International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 7 (2016) pp 5338-5341
 https://dx.doi.org/10.37622/IJAER/11.7.2016.5338-5341

Survey on Web Content Extraction

Jincymol Joseph
Assistant professor Department of Computer Science,
St.Pius X College Rajapuram Kasargod, Kerala, India.

Dr. J.R. Jeba
Associate Professor & HOD, Department of Computer Applications,
Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India.

Abstract
World wide web has become one of the most significant resources nowadays. It brings the information mainly in the form of web pages. It may contain informative contents as well as non-informative contents. The non-informative contents like advertisements, header, footer, copyright statements, etc are called noisy parts. It has been proved that almost 40-50% contents are these types of noisy data. Web mining is an application of data mining technique to extract informative contents from non-informative contents. Web content mining is a subdivision under web mining. It is defined as the process of extracting informative content from non-informative contents known as noise. The advantage of eliminating non-informative content will saving in storage and indexing. This paper describes various methods for extracting web information from the huge volume of data present in world wide web.

Keywords: Hierarchical pattern based clustering, vision based approach, support vector, noisy data

Introduction

Web content mining is the process of extracting core contents from web documents. The term content extraction was introduced by Rahman[13]. As the internet grows rapidly, anyone can upload or download any information at any time. This leads to the continuous expansion of irrelevant, redundant, structured and unstructured information on the web pages. A web document may contain audio, video, text, images, tables etc. Extracting useful information from these types of unstructured data is a complex task. Some algorithms were developed for this purpose and each one has its advantages and disadvantages.

Content extraction has many advantages. It is easier for accessing useful information in a timely and efficient manner. Irrelevant and redundant information is removed. Since it will not waste their time and memory for indexing and storing irrelevant content, the performance of search engine is increased. So it can be considered as a pre-processor for search engine. It also helps users to browse internet through small screen devices. It also helps in generating rich site summary from blogs or articles.

Normally a web content extractor, extracts all the information on the web pages including text, graphics, audio, video, links, advertisements, contents, etc. During the extraction process, noisy data are discarded and useful information is preserved. Many algorithms were developed for eliminating these noisy information and extracting the core contents of the web pages.

Literature Review

Many authors have tried to exploit content extraction tools for web documents. Some highlights of the relevant work are outlined here.

Sandip et al [1] proposed the automatic identification of informative sections of web pages. Here four simple yet powerful algorithms called Contentextractor, FeatureExtractor, K-FeatureExtractor an L-Extractor were proposed to identify and separate content blocks from non-content blocks. FeatureExtractor is based on the characterization and uses heuristics based on the occurance of certain features to identify content blocks. K-FeaureExtractor is a special modification of FeatureExtractor which perform better in a wide variety of web pages. ContentExtractor identifies non-content blocks based on the appearance of the same block in multiple web pages. L-Extractor uses various block features and train a support vector(SV) based classifier to identify a informative block versus a non-informative block. First, the algorithm partition the web page into blocks based on heuristics. Second, the algorithm classifies each block as either a content block or non-content block. It has the advantage that both K-FeatureExtractor and ContentExtractor produce excellent precision and recall values and runtime efficiency. It also reduces the complexity and increases the effectiveness of the extraction process. It has the disadvantage that it will increase the storage requirement for indices and the efficiency of the markup algorithm are not improved.

Yinghui et al[2] proposed a methodology called Hierarchical pattern based clustering algorithm. Based on using item sets to represent patterns in web transactions, Greedy Hierarchical item set based clustering (GHIC) has been presented. At first, the set of frequent item sets in the unclustered data is obtained. After that, a new dataset (Binary item sets dataset) was generated where the rows represent the original transactions and the columns represent the presence or absence of a frequent item set. This is represented as the new set of transactions. The problem was converted into clustering these binary vectors. Then GHIC is presented to solve the clustering problem in the new set of transactions. It has the

advantage that a set of item sets was allowed to describe a cluster instead of just a set of items and it has the ability to explain the clusters and the differences between clusters. Difference or similarity matrix was not considered here.

Jinbeom Kang et al[3] proposed a new method of web page segmentation by recognizing tag patterns in the DOM tree structure of a page. These repetitive HTML tag patterns are called key patterns. Repetition based page segmentation(REPS) algorithm is proposed to detect key patterns in a page and to generate virtual nodes to correctly segment nested blocks. REPS proceeds in four phases. First a web page is represented by a DOM tree structure after removing less meaningful tags such as <a>, <b>, <script>, etc from the HTML source of the page. In the second phase, REPS generates a sequence from the DOM tree by using the tags in the child nodes of the root node. The third phase is to find the key patterns from the sequence and recognize candidate blocks by matching the sequence with the key patterns. The final phase of REPS is to generate blocks in a page by modifying the DOM tree into a more deeply hierarchical structure by introducing virtual nodes.

Chia-Hui Chang[4] done a survey of web information extraction systems. They noticed some points like to automate the translation of input pages into structured data, a lot of efforts have been devoted in the area of information extraction(IE). IE produces structured data ready for post processing, which is critical to many application of web mining and searching tools. The web IE processes online documents that are semi-structured and usually generate automatically by a server-side application program. Web IE usually applies machine learning and pattern mining techniques to exploit the syntactical patterns of the template based documents. They found disadvantages like the extraction precision is greatly decreased in case of missing or multiple order attributes.

Wei Liu et al [5] proposed a method called Vision based approach for deep web data extraction. It is primarily based on the visual features human users can capture on the deep web pages while also utilizing some simple non-visual information such as data types and frequent symbols to make the solution more robust. It consists of two main components, vision based data record extractor(ViDRE) and vision based data item extractor(ViDIE). First, given a sample deep web page from a web database, obtain its visual representation and transform into a visual block tree. Second, extract data records from the Visual Block tree. Third, partition extracted data records into data items and align the data items of the same semantic together. Fourth, generate visual wrappers (a set of visual extraction rules) for the web database based on sample deep web pages. ViDE can easily distinguish the misaligned data items due to their different fonts or positions. Visual information of web pages helps to implement web data extraction. Demerits like either precision or recall is not 100 percent. Also this measure indicates the percentage of web databases the automated solution fails to achieve perfect extraction.

Badr Hssina et al[14] proposed a method to extract required pattern by removing noise that is present in the web document using hand-crafted rules. Hand-crafted rules use string manipulation functions to extract information from HTML. Since the source of information is a mixture of image, audio, presentation, etc, it is not easy to separate out the informative content effectively and intelligently. First, connect to any website and get data from that site. Then choose options like extract links, extract image, extract media, extract HTML schemas, extract content.

R.Gunasundari[15] developed a method to extract content from web pages, which is based on links present in a web site. In this method, algorithm judges the contents by several parameters in the nodes. They are Link Text Density(LTD), Link Amount(LA), Link Amount Density(LAD) and Node Text Length(NTL).LTD and NTL are very important parameter for content location judgement and LA & LAD are indicators for accurate content judgement. The following methods are used for extracting the main contents. First, standardize the web page tags. Second, pre-processing the web page tags. Third, judging the location of the content. Four, extracting the content. Five, Adjusting the extraction results. This method can reduce quantity of data transmission and complexity. Also it is suitable for data collection workers and other professionals. Concept retrieval and the expansion of semantic and synonyms are needed for further work.

According to S.S. Bhamare [16] noise on the
web pages are not the part of the main content and this irrelevant information in web pages can really affect web mining task. Two categories of noise group are formed. They are global noise and local noise. There are web cleaning techniques or methods.

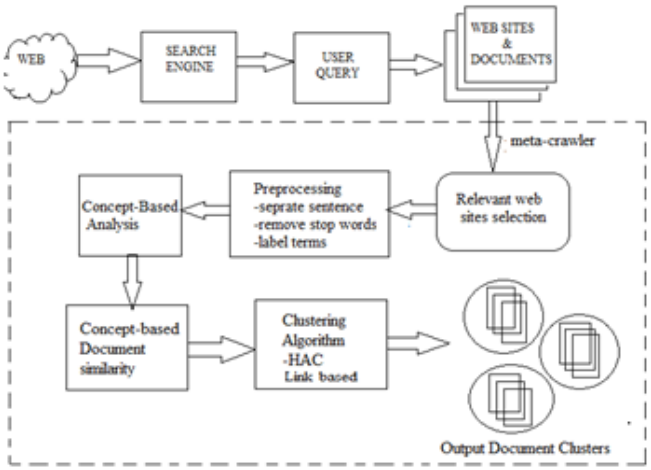
1) Page segmentation manually or automatically segments a web page into small blocks focusing on coherant subtopics.
2) Block matching identifies logically comparable blocks in different web pages.
3) Importance evaluation measures the importance of each block accrding to different information or measurements.
4) Noise determination distinguishes noisy blocks from non-noisy blocks based on the importance evaluation of blocks

Pralhad S Gamre et al[17] organize a set of documents into categories through clustering. Grouping of similar documents into clusters will help the users to find the information easily. Objects in the same cluster should be similar. Also, objects in one cluster should be dissimilar from objects in other cluster. Hybrid approach uses concept based mining model. In this model, it analyses terms on the sentence, document, corpus level and Hierarchical Agglomerative Clustering(HAC) to group similar documents in clusters and the documents are arranged in hierarchical structure to make easy access of web documents.

Requirements for document clustering methods are identified. They are extraction of informative features, overlapping cluster model, scalability, noise tolerance, incremental and result presentation[17]. Some properties of clustering algorithm are data model, similarity measure and cluster model. The proposed system works in the following manner.

1. Retrieve the results obtained for a search query from search engine.

2. Select the most important results from all the retrieved URLs.
3. Pre-process the documents using concept based model.
4. Measure the importance of each concept with respect to semantics of sentences.
5. Use Hierarchical Agglomerative Clustering (HAC) to group the similar documents in clusters and the documents are arranged in hierarchical structure.



Categorizing similar documents together into clusters will help the users to find useful information quicker. Each cluster contains documents that are very similar to each other and very dissimilar to the documents in other clusters. Clustering can increase the efficiency of information retrieval. So, it will reduce the time and get high precision. An important issue is incrementality, because web pages changes frequently and new pages are added frequently.

Conclusion

Each of these model examines web content present in the internet and extract information using various methods. All these methods have some pros and cons. The review on several existing models is examined and the pitfalls explored are identified during the review. As future work, research is to be continued on alternative method for extracting core contents from web pages.

References

[1] Sandip, prasenjit, Nirmal Pal and C.Lee Giles, "Automatic Identification of Informative Sections of web pages", IEEE Transactions on knowledge and data Engineering, Vol 7, No 9, 2005.

[2] Yinghui Yang and Balaji Padmanabhan "A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions", IEEE Transactions on knowledge and data Engineering, Vol 7, No 9, 2005.

[3] Jinbeom Kang, Jaeyoung Yang, Nonmember and Joongmin Choi, "Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices", IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010

[4] Chia-Hui Chang, Moheb Ramzy Girgis, "A Survey of Web Information Extraction Systems", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 10, October 20006.

[5] Wei Liu, Xiaofeng Meng and Weiyi Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2010

[6] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali, Belaid Bouikhalene "An implementation of web content extraction using mining techniques", Journal of Theoretical and Applied Information Technology 31st December 2013. Vol. 58 No.3, ISSN: 1992-8645

[7] Ms. Pranjali G. Gondse, Professor Anjali B. Raut "Main Content Extraction From Web Page Using Dom", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014

[8] K. Nethra1, J. Anitha2 and G. Thilagavathi "Web Content Extraction Using Hybrid Approach", ICTACT Journal On Soft Computing, January 2014, VOLUME: 04, ISSUE: 02

[9] Rajni Sharma, Max Bhatia, "Eliminating the Noise from Web Pages using Page Replacement Algorithm ", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3066-3068

[10] J.R jeba, S.P.Victor, "A novel approach for finding item sets with hybrid strategies", International Journal of Computer Applications., Vol.17, No.5, 2011

[11] J.R.Jeba, S.P.Victor, " Comparison of frequent item set Mining algorithms", International Journal of Computer Science and Information Technologies, Vol 2 (6), 2011

[12] J R Jeba, S.P.Victor, "Effective measures in Association Rule mining", International Journal of Scientific and Engineering research, Vol 3, Issue 8, 2012.

[13] A.F.R Rahman, H.Alam and R.Hartono, "Content extraction from HTML documents", International workshop on Web document Analysis, pp.7-10, 2001.

[14] Badr Hssina, " An implementation of web content extraction using mining techniques", Journal of Theoretical and Applied Information Technology, Vol. 58 No.3

[15] R.Gunasundari, "A study of content extraction from web pages using links"International Journal of Data Mining & knowledge management process, Vol.2, No.3, May2012

[16] S.S. Bhamare, Dr.B.V., "Survey on Web Page Noise Cleaning for Web Mining", International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013

[17] Pralhad S. Gamare, G. A. Patil, "Web document clustering using hybrid approach in data mining"

International Journal of Advent Technology, Vol.3, No.7, July 2015

[18] Lambodar Jena, Narendra Kumar Kamila "Data Extraction and Web Page Categorization using Text Mining", International Journal of Application or Innovation in Engineering & Management, Vol 2, Issue 6, June 2013.