Web Mining for Cyber Monitoring and Filtering

Tien Dung Do
School of Computer Engineering
Nanyang Technological University
Singapore
pa0001852a@ntu.edu.sg

Kuiyu Chang
School of Computer Engineering
Nanyang Technological University
Singapore
askychang@ntu.edu.sg

Siu Cheung Hui
School of Computer Engineering
Nanyang Technological University
Singapore
asschui@ntu.edu.sg

Abstract—Like any self-regulating environment, the Internet is fertile ground for all kinds of potential abuse ranging from get-rich-quick scams, touting of illegal or adult-oriented material, promotion of extremist/ anarchist views, to online pimping, etc. Consequently, the ability to discreetly intercept and analyze Internet access has tremendous potential in shielding users, especially our youngsters, from inappropriate content. This paper proposes one such system, the Web Access Monitoring and Filtering (WAMF) system. The WAMF system comprises two main decoupled components, one for online monitoring and filtering, and the other for offline Web classification and data analysis. The former tracks, tallies, and selectively blocks user Web access in real-time, whereas the latter employs Web mining techniques to classify Web pages into pre-defined user categories and analyze user Web access data for user behavior patterns. In this paper, we will discuss the WAMF system, and in particular, Web mining techniques for adaptive Web page categorization.

Keywords—Web mining; Cyber monitoring; Cyber filtering; Web categorization

I. INTRODUCTION

According to a 2001 IDC survey, up to 40 percent of worker productivity is lost due to non work-related Internet access. Outside of work, a 2000 US congressional study showed that one in five of the 24 million children with Internet access have been solicited online for sex. In order to fight off the rising tide of Internet abuses, we have come up with the Web Access Monitoring and Filtering (WAMF) system, which is capable of monitoring, tracking, analyzing, and selectively blocking individual or aggregated user Web access. The WAMF is comprised of two components, one for online monitoring and filtering, and the other for offline Web classification and data analysis. The former tracks, tallies, and selectively blocks user Web access in real-time, whereas the latter employs Web mining techniques to classify Web pages into pre-defined user categories and analyze user Web access data for user behavior patterns.

The WAMF employs Web mining techniques for offline Web classification and data analysis. More specifically, the Web content/structure mining technique is used for classifying Web pages, whereas Web usage mining technique is used for characterizing and analyzing users’ access behavior. The use of Web mining techniques makes it a reasonably robust tool for analyzing user Web access behavior. In addition, what makes the WAMF stand out is its adaptive classification engine, which assigns a category to each URL, thereby making interpretation and analysis significantly more intuitive and interpretable.

In this paper, we will discuss the WAMF system. In particular, we will discuss in detail the use of Web mining techniques for Web categorization, which implements the adaptive Web classification engine of the WAMF system. The rest of this paper is organized as follows. Section II discusses the related techniques for cyber monitoring and filtering. Section III gives the overall system architecture of the WAMF system. Section IV discusses the adaptive Web page classification process. The performance results will also be given. Finally, Section V concludes the paper.

II. CYBER MONITORING AND FILTERING

A cyber monitoring and filtering system usually sits between the WWW and Web browser, with the sole objective of monitoring and/or selectively blocking Web accesses. Web pages can be blocked based on URL alone even before retrieval, or after the requested HTML page is returned and analyzed.

The system usually includes a monitoring and filtering component. The network-monitoring component monitors network traffic between Web browsers and the Internet. The filtering component then analyzes each incoming Web page to decide if blocking is necessary. Current Web filtering systems, such as Cyber Patrol [1], Cyber Snoop [2], and WebChaperone [3] use four major filtering approaches including the Platform for Internet Content Selection, URL blocking using reference lists, keyword filtering, and intelligent content analysis.

Platform for Internet Content Selection (PICS) [4] entails a set of specifications created by the World Wide Web Consortium [4] to define a platform for the creation of a Web content rating system. It enables Web publishers to associate rating labels or metadata with Web pages targeted at a particular audience, e.g. adult Web pages. However, the adoption of PICS is not regulated and it is possible for some publishers to mislabel their Web content either by intent or mistake. PICS should therefore only be used as a supplementary tool in any Web filtering system.

URL blocking is used to restrict or allow access to a requested Web page by comparing its URL (and equivalent IP address) with those in a reference URL list. The reference URL list may be a “blacklist” of disallowed Web pages or a “white-list” of allowable Web pages. Most systems that adopt the URL blocking approach use a blacklist as reference. At any rate, human experts are needed to compile the reference list.
Once a reference list is available, this method can effectively filter out blacklisted Web pages as soon as the user specifies a URL, without even having to download the page. Thus, it can be a very fast approach. However, with the dynamic nature and explosive growth of the Internet, it is very difficult to keep the reference list up-to-date, much less comprehensive. This means the effectiveness of this approach will deteriorate over time unless there is an efficient way of maintaining the reference list.

Keyword filtering blocks access to Web pages based on the occurrence of words and phrases in the Web content. When a Web page has been successfully retrieved from the remote Web server, every word or phrase is compared against those in a keyword dictionary of prohibited words and phrases. If the number of matches reaches a predefined threshold, access to that Web page is blocked. Like URL blocking, this is an intuitively simple technique to implement. However, its accuracy rests on the quality of the dictionary of keywords and phrases. In particular, this approach is prone to “over-blocking” due to a lack of semantic understanding of the context in which certain keywords appear. For example, a health-related Web page that contains many occurrences of the keyword “sex” might be misconstrued as being pornographic.

Intelligent content analysis is an attempt at achieving semantic understanding of the context in which certain keywords appear. In particular, intelligent classification techniques can be used to categorize Web pages into different groups (e.g. pornographic and non-pornographic) according to the statistical occurrence of sets of features. Statistical methods such as K-Nearest Neighbor (KNN) classifier [5, 6], Linear Least Squares Fit (LLSF) [7, 8], Linear Discriminant Analysis (LDA) [9, 10], and Naïve Bayes (NB) classifier [11] have been introduced in this field of research. In addition, Neural Network (NN) models [12, 13, 14] have also been found to be particularly suitable for categorizing real-world data characterized by incomplete and noisy data. However, the use of statistical and NN techniques in a real-time filtering system can be computationally intensive and often incur intolerable latency.

In this research, we have adopted an integrated approach for Web monitoring and filtering. The proposed approach uses the Web mining techniques to classify Web pages into categories (e.g. pornography, games, finance, entertainment, etc.) based on both the content and structure of Web pages. This process is done in an offline mode. The generated categories are then used by the online filtering component for blocking inappropriate Web accesses according to a reference list of categories. The proposed integrated approach is much more efficient than most current systems, as the categories can be updated automatically, and the online filtering component does not need to spend precious CPU time to perform online classification.

III. WAMF SYSTEM ARCHITECTURE

Figure 1 shows the architecture of the WAMF system. It is made up of two major components for online logging, monitoring and filtering, and offline Web classification and analysis. The online component comprises the following modules: Online Logging and Filtering and Online Monitoring, while the offline component contains the Adaptive Classification Engine and Web Access Analysis module. The operations of the various components are described as follows.

A. Online Logging and Filtering

The Online Logging software module can be planted in individual client computers or installed on a gateway/proxy server.

In client side monitoring, Web access activities are collected from individual PCs and sent to a centralized Log Server for aggregated storage as User Web Access Logs. Each access will be tagged with the PC’s unique IP address so that analysis can be done on the individual or aggregated level. We have used Microsoft’s Browser Helper Object (BHO), which is a component that integrates into the Internet Explorer Web browser, for recording Web access on the client/user side. Online Filtering is performed individually at the client side, and is thus more susceptible to user bypassing and hacking.

In proxy side monitoring, Web access activities are collected at the proxy/gateway and also sent to the Log Server. Moreover, the proxy server can also perform Online Filtering to filter inappropriate user Web accesses based on a list of pre-classified categories of URLs on pornography, games, etc. Different from client side monitoring, this arrangement cannot be easily circumvented.

The client side monitoring is more suited for the home PC users whereas the proxy side monitoring is more applicable for large organizations, ISPs, and schools, where administrators/teachers can graphically monitor and filter employees/students Internet usage in real-time.

B. Online Monitoring

The Online Monitoring module, implemented as a Java applet, can display Web access statistics over a Web browser in real-time. Authorized users just need to log on to view in real-time the historical Web access statistics of a specific user or group of users. Further, it can be configured to send out
alerts via email or mobile SMS (short messages) if a prescribed trigger event such as a user accessing pornographic sites for extended period of time is fired.

C. Web Access Analysis

After a reasonably representative set of User Web Access Logs has been collected, the Web Access Analysis module can be invoked to analyze user behavior and identify surfing trends tied to categories. For example, Figure 2 shows the average number of visits to various types of websites over a two-week period at a major Internet Service Provider. From the figure, it can be seen that most people visit weather sites twice per day (observe the peak at the 8-hour interval), followed by those who check for weather updates once (peak at the 22-hour interval) daily.

More advanced Web usage mining techniques [15] based on association rule mining [16] and sequential pattern mining [17] are currently under investigation. These advanced Web usage mining techniques are capable of extracting frequent user access sequence patterns or behavior according to the time-of-day or day-of-week criteria.

D. Adaptive Classification Engine

The key component of the WAMF is the Adaptive Classification Engine (ACE), which automatically classifies Web pages into multiple pre-defined and/or customized categories. ACE analyzes each Web page’s content and context using Web mining techniques, more details of which will be given in Section IV.

IV. WEB MINING FOR WEB CATEGORIZATION

Web mining techniques have been applied to develop the Adaptive Classification Engine (ACE) that classifies Web pages into categories such as pornography, games, sports, finance, etc. These categories can be “allowable” or “objectionable” according to some criterion. The classification information is then used by the Online Filtering module to block objectionable Web pages if so configured.

Figure 3 shows the classification model of ACE. ACE currently employs three complementary classifiers, namely Kohonen’s Self-Organizing Map (KSOM), Fuzzy Adaptive Resonance Theory (Fuzzy ART), and the Associative Classifier (AC) for Web page categorization. All three classifiers work on the same set of pre-processed input data, which comprises largely of informative terms extracted from Web pages. This section will first discuss the data preparation task for the classifiers. The implementation techniques for the three classifiers will then be discussed. And the performance of the three classifiers in comparison with some of the existing filtering systems will also be given.

A. Data Preparation

Data cleansing is the most important step of any data-mining endeavor. For ACE, text features are extracted from Web pages as content and contextual information.

Content information is obtained from semantic content of Web pages. A Web page dedicated to a particular subject/discussion (called domain) usually contains a representative set of words and phrases, known as category keywords that characterize that topic. This set of terms is considered to be high content information and usually shared by other Web pages dealing with the same domain. Therefore, the category keywords can be viewed as a unique collection of features characterizing Web pages belonging to a domain. This gives rise to the idea of using a unique set of content-rich terms to distinguish a particular type of Web page from others.

For example, consider pornographic Web pages, which contain many sexually explicit terms such as “xxx” and “erotic”. In order to make use of such sexually explicit terms in the content analysis process for Web filtering, it is necessary to compile a list of such terms that appear most frequently among pornographic Web pages. To avoid introducing too much noise (irrelevant words) to this list of indicative terms, we need a systematic approach to determine the inclusion of a specific term in the list. We achieve this by collecting and analyzing the statistical data on the usage of indicative terms commonly found in pornographic Web pages. The indicative terms identified in pornographic Web pages can be classified into two major groups according to their meanings and usage. The majority group comprises sexually explicit terms with sexual connotations or related to sexual acts, and the other group mostly comprises legal terms, which are used to establish legitimacy and renounce liability. The reason why legal terms are found in the pornographic Web pages is because they tend to have a warning message block in their entry page that states the legal conditions governing the access to the sexually explicit materials contained therein.

Figure 2. An example of a categorized Web access interval analysis of 4000 users over a 2-week period.

Figure 3. Classification model of the Adaptive Classification Engine.
Contextual information is usually derived from the HTML structure and metadata. It has been shown to contain additional discriminative information useful for classification [18]. For example, terms appearing in the title of a Web page are deemed more important than those included in the content body. In our case, we look at the following HTML tags: title, key word/description, graphical text, image tooltip, warning block, and viewable text content.

The complete data preparation process includes the following steps. In the first step, domain knowledge on the content and context information is gathered. A list of indicative terms with respective weights (called content weights) is defined for each domain. Only indicative terms are extracted and processed from each Web page. A list of context tags (i.e. title, key word/description, graphical text, etc.) with corresponding weights (called context weights) is also specified. The context weights may be similar over different domains as it formulates common layouts and structures for Web pages. For example, terms appearing in the title tag are relatively more relevant to a domain compared to those included in the body. Therefore, the title tag’s context weight should be set much higher than the other tags. Moreover, the context weight list also depends on the domain. In pornographic Web pages, terms appearing in warning messages or image tooltips may be more relevant to the domain than those in game Web pages, for example.

In a typical bag-of-word approach, a Web page is represented by terms that occur in the Web page with corresponding frequencies. In our system, both content and contextual information are attached to the extracted terms. While the content weight is the same for a term in a domain, the context weight of each term depends on its position in the Web page (i.e. context tag).

After the domain information has been collected, indicative terms are extracted from Web pages using feature extraction and pre-processing. In feature extraction, the content of a Web page is parsed, taking into account the various contextual blocks such as title, warning message block, metadata, and image tooltips. The extracted features then collectively form a raw signature of the Web page. While in pre-processing, the raw signature of the Web page is further tokenized and processed to extract indicative terms. That is, only terms contained in the indicative term lexicon are kept, and the rest are discarded. Indicative terms are keywords or phrases representative of a particular category or domain. Each term is recorded with (i) a simple frequency count (number of occurrences); (ii) a context tag in which it was contained. Notice that the content weight is constant for each term with respect to a domain and can be obtained from the domain knowledge.

Finally, data transformation is performed to convert the extracted data into a suitable form for training the ACE classifiers. The inputs to the neural network and AC models are slightly different, and will be discussed in subsequent subsections.

B. Artificial Neural Network (ANN) Classifiers

The ANN models used in ACE include KSOM [19, 20, 21] and Fuzzy ART [21, 22]. For each Web page, the indicative term frequencies tallied from the pre-processing step are sorted and packed into a vector representation. The vectors are then fed as training inputs to a neural network.

Each indicative term is represented as a dimension in the input vector. The weight for each dimension is calculated as the product of the corresponding term frequency and associated content weight. The content weight attempts to boost the influence of relevant terms and reduces sensitivity to less relevant terms, which help improve classifier accuracy. In addition to the terms, context tags are also represented as dimensions in the vector.

Before we can use ACE to classify a Web page, we have to first train each of the classifiers. The training process allows each classifier to learn by example the distinction among Web page categories.

The ANN classifiers comprise processes from training to categorization as follows:

1. ANN training (model generation) – Indicative vectors of Web pages, along with their known category labels, are fed into the ANN for model adjustment and generation.
2. Category Assignment – Each cluster generated from the ANN are labeled depending on which category is the majority member. However, clusters with narrow winning margins (difference between winning and loosing category below a certain threshold) are discarded.
3. Categorization – A new Web page is classified using the trained ANN.

C. Associative Classifier

In Associative Classification (AC) [23, 24], association rules are used to construct the classifier. An association rule in the form of \( \text{itemset } \Rightarrow c \) (where itemset is a set of items, \( c \) is a class) implies that if a transaction contains the itemset, then it probably belongs to the class \( c \). In order to implement AC, indicative terms extracted in the pre-processing step are considered as items, whereas sets of terms representing Web pages are considered as transactions.

While ANN models use content and context weights to form input vectors, the AC model factors term weights into the calculation of support and also in rule sorting. In other words, instead of computing the support, \( \text{supp}(s) \), of an itemset \( s \) as the frequency (number of occurrences of the \( s \) in the data set), we compute it as the product of the frequency of \( s \), \( \text{freq}(s) \), and the minimal value of content weights among items in \( s \), i.e.

\[
\text{supp}(s) = \text{freq}(s) \times \min\{\text{content-weight}(i) | i \in s\} \tag{1}
\]

Equation (1) assigns a higher support for itemsets containing indicative terms so that “high indicative itemsets” are more likely to surface to the top in the ranking. The calculation also satisfies the anti-monotone constraint [25], which means that the association rule mining process can be implemented using any Apriori-like algorithm.
In order to distinguish between the various context tags, we constrain the rule mining process to operate only on itemsets in the same context tag. That is, we only keep rules $itemset \Rightarrow c$ such that all items in $itemset$ belong to the same context tag. Consequently, this segregation of rules allows the context tag weights to be factored into support counting, which would emphasize rules from more important context tags.

Typically, for associative classifiers, the mined association rules are sorted in decreasing order of confidence. During the classification phase, a new Web page will be assigned to the class of the first rule (in the order) in which the itemset (antecedence of the rule) is contained in the transaction representing the Web pages. The order of the association rules, therefore, strongly affects the classifier outcome. In our AC model, we use term weights to bias the confidence ranking of the association rules. Thus, rules with more indicative itemsets are more likely to have higher ranking and, therefore, contribute more to the classifier.

Before we can use the AC to classify any Web pages, we have to first train the model, i.e. mine the rules from a training set. The associative classifier comprises processes from training to categorization as follows:

1. Association Rules Generation – Association rules are mined from the training set in the form of $itemset \Rightarrow c$ where all items in $itemset$ belong to the same context tag and class $c$ is a web page category.
2. Classifier Construction – The association rules obtained from the last step are used to form the classifier.
3. Categorization – A new Web page is classified using the mined association rules.

D. Performance Evaluation

To measure the classification performance of the Adaptive Classification Engine, we have implemented KSOM and Fuzzy ART classifiers and the associative classifier. We have used the domain on pornographic Web pages to test the classifiers. We conducted the experiment using a total of 4786 Web pages including 1009 pornographic and 3777 non-pornographic Web pages. We then tested the classifiers using a test exemplar set which consists of 535 pornographic Web pages and 523 non-pornographic Web pages. The results are summarized in Table 1.

Table 1 shows that our KSOM approach is able to correctly classify 96% of all test Web pages, whereas Fuzzy ART and associative classification accuracy rates are of 91.6% and 88% respectively. The high accuracy rates of the three classifiers show that the Web mining approach is able to utilize thoroughly the Web page information. The performance also illustrates the learning capability of ANN models with high dimensional inputs. The accuracy rate of the associative classification is not as good as the ANNs due to the various sizes of Web pages. Some Web pages contain a small number of indicative terms. The consequence is that very few (or none) rules of the classifier are covered by those Web pages, thereby leading to an inaccurate classification.

Using the same collection of Web pages, we have also tested five commercially available Web filtering systems. Web Chaperone is the only system in our test that employs some form of machine intelligence in content analysis. It utilizes a proprietary mechanism called Internet Content Recognition Technology (iCRT), which performs a heuristic analysis on an incoming Web page. iCRT analyzes a Web page’s attributes which include word count ratios, length of page, structure of page, and contextual phrases. The results of each attribute analysis are then aggregated according to the attribute weights. Based on the overall results obtained, WebChaperone determines whether the Web page contains pornographic material.

Clearly, our results have verified the effectiveness of using Web mining techniques for Web page categorization and filtering. We are currently investigating a majority voting method to combine the results from the three classifiers. For example, to classify a Web page, all three classifiers are used to assign appropriate classes individually. The final assigned class is the one garnering the most number of votes. It is well-known that an ensemble of classifiers considerably increases the classification accuracy [26]. We believe that the ensemble of classifiers will considerably increase the accuracy of the overall system.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Major approach</th>
<th>Web page</th>
<th>Pornographic</th>
<th>Non-pornographic</th>
<th>Overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correctly</td>
<td>Incorrectly</td>
<td>Correctly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>classified</td>
<td>classified</td>
<td>classified</td>
</tr>
<tr>
<td>WAMF</td>
<td>KSOM</td>
<td></td>
<td>95.33%</td>
<td>4.67%</td>
<td>96.75%</td>
</tr>
<tr>
<td>WAMF</td>
<td>Fuzzy ART</td>
<td></td>
<td>88.60%</td>
<td>11.40%</td>
<td>94.65%</td>
</tr>
<tr>
<td>WAMF</td>
<td>Associative Classification</td>
<td></td>
<td>90.0%</td>
<td>10.0%</td>
<td>86.0%</td>
</tr>
<tr>
<td>Cyber Patrol</td>
<td>URL blocking with referent list</td>
<td></td>
<td>81.5%</td>
<td>18.5%</td>
<td>94.0%</td>
</tr>
<tr>
<td>SurfWatch</td>
<td>URL blocking with referent list</td>
<td></td>
<td>85.5%</td>
<td>14.5%</td>
<td>95.7%</td>
</tr>
<tr>
<td>Cyber Snoop</td>
<td>Keyword filtering</td>
<td></td>
<td>93.5%</td>
<td>6.5%</td>
<td>82.3%</td>
</tr>
<tr>
<td>CYBERsitter</td>
<td>Context-based key-phrase filtering</td>
<td></td>
<td>91.5%</td>
<td>8.5%</td>
<td>85.0%</td>
</tr>
<tr>
<td>WebChaperone</td>
<td>iCRT</td>
<td></td>
<td>88.5%</td>
<td>11.5%</td>
<td>94.7%</td>
</tr>
</tbody>
</table>
V. CONCLUSIONS

In this paper, we have described an effective Web monitoring and filtering system, the purpose of which is to develop a system so that objectionable Web pages can be effectively identified and blocked. We have developed a system that decouples the classification and analysis process from the monitoring/blocking process. The former is performed by the Adaptive Classification Engine and Web Access Analysis module, whereas the latter is done by an Online Logging and Filtering Agent, and the Online Monitoring module.

We have analyzed the characteristics of the targeted Web pages coupled with the learning capability of the artificial neural networks and associative classifier. Web pages are processed using the popular bag-of-words approach, supplemented with extra content and contextual information. This approach is domain-specific as the content information includes a list of indicative terms for each domain. This utilization of relevant terms increases the accuracy of the classification; while at the same time reduces the effect of noise from less relevant terms. The approach also makes full use of the structure of Web documents by extracting the contextual information in the form of different weighted HTML tags. We are currently investigating a majority voting method to combine the results from the three different classifiers in order to achieve a much more accurate and robust classifier system.

The ability to intercept and analyze Internet traffic via the very same network has enormous potential. Apart from monitoring and filtering out objectionable Web content, it can also be used to monitor activities of employees at the workplace to reduce the amount of non-work-related usage, or for parents to monitor online chat activities between their children and strangers. Ultimately, if our system is deployed at a national gateway, it can be used as a form of security surveillance to fight cyber crime and terrorism. Moreover, privacy is a sticky issue if the system is to be deployed at the ISP side. Fortunately, there are no privacy concerns at the workplace, school, and home (user initiated) environments, which currently stand to benefit the most from our system.

ACKNOWLEDGEMENT

We would like to thank Professor Joydeep Ghosh of the University of Texas at Austin for his invaluable guidance in computing the graph of Figure 2.

REFERENCES


