A Survey: Techniques of an Efficient Search Annotation based on Web Content Mining

Sobana.E
M.E Computer Science and Engineering,
K.S.Rangasamy College of Technology,
Tiruchengode.

Muthusankar.D
Assistant Professor,
Department of CSE,
K.S.Rangasamy College of Technology,
Tiruchengode.

ABSTRACT
In the World Wide Web, or simply the web, the content of information is changing everyday and it is known as dynamic environment. There is more information are uploaded in web and it has grown steadily in recent years. Therefore the several billions of HTML documents, pictures and another multimedia files available on the Internet. Due to the overloaded of information in web, the information extraction is not effectively based on user needs. To overcome the above problem, there is a need of methods to help us extract information effectively from the content of web pages. Nowadays, various web content mining techniques are developed to mine the information and serve people in a simple way: These techniques focuses on the discovery/retrieval of the useful information from the Web contents/data/documents. This paper focus on how to extract the information effectively based on classification and clustering, and detecting phishing websites.

Keywords
Web Content mining, classification, clustering, phishing Websites.

1. INTRODUCTION
The Web Content mining is used to discover the useful information from web content such as text, images videos etc. The content data corresponds to the collection of facts which may consist of text, images, audio, video, or structured records such as lists and tables. The Web content mining mainly focuses on how the user can retrieve the desire information based on web content.

The rest of this paper is organized as follows: The overview of Web mining, process of Web mining and taxonomy of Web mining are described in Section 2, and the Overview of Web content mining and it's techniques are described in Section 3. Finally, conclude this paper in Section 4.

2. WEB MINING

2.1 Overview
In data mining techniques, web mining is one of the applications and it's used to discover patterns from the web. The research of the web mining is based on interdisciplinary field and it used techniques from data mining, text mining, databases, statistics, machine learning, multimedia, etc. Web mining has interest based on three categories such as clustering i.e. finding natural groups of users, pages, etc., next one is associations i.e. which URLs tend to request together and finally sequential analysis i.e. the order in which URLs tend to be accessed.

2.2 Web Mining Process
The Fig.1 shows the process of web mining and it has following subtasks,

(1) Resource Finding:

This task is used to retrieving intended web documents

(2) Information Selection & Preprocessing

In this task the specific information can be selected automatically and preprocessing from information retrieved web resources.

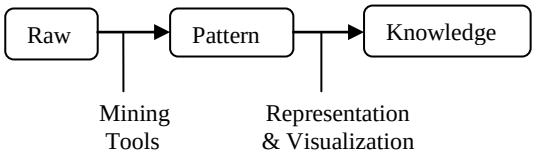
(3) Generalization:

Here, the general patterns are discovering automatically at individual websites as well as multiple sites

(4) Analysis:

In this task, validation and interpretation of the mined patterns can be done.

Fig 1: Process of Web Mining



2.3 Web Mining Taxonomy
Web mining can be divided into three categories,

- Web usage mining.
- Web content mining.
- Web structure mining.

The Fig.2 shows the taxonomy of web mining.

2.3.1 Web Usage Mining
Web Usage Mining is the application of data mining techniques which is used to discover interesting usage patterns from Web data. Based on this, the web can be provides understand and better serve the needs of Web-based applications. The identity or origin of Web users can be captures related to usage data along with their browsing behavior at a Web site. Its technique mainly used to predict the user behavior while user interacts with the web and it uses the secondary data on the web to discover the useful information. It consists of three phases namely,

- Pre-processing.
- Pattern discovery.
- Pattern analysis.

The web usage data can be captured easily based on Web servers, proxies and client applications. It is also referred as Web log mining which is used to analyze the behavior of website users. The discovery of user access patterns can be done from Web usage logs, which record every click made by each user. In web usage mining, the kind of usage data can be classified as follows,

(i) Web Server Data:

Based on the web server the user logs are collected. The IP address, page reference and access time are included in this data.

(ii) Application Server Data:

The Web logic, Story Server is called as Commercial application servers which have significant features to enable E-commerce applications. The main goal of application server data is to track various kinds of business events and log into an application server logs.

(iii) Application Level Data:

In application level data, the defined events are defining new kinds of events in an application, turned logging and generating histories.

2.3.2 Web Content Mining

Web content mining is the process of mining, extraction and integration of useful data, information and knowledge from the contents of Web documents. For example, the Web pages can be automatically classifies and cluster according to their topics. The content data corresponds to the collection of facts which may consist of text, images, audio, video, or structured records such as lists and tables. Web content mining is closely related to data mining and text mining because many of the techniques are applied for mining the Web, where most data are in text form. The data is examined based on content mining which is collected by search engines and Web spiders. The NLP (Natural language processing) and IR (Information Retrieval) technologies are normally used in web content mining. Web content mining also distinguishes personal home pages with other web pages. Based on research work, the web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. This Survey work represents how to retrieve the information efficiently based on applying some techniques. In section 3, the overview of Web content mining and its techniques are explained briefly.

2.3.3 Web Structure Mining

Web Structure Mining is used to analyze how the pages are written and discovers useful knowledge from hyperlinks, which represent the structure of the Web. For example, the important Web pages can be discovers from the links which is a key technology used in search engines. The communities of users who share common interests also can be discover. In web structure mining, the generation of structured summary about websites and web pages is the main goal. The tree-like structure is used to analyse and describe HTML or XML. The web structure mining is classified into two types namely,

- Intra-page structure.
- Inter-page structure

The existence of links within a page is known as Intra-page structure. There is no need to open the new page here. The connection of one page with the other page is known as the Inter-page structure. Based on the kind of structure information used, the web structure mining can be further divided into two types.

(i) Hyperlinks:

A Hyperlink is a structural unit which is used to connect the location of one Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.

(ii) Document Structure:

The content within a Web page can also be described in a tree-structured format, based on the various HTML and XML tags within the page. The document object model (DOM) structures are extracting automatically out of documents by mining efforts.

3. WEB CONTENT MINING

3.1 Overview

In traditional technique the searching was done via contents in the web. The extended work performed by search engines known as Web Content mining. Web Content mining is used to discover the useful information from web content such as text, images videos etc. In web content mining, there are two approaches are used namely Agent based approach and database approach.

(i) Agent based Approach:

The following three types of agents are used in Agent based approach.

- Intelligent search agents.
- Information filtering/Categorizing agent.
- Personalized web agents.

According to a particular query using domain characteristics and user profiles, the Intelligent Search agents automatically searches for information. The number of techniques are used in information agents to filter data according to the predefine instructions. The learning of user preferences and discovers documents related to those user profiles are done by Personalized web agents.

(ii) Database approach:
The Database approach consists of well formed database containing schemas and attributes with defined domains. The mining of unstructured, structured, semi structured and multimedia data is more complicated in Web content mining.

The web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. There are two main approaches are used in Web Content Mining namely,

- Unstructured text mining approach.
- Semi-Structured and Structured mining approach.

(a) Unstructured Text Data Mining:

The most of the web content data is unstructured text data. The data mining techniques is used to represent the unstructured text into termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence, most of them consider the text mining as an instance of Web content

mining. The pre-processing step for any structured data is done by means of information extraction, text categorization, or applying NLP techniques which is used to provide efficient results.

(b) Semi-Structured and Structured Data Mining:

To represent the host pages on the web, the structured data is important, due to this reason it is consider as important and popular. Semi-structured data is a point of convergence for the Web and database communities: the former deals with documents, the latter with data. In the Object Exchange Model (OEM), the emergent representations for semi-structured data (such as XML) are different. In OEM, the data should be representing in the form of atomic or compound objects. The atomic objects may be integers or strings and the compound objects are used to refer the other objects through labeled edges.

3.2 Techniques

3.2.1 Classification

In the Web there are large number of XML documents are exist and XML has become the universal data format for a wide variety of information systems. Therefore the classification task is important for the information storage system. XML documents have both structures and contents as a typical type of semi-structured data. This paper presents a novel complete framework for XML document classification. The main goal of XML document classification is to build a classifier model that can automatically assign XML documents to some existing categories. Based on information retrieval methods, the most of the existing XML document classifiers work solely. The pre-processing phase of feature reduction and feature extraction has been successfully utilized to reduce the complexity of classification models and to improve the accuracy of the classification process. There are two major techniques in XML document classification includes:

- Content based Techniques.
- Structure based Techniques.

Typically, in the image point of view, the collected web images are useful for many applications which is classified into object classification and animal classification based on some classification techniques. Here, some of the classification techniques are explained in details in order to retrieve the information effectively from the web.

(i) Decision-tree Learning Algorithm

The decision-tree learning algorithm is driven by the precision/recall (PRDT) heuristic for XML document classification. In XML document classification, the motivation of choosing the precision and recall heuristic is mainly that XML documents have strong connections with text documents.

In this algorithm, the precision is represent as the percentage of the documents correctly classified to the positive class among all documents being classified to the positive class and the recall is defined as the percentage of the documents correctly assigned to the positive class among all documents of the positive class. Precision measures the "soundness" of the classifier, and recall measures the "completeness" of it. The main goal of this algorithm is to find a tree that can produce the best BEP value.

When precision and recall are equal or very close, this point is called the precision/recall-breakeven point (BEP) of the system.

(ii) Transductive and Semi-Supervised Inductive Classification Algorithm

The transductive classification algorithm is applied for image classification and it's worked based on the label propagation. The graph-based transductive algorithms have been applied to many applications, such as cartoon gesture recognition, content-based multimedia retrieval, image annotation and video annotation in the field of multimedia. This algorithm is more suitable for static image databases. But, both web and personal image databases are dynamic that is the number of web and personal images keeps increasing. Because of this problem, the transductive classification algorithms are not applicable to web and personal image annotation. Therefore, the semi-supervised inductive algorithm is mostly used for web and personal image databases.

The inductive algorithm is able to predict the labels of unseen data, which are outside the training set compared with transductive learning. It is therefore more suitable to apply the algorithm to dynamic image database annotation. A new inductive algorithm is used for image annotation by integrating label correlation mining and visual similarity mining into a joint framework.

3.2.2 Clustering

The text clustering is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering. It involves the use of descriptors and descriptor extraction. The descriptors are sets of words that describe the contents within the cluster. In general, there are two common algorithms in clustering. The first one is the hierarchical based algorithm, which includes single link, complete linkage and group average. The documents can be clustered into hierarchical structure by aggregating or dividing which is suitable for browsing. These algorithms can be further classified into,

- Hard clustering.
- Soft clustering.

The hard clustering computes a hard assignment that means each document is a member of exactly one cluster. The soft clustering computes a soft assignment that means a document's assignment is a distribution over all clusters and document has fractional membership in several clusters. Based on facial image, a clustering algorithm is used to estimate the relationship between the facial images and the names in their captions.

(i) Tag Path Clustering (TPC)

In a web document, the Tag Path Clustering (TPC) is used to extract all of the data records from all of the data regions. It is based on the hypothesis that a data region contains multiple contiguous or noncontiguous data records. The TPC algorithm works as follows:

- To build the DOM tree first and calculate the DOM paths of every node.
- Based on the DOM paths, the algorithm works on mining visually repeating information.
- The authors defined a visual signal as a triple (p,s, O), where p is a DOM path, s is its visual signal vector, and O is a collection that contains the individual occurrences.

- Then collect the visual signals and builds a similarity matrix between the visual signals using a similarity function.
- It discards clusters that contain less than three visual signals because it considers that a data record must contain at least three HTML tags.
- The ancestor and descendant relationships between visual signals in each cluster are introducing, and determine the visual signals that are the maximal ancestors.
- If there is a single maximal ancestor in a cluster, then the algorithm considers that some of the nodes in the occurrence collection in the maximal ancestor are data records and that some of these nodes may contain multiple data records.
- Finally, the algorithm tries to detect nested data records by applying some heuristic.

(ii) Clustering-based Approximation

A clustering-based approximation algorithm is used to improve the scalability and efficiency for large-scale problems. The clustering strategy could be applied in two different levels, namely:

- Image level.
- Name level.

In image level, all the 'n' facial images are directly separate into a set of clusters. In name level, first separate the 'm' names into a set of clusters, then to further split the retrieval database into different subsets according to the name-label clusters. Here, 'n' is the number of facial images in the retrieval database and 'm' is the number of distinct names (classes). The Bisecting K-means clustering based algorithm (BCBA) and the divisive clustering based algorithm (DCBA) are mainly used to improve the scalability and efficiency.

3.2.3 Phishing Websites

The phishing website fraud is a relatively new Internet crime which is a form of online fraud. The malicious people also known as phishers can create the phishing web pages that is forgeries of real web pages, to steal individuals' personal information such as bank account, password, credit card number, and other financial data. The methods for detecting phishing web pages can be classified into following,

- Industrial toolbar based anti-phishing
- User-interface-based anti-phishing
- Web page content-based anti-phishing.

Bayesian approach

For the detection of content-based phishing web page, the Bayesian approach is used and it is intelligence-based. The main goal of this algorithm is used to estimate the threshold, which is required in classifiers to determine the class of web page. The Bayesian approach has the following steps:

- A text classifier using the naive Bayes rule for phishing detection
- A Bayesian approach to estimate the threshold for either the text classifier or the image classifier such that classifiers enable to label a given web page as "phishing" or "normal."
- A novel Bayesian approach to fuse the classification results from the text classifier and the image classifier.
- A Bayesian approach that directly fuses the classification results instead of the similarity measurements.

4. COMPARATIVE STUDY

Table 1. Comparative Study of Algorithms

Techniques	Methods /Algorithm	Performance	Accuracy
Classification	Decision-tree Learning	High	High
	Transductive and Semi-Supervised	High	Medium
Clustering	Tag Path Clustering	Low	Medium
	Clustering-based	Medium	Medium
Phishing Websites	Bayesian	High	High

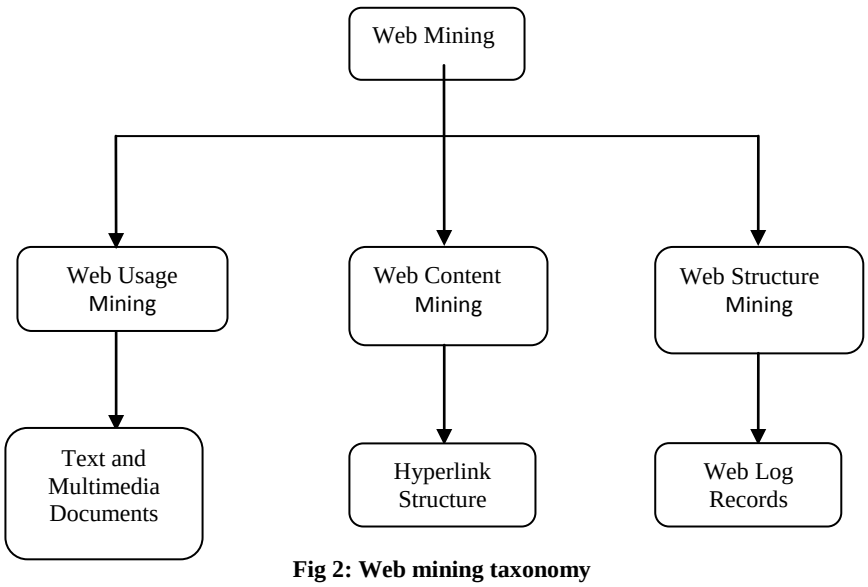


Fig 2: Web mining taxonomy

5. CONCLUSION

This paper discusses the techniques of the web content mining for retrieve the information effectively. Normally, the World Wide Web (W3C) consortium is the dynamic environment that is the content of the data is changing day by day. Because of this problem, the user should not get the desired information effectively. The Web content mining solves this problem and helps the users to fulfill their needs. The classification and clustering algorithm is used to retrieve the information effectively from the web and also it is used to improve the scalability. Based on the phishing websites, the user can get the irrelevant information. The phishing websites are detected and removed by using mentioned above techniques. Therefore, the user can get the desired information efficiently.

6. ACKNOWLEDGMENTS

My sincere thanks to Mr.D.Muthusankar for his valuable comments and contributed towards development of the template.

7. REFERENCES

[1] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, 2013 "Overview of Web Content Mining Tools," The International Journal of Engineering And Science.

[2] Claudia Elena Dinucă, Dumitru Ciobanu, 2012 "Web Content Mining," Annals of the University of Petroşani, Economics, PP.85-92.

[3] Danushka Bollegala,Yutaka Matsuo, and Mitsuru Ishizuka, 2013 "Minimally Supervised Novel Relation Extraction Using a Latent Relational Mapping," IEEE Transactions On Knowledge and Data Engineering, Vol. 25, No. 2, pp. 419-432.

[4] Darshna Navadiya, Roshni Patel, 2012 "Web Content Mining Techniques-A Comprehensive Survey," International Journal of Engineering Research & Technology," Vol. 1.

[5] Dayong Wang, Steven C.H. Hoi, Ying He, and Jianke Zhu, 2014 "Mining Weakly Labeled Web Facial Images for Search-Based Face Annotation," IEEE Transactions On Knowledge and Data Engineering, Vol. 26, No. 1, pp. 166-179.

[6] Dayong Wang, Steven C.H. Hoi, Ying He, Jianke Zhu,Tao Mei, and Jiebo Luo, 2014 "Retrieval-Based Face Annotation by Weak Label Regularized Local Coordinate Coding," IEEE Transactions On Pattern Aanalysis and Machine Intelligence, Vol. 36, No. 3, pp. 550-563.

[7] Govind Murari Upadhyay, Kanika Dhingra, 2013 "Web Content Mining: Its Techniques and Uses," Vol. 3, PP. 610-613.

[8] Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, 2011 "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach," IEEE Transactions on Neural Network, Vol. 22, No. 10, pp. 1532-1546.

[9] Hao Ma, Irwin King, and Michael Rung-Tsong Lyu, 2012 "Mining Web Graphs for Recommendations," IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No. 6, pp. 1051-1064.

[10] Hassan A. Sleiman and Rafael Corchuelo, 2013 "A Survey on Region Extractors from Web Documents," IEEE Transactions On Knowledge and Data Engineering, Vol. 25, No. 9, pp. 1960-1981.

[11] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions," Department of Computer Science, PP.51-71.

[12] Jemma Wu, 2012 "A Framework for Learning Comprehensible Theories in XML Document Classification," IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No. 1, pp. 1-14.

[13] Mohammad Khabbaz, Keivan Kianmehr, and Reda Alhajj, 2012 "Employing Structural and Textual Feature Extraction for Semistructured Document Classification," IEEE Transactions on Systems, Man, Cybernetics-PartC: Applications and Reviews, Vol. 42, No. 6, pp. 1556-1578.

[14] Niki R. Kapadia, Kinjal Patel, 2012 "Web Content Mining Techniques – A Comprehensive Survey," International Journal of Research in Engineering & Applied Sciences, pp. 1869-1877.

[15] Weiwei Zhuang, Yanfang Ye, Yong Chen, and Tao Li, 2012 "Ensemble Clustering for Internet Security Applications," IEEE Transactions On Systems, Man,and Cybernetics, Vol. 42, No. 6, pp. 1784-1796.

[16] Yan Wang, 2000 "Web Mining and Knowledge Discovery of Usage Patterns," CS 748T Project (Part I), PP. 1-25.

[17] Yi Yang, Fei Wu, Feiping Nie, Heng Tao Shen, Yueting Zhuang, and Alexander G. Hauptmann, 2012 "Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding," IEEE Transactions On Image Processing, Vol. 21, No. 3, pp. 1339-1351.