International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 4, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Ontology Extraction by Crawling HTML Pages

Hamed Zeinali
Department of Computer
Science and Research Branch
Islamic Azad University, Sirjan, Iran
Email: zeinali.hamed@yahoo.com

Reza Nourmandi-Pour
Department of Computer,
Sirjan Branch, Islamic Azad University,
Sirjan, Iran
Email: noormandi_r@iausirjan.ac.ir

Arash Azizi Mazreah
Department of Computer
Sirjan Branch, Islamic Azad University
Sirjan, Iran
Email: aazizi@iausirjan.ac.ir

Abstract **– Web is great resource of general and especial data about any things. Through establish relation between many of web pages we can provide new web service of them. So information managementof web pageshelpsus to applied web document for novel queries, links and other applications. Semantic web as a hot topic in the past decades, has tried to provide more computer recognizable data of web and upgrade web for increasable need of users. Ontology extraction is start part of create semantic web layer. In this paper we survey how we get the more utility of web capacity and its information capital by ontology extraction. Our result has been shown via a typical of web, free and accessible information; where next of ontology extraction provide new service for users.**

Keywords **– Component, Semantic Web, Ontology Extraction, Information Management, Web Service, Persian Web.**

I. INTRODUCTION

Web is mass repository of data such that are mostly managed by relational database (RDB). Today web data are more increased permanently. Many group pages as weblogs, personal home pages, news, electronic publications, product informationin the electronic stores and etc. are received new documents per any day, hour, minutes or secondly. This information almost has writing by non-official and without any rent for their produce. Making machine understandable and semantic access to this huge data is necessary. In this regard semantic web technology has been suggested in order to process web data.

Semantic web has focuses on the ontology extraction. Ontology extraction is considered as attractive studies in the field of semantic web. Nevertheless, there is no considerable attention on the Persian web and extract ontology from Persian web. This paper study aims to introduce how we can extract ontology from Persian web. Then, ontology extraction will run for a specific domain.

A. Persian Web

Persian web is a certain platform of information in Persian language. It has been expanding in recent years and makes the diversity of information, information volume, and applying standard domains of Persian web dot ir important [1]. Today, different gateways such as online stores, virtual communities, weblogs, numerous news sites and professional and different organizational portals are active in Persian web. However, what makes the exploitation of such platform weak and sometimes impossible, is lack of effective processing and processing tools[2]. Considerable knowledge is produced freely and without any fee in Persian web, but this web has no capability to process these information and knowledge, hence they will exploit the minimum of their property. While the World Wide Web has endowed making intelligent and upgrading the processing, it seems necessary to consider semantic web in the Persian web platform as fast as possible. This help to spur Persian cyberspace web which will provide the implementation of different IT projects with the aim of commercial, scientific, and cultural[3,4].

A lot of knowledge produces and publishes in Persian website, but there is no capability to process these knowledge and information in the web. Even capacity of search engines for Persian language is very weak and most of the web pages are cough in spam pages and irrelevant results. Unlike Persian web, semantic web has been able to present the intelligent and upgrade processing to World Wide Web. Progress of semantic web for languages like English and Chinese has been passed from research and some applications are commercialized. Therefore, it is necessary to consider the semantic web study in Persian language platform as fast as possible. This help to flourishing the Persian web cyberspace which will provide the possibility of implementation of numerous IT projects with commercial, scientific, and cultural purposes [5]. Localization of semantic web knowledge for processing the Persian web is a virgin research area which its creation can have many changes in web and using it.

In this paper we have attempted to follow the creating these mantic web layer based on a specific case study. In the current study, deficiencies to create Persian domains ontology, benefits, and capabilities which are today available by semantic web for these domains will be investigated. It is tried to introduce a method for ontology extraction from Persian web in this investigation.

The remainder of this paper is organized as follows: In Section II, we have review of previous worksfor ontology extraction. In section III, we discuss about management of web information. In Section IV we proposed new method for ontology extraction of HTML pages. Section V consists of some examples and the result of our work and comments on some future work.

II. RELATED WORKS

Ontology extraction methods for relational databases often are effective when they have access to a relational database schema [6,7,8]. In a general view, this will make many limitations in semantic web. For example, if extract ontology domain of Iran universities and form its semantic web layer, it must provide access to relational database scheme for more than 1000 different universities, while this

is impossible. Although, even if somehow managing this problem resolve, ontology extraction approaches based on relational scheme haven't high performance to present answer at the level of a semantic web service [9]. Perhaps the most important problem to represent a practical approach to create ontology domain is that most of databases information are not available. However, this is not the only problem. Irrelevant information, permanent change in pages, and high processing overhead are problems which make the semantic web ontology extraction hard on available RDBs. Concentration on the part of web which its documents are modeling doesn't mean the capability of suggested approach in ontology extraction, even most of today's web is so [10].

III. Manage of Web Information

Software engineers have always viewed business process management and integration from their angle. Attempts to bridge the gap between business and IT have been initiated from the IT side at large. As a result, software engineers have not done enough to enable business experts to describe business semantics.

The need for better alignment of IT with business has been articulated uncountable times. Lessons from experience seem to indicate that the only way to fulfill this need is to equip business experts with the means to express business knowledge in a language, they are able to understand and which can also be spoken by IT experts.

Such kind of business language needs syntax and semantics of concepts and relationships among them must be clearly defined. That is what ontology provides. To be able to make business knowledge persistent, there must be some kind of repository. A repository is represented by a knowledge base.

Ontology can be compared to a relational database schema, which is the structure for a database. a relational database does not contain any user data after a database schema has been defined. also, an ontology contains no user data. When we instantiate an ontology, that is when information is entered and stored, we actually create a knowledge base. Like a relational database with tuples, a knowledge base contains structure and data. with the advent of web ontology language (OWL), the term ontology is now used to mean both ontology and knowledge base.

Depending on the domain, developing an ontology can take several months and even several years and may require ongoing involvement of domain experts. Since organization tend to eschew engaging domain experts in long running endeavors and encoded in computer applications and databases, it might be more practical to derive this knowledge from existing systems. This derived knowledge could then be used to generate a putative ontology, which can then be revised or further augmented. The rift between business and IT can be eliminated, thanks to semantic bridge that ontology represents.

IV. Ontology Extraction from HTML Pages

Nowadays, a considerable part of web content is saved in web database and after a retrieval cycle, broadcasts for user in the web in formats available for showing. Several databases with different schema are broadcasting different contents in similar websites, which are mostly not accessible directly. From the other side, accessing these databases is necessary to create semantic web. Ontology production and creating metadata tags is only available in this way.

A. Content Management Systems

Through development of content management software's content management systems (CMS) formed. A content management system (CMS) in the web is software on the server which creates the contents of web pages through dynamic connection to database. Pages with constant and non-transformative content are called constant (static). Today, websites are so designed that extract the raw information from database after connecting to a database and display it in the webpages and CMSes do this. A CMS characterize following applications:

B. Separate the content from exhibition

A report is edited only after content for broadcast in the web, and cases such as outstanding titles, figures position, broadcast location and other cases related to show in the web, are performed by CMS.

C. Publish cycle management

Management of content, before and after its production is completely implemented by CMS. Who can edit the contents and when it continues to broadcast it, are samples of this management.

D. Content evaluation

Possibility of the evaluation of contents is one of important CMS capabilities which are prepared automatically and make available for content management. Ability of a CMS to present reports about available contents is dependent on the structure of websites database and level of saving events around contents.

Main elements of CMS are presented in the following figure. According to the figure content management adds its content through database intermediate, then when the user sends a request from domain, available contents in the database are placed in the exposed format, and are represented as final output for user.

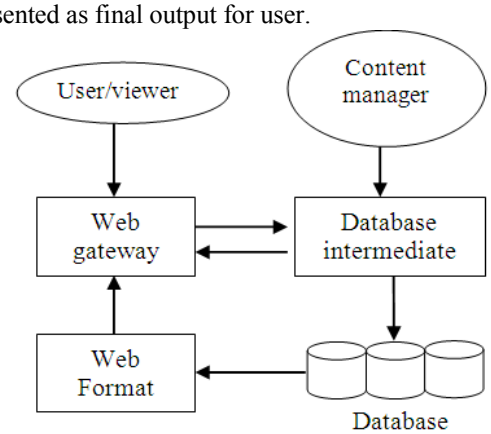

Fig.1. Overall structure of a CMS

There are two types of content management systems: page base CMS and future base CMS.

1) Page base CMS

Information unit of website is page in these CMSes, and website is considered as a collection of pages, basically. Therefore, these types of CMSes have a high ability in management of pages in one of htm, jsp, php, html, and such formats. Searching these types of websites leads to presentation of a list of links including pages having word searched by manager. These types of CMSes don't use database basically.

2) Future base CMS

All the capabilities needed for customers are inserted in future base CMSes. Capabilities include menu management, management of pictures gallery or picture album, member's management, links management and so no. After creation of various abilities for website, manager of site would only insert the information into databank structure. Organizing the pages and necessary editions to create website output is done by CMS. Page creation is automatically in this method.

V. Benefits of Ontology Extraction form PCMS

Suppose that each of the below cases with information in the PCMS are followed:

Search to find a professor with specified special field Suppose that we want to follow a specialist professor about processing the pictures having experience of working on MRI pictures (or satellite pictures, infrared pictures and ...).

Scenarios: following scenarios are stated for this:

a. Referring to special page of about 1000 professors in the fields of computer and electronics in different universities to find person having regarded properties.
b. a simple query in search engines like Google
c. query in a special directory such as database of journals related to image processing, or societies specialized about processing MRI images or one of both.
d. Refer to highly experienced professors which are relative to this field to find people active in these fields.

Each scenario has some problems:

1. Suggestion *a* cannot be a real choice, because in the simplest search condition for 1000 set of pages it needs 4 hours, if each page loading time be 10 seconds. Therefore, such a query is not possible. From the other side, if we generalize it for the same query such as teachers working in the field of Web GIS, it needs similar time (a combination of Web and GIS). However, many errors occur in this method.
2. Suggestion *b* may be able to provide a negligible part of needed information, but it will never help us to find suitable answer, because firstly ranking mechanism of pages in Search Engine more than correlation with meaning and keywords content it depends on other parameters such as site credit, rate of visits and etc. Therefore it might index the targeted content in very low rankings that brows all topics to reach the considered topic are difficult. Second, if all the listed results are aimed contents, gathering these results and revealing identification in this regard is never easy. For example, person expertise in processing MRI images might:
   - Explicitly has referred to this in his research interests.
   - He has just referred to image processing in research interests as interesting research topic, but has published some papers in the field of working on MRI images.
   - He has conducted a project about processing MRI images, in the projects part.
   - He may present a course or special workshop about processing MRI images.
   - May have participated in a conference on the subject, and etc.

   These reveal that the individual is familiar with the field. Nevertheless, someone may has all these cases in his profile or part of them that determination of this along with error from incorrect interference is considered as weakness of approach *b*.
3. However, suggestion *c* and query in a specific directory is better than last two choices, but there are two problems: first, finding that directory or page of forums need query, if existed. Second, query in such domains provide part of information, itself. For example, we may inform someone who has paper in the field of image processing by searching in articles database. Although, it is not possible to find that he had research about processing satellite images or medical images.
4. Approaches *d* don't give us a documentary response and solely according to impressions which are probably blind and are not informed from most of the cases lead to a partially correct answer.

As it is not possible to accurately query from specific field of a professor in the current condition, a great domain of queries like followings have also these problems:

- Determination of the all articles of a professor in a specific field (like image processing) in 1390
- Which professor has presented the "image processing" course in first semester of 1392
- Which individuals are presented a research article about "semantic web" with identified person X?
- Which professors are participated in "international conference of semantic web" in 2013
- Which professors have presented paper in the field of " biometric security systems"
- Which professors have published a book in the field of "database" and etc.

If we are going to answer each of these requests in the current web, we should use approaches more than having complexities have no efficiency to present effective answer. For example, suppose that each of professors A and B have brought the title of improving the process of recognition of the pattern of liver cancer on cardiographic images without mention of authors. How a search engine can understand this is a joint paper between the two?

VI. Ontology Extraction by Focused Search

In this chapter we will illustrate a new approach of domain ontology extraction when there is no access to

relational database scheme. The method can support ontology domain extraction semi-automatically. Efficiency of the approach is limited to type of managing HTML assignments related to a domain. In the suggested approach, when dynamic construction of HTML pages follows a template (same or different) with content of data available in relational web database, extracted ontology will provide a high assessment features. Today, modeling of HTML documents structures is very popular. Due to high rate of webpages construction, currently all the process of creation, publication and change of HTML documents is carried by content management software (CMS). Therefore, many different domains of Web pages have a common pattern. Hence, based on the features CMS provide for management of webpages, mapping method of CMS to ontology is introduced as making the ontology by scheme-like RDB and HTML tree (PSTH)[1]. Today, a large portion of databases available in web are managed by portals, which CMS is part of it.

We will study the structure of management of HTML pages in general condition and when are managed by CMSes. We will indicate how to find a way to extract ontology automatically from databases based on CMSes. Then, we will suggest a new architecture to create a semi-automatic ontology from webpages. Investigation of the benefits of suggested approach in contrast to other approaches will state the possible platforms for evaluation of this approach.

In order to create ontology automatically from web pages, first we try to establish available relational database using a specialized crawler. For this, we use Sphider-plus which is an open-source language search engine and based on PHP language and Mysql database. The software provides specialization of crawler behavior together with its codes without high complexities, and also is capable of indexing and ranking the web pages based on identified primary keys. Although, we don't consider the indexing capability and ranking of webpages by this crawler, and just seek to extract desired information using web pages and insert them as a new record in database.

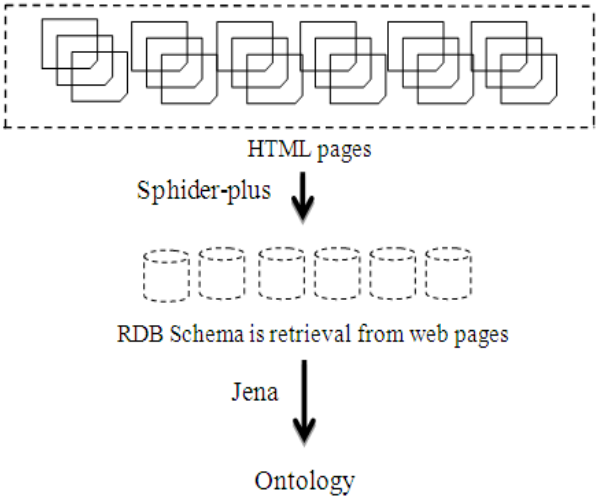

Fig.2. Ontology Extraction by Focused Search

[1] Pseudo-schema of relational DB and tree of HTML

Nextwe extract ontology from RDB that is retrieval from html pages. This stage is doing with HTML tags is various information sections of HTML pages and RDBschema where is created by Sphider-plus crawler. We can even effectively extracting domain ontology.

A. Retrival RDB schema by focus search

Sphider-plus crawler implementation process and establishing referencedatabase from pages in the libraries of universities.One way to run our crawler in windows environment is as follow.

1) Install the Wamp program on computer. Installation of this software will automatically implement the PHP, Apache, and Mysql software's on computer.
2) After installation of Wamp software in drive (C:\), we will transfer the crawler Sphider-plus to C:\wamp\www path.
3) If Mysql database has settings such as user password database.php file will open from C:\wamp\www\sphider-plus\settings path and settings required are performaed as follow: $database, $mysql_user, $mysql_ password, $mysql_host
4) We will run the install_all.php program from path http://localhost/sphider-plus/admin/ install_all.php.
5) Now, we proceed to run Sphider-plus crawler. Here, from ./admin/admin.php path go to the management section of crawler and introduce the desired initial list to search engine which include a list of link of libraries of universities (in Persian) and is located in reference address.

Now, crawler begins to search and save our desired information in its database.

Table 1: Search Index

Name	URL
University of Tehran	http://dlib.ut.ac.ir/
Tarbiat Modares University	http://modares.ac.ir/library
Amir Kabiruni. of Technology	http://library.aut.ac.ir/
Sharif uni. of Technology	http://sharif.ir/~library/

- benefits of index for specialized searching

The index has two effects on performance of search engine:

1-it put the specific key words in the file common.txt and make it available for crawler. This makes the crawler only regard the parts of webpages containing these words.

2-Also, Rie-index intelligent module is used to prevent presented pages.

B. Ontology Extraction

Persian Web Ontology Extraction problems have not been followed, like a lack of specific vocabulary in different domains and computing weak support web standards Persian useful results. In this study, we first focused on the development of Persian extraction, Semantic Web, Web Ontology, Web slopes of Persian. The first study of the relational database web directly transferred to the ontology.

We have reviewed the approaches of creating ontology from HTML pages in part 2-9-3. As it was stated there, using HTML pages can help to improve constructed

ontology from relational schemes. However, using HTML documents have benefits, which ultimately can lead to improvement of extracted ontology. Overview of tree structure of HTML is as follow:

```
<html>
<head>
.....<!—page information tags—> .....
</head>
<body>
..... <!—The main bodytags—>.....
</body>
</html>
```

Each HTML file creates with <html></html> tag and the content are placed between the tag. HTML pages are composed of two <head> and <body> parts. Information of the page that don't render are placed in <body> part. In this part, important tags such as Title, Style, Script, Nextid, Meta, Link, and Base are applied. We can consider following advantages for work with HTML pages:

- Each HTML page has a unique URL which parts of the URL states its content and makes possible the class determination of that data.
- Information is placed in a tree structure in HTML page. Therefore, information available in sub-nodes of a node is considered as part of main node concept or a subclass of it.
- Information is placed in each location of HTML page in a tag. Therefore, it can identify the existing relationship between data placed in the same tag in a set of different pages. General structure of a tag in HTML is as follow:

```
<tagname id=idname option1 =value1… option k =value k>Tag Value </tagname>
```

If each tag id is determined accurately, it can simply identify the considered tag through it. <title> tag specifies the title of a page. <meta> tag is used for various applications such as determination of page language, page keywords, page distribution, and such cases. We have brought some of these applications.

```
1. <meta content=fa http-equiv=Content-Language/>
<meta content="text/html; charset=windows-1256" http-equiv="Content-Type"/>
2. <meta content="semantic web laboratory " name="keywords"/>
3. < meta content="shahidchamran university of Ahvaz" name="description">
```

VII. Result and Future Work

In this paper, we survey of semantic web and ontology extraction for web information. This study was conducted a practical evaluation of Persian web mining for manage its information and provide new web services. Ontology Extraction for Persian web as main problemin the use of this mass repository is determined.

However, Persian web is consist of mass data but we don't have define , like a lack of specific vocabulary in different domains and computing weak support web standards Persian useful results. In this study, we first focused on the development of Persian extraction, Semantic Web, Web Ontology, Web slopes of Persian. The first study of the relational database web directly transferred to the ontology.

There are different ontology because of difference in ontology, is inevitable. The Semantic Web ontology building and integration problems communicate between its faces. The application of ontology matching to detect similarities and correspondence between ontology, where is a practical solution to solve the problem of heterogeneity. Here's a basic comparative methods have been introduced. Finally, the proposed method on a specific domain we have studied in the Persian web.

References

[1] B. Omelayenko ,and D. Fensel, "A Two-Layered Integration Approach for Product Information in B2B E-Commerce", K. Madria ,and G. Pernul (Editors), In *Proceedings of the Second International Conference on Electronic Commerce andWeb Technologies (EC WEB-2001)*, Springer, Vol. 2115, pp. 226-239, Münich, Germany, 4-6 September 2001.

[2] M. Paolucci, K. Sycara, T. Nishimura ,and N. Srinivasan, "Toward a Semantic Web E-Commerce", W. Abarmowicz, G. Klein (Editors), In the proceeding of 6th the international conference on Business, Information and Systems (BIS) , pp. 1-9, Colorado, Springs, USA, 4-6 June 2003.

[3] S. L. Huang ,and F. R. Lin, "The design and evaluation of an intelligent sales agent for online persuasion and negotiation," Electronic Commerce Research and Applications, vol. 6, pp. 285–296, 2007.

[4] B. VijayaLakshmi, A. GauthamiLatha, Y. Srinivas , and K.Rajesh, " Perspectives of Semantic Web in E- Commerce", International Journal of Computer Applications (IJCA), Vol. 25, No.10, July 2011.

[5] T. Rasovic, D. Milosevic, Z. Zoric ,and M. Milosevic, " Enhancing E-Mail Marketing by Semantic Addressing", TEM Journal, Vol. 1, N. 3, pp. 131-135, 2012.

[6] V. Kashyap, "Design and Creation of Ontologies for Environmental Information Retrieval", The 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99), Banff, Canada, October 1999.

[7] L. Stojanovič, N. Stojanovič, and R.Volz, "Migrating Data-Intensive Web Sites into the Semantic Web", In Proceeding of 2002 ACM Symposium in Applied Computing, Madrid, Spain, pp. 1100-1107, Madrid, Spain, 11-14 March 2002.

[8] S. Zhou, "Relational Database Semantic Access based on Ontology", Advances in Wireless Networks and Information Systems, Lecture Notes in Electrical Engineering (LNEE), Vol. 72, pp. 537-545, Springer, Berlin, Germany, 2010.

[9] M. D'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez, and D. Guidi, "Towards a New Generation of Semantic Web Applications", IEEE Intelligent Systems, Vol. 23, No. 3, pp. 20-28, May-June 2008.

[10] M. Sabou, "Extracting Ontologies from Software Documentation: A Semi-Automatic Method and Its Evaluation" in Proc. Workshop on Ontology Learning and Population, Valencia, Spain, August 2004.

AUTHOR'S PROFILE

Hamed Zeinali
received the B.Sc. degree in computer software engineering at Erfan University of Kerman, Iran in 2008 and M.Sc. degree in computer engineering atscience and research branch, Islamic Azad University of Sirjan, Iran in 2014. He is interestedin web mining and intelligence system.

Reza Nourmandi-Pour
is faculty member at department of computer engineering, Islamic Azad University, Sirjan, Iran, August 2010 – Preset. He is interested in computer hardware, digital circuit test, embedded memory test and interconnection test.

Arash Azizi Mazreah
received the B.S. degree in computer hardware engineering and M.S. degree in computer system architecture engineering in 2005 and 2007 respectively. Currently he is a Ph.D. candidate in computer system architecture engineering at Islamic Azad University, Science and Research Branch and faculty of Islamic Azad University, Sirjan branch. His major research experiences and interests include low power digital system design, high speed SRAM, and VLSI testing and high density VLSI system design.