Pornography Detection with the Wisdom of Crowds

Cheng Luo*, Yiqun Liu, Shaoping Ma, Min Zhang,
Liyun Ru, and Kuo Zhang

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China
{c-luo12@mails.thu.edu.cn,{yiqunliu,z-m,msp}@mail.thu.edu.cn
lyru@vip.sohu.com,zhangkuo@sogou-inc.com
http://www.thuir.cn

Abstract. With rapid development of the Internet, much attention has been paid to the problem of children exposed to Internet pornography. Existing detection techniques, which mainly focus on pornography content analysis have gained much success. However, they still meet challenges in practical Web environment due to the great computational costs and the difficulties in dealing with various pornography forms. We attempt to solve this problem from a new perspective with the wisdom of crowds in search engine click-through logs. Inspired by the idea that different pornography Web pages may be oriented by similar search keywords, a label propagation method on click-through bipartite graph is proposed which can locate pornography Web pages from a small set (a few hundreds) of manually labeled seed pages. Experiments performed on datasets collected from both English and Chinese search engines show that the proposed algorithm can identify different forms of Internet pornography both effectively and efficiently.

Keywords: Pornography Detection, Click-through Graph, Semi-supervised Learning

1 Introduction

The Internet is increasingly prominent in young people's lives. A global Internet usage survey in the year of 2008 revealed that, among youths between 12 and 14 years old in the United States, 88% used the Internet; the percentage of Internet users in this age group was 100% in the United Kingdom, 98% in Israel, 95% in Canada, and over 70% in Singapore[1].

Coupled with the very large number of pornographic Web pages, concern has been raised that increasing accessibility could lead to a rise in pornography seeking among children and adolescents, with potentially serious ramifications for their sexual development. Ropelato's statistics shows that there are 420 million pornographic Web pages on the Internet, and 42.7% of Internet users saw pornographic content in 2006[2].

* This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205) of China.

In many countries, sexual materials on the Internet are subject to censorship and legal restraints on their publication on the grounds of that they are obscene and that adolescents must be protected from inappropriate information [3][4]. However, the Internet has no boundaries, which means that pornography from a place where there are no effective restrictions imposed on Internet pornography can be easily accessed.

There have been several proposals to protect children from pornographic information on the Web. Traditional filtering techniques regard it as a classification problem which relies on features extracted from contents. Different types of page content are taken into account, such as texts, images and videos. Many approaches, including neural networks [5], statistical natural language processing [6] and pattern recognition [7], have been used to train a classifier to identify Web pages with pornographic content. Most of these content-based methods are usually dependent on the form of pornographic material and are limited by the efficiency of the algorithms. It is sometimes difficult to adopt these methods on practical Web environment due to the costs of a large amount of computational resources, especially for those with multimedia contents.

With explosive growth of Web resources, search engine become one of the most important portals for all kinds of Web pages including the pornographic ones. The click behaviors on pornographic pages usually reflect the 'search for porn' intent of the users. Because of this similar search intent, different users often use similar queries to search for pornography on the Web. Therefore, the aggregation of a large number of user clicks is likely to provide valuable implicit evidence of whether one page being pornography or not. We can utilize the correlations between queries and URLs to detect pornographic Web pages. In other words, this approach can be regarded as a type of wisdom of crowds.

Inspired by this idea, we try to utilize the correlations between queries and URLs to detect pornographic Web pages. We construct a bipartite graph from click-through data with queries/URLs as nodes and user clicks as edges. After that, a propagation based algorithm is performed on the graph to estimate the possibility of a Web page containing pornography.

The major contribution of this work is that we propose a highly efficient method to identify pornography based on a click-through bipartite graph. In this way, there is no need to crawl Web pages and perform time-consuming content analysis on them. To the best of our knowledge, we are among the first to address the problem of pornography detection using only click-through data.

The remainder of this paper is organized as follows: Section 2 provides a brief review of the related literature. Section 3 presents our motivations for detecting pornography, discusses our algorithm in detail and gives a proof of its convergence. In Section 4, an experimental validation is conducted, and the analysis of the results demonstrates that our method can detect pornography both effectively and efficiently. Our conclusions are given in Section 5.

2 Related Work

To fight against pornographic content on the Web, there are a number of major content-filtering approaches that are adopted, including the Platform for Internet Content Selection (PICS), URL blocking, keyword filtering, and intelligent content analysis.

PICS is a voluntary labeling system that allows Web publishers to associate labels or metadata with Web pages[8]. RSACi and SafeSurf are the two most popular PICS systems. Currently, Microsoft Internet Explorer and several other popular Web browsers offer PICS support with embedded PICS rating labels. However, PICS is not adopted by all major content providers and is not very reliable because of the mislabeling problem, either by negligence or by intent.

The second common approach focuses on URL blocking systems that restrict or allow access by comparing the requested Web page's URL with URLs in a stored list. Usually, two types of lists are maintained, namely a black-list and a white-list[9]. Lee et al. proposed an inverse chi-square based classification method and an incremental updating mechanism[10]. It's not necessary for URL blocking technology to consume a large amount of computational resources to perform content analysis. It also avoids the risk of virus infections. However, it is quite difficult to maintain a black-list for the practical Web environment because of the rapid growth of the Web pages.

Keyword filtering[11] blocks access to Web pages when the occurrence of harmful words or phrases exceeds a predefined threshold by comparing the text in the retrieved Web page to a predefined dictionary of prohibited words and phrases. Gui-yang et al. proposed a keyword-matching method to filter harmful text[12]. One of the greatest challenges of keyword blocking is over-blocking. Nevertheless, it can be adopted to decide whether further content analysis is needed, which might require more time.

Intelligent content analysis attempts to gain a semantic understanding of the context on a Web page. Existing methods usually train a model using statistical computing of the discriminative features extracted from texts or images to make a decision. Lee et al. used the frequency that keywords appear in a text and the relevant Web page feature as the input to a Kohonen self-organizing neural network (KSOM) to train a classifier[5]. Ho and Watters used a Bayes classifier and considered the difference of a word's weight and the positions in which the word appeared[6]. Polpinij et al. proposed a filtering system that combines both text and images[7].

To summarize, while adopted to practical Web environment, PICS and URL blocking meet the problem of keeping prior information both credible and up-to-date. Keyword filtering can be regarded as a kind of content analysis method because they both rely on content features and suffer from the problem of obtaining and dealing with contents from Web pages that are usually noisy, ill-formed and incredible. For those methods which adopt multimedia content features, they are further constrained by limited computational resources and high-efficiency requirement. Different from these detection methods, we utilize user behavior information stored in search engine click-through logs and therefore avoid the problem of (multimedia) content analysis of numerous Web pages. Because search engines have oriented a large part of user visits for most Web sites including pornography ones, we believe that this method can deal with most pornography problems on the Web although it doesn't require crawling these pages.

3 Pornography detection with click-through data

3.1 Motivation

Traditional pornography detection systems usually focus on various content analysis techniques. The major limitation of the approaches mentioned above might be the computational cost. The behavior of search engine users offers some information that is helpful for pornography detection. Pornographic Web pages usually select some attractive keywords that reflect the 'search for porn' intent to boost the ranking of their pages/sites in corresponding search results lists. In the other way, the users who want to access pornography by search engine most likely issue similar queries. Therefore, our basic assumption is that users share similar queries to search for pornography, which are most likely to be popular on Web pages that contain pornographic materials. The collection of queries that are related to a certain URL can be considered to be a profile of a Web page. If a large percentage of queries contain implicit pornographic intent, then the reason is most likely that there was pornographic content on the Web page.

By noticing the relationship between pornographic Web pages and porn-intent queries, we designed a label propagation algorithm on click-through data. First, a small number of seed pages are selected and each labeled with a pornographic score. Then, their labels are propagated on the click-through bipartite graph to identify other possible instances of pornography. The input comprises a set of labeled URLs, a set of unlabeled URLs and a set of constraints between URLs and the queries in the log. The goal is to find unlabeled pornographic pages from labeled pages.

3.2 Problem Formulation

Before formulating our problem, some definitions should be given.

I. Click-through data C and bipartite graph G.

The click log is a set of triples $\langle q, u, f_{qu} \rangle$, where q is a query, u is a URL, and f_{qu} is the times URL u is clicked when query q is issued. Define $Q = \{q|q \ appears \ in \ C\}$,and $U = \{u|u \ appears \ in \ C\}$. Click-through data C can be presented as another equivalent form – a click-through bipartite graph $G = (Q, U, E)$. There are two types of nodes in the graph, queries and URLs. For a certain edge(q, u), each q/u is assigned a score p_q/p_u, which denotes how likely this query/URL is to be a pornographic query/page. A sample portion of a bipartite graph constructed with search engine log, as shown in Figure 1.

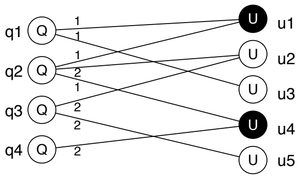

Fig. 1. An example to of query-URL bipartite graph

The click-through graph can be constructed either at the page level or at the site level. In this study, we choose the URL itself to construct graph because the structure and content of a Web site might be very comprehensive.

II. Labeled Seed URL set L.

All of the URLs in L are selected from C(or G) and are manually labeled as porn. Formally, $L = \{u|u \ is \ labeled \ as \ a \ pornographic \ page\}$. We will discuss the construction details of L in Section 4.1.

III. URL result set RU and query result set QU.

Respectively, RU and QU contain all of the $\langle u, p_u \rangle$ and $\langle q, p_q \rangle$ pairs. After the algorithms ends, each URL u or query q in C(or G) will receive a score p_u/p_u, which denotes the possibility that this URL or query is a porn one.

Given G = (Q,U,E) and $L \subset U$, the goal of this problem is to obtain the results set RU, which contains all of the possible pornographic pages in G.

3.3 Algorithm design

First we will propose a label propagation algorithm for the the detection of pornography. Specifically, for every URL u, we could calculate the probability p_u of a certain URL u by incorporating all of the label information of its adjacent query nodes. Similarly we could calculate p_q for every query q. This procedure can be described formally as follows.

For $\forall$q/u, l_q/l_u denotes its label, which is **P** for pornography and **N** for non-pornography. Thus, every URL u in L would receive a label, such as **P** or **N**, which means that P(l_u=**P**)=1 or P(l_u=**P**)=0 initially while every URL u in the set $U - L$ has P(l_u=**P**)=0. Then, we have

$$P(l_u = \mathbf{P}) = \sum_{q:(q,u) \in E} (t_{qu} P(l_q = \mathbf{P})) \quad (1)$$

where

$$t_{qu} = \frac{\omega_{qu}}{\sum_{q':(q',u) \in E} \omega_{q'u}} \quad (2)$$

and

$$\omega_{qu} = f_{qu} \quad (3)$$

t_{qu} can be interpreted as the transition probability from query q to URL u and $\omega_{qu}$is the weight of edge (q, u) in the bipartite graph. From equations (1) and (2) that q's label is determined by both its neighbors' labels and the relationship of the connection. The larger the value of ω_{qu} is, the more influence its corresponding node has on the label determining the label of q.

Similarly, for each query q in Q, the probability P(l_q=**P**) is computed as

$$P(l_q = \mathbf{P}) = \sum_{u:(q,u) \in E} (t_{uq} P(l_u = \mathbf{P})) \quad (4)$$

where

$$t_{uq} = \frac{\omega_{qu}}{\sum_{u':(q,u') \in E} \omega_{qu'}} \quad (5)$$

t_{uq} can be interpreted as the transition probability from URL u to query q.

Using the equations above, we can obtain P(l_q=**P**) and P(l_u=**P**) recursively for all of the queries and URLs in the click-through bipartite graph. A concise representation of this iterative process is stated as follows.

Suppose that there are $|Q|$ queries and $|U|$ URLs. Define possibility vectors as follows:

$$\mathbf{P_Q} = (P(l_{q1} = \mathbf{P}), P(l_{q2} = \mathbf{P})...P(l_{q|Q|} = \mathbf{P}))^T \tag{6}$$

$$\mathbf{P_U} = (P(l_{u1} = \mathbf{P}), P(l_{u2} = \mathbf{P})...P(l_{u|U|} = \mathbf{P}))^T \tag{7}$$

and the transition probability matrixes as:

$$\mathbf{T_{qu}} = (t_{qu})_{|Q|\times|U|} \quad and \quad \mathbf{T_{uq}} = (t_{uq})_{|U|\times|Q|} \tag{8}$$

Then, in the i^{th} iteration,

$$\mathbf{P_Q^i} = \mathbf{T_{qu}P_U^{i-1}} \quad and \quad \mathbf{P_U^i} = \mathbf{T_{uq}P_Q^{i-1}} \tag{9}$$

It should be noted that the possibility of the labeled nodes should be clamped before each round of iteration, which means that all of the URLs in the seed set L should be re-assigned their initial label. In this way, the algorithm converges. The convergence will be proved in Section 3.5.

3.4 Bidirectional Edge Weight Definition

In the naive definition of the edge's weight shown in Equation(3), the weight is related only to the amount of clicks in click-through log. However, we find that in experiment the naive definition has to face the positive feedback problem and the reliability problem. More specifically, the problems that naive definition faces are stated as follows:

The Positive Feedback Problem Label propagation algorithms or random walks on click-through bipartite graphs usually face positive feedback problems. For example, u_3 in Figure 1 is connected only to q_1. If we use the naive weight definition for iteration, after the first iteration, P(l_{u3}=**P**)=0.5. Before the second iteration begins, we will set P(l_{u1}=**P**) to be 1 because u_1 is a seed URL. It is easy to see that P(l_{u3}=**P**) is 0.75 after the second iteration and converges to 1 as the iteration process proceeds. The reason is that u_3 is a 1-degree node, which means that the score of u_3 will flow back to q_1; from this process, it obtains its original pornography score. We call this effect the positive feedback problem, which would magnify the noise in the click-through bipartite graph and distort the final results.

Suppose that a non-typical user issues a query q to the search engine and then clicks a pornographic page while most of the other users use this query to navigate to ordinary sites. After applying this label propagation algorithm, all of the pornography scores of the URLs will converge to 1 after our algorithm is applied on the graph. Although the 1-degree nodes will be removed from results, it will most likely misinterpret the real explanation of other pages.

The Reliability Problem Another important challenge that we must face is the reliability problem. In a naive label propagation algorithm, the score of one specific node comes from its adjacent nodes, and the weight of each score is related only to the weight of the edge, which is click times in the query log. However, the correlation between a query node and an URL node should be determined by the click distribution of both of these nodes jointly. The naive definition fails to take this factor into account. Consider another sample portion of a bipartite graph, as shown in Figure 2.

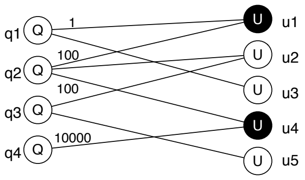

Fig. 2. A case for the reliability problem

Based on the naive weight definition, for query node q_2, the score of u_1 and u_4 are equally weighted. Let us focus on u_1 and u_4; u_1's clicks are almost all from q_2 while u_4 has the same clicks from q_2, but most of its clicks are from q_4. Obviously, the score from u_1 is more convincing for q_2 than for u_4.

Based on these two observations, we take the click distribution of both the query and the URL into account and propose a novel bidirectional edge weight definition. For edge $(q, u) \in E$, the weight is defined as:

$$\omega_{qu} = \frac{f_{qu}}{\left(\sum_{q':(q',u)\in E} f_{q'u}\right)\left(\sum_{u':(q,u')\in E} f_{qu'}\right)} \tag{10}$$

Essentially, our bidirectional edge weight definition will help the iterative process to magnify the influence from nodes with a close relationship and to minimize the effect of noisy nodes(e.g., a query seldom issued or URLs with few clicks). To summarize, the outline of our algorithm is as follows:

Algorithm 1 Pornography detection algorithm

```
Input: labeled seed set $L$,click-through data $C(G)$
Output: P($l_u$=P) for all URLs in $G$
repeat
    for $u \in L$, set P($l_u$=P) = 1 due to they are in seed set
    for all $q \in Q$, calculate P($l_q$=P) as $\mathbf{P_Q} = \mathbf{T_{qu}P_U}$
    for all $u \in Q$, calculate P($l_u$=P) as $\mathbf{P_U} = \mathbf{T_{uq}P_Q}$
until Algorithm converges
Output P($l_u$=P) for every URL $u$ in $U$
```

3.5 Convergence of the algorithm

Let us look into $\mathbf{M_{qu}}$ and $\mathbf{M_{uq}}$, each of whose rows is composed of nonnegative real numbers, with each row summing to 1. They are right stochastic matrixes. Consider $\mathbf{M_{uu}} = \mathbf{M_{uq}}\mathbf{M_{qu}}$.

For each element t_{ij} in $\mathbf{T_{uu}}$,$\omega_{ij} = \sum_j \omega_{ik}\omega^{'}_{kj}$, where $\omega_{ik} \in \mathbf{T_{uq}}$ and $\omega^{'}_{kj} \in \mathbf{T_{qu}}$. Thus, we have

$$\sum_j t_{ij} = \sum_j \sum_k \omega_{ik}\omega^{'}_{kj} = \sum_k \sum_j \omega_{ik}\omega^{'}_{kj} = \sum_k \omega_{ik} \sum_j \omega^{'}_{kj} = \sum_k \omega_{ik} = 1 \quad (11)$$

$\mathbf{T_{uu}}$ is also a right stochastic matrix. Next, look into $\mathbf{P_U}$; the iteration process can be represented as,

$$\mathbf{P^i_U} = \mathbf{T_{uu}}\mathbf{P^{i-1}_U} = \mathbf{T_{uq}}\mathbf{T_{qu}}\mathbf{P^{i-1}_U} \quad (12)$$

Let T_l be the top l rows of T(the labeled pages), and let T_u be the remaining u rows. Note that T_l never really changes because it is re-assigned in every iteration.Define the probability vector $\mathbf{P_U} = \begin{pmatrix}\mathbf{P_L} \; \mathbf{P_R}\end{pmatrix}$, where $\mathbf{P_L}$ are the top l rows of $\mathbf{P_U}$(the labeled pages) while $\mathbf{P_R}$ are the remaining rows. We can split $\mathbf{T_{uu}}$ into 4 sub-matrixes

$$T_{uu} = \begin{pmatrix} T_{ll} & T_{lr} \\ T_{rl} & T_{rr} \end{pmatrix} \quad (13)$$

It is noted that $\mathbf{P_L}$ never really changes. Zhu et al. proved that P_L converges to $(\mathbf{I} - \mathbf{T_{rr}})^{-1}\mathbf{T_{rl}}\mathbf{P_T}$ if $\mathbf{T_{uu}}$ is a right stochastic matrix[13]. Thus, the initial value of $\mathbf{P_L}$ is inconsequential. Using the same approach, we could prove that $\mathbf{P_Q}$ also converges.

4 Experiments and Discussion

4.1 Experiment Setups

The goal of our experiments is to evaluate whether our algorithm is effective in detecting pornographic Web pages. Given a labeled seed set L, our algorithm will return a list of pages that are ranked according to the possibility of being pornography. Seed pages are not included in this list, and the pages connected with only one query are also removed from this list because we think it is arbitrary to make a decision from only one query.

We use two datasets to build the bipartite graphs separately. From both datasets, we extract the information of the query, URL and timestamp. Private information is reduced as much as possible without introducing any ID or IP information.

The first query log dataset was collected from May 1, 2012 to May 14, 2012, with the help of a popular commercial search engine company in China. We pruned all of the query-URLs that appear only once in dataset because they could be noisy and potentially private. After that, the query log comprised 2,625,029 unique queries, 4,699,150 unique URLs and 72,106,874 query-URL pairs, which involve 717,916,107 individual clicks.

The second dataset is the America Online(AOL) query logs released in 2006 for research[1][14].This dataset contains 16,946,938 unique(normalized) queries, which were collected from March 1, 2006 to March 31, 2006.

[1] More information about the AOL dataset : http://www.gregsadetsky.com/aol-data/

4.2 Seed set selection and labeling criteria

The seed set contains labeled pages for our detection algorithm. On the Chinese dataset, we obtained a pornographic page list that contains 700 popular Web pages with the help of the same search engine that provides click-through data. This list was annotated by professional assessors, and each of the pages was double checked by us. A total of 691 pages appear in our click-through data, and we use them as the seed set for our algorithm.

For the English dataset, we picked out the URLs which contains 'sex' or 'porn' in the domain name to generate a candidate set. From the candidates, we randomly select a group of Web pages and have three human annotators with professional skills to label them as pornography or not. The labeling process stops when there are 500 pornographic pages in the seed set. The labeling criteria is stated as follows:

- NONPORN - The page contains no porn materials.
- BORDERLINE - The page contains some sexual material but no pornography.
- PORN - The page contains pornographic material.
- CAN NOT CLASSIFY - The page can not be accessed or the accessor could not classify it.

We also use these criteria to evaluate the results of our algorithm. It should be noted that we adopt a a relatively strict judgment rule on the pornographic pages in the seed selection step. All of the "BORDERLINE" pages are as "NONPORN", because the cost of mislabeling a normal page as pornography seed is much higher than the opposite situation. All of the "CAN NOT CLASSIFY" pages are removed in both the seed selection step and the result evaluation step.

When we labeled the seed set and results, some of the pages could not be accessed for different reasons. We attempted to label them according to the snapshots obtained from commercial search engines. If snapshots could not be obtained, we labeled them as "CAN NOT CLASSIFY".

4.3 Performance Comparison

We conducted our algorithm on both the Chinese dataset and the English dataset and compared their performances by the area under the receiver operating characteristic curve (AUC) and precision score.

We observed that the possibility of pornography changed little after 20 iterations. Specifically, we ran the iteration process 20 times in our experiment and then output the results. On a PC with a Intel CPU of 3.3 GHz and 32 GB RAM, the algorithm finished 20 iterations in 45 minutes on Chinese dataset and in 20 minutes on English dataset.

After the algorithm ends, we rank the URLs by the possibilities of pornography in descending order. Similar with the annotation approaches adopted by Gyöngyi et al in [15] , we separate the URLs into ten buckets sequentially and make sure that each bucket has an equal sum of possibilities that belong to the URLs in it. From each bucket, we randomly label 50 URLs with the criteria in Section 4.2, and we rank all of the URLs by their probabilities in descending order to generate the results list. We evaluate this list with both AUC and the precision, which are calculated based on the list. Content analysis based methods were not used for comparison in this step because their methods mainly focus on the classification of specific Web page sets while our method addresses the pornography detection within a large set of Web pages.

The experiment results are shown in Figure 3. From this figure, we can see that the AUC values are greater than 0.83 and the precision values are greater than 0.91, which suggests that our algorithm is effective in detecting pornography. The performance on the Chinese dataset is slightly better than on the English dataset, probably because the size of the English dataset is smaller and the Chinese dataset is more up-to-date.

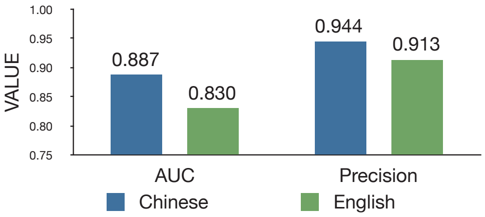

Fig. 3. Performance comparison between the Chinese dataset and the English dataset

We also want to see our algorithm's performance on detecting various forms of pornography. We randomly selected 280 URLs from the Chinese results that represent pornography on the Web page, and we manually classified their main pornographic forms into 4 categories: Text, Image, Video and Others(e.g., pornographic audio, pornographic chatting service). Of all the Web pages, 42% represent pornography information with text, most likely because this venue is the cheapest way to attract traffic from a search engine. Other research regarding anti-spam[16] shows that pornographic terms are one of the most important categories that lead to spam pages on Chinese Web pages. In our experiment, we also find that many of the porn pages are spammy with the purpose of cheating.

4.4 Disscussion: Pornography Score as Feature

In practical Web application, it is difficult to identify pornography Web pages only with the pornography scores given by our algorithm. However, we can use this method to generated candidates for further context analysis. This will help to reduce the number of pages to be analyzed. Also, we can use the pornography score as a feature to classify weather the Web pages contains inappropriate material. We implement a text-analysis method and compare the performance by adding the pornography score(PS) as a feature. In our implementation, each Web page is represented as a vector of TF-IDF values. Classification results on 812 Web pages(4-fold cross-validation) is shown in Table 1.

Experiment result shows that we can get better classification performance by adding the feature of pornography scores. However, the improvement is limited because classifiers have already reach a quite good performance with only text features.

	Precision	Recall	F-measure
SVM	0.924	0.924	0.924
SVM+PS	**0.963**	**0.962**	**0.962**
Naive Bayes	0.919	0.909	0.907
Naive Bayes+PS	**0.952**	**0.952**	**0.952**

Table 1. Performance Comparision by Adding Pornography Score Feature

4.5 Disscussion: Algorithm Robustness

Seed selection is very important in semi-supervised algorithms. Therefore, an experiment was conducted to see how robust our algorithm is. We only conducted this experiment on the Chinese dataset because it is newer and much larger than the English dataset. We randomly split the pornographic page seed set into 14 subsets(each subset contains approximately 50 seed sites) and then gradually added the subsets into the seed set to observe the influences on the algorithm's performance. The results are summarized in Figure 4.

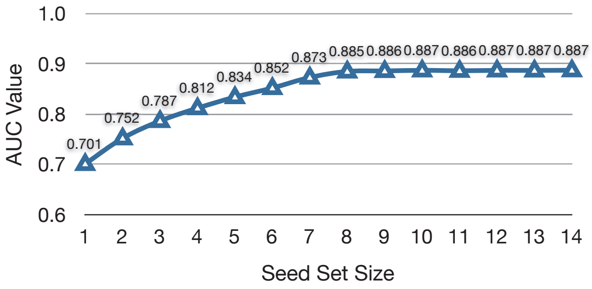

Fig. 4. Performance on different sizes of seed sets

It can be observed that our algorithm is very robust because it can achieve a relatively stable AUC value after only 400 pornographic pages are added. This experiment shows that our algorithm gained a stable performance in detecting pornography from a small number of seed set.

5 Conclusions

The very large number of pornographic Web pages has raised concerns about protecting children and adolescents. Traditional pornography detection methods focus on the extraction of textual or multimedia features from Web content, which consumes a large amount of computational resources. This paper attempts to solve the problem from a new perspective by proposing a novel method that is based on label propagation on a large scale bipartite click-through graph. First, a bidirectional edge weight definition is introduced to measure the correlation between the query and the URL reasonably. Then, we propagate the pornographic possibilities for all of the URLs iteratively on the click-through graph from a small set of seed URLs. The experiment that was conducted on both the Chinese and English datasets indicates that our method can detect pornography in different forms both effectively and efficiently. We hope that this method will be useful for protecting children and adolescents from pornography.

For future work, we plan to combine our pornography detection method with traditional content-based methods to improve performance. More specifically, our algorithm could return a candidate set for further content analysis efficiently.

References

1. S.S.A. Guan and K. Subrahmanyam. Youth internet use: risks and opportunities. Current opinion in psychiatry, 22(4):351C356, 2009.
2. J. Ropelato. Internet pornography statistics. TopTenReviews. com, internetfilter-review. toptenreviews. com/internetpornographystatistics. html, accessed Dec, 3:2012, 2006.
3. M.L. Ybarra and K.J. Mitchell. Exposure to internet pornography among children and adolescents: A national survey. CyberPsychology & Behavior, 8(5):473C486, 2005.
4. M.P. Goldstein. Congress and the courts battle over the first amendment: Can the law really protect children from pornography on the internet. J. Marshall J. Computer & Info. L., 21:141, 2002.
5. PY Lee, SC Hui, and ACM Fong. An intelligent categorization engine for bilingual web content filtering. Multimedia, IEEE Transactions on, 7(6):1183C 1190, 2005.
6. W.H. Ho and P.A. Watters. Statistical and structural approaches to filtering internet pornography. In Systems, Man and Cybernetics, 2004 IEEE International Conference on, volume 5, pages 4792C4798. IEEE, 2004.
7. J. Polpinij, C. Sibunruang, S. Paungpronpitag, R. Chamchong, and A. Chotthanom. A web pornography patrol system by content-based analysis: In particular text and image. In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, pages 500C505. IEEE, 2008.
8. P. Resnick and J. Miller. Pics: Internet access controls without censorship. Communications of the ACM, 39(10):87C93, 1996.
9. P.Y. Lee, S.C. Hui, and A.C.M. Fong. Neural networks for web content filtering. Intelligent Systems, IEEE, 17(5):48C57, 2002.
10. L.H. Lee and C.J. Luh. Generation of pornographic blacklist and its incremental update using an inverse chi-square based method. Information Processing & Management, 44(5):1698C1706, 2008.
11. R. Du, R. Safavi-Naini, andW. Susilo. Web filtering using text classification. In Networks, 2003. ICON2003. The 11th IEEE International Conference on, pages 325C330. IEEE, 2003.
12. G. Su, J. Li, Y. Ma, and S. Li. Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model. Journal of Zhejiang University-Science A, 5(9):1106C1113, 2004.
13. X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD- 02-107, Carnegie Mellon University, 2002.
14. G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In Proceedings of the 1st international conference on Scalable information systems, page 1. Citeseer, 2006.
15. Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pages 576C587. VLDB Endowment, 2004.

16. Chao Wei, Yiqun Liu, Min Zhang, Shaoping Ma, Liyun Ru, and Kuo Zhang. Fighting against web spam: a novel propagation method based on click-through data. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR 12, pages 395C404, New York, NY, USA, 2012. ACM.
17. P.J. Carnes, D.L. Delmonico, and E. Griffin. In the shadows of the net: Breaking free of compulsive online sexual behavior. Hazelden, 2007.