Intelligent Websites' Content Analyzing and Classification system

Abbas Mohsen Al-Bakry
University of Babylon College of Information Technology
abbasmoh67@yahoo.com

Talib Turki Hashush
University of Babylon College of Science Dept. of Computer Science
talib_alfatlawii@yahoo.com

Abstract

The Web can consider as infinite repository of information because of the vast amount of information that they can contain, but with this advantage of Web, it also some drawbacks, because of unregulated nature of the Web, it is led to make any one make any contents available on the Web even though this content is objectionable for people like violence and pornographic contents. Therefore existing of Web content classification and filtering systems is necessary. In this paper we produce Website classification system using Fuzzy C – Means based on textual features, to classify the Website to two categories: (white) that not contain pornographic materials and (black) that contain pornographic materials.

key wrds: Intelligent Websites , fuzzy c-means , Content Analyzing

1. Introduction

With the rapid development of information technology, people became more independent on internet in their life, where surfing the internet takes more attentions from people, and may be they using the internet more than watching TV. However the internet Websites can contain harmful material like pornographic materials, terrorism, racism violence,...etc. For this reason existing of systems that could diagnose and filtering the Website that contain the offensive material became necessary especially with the flooding of pornographic materials on the Web with reference to exist (4,200,000) pornographic Websites in the world in 2012[Akbulut2012]. In this paper we proposed intelligent Website analysis and content classification, using the Fuzzy C – Means for classification Website, and taking the pornographic Website as a case study.

This paper organized as follows section 2 explains the related work, section 3 explain the Fuzzy C – Means, section 4 explains the proposed systems, and finally section 5 presents the experimental results.

2. Related Works

In this section we show some of researches in this field.

- P.Y. Lee, et al.(2005): present bilingual categorization engine to perform offline Web page analysis and near – instantaneous online filtering to handle both English and Chinese Web pages, taking a pornographic Website as a case study of
objectionable contents, and employing artificial neural network to perform intelligent clustering of Web pages\cite{Lee2005}.

- Deepak Agrawal (2008): presented Web data clustering using FCM and Proximity Hints from textual contents and Hyperlink – structure to improve the clustering done by FCM, the design algorithm consists of two phases that are realized in interleaved manner. The first phase is data driven and is primarily the standard FCM applied to the parents. The second concerns an accommodation of the proximity – based hints and involves some gradient oriented learning\cite{Agrawal2008}.

- A.C.M. Fong, et al. (2010): Describe intelligent offline filtering agent, which consist of two main modules: automated Webpage crawling to collect URL list from search engine and download all the Web pages of the collected URL and intelligent classification module to classify these Web pages based on ANN, and send the result to online filtering agent\cite{Fong2010}.

- Ali Ahmadi, et al. (2010): design intelligent Web pages classification using contextual and visual features to detect pornographic Web pages. It presented a Web page filtering system based on combination of textual, profile, and visual features. They employs a hierarchical set of classifiers, an ID3 classifier for textual and profile features, and neural networks model for skin color and visual features\cite{Ahmadi2010}.

3. Fuzzy C – Means

In this section we explain the fuzzy clustering algorithm. The fuzzy clustering plays important roles in many application domains such as pattern recognition, machine learning, data mining, computer vision and computational biology have used clustering algorithms\cite{Pimental2013}. The Fuzzy c – means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. The main objective of FCM is to minimize the following objective function:

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^m \ d_{ij} \hspace{1cm} 1 \leq m < \infty$$

$$d_{ij} = \|x_i - v_j\|$$

Where m is any real number greater than 1, $\mu_{ij}$ is the degree of membership of $x_i$ in the cluster $j$, $x_i$ is the $i^{th}$ of $d$ – dimensional measured data, $v_j$ is the d – dimension center of the cluster, and $\|\|\|$ is any norm expressing the similarity between any measured data and the cluster center. The membership must have the following characteristics:

$$\mu_{ij} \in [0,1] \hspace{1cm} \forall \ i = 1,2,\ldots,n; \hspace{1cm} \forall \ j = 1,2,\ldots,c.$$ 

$$\sum_{j=1}^{c} \mu_{ij} = 1 \hspace{1cm} \forall \ i = 1,2,\ldots,n.$$ 

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \hspace{1cm} \forall \ j = 1,2,\ldots,c.$$
Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $\mu_{ij}$ and the cluster centers $v_j$ by:

$$
\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{m-1}}, \quad v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m x_i}{\sum_{i=1}^{n} \mu_{ij}^m}
$$

This iteration will stop when $\| U_{(k+1)} - U_{(k)} \| < \varepsilon$, where $\varepsilon$ is a termination criterion between 0 and 1, where $k$ is the iteration steps. This procedure converges to a local minmum or a saddle point of $J_m$.

The algorithm is composed of the following steps[URL12013]:

1. Initialize $U=[U_{ij}]$ matrix, $U^{(0)}$.
2. At $k$ – step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$.

$$
v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m x_i}{\sum_{i=1}^{n} \mu_{ij}^m}
$$

3. Update $U_{(k)}$, $U_{(k+1)}$.

$$
\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{m-1}}
$$

4. If $\| U_{(k+1)} - U_{(k)} \| < \varepsilon$ then Stop, otherwise return to Step2

4. The Proposed System

The proposed system using fuzzy C-Means algorithm to classify the Websites depending on the textual feature that exist on the Webpage and other pages that linked to that Website either the Webpage that composed the Website or that belong to other Websites. In our work we classified the Web page to two groups(clusters), each page is implementing as a vector of indicative terms frequencies, (indicative term is the word that frequently appeared in pornographic Web pages). Where each page is analyzing and converting to indicative terms vector then pushed to classification model. the system is consist of two stage:
The first stage is used to building the classifier model. In this stage we take 100 Webpage, 50 is normal Web page and 50 is pornographic Web page. Converting it to indicative terms vector and assign to each vector two random membership value, because we have two clusters. After that we applying fuzzy clustering algorithm to find the value of cluster center to each cluster(group).

The second stage is classification stage. In this stage the system reads the URL from a given database, download it from Internet and parsing it to get its Webpage and Website that linked to it.

Then the system check the following things in the page:
1) Checking **Title**: where the title of the Web page usually contains the subject and contain word that describe the Web page and is found under the `<title>` tag[Lee05].
2) Checking Metadata: Meta data contain information about the Web page, for example **Meta description** contain a short description about the Web page, it allows the developer to summarize the content that can be found on the page, also **Meta Keywords** contain the keywords that provide short and accurate information about the Web page. [Han2010,URL2013].
3) Checking Warnings: the warnings message block is used to alert visitors to the explicit contents of the Website and the legal aspects involved. Is used to relieve any legal responsibility resulted from visitors viewing the pornographic contents. It is contain number of legal terms like "Warning This Site Contains Adult Content", "sexually explicit material".After checking these things and ensuring the Web page is not refer to pornographic material then classifying the Web page using FCM, in FCM convert the Web page to indicative terms vector and computing its membership to each cluster, using the cluster center computing in stage (a).

If the main page classified as pornographic or the Website contain two pages that classified as pornographic, then the Website at all considered as pornographic Website, otherwise it consider normal Website. Figure (4.1) shows block diagram of the proposed system.
5. Conclusion

Due to the self–regulating nature of the Internet and the explosive growth in the number of websites, also the openness of the web allows user to easily get information anytime and anywhere. However these things have disadvantage, for example, children can get inappropriate information like pornographic contents in the web. according to the Korea Communications Standards Commissions, 83.4% of respondents encounter harmful contents on the web. Therefore there is an urgent need for effective web content filtering tools to protect people from the adverse effects of exposure to harmful web contents, especially with flooding of pornographic Websites in the Internet. An effective Websites categorization mechanism lies at the heart of a
web content filtering tool. Using Websites classification increasing the efficiency of the Web Content Filtering System.

6. Implementing and Results

The system implementing in java language, using MySql language to create database, and using external library called jsoup to parsing the Webpage. In this work we taking (700) Websites, (327) pornographic Websites and (373) normal Websites. And when applying our system to classify these Websites, and measure the following criteria.

- **False rejection**: implement count of the white Websites that classified as black.
- **False acceptance**: implement count of the black Websites that classified as white.
- **Correct rejection**: implement count of the black Websites that classified as back.
- **Correct acceptance**: implement count of the white Website that classified as white.

Table (6.1) implement the result of this system.

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Count</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>False rejection</td>
<td>26</td>
<td>7.58</td>
</tr>
<tr>
<td>False acceptance</td>
<td>10</td>
<td>2.80</td>
</tr>
<tr>
<td>Correct rejection</td>
<td>317</td>
<td>92.42</td>
</tr>
<tr>
<td>Correct acceptance</td>
<td>347</td>
<td>97.2</td>
</tr>
</tbody>
</table>

The overall system has the following results (94.85) correct classification and (5.15) is false classification.

Reference

Akbulut Akhan Akbulut, et al., 2012, "Agent Based Pornography Filtering System", IEEE.


Agrawal Deepak Agrawal, 2008, "Web Data Clustering using FCM and Proximity Hints from Text as well as Hyperlink – structure", IEEE.


Fong A.C.M. Fong, et al. 2010,"An intelligent offline filtering agent for Website analysis and content rating", IEEE.

URL1"http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html", Date of access 7/3/2013.

