ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

A Survey on Detection of Mining Service Information Discovery Using SASF Crawler

R.Eswaramoorthy[1], M.Jayanthi[2]

PG Scholar, Dept. of Computer Science and Engineering, Kalaignar Karunanidhi Institute of Technology, Coimbatore, TamilNadu, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, Kalaignar Karunanidhi Institute of Technology, Coimbatore, TamilNadu, India[2]

ABSTRACT: The internet has occupies the largest place in online shopping and online transaction systems. At present the online shopping has became more popular by the service advertisements with various industries. The mining service information are effectively carried out through the mining service advertisements. During this process there exist three major issues such as heterogeneity, Ubiquity and Ambiguity. The proposed work of this paper is using Self Adaptive Semantic Focused Crawler –SASF crawler, this technique presents the efficient discovering of information, Formatting and indexing mining service information discovery. Here incorporating the technique of semantic focused crawling and ontology based learning to maintain the performance of the crawler. The idea of this paper is based on the design of an unsupervised framework for vocabulary-based ontology learning, and also a hybrid algorithm is used for matching semantically relevant concepts and metadata.

KEYWORDS: semantic focused crawler, ontology learning, service advertisement, service information discovery.

I.INTRODUCTION

The internet has becoming the most largest unstructured database for accessing information over the documents. And also internet has became the largest market place in the world and online advertising is very popular with numerous industries, where mining service advertisements are effective carriers of mining service information. Customers are able to browse more number of products with differranges and get service through various service advertisements. Users can buy a product with the service advertisements through online transaction systems. There occur three issues during the service advertisements (a) Heterogeneity, (b) ubiquity, and (c) Ambiguity.

A. Heterogeneity

Heterogeneity has been proposed to retrieve the service information efficiently from the websites. The services has been classified under various perspectives, including the ownership of service instruments, the effects of services, the nature of the service act, delivery, demand and supply, and so on.

B.ubiquity

Service advertisements can be registered by service providers through various service registries. All the services which are registered under one service provider will be geographically distributed over the Internet.

C.Ambiguity

A Web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. Usually for indexing, crawler-based engines consider much more factors than those they can find on the web pages. The crawler puts every web page into an index and look for other pages in the index that are linking to the current web page, The query which the user requested is present on the links will be in some other directories under related categories, These “off-the-page” part has more weight when the page is evaluated by a crawler-based engine. The web page developer can increase the web page relevance for keywords by adjusting the corresponding areas of the HTML code, user still have much less control over other related pages in the internet that are linking to the user. The major challenges in the web crawler are (a) Collaborative web crawling,(b)Crawling the deep web,(c)Crawling the multimedia content,(d)Future directions.

(a)Collaborative web crawling

Crawling node is responsible for a specific portion of the web. The goal is to collect web pages about specified geographic locations, by considering features like URL address of page, content of page, extended anchor text of link, and others. Later, various evaluation criteria to qualify the performance of such crawling strategies have been proposed. More precisely, features like URL address of page and extended anchor text of link are shown to yield the best overall performance for the geographically focused crawling.

(b)Crawling the deep web

Several form query languages (e.g., DEQUEL) have been proposed that, besides issuing a query, also allow extracting structured data from result pages. The Sitemap Protocol (first developed, and introduced by Google in 2005) and mod oai are mechanisms that allow search engines and other interested parties to discover deep web resources on particular Web servers. Both mechanisms allow Web servers to advertise the URLs that are accessible on them, thereby allowing automatic discovery of resources that are not directly linked to the surface Web.

(c)Crawling the multimedia content

Web is now multimedia platform if there are images, audio, videos those are integral part of web pages. There will be apparent copyright issues, more resources required from the crawler (ex: Bandwidth, Storage place), more complicated duplicate resolving and revisiting policy are the most important challenges in the multimedia crawling environment.

(d)Future directions

Collaborative crawling, mixed push model, understanding site structure, deep web crawling, media content crawling, social network crawling are the challenges in the future directions of the web crawler.

II.RELATED WORK

In this section we introduced the mining service information with the SASF crawler. Finally we give a rapid analysis of the XML, HTML, RDF, Ontology learning.

(a)XML

XML is used in many aspects of web development, often to simplify data storage and sharing. With XML, data can be stored in separate XML files. This is the way to concentrate on using HTML/CSS for display and layout, and be sure that changes in the underlying data will not require any changes to the HTML. XML data is stored in plain text format. This provides a software and hardware independent way of storing data. This makes it much easier to create data that can be shared by different applications. One of the most time-consuming challenges for developers is to exchange data between incompatible systems over the Internet. Exchanging data as XML greatly reduces this complexity, since the data can be read by different incompatible applications. There are some drawbacks which have hindered it from gaining widespread use since its inception. There are no XML browsers on the market yet. Thus, XML documents must either be converted into HTML before distribution or converting it to HTML on-the-fly by middleware. Barring translation, developers must code their own processing applications. XML isn't about display -- it's about structure. This has implications that make the browser question secondary. So the whole issue of what is to be displayed and by what means is intentionally left to other applications

(b)HTML

HTML, or Hypertext Markup Language, is used to create web pages. Site authors use HTML to format text as titles and headings, to arrange graphics on a webpage, to link to different pages within a website, and to link to different websites. HTML is a set of codes that a website author inserts into a plain text file to format the content. The creator inserts HTML tags, or commands, before and after words or phrases to indicate their format and location on the page. It can create only static and plain pages so if we need dynamic pages then HTML is not useful. Need to write lot of code for making simple webpage. Security features are not good in HTML. If we need to write long code for making a webpage then it produces some complexity.

(c)RDF

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats. It is also used in knowledge management applications. The RDF model is based on a simple idea, but it has problems that make it unnecessarily complicated, thus decreasing its value. These problems can be divided into three categories:

- The existence of nodes that have no name.
- Problems associated with the literals.
- The lack of a unique concept of the node.

Due to various drawbacks in the above versions, here ontology learning and semantic focused crawler has been used.

(d)ONTOLOGY

Ontology learning (ontology extraction, ontology generation, or ontology acquisition) is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts from a corpus of natural language text, and encoding them with an ontology language for easy retrieval. As building ontologies manually is extremely labor-intensive and time consuming, there is great motivation to automate the process. Typically, the process starts by extracting terms and concepts or noun phrases from plain text using linguistic processors such as part-of-speech tagging and phrase chunking. Then statistical or symbolic techniques are used to extract relation signatures, often based on pattern-based or definition-based hyponyms extraction techniques.

A (Semantic Web) **vocabulary** can be considered as a special form of (usually light-weight) ontology, or sometimes also merely as a collection of URLs with an (usually informally) described meaning.

To solve a problem using ontology, Define an ontology to be a set of concepts *C* and relationships *R*. The relationships in *R* can be either taxonomic or non-taxonomic. For example, A simple college ontology consisting of a set of concepts
Ccollege = {Person, Faculty, Staff, Student, Department, Project, Course}, and a set of relationships.
Runiv={Department_Of(Person, Department), Member_Of(Person, Project), Instructor_Of(Course, Person), Superclass_Of(Faculty, Person), Superclass_Of(Staff, Person), Superclass_Of(Student, Person)}.

*Superclass_Of*represents the taxonomic relationship while the rest are not. With this definition, the instances of ontology refer to the instances of its concepts and relationships. If each concept instance exists in the form of a Web page, a relationship instance will then exist in the form of a Web page. On the other hand, if each concept instance exists in the form an HTML element, a relationship instance will then exist in the form of an HTML element pair. This alternative view is usually adopted in Web extraction research. It is noted that other forms or hybrid forms of concept instances may also exist for some Websites.

III.ONTOLOGY-BASED WEB MINING

Overview of Web Mining

Web mining refers to the discovery of knowledge from Web data that include Web pages, media objects on the Web, Web links, Web log data, and other data generated by the usage of Web data. Web mining is classified into: (a) *Web content mining,* (b) *Web structure mining* and (c) *Web usage mining.* Web content mining refers to mining knowledge from Web pages and other Web objects. Web structure mining refers to mining knowledge about link structure connecting Web pages and other Web objects. Web usage mining refers to the mining of usage patterns of

Web pages found among users accessing a Website. Among the three, Web content mining is perhaps studied most extensively due to the prior work in text mining. The traditional topics covered by Web content mining include:

Web page classification

This involves the classification of Web pages under some pre-defined categories that may be organized in a tree or other structures.

Web clustering

This involves the grouping of Web pages based on the similarities among them. Each resultant group should have similar Web pages while Web pages from different resultant groups should be dissimilar.

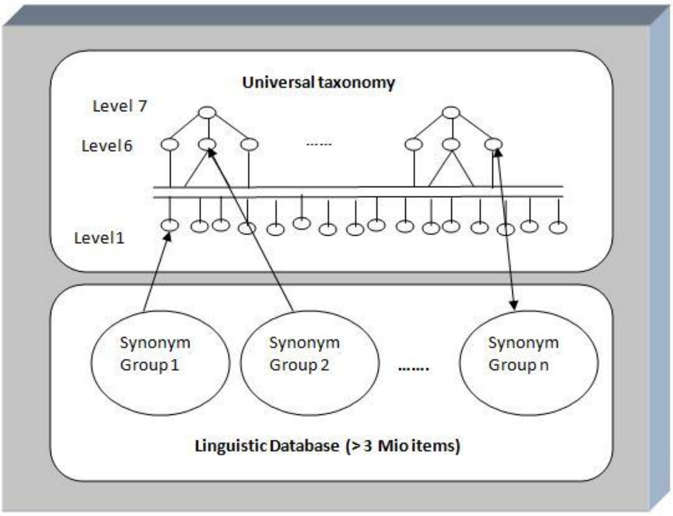
Web extraction

This involves extracting HTML elements, term phrases, or tuples from Web pages that represent some required concept instances, e.g., person names, location names, book records, etc.

Web Mining and Ontologies

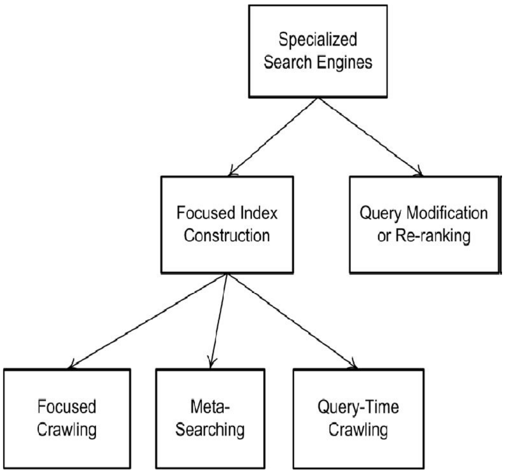
In all the above types of Web mining, ontologies can be applied in the following two general approaches:
If both ontology and the instances of ontology entities are known, then it usually applies to cases where instances of ontology have been identified among the input Web data. With this additional data semantics, Web mining techniques can discover knowledge that is more meaningful to the users. For example, ontology-based Web clustering can use HTML elements corresponding to concept instances as features to derive more accurate clusters. If Web pages are concept instances, ontology-based Web structure mining can derive linkage pattern among concepts from Web pages for Website design. The example given below shows the concept of ontology.

Words or expressions (groups of words such as "European Union", "Enterprise Search Engine" etc.) present in the text are matched against InfoCodex's linguistic database, which comprises more than 3 million words/expressions which are structured into cross-lingual synonym groups. These synonym groups point to a node in a universal taxonomy (ontology) characterizing the meaning of the matched synonym. This language-independent information is then used to extract the content of the document.



The central multi-lingual knowledge base (linguistic database) is the result of combining and harmonizing knowledge repositories from established works such as the WordNet, the International Classification for Standards (ICS), financial taxonomies and many other sources.

If only ontology is available as input semantic structures, then Ontologies can also be used asbackground semantic structures for Web mining. For example, instead of categorizing Web pages into categories, ontology-based Web page classification may classify Web pages as concept instances and Web page pairs as relationship instances. This allows Web pages to be searched using more expressive search queries involving search conditions on concepts and/or relationships. In ontology-based Webextraction, one may address the problem of extracting both HTML elements as concept instances and finding related pairs of HTML elements.


Specialized Search Engines
Focused Index Construction
Query Modification or Re-ranking
Focused Crawling
Meta-Searching
Query-Time Crawling

IV.LITERATURE SURVEY

Hai dong has proposed a Focused Crawling for Automatic Service Discovery, Annotation and Classification in Industrial Digital Ecosystems to deal with the pre-existing service information that has becomes a crucial issue in Digital Ecosystems. This can be solved by presenting a conceptual framework for a semantic focused crawler, with the purpose of automatically discovering, annotating and classifying the service information with the Semantic Web technologies.

FarookhKhadeerHussain has proposed an A framework for discovering and classifying ubiquitous services in digital health ecosystems to find the difficulties for a service consumer to precisely and quickly retrieve a service provider for a given health service request to solve the framework that incorporates with the technology of semantic focused crawler and social classification.

Farias Pereira has proposed the A Model of Service Classification Based on Consumer Needs to find the difficulties that customers face in the online shopping such as Most service classifications do not consider the client's needs as a crucial parameter either to the quality of the service or to the success of the strategy and this can be solved by using classifying services into categories which can facilitate the understanding of the client's needs and generate support for the development of strategies. Efforts should be made in order to verify the model's usefulness and examine the relations among the proposed dimensions based on the research developed with clients of different services.

Swati Ringe, has proposed the technique called Ontology Based Web Crawler and finds the problems that Users must either browse through a hierarchy of concepts to find the information they need or submit a query to a Search Engine and wade through hundreds of results most of them irrelevant and eliminate those irrelevant items by makes use of Semantics which helps to downloaded only relevant pages. Semantics can be provided by ontologies.

Robert D. Winsora has described in Differentiating goods and services retailing using form and possession utilities that The ‘‘goods–services continuum’’ provides little clarification as to the issues of retail classification or strategy development and it will be solved by based upon the utilities that is provided to consumers by Differentiating goods and services retailing using form and possession utilities.

V.CONCLUSION

From the survey, the concept of ontology learning based semantic focused crawler is available over the internet to solve the issues of heterogeneity, ubiquity and ambiguity.Preprocessing helps to process the contents of each concept in theontology before matching the metadata. Adaptive crawling and extraction is to download Web pages from the Internet at one time, and to extract the required information from the downloaded Web pages, according to the mining service metadata schema. Semi supervised based ontology string matching algorithm automatically obtains the statistical data from the Web pages, in order to compute the semantic relevance between a service description and a concept description of a concept. Performance measure evaluate the performance of the proposed approach based on the SASF crawler with semi supervised ontology method with the previous approaches.

REFERENCES

1. H. Dong and F. K. Hussain, "Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2106–2116, Jun. 2011,

2. Hai Dong, FarookhKhadeerHussain, Elizabeth Chang, A framework for discovering and classifying ubiquitous services in digital health ecosystems, *Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology, Perth, WA 6845, Australia*

3. Swati Ringe#, Nevin Francis, PalanawalaAltaf H.S.A., Ontology Based Web Crawler,
Swati Ringe, Fr.Conceicao Rodrigues College Of Engineering, Fr.Agnel Ashram, BandStand, Bandra-w, Mumbai-400050

4.fariaspereira , Suzana Carla, A Model of Service Classification Based on Consumer Needs, POMS-2001 ,March 30 – April 2, 2001 , Orlando Fla

5. Robert D. Winsora,*, Jagdish N. Shethb, Chris Manolisc, Differentiating goods and services retailing using formand possession utilities, R.D. Winsor et al. / Journal of Business Research 57 (2004) 249–255

6. J. Rennie and A. McCallum, "Using reinforcement learning to spider the Web efficiently," in *Proc. 16th Int. Conf. Mach. Learning (ICML'99)*, Bled, Slovenia, 1999, pp. 335–343.

7. M. Ruta, F. Scioscia, E. Di Sciascio, and G. Loseto, "Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 731–739, Nov. 2011.

8.S. Runde and A. Fay, "Software support for building automation requirements engineering—An application of semanticweb technologies in automation," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 723–730, Nov. 2011.

9. I. M. Delamer and J. L. M. Lastra, "Service-oriented architecture for distributed publish/subscribe middleware in electronics production,"*IEEE Trans. Ind. Informat.*, vol. 2, no. 4, pp. 281–294, Nov. 2006.

10. H. L. Goh, K. K. Tan, S. Huang, and C. W. d. Silva, "Development of Bluewave: A wireless protocol for industrial automation," *IEEE Trans.Ind. Informat.*, vol. 2, no. 4, pp. 221–230, Nov. 2006.

11. P. Plebani and B. Pernici, "URBE:Web service retrieval based on similarity evaluation," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1629–1642, Nov. 2009.

12. P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell.Res.*, vol. 11, pp. 95–130, 1999.

13. H.-T. Zheng, B.-Y.Kang, and H.-G. Kim, "An ontology-based approach to learnable focused crawling," *Inf. Sciences*, vol. 178, pp. 4512–4522, 2008.

14. T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, pp. 199–220, 1993.

BIOGRAPHY

R.Eswaramoorthy is a PG Scholar of Kalaignar Karunanidhi Institute of Technology, Coimbatore.He received B.E Computer Science and Engineering Degree in Kalaignar Karunanidhi Institute of Technology, Coimbatore. He is doing Project in the field of Web Mining.

Mrs.Jayanthi M is working as Assistant Professor in Kalaingnar Karunanidhi Institute of Technology. She had received his Bachelor of Technology from Anna University Chennai, Master of Engineering from Anna University Coimbatore. Her field of Interest is Network Security and Data Mining.