International Journal For Technological Research In Engineering
Volume 1, Issue 9, May-2014 *ISSN (Online): 2347 - 4718*

WEB CONTENT MINING USING RULE BASED CLASSIFIER

Patel Archana J[1], Asst. Prof Mukti Pathak[2]
Department of computer engineering
Hasmukh Goswami College of engineering
Gujarat, India

Absrtact: The rapid growth of web resources leads to a need of enhanced Search scheme for information retrieval. Every single user contributes a part of new information to be added to the web every day. This huge data supplied are of diverse area in origin being added, without a mere relation. So there is a need to develop such an algorithm through which web content mining can be in corporate.
The paper focuses on development and implementation of effective data mining algorithm of web content mining for classification and categorization concept.
Index Terms: Data mining, clustering, web mining, Object Exchange Model (OEM).

I. INTRODUCTION

Data mining is the process of extracting hidden useful information from volume of database. Data mining is a process that consists of applying data analysis and discovery algorithms that under acceptable computational efficiency limitations produce a particular enumeration of patterns over the data. Data mining and knowledge discovery is about creating a comprehensible model of the data. Such a model may take different forms from simple association rules to complex reasoning system. This aspect aims at making the process of knowledge extraction continually maintainable and update as new data become available. We refer to this process as knowledge learning.

Data mining process has two major Techniques: classification and clustering. Classification is the derivation of function or model which determines the class of an object based on its attributes. Classification is a data mining technique to predict group membership for data instance. For Example you may wish to use Classification to Predict Whether the Weather on a Particular day will be "sunny", "rainy" or "cloudy". Clustering is the identification of classes, also called cluster or groups for a set of object whose classes are unknown. The goal of data mining classifiers is to predict the class value of a new or unseen instance, whose attribute values are known but the class value is unknown. Classification consists of assigning a class label to a set of unclassified class. Supervised Classification is the set of possible classes is known in advance.

In an Unsupervised Classification the set of possible classes is not known. It is also known as a Clustering. Data mining is the process of extracting hidden useful information from large volume of database. Data mining is the process of discovering knowledge from large amounts of data stored either in databases or warehouses. Data mining is becoming an increasingly important tool to transform these data into information. Data mining can also be referred as knowledge mining or knowledge discovery from data [3]. Data mining has provided so many tools that allow a great variety of analysis techniques. Data mining and knowledge discovery is about creating a comprehensible model of the data.

II. DATA MINING PROCESS

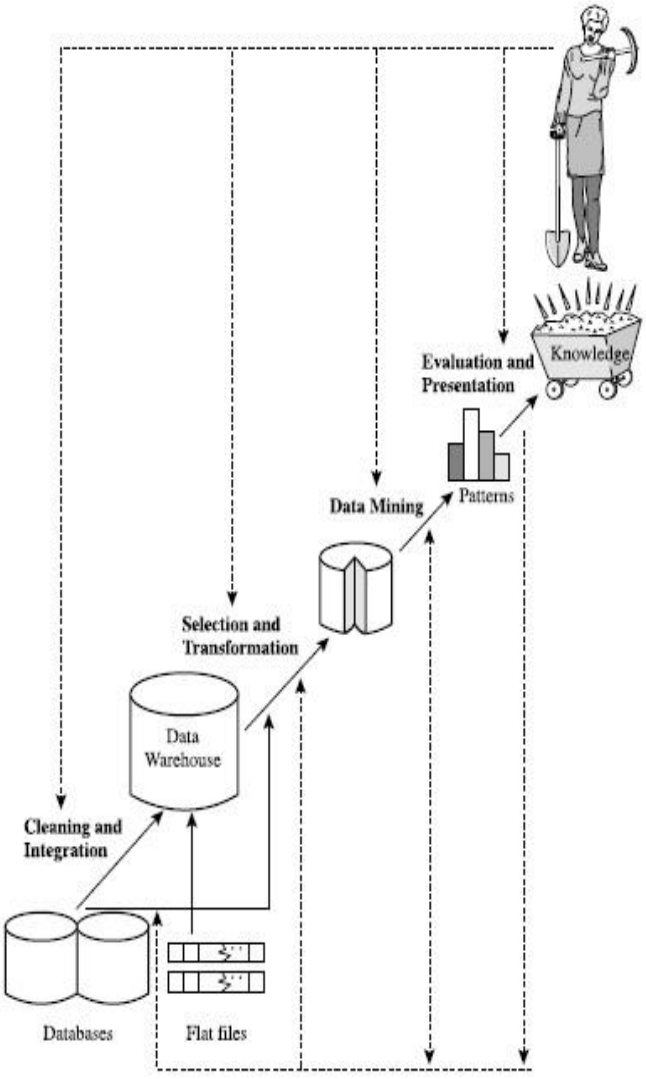


Fig. 1. Data mining process

- **Data Cleaning:** Data cleaning is use to remove noise and inconsistent data.
- **Data Integration:** Where multiple data source may be combined.
- **Data Selection:** Where data relevant to the analysis task are retrieved from the database.
- **Data Transformation:** Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining:** Data mining is an essential process where intelligent methods are applied to extract data patterns.
- **Pattern Evaluation:** To identify the truly interesting patterns representing knowledge
- **Knowledge Presentation:** Where visualization and knowledge representation techniques are used to present mined knowledge to users.

III. TYPES OF WEB MINING

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Web mining has been explored to different techniques have been proposed for the Variety of the application. Web usage mining, Web structure mining and Web content mining are the types of Web mining.

A. Web Content mining tools to Improve techniques of web data mining

Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions, in short, mining the Web log data. Web mining consists of the different essential tasks, which are described in a fig. below.

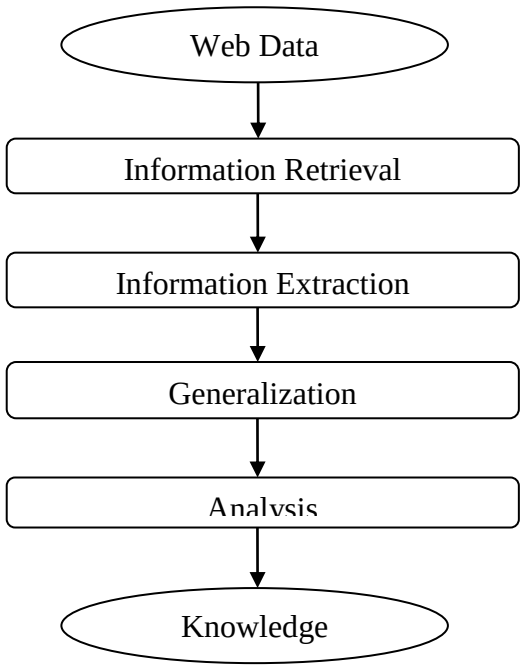


Fig 2-Web Mining Tasks

Information Retrieval
It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web.
Information Extraction
It is the task of automatically selecting and pre-processing specific information from retrieved Web resources.
Generalization
It is the task to automatically discover general patterns of individual Web sites as well as across multiple sites.
Analysis
It is the task of analyzing, validating and interpreting the mined patterns.

B. Web Content Mining Techniques

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. The several approaches in web content mining are represented. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

1. Unstructured Data Mining Techniques
Content mining can be done on unstructured data such as text. Mining of unstructured data give unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic Content Mining is a type of text mining. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

Information Extraction
To extract information from unstructured data, pattern matching is used. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. This technique is very useful when there is large volume of text. IE is the basis of many other techniques used for unstructured mining?

Topic Tracking
Topic Tracking is a technique in which it checks the documents viewed by the user and studies the user profiles. According to each user it predicts the other documents related to users interest. In Topic Tracking applied by yahoo, user can give a keyword and if anything related to the keyword pop up then it will be informed to the user.

Summarization
Summarization is used to reduce the length of the document by maintaining the main points. It helps the user to decide whether they should read this topic or not.

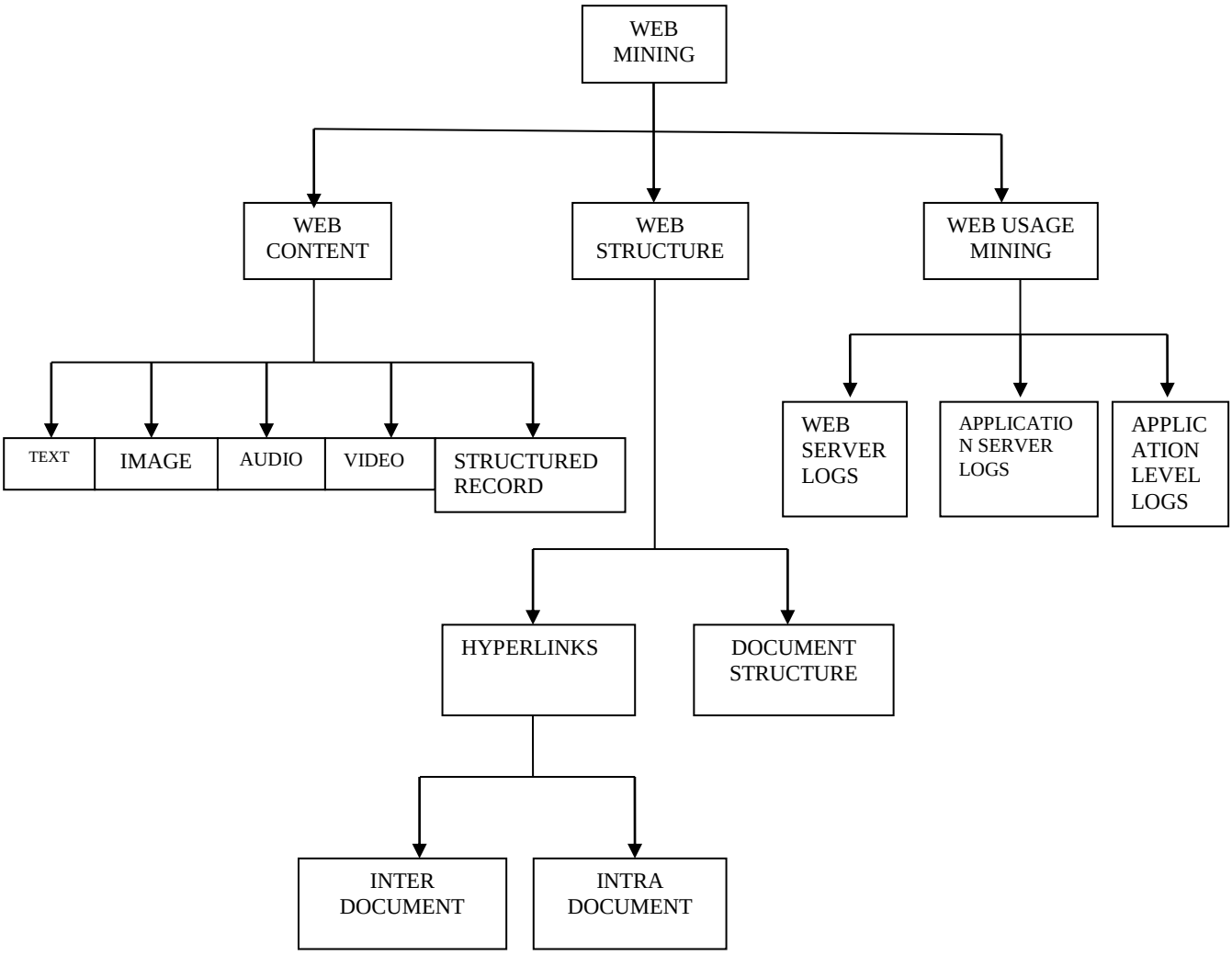


Fig 3- WEB MINING

Categorization

Categorization is the technique of identifying main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does not process the actual information. It decides the main topic from the counts. It ranks the document according to the topics.

Clustering

Clustering is a technique used to group similar documents. Here in clustering grouping is not done based on predefined topic. It is done based on fly. Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering helps the user to easily select the topic of interest. Clustering technology is useful in management information system.

Information Visualization

Visualization utilizes feature extraction and key term indexing to build a graphical representation. Through visualization, documents having similarity are found out.

2. Structured Data Mining Techniques

The techniques used for mining structured data are Web Crawler, Wrapper Generation, Page content Mining.

Web Crawler

There are two types of Web Crawler which are called as External and Internal Web crawler. Crawlers are computer programs that traverse the hypertext structure in the web. External Crawler crawls through unknown website. Internal crawler crawls through internal pages of the website which are returned by external crawler.

Wrapper Generation

In Wrapper Generation, it provides information on the capability of sources. Web pages are already ranked by traditional search engines. According to the query web pages are retrieved by using the value of page rank. The sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, statistics, index look up about the sources.

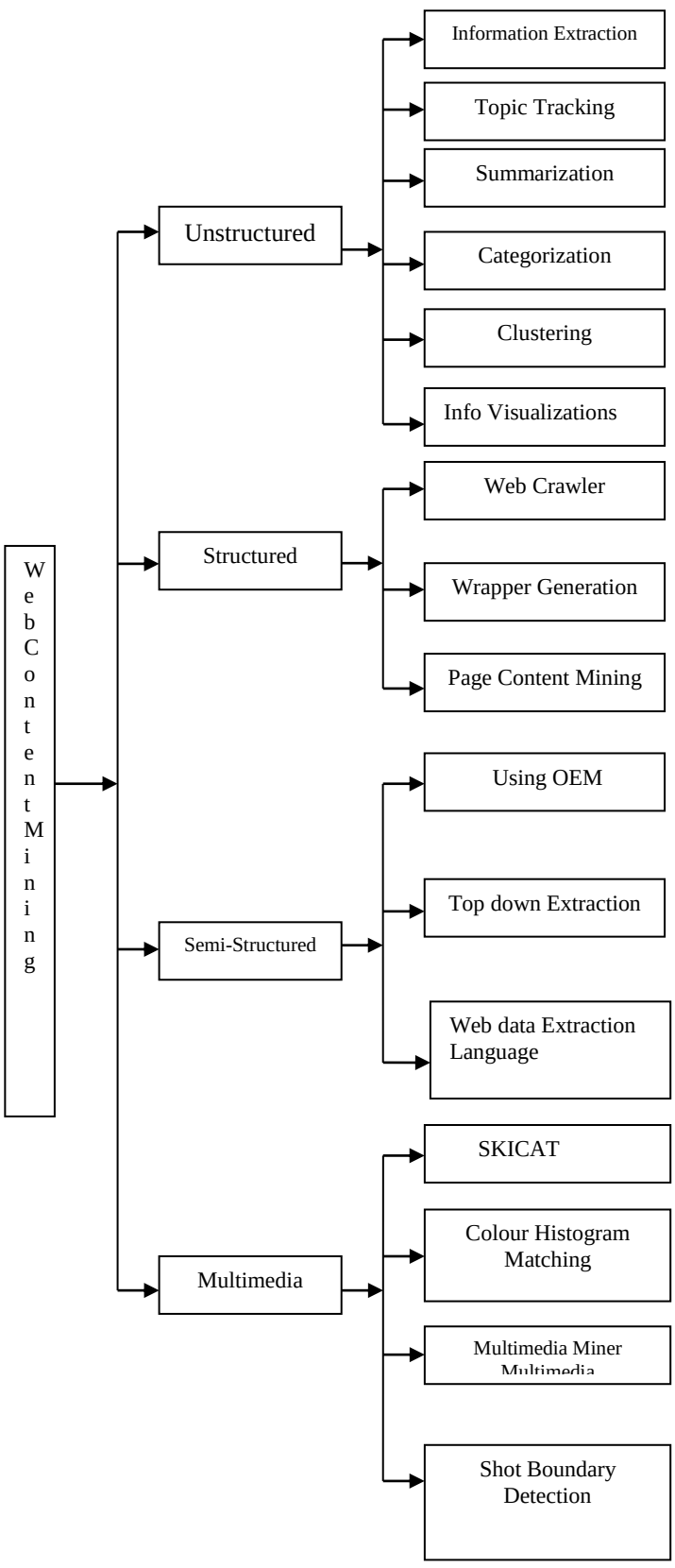


Fig 4-Web Content Mining Techniques

Page Content Mining

Page Content Mining is structured data extraction technique which works on the pages ranked by traditional search engines. By comparing page Content rank it classifies the pages.

3. Semi-Structured Data Mining Techniques

The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language.

Object Exchange Model (OEM)

Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange model (OEM). It helps the user to understand the information structure on the web more accurately.

Top down Extraction

In top down extraction, it extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted.

Web Data Extraction Language

In Web data extraction language it converts web data to structured data and delivers to end users. It stores data in the form of tables.

4. Multimedia Data Mining Techniques

Some of the Multimedia Data Mining Techniques are SKICAT, color Histogram Matching, Multimedia Miner and Shot Boundary Detection.

SKICAT

It is a data analysis and cataloging system which produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set.

Color Histogram Matching

It consists of Color histogram equalization and Smoothing. Equalization tries to find out correlation between color components. The problem faced by equalization is sparse data problem which is the presence of unwanted artifacts in equalized images. This problem is solved by using smoothening.

Multimedia Miner

It is Comprises of four major steps. Image excavator for extraction of image and Video's, a preprocessor for extraction of image features and they are stored in a database, a search kernel is used for matching queries with image and video available in the database. The discovery module performs image information mining routines to trace out the patterns in images.

Shot Boundary Detection

It is a technique in which automatically the boundaries are

detected between shots in video

IV. SEMANTIC WEB ARCHITECTURE

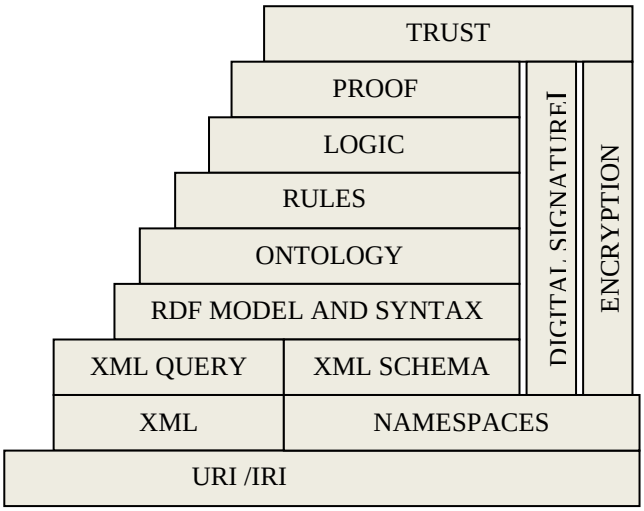


Fig 5-Semantic Web Architecture

The idea of searching optimized information on the web using Capture-Recapture method. This paper also explains an optimized semantic searching of keywords and detailed explanation and simulation on Ontology of Indian Universities. Ontology is the bases of semantic web and can be further expanded. Example detailed ontology was simulated using protégé and results were analyzed. We have also addressed Capture-Recapture method for estimating the Key sets probabilities in Semantic Searching Technique. This web searching technique by assigning priority to the web pages based on the high probabilities of the Search Keysets.

Proposed algorithm

Step 1: Input for the proposed work will be URL/URL's (Uniform Resource Locator) name given by User. For the same input URL is scanned.

Step 2: The scanned result is processed and frequency of words is calculated.

Step 3: The Initial Phase is to apply mapping of concepts through Semantic mapping. It can be done to turn the conventional web services to Semantic web services which could favour reasoning. It is followed by a simple classification that organizes the mapped concepts in to the specified categories. The Classification can be done based on a pre-defined order. Now, the web services are arranged in an order. In order to check the semantic discovery of results from the underlying concepts that are standardized, a query interface be designed.

Step 4: The Semantic Similarity measure between words supplied in the query is observed. The relevant results matching the given query are retrieved. For more efficient results; these matched results are sorted in an order of ranking applied to them. It can be achieved through matching the key words from query to the similar words Synonyms from the Library being added. For Example: If the key word from query supplied be 'Money'. It can be matched to the similar words like 'Currency'.

Step 5: The related result found with measure of semantic similarity can be ranked. More relative result and most viewed results for the supplied query can be found. It can also be ordered according to their computed value of importance.

Step 6: As an outcome of entire process URL can be classified.

V. IMPLEMENTATION METHOD

A. Technology Used

Operating System
- Window XP /Later Version

Tool
- Eclipse Tool

Language of Implementation
- java

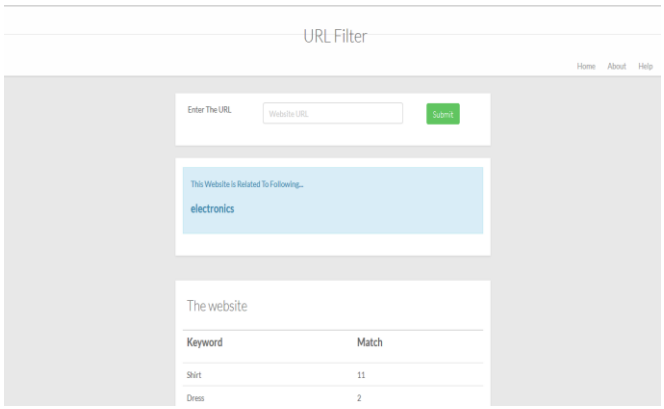
B. System Requirements

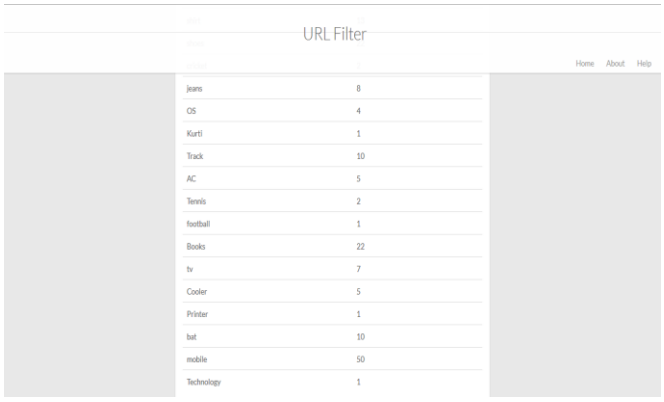
Supported Operating Systems: Windows XP, Windows Vista, Windows 7

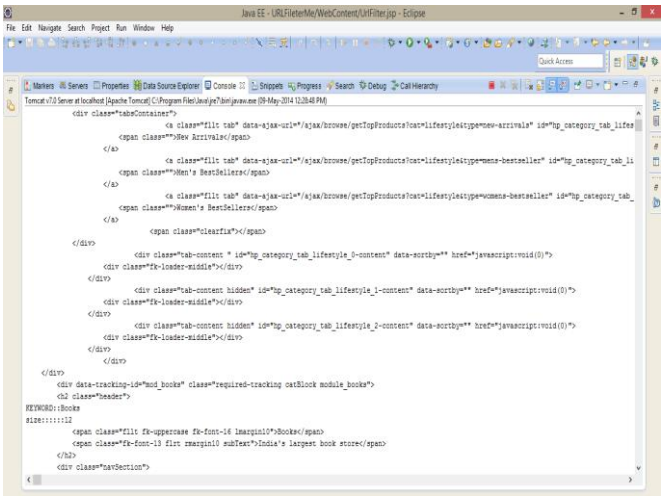
Hardware Requirements:

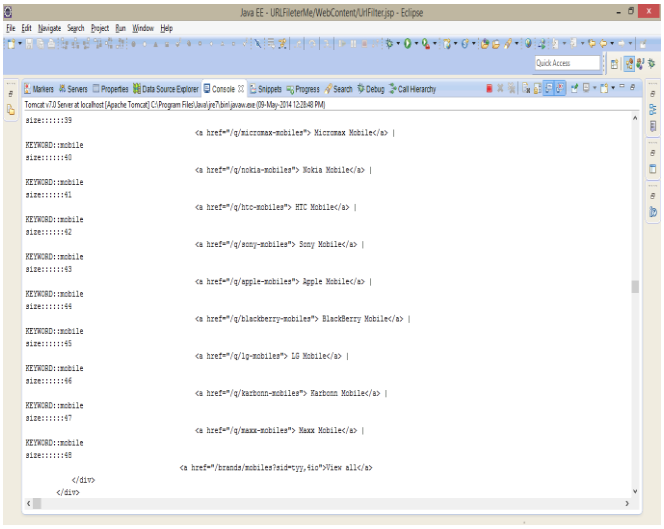
Recommended Minimum: Pentium 1 GHz or higher with 512 MB RAM or more Minimum disk space: x86 – 850 MB, x64 – 2 GB

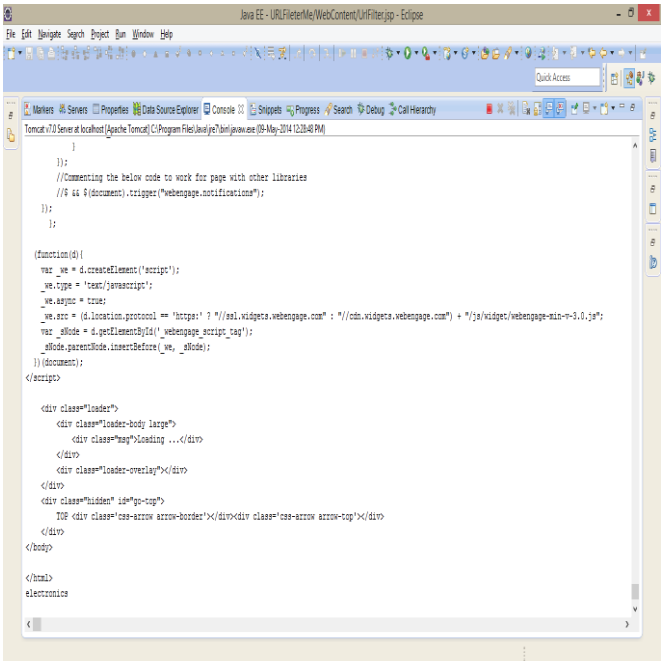
C. Experimental Results











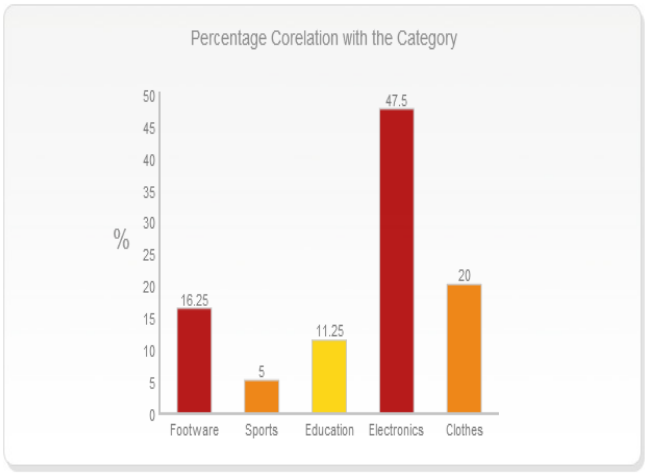


Fig. 5. Percentage correlation with the category (Flipkart)

VI. CONCLUSION AND FUTURE WORK

The rapid growth of web resources leads to a need of enhanced Search scheme for information retrieval. Every single user contributes a part of new information to be added to the web every day. This huge data supplied are of diverse area in origin being added, without a mere relation. So there is a need to develop such an algorithm through which web content mining can be in corporate. To develop an effective data mining algorithm of web content mining for classification and categorization concept.

REFERENCES

[1] Amit Biswas, Dewan Md. Farid and Chowdhury Mofizur Rahman, "A New Decision Tree Learning Approach for Novel Class Detection in Concept Drifting Data Stream Classification," Journal of Computer Science and Engineering, Volume 14, Issue 1, July 2012.

[2] Nikita Jain, Vishal Srivastava Data Mining Techniques: A Survey Paper IJRET.

[3] Jiawei Han. Data Mining: Concepts and Techniques. Department of Computer science, University of Illinois at Urbana- Champaign, 2006.

[4] Tan, Steinbach, Kumar, "Introduction to Data mining",http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap6_basic_association_analysis.pdf

[5] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter1/sld019.htm.

[6] Monika Yadav, Mr. Pradeep Mittal, Web Mining: An Introduction, ijarcsse Volume 3, Issue 3, March 2013.

[7] Abdel hakim Herrouz, 2Chabane Khentout Mahieddine Djoudi, Overview of Web Content Mining Tools (IJES) ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[8] Arvind Kumar Sharma, P.C. Gupta Study & Analysis of Web Content Mining tools to improve techniques of Web data mining. IJARCET, Volume 1, Issue 8, October 2012.

[9] Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki, Web Personalization Using Web Mining: Concept and Research Issue, International Journal of Information and Education Technology, Vol. 2, No. 5, October 2012.
[10] Siddharth Gupta1, Narina Thakur2, Using Capture-Recapture Method for Web Intelligence, 2010 IEEE.
[11] Sonal Tiwari, Prashant Richariya, A Web Usage Mining Framework for Business Intelligence, 2011 IEEE.
[12] Lynnda Wagner & Jean-Paul Van Belle, Web Mining for Strategic Intelligence, *Online, 25 (2), 27-32.*
[13] Abdul-Aziz Rashid Al-Azmid, DATA, TEXT, AND WEB MINING FOR BUSINESS INTELLIGENCE: A SURVEY, (IJDKP) Vol.3, No.2, March 2013.
[14] Chao Wang, Jie Lu, Guangquam zhang, Mining Key information of web Pages: A method and its Application. NSW 2007.
[15] Jenice Aroma R, Mathew Kurian,*A Semantic web: Intelligence in Information Retrieval*, 2013 IEEE International Conference on Emerging Trends in Computing.
[16] Gerd Stumme, Andreas Hotho, Bettina Berendt Semantic Web Mining February 2006.
[17] http://www.ontotext.com/kim/architecture
[18] Dunham, M.H.2003. Data Mining Introductory and Advanced Topics. Person Education

