ISSN: 2321-7782 (Online)
Volume 3, Issue 1, January 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Techniques for Web Mining of Various Forms of Existence of Data on Web: A Review

Asha Joy[1]
PG Scholar, Department of Computer Science
College of Engineering Perumon
Kerala, India

Remya R[2]
Assistant Professor, Department of Information Technology
College of Engineering Perumon
Kerala, India

Abstract: ***Information extraction is known to be the task of extracting knowledge from any meaningful text automatically. It provides services to the users who retrieve the information by firing query on Internet. When huge amount of information is extracted from large corpora like Web, it is called Web mining and there has been huge interest towards it .The task of Web mining and the strategy or technique used for it depends upon the input. The input can be natural language text or documents arranged in semi structured formats such as web tables or enumerated lists. This paper presents a survey on web mining strategies used for mining different forms of existence of data on Web. It also presents an emerging technique used for data extraction from web pages called top-k web pages that describes top-k instances of any particular topic of interest which is very useful in search or fact answering systems.***

Keywords: Web mining, information extraction, data, tag path.

I. INTRODUCTION

Researchers have focused on extracting large amount of structured information from large corpora like the Web. The process of collecting such a large body of information is always tedious. The application of data mining techniques to extract knowledge from Web data is called Web mining. The huge dataset of Web data includes many different kinds of information, including, web documents data, web structure data, and user profiles data.

The emergence of web mining is due to many reasons. Firstly, Web data is a huge and important source for data mining and data warehousing. The dataset is large and still grows rapidly. Also the web page contents are much more complex compared to other text documents. They lack the standard structure nowadays demanding more efficient strategies. In addition to these, only a small portion of the content is useful or relevant. All the above mentioned challenges and reasons necessitate a research on different strategies of extracting knowledge or information from Web.

The techniques of extracting information [1], [4], [5], [8], [10] always depend on the type of input to be processed and the extracted target. The input can be unstructured text which exists in natural language or documents in semi-structured or table-like formats. Mostly it is in unstructured format and processing such texts in natural language is very difficult. Also information still exists in the form of web tables which are specified using html tags such as <tables> on html pages. Many studies have focused on getting knowledge from such formats. In addition to such formats or styles of information, there are certain web pages called top-k web pages which are intended to specify the top-k instances of any particular matter of interest. Those web pages are found to be useful in many applications such as search or fact answering systems and are also useful in adding instances to knowledgebase such as Probase. Thus top-k web pages are also considered to be important sources of input in web mining.

There are many techniques for processing and finding information from all these types of data on web and Section III describes them in detail.

II. STRATEGIES OF INFORMATION EXTRACTION

A. Mining of Data Records

Data Records[5] are regular structures of information found on the Web. It is useful to mine such data records in order to obtain useful information from the corresponding web pages. There are different existing techniques for extracting them.

One of the important strategies or techniques[5] is based on two observations about web pages. In web pages the data records that describe some similar data items are found in a particular region known as Data Region. From the tag tree, i.e , the nested structure of html tags of the web page, these data regions can be easily identified as they will be under one single parent node as shown. Thus the technique is to build the tag tree of the page first and then the data region itself is mined from the tree. Then the data records within the region can be easily obtained.

Another technique for mining data records[6] from web pages is based on tag path clustering. The data records in an HTML web page are usually visually aligned and have a similarity in their appearance. The method identifies the tag path that appears repeatedly in the tag tree of the web document. In this approach, a pair of tag path occurrence patterns known as visual signals is compared to determine whether they represent the same list of objects. The similarity measure obtained from the comparison can be used for clustering the tag paths that represents the similar objects. And the data record can be extracted by the clustering of tag paths.

B. Mining of Web tables

Web is mostly found as huge corpora of unstructured documents, but it also have data organized in structured formats such as web tables that represent relational data. The corpus of such web tables is larger and many researchers have focused on the problem of identifying useful and relevant web tables rich in information and extracting knowledge from them. One of such systems is called *WebTables system*[4] that provides an efficient technique for performing keyword search over a corpus of tables.

The corpus of html tables is large as most of them are used for content or page layout. Finding meaningful or related relations from them is important as they are rich in valuable information. The WebTables system extracts such related relations from the corpus and ranks the relations by relevance using a keyword query as input in a search-engine-style. Many ranking functions such as NaiveRank , FilterRank are used for performing the ranking of relations.

Another technique[8] of mining html tables is independent of the tree-based representation of web pages and focuses on the domain-independent information extraction. The approach makes use of the visual box model used by web browsers to display information for extracting information. The html documents when they are rendered by a browser are represented by rectangular boxes and it is called CSS(Cascading Style Sheet) box model or Visual box model. The web tables topologically form a frame in the visual box model. Firstly, the task involves identifying or locating these frames representing the tables in web documents. Then the relative spatial relationship between the logical cells of a table are recognized is identified. Meaningful information contained in all available visual relationship can then be extracted from by interpreting the obtained topological structure of the table.

C. Mining of Top-k Web pages

Top-k web pages are those web pages that describe k items of a particular topic of interest. They are rich and valuable source of information. Many recent researches have focused on extracting information from these top-k web pages in order to enrich knowledge bases for supporting many applications like fact or search answering.

There is an approach[1] for extracting information called *top-k lists* from these web pages. Compared to other structured information, these top-k lists are more useful as they are cleaner, larger and richer with high quality and precise information. In

all the above mentioned data mining approaches, the focus was on structured data such as web tables or data records. But this one focuses on less structured or almost free-text information in top-k web pages and guides its extraction.

The system first recognizes top-k web pages by using a title classifier. The classifier is trained using a labeled dataset of positive and negative titles. Once it is recognized as a top-k web page, the system extracts all potential top-k lists. They are initially considered as candidate lists. Then obtained candidate lists are scored or ranked by using ranking functions or algorithms and the best one is chosen as top-k list. The extracted list is further enriched with attribute values and only the essential information is obtained from the top-k web page in the form of a top-k list.

III. CONCLUSION

Huge amount of data is available on the web and its size is rapidly growing. The data maintained by the sources can be efficiently extracted by various web mining techniques accurately based on the requirements of the user. In this paper, a detailed survey is performed on the existing techniques or strategies of information extraction from different forms of existence of data on the web. Some techniques are based on HTML structure; some extracts data records on web pages; while some are based on visual appearance of data on web. It also presents a technique for extracting information from top-k web pages.

ACKNOWLEDGEMENT

Our sincere thanks go to all the teaching and non-teaching staffs in the Department of Computer Science and Engineering, College of Engineering Perumon, for their help and co-operation throughout the work.

References

1. Zhixian Zhang, Kenny Q. Zhu, Haixun Wang , and Hongsong Li , “Automatic Extraction of Top-k Lists from the Web”, in IEEE Transactions On Knowledge And Data Engineering, 2013
2. Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. ACM Transactions On Internet Technology (TOIT), 3(1), 1-27.
3. Sadegh Kharazmi, Ali Farahmand Nejad, Hassan Abolhassani.” Freshness of Web Search Engines: Improving Performance of Web Search Engines Using Data Mining Techniques” IEEE 2009.
4. M. J. Cafarella, E. Wu, A. Halevy, Y. Zhang, and D. Z. Wang, “Webtables: Exploring the power of tables on the web,” in VLDB, 2008.
5. Liu, R. L. Grossman, and Y. Zhai, “Mining data records in web pages,” in KDD, 2003, pp. 601 606.
6. G. Miao, J. Tatemura, W.-P. Hsiung, A. Sawires, and L. E. Moser, “Extracting data records from the web using tag path clustering,” in WWW, 2009, pp. 981–990.
7. Shiqun Yin; Yuhui Qiu; Chengwen Zhong; Jifu Zhou, “Study of Web Information Extraction and Classification Method ”, Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007.
8. W. Gatterbauer, P. Bohunsky, M. Herzog, B. Kr¨upl, and B. Pollak, “Towards domain-independent information extraction from web tables,” in WWW. ACM Press, 2007, pp. 71–80.
9. F. Fumarola, T. Weninger, R. Barber, D. Malerba, and J. Han, “Extracting general lists from web documents: A hybrid approach,” in IEA/AIE (1), 2011, pp. 285–294.
10. J. Wang, H. Wang, Z. Wang, and K. Q. Zhu, “Understanding tables on the web,” in ER, 2012, pp. 141–155.
11. W. Wu, H. Li, H. Wang, and K. Q. Zhu, “Probase: A probabilistic taxonomy for text understanding,” in SIGMOD, 2012.
12. C.-H. Chang, and S.-C. Lui, “IEPAD: Information extraction based on pattern discovery,” Proceedings of the Tenth International Conference on World Wide Web (WWW), Hong-Kong, pp. 223-231, 2001.
13. Crescenzi, V., Mecca, G. and Merialdo, P., RoadRunner: towards automatic data extraction from large Web sites. Proceedings of the 26th International Conference on Very Large Database Systems (VLDB), Rome, Italy, pp. 109-118, 2001

