International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 10, December- 2012

Web Content Mining Techniques-A Comprehensive Survey

Darshna Navadiya
Government Engineering collage,Modasa

Roshni Patel
Jodhpur National University

ABSTRACT

With flooding of information on WWW it has become necessary to apply some strategy so that valuable knowledge can be extracted and consequently returned to the user. Data mining techniques find their applicability in these scenario. Data mining concepts and techniques when applied to WWW with its existing technologies are known as web mining.The paper contains techniques of web content mining, review, various algorithms, examples and comparison. Web mining is one of the well-known technique in data mining and it could be done in three different ways (a)web usage mining, (b)web structure mining and (c)web content mining. Web usage mining allows for collection of web access information for web pages. Web content mining is the scanning and mining of text, pictures and graphs of web page to determine relevance of content to the search query. Web structure mining is used to identify the relationship between the web pages linked by information.The paper presents various examples based on web content mining techniques in detail, results and comparison to extract necessary information effectively and efficiently.

Keywords: Data Mining, Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Classification

1. INTRODUCTION

Web mining is the Data Mining technique that automaticallydiscovers or extracts the information from web documents. It isthe extraction of interesting and potentially useful patterns andimplicit information from artifacts or activity related to the WorldWide Web.Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web.Web-mining is a multi-disciplinary effort that draws techniques from fields like in-formation retrieval, statistics, machine learning, natural language processing, and others.

2. WEB MINING PROCESS

The figure given below shows the process of Web mining:

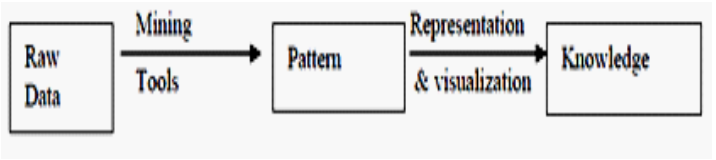

Fig.1 Web Mining Process

The Following are the subtasks of Web Mining process:

I. Resource finding: It is the task of retrieving intended webdocuments.
II. Information selection and pre-processing: Automaticallyselecting and

pre- processing specific from informationretrieved Web resources.

III. Generalization:Automatically discovers general patterns atindividual Web site as well as multiple sites.

IV. Analysis: Validation and interpretation of the minedPatterns.

3. WEB MINING CATEGORIES

Web Mining can be categorized into three types:

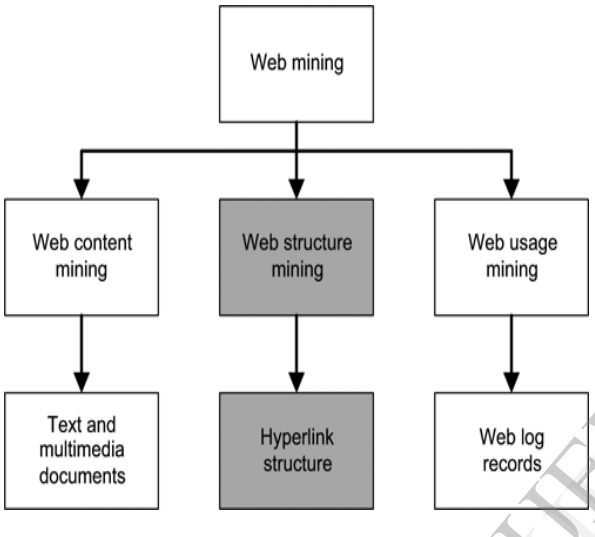

Fig.2 Web Mining Categories

I. Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. It is related to data mining. It is related to text mining because much of the web contents are textbased. Text mining focuses on unstructured texts. Web content mining is semi-structured nature of the web. Web content mining can be of two types:Those that directly mine the content of documents and those that improve on the content search of other tools like search engines. Content mining is used to examine data collected by search engines and Web spiders. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information Retrieval).

II. Web Structure Mining

Web Structure Mining tries to discover useful knowledge from the structure and hyperlinks. The goal of web structure mining is to generate structured summery about websites and web pages. It is using tree-like structure to analyse and describe HTML or XML.

III. Web Usage Mining

Web usage mining is the process by which we identify the browsing patterns by analysing the navigational behaviour of user. It focuses on technique that can be used to predict the user behaviour while user interacts with the web. It uses the secondary data on the web. This activity involves automatic discovery of user access patterns from one or more web-servers. It consists of three phases namely: pre-processing, pattern discovery, pattern analysis.web usage miningitself can be classified further depending on the kind of usage data considered: Web Server Data, Web Server Data, Application Level Data.

4. WEB CONTENT MINING METHODS

This section gives an introduction to some of the current Web content mining tasks:

I. Structured data extraction:

Structure data extraction is most widely used in web content mining. Structured data is easier to extract compare to unstructured data. There are several approaches to structure data extraction, called wrapper generation.

The first approach is to manually write an extraction program for each web site based on observable format patterns of the site. This approach is time consuming. It doesn't scale for large number of sites.

The second approach is wrapper induction/wrapper learning. The user first manually labels set of trained pages. A learning system then generates rule from the training pages. The resulting rules are then applied to extract target items from web pages. E.g.: WIEN, stalker, BWI, WL^2.
The third approach is automatic approach. Structured data objects on the web are normally retrieved from database and displayed in the web pages with fix templates.E.g. MDR, Roadrunner, EXALG etc.

II. Unstructured text extraction:

Most Web pages can be seen as text documents. Extracting information from Web documents has also been studied by many researchers. The research is closely related to text mining, information retrieval and natural language processing. Current techniques are mainly based on machine learning and natural language processing to learn extraction rules. Recently, a number of researchers also make use of common language patterns (common sentence structures used to express certain facts or relations) and redundancy of information on the Web to find concepts, relations among concepts and named entities. The patterns can be automatically learnt or supplied by human users. Another direction of research in this area is Web question-answering. Although question-answering was first studied in information retrieval literature. it becomes very important on the Web as Web offers the largest source of information and the objectives of many Web search queries are to obtain answers to some simple questions.

IV. Web Information Integration.

Due to the sheer scale of the Web and diverse authorships, various Web sites may use different syntaxes to express similar or related information. In order to make use of or to extract formation from multiple sites to provide value added services, e.g. metasearch, deep Web search etc., one needs to semantically integrate information from multiple sources. Recently, several researchers attempted this task. Two popular problems related to the Webare (1) Web query interface integration, to enable querying multiple Web databases (which are hidden in the deep Web) and (2) schema matching, e.g. integrating Yahoo and Google's directories to match concepts in the hierarchies. The ability to query multiple deep Web databases is attractive and interesting because the deep Web contains a huge amount of information or data that is not indexed.

V. Building concept hierarchies.

Because of the huge size of the Web, organization of information is obviously an important issue. Although it is hard to organize the whole Web, it is feasible to organize Web search results of a given query. A linear list of ranked pages produced by search engines is insufficient for many applications. The standard method for information organization is concept hierarchy and/or categorization. The popular technique for hierarchy construction is text clustering, which groups similar search results together in a hierarchical fashion. Several researchers have attempted the task using clustering. A different approach is proposed which does not use clustering. Instead, it exploits existing organizational structures in the original Web documents, emphasizing tags and language patterns to perform data mining to find important concepts, sub-concepts and their hierarchical relationships. In order words, it makes use of the information redundancy property and semi-structure nature of the Web to find what concepts are important and what their relationships might be. This work aims to compile a survey article or a book on the Web automatically.

VI. Segmenting Web pages & Detecting noise.

In web data mining, classification and clustering are used to remove noisy blocks and enables to produce much better results. Another application is web browsing using a small screen device called PDA.

VI. Mining web opinion sources.

Web was available. Companies usually conduct consumer surveys or engage external consultants to find such opinions about their products and those of their competitors. Now much of the information is publicly available on the Web. There are numerous Web sites and pages containing consumer opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This online word-of-mouth behaviour represents new and measurable sources of information for marketing intelligence. Techniques are now being developed to exploit these sources to help companies and individuals to gain such information effectively and easily. For instance, proposes a feature based summarization method to automatically analyse consumer opinions in customer reviews from online merchant sites and dedicated review sites. The result of such a summary is useful to both potential customers and product manufacturers.

5. WEB CONTENT MINING TECHNIQUES

The two common tasks through which useful information can be mined from Web are Clustering and Classification. Here In this paper, I present various classification algorithms used to fetch the information.

Classification is often posed as a supervisedlearning problem in which a set of labeled data is used to train a classifier whichcan be applied to label future examples.

I. Decision Tree:

Decision tree is a powerful classification technique. The decision trees, take the instance described by its features as input, andoutputs a decision, denoting the class information in our case. Two widely known algorithmsfor building decision trees are Classification and Regression Treesand ID3/C4.5. The tree tries to infer a split of the training data basedon the values of the available features to produce a good generalization. Thesplit at each node is based on the feature that gives the maximum informationgain. Each leaf node corresponds to a class label. A new example is classifiedby following a path from the root node to a leaf node, where at each node atest is performed on some feature of that example. The leaf node reached isconsidered the class label for that example. The algorithm can naturally handlebinary or multiclass classification problems. The leaf nodes can refer to eitherof the K classes concerned.

II. k-Nearest Neighbor:

kNN is considered among the oldest non-parametric classification algorithms.To classify an unknown example, the distance (using some distance measuree.g. Euclidean) from that example to every other training example is measured.The k smallest distances are identified, and the most represented class inthese k classes is considered the output class label. The value of k is normallydetermined using a validation set or using cross-validation.

III. Naive Bayes:

Naive Bayes is a successful classifier based upon the principle of Maximum APosteriori (MAP). Given a problem with K classes {C1, . . . ,CK} with so-calledprior probabilities P(C1), . . . ,

P(CK), we can assign the class label c to an unknownexample with features x = (x1, . . . , xN) such that c = argmaxcP(C =ckx1, . . . , xN), that is choose the class with the maximum a posterior probabilitygiven the observed data. This aposterior probability can be formulated, usingBayes theorem, as follows: P(C = ckx1, . . . ,xN) = P(C=c)P(x1,...,xNkC=c)P(x1,...,xN) . Asthe denominator is the same for all clases, it can be dropped from the comparison.Now, we should compute the so-called class conditional probabilities of thefeatures given the available classes. This can be quite difficult taking into accountthe dependencies between features. Thenaive bayes approach is to assumeclass 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 classesc = 1, . . . ,K.

IV. Support Vector Machine:

Support Vector Machines are among the most robust and successful classificationAlgorithms.It is a new classification method for both linear and nonlinearData.It uses a nonlinear mapping to transform the originaltraining data into a higher dimension.With the new dimension, it searches for the linear optimalseparating hyperplane(i.e., “decision boundary”). With an appropriate nonlinear mapping to a sufficientlyhigh dimension, data from two classes can always beseparated by a hyperplane. SVM finds this hyperplane using support vectors(“essential” training tuples) and margins (defined by thesupport vectors)

V. Neural Network:

The most popular neural network algorithm is backpropagation 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 foreach training tuple. Inputs are fed simultaneously into the units making up the inputlayer. 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 unitsmaking up the output layer, which emits the network's prediction. The network is feed-forward in that none of the weights cycles backto an input unit or to an output unit of a previous layer.

6. CONCLUSION

This Paper includes the techniques for classifications which are commonly used to mine the information from the Web, Each having its own advantages and disadvantages. The selection of the technique depends on the application. For the future work other classification techniques can be considered to improve performance.

7. REFERENCES

1 Jiawei Han and MichelineKamber, Data Mining: Concepts and Techniques, Morgan kaufmann, 2nd Edition.
2 Charu C. Aggarwal, ChengXiangZhai, Mining Text Data,springler.
3 http://en.wikipedia.org/wiki/Data_mining
4 http://en.wikipedia.org/wiki/Web_mining
5 http://www.web-datamining.net/structure
6 http://www.web-datamining.net/content
7 http://www.web-datamining.net/usage
8 http://www.web-datamining.net
9 Bing Liu, Kevin Chen-chuan Chang...”Editorial Issues on Web

content Mining".SIGKDD Explorations –Volume

10 Bharat BhusanAgrawal, Dr,MHKhanandShivangiDhall,"WEB MINING: INFORMATION AND PATTERN DISCOVERY ON THE WORLD WIDE WEB", International Journal of Science, Technology & Management, December-2010.

11 Tamanna Bhatia," Link Analysis Algorithms For Web Mining", International Journal of Computer Science and Technology,issue-2,June-2011