13—14

Image Rating System for Filtering Web Pages with Inappropriate Contents

Yoshinori Musha, Atsushi Hiroike, Yasutsugu Morimoto, and Jyunichi Matsuda
Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan.
E-mail: {sha, he, y-morimo, j-matsud}@crl.hitachi.co.jp

Abstract

We have developed a prototype system with image discrimination for the filtering and rating of web pages displaying inappropriate content. We used the SafetyOnline rating standard for the system. The rating standard defines five categories having five levels. The system rates web pages and classifies them into five levels of inappropriateness for each category according to the rating standard. The filtering process is different from the rating process. To enable the filtering of web pages, a user pre-selects a level of inappropriateness in each category in advance and the system filters out web pages whose the level of inappropriateness is higher than the level specified by the user. In the rating process, the system classifies web pages into five levels of inappropriateness for each category. The system performs both image and text classification to filter web page content. In this paper, we focused on image classification. We developed a method that enables the system to rate an image by integrating results of discriminating the image at all levels based on the Bayesian theory. The system also rates and filters a certain web page by integrating the results of images attached to the web page. We examined the process of image discrimination for the filtering of inappropriate content and the process of image classification for the rating of that.

1 Introduction

With the worldwide spread of the Internet, the number of web pages is increasing every year. Anyone can create and run a web page without disclosing his or her identity. Even if users create their web pages with malicious intent, the inappropriate contents of these pages will be available to the general public. Filtering systems, however, should unreasonably not interfere with the right of web page owners to free speech. The WWW Consortium (W3C) proposed a platform for Internet content selection (PICS). In the PICS[1], users can send web pages without restrictions and those who receive those pages can filter them according to their ratings assigned by a third party. However, the rating of web pages is a very labor-intensive task. It is becoming increasingly difficult to rate web pages, the number of which just in the ".jp" domain increases by 15—20 million every year. Therefore, we need a system that would filter web pages without assigned ratings automatically and can assist rating operators in rating web pages. A lot of research is being done on the subject, with many studies focusing on systems that can automatically categorize web pages by using keywords and text information[2].

We have developed a system that categorize and rate web pages by using text and image information[3]. The system can automatically filter the web pages containing inappropriate contents without assigned ratings and can assist rating operators in rating the web pages more effectively. In this paper, we describe our system and, in particular, its image discrimination method for filtering web pages. We show our experimental results obtained for automatic filtering and rating based on image discrimination.

2 Prototype system for filtering web pages

Our prototype system consists of three components, a URL-based filtering system, a content-based filtering system, and a rating estimator for assisting rating-operators. Figure 1 shows the data flow in the system. Users including parents of young children register the desired level of filtering per category in the system before the filtering begins. The categories, such as (n), (s), (v), (l), and (e), and the criterion for choosing an appropriate level of filtering on each category are based on the SafetyOnline rating-standard in Japan, which is the basis for the RSACi rating-standard. In the criterion, a low level means "tight filtering" and acceptable contents to all users including children, and a high level means "loose filtering" and unacceptable contents to children and others. When a user accesses a web page, the URL-based filtering system checks the URL of the web page. If the ratings of the URL are lower than the user-registered levels, the user can access it. Otherwise, the web page is blocked. If the URL is not labeled, the content-based filtering system checks the text and image data of the web page and judges if the web page is acceptable according to the user-registered levels. These web pages are stored into a URL rating database. After that, the rating

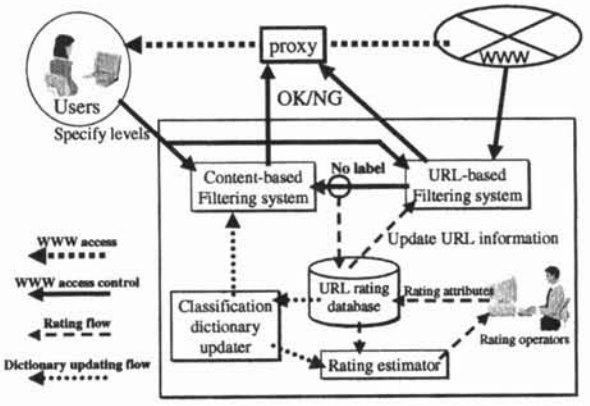

Figure 1: Data flow in our prototype system

estimator rates these pages and shows the results to rating-operators. The operators register the rating attributes in the database. The attributes are used for URL-based filtering as well as to improve content-based filtering and rating estimation. In particular, the method improves the capabilities of filtering and rating similar web pages to ones users have accessed.

3 Filtering system and rating estimator based on image discrimination

Figure 2 shows the structures of the content-based filtering system and the rating estimator based on image discrimination. The two systems use eight level-boundary discriminators (image discriminators): four for the (n)-category and four for the (s)-category. Each discriminator has a feature vector of 200 dimensions per image and provides image discrimination by comparing the feature vector of an input image with labeled feature vectors in the dictionary. It's described in detail in the next session. First, in both systems, input images attached to an HTML file are filtered so that only images larger than 64x64 pixels could be used. Next, the user-selected level-boundary discriminator in the filtering system then judges whether the images are acceptable. Finally, the results for all the filtered images in the HTML file are integrated so that the filtering system outputs a judgment that the HTML file is appropriate if all the results are appropriate and the file is inappropriate if otherwise. Therefore, the rate of precisions for filtering HTML files should be higher than for filtering image files and the rate of recalls of inappropriate contents for HTML files should also be higher than for image files. In contrast, in the rating estimator, after images are filtered according to the size, all discriminators in a category judge the appropriateness of each filtered image. Then reliabilities of five level-candidates for each image are calculated by using the Bayesian theory and the level of the image is determined by adopting the level with highest reliability among the level-candidates. Finally, the levels for all the

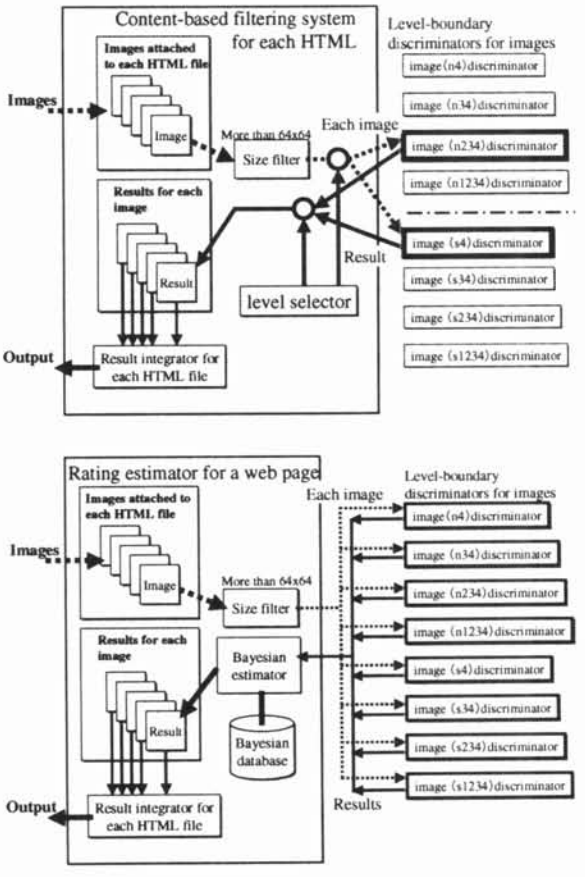

Figure 2: Image discrimination: (1) the upper figure shows a content-based filtering system that has the level selector, with which the user can control the level of categories. (2) the lower figure shows a rating estimator that assists rating operators in rating web pages easily.

filtered images in the HTML file are integrated so that the rating estimator outputs the highest level in all the images as the level of the HTML file.

4 Level-boundary discriminator for images

The system has an image discriminator that determines a class showing whether a certain image is appropriate depending on its level-boundary. When the discriminator [in Fig. 2] receives an image, it extracts a vector having 200 dimensional features from the image and scans feature vectors stored in the classification dictionary that is made from sample images with the classes in advance. Then the discriminator finds the nearest feature vector to the extracted vector from the stored feature vectors and outputs the class attached to the nearest feature vector.

The feature includes color features and differential orientation features. In the color feature, we prepared 4 x 4 x 4 boxels by dividing an RGB color space beforehand. The system sorts all the pixels in an image into the 64 boxels, according to

the quantized color vectors for the pixels, makes a histogram using the 64 boxels counts, and finally normalizes the histogram. In the differential orientation feature, the system transforms color images to gray scale images, calculates the orientation and power of the changes of brightness, adds the powers of all the pixels in the image to one of 8 bins according to the quantized orientation of each pixel, makes a histogram using the 8 bins counts, and finally normalizes the histogram. The system applies the above feature extractions to each 4 x 4 area and each of four resolutions. As a result, the number of color feature elements is 1,024 and the number of differential orientation feature elements is 512. We compress the feature dimensions by applying principal component analysis (PCA) to the features and get a 100-dimension feature vector in each feature[3][4].

The classification dictionary updater [in Fig.1] creates a dictionary as follows. (1) The updater extracts feature vectors from sample images with known classes and stores the center of gravity of the feature vectors of the images in each class as templates in the dictionary. (2) It determines the class of all the sample images, just like the image discriminator does. (3) It groups together all error sample images, for which the use of each same template in the dictionary result in errors, and stores in the dictionary a new center of gravity of the feature vectors of each group of images. (4) It removes the error sample images from the sample images compounding the old center of gravity and then creates a new center of gravity of feature vectors of the remained sample images. As a result, the number of templates in the dictionary increases. The updater repeats all the steps from step (2) until the error rate is less than a specified threshold or until the loop counter reaches a specified count. Using this method, the updater does not always completely eliminate errors in the sample images, but the number of errors does decrease very quickly.

5 Rating estimator for image rating

By applying the Bayesian theory to the output of image discriminators, the rating estimator classifies given images into five levels of inappropriateness for each category. This method has the following three advantages.

(1) Improving the filtering system also improves the rating estimator because both systems use the same image discriminators. It becomes easy to maintain the systems.

(2) The rating estimator has a simple structure and can quickly be carried out. It assigns the level C_1 to an image that allows conditional probability $P(l|D_{n4}, D_{n34}, D_{n234}, D_{n1234})$ to be the maximum by using Eq.1.

$$C_1 = \left\{ l \middle| \max_{l \in L} P(l|D_{n4}, D_{n34}, D_{n234}, D_{n1234}) \right\} \quad \text{(Eq.1)}$$

where $D_i = \{0,1\}$ $(i = n4, n34, n234, n1234)$ represents four outputs of four image discriminators, i is each image discriminator, classes 0 and 1 of the D_i are, respectively, acceptable and unacceptable content, and $l \in L = \{n0, n1, \cdots, n4\}$ represents the level of inappropriateness of the image. The conditional probability is calculated using Eq. 2, and it is stored in the Bayesian database [in Fig. 2].

$$P(l|D_{n4}, D_{n34}, D_{n234}, D_{n1234}) = \frac{P(l)P(D_{n4}, D_{n34}, D_{n234}, D_{n1234}|l)}{\sum_{k \in L} P(k)P(D_{n4}, D_{n34}, D_{n234}, D_{n1234}|k)} \quad \text{(Eq.2)}$$

Here, $P(l)$ can be obtained by counting the sample images in each level. The $P(D_{n4}, D_{n34}, D_{n234}, D_{n1234}|k)$ can be obtained by inputting sample images of level k into the four image-discriminators and counting the images in each combination of the four outputs.

We also used the level C_2 in Eq.3 as the output of the rating estimator with a reject function.

$$C_2 = \left\{ l \middle| \max_{l \in L} f_l(P(l|D_{n4}, D_{n34}, D_{n234}, D_{n1234})) \right\} \quad \text{(Eq.3)}$$

$$f_l(x) = \begin{cases} x \cdots\cdots\cdots\cdots (x \geq T_l) \\ 0 \cdots\cdots (otherwize) \end{cases} \quad \text{(Eq.4)}$$

The $f_l(x)$ in Eq.4 is a threshold function that has a threshold value, T_l, in each level. The C_2 can cut off an output with low reliability and can control the output frequency in each level because the levels can have different thresholds.

(3) The rating estimator has a mechanism, in which the output is influenced by the ranking lying in all the levels of image inappropriateness. The ranking shows, for instance, the difference between $n0$ and $n1$ is much smaller than the difference between $n0$ and $n4$. If a rating estimator is based on the other method classifying images into five levels, the outputs may be independent of the ranking because the rating estimator may not discriminate the difference between $n0$ and $n4$ from the difference between $n0$ and $n1$. In contrast, our rating estimator uses results of an image discriminator that classify images into levels higher or lower than the level specified for each image discriminator, so that the rating estimator includes the relationship between $n0$, $n1$, and $n4$. The rating estimator therefore reduces the possibility of fatal mistakes of rating images.

6 Evaluating the filtering accuracy and the rating estimation accuracy

We evaluated the filtering accuracy (1) and the rating estimation accuracy (2) based on image discrimination. In this paper, we show (1) the effects of integrating results of discriminating images attached to a HTML file [in Fig. 2] and (2) the application of the reject function in Eq. 3.

(1) We prepared 6,383 images, which were

attached to HTML files and had rating attributes of five levels of an (n)-category. We divided the images into two sets of images and used one set to train the discriminators, and the other one to evaluate the effectiveness of the discriminators. Figure 3 shows the recall rates of image discrimination for the (n)-category. The recall rates of all four image discriminators ranged from 43 to 71% according to the order of the discriminators: n4, n34, ..., n1234. The recall rates of the result integrator [in Fig.2] for HTML files ranged 79 to 88%. The recall rates improved dramatically as a result of integrating the results for images attached to the same HTML file. Incidentally, the precision rates also ranged from 40 to 68% (There is no figure in this paper).

(2) We divided the images into three sets of images and used the first set to train the discriminators, the second one to make a database based on the Bayesian theory, and the third one to evaluate the effectiveness of the discriminators. We performed six experiments with different combinations of the three sets. In the experiments, we used the reject function in Eq.3 in the rating estimator. The rating estimator blocked level candidates with relatively low reliability in each level, which resulted in acceptance rates of less than 16%. Figure 4 shows the results of rating estimation for the HTML files. The precision rates without the reject function for levels n0, n1, n2, n3, and n4 were 80, 0, 19, 13, and 49%, respectively. The precision rates with the reject function for these levels were 92, 10, 19, 10, and 36% and the reject rate was 28%. Therefore, we found that the use of the reject function increased the precision rates for the more acceptable levels, n0 and n1 and reduced the precision rates for the less acceptable levels, n3, and n4. In particular, for level n1, the acceptance rate increased from 0 to 12%, so the precision rate increased as a result of the use of the reject function from 0 to 10%. Incidentally, we used the threshold values 0.811, 0.095, 0.22, 0.19, and 0.093 for levels n0, n1, n2, n3, and n4. The highest rate was obtained for level n0. The highest ratio of the precision rate to the random rate $P(l)$ was obtained for level n4.

7 Conclusion

We have developed a prototype system for filtering web pages. The system filters web page content and assists rating operators in rating web pages by using a rating estimator, and it can improve both the filtering accuracy and the rating estimation accuracy because it uses the same level-boundary discriminators. In particular, the system can improve the accuracy of filtering and rating web pages similar to those that users have already accessed. We evaluated the filtering effectiveness and the rating estimation effectiveness based on image discrimination only. The system improved the filtering precision for each HTML file by integrating the results of the images attached to each HTML file. By using the reject function that allows for different thresholds in every level, the precision rates for the levels were able to be controlled in a measure. We need to further improve the filtering and estimation functions of our system.

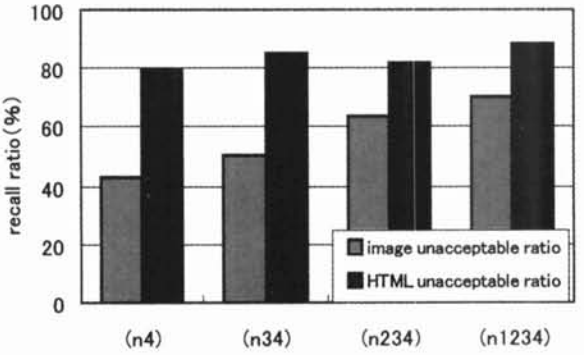

Figure 3: Recall rates for four image discriminators, n4, n34, n234, and n1234.

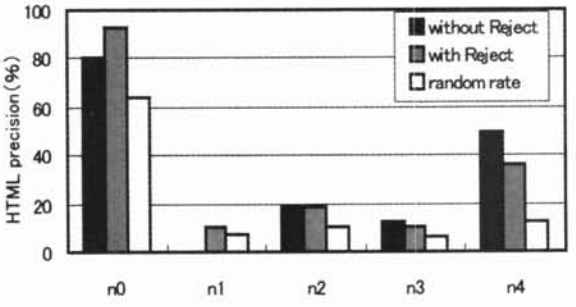

Figure 4: The precision rates of the rating estimator: The rating estimator for images has a threshold of reliability for the reject function in each level and then we determined the accept rates to be less than 16%.

This research, which this paper is based on, was performed through a contract with the Telecommunications Advancement Organization of Japan (TAO) authorized by the Ministry of Public Management, Home Affairs, Posts and Telecommunications.

References

[1] Internet Content Rating Association: http://www.icra.org/

[2] N. Inoue, K. Hoashi and K. Hashimoto: Development of filtering software of hazardous WWW information by using document classification method (in Japanese), *IEICE, Vol.J84-D2, No.6*, pp.1158--1166, 2001.

[3] Y. Musha and A. Hiroike: Image discriminant method for assisting with rating of potential harmful information on WWW(in Japanese), *General conference of IEICE, D-12-33*, p. 209, March, 2002.

[4] Y. Musha and A. Hiroike: Visualization system displaying retrieved images in a 2-D semantic space, *Proc. of MVA2000, 3-19*, pp.103—106, 2000.