www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 9 Sept., 2013 Page No. 2659-2662

A Study on Semantic Web Mining And Web Crawler

S.Balan[1], Dr.P.Ponmuthuramalingam[2]

[1]Ph.D Research Scholar, Department of Computer Science,
Government Arts College, Coimbatore, Tamilnadu, India.
balan.sethuramalingam@gmail.com

[2]Associate Professor & Head, Department of Computer Science,
Government Arts College, Coimbatore, Tamilnadu, India
Ponmuthu.journal@gmail.com

Abstract: ***The main purpose of the research is to study and understand the concepts of semantic web mining and web crawlers. Everyday tens to hundreds of millions of web information are generated and they answer tens of millions of queries. The most important thing is search engine to search the performance, quality of the results and ability to crawl, and index the web efficiently. The primary goal is to provide high quality search results over a rapidly growing World Wide Web in semantic web mining and to download the information from web crawling is needed. This research work aimed at how the information's are extracted from the web using crawler method and studying the research areas of semantic web mining.***

Keywords: Web Crawling Techniques, Semantic Web Mining, Ontology Learning, Challenges

1. INTRODUCTION

The World Wide Web is officially defined as a "wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents." In simpler terms, the Web is an Internet-based computer network that allows users of one computer to access information stored on another through the world-wide network called the Internet. The Web's implementation follows a standard client-server model.

In this model, a user relies on a program (called the client) to connect to a remote machine (called the server) where the data is stored. Navigating through the Web is done by means of a client program called the browser, e.g., Netscape, Internet Explorer, Firefox, etc. Web browsers work by sending requests to remote servers for information and then interpreting. The returned documents written in HTML and laying out the text and graphics on the user's computer screen on the client side.

The two fast-developing research areas Semantic Web and Web Mining build both on the success of the World Wide Web (WWW). They complement each other well because they address one a new challenge posed by the great success of the current WWW: The nature of most data on the Web is so unstructured that they can only be understood by humans, but the amount of data is so huge that they can only be processed efficiently by machines [7].

The Semantic Web addresses the first part of this challenge by trying to make the data (also) machine understandable, while Web Mining addresses the second part by (semi-)automatically extracting the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions. This paper is organized into four sections. Section-1 contains introduction of World Wide Web, Section-2 contains overview of the web crawling method, Section-3 contains presents overview of Semantic Web Mining and Section-4 includes conclusion and future work while references are shown in the last section.

2. WEB CRAWLING

A Web crawler is a program that automatically traverses the Web's hyperlink structure and downloads each linked page to a local storage. Crawling is often the first step of Web mining or in building a Web search engine. Although conceptually easy, building a practical crawler is by no means simple. Due to efficiency and many other concerns, it involves a great deal of engineering. There are two types of crawlers: universal crawlers and topic crawlers [7].

A universal crawler downloads all pages irrespective of their contents, while a topic crawler downloads only pages of certain topics. The difficulty in topic crawling is how to recognize such pages. Web crawler is an Internet that systematically browses the World Wide Web, typically for the purpose of Web indexing. It also called as Web spider, an ant, an automatic indexer, Web Scutter.
Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search

them much more quickly [19]. The figure 1 shows that the architecture of a web crawler.

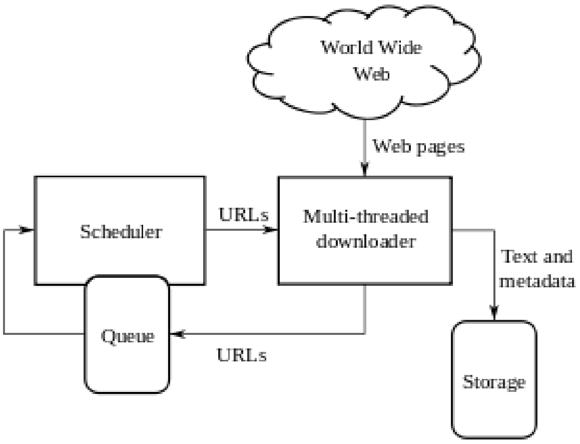

Figure 1: Web Crawler Architecture

2.1 WEB CRAWLING TECHNIQUES

A. Distributed Crawling

Indexing the web is a challenge due to its growing and dynamic nature. As the size of the Web is growing it has become imperative to parallelize the crawling process in order to finish downloading the pages in a reasonable amount of time. A single crawling process is insufficient for large – scale engines that need to fetch large amounts of data rapidly.

When a single centralized crawler is used all the fetched data passes through a single physical link. Distributing the crawling activity via multiple processes can help build a scalable, easily configurable system, which is fault tolerant system. Splitting the load decreases hardware requirements and at the same time increases the overall download speed and reliability. Each task is performed in a fully distributed fashion, that is, no central coordinator exists [1].

B. Focused Crawling

A general purpose Web crawler gathers as many pages as it can from a particular set of URL's, Where as a focused crawler is designed to only gather documents on a specific topic, thus reducing the amount of network traffic and download. The goal of the focused crawler is to selectively seek out pages that are relevant to a pre-defined set of topics.

The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the web.

This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date. The focused crawler has three main components: a classifier, which makes relevance judgments on pages, crawled to decide on link expansion, a distiller which determines a measure of centrality of crawled pages to determine visit priorities, and a crawler with dynamically reconfigurable priority controls which is governed by the classifier and distiller [6].

2.2 WORKING OF A WEB CRAWLER

Web crawlers are an essential component to search engines; running a web crawler is a challenging task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers, which are all beyond the control of the system [14]. The figure 2 shows that the working of a web crawler.

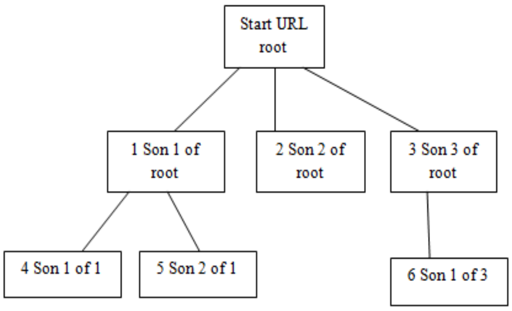

Figure 2: Web Crawler Working

Web crawling speed is governed not only by the speed of one's own Internet connection, but also by the speed of the sites that are to be crawled. Especially if one is a crawling site from multiple servers, the total crawling time can be significantly reduced, if many downloads are done in parallel. Despite the numerous applications for Web crawlers, at the core they are all fundamentally the same. Following is the process by which Web Crawler's work [14].

1. Download the Web page.
2. Parse through the downloaded page and retrieve all the links.
3. For each link retrieved, repeat the process.

2.3 WEB CRAWLING CHALLENGES

The basic web crawling algorithm is simple: Given a set of seed Uniform Resource Locators (URLs), a crawler downloads all the web pages addressed by the URLs, extracts the hyperlinks contained in the pages, and iteratively downloads the web pages addressed by these hyperlinks. Despite the apparent simplicity of this basic algorithm, web crawling has many inherent challenges [6]:

i). Scale. The World Wide Web is very large and continually evolving. Crawlers that seek broad coverage and good freshness must achieve extremely high throughput, which poses many difficult engineering problems.

ii). Content selection tradeoffs. Even the highest-throughput crawlers do not purport to crawl the whole web, or keep up with all the changes. Instead, crawling is performed selectively and in a carefully controlled order..

iii). Social obligations. Crawlers should be "good citizens" of the web, i.e., not impose too much of a burden on the web sites they crawl.

3. SEMANTIC WEB MINING

"The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines –not just for display purposes, but for using it in various applications. "It is defined by Tim Berners-lee.

Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. The idea is to improve the results of Web Mining by exploiting the new semantic structures in the Web. Furthermore, Web Mining can help to build the Semantic Web [3].

The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. Berners-Lee suggests enriching the Web by machine-process able information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still return too often too large or inadequate lists. Machine-process able information can point the search engine to the relevant pages and can thus improve both precision and recall. To reach this goal the Semantic Web will be built up in different levels: Unicode/Unified Resource Identifiers, XML, RDF, Ontologies, logic, proof, trust [5].

3.1 Extracting Semantics from the Web

The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way. Web Mining can help to learn definitions of structures for knowledge organization (e. g., ontologies) and to provide the population of such knowledge structures.

i) Ontology Learning

Extracting ontology from the Web is a challenging task. One way is to engineer the ontology by hand, but this is quite an expensive. The expression towards Semantic Web Mining 269 Ontology Learning was coined for the semi-automatic extraction of semantics from the Web in order to create ontology [10]. There, machine learning techniques were used to improve the ontology engineering process.

Ontology learning exploits a lot of existing resources, like text, thesauri, dictionaries, databases and so on. It combines techniques of several research areas e. g., from machine learning, information retrieval (cf. [12]), or agents [18], and applies them to discover the 'semantics' in the data and to make them explicit.

ii) Mapping and Merging Ontologies

With the growing usage of ontologies, the problem of overlapping knowledge in a common domain occurs more often and becomes critical. Domain-specific Ontologies are modeled by multiple authors in multiple settings. These ontologies lay the foundation for building new domain-specific ontologies in similar domains and extending multiple ontologies from repositories. The process of ontology merging takes as input two (or more) source ontologies and returns a merged ontology based on the given source ontologies. Manual ontology merging using conventional editing tools without support is difficult, labor intensive and error prone. Therefore, several systems and frameworks for supporting the knowledge engineer in the ontology merging task have recently been proposed [10, 8, 13, and 15].

iii) Instance Learning

It is probably reasonable to expect users to manually annotate new documents to a certain degree, but this does not solve the problem of old documents containing unstructured material. In any case we cannot expect everyone to manually mark up every produced mail or document, as this would be impossible. Moreover some users may need to extract and use different or additional information from the one provided by the creator. For the reasons mentioned above it is vital for the Semantic Web to produce automatic or semi-automatic methods for extracting information from Web-related documents, either for helping in annotating new documents or to extract additional information from existing unstructured or partially structured documents.

3.2 Mining the Semantic Web

As the Semantic Web enhances the first generation of the WWW with formal semantics, it offers a good basis to enrich Web Mining: The types of (hyper) links are described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, allowing her to apply mining techniques which require more structured input.

i) Semantic Web Content and Structure Mining

In the Semantic Web, content and structure are strongly interred twined. Therefore, the distinction between content and structure mining vanishes. However, the distributions of the semantic annotations may provide additional implicit knowledge. We discuss now first steps towards semantic Web content/structure mining.

An important group of techniques which can easily be adapted to Semantic Web content/Structure Mining are the approaches discussed as Relational Data Mining (formerly called Inductive Logic Programming (ILP); see [17] for an introductory collection of articles). Relational Data Mining looks for patterns that involve multiple relations in a relational database. It comprises techniques for classification, regression, clustering, and association analysis. It is quite straightforward to transform the algorithms so that they are able to deal with data described in RDF or by ontologies. There are two big scientific challenges in this attempt. The first is the size of the data to be processed (i. e., the scalability of the algorithms), and the second is the fact that the data are distributed over the Semantic Web, as there is no central database server.

ii) Semantic Web Usage Mining

Usage mining can also be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of ontology. Semantic Web usage mining can for instance be performed on log files which register the user behavior in terms of ontology. A system for creating such semantic log files from a knowledge portal [9] has been developed at the AIFB [16]. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

4. CONCLUSION & FUTURE WORK

This research work is concerned with the study and analysis of web crawling methods, techniques, Semantic web Mining and research challenges in the web. To extract information from the web, crawling techniques are discussed in this paper. It can be further extended in the following directions such as different types of Web Crawling methods, exploiting the information based on semantics from the web and learning the domain ontologies, applications of semantic web mining and its features.

REFERENCES

1) Baldi, Pierre, Modeling the Internet and the Web: Probabilistic Methods and Algorithms, willey Publications, pp-158, 2003.
2) Bing Liu, Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, ACM Computing Classification, Springer (1998).
3) Bettina Berendt, Andreas Hotho, and Gerd Stumme, Towards Semantic Web Mining, Horrocks and J. Hendler (Eds.): Springer-Verlag Berlin Heidelberg ISWC 2002, LNCS 2342, pp. 264–278.
4) Chakrabarti, Soumen., Mining the Web: Analysis of Hypertext and Semi Structured Data, Morgan Kaufmann Publications, Elsevier, 2003
5) Chakrabarti S., Data mining for hypertext: A tutorial survey, SIGKDD Explorations, 1:1–11, 2000.
6) Christopher Olston and Marc Najork., Web Crawling, Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010) 175–246
7) Dhiraj Khurana, Satish Kumar., Web Crawler: A Review IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012 ISSN (Online): 2231 –5268
8) Hans Chalupsky., Ontomorph: A translation system for symbolic knowledge. In Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000), pages 471–482, 2000.
9) Hotho, Maedche, Staab, and Studer. SEAL-II — The Soft Spot Between Richly Structured and Unstructured Knowledge. Journal of Universal Computer Science (J.UCS), 7(7):566–590, 2001.
10) Hovy E.H., Combining and standardizing large-scale, practical ontologies for machine translation and other uses, In Proc. 1st Intl. Conf. on Language Resources and Evaluation (LREC), Granada, 1998.
11) Maedche and Staab S., Ontology learning for the semantic web. IEEE Intelligent Systems, 16(2):72 –79, 2001.
12) Maedche., Ontology Learning for the Semantic Web, Kluwer, 2002.
13) McGuinness D, Fikes R, Rice J, and Wilder S., An environment for merging and testing large ontologies, In In the Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), pages 483–493, Breckenridge, Colorado, USA, 2000.
14) Monica Peshave Kamyar Dezhgosha., How Search Engines Work And A Web Crawler Application, University of Illinois at Springfield, IL 62703,1-15, 2002
15) Noy N and Musen M., Prompt: Algorithm and tool for automated ontology merging and alignment. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 450–455, Austin, Texas, 2000.
16) Oberle D., Semantic Community Web Portals - Personalization, Studienarbeit, Universit¨at Karlsruhe, 2000.
17) Saso Dzeroski and Nada Lavrac, editors. Relational Data Mining. Springer, 2001.
18) Williams A Band Tsatsoulis C., An instance-based approach for identifying candidate ontology relations within a multi-agent system, In Proceedings of the First Workshop on Ontology Learning OL'2000, Berlin, Germany, 2000. Fourteenth European Conference on Artificial Intelligence.
19) www.wikipedia.com/web crawler [Accessed on 02.07.2013]