO extraction process. In general terms, the Key Objects can be represented as an order relation from the end-user perspective, where each object's relevance is inferred from the usage of the given Website.

Key Objects core ontology:- A Website Key Objects' core ontology is represented by the tuple. The proposed ontology sets a common ground to characterize each object, independently from the original format from which it was created. By using this ontology, it is possible to make a pair wise conceptual comparison between objects, disregarding their original format. Overall, Web Concepts can be represented as keywords used by the Website administrator to define a given object. Likewise, Web Meta Concepts will be considered as groups of keywords or categories, associated with a more general concept than the one's used in the set of Web Concepts. However, Web Meta Concepts will not be considered in this work as part of the ontology and will be used as categories.

Sessionization and approximated time spent in Web Object: - A normal Weblog considers,

among other information, the page a host requested and a timestamp for the request. By reconstructing the user sessions, it is possible to determine how much time each user spends on a given Webpage. However, it is not possible to determine how much time that user spends in a certain object within that page. The analysis should be made under the assumption that every user spends an equal amount of time in each object that defines a page. If this assumption is made, the analysis that could define Website Key Objects would be merely the analysis of pages browsed by users and a definition drawn on the basis of the most popular objects. To avoid this, an approximation of the time spent by each user is obtained by making a survey over a controlled group of users. The purpose of this survey is to analyze which objects were more appealing to users in each page, so a grade was awarded to every object in a given Webpage.

User sessions clustering: - The SOFM machine learning algorithm is a special type of neural network where a typically two-dimensional grid of neurons is ordered so it reflects changes made in the n-dimensional vector that neurons represent. In this particular case, these vectors can be considered as IOVs. SOFM works with the concept of neighborhoods among neurons, where within the grid, some neurons are considered as neighbors and further more changes in one neuron will affect their neighbors.

III. Result



Fig 1: First Look



Create An New Account

First Name
 Manikandan
Last Name
 mani
Email Id
 sample@gmail.com
Re-Enter Email Id
 Eg: balaji@gmail.com |
Date Of Birth
 Eg: 05-02-1986
Location
 Eg: Chennai
Job
 Eg: Student or designation
Name Of School/College or Company
 Eg: Mary's or Wipro
Gender Male Female
 I Agree to the All Terms of Service and Privacy Policy.

Fig 2: Registration Form

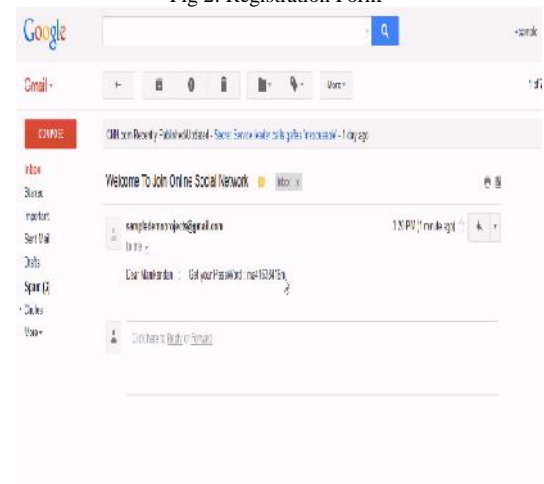


Web content Extractions Home Provider Logout

Keywords of the Web site

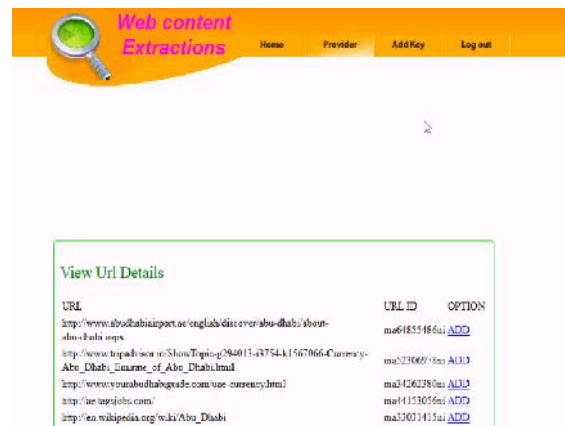
- Ahmed
- Aba
- Dhabi
- Dhaker
- Ah
- Dhabi
- Ah
- Dhabi
- International
- Airport
- var
- pmg
- pmg
- pmg

Fig 6: Search Keyword



Google search bar and Gmail interface showing a confirmation message: "Welcome To Join Online Social Network".

Fig 3: Confirm Registration



Web content Extractions Home Provider Add Key Logout

View Url Details

URL	URL ID	OPTION
http://www.abudhabiairport.ac/english/dia-over-abu-dhabi/about-us.html	ma64855486a	ADD
http://www.inpads.gov.ae/Share/Topic/264011-43754415670664/Currency-Abu-Dhabi_Currency_of_Abu-Dhabi.html	ma22306938a	ADD
http://www.youabudhabiguide.com/uae-currency.html	ma34262380a	ADD
http://ae.spajobs.com/	ma41153056a	ADD
http://en.wikipedia.org/wiki/Abu_Dhabi	ma33034151a	ADD

Fig 7: URL With id Ready for position



User Name:

Password:

[New User Signup](#)

Fig 4: Login



ma6485486a

ma6485486a

Fig 8: Given Link Position



Web content Extractions Home Provider Logout

Fig 5: URL Section Search Keyword



Fig 9: Keyword Search and according this result

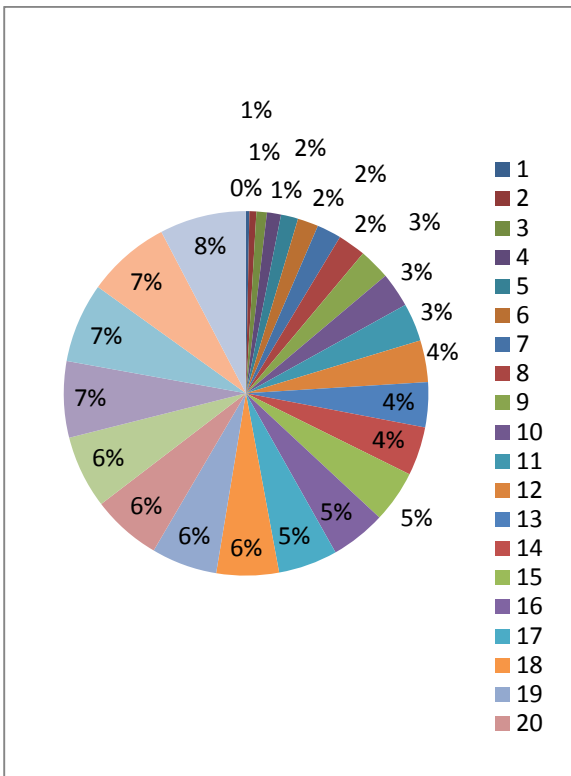


Fig. 10 Expected keyword by users

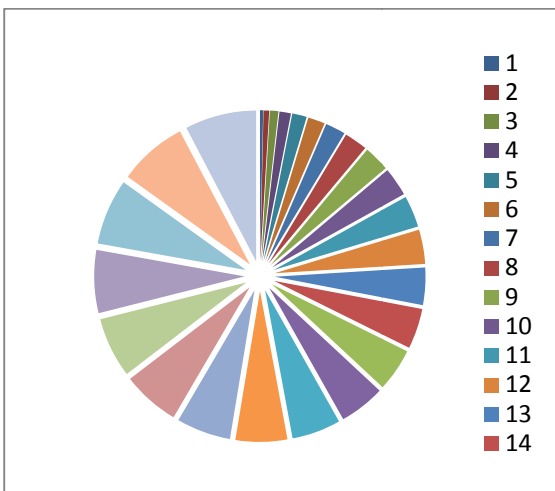


Fig. 11 Get keyword by ontology based webmining

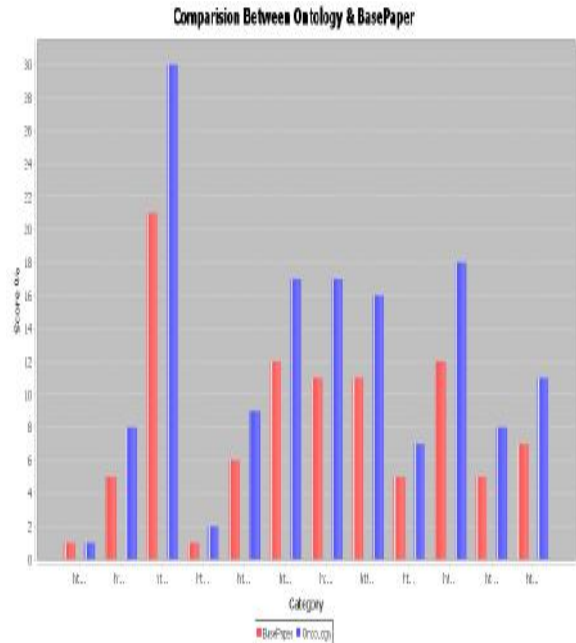


Fig. 12 comparison between base paper & Ontology based webmining

S.NO.	Result	Keyword Webmining
1.	Base Paper	30%
2.	Proposed	70%

IV. Conclusion

In this work, a methodology for identifying Website Key Objects is introduced. Website Key Objects are the most appealing objects for users within a Website. This methodology is a generalization of a prior developed by Velasquez et al. (2005) for identifying Website Keywords. Our approach is based on the fact that there is a correlation between the time spent by a user in a certain page during a session and the interest the user has in its content.

In order to develop this methodology a definition of a Web Object was created, and particularly a definition for Website Key Objects, which are those objects in a Website that drives the attention of users. The definition of these objects enables the characterization of the conceptual content represented by simple core ontology. The Website Key Object Extraction Problem (WSKOP) is aimed towards the definition of a new relation between the core ontology's WebObjects. This relation is an ordered list of Website Key Objects, according to the Web user preferences.

Reference

- [1] Baeza-Yates, R.A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- [2] Berendt, B., Hotho, A., Stumme, G., 2002. Towards semantic web mining. In: ISWC '02: Proceedings of the First International Semantic Web Conference on the Semantic Web. Springer-Verlag, London, UK, pp. 264–278.
- [3] Berendt, B., Spiliopoulou, M., 2001. Analysis of navigation behavior in web sites integrating multiple information systems. The VLDB Journal 9, 56–75.
- [4] Berry, R., 1998. Common user access—a consistent and usable human–computer interface for these environments. IBM Systems Journal 3 (27), 281–300.
- [5] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.
- [6] Burget, R., Rudolfova, I., 2009. Webpage element classification based on visual features. In: Proceeding of the 1st Asian Conference on Intelligent Information and Database Systems ACIIDS2009, Dong Hoi, VN. IEEECS.
- [7] Chambers, N., Allen, J., Galescu, L., Jung, H., Taysom, W., 2006. Using semantics to identify web objects. In: AAI'06: Proceedings of the 21st National Conference on Artificial Intelligence. AAAI Press, pp. 1259–1264.
- [8] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. Ai Magazine 17, 37–54.
- [9] Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal Wendy Hall, Paul H. Lewis, Nigel R. Shadbolt et. al. “Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web”, IEEE Intelligent Systems, 2003.
- [10] Juan D. Velasquez, Luis E. Dujovne, Gaston L'Huillier et. al. “Extracting Significant Website Key Objects: A Semantic Web Mining Approach”, Journal of Eng. Appl. Artif. Int July 29, 2012.
- [11] Milos Kudelka, Vaclav Snasel, Zdenek Horak, Aboul Ella Hassanien, Ajith Abraham, Juan D. Velásquez et. al. “A novel approach for comparing web sites by using Micro Genres”, 2014 Elsevier Ltd.
- [12] Juan D. Velasquez et. al. “Combining eye-tracking technologies with web usage mining for identifying Website Key objects”, 2013 Elsevier Ltd.
- [13] Carlos Vicent, David Sanchez, Antonio Moreno et. al. “An automatic approach for ontology-based feature extraction from heterogeneous textual resources”, 2012 Elsevier Ltd.
- [14] Juan D. Velasquez Silva et. al. “Improvement of a Methodology for Website Key object Identification through the Application of Eye-Tracking Technologies”, University of Chile Santiago, IEEE 2012.