

Anti-Index: Against Privacy Mining via Search Engines

Xiaofeng Meng Jing Ai Zhongyuan Wang

*School of Information, Renmin University of China
Beijing, China 100872*

{xfmeng, aijingruc, zhywang}@ruc.edu.cn

ABSTRACT

With the flourishing of Web 2.0, Internet has become the largest information depository, which contains huge personal information of Web users. The information related to a specific person is usually scattered on various pages in different websites. However, today's Internet has been highly crawled and indexed by search engines. The malicious attacker may collect a specific person's information via search engines and obtain some privacy-sensitive information. Therefore, we observe a new type of privacy problem on the Internet: *Privacy Mining via Search Engines*. Our experiment shows this problem is serious yet easily ignored; it is a potential threat to Web users. We present a privacy mining model for describing the process of privacy leakage via search engines. To prevent this kind of privacy attack, we propose a method called *Anti-Index* and extend robots.txt to ERobots.txt with the fine-grained access control policies from the perspective of website constructors. In addition, we suggest a new service: automatically detecting potential dangers of privacy leakage for Web users. We also discuss several challenging problems about the privacy mining via search engines.

1. INTRODUCTION

With the rapid development of information technology, Internet has been developing at a tremendous pace. With the sharp rise of the amount of information, Internet has become the biggest information space. More and more information has been published on the Internet. Some is personal information related to web users. e.g. employees' identity data in business enterprise's webpage, and students' archives (including grade, record and identity, etc.) in school's database, etc. In the last several years, with the flourish of Web 2.0, websites aren't the only information publishers. Web users are allowed to publish their articles and views on the Internet. It is very convenient and easy to publish personal information on web 2.0 websites for web users, e.g. blogs, forums, etc. Therefore, publishing personal information on the Web becomes popular. Many web users would like to put some personal information including privacy-sensitive information on the Internet for seeking help and exchanging ideas.

Usually, the information of a specific person is distributed on

different webpages. However, today's Internet has been indexed highly by search engines. Therefore, it is possible to collect the decentralized information of a specific person together via search engines because the clues which link up this decentralized information are also indexed by search engines. This causes a serious problem which is usually ignored by people. A malicious user can collect the decentralized information about a specific person via search engines. Obviously, such information collecting may mine web users' personal privacy and bother person's life. People usually put too much confidence in the privacy preserving of websites where they publish personal information. But some websites provide their page hyperlink for search engines to improve website traffic. Even if some web2.0 websites have established protective mechanisms for privacy information publishing (e.g. web users must access these sites through a username and password), they block information communication and sharing for fresh users. We focus on how neither harm the exchange and sharing of information, while protecting the privacy of users.

To preserve personal privacy on the searchable Internet, we can solve it from several perspectives. In this paper, we propose a solution from the perspective of website construction in subsection 4.1. This method prevents some information from being indexed by search engines. That is just why it is called *Anti-Index*. We also propose a service to solve this problem from the perspective of web users. We call this service *S-Mining*, which will be elaborated in subsection 4.2.

1.1 Motivation Example

Although personal information is distributed throughout the Internet, the malicious person may mine privacy information via search engines. If the bad guy obtains both privacy-sensitive information and identity information related to one specific person, his privacy is leaked.

To verify how serious the problem is, we make an experiment (you can find more details in subsection 3.2). We collect one thousand users' privacy-sensitive information from a diabetes forum. And as Fig. 1 shows, about 46% of users' identity information is found out via search engines. Thus the problem is serious, widespread yet easily ignored. In the experiment, we find many users also have accounts of MySpace, Flickr, YouTube, etc., where we may find their identity information.

We now look at a concrete example in order to give a sense of privacy mining via search engines. Although the victim Alice's personal information is distributed over the Internet, the attacker Trudy may obtain both privacy-sensitive information and identity information related to Alice. In Fig. 2, each data item represents a message of Alice, e.g. Name, Email, Add., etc. Each page contains several data items and these pages are indexed by search engines, e.g. the page *a* can be regarded as an identity webpage which comes

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the WAMDM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the WAMDM Lab. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the author or WAMDM Lab.

WAMDM Technical Report, WAMDM-TR-2009-001, January, 2009, Renmin University of China, Beijing, China

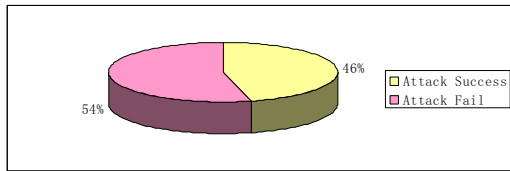


Figure 1: The Privacy Mining Experiment

from the website of Alice’s company, and the page *b* comes from a forum about diabetes where Alice asked for help.

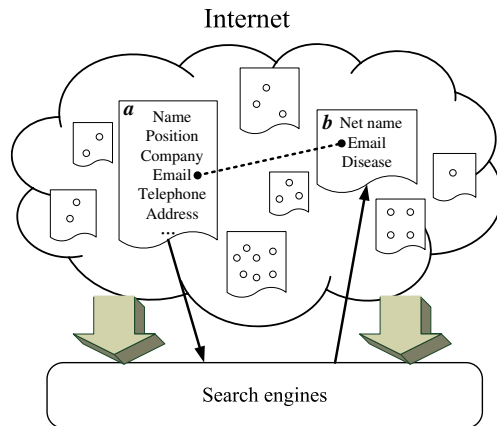


Figure 2: Motivation Example

The two pages have been crawled and indexed by search engines. All of data items in the two pages are stored in the index databases of search engines (we use index database loosely to include Google File System and other search engines’ index system). Search engines “know” all the relationships among data items on different pages on the Internet.

Trudy has Alice’s email “Alice***@gmail.com”, and searches can be launched for keywords, like “Alice***@gmail.com” or “Alice***”. In returned results, there are many pages containing these keywords. Trudy obtains these related webpages and examines all data items. She just holds data items related to Alice and uses some of them as new keywords for next searches. Then Trudy can discover new related pages. It looks as if there are link bridges among these webpages. Even if Trudy just has little information about the victim Alice at first, the privacy may be collected and mined successfully via search engines.

1.2 Challenges

Challenge 1: The problem of privacy mining via search engines is a serious but easily ignored problem on indexed Internet. It is a potential threat to Web users. But up to now, this problem has not been clearly and sharply defined yet. In order to remind users and researchers to notice this new problem, it is very important to define this new kind of threat and to model the attack process of privacy mining via search engines.

Challenge 2: Most of the existing information protection mechanisms are usually based on visitors’ identity authentication through the login mechanism of username/password before they access the page. However, we argue that this method may block the normal information sharing and exchanging, and it’s not conducive to newcomers. The second challenge is how to preserve web users’ privacy in the premise of guaranteeing normal information sharing

and exchanging on the Internet.

Challenge 3: Furthermore, if a user’s information has been indexed, he may want to know what information can be collected together and which information plays the key role in the attack. The problem is how to detect the potential dangers of privacy leakage for users automatically, and how to detect the linkage information in the process of mining.

1.3 Contributions

In summary, this paper makes the following contributions:

1. In this paper, we show a new kind of privacy threat to web users: privacy mining via search engines. We give the clear and specific definition of it and propose a model for this new privacy problem. Our model can describe the principles and process of privacy information mining via search engines.

2. Based on the model, we propose an effective method *Anti-Index* to solve the problem from the perspective of website constructors. We extend robots.txt to the new standard: ERobots.txt. Our method not only can defend users’ privacy against the threat of information collecting and mining via search engines, but also doesn’t affect normal information sharing and communication on the Internet.

3. Regarding challenge 3, we think a service should be provided from the perspective of Web users. We call it *S-Mining* service. This service can help Web users identify whether their privacy may be leaked. There are two crucial problems according to our model. One is *Linkage Information Ranking* which refers to how to identify which is linkage information and how to rank the linkage information for different users. The other is *Result Pages Sensitive*.

The rest of the paper is organized as follows. Section 2 presents the problem definition and introduces the privacy information mining model. Section 3 discusses the relationship of information preserving and sharing. Section 4 proposes solutions from different perspectives and gives some challenging problems. Section 5 reviews the related work. Finally, Section 6 concludes the paper.

2. PERSONAL PRIVACY MINING MODEL

In this section, we define the problem and introduce a privacy mining model. It describes how the attacker mines the privacy information via search engines, e.g. Google, Yahoo, Live.com, etc. Then we analyze the possible attack paths.

2.1 Problem Definition

For better describe the problem of privacy leakage on the Internet, we define several kinds of special information according to their characteristics:

Identity information (I): *One person’s open social identity.* E.g. SSN, ID number, name, occupation and company which he is affiliated with, etc. This kind of information can be used to confirm a person’s identity uniquely.

Privacy-Sensitive information (S): *Privacy-sensitive topic or personal privacy.* Users may publish it in special cases. Usually they don’t want to link it with their identity information. E.g. the information about seeking help for a debilitating disease, recovering alcoholics, gambling online, etc.

Other information (O): *Some inessential information, e.g. interest, educational level, marriage state, etc.* It can’t help identify people’s identity, and doesn’t refer to privacy-sensitive information.

Universal Set (U): *The set containing all the information of a specific person.* $U = I \cup S \cup O$.

I and *S* information seldom appear on the same page. If *I* information and *S* information related to the same user are collected together by the attacker, this user’s privacy is threatened. Thus we

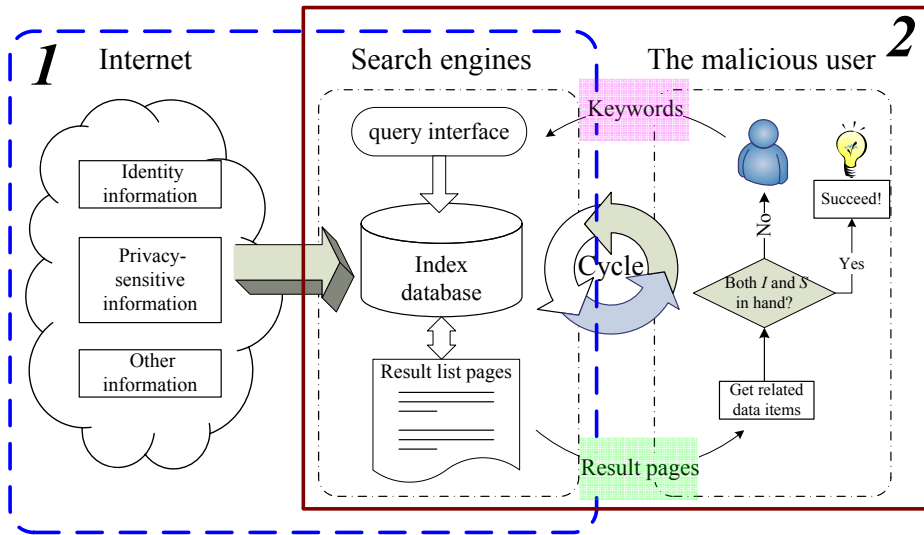


Figure 3: The framework of privacy mining model

make the following definition:

The Success of Privacy Mining Attack: *The bad guy has obtained both identity information(I) and privacy-sensitive information(S) related to one person.*

Why the scattered information can be collected together? There are common parts among different webpages' information, such as the same keyword or value. This is the essential reason why the malicious attacker can collect scattered personal information together via search engines.

Linkage information (L): *The common information contained by different webpages. It plays the role of bridges among pages.*

The linkage information plays a significant role in privacy mining via search engines. In our motivation example, the email and the email prefix is the linkage information. It bridges between the identity page and the privacy-sensitive page.

2.2 Framework of The Model

Fig. 3 shows the whole process of the malicious user relating the victim's identity information with his/her privacy-sensitive information via search engines. We consider all information about a specific person on the Internet as a data set, including three kinds of information: I , S , and O . All of them have been indexed by search engines.

The malicious user has obtained some information about the victim beforehand. The information may be I , S , or O . Even it may include I with O , or S with O .

Without loss of generality, we can suppose the malicious user just has I information on hand. He can launch queries with keywords from this information. Then he will get many returned webpages as query results. From these webpages, the attacker may find new unknown information about the victim. This new information may be I , S , or O .

(1) If the new information is S , the attacker has obtained I and S now. At that time the victim's I information is linked up with his S information! User's privacy has been predominated by the attacker.

(2) If the new information is I or O , the attacker can use the new information or a part of it as keywords for the next query. He can also use the combination new information with original information as keywords. Then, the attacker sequentially repeats above process until he finds S information of the same victim from returned web-

pages. This is a circular process. The results of the last query are used as the input keywords of the next query. According to relationships among information, the decentralized information related to the same user may be collected together in this circulation. Along a specific mining path, S will be found out finally.

The relationships among the information of the same user is the fundamental causes of privacy mining based on available information on the Internet via search engines. The attacker can obtain new information from old information according to the relationships among them. And search engines with powerful retrieval capability provide a great convenience and help for attackers.

2.3 Privacy Mining Paths

As mentioned above, all the information about a specific person usually doesn't display on the same page of Internet, especially the identity information and the privacy-sensitive information. If the attacker wants to reveal the users' privacy, he must find the same person's privacy-sensitive and identity information on the Internet via search engines.

Mining paths: *The process of mining the personal scattered information throughout the Internet via search engines.*

Firstly, the attacker must have partial information in hand. That may be I , S or O information. If it is I , the attacker must find the corresponding S . If it is S , he must find the corresponding I . And if it is O , he must find the corresponding I and S which have common O as the bridge. He can submit queries to search engines, using the partial information in hand as keywords. He may find some more information about this user, and he can use them to do further queries to get more useful information. He will search information about this person on the Internet via search engines' searching ability until he has both I and S information in hand. At that time, the attack has succeeded. As showed in Fig. 3, this process runs circularly. The searching process seems like a path that the attacker kept to on the Internet. Certainly, there are many steps in the process. We define them as the attack paths.

After carrying on many observation and experiments, we define five kinds of attack paths to help people understand the attack process. Respectively, they are shown in Table 1.

The asterisk in mining paths of Table 1 indicates there are zero or more of the preceding item found in the search results. The ta-

Table 1: Privacy Mining Paths

Begin with	Via	End with	Mining Path
Identity Info	Identity Info	Sensitive Info	$I \rightarrow \{I\}^* \rightarrow S$
Identity Info	Identity Info Other Info	Sensitive Info	$I \rightarrow \{I\}^* \rightarrow O \rightarrow \{O, I\}^* \rightarrow S$
Sensitive Info	Sensitive Info	Identity Info	$S \rightarrow \{S\}^* \rightarrow I$
Sensitive Info	Sensitive Info Other Info	Identity Info	$S \rightarrow \{S\}^* \rightarrow O \rightarrow \{O, S\}^* \rightarrow I$
Other Info	Other Info Identity Info or Sensitive Info	Identity Info and Sensitive Info	$O \rightarrow \{O\}^* \rightarrow \{O, I\}^* \rightarrow \{I \& S\}$

ble includes all possible paths of mining information at the privacy attack. Only if through these paths, the malicious person can mine one person’s privacy successfully.

3. INFORMATION PRESERVING VS. SHARING

As elaborated in Section 2, there are some special relationships existing among the scattered webpages containing information related to the same person. These pages can be linked together according to such relationships. Therefore the attacker can collect the scattered information together via search engines. The relationship between webpages is the linkage information that they all contain. The cause of privacy leakage based on accessible information is that linkage information is indexed by search engines. And linkage information is necessary for forming privacy mining paths in the attack process.

3.1 Block Pages or Linkage?

To preserve users’ privacy on the searchable Internet, it can be addressed from several perspectives.

Some websites have established encryption and user’s identity authentication mechanism. Users must login with username and password, and then can access the content of the website. This method does can protect users’ information, but all information of the website is unable to be indexed by any search engine. Thus it is contrary to the thinking of information communication and sharing among Web users. And this mechanism may make newcomers (e.g. the person who have a disease recently) confused, because they can only search some keywords in search engines and may not find the website with authentication.

E.g., the topic of a forum is about diabetes. The forum takes the user authentication mechanism. If a new user wants to learn something about diabetes, the forum is very helpful because there are many experiences and true feelings. But the new person didn’t know the website of the forum before. The only way that he can take to search information on the Internet is search engines. But when he searches information about diabetes, he will not get any information about the forum.

Therefore, blocking webpages is not a good solution for our problem. We need to find solutions which can preserve users privacy against the threat of information collecting and mining via search engines without blocking normal sharing and communication of information on the Internet. Because it’s linkage information, not pages, that causes privacy leakage. If the linkage information isn’t indexed into the search engines’ database, it is impossible to form the attack paths via search engines. We propose an effective

solution in subsection 4.1.

3.2 A Statistical Experiment

We conduct a statistical experiment to verify the seriousness of this problem and observe the importance of different linkage information. We pretend ourselves to be the attackers, and attempt to collect users’ *I* and *S* information together via search engines. We begin with privacy-sensitive information, and attempt to find out the identity information which belongs to the same user.

Firstly, We randomly choose three web forums which contain multiple users’ *S* information. They are web forums or communities for some sensitive diseases: cancercompass.com, diabets-daily.com and aidscommunitieservices.com. Then, we choose users randomly from each forum as experimental objects. We check the content of the information he published beforehand, and we just take users who referred to himself suffering from this disease as experimental objects. We obtain the initial data set from these three forums: 1000 users’ *S* information.

Then we search these users’ identity information via search engines. Each user has a unique username on the forum, and some users put their email addresses. We use these data with other information on these forums (such as location, gender and so on) as initial keywords for search. In the experiment, we discover that social websites such as MySpace, Facebook, Flickr, often play the role of intermediate bridges. The information users published on these websites is usually related with them and offers facilities for attackers. Through multiple steps, we find some users’ *I* information on some companies’ or social organizations’ pages.

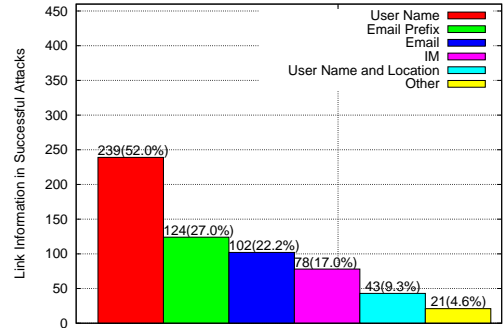


Figure 4: Statistics of linkage information Frequency

Experimental results show: 460 (success rate of 46%) users’ identity information and privacy-sensitive information could be gathered up successfully. In Fig. 4, the Y-axis represents the times of each data item playing the role of linkage information in all accomplished cases, and the rate represents the ratio of this number to the number of accomplished cases.

As the statistics of experimental results shows, user name, email address, email prefix, IM account and the combination of user name and location are the top five data items which are the most possible to play the role of linkage information in actual mining processes. Especially, Email address and IM account can be used to confirm uniquely one web user. Although some data items, e.g. user name and email prefix etc., can’t confirm one person uniquely, they still played an important role of searching out new webpages containing linkage information as the keywords of queries. These data items and other data items can be used as the assistant evidences to identify whether the information on the newly returned pages belongs to the same person.

4. DISCUSSION

From Fig. 3, we know the model can be divided into two parts: *part 1* and *part 2*. These two parts represent different perspectives. The *part 1* represents the process of three kinds of information indexed by search engines. In this process, the linkage information indexed by search engines causes the privacy leakage on the searchable Internet. Therefore, we propose an effective solution called *Anti-Index* in subsection 4.1. The *part 2* represents the process of privacy mining via search engines. If this process can be implemented automatically, we can provide a service for people who care about their privacy on the Internet and suggest them which information is their linkage information on the internet. This service called *S-Mining* and some challenging problems are discussed in subsection 4.2.

4.1 Anti-Index by Servers

Subsection 3.1 indicates that the linkage information bridges different webpages and forms privacy mining paths. So we discuss an effective solution from the perspective of websites' constructors. The essence of this solution is to filter linkage information at the arrowhead of *part 1* in Fig. 3.

Theoretically, search engines' crawling activities can be controlled from the server side by deploying the Robots Exclusion Protocol in a file called robots.txt. The file called robots.txt [8] contains robot access policies, and it is deployed at the root directory of a website and accessible to all robots. Moral robots read this file and obey the rules during their visit to the website.

But, the expressive granularity of restrictions in general robots.txt file is too rough to make sophisticated regulations for search engines' crawling and indexing. The minimum granularity of the restrictions in Disallow field is page-level.

<pre>User-agent: * Allow: /content/ Disallow: /privacy/ Disallow: /tmp/*.htm</pre>	<pre>User-agent: * Anti-Index: /privacy/ Anti-Index-Id: Email, Phone Anti-Index-Update: /privacy/</pre>
(a) robots.txt	(b) ERobots.txt

Figure 5: robots.txt and ERobots.txt

To solve above problems, we propose to extend the robots.txt. We add three fields to promote robots.txt from page-level to object-level which are showed in Fig. 5(b). The following fields are added:

Anti-Index: *The value of this field specifies a partial URL that is needed to anti index.* This can be a full path, or a partial path; any URL that starts with this value will not be retrieved. E.g. "Anti-Index: /privacy/" in Fig. 5(b) means all pages under "privacy" directory need to filter some link information for robots.

Anti-Index-Id: *The value of this field specifies the unique identifier to an element on Anti-Index webpages.* It may be the id name of a div, a table or other HTML tags. E.g. "Anti-Index-Id: Email, Phone" in Fig. 5(b) disallows robots index tags whose id is "Email" or "Phone".

Anti-Index-Update: *The value of this field specifies a partial URL that is needed to be updated by search engines immediately.* For various reasons, some linkage information may already be indexed by search engines and it may be potential threat to Web users. This field requires search engines update these specified pages im-

mediately and apply new rules specified in robots.txt to update their index databases.

By these new fields, the robots.txt can express some rules based on objects on webpages. It's promoted from page-level to object-level. We call this new robots.txt *ERobots.txt*. With *ERobots.txt*, website administrator can make the linkage information not be indexed by robots easily. Thus the mining paths can't be formed because there are not bridges among different webpages.

We develop a Apache module called Anti-Index Module. This module controls webpages' access. It reads rules of *ERobots.txt* and enforces the policy. We add this module to a real web server and observe it over a period of time. The result shows it's effective without noticeable decline in performance.

4.2 S-Mining Service

Although the *Anti-Index by server* method can prevent *l* information from being indexed by search engines, it is impracticable to demand all web servers to adopt the *ERobots.txt* standard and add an enforcement. In this section, we try to solve this problem from users' perspective.

Let's recall the motivation example(subsection 1.1). Alice is a department manager of a famous IT company and has an introduction page on the company's website. Unfortunately, she also suffers from diabetes. She ever posted articles for help on some diabetes forums.

But Alice cares about her privacy. She wants to know if her introduction page on the company's website and the privacy-sensitive page where she referred to her disease could be collected together via search engines.

This is a challenging problem: *how to detect potential dangers of privacy-sensitive information leakage for Web users automatically.* We call it *S-Mining*(*S* denotes privacy-sensitive information) for short.

For providing such service, it is necessary to run the collecting process automatically described in the *part 2* of privacy mining model(Fig.3). There are two crucial problems in the *part 2*, which have different background colors.

4.2.1 Linkage Information Ranking

The first problem is how to identify which information is linkage information(*l*) and how to rank this linkage information. It has an important impact on the efficiency of *S-Mining*.

By subsection 3.1, the scattered information related to one specific person is joined together by linkage information on webpages. In *S-Mining* service, every search needs to choose some keywords and the linkage information is a subset of all query keywords. If the query keyword is *l* information, it can obtain new valuable information and join webpages together. Therefore choosing effective keywords for search is very important to the efficiency of *S-Mining* service.

Although we make a statistical experiment in subsection 3.2 and obtain a ranking list of linkage information, the probability of a keyword as linkage information is not immutable for different users. For example, Alice is a department manager of an IT company and her office number may be *l* information; while Jim is a student of a college and his college name and student ID may become *l* information. Therefore, rank of linkage information (*l*) for different users automatically can make detecting service more efficiently.

4.2.2 Result Pages Sensitive

The second problem is how to identify the sensitive level of result pages, which decides whether the attack is successful.

For different users, different situations, even different time, the

meaning and implication of privacy-sensitive information may be different. But there is always some special information, which most ordinary people consider as sensitive information, such as AIDS, drug addicting and so on. The information such as political inclination, family status may be defined as lower level. How to correctly define the sensitive level of a specific webpage according to its content is a very challenging problem. Perhaps the Web user needs to input his privacy demands, and then the service system should be able to translate the demand to appropriate levels.

5. RELATED WORK

Until now, the personal privacy mining problem on the Internet is not paid enough attention by researchers and Web users.

In traditional research, the most essential and widely applicable privacy model was the k-anonymity model [16, 13]. It can resist the privacy attacks that identifying individuals by joining the published identity information table with some external tables containing privacy-sensitive information. However, the kind of privacy model doesn't apply to our problem. The goal of attacks discussed in k-anonymity model is to join tables published by certain organizations. By contrast, the attacks in our problem may involve all the information published on the Internet. Our problem's scope is broader and more complicated.

There are many privacy concerns on the social networks. The environment of this kind of researches is social networks or popular Instant Messaging (IM) networks. Generally speaking, these websites provide such mechanism that can enable users to restrict access to *friends' list* or *circle of trust* [11]. It is a selected group of contacts. For secure personal web content sharing in such networks or personal profile, [3, 4, 11, 1, 6] proposed many methods to protect users' information from being accessed by unexpected visitors except *friends' list* or *circle of trust*. Most of these methods are based authentication mechanism and encryption. But because of the differences of environment, the protection mechanism of these works is different from ours. The users' social network data is completely forbidden to access for strangers not in *friends' list*. The topic of that is to prevent attackers not in *circle of trust* from accessing or obtaining users' published information in this user's Instant Message, social networks or personal website. However, we focus on all searchable information about users on the Internet. Our work aims at protecting user's identity information and privacy-sensitive information from being collected together by malicious attacker.

In addition, social network can be regarded as a part of Internet, and the information just comes from the social network, while our work is to solve the problem of information mining and collecting on the whole Internet. The quantity and scope of information in our problem are greater.

There is also some work [9, 14, 10] on finding entities' relationship on the web and [12] proposed a set of techniques for finding terms that are correlated to one or more query terms. In this kind of research work, it finds the relationship between two entities via search engines. In contrast with our problem, from the perspective of the attacker, it collects all information about the victim together.

The robot [5] or spider access policies are communicated in the file called robots.txt, and it is deployed at the root directory of a website and accessible to all robots. There is also some work [2, 7, 8, 5] focus on this problem. On the problem of robots.txt, [15] is a very valuable work. It presented a survey of the use of the Robots Exclusion Protocol on the Web through statistical analysis of a large sample of robots.txt files. In our solution, we discuss the problem of current robots.txt and extend the page-level robots.txt to object-level ERobots.txt.

6. CONCLUSIONS

In this paper, we propose a new kind of privacy threat on the Internet: *privacy mining via search engines*. We proposed a privacy mining model to describe the process and principles of mining information via search engines, and the most popular five privacy mining paths are formalized. We propose methods from two perspectives (server-based and service-oriented solutions) and make some discussions about challenging problems need to be addressed in the future.

7. REFERENCES

- [1] M. Bellare and C. Namprempre. Authenticated encryption: Relations among notions and analysis of the generic composition paradigm. In *AsiaCrypt*, 2000.
- [2] M. Drott. Indexing aids at corporate websites: The use of robots.txt and meta tags. In *Information Processing and Management*, volume 38(2), pages 209–219, 2002.
- [3] C. Dwyer and S. R. Hiltz. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In *Proceedings of the Thirteenth Americas Conference on Information Systems*, Keystone, Colorado, August 2007.
- [4] R. Feizy. An evaluation of identity on online social networking: Myspace (poster). In *ACM Hypertext and Hypermedia (HT)*, 2007.
- [5] P. S. G. Pant and F. Menczer. Crawling the web. In *chapter Web Dynamics*. Springer-Verlag, 2004.
- [6] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2005.
- [7] B. Kelly and I. Peacock. Webwatching uk web communities: Final report for the webwatch project. In *British Library Research and Innovation Report*, 1999.
- [8] M. Koster. A method for web robots control. In *In the Internet Draft, The Internet Engineering Task Force (IETF)*, 1996.
- [9] G. Luo, C. Tang, and Y. li Tian. Answering relationship queries on the web. In *Proceeding of the 16th international conference on World Wide Web (WWW'07)*, pages 561–570, Banff, Canada, May 2007.
- [10] D. Mahler. Holistic query expansion using graphical models. In *New Directions in Question Answering 2004*, pages 203–214, 2004.
- [11] M. Mannan and P. C. van Oorschot. Privacy-enhanced sharing of personal content on the web. In *Proceeding of the 17th international conference on World Wide Web (WWW'08)*, pages 487–496, Beijing, China, 2008.
- [12] V. K. P. Tan and J. Srivastava. Indirect association: Mining higher order dependencies in data. In *PKDD'00*, pages 632–637, 2000.
- [13] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS98*.
- [14] F. L. Sanda Harabagiu and A. Hickl. Answering complex questions with random walk models. In *SIGIR'06*, pages 220–227, 2006.
- [15] Y. Sun, Z. Zhuang, and C. L. Giles. A large-scale study of robots.txt. In *Proceeding of the 16th international conference on World Wide Web (WWW'07)*, pages 1123–1124, Banff, Canada, May 2007.
- [16] L. Sweeney. K-anonymity: A model for protecting privacy. volume 10, pages 557–570, 2002.