Access control based on content

Sebastian Monte
Helsinki University of Technology
smonte@cc.hut.fi

Abstract

Internet has become a significant platform for sharing personal information and opinions. Web-based social networks (WBSN) are rapidly getting more users to their community, and unfortunately these users are usually not aware of the privacy risks related to sharing personal content. Current access control mechanisms in WBSNs rely heavily on the users to make correct access control policies and this might result in information reaching to people not intended to. Content based access control (CBAC) takes a different approach. In CBAC system, the access to an object is based on the content of the object in the system. So rather than requiring the user to specify security policies for each object, such as an image or a blog post, CBAC automatically applies a security policy based on the object's content. CBAC uses some well known technologies such as computer vision and natural language processing for identifying and analyzing the content. Currently no WBSN uses access control based on content. However, CBAC-enabled system such as PLOG (A Privacy/Policy aware bLOGging) and POG (Privacy/Policy aware Online Gallery) are being developed for research purposes.

KEYWORDS: Content-Based Access Control, Access Control Based on Content

1 Introduction

The Internet has evolved greatly from the time of static web pages. The publishers are no longer only IT professionals, but nowadays a wide variety of users are publishing and sharing their experiences in the Web. The content also varies greatly. People might write their experiences on blogs, share their pictures in Facebook and upload videos to YouTube. Regardless of the underlying service, one of the main reasons for participating in social networks is to share information and experiences with other users.

As people are sharing more and more information of themselves, the information might reach to persons not intended to. There are numerous examples from the media where the shared content is causing problems for the publisher or other people involved. For example a teen was fired from her job after calling it boring on Facebook[5]. A more scary scenario is where sexual predators are using social networking sites to find their victims[7].

Even though it is possible to configure various privacy settings, the users often have problems setting restrictions on who can and who cannot access the digital content. A research[8] concludes that access control mechanisms interfered with users' work flow and that they had problems with setting the correct permissions for objects. So there is clearly a need for more usable access control mechanisms.

With the increased content sharing, we need better ways to control access to the content. One promising approach is content based access control (CBAC). In CBAC system, the access to an object is based on the content of the object in the system. CBAC lets the users to specify a a single access control policy based on object features and then automatically apply that policy to new objects[3]. In this paper we discuss how CBAC can help people to protect their privacy in a convenient way in web-based social networks (WBSN). Carminati et al.[1] defines WBSNs as "online communities that allow users to publish resources and to record and/or establish relationships with other users, possibly of different type ("friend", "colleague", etc.) for purpose that may concern business, entertainment, religion, dating, and the like".

The structure of the paper is the following. We first look at the characteristics of some well-known WBSNs and how access control is implemented in them. Then we discuss how CBAC could be used in the WBSNs and its advantages and disadvantages. Lastly, we discuss about the future work of CBAC.

2 Current Solutions for Access Control in Social Networks

When a user registers to a WBSN, the system creates an account, to which the user will be able to add personal information about himself such as a name, age and gender. The user is also able to specify relationships with other users and manage personal resources such as blogs, videos and pictures. Of course once the content is added to the system, the users need to specify who can actually access that content. The very basic access setting is to mark the content as private (only the content owner can see), public (everyone can see) or accessible by direct contacts. Some sites offer additional access settings, for example Facebook[1] supports the "selected friends" option and Friendster[2] supports the option "friends of friends". Some WBSNs even support the option "my network", that grants access to whomever the user is in contact (indirect or direct).[1]

Michael Hart et al.[3] have surveyed 23 blogging and social networking sites to understand which types of access control and privacy features have been implemented already.

[1] www.facebook.com
[2] www.friendster.com

The survey included various types of WBSNs, such as famous blogging sites Blogger[3] and LiveJournal[4], the most popular social network Facebook and video-sharing sites such as YouTube[5] and Google Video[6]. From the survey, it can be concluded that the access control features fall into few broad categories (some sites offering minor extensions):

- **Private/public.** This is the simplest access control. Private objects can only be viewed by the owner, while public content is available to anyone, humans and search engines for example.
- **Friends.** Some WBSNs enhance the private/public separation by allowing users to create a list of friends. Once the user has specified which users are his friends, he may restrict the content to be only accessible to his friends. There are various extensions to this as discussed earlier.
- **Other users.** Some sites such as Xanga[7] provides users an option to deny access to content to users who are not registered to the service. It seems to be intended to block search engines to index some pages, as Xanga is free for anyone to register.
- **Search engines.** The Wordpress blogging site allows the author to deny access for search engines by using the robots.txt mechanism[8].
- **Password-protected posts.** Some blogging sites allow the author to create password-protected posts, so only the persons who know the password will be able to read them.

Table 1 shows a good summary of the protection options provided by some well known WBSNs. As can be seen from table, the "friends" relationship dominates the WBSNs. Every WBSN except RepCheck provides an option for classifying the content as public or private. Depending on the WBSN, the content can be protected to the n*th*-degree of contacts (for example friend's friends). Some WBSNs even allow you to categorize between online and real-world contacts, which is quite useful as people rarely want to share exactly the same content to different contact types.

Some WBSNs also allow the users to specify how much they trust each others, thus establishing trust relationships. For example LinkedIn provides the means to associate users with text labels explaining why he is recommended by another user. Orkut on the other hand provides a simple rating system to rate personal trust.

Although the current access control mechanisms seem to provide quite a decent amount of options, they are very inflexible and lack some common needs. First, the current WBSNs do not allow to segregate users to different social groups. For instance some of the friends might be real-world friends and some just people you have met online. A user might for example want to share certain information with his work-colleagues, but not to share it with online-buddies.

Friends is a very broad term for describing persons having some sort of a relationship. Hart et al.[3] have found that an average MySpace user has 115 friends. It is clear that the level of intimacy encompasses many levels, some being closer friends than others. It must also be noted that some people perceive having many friends as a good characteristic, so there is a certain balance between privacy and popularity. Some WBSNs allow users to group good friends over normal friends. For example Facebook allows a user to select some friends to his "entourage" and MySpace provides an option to display "Top Friends" above normal friends in the profile page. The term "friend" becomes quite ambiguous, so there is really no sense to build an access control mechanism around "friend" relationships.

All of the above access control schemes require users to manage their access control policy on a per-object basis. This is usually frustrating to users, who are unwilling to spend much time on specifying access control policies. WBSNs also have very heterogeneous collections of objects, such as text, music and pictures and all of these should be given some access control policies. To make matters even worse, there is a constantly changing set of people who might desire access to the content, therefore the policies must be updated accordingly. The owner is usually not an expert on security issues, so mistakes are common.

To overcome the shortcomings of standard access control polices in WBSNs, we need high-level policies that can be expressed in simple ways, applied automatically and updated easily. This is the goal of CBAC systems, of which we will discuss in the next chapter.

3 Control Based Access Control

In CBAC system, the access to an object is based on the content of the object in the system. This in in contrast with Role-Based access control in which the decision to grant or deny access is based on the role of an individual trying to access the content. In simple CBAC, the content of an object can be used to tag it with labels from a set of features. The access control policy of those features are specified by the user. Once the basic CBAC infrastructure is present, it is easy to extend it to support other natural access control policies.[3]

Hart et al.[3] describes several relevant access control schemes that really illustrate the power of CBAC systems.

- **Blog posts on topic T are visible to users in group G.** Since managing groups is usually easier than managing individual objects, this policy let's the user to classify individuals to groups as with some of today's WBSNs. However, this policy is applied automatically to new posts by computing a topic for each post and then applying the above rule. There already exists these sort of systems, for example adult-content filters try to recognize mature topics and then deny access to that content if the individual is not in the "adults" group.
- **Group G consists of all users that I mention in my post.** To easy the burden of defining and managing

[3] `http://www.blogger.com`
[4] `www.livejournal.com`
[5] `www.youtube.com`
[6] `video.google.com`
[7] `www.xanga.com`
[8] The robots.txt files are used to give a web crawler instuctions on how to index the. More information can be found for example in `www.robotstxt.org`

WBSN	Purpose	Relationships	Trust	Protection Options
Bebo	general	friend	none	public, private, 1st-degree contacts, selected contacts
Facebook	general	friend	none	public, private, 1st-2nd-degree contacts, selected contacts
Friendster	general	friend	none	members from selected continents, private, 1st-2nd-degree contacts
MySpace	general	friend	none	public, members > 18 years old, private, 1st-degree contacts
Multiply	general	various	none	public, private, 1st- and *n*th-degree contacts, 1st-degree but not online contacts, selected contacts
Orkut	general	friend	personal	public, private, 1st-2nd-degree contacts
Flickr	photos	friend/family	none	public, private, 1st-degree contacts (friends or family)
Last.fm	music	friend	none	public, private, 1st-degree contacts (and profile neighbours)
Xing	business	generic	none	public, private, 1st-4th-degree contacts
LinkedIn	business	various	business	public, private, 1st- and *n*th-degree contacts
RepCheck	reputation	generic	personal, business	none

Table 1 Protection options in WBSNs. Source: [1]

groups, this policy automatically classifies which users belong to the group of individuals which are allowed to view the post.

- **Group G consists of all users that demonstrate sincere interest in topic T through their posts.** This policy is probably the most powerful one concerning the nature of WBSNs. As people are commonly sharing their personal information and opinions in WBSNs, this policy might prevent employers from evaluating current and potential employees based on their personal opinions. Of course the employer could create a bogus blog for example, but this requires substantially more work than just simply looking up at the employees blog looking for disagreeable opinions. More over, this policy encourages people to meet like-minded people in WBSNs, which is one of the primary goals of WBSNs.

Hart et al.[3] also present two other scenarios where CBAC can be useful: inference control and integrity control. Inference control prevents stalkers from getting information about user's location from his WBSNs. For the user it might be difficult to check every post (also prior) and consider whether people might be able to infer his hometown from the information. By using natural language processing techniques, the system could detect a violation of this policy and offer the user an option to change the content. Of course this sort of review is not bulletproof, as people might still be able to infer a location even though the system cannot. However, if the system can detect a violation in the policy, other people will surely also be able to infer the user's location.

Collaborative knowledge-based sites such as Wikipedia[9] allows users to add information freely in hope of increasing wisdom in the community. Wikipedia operates on user trust and therefore, is subject to repeated acts of vandalism. By using natural language processing technology for ensuring information integrity, we could for example detect when a text document switches from one topic to another[2]. This way many acts of vandalism can be detected automatically, thus letting Wikipedia editors to devote more time on checking facts of new posts.

CBAC uses different technologies for analyzing the content, such as text summarization and computer vision. Naturally, the flaws and drawbacks of these technologies also affect on the CBAC system's features. However, no access control system is perfect so we should not expect CBAC to be either. Users with strong privacy requirements can always review the automated decisions as well. This way they will benefit from the security options provided by a manually administrated system, but at the same time allowing them to enjoy the convenience of automatically applied security policies.

4 Architecture

Figure 1 shows an architecture for implementing a CBAC system[3]. First the administrator would select certain security schemas. The choice of these rules will focus on content type (e.g. video, image, sound etc.), on people and roles (e.g. "Mike", my parents) and on the relationships that these people and the content have. For example a user might only want for his brother to see his holiday pictures, so he would specify "my brother Mike can see my holiday photos". An access control matrix (ACM) is derived from the identified content, roles and policies the user has set up. ACM is an abstract security model, where each row is a "subject" (a role for example) and each column is an "object" (a file for example).

When content is loaded into the system, it is first categorized according to the content type by using some content taxonomy. The user is able to add new content types to the taxonomy and may change the system's content classification if needed. The system is hopefully smart enough to

[9] wikipedia.org/

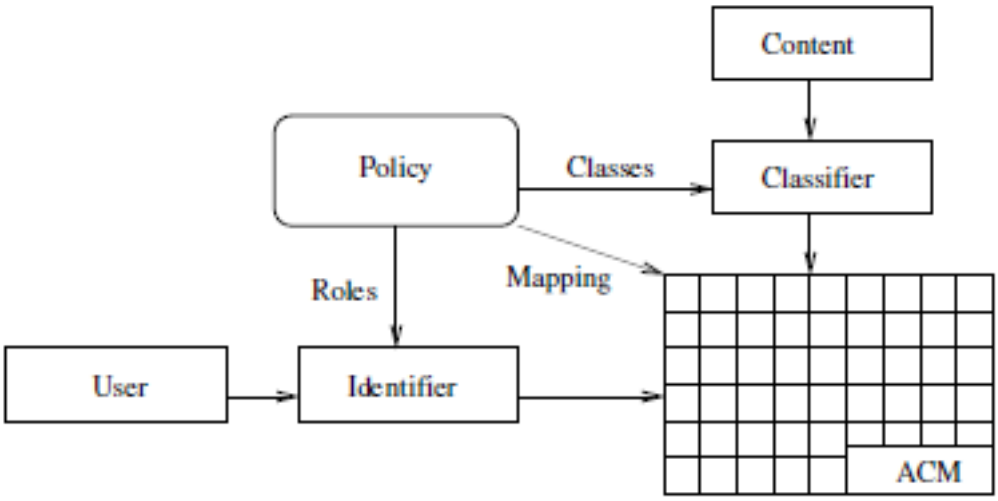

Figure 1: An architecture for implementing CBAC[3].

learn from the users previous choices so the content classification becomes more accurate as time goes on. Once the content is successfully classified, people and their roles are identified. Once the system knows the people and their roles related to the content, relationships between content and people are identified using for example patterns that to look for sentences like "my brother Mike can see my holiday photos". Once the relationships are computed, they are used to create access control policies.

5 Open Issues

Today's access control mechanisms rarely provide they users what they actually need in order to protect their privacy. The privacy controls are too labor intensive or people might just be unaware of privacy risks. For example Facebook has 37 privacy options for the user to choose from and this computes to more that 10^{14} possible security combinations[3]. Though the security configurations provide a huge amount of options to confuse the user, Facebook has millions of users. This said, we can confidently come to a conclusion that a system can be imperfect access control wise and still attract a large amount of users. Therefore, we should not expect CBAC to solve every problem perfectly when concerning access control. The goal is to improve the access control mechanisms, so that users can with relative confidence publish personal information as well, without constantly worrying about information misuse.

Even though the technology behind a CBAC system is not perfect, the areas are actively researched. For example Microsoft is investing in natural language technology to provide better user experience for their customers[9]. Fields related to computer vision such as face recognition receive also much attention these days. Face recognition is extremely important in CBAC WBSNs, as much of the users' content are images. Lee et al.[4] have proposed an efficient method for face recognition in different illuminations. This should be quite beneficial in CBAC as users' image environments vary greatly. Li et al. [6] proposes a fully automatic video retrieval system that integrates head tracking, face alignment and face recognition algorithms. This sort of technology would also certainly benefit CBAC as users often upload videos.

6 Future Work

Currently there is no WBSN that uses content based access control to determine who has access to which content. However, Hart et al.[3] are developing a system called PLOG (A Privacy/Policy aware bLOGging) for testing CBAC. PLOG should be able to provide automatic access control that is expressive and usable. This will help users to decide how they really want to protect their content, for example "share my entities related to food with people who have sincere interest in cooking".

The PLOG system is still under development. Based on its results, the same people are also planning to develop CBAC-based system called POG (Privacy/Policy aware Online Gallery). Since people like to share pictures of their lives, the developed system should be of high interest to normal users. For example some users do not want their family and other random web users to see pictures of themselves in certain activities. The system will use computer vision to identify the places and people, and this information is presented to the user in an intuitive way so the user can create photo-sharing policies with less effort.

7 Conclusion

Internet has become a significant platform for sharing personal information and opinions. This information usually has a clear target group, for example only your friends should see some of your personal messages posted in Facebook.

Since the users are commonly unaware of the security risks related to publishing, the information might reach persons unintended to.

Nowadays WBSNs provide a rich selection of security options, but the relationships are somewhat simple for the real world. As mentioned earlier, "friends" can mean many things to different people. CBAC tries to provide the users an easier and a more convenient way to handle privacy. Rather than requiring the user to specify security policies for each object, such as an image or a blog post, CBAC automatically applies a security policy based on the object's content. It bridges the gap between user's privacy goals and current access control mechanisms by providing a convenient and easy to use privacy controls. This is extremely important in WBSNs as they seem to achieve ever higher user population.

Though CBAC still has a long way to go before it can be successfully implemented, some practical testing is been done. Systems like PLOG and POG provide valuable information on how users might use CBAC-enabled systems. If positive feedback is received, this probably encourages other researchers and companies to contribute to the subject. Overall, CBAC seems to be more a more user friendly way for people to protect their privacy, creating and updating countless security combinations can be really frustrating.

References

[1] A. P. Barbara Carminati, Elena Ferrari. Enforcing access control in web-based social networks. *Transactions on Information and System Security (TISSEC)*, 13(1), October 2009.

[2] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[3] Michael Hart and Rob Johnson and Amanda Stent. Content-based access control. 2006.

[4] C. H. P. Moonhwi Lee. An Efficient Image Normalization Method for Face Recognition Under Varying Illuminations. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 128–133, Vancouver, British Columbia, Canada, 2008.

[5] F. News. Teen fired for calling job 'boring' on facebook, 2009 (accessed February, 2010). http://www.foxnews.com/story/0,2933,501025,00.html.

[6] Y. L. C. H. Pengxu Li, Haizhou Ai. Video parsing based on head tracking and face recognition. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 57–64, Amsterdam, The Netherlands, 2007.

[7] K. Poulsen. Myspace predator caught by code, 2006 (accessed February, 2010). http://www.wired.com/science/discoveries/news/2006/10/71948.

[8] Thara Whalen and Diana Smetters and Elizabeth Churchill. User Experiences with Sharing and Access Control. In *Extended abstracts CHI 2006*, pages 1517–1522, Montréal, Québec, Canada, 2006.

[9] A. Woodie. Microsoft buys colloquis for natural language technology, 2006 (accessed March, 2010). http://www.itjungle.com/two/two101806-story05.html.