Volume 5, Issue 4, 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Content Mining: Techniques and Algorithms

Govind Murari Upadhyay, Kanika Dhingra
Assistant Professor, IITM, Janakpuri,
New Delhi, India

Abstract: The Web has growing continuously with respect to the volume of information, in the complexity of its topology, as well as in its diversity of content and services. This phenomenon was transformed the web in spite of his young age to an obscure media to take useful information. If we search a specific keyword at a time than it is not necessary that when again search that keyword on the internet will revert the same result with respect to files, links and services. As the content on the web are updating so frequently that the volume of information become so large. Today, they are billions of HTML documents, images and other media files on the Internet. In that case the accessed data may or may not be relevant, unnecessary data or the volume of data may be large. Taking into account the wide variety of the web, the extraction of interesting content has become a necessity. Web mining came as a rescue for the above problem. Web content mining is a part of web mining, which is defined as "the process of extracting useful information from the text, images and other forms of content that make up the pages" by eliminating noisy information . The objective of web content mining is to extract the exact information from the web, which we want, no noisy data will be extracted.

Keywords: Web Mining, Content Mining, Extraction, Clustering, Classification

I. INTRODUCTION

In the modern age we are dependent on the web. We seek all the information from web, but it is becoming challenging task to retrieve the required web pages/information very effectively and efficiently on the web. Today, they are billions of HTML documents, images and other media files on the Internet. In that case the accessed data may or may not be relevant, unnecessary data or the volume of data may be large. Taking into account the wide variety of the web, the extraction of interesting content has become a necessity. Web mining came as a rescue for the above problem. Mining techniques are in detail, results and comparison to extract necessary information effectively and efficiently. Web content extraction is concerned with extracting the relevant text from Web pages by removing unrelated textual noise like advertisements, navigational elements, contact and copyright notes. Web crawling involves searching a very large solution space which requires a lot of time, hard disk space and lot of usage of resources. The research done in Web content mining from two different points of view: IR and DB views. IR view is mainly to assist or to improve the information finding and filtering the information to the users usually based on either inferred or solicited user profiles. DB view mainly tries to model the data on the Web and to integrate them so that more sophisticated queries other than the keywords based search could be performed. The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized web agents. 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 instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well formed database containing schemas and attributes with defined domains. The algorithm proposed is called Dual Iterative Pattern Relation Extraction for finding the relevant information used by search engines. The content of web page includes no machine readable semantic information. Search engines, subject directories, intelligent agent, cluster analysis and portals are employed to find what a user must look for. [1]

II. WEB CONTENT MINING TECHNIQUES

There are two types of web content mining techniques, one is called clustering and other is called classification.

2.1. Clustering:

Clustering is one of the major and most important preprocessing steps in web mining analysis. In this context (Web Usage/Context Mining) items to be studied are web pages. Web page clustering puts together web pages in groups, based on similarity or other relationship measures. Tightly-couple pages, pages in the same cluster, are considered as singular items for following data analysis steps. A complete data mining analysis could be performed by using web pages information as it appears in web logs, but when the number of pages to take into account increases (i.e., in a corporative large scale web server or a server using dynamic web pages) this process could be quite hard or even unbearable. In order to deal with this issue, web page clustering appears as a reasonable solution. These techniques group pages together based on some kind of relationship measure. Pages in the same cluster will be considered as a single item for further data analysis steps [8].

Clustering techniques

Web page clustering deal with a set of web pages hosted on a web server to obtain a collection of web page sets (clusters). These clusters are applied in the following steps of the mining process instead of original pages. There are three web clustering criteria: semantic, structure, and usage based.

2.1.1. Semantic Clustering

Semantical web page clustering are based on the concept of web page hierarchies. The lowest level leaves in these hierarchies are web pages, that are grouped in higher level nodes based on semantical affinities. For example, product web pages are clustered in several product families that are later grouped in a cluster for all products, beside other clusters of corporative or support information can also be defined. Semantical hierarchies can be defined following many different criteria, depending on the objectives and strategies of this analysis, and, hence, many different collections of clusters can be provided. This web page clustering techniques requires, anyway, some domain information, either from the domain experts or retrieved by any semantic repository. In this later case, there is a range of possible paths, from META-like information provided on the page contents, to Semantic Web principles, including also CMS-based web sites.

2.1.2. Graph Partitioning

Structure and usage page clustering are both very similar. These two approaches build a web page graph, in which nodes are the different web pages and arcs are the links among these pages. These links can be defined by the actual web links, in the case only web structure is considered or may be weighted by the usage of these transitions. In this last case, web log file is scanned to analyze the frequency of the transitions. In these entire cases web clustering problem is translated in what is called graph partitioning. The graph partitioning problem is NP-hard, and it remains NP-hard even when the number of subsets is 2 or when some unbalancing is allowed [9]. For large graphs (with more than 100 vertices), heuristics algorithms which find suboptimal solutions are the only viable option. Proposed strategies can be classified in combinatorial approaches based on geometric representations [11], multilevel schemes, evolutionary optimization and genetic algorithms [10]. We can find also a hybrid scheme that combines different approaches. There are a lot of graph partitioning algorithms and, as we cannot describe every single algorithm, we have selected those that we consider more relevant. We will start talking about four graph partitioning heuristics we have used in this study and then we will give a brief description of other clustering algorithms we find interesting.

2.2. Classification

Classification Techniques:

The Classification algorithms are discussed under this section. The need and requirement of the user's of the websites to analyze the user preference become essential due to massive internet usage. Classification techniques are to be applied on the web log data and the performance of these algorithms can be measured. Here, in the following several classifiers are being discussed.

2.2.1. Decision Tree:

Decision tree is a powerful classification technique. The decision trees, take the instance described by its features as input, and outputs a decision, denoting the class information in our case. 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. The split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. 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[2].

2.2.2. k-Nearest Neighbor:

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[2].

2.2.3. 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), we can assign the class label c to an unknown example with features x = (x1,. . . , xN) such that c = argmaxcP(C =ckx1, . . . , xN), that is choose the class with the maximum a posterior probability given the observed data. This a posterior probability can be formulated, using Bayes theorem, as follows:

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 available classes.

This can be quite difficult taking into account the dependencies between features. The naive bayes approach is to assume class conditional independence i.e. x1, . . . ,xN are independent given the class. This simplifies the numerator to be 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.[2]

2.2.4. 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. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane(i.e., "decision boundary"). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. SVM finds this hyper plane using support vectors("essential" training tuples) and margins (defined by the support vectors).[2]

2.2.5. Neural Network:

The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. it consists of 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. Inputs are fed simultaneously into the units making up the input layer. They are then weighted and fed simultaneously to a hidden layer. The number of hidden layers is arbitrary, although usually only one. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction. The 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[2].

III. CONCLUSION

In this paper we have discussed about the techniques of web content mining, one is clustering and other is classification. Various algorithms regarding clustering and classification are discussed. Future work will be how can we extract data more efficiently using these algorithms.

REFEREANCE

[1] Badr Hssina, 2abdelkarim Merbouha,3hanane Ezzikouri, 4mohammed Erritali, , 5belaid Bouikhalene, An Implementation Of Web Content Extraction Using Mining Techniques, JATIT, 517-519, 2013.

[2] Web Content Mining Techniques-A Comprehensive Survey- Darshna Navadiya, Roshni Patel- (IJERT) December-2012.

[3] Text Classification Using Data Mining-S. .Kamruzzaman,Farhana Haider,Ahmed Ryadh Hasan-ICTM2005.

[4] A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms-Preeti Chopra, Md. Ataullah-(IJEAT)Feb2013.

[5] Patel Archana J, Mukti Pathak, WEB CONTENT MINING USING RULE BASED CLASSIFIER, *International Journal For Technological Research In Engineering Volume 1, Issue 9, May-2014*

[6] Web Content Mining, The 14th International World Wide Web Conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[7] R.Agrawal and R.Srikant. Fast algorithms for mining association rules.In VLDB'94, pp.487 {499.

[8] Antonio LaTorre, Jos´e M. Pe˜na, V´ıctor Robles, Mar´ıa S. P´erez A Survey in Web Page Clustering Techniques

[9] T.N. Bui and C. Jones. Finding good approximate vertex and edge partitions is np-hard. *Information Processing Letters*, 42:153–159, 1992.

[10] T.N. Bui and B. Moon. Genetic algorithms and graph partitioning. *IEEE Transactions on Computers*, 45(7):841–855, 1996.

[11] J. Gilbert, G. Miller, and S. Teng. Geometric mesh partitioning: Implementation and experiments. In *Proceedings of the 9th International Parallel Processing Symposium*, pages 418–427, 1995.