A structural and content-based analysis for Web filtering

P.Y. Lee
S.C. Hui and
A.C.M. Fong

The authors

P.Y. Lee is a Research Student and **S.C. Hui** is an Associate Professor, both in the School of Computer Engineering, Nanyang Technological University, Singapore.
A.C.M. Fong is a Lecturer at the Institute of Information and Mathematical Sciences, Massey University, Auckland, New Zealand.

Keywords

Web sites, Filters, Classification, Neural networks, Content analysis

Abstract

With the proliferation of objectionable materials (e.g. pornography, violence, drugs, etc.) available on the WWW, there is an urgent need for effective countermeasures to protect children and other unsuspecting users from exposure to such materials. Using pornographic Web pages as a case study, this paper presents a thorough analysis of the distinguishing features of such Web pages. The objective of the study is to gain knowledge on the structure and characteristics of typical pornographic Web pages so that effective Web filtering techniques can be developed to filter them automatically. In this paper, we first survey the existing techniques for Web content filtering. A study on the characteristics of pornographic Web pages is then presented. The implementation of a Web content filtering system that combines the use of an artificial neural network and the knowledge gained in the analysis of pornographic Web pages is also given.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1066-2243.htm

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · pp. 27-37
 · ISSN 1066-2243
DOI 10.1108/10662240310458350

Introduction

The World Wide Web (WWW or Web) has become an extremely popular communications medium due to its wide geographical coverage and continuous availability. In addition, it is very easy for anyone to put up information on the Web to reach a wide audience that continues to grow in size. However, the self-regulating nature of the Web community, coupled with the ease of making information available on the Web, has led some individuals to abuse their freedom of expression by putting up harmful materials on the Web. These include violence, gambling, drugs and pornographic materials. It is therefore urgently important to provide effective Web content filtering to protect children and other unsuspecting users from the adverse effects of such harmful materials.

Web content filtering would also benefit companies that want to reduce the overheads associated with their employees' non-work-related access to the Internet. Unauthorized access to the Internet not only adds costs to the company in terms of increased data traffic charges, but also loss of productivity by the offending employees. Sometimes, when valuable bandwidth resources are used for unauthorized access to the Web, other employees who have a legitimate need to access the Internet might also be affected.

In this paper, we survey existing methods and systems that attempt to provide Web filtering to users. Our survey reveals that it is necessary to analyze and characterize the offending Web pages in order to develop effective filtering techniques against them. Using pornographic Web pages as a case study, we present the results of our study of the distinguishing features of this type of offending Web materials. With this knowledge, we have developed an intelligent classification engine for effective Web filtering.

The rest of this paper is organized as follows. We first present a survey of the existing Web filtering approaches. This is followed by a thorough analysis of the characteristics of pornographic Web pages. We then describe our Web filtering system that combines the use of an artificial neural network (NN) and the knowledge gained in the analysis of

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · 27-37

pornographic Web pages. Finally, we conclude the paper.

Web filtering approaches

Four major approaches have been developed for Web content filtering:
(1) platform for Internet content selection (PICS);
(2) uniform resource locator (URL) blocking;
(3) keyword filtering; and
(4) intelligent content analysis.

PICS (W3C, 1997) is a set of specifications created by the World Wide Web Consortium (W3C) to define a platform for the creation of content rating systems. It enables Web publishers to associate labels or meta data with Web pages to limit certain Web content with explicit nature targeted at adult audiences from reaching other groups of Internet users. 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 the 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 "white-list" of allowable Web pages or a "blacklist" of disallowed pages. Most systems that adopt the URL blocking approach use a blacklist as reference. At any rate, the reference list is compiled by human experts, making this a laborious approach. Once a reference list is available, URL blocking can effectively filter out blacklisted Web pages as soon as the URL is specified by the user without having to load the page for analysis. Thus, it can be a very fast approach. However, with the explosive growth of the Internet, it is very difficult to keep the reference list current and complete all the time. This means the effectiveness of this approach will fall over time unless there is an efficient way of maintaining the reference list.

Keyword filtering is a filtering approach that blocks access to Web pages based on the occurrence of offensive 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 has reached a predefined threshold, access to the Web page is blocked. Like URL blocking, this is an intuitively simple technique to implement. However, its accuracy rests on the effectiveness of identifying 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 misidentified as 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) classification (Yang, 1994, 1999), linear least squares fit (LLSF) (Yang and Chute, 1992, 1993), linear discriminant analysis (LDA) (Fukunaga, 1990; Koehler and Erenguc, 1990) and naïve Bayes (NB) probabilistic classification (McCallum and Nigam, 1998) have been introduced in this field of research. In addition, NN models (Dalton and Deshmane, 1991; Lippmann, 1987; Salton, 1989) are well suited to providing categorization on real-world data characterized by incomplete and noisy data. However, the use of statistical and NN techniques can be computing-intensive and often incur intolerable latency.

In this research, we decouple the Web page classification process from the filtering process to achieve fast and effective filtering from the users' point of view. The classification process is conducted offline by NN, whose learning capabilities allow it to adapt to noisy data and acquire human-like intelligence in distinguishing the nature of a Web page by semantic understanding of the context in which keywords appear. We investigate two popular NN models that are effective classifiers:
(1) Kohonen's self-organizing map (KSOM) (Kohonen, 1995; Flexer, 2001); and

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · 27-37

(2) fuzzy adaptive resonance theory (fuzzy ART) (Carpenter *et al.*, 1991).

This offline intelligent classification process is used to create and maintain a knowledge base of prohibited URLs without the need for human expertise or supervision. This means the online filtering process can be very fast and effective. To achieve this, we need to study the characteristics and distinguishing features of pornographic pages.

Analysis of pornographic Web pages

We attempt to identify the characteristics of pornographic Web pages by analyzing the textual and page layout information contained in such Web pages. We also investigate the adoption rate of PICS, with an understanding that Web page classification could not totally rely on it. PICS could be used for positive identification of pornographic Web pages. We present the results of our analysis on a sample of Web pages from 200 different pornographic Web sites.

Page layout

Like other Web pages, pornographic Web pages can be classified into two layout formats: single-frame and multi-frame. When viewing a Web page with single-frame layout in a Web browser, the browser only needs to download one HTML document from a single URL address to construct the entire Web page content from that document. On the other hand, to view a multi-frame Web page, an HTML document containing information of other HTML documents that make up the contents of the whole Web page is downloaded first. The information includes the URL addresses of the HTML documents, as well as the position data of where to display the specific document in the browser window. According to this piece of information, the browser fetches all the necessary HTML documents and constructs the Web page.

Consequently, all the HTML documents used in constructing a multi-frame Web page must be treated as a single entity. This is an important consideration because when multi-frame Web pages are encountered in the data collection and analysis process, the statistics obtained from any aspect of the Web page should be derived as a whole from the aggregation of data collected from every individual HTML document making up the whole Web page.

We have studied how widespread multi-frame Web pages are among pornographic Web sites in order to determine whether it is necessary to gauge the importance of incorporating processing capability of such Web pages in our system. We found that while an overwhelming majority of pornographic Web pages adopt the single-frame format (86 per cent), 13 per cent used a two-frame format and 1 per cent used a three-frame format (Lee, 2002). The sizable minority meant that it was necessary to incorporate multi-frame processing capability into our system development.

PICS usage

We have collected statistical data to gauge the adoption rate of PICS among pornographic Web sites. We found that PICS is only adopted by 11 per cent of the pornographic Web sites surveyed (Lee, 2002). Since many of the publishers of pornographic Web sites may not want their contents to be filtered out by a Web filtering system, they are reluctant to provide such support in their Web sites. We therefore focus on the textual analysis of pornographic Web pages instead.

Indicative key terms in textual context

A Web page that focuses on a major subject carries a specific set of words and phrases that characterize the subject of discussion in the contents provided. This set of terms is usually found to be common among other Web pages on the same subject. Therefore, a specific set of terms can be viewed as a unique collection of features characterizing Web pages that emphasize the same subject and lead to a similar theme related to that set of terms. This gives rise to the idea of using a unique set of terms to distinguish a particular type of Web page from others.

This observation is applicable to pornographic Web pages, since these Web pages contain many sexually explicit terms such as "xxx" and "erotic". In order to make use of such sexually explicit terms in the content

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · 27-37

analysis process of the Web filtering system, it is necessary to compile a list of such terms that appear most frequently among the pornographic Web pages. To avoid introducing too much noise to this list of indicative terms, we need a systematic approach to determine the inclusion of a specific term in the list. We do 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 group of the majority comprises sexually explicit terms which are those with sexual meanings or related to sexual acts, and the other group mostly consists of legal terms which are terms used to establish legitimacy. 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.

Indicative terms may be found in both the displayed and non-displayed textual contents. The displayed indicative terms may be found in the Web page title, warning message block, graphical text and other viewable textual contents. Non-displayed items are stored in the URL, meta data of "description" and "keywords" and image tooltip. We can determine whether some textual contents are displayed or not by checking the markup language tags. For HTML code, displayed textual contents are those not contained within an HTML tag while non-displayed textual contents are found inside an HTML tag. An HTML tag is defined as a block that begins with a "<" and subsequently ends with a ">" in the HTML code of a Web page. For example, the text in <B>Hit control-D to bookmark this site!</B> is displayed, whereas the text in <A HREF="http://www.ntu.edu.sg/"> is not displayed.

Web page title

The title of a Web page usually contains the subject of the Web page and thus may consist of important indicative terms describing the nature of the Web page. It is displayed in the title bar at the top of the Web browser window. When analyzing the HTML code, the textual information on the title is found between the opening tag <TITLE> and closing tag </TITLE> in the header section of the code. For example, the indicative terms contained in <TITLE>xxx adult entertainment</TITLE> can easily be extracted.

Warning message block

A warning message block is commonly found in the entry page of pornographic Web sites to relieve the publishers from any legal responsibility when users access the pornographic contents. It is usually located in the body section of the HTML source code of the page. Typically, there are quite a number of legal terms such as "age of 18" and "adults" as well as sexually explicit terms such as "sexual".

Other displayed textual contents

In addition to the title and warning message block, pornographic Web pages also contain other viewable textual contents rich in sexually explicit terms. Similar to the warning message block, these displayed textual contents are located in the body section of the HTML code of the Web page.

Meta data of "description" and "keywords"

A Web site can provide two meta data fields in the HTML header section of their Web pages which facilitate the indexing process of some search engines. They are meta data "description" and meta data "keywords". These meta data are not visible in a Web browser window since they are contained inside a HTML tag. The "description" meta data carries a string of text describing the contents of the associated HTML document e.g. <META NAME="keywords" CONTENT="adult, porn, sex, video">. A search engine can use this string to provide a document summary to the users in the results of a query. Similarly, the "keywords" meta data has a string of text which consists of comma-delimited words or phrases related to the contents of the Web page, e.g. <META NAME="description" CONTENT="XXX Adult Entertainment is like nothing you have ever seen before.">. A search engine can record these terms during the indexing process and allow a user to perform a keyword search. The contents of both "description" and "keywords" meta data

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · 27-37

contain terms directly related to the subject of the associated Web page. Thus, they can be used to determine the content nature of the Web page since they carry important indicative terms. However, not all the Web pages of a Web site have these two meta data fields. In fact, they are usually found only in the entry page of a Web site and both tend to appear together on the same page.

URL

A Web page is uniquely identified by an address called a URL, which serves the purpose of locating the specific Web page on the Web. A URL can be represented in either absolute form (e.g. http://www.w3c.org/rfc/index.htm) or relative to some known base URL (e.g. /rfc/index.htm). These two forms can be distinguished by the fact that an absolute URL always begins with a scheme name followed by a network DNS name which points to a hosting server. Examples of scheme name are "http://", "ftp://" and "gopher://". Both the absolute and relative URLs may contain a directory pathname and a filename to identify the location of the specific document on the hosting server.

A Web page usually contains hyperlinks, which are the URL addresses pointing to other Web documents hosted either by the same Web server (absolute or relative URL) or other remote servers (absolute URL). It is therefore reasonable to assume that the text extracted from a URL can sometimes identify content nature of the associated Web page. This information can be from the server DNS name, the directory pathname, as well as the filename in the URL string. For example, the URL above contains the substring "rfc" as its directory pathname indicating that the Web page contains information related to Request For Comments documents. Furthermore, the Web page can have multiple hyperlinks embedded in the body section of its HTML code, which usually point to other related Web pages with similar subjects. We therefore analyze the indicative terms in the sample Web page's URL as well as those URLs embedded in the HTML code.

Image tooltip

Sometimes, an image of a Web page shown in a Web browser has a string of text associated with it, which will only be displayed with user interaction. This string of text is defined as an image tooltip. It is used to describe the image associated with it and is shown to the user as a popup window when the mouse pointer hovers over the image. It is usually found inside the image HTML tag <IMG> located in the HTML body section of the Web page, e.g. <IMG HEIGHT="220" WIDTH="332" SRC="http://www.viewporn.com/sasv/images_front/table3_sale_animation.gif" BORDER = "0" ALT="hidden cameras, live shows! enter here!">.

Since the image tooltip describes the associated image, it provides useful information on the nature of the image. For Web sites that heavily use graphical text, an image tooltip can prove valuable for content filtering systems that primarily rely on textual content analysis.

In the above example, the image tooltip of an associated image is actually found in the ALT attribute of an HTML <IMG> tag. Thus, the data collection process should look at the text string of this attribute for indicative terms.

Indicative terms in graphical contents

Occasionally, indicative terms can be extracted from the contents in graphical or image form, which is defined as graphical text in this analysis. Graphical text is used by some of the pornographic Web sites to replace a portion of their entry page contents, with the intention of either to beautify the overall design of the Web page or to jeopardize a Web filtering system that depends on textual content analysis.

Like other Web filtering systems, our approach depends primarily on textual content analysis. Thus, its effectiveness will be adversely affected by widespread use of graphical information. Fortunately, our statistical analysis (see the section below) has revealed that only 1.74 per cent of the pornographic pages use graphical text.

Statistical analysis

We have collected a sample of 200 pornographic Web sites for statistical analysis based on the discussion of indicative terms given in the above section. A pornographic Web site is one whose contents satisfy at least one of the following conditions:

Internet Research: Electronic Networking Applications and Policy
Volume 13 · Number 1 · 2003 · 27-37

- sexually oriented contents;
- erotic stories and textual descriptions of sexual acts;
- images of sexual acts, including inanimate objects used in a sexual manner;
- erotic full or partial nudity; and
- sexually violent text or graphics.

In our study, we are interested in the number of unique indicative terms found in the sample Web pages, as well as their frequency of occurrence which is given by the number of times a specific term appears in the Web pages. In this research, we focus on English Web sites, and English is rich in terms of morphology. So, there are often similar words that can be found among the contents of the sample pages, e.g. "pornography" and "pornographic" as well as "sexy" and "sexiest" are not uncommon. Thus, morphed versions of a base word are treated as the same as the base word and contribute to the frequency of occurrence of the base word in the sample pages when collecting the statistical data. Of course, there are a few exceptions. Words such as "sexual" and "sexy", although they are from the base word "sex", are treated as different terms in the data collection process for two reasons:

(1) The occurrence of the base word is not only common among pornographic Web sites, but also among others. One example is "sex" which can also appear very frequently in a health related Web site. Viewing "sexy" and "sexual" differently from "sex" can actually reduce such type of noise incurred by treating all the three terms similarly.
(2) The morphed version of a base word may have significance as well as density difference from the base word in different portions of a Web page content. Studying such differences can contribute to more accurate identification of targeted Web pages.

In addition, apart from single-word indicative terms, phrasal terms that comprise two or more words need to be studied. This is because a phrase can give a more specific meaning and more direct indication of the Web page contents than the individual words that make up the phrase. When collecting data of phrasal terms, a phrase containing words similar to another single-word term does not contribute to the statistical data of that term. For example, the phrasal term "adult material" does not contribute to the statistics of the single-word term "adult".

Our study has identified a total of 55 indicative terms comprising 42 sexually explicit terms (e.g. "porn") and 13 legal terms (e.g. "of legal age") with 88.9 per cent and 11.1 per cent frequencies of occurrence. Among the 55 indicative terms, 95.9 per cent are single-word with the remainder being phrasal terms. Table I summarizes the usage of indicative terms in the eight locations described.

From Table I, more than half of the 55 indicative terms can be found in all of the eight locations except graphical text, with other viewable textual contents and meta data "keywords" containing more than 81 per cent and 70 per cent of the indicative terms, respectively. Further, more than 55 per cent of the indicative term occurrences are found in these two locations. This indicates that the two locations are quite densely populated with indicative terms. The other six locations have their indicative term occurrences ranging from about 1 per cent to 11 per cent. Among these six locations, the graphical text location contributes the least occurrences of indicative terms. Therefore, the probability of graphical text affecting the effectiveness of textual content analysis capability of a Web filtering system is negligible. Also, only about half of the indicative term occurrences are in the displayed textual contents portion of the sample Web pages, namely, title of Web page, warning message block, and other viewable textual contents. Thus, features in the non-displayed textual contents, which are meta data "description" and "keywords", URLs, and image tooltip, also provide important information on the nature of the Web pages.

Discussion

We have investigated the characteristics of pornographic Web pages based on three attributes: page layout formats, PICS usage and indicative terms in textual contents. We have also conducted statistical analysis on a sample of 200 pornographic Web sites based on these attributes.

Although there are not many Web pages that use the multi-frame layout format, the Web

Table I Usage of indicative terms in eight locations

Location	Number of unique terms (total: 55 indicative terms) *n*	Per cent	Frequency of occurrence (per cent)
Title of Web page	30	54.55	6.61
Warning message block	31	56.36	10.95
Other viewable textual contents	45	81.82	32.19
Meta data "description"	28	50.91	7.76
Meta data "keywords"	39	70.91	23.66
URL	30	54.55	9.91
Image tooltip	33	60.00	7.18
Graphical text	23	41.82	1.74

filtering system should still be capable of analyzing multi-frame Web pages. Since most of the multi-frame Web pages use the two-frame layout format, the system should target the support of processing of two-frame Web pages.

Although PICS could be incorporated into the Web filtering system as a supplementary approach for determining the content nature of a Web page, currently the percentage of PICS support among the Web sites is still not large enough to justify it.

Indicative terms are found in abundance among pornographic Web sites, providing good indication on the content nature of the Web sites. We have identified a total of 55 frequently occurring indicative terms among the pornographic Web pages, which include 42 sexually explicit terms, and 13 legal terms. The legal terms usually appear in the displayed textual contents portion of a Web page, which consists of the title of the Web page, warning message block and other viewable textual contents, with a high density of occurrence in the warning message block. The sexually explicit terms are found in not just the displayed textual contents portion but also the non-displayed textual contents portion of a Web page, i.e. meta data "description" and "keywords", Web page's URL and embedded URL, as well as image tooltip. Thus, two sets of indicative terms should be constructed, with one set of sexually explicit terms, and the other set of legal terms. These two sets of indicative terms should be used by the Web filtering system in the textual content analysis process to extract indicative terms from the textual contents of the Web page.

Web filtering system implementation

We have used the knowledge gained from the analysis of the characteristics of pornographic Web pages to develop an effective Web filtering system. The system architecture is illustrated in Figure 1. It consists of two major processes:
(1) the offline training process; and
(2) the online classification process.

The training process learns from the sample of both pornographic and non-pornographic Web pages to form a knowledge base of the NN models. The classification process then classifies incoming Web pages from the Web according to the nature of the contents.

The training process consists of the following steps:
(1) feature extraction;
(2) pre-processing;
(3) transformation; and
(4) NN model generation and category assignment.

The classification process also performs feature extraction, pre-processing and transformation. In addition, the categorization step is needed to classify the incoming Web pages based on the results given by the NN models. The meta content checking step performs post-processing to enhance the classification results. These steps are briefly described in the following sections.

Feature extraction

A Web page is parsed and the contents in various locations such as the title of the Web page, warning message block, meta data contents of "description" and "keywords", and

Figure 1 System architecture

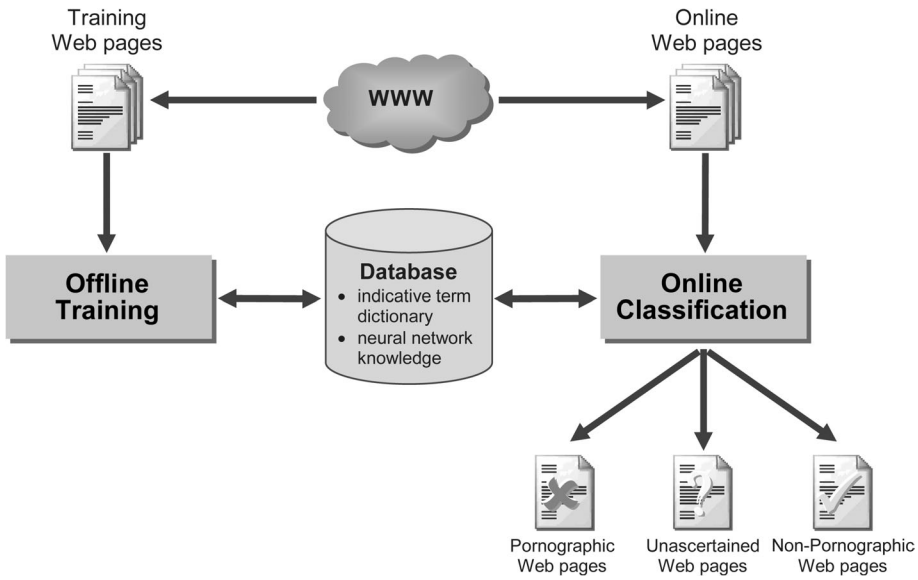
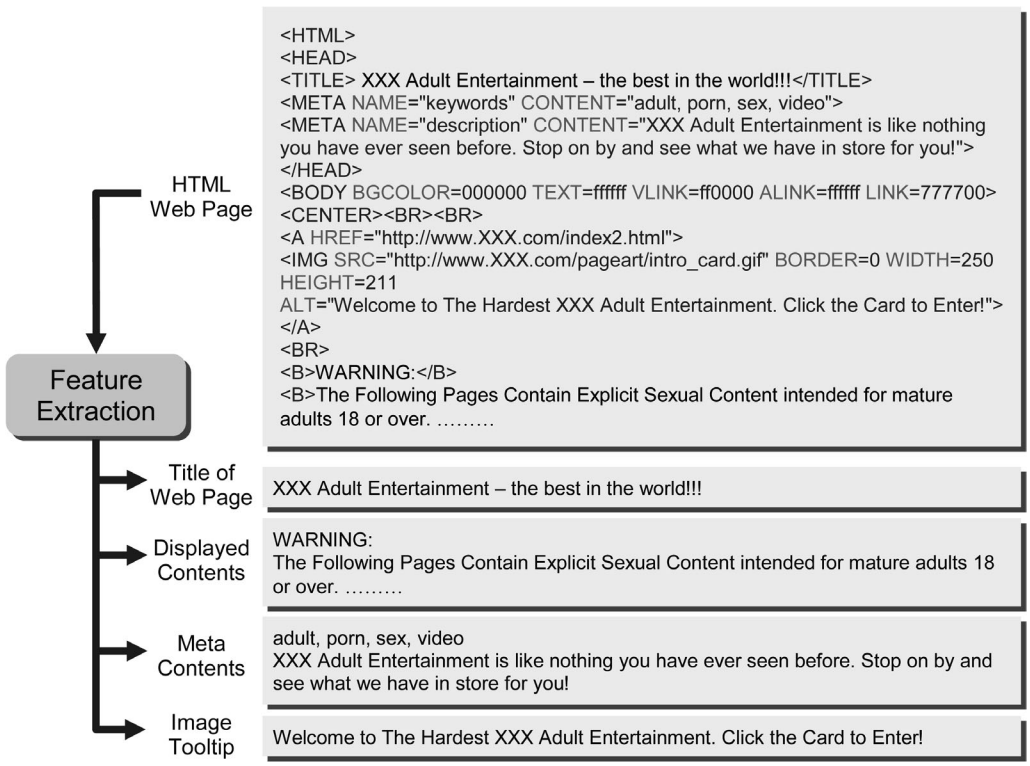


image tooltips, are extracted as the features to represent the Web page. However, we have decided to exclude URLs due to the difficulties of identifying indicative terms in a URL address. Figure 2 illustrates the feature extraction step.

Pre-processing

This step converts all the raw textual contents extracted from the feature extraction step into numeric data representing the frequencies of occurrence of indicative terms. It consists of the tokenization of words, and indicative term

Figure 2 Feature extraction illustration (with most explicit words and company name discarded)



identification and counting using an indicative term dictionary as shown in Figure 3. Tokenization produces four word lists that correspond to the Web page title, displayed contents, meta contents of "description" and "keywords", and image tooltip. As each list represents a different degree of relatedness to the nature of the Web pages, they will carry different weights when training the NN. As we use frequencies of occurrence of indicative terms in a Web page to judge its relevance to pornography, an indicative term dictionary is employed to support the identification of such terms. The dictionary is compiled according to the results of the statistical analysis. There are two types of indicative terms in the dictionary: sexually explicit terms and legal terms, which together give 55 sets of indicative terms. Finally, the indicative term identification and counting step uses the indicative term dictionary to identify the indicative terms in the four word lists from tokenization, and collects the occurrences for each set of indicative terms in the dictionary.

Transformation

The frequencies of occurrence of the respective indicative terms resulting from the pre-processing step are then sorted and converted into vectors representing the Web pages which are fed as the inputs to a neural network.

Neural network (NN) model generation

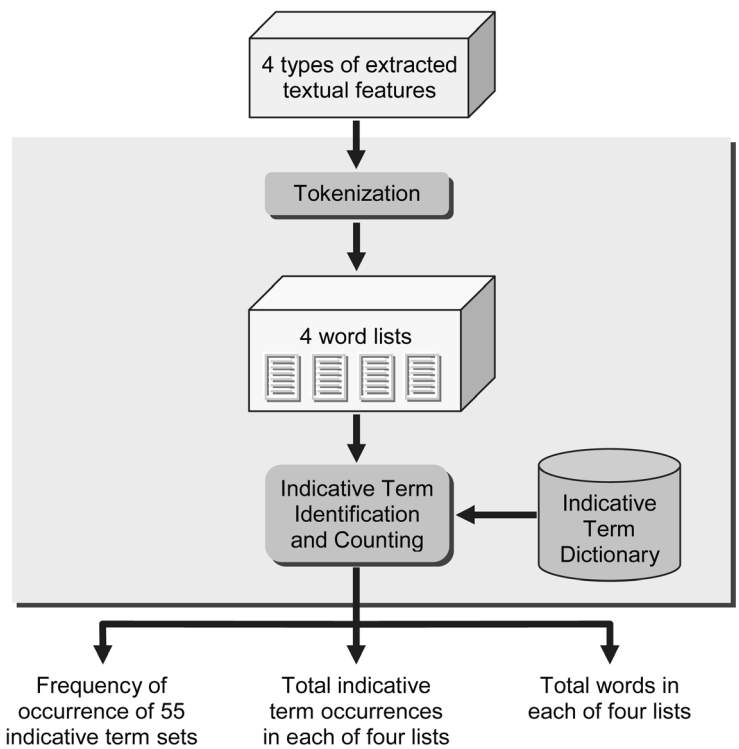
Since the engine makes use of NN models (KSOM and fuzzy ART) for classification purpose, the networks need to be trained before being used for classification. We have collected a total of 1,009 pornographic and 3,777 non-pornographic Web pages to be used as the training exemplar set for offline NN training. The non-pornographic Web pages are collected from 13 categories of the Yahoo! search engine to cover a wide range of topics. Once the training is complete, the NN generated knowledge is stored in a database together with the dictionary of indicative terms.

Category assignment

The clusters generated from the NN model generation are assigned to one of three categories, based on a pre-defined assignment strategy:

(1) pornographic;
(2) non-pornographic; and
(3) unascertained.

Figure 3 Pre-processing



In particular, if a cluster contains at least 80 per cent of Web pages labeled as "pornographic", then the cluster is considered "pornographic". On the other hand, if a cluster contains at least 80 per cent of Web pages labeled as "non-pornographic" then it is considered "non-pornographic". The remaining few clusters are considered unascertained.

Categorization
In this step, the incoming Web pages are classified using the trained NN into one of the three pre-defined categories: pornographic, non-pornographic and unascertained.

Meta content checking
This step checks each Web page classified as unascertained using the contents of the meta data of "description" and "keywords" to determine its nature. The purpose is to further reduce the number of unascertained Web pages. The keywords used are the indicative terms contained in the indicative term dictionary. By analyzing and searching for indicative terms within the meta contents, it determines whether a Web page belongs to the pornographic category. If at least one indicative term is found within the meta contents, the associated Web page is classified as pornographic. On the other hand, if no indicative terms are found inside the meta contents, the Web page is identified as a non-pornographic Web page. If the meta contents cannot be found or do not exist in the Web page, the Web page will remain as unascertained.

Performance evaluation

We conducted a number of experiments to evaluate the effectiveness of our approach. All results obtained were based on the experiments conducted on an Intel Pentium III 866MHz computer running Microsoft Windows 2000 operating system.

Pre-NN processing
First, we measured the performance of the pre-NN steps including feature extraction, pre-processing and transformation, which are used for both the training and classification processes. These three steps are responsible for converting a Web page into the corresponding Web page vector. To evaluate their efficiency, we measure the total processing time for the entire training set of 4,786 Web pages. Table II shows the statistics on efficiency measured for the pre-NN processing.

From Table II, the pre-NN processing steps convert 29 Web pages or 547Kbytes of data in one second on average. This result shows that a Web page can go through the pre-NN processing in less than 35 milliseconds, which is considered very fast. It is important to understand that an efficient pre-NN processing not only helps to improve training efficiency, but also reduces the processing latency in classification.

Online classification
From the above discussion, we observe that each Web page requires an average of 35 milliseconds before reaching the NN and subsequent stages. Using a testing exemplar set comprising 535 pornographic and 523 neutral Web pages, we tested the classification accuracy and efficiency using the two NN models, namely KSOM and fuzzy ART. Tables III and IV show the classification accuracy of the KSOM and fuzzy ART models respectively.

Tables III and IV show that highly accurate classification is possible by combining our knowledge gained in analyzing the characteristics of target Web pages (pornographic in this study) with the learning capability of NNs. KSOM, in particular, provides an exemplary classification accuracy of 95 per cent, which is much better than the ten popular commercial Web filters that we have surveyed (Lee *et al.*, 2002). In addition, the total online processing time is less than 40 milliseconds per Web page on average. This translates to near instantaneous response for highly accurate Web content filtering from the user's viewpoint.

Conclusion

With the proliferation of objectionable materials (e.g. pornography, violence, drugs, etc.) available on the WWW, there is an urgent

Table II Pre-NN processing efficiency

Measure	Number
Number of Web pages	4,786
Total size of Web pages (bytes)	93,578,232
Total processing time (seconds)	167

Table III Classification accuracy of KSOM

Web page	Correctly classified		Incorrectly classified		Unascertained		Total
Pornographic	508		23		4		535
Non-pornographic	497		7		19		523
Total	1,005	(95.0%)	30	(2.8%)	23	(2.2%)	1,058

Table IV Classification accuracy of fuzzy ART

Web page	Correctly classified		Incorrectly classified		Unascertained		Total
Pornographic	460		47		28	535	
Non-pornographic	483		16		24	523	
Total	943	(89.1%)	63	(6.0%)	52	(4.9%)	1,058

need for effective countermeasures to protect children and other unsuspecting users from exposure to such materials. Indeed, a number of commercial Web filtering systems have been developed. However, they tend to lack the accuracy for effective filtering.

In this paper, we have discussed the four major techniques employed in Web content filtering systems. We have concluded that it is necessary to take a new approach based on analyzing the characteristics of the targeted Web pages coupled with the learning capability of the artificial neural networks (NN). Using pornographic Web pages as a case study, we have presented a thorough analysis of the distinguishing features of such Web pages. We have also used this knowledge to develop an effective filtering system that can achieve a high accuracy of 95 per cent with virtually no added processing latency from the user's perspective.

The same approach could be adapted to filtering other objectionable Web materials. The key attribute is a thorough understanding of the distinguishing features of the selected materials to be filtered. In fact, future research is underway to develop effective filtering tools for other types of objectionable materials, as well as multilingual Web pages.

References

Carpenter, G.A., Grossberg, S. and Rosen, D.B. (1991), "Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system", *Neural Networks*, Vol. 4 No. 6, pp. 759-71.

Dalton, J. and Deshmane, A. (1991), "Artificial neural networks", *IEEE Potentials*, Vol. 10 No. 2, pp. 33-6.

Flexer, A. (2001), "On the use of self-organizing maps for clustering and visualization", *Intelligent Data Analysis*, Vol. 5 No. 5, pp. 373-84.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, NY.

Koehler, G.J. and Erenguc, S.S. (1990), "Minimizing misclassifications in linear discriminant analysis", *Decision Sciences*, Vol. 21 No. 1, pp. 63-85.

Kohonen, T. (1995), *Self-Organizing Maps*, Springer-Verlag, Berlin.

Lee, P.Y. (2002), "Intelligent Web content filtering", MEng thesis, School of Computer Engineering, Nanyang Technological University, Singapore.

Lee, P.Y., Hui, S.C. and Fong, A.C.M. (2002), "Neural networks for Web content filtering", *IEEE Intelligent Systems*, Vol. 17 No. 5, pp. 48-57.

Lippmann, R.P. (1987), "An introduction to computing with neural networks", *IEEE ASSP Magazine*, April, pp. 4-22.

McCallum, A. and Nigam, K. (1998), "A comparison of event models for naïve Bayes text classification", AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, AAAI Press, Menlo Park, CA.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

World Wide Web Consortium (W3C) (1997), "Platform for Internet content selection", available at: www.w3.org/PICS/.

Yang, Y. (1994), "Expert network: effective and efficient learning from human decisions in text categorization and retrieval", *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, Vol. 1, pp. 11-21.

Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1/2, pp. 69-90.

Yang, Y. and Chute, C.G. (1992), "A linear least squares fit mapping method for information retrieval from natural language texts", *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, Vol. 2, pp. 447-53.

Yang, Y. and Chute, C.G. (1993), "An application of least squares fit mapping to text information retrieval", *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, pp. 281-90.