International Journal of Computer Applications Technology and Research
Volume 3– Issue 6, 345 - 348, 2014, ISSN: 2319–8656

Data Retrieval from the Discrete Media Available on Web and its Diversified Employment Perspectives

Kulvinder Singh
Department of Computer Science and Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Sanjeev Dhawan
Department of Computer Science and Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Pratibha
Department of Computer Science and Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Abstract: Internet has provided its users with an efficient enough and quality-driven source of information in form of the interconnected web of documents. Apart from that it also facilitates its users with some of its really interesting and interactive features served in the package of all these prevailing commercial and social networking websites like facebook, Twitter, flipcart, amazon.com etc. In that particular concern of availability of information and processing it to derive some more interesting results this paper has been written to present the review of some already performed and worth mentioning contributions made by the experts in the field of information on web.

Keywords: Information Retrieval; Query Processing System; Corpora; Focused Crawling; Precision; Recall; PageRank

1. INTRODUCTION

Information we retrieve comes from a huge source of it that we call the World Wide Web. To make this retrieval to acquire its most accurate and relevant version becomes the responsibility of the programs performing it known as Web Crawlers. Their designs equip them to do all the required tasks related to information retrieval, whether it is the link crawling strategy (Breadth First or Depth First approach) or their inside algorithm which teaches them to schedule the resultant links or pages as per their prioritized and calculated ranks. This retrieved information could be mined further and utilized in any possible ways. The way we get that information can possibly suggest how proficient the crawler was while it got the required information through the links. Such information can help to improve the crawling strategies over time. It can also be implemented for creating the relationship graphs on the web. We can also get to know about the interests of people and can help them direct to their desired products and services which would enhance the interface quality as per the user's perspective. The paper is organized in 3 sections, a general introduction of the concerned topic has been given in the current section, the next section provides the brief review of the previously done related work in the respective field, third section presents the analysis and derived conclusions and the future research perspectives are provided in the last section of this paper.

2. RELATED WORK

Gupta *et al.* [1] worked upon the information retrieval system for local databases by searching the web both syntactically and semantically. They created both kinds of information retrieval systems and then compared them based upon Precision and Recall. Different databases were created for each of them, out of those comparisons it was inferred that semantic web was found to be more futuristic. Based upon the results they created an ontology-based focused Information Retrieval (IR) system for learning styles of autistic people. Such ontology can be designed for other disabilities as well. They suggested that using a larger warehouse and implementing search engine combined with properties of various Information Retrieval (IR) systems could draw some better results. As a succession to this data retrieval approach Wan *et al.* [2] worked upon-"URL Assignment Algorithm of Crawler in Distributed System Based on Hash". They described the function of every module and established some rules which parallel crawlers must follow to maintain the equilibrium of load and robustness of system when they perform the simultaneous search on the web. They further designed and implemented a new URL assignment algorithm based on hash for partitioning the domain to crawl and discuss the complete decentralization of every task. They presented an improved Hash method used by the distributed crawler system to assign URL and decentralize the different tasks of crawlers to guarantee the load balancing of the system. As a future concern of research work they stressed upon the detailed study of the scalability of the system and the behavior of its components. A further refinement has been made by Lim *et al.* [3] in the concerned arena by designing a commercial search engine based Query Processing System (QPS) which is capable of answering 5 million user queries against over 6.5 million web pages per day. For making it really fast, they employed more than one server to work in parallel even to solve a single query. They used to implement such server clusters by connecting them via high-speed LANs. Regarding memory requirements for such fast and efficient QPS, they used hierarchical 4-level cache; in its topmost level they stored the recent search result pages and more query results in its lower levels. Such QPS with its multi-level cache could save 70% of server cost for query processing. In the allied concern of fetching a prioritized list of search results, Rafiq *et al.* [4] had designed a new URL ordering algorithm which overcame the flaws of conventional PageRank algorithm. They worked upon some of the determining factors about the performance of web crawlers and then created their algorithm.

For producing an efficient site ordering, they designed a formula for computing the site score:

Final Rank = 0.223(Public popularity score using server logs) + 0.2387(Site updating frequency) + 0.35(Content similarity).

Castillo *et al.* [5] had made a further progression to the ranking task. While working on a sample web (Chilean web) they used the breadth first search approach to traverse all the relevant links and from that information they managed to perform the ranking job. Along with this they also had a serious focus on parallelism encountered while making multiple page requests simultaneously in the same session. Major concerns of their work involved (i) The time interval (w) between successive connections to same website, initially they put w=60 seconds but then reduced it to 15, (ii) The keep-alive header in HTTP/1-1 so as to download k (no. of pages downloaded per session) ≥ 1 which indeed required $w \leq 15$, and (iii) The number of simultaneous requests (r) which depends completely upon the availability of relevant website. They worked upon discovering the algorithms for long-term and short-term scheduling and found that for long-term scheduling, a strategy named "length" that focuses upon the longevity of any website was decided to be efficient enough. They also discovered a few issues: a.) Page detection based on higher page ranks creates inefficiencies with the advancement in the search because some new websites and pages might keep on getting discovered and disturb the previously accomplished rankings. b.) Waiting time for retrieving websites gets amortized if we keep k>>1 for page retrieval. The presence of duplicate and near duplicate web documents on the web creates additional overhead for the search engines that critically affects their performance. In that arena Narayana *et al.* [6] presented their work in the paper "A Novel and Efficient Approach for Near Duplicate Page Detection in Web Crawling". Their proposed method used to first extract the keywords from the crawled pages and then compute the similarity scores between the pages. Documents that had a similarity score greater than some defined threshold were considered as near duplicates and rejected. This in-turn reduced the memory requirements for repositories and improved search engine quality. In the similar concern Sharma *et al.* [7] discovered some important issues regarding the lack of topic relevancy in the information retrieved in response to the quoted keywords. They paid the major concern of their work to the information retrieval evaluation measures-Precision and Recall. They also proposed the formulae to calculate these parameters. They figured it out how to model the user's efforts using gain function and discount function of their formulation called Discount Cumulative Gain (DCG). They introduced some future research arenas like predicting the ways to determine if the retrieved pages are the result of exhaustive search from the web and how to uniformly sample web pages on a website if it does not have complete list of web pages. They also stated the need to design more effective and precised algorithms that could detect the duplicity of documents available on web. A furthermore research in the field of detecting and analyzing irrelevancy in information retrieval has been made by Moshchuk *et al.* [8] who accomplished it through large-scale, longitudinal study of the web. They worked upon both the drive-by download malicious executables and the scripted page content that is capable of disrupting the end user's system by playing with his browser settings. They conducted such a crawl in May 2005, involving 18 million URLs, and from that they discovered that spyware was embedded in 13.4% of the 21,200 executables they identified. Additionally they discovered scripted "drive-by download" attacks in some other 5.9% of the web content processed by them. They also studied the frequency pattern of this spyware detection. They conducted the same experiment later in October 2005 and detected a substantial reduction in the drive-by download attacks. They made their initial studies by actually sniffing into the Internet connection at the University of Washington. It was a three step analysis i.) determining the presence of some executable software in the content extracted from web, ii.) downloading, installing and executing that software within a virtual machine and iii.) analyzing the post installation spyware infections made by that software. The prime aspect of retrieving information from a web of documents is to retrieve the most valuable web information by utilizing the least system resources and filter the useless information to the maximum extent. To achieve that Gao *et al.* [9] designed a special crawler for Internet forums. Different from General Crawler and Focused Crawler, it could get structured information directly. This crawler adopted template based processing method which is actually made to use regular expressions to extract structured information. Their forum crawler also proved to be suitable for news and blog sites. It can be applied in the field of public-opinion monitor, news collection, and search of special information such as house rent or recruitment information. However, the configuration of template files is somehow the most complicated part; hence they suggested improving it as a future perspective. Proceeding in the identical research domain, Zhai *et al.* [10] presented their work in the paper "Structured Data Extraction from the Web Based on Partial Tree Alignment" in which they proposed about the problem of structured data extraction from arbitrary web pages. They gave a novel and effective technique called Data Extraction based on Partial Tree Alignment (DEPTA) via which they performed the automatic web data extraction. It was a two step technique: i) identifying individual records in a page and ii) aligning and extracting data items from the identified records. Experimental results proved the worth of this technique. The web source which provides the relevant and most reliable information becomes prior to be visited the most, working in that direction Forsati *et al.* [11] extended the traditional association rule problem by allowing a weight to be associated with each item in a transaction so as to reflect the interest/intensity of each item. They assigned a significant weight to each page based on the time spent and visiting frequency of user for that page, taking into account the degree of interest instead of binary weighting. They presented a new personalized recommendation method based on the proposed weighted association-mining technique. The experimental results showed that the Weight Association Rule (WAR) based model could significantly improve the recommendation effectiveness.

In continuation to the above recapitulation, information extraction and its usage in concern with Social Media is an equally significant and related domain of research. In that particular context, Nemeslaki *et al.* [12] explained the process of mapping business relationships using social media information. They made a study over the business ecosystems in Hungary by examining 5000 out of 15000 facebook users who publically displayed their employers. Then they depicted the complexity of connections graphically through a simulator. Also they transformed the overall graphical network into a relationship graph of employers. The more individuals it showed related to each other in the network between firms, the stronger the relationships became. While making advancement to the same context Neunerdt *et al.* [13] presented their work in the paper "Focused Crawling for

Building Web Comment Corpora". They proposed two algorithms for collecting and processing web comments in context of social blogging. The first approach proposed by them was a combination of a link-based and a content-based focusing algorithm. A relevance classifier was also combined with the link-based Online Page Importance Computation (OPIC) scoring algorithm. They compared OPIC combined with comment detection (OPIC + COMMD) and usual comment detection (COMMD) focused algorithm to the standard breadth-first search (BFIRST) crawling approach. These algorithms allowed for type-specific focused crawling to build web comment corpora. For future endeavors they proposed topic-specificity in the web comment corpus, topic classification in the relevance scoring of web pages, ontology-based corpus generation tool for further refinement and the need for a more improved and sophisticated web comment detection approach. Working on the similar perspective Agarwal [14] proposed his remarkable research work on "Prediction of Trends in Online Social Networks". He used the "directed links of following" in the social media of Twitter to determine the flow of information. This strategy indicates a user's influence on others that could decide if the topic is trendy or viral in the social network. His automated system takes raw tweets, processes them using NLP to filter out noise (spams or chats) and extract informative tweets and then mine them further to predict the trending topics in their early stages. An aggregate score for each tweet is calculated by that system. But since Twitter is really spontaneous and dynamic therefore extracting the complete Twitter graph and thence resolve all the relations for the users is impossible. So, as the future perspective he proposed to work on some more effective data structures and algorithms to deal with the dynamic Twitter graphs. To make it more reasonable Nooralahzadeh *et al.* [15] presented their work in the paper entitled "2012 Presidential Elections on Twitter - An Analysis of How the US and French Election were Reflected in Tweets". In their analysis they studied the sentiments that prevailed before and after the presidential elections, held in both US and France in the year 2012. To achieve it, they retrieved the content information from the social medium-Twitter and used the tweets from electoral candidates and the voters, collected by means of crawling during the course of election. In order to gain useful insights about the US elections, they scored the sentiments for each tweet using different metrics and performed a time series sentiment analysis for candidates and different topics retrieved as per the quoted keywords by the formula:

Number of Positive Tweets - Number of Negative Tweets/Total Number of Tweets

In addition to this, they compared some of their insights obtained from the US election with their observations from the French election. They made these observations to understand the inherent nature of elections and to bring out the influence of social media on elections.

3. CONCLUSIONS

In this paper we presented the review of the work done in the domain of information extraction, mining and processing it, in order to derive some meaningful results. It summarized the explorations made by some very proficient researchers in the respective field. The paper encloses the explorations made about the relevant arenas of information, ranging from usual information retrieval to focused crawling, derivations about information retrieval from usual commercial websites to social blogging sites, from ontology-based Information Retrieval (IR) to spyware detection embedded in scripted documents. We also presented the work done about more efficient and improved ranking algorithms, more promising Query Processing Systems (QPS), duplicity detection methods, data mining and processing from social blogging sites as well. The valuable findings, flaws and future endeavors about every referred paper are well presented in this paper.

4. SCOPE FOR FURTHER RESEARCH

For the future perspectives, we would work on the real-time data retrieval from the social networking websites like Twitter and perform the analysis of that data to derive the results regarding the trending and viral issues. We would work upon extracting and observing the social graphs of relationships on such SNSs to understand the plotting criteria and then employ those results for plotting the results obtained from the analysis of the data retrieved.

5. REFERENCES

[1] Dr. Deepak Garg and Sanchika Gupta, "A Frequent Pattern Based Approach to Information Retrieval", 2011.

[2] Yuan Wan and Hengqing Tong, "URL Assignment Algorithm of Crawler in Distributed System Based on Hash", Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference, April 2008, pp 1632-1635, 2008.

[3] Sungchae Lim and Joonsen Ahn, "A Hierarchical Cache Scheme for the Large-scale Web Search Engine", Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp 1643-1647, 2011.

[4] Sandhya and M. Q. Rafiq, "Performance Evaluation of Web Crawler", International Journal of Computer Applications® (IJCA)/International Conference on Emerging Technology Trends (ICETT), pp 43-46, 2011.

[5] Carlos Castillo, Mauricio Marin, Andrea Rodriguez and Ricardo Baeza- Yates, "Scheduling Algorithms for Web Crawling", 2010.

[6] V.A. Narayana, P. Premchand and Dr. A. Govardhan, "A Novel and Efficient Approach For Near Duplicate Page Detection in Web Crawling", 2009 IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pp 1492-1496, 2009.

[7] Dr. Deepak Garg and Deepika Sharma, "Information Retrieval on the Web and its Evaluation", International Journal of Computer Applications (0975-8887), pp 26-31 Issue 3/Volume 40, 2012.

[8] Alexander Moshchuk, Tanya Bragin, Steven D. Gribble and Henry M. Levy, "A Crawler-based Study of Spyware on the Web", Department of Computer Science & Engineering, University of Washington, pp 9-13, {anm, tbragin, gribble, levy}@cs.washington.edu.

[9] Qing Gao, Bo Xiao, Zhiqing Lin, Xiyao Chen and Bing Zhou, "A HIGH-PRECISION FORUM CRAWLER BASED ON VERTICAL CRAWLING", Proceedings of IC-NIDC, pp 362-367, 2009.

[10] Yanhong Zhai and Bing Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, pp 1614-1628, VOLUME 18/NO. 12, 2006.

[11] R. Forsati, M. R. Meybodi and A. Ghari Neiat, "Web Page Personalization Based on Weighted Association Rules", International Conference on Electronic Computer Technology, pp 1317-1321, 2009.

[12] Nameslaki, Andràs; Pocsarovszky and Kàroly, “Web crawler research methodology”, 22nd European Regional Conference of the International Telecommunications Society (ITS2011), 2011. Available at http://hdl.handle.net/10419/52173 .

[13] Melanie Neunerdt, Markus Niermann, Rudolf Mathar and Bianka Trevisan, RWTH Aachen University, “Focused Crawling for Building Web Comment Corpora”, The 10th Annual IEEE CCNC- Work-in-Progress, pp 761-765, 2013.

[14] Pranay Agarwal, Department of Computer Science and Engineering, IIT Delhi, Thesis on “Prediction of Trends in Online Social Network”, 2013.

[15] Farhad Nooralahzadeh, Viswanathan Arunachalam and Costin Chiru, “2012 Presidential Elections on Twitter - An Analysis of How the US and French Election were Reflected in Tweets”, 2013 19th International Conference on Control Systems and Computer Science, pp 240-246, 2013.