www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 20003-20006
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.29

A Study on Web Content Mining

Anurag kumar[1], Ravi Kumar Singh[2]
[1]Dr. APJ Abdul Kalam UIT,
Jhabua, MP, India
Anurag.davv@gmail.com
[2]Prestige institute of Engineering Management and Research,
Indore, MP, India
Ravi.singh1308@gmail.com

Abstract: Web Mining is extracting information from the web re-sources and finding interesting patterns that can be useful from ever expanding database of World Wide Web. Whenever we talk about data, we conclude that there is a huge range of data on World Wide Web. Due to heterogeneity and unstructured nature of the data available on the WWW, Web mining uses various data mining techniques to discover useful knowledge from Web hyperlinks, page content and usage log. Web Content Mining is a component of Data Mining. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the Web to the user requesting the information. This paper deals with a preliminary discussion of Web content mining, contributions in the field of web mining, the prominent successful tools and algorithms.

Keywords: Web content mining, structured data mining, unstructured data mining, semi-structured data mining.

1. Introduction

Internet is a network of worldwide level, constantly changing and non-structured [1]. The Web is the largest data source in the world. Web mining aims to extract and mine useful knowledge from the Web. It is a multidisciplinary field involving data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc. The amount of information on the Web is huge, and easily accessible. The knowledge does not come only from the contents of the web pages but also from the unique feature of Web, its hyperlink structure and the diversity of contents. Analysis of these characteristics often reveals interesting patterns and new knowledge which can be helpful in increasing the efficiency of the users, so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques help to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process.

Web Content mining
Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. Webpage can be in traditional text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents. Web content mining could be differentiated from two points of view the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [1]:

***Intelligent Search Agents*:** These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
***Personalized Web Agents*:** These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

Web structure mining
The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level [6, 8]. It also helps in discovering the structure of document which is used in revealing the structure the structure of web pages and it's possible to compare the web page schemes

Web usage mining
It is also known as Web log mining, is used to analyze the behavior of website users. It tries to discover useful information secondary data derived from the interaction of users while surfing web [8]. Web usage mining collects the data from Web log records to determine user access patterns of Web pages.
.

2. Web Content mining

Web Content Mining is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [3].

2.1 Web Content Mining Strategies

Web Content Mining Approaches: Two approaches used in web content mining are Agent based approach and database approach [4],[5]. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, and personalized web agents [6]. Intelligent Search agents

automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles [4], [5].

Web content mining has the following approaches to mine data

1 Unstructured text mining,
2 Structured mining,
3 Semi-structured text mining, and
4 Multimedia mining.[8]

Unstructured Text Data Mining: Most of the Web content data is of unstructured text data. Content mining requires application of data mining and text mining techniques [7]. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are

- Information Extraction,
- Topic Tracking,
- Summarization, Categorization,
- Clustering and
- Information Visualization.

Structured Data Mining: The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler,
- Wrapper Generation,
- Page content Mining.

Semi-Structured Data Mining*:* Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intradocument structure. The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top Down Extraction, and
- Web Data Extraction language.[8]

Multimedia Data Mining: The techniques of Multimedia data mining are;

- SKICAT,
- Color Histogram Matching, Multimedia Miner and Shot Boundary Detection.

2.2 Web Content Mining Algorithms

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to fetch the information are described

(i) Decision Tree: The decision tree is one of the powerful classification techniques. Decision trees take the input as its features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5.

The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [12].

(ii) k-Nearest Neighbour: KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation [15].

(iii) Naive Bayes: Naive Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP).Given a problem with K classes {C1, ..., CK} with so called prior probabilities P (C1), ... , P(CK), can assign the class label c to an unknown example with features such features x=(x1,.....,xN) such that **c=argmaxcP(C=ckx1.....xN),** is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, that is choosing the class with the maximum a posterior probability given the observed data. This posterior probability observed data. This posterior probability can be formulated,

P(C=ckx1,..,xN)=p(C=c)P(x1,....,xNkC=c)P(x1.....xN)

As the denominator is the same for all classes, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the accessible classes. This may be quite difficult taking into account the dependencies between features. This approach is to assume conditional independence i.e. x1,. , xN are independent. This simplifies numerator as P(C = c) P (x1kC = c). P (xNkC = c), and then choosing the class c that maximizes this value over all the classes c = 1. . . K [12].

(iv) Support Vector Machine: Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear data and uses a nonlinear mapping to transform the original training data into a higher dimension. Among the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyper plane [15].

v) Neural Network: The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer. Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [12]. As network is feed-forward in that

none of the weights cycles back to an input unit or to an output unit of a previous layer.

vii) Cluster Hierarchy Construction Algorithm (CHCA)
The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed). Using the reduced table, we create a cluster hierarchy by examining each row, starting with those with the fewest terms (fewest number of 1's); these will become the most general clusters in our hierarchy.
The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster. Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above. If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it. This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created. The next step in the algorithm is to assign the web pages to clusters in the hierarchy. In genera there will be some similarity comparison between the terms of each web page (rows in the original table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page. Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

3. Web Content mining Tools

Web content mining tools helps to download the essential information. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda and Web Content Extractor, Rapid Miner.
1. Rapid Miner: Rapid Miner is open source software and it is a tool for extracting information from web, Contains inbuilt algorithm. It can generate algorithm by itself.
Features:

- Easy to use.
- Reduce time.
- Open source software.

2. Screen-scaper: Screen-scraping is a tool for extracting/ mining information from web sites [11]. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. Features: Screen-scraper present a graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data. Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked. This also facilitates scraping of information at periodic intervals. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time after which the results are displayed in a single interface.

3. Automation Anywhere: It is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining [14].
Features:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly. Web record and Web data extraction.

4. Web Info Extractor: This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.
Features:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file
- Can deal with Web page in all language
- Running multi-task at the same time
- Support recursive task definition.

5. Mozenda: This tool enables users to extract and manage Web data [15]. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, the data to be used in other applications or as intelligence. There are two parts of Mozenda's scraper tool:
i. Mozenda Web Console: It is a Web application that allows user to run agents, view & organize results, and export publish data extracted.
ii. Agent Builder: It is a Windows application used to build data extraction project.
Features:

- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence.

6. Web Content Extractor: It is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet[13].It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner. This tool allows users to extract data from various websites such as online stores, online actions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT,

HTML, XML, SQL script, MySQL script and to any ODBC data source.

Features:

- Helps to extract/collect the market figures, product pricing data, or real estate data.
- Helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- Assists to Journalists extract news and articles from news sites.
- Helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

4. Conclusion

The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in web content mining but this paper tried to expose the various web content mining strategy and explore some of the techniques. Then we described some tools web content mining.

References

[1] Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).

[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[3] Han, J., Kamber, M. Kamber. "Data mining: concepts and techniques". Morgan Kaufmann Publishers, 2000.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668–677,1998

[5] R.Cooley,. B.Mobasher,.; J.Srivastava,.; "Web mining: information and pattern discovery on the World Wide Web". In Proceedings of Ninth IEEE International Conference. pp. 558 – 567, 3-8 Nov. 1997.

[6] Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[7] V. Bharanipriya & V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[8] Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).

[9] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.

[10] R. Baeza-Yates and e. Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Longman Publishing Company, 1999.

[11] screen-scraper,http://www.screen- scraper.com Viewed 19 February 2013

[12] Darshna Navadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue10,December- 2012 ISSN: 2278-0181

[13] Web Content Extractor help. WCE ,http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013.

[14] Automation Anywhere Manual. AA, http://www.automationanywhere.comViewed 06 February 2013.

[15] Mozenda,http://www.mozenda.com/web- mining-software Viewed 18 February 2013.

[16] Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

Author Profile

Anurag Kumar received the M.Tech. degree in Information Architecture & Software Engineering from School of Computer Science & IT, DAVV Indore in 2016. He is now assistant professor at Dr. APJ Abdul Kalam UIT Jhabua, MP.

Ravi Kumar Singh received the M.Tech. degree in Information Architecture & Software Engineering from School of Computer Science & IT, DAVV Indore in 2016. He is now assistant professor at Prestige institute of Engineering Management and Research,Indore MP.