Position Paper for W3C workshop on “Privacy for Advanced Web APIs”

Simon Moritz
Ericsson Research, 16480 Stockholm, Sweden
Simon.Moritz@ericsson

Abstract

Privacy has become critical for many Web APIs. Particular challenging are those APIs used in relation to techniques referred to as data mining, which task is to gain knowledge from personal information. In this paper some of these challenges are addressed, and it is our intention that this position paper may highlight the importance of common terms of use for handling such information. Our prime interest in participating in the workshop is to learn about the rising privacy challenges in data mining, so that we can become a part when creating the necessary technology enablers that has to be in place to protect the end users. As food for thought and as a step to offer user control over personal information we will in this paper share one possible implementation. We believe that the only way something similar may be rolled out, is when several actors work together and agrees on a potential solution.

1. Introduction

In the future we believe that everything will be connected, everything will communicate, and everything can be connected everywhere since connection is wireless. All these connections creates vast amount of data. Data that users will be encouraged to offer to cloud services, which hosts the data through a multitude array of applications and APIs.

In a world where your data is no longer stored on your personal device, but in the cloud, the personal control may fade away. This paper addresses some of these issues that will arise in respect to Web APIs used for data mining purposes. Among those are; information overflow vs. user control, information spread to more than the original thought destination (forwarding to third party), too much knowledge by merging several information sources (inductive learning), data ownership, need of knowledge competition, constant need of user consent, and how can a user can withdraw a given user consent in such way that it restricts further use of her personal information that may have been distributed.

In this paper we will first described the identified problems, present a potential way forward as food for thoughts and conclude with some key challenges that has to be met. Our interest in this workshop is to get to know the other actors in this area and see what we together can do that offers end users control over their personal information.

2. Problem

Below follows a short description of six challenges we identified in a data mining context.

2.1. Information Overflow vs. User Control

Still in a world of vast data there exist great needs of help for users to find what they are looking for. Sometimes the users may be able to specify a search questions, other times the users may need to rely on techniques to provide the information that they want. These data mining techniques base their knowledge on e.g. the users’ history, behavior, consumptions and ratings. This data, one’s digital footprints, may create patterns in the daily life, revealing who the user really is.

2.2. Forwarding to Third Party

When sharing information, explicitly or implicitly, by leaving digital footprints it is easy for an individual to loose control. Either one forgets what has been shared, or to whom it has been sent to. These footprints may also get wings, meaning that they may be spread from the thought destination to other contexts, when a service provider offers the information to a third party.

2.3. Inductive Learning

Inductive Learning is learning from preparatory information. An example of this is when a third party has been granted access to obscured digital footprints.

This data may be correlated with other data revealing more about an individual than originally imagined, which may indirectly identify a user. In this way the third party has gone outside the original user consent and, in effect, violated the user's privacy. This is hard to control and thereof a serious threat against privacy.

2.4. Data Ownership and Knowledge Distribution

From a global perspective it is good to keep as much information as possible distributed on several hosts rather than creating one single point of information. With information comes knowledge and with knowledge a lot of power. Hence, the one hosting all personal information will automatically possess a great deal of knowledge and through that also a great deal of power. This creates new risks and threats to privacy.

Today there exist no solutions for the user to know where her data is stored, but according to the directive [1], European laws has to be followed for all data generated from Europeans, both when stored within the borders and outside.

2.5. Constant Need of Consent

User consent is a way for the user to be in control of the personal information collected about them. Examples of such information are: call logs, texts, time stamps, duration of call, start and finish time, what the user have been using etc. Simply put, all information that an operator can collect from one single user.

According to the third principle in the EU data protection directive [1] data should "not be disclosed without the data subject's consent".

Though consent is a widely used approach for most Web Services and APIs, it is a tedious task to design an accurate one. To mention one challenge; how to write a consent that explains the purpose of the information usage in such way that the subject understands.

In relation to data mining the issue above is particularly tricky since the very essence of data mining is that one cannot fully know the outcome in advance. Data mining is a secondary future use while the primary purpose of the collection must be clearly understood and identified at the time of the collection [2].

2.6. How to Withdraw a Given Consent

The subscribers should be given simple means to withdraw their consent for the processing of traffic data at any time [3], and for location data free of charge [4].

However, there are no technical solutions that fully make sure that such data is prevented from being used in the future for data mining purposes after consent has been withdrawn. Consequently, when consent has been withdrawn, the models created by data mining algorithms, has to be recalculated too.

3. Scenarios

In order to exemplify where these issues may emerge are here different scenarios presented that relates to the work we do at Ericsson research within the data mining area.

3.1. Calculation as a Service – Cluster Constructor

At Ericsson we are studying ways to enable end user to find their way through all the content while still preserving their privacy. A lot of these enablers are offered through Web APIs and some of these have we decided to share publicly on our LABS portal [5].

One example when this is used is for the Cluster Constructor API. The API is composed of basically three calls. First a CSV file has to be uploaded to the server, then trigger the data mining clustering process, and thirdly request the result from the clustering process using GET requests. In all of these calls an API key is used. This key authenticates the user as well as authorizes her to perform operation only on her own data that may be stored on the same server as many other users' data.

The Cluster Constructor web service offers calculation process power and clustering know-how to a developer who otherwise maybe would not be able to calculate this on her own. However, just as in any web service and cloud computing setup, the users have to entrust the web service provider with their data. The only guarantee offered in return is normally the terms of use and the risk of bad will for the service provider if they break those terms. There are seldom any technical protections which the end user may use when they for instance would like to withdraw their data.

Hence, information that has been uploaded through the API, is protected from unauthorized users, but may still risk being used for various data mining purposes without the user's ability to restrict if the models are not recalculated.

3.2. Recommender Service

In an IPTV system users are challenged to find something that personally interests them among all the content available at any time. It is often hard to exactly formulate what to look for, except for those occasions when looking for e.g. a favorite movie. For this reason users normally prefers to be suggested something by the system rather than writing a search query. More, and perhaps stronger, reasons are the lean back approach that those users tend to have while watching TV. It is more relaxed to be suggested that actively search for something yourself.

To suggest to a user what to watch without stating a question can be solved by letting a recommender system suggest something to watch. The recommender system collects user consumption logs, such as star rating, watch or not watched etc, to learn what the user is interested in. Based on this knowledge the system predicts what to watch next, e.g. by using collaborative filtering techniques.

The problem in such techniques is that the data has to be uploaded to the server in order to find the semantic connections between users of same taste. If then the user data then is forwarded to a third party, as in the Netflix case, there are risks of inductive learning that may reveal things about the user which was not expected in the first place. This is why Netflix was sued for privacy invasion in 2009 [6].

3.3. Location Based Service

Location based services (LBS) are sometimes used to provide users with interesting information in reachable geographical distance. Such information could be restaurants, shops or local campaigns.

Other reasons for location based services that we have worked with are the reversed, to enable advertisers to pin point the users in a particular area.

Location based information could either be collected on a mobile terminal or in the network. E.g. when a user makes a phone call, or makes any other interaction with the telecom network the position is recorded in e.g. a Visiting Location Register (VLR).

The knowledge of where a user is located is very sensitive. An article of Limits of Predictability of Human Mobility [7] shows that with more than 80 percent accuracy it is possible to predict where an individual will be located next by looking at her past.

Collecting this data creates huge possibilities for better tailored services on one hand, but at the same time it reveals some of our most sensitive information, our current and future location.

3.4. Data Mining on Telecom Data

Telecom operators have access to more than location data. Many operators have started to mine their customers' service behavior. Some to predict churn behavior with hope to win back customer trust. Others look at customers' social network graph to tailor services and optimize package deals.

There is a huge amount of advantages, cost savings, efficiency making and customer adaptations, which an operator can find when learning more about its' customer. As the world turn more and more mobile the need of such understanding will emerge.

As a company providing technology to host more than 40 percent of all mobile calls and as an enabler for the larges operators in the world, are we on Ericsson are keen on helping our customers, the operators, with their tedious task of improving their services. Some of the operators are happy as a bit pipe, while others hope for additional services to flourish their business. In doing so they need to understand what their customer needs and it is our intention while creating the necessary underlying technology to help them while including all necessary techniques to keep the privacy of the end users.

4. Possible Ways to Protect Personal Information

We have in this paper addressed a number of problems in relation to data mining and Web services that all can be mapped in to a telecom context.

To solve the privacy challenges mentioned there is no single solution, but there may exist different ways and approaches to offer the end users tools that make them more in control of their originated data.

One of our aims, on Ericsson, has been to create a new technique that gives the end users a mean to hinder further use of their personal information. This can be achieved by creating a privacy policy framework which restricts use of personal information in a similar manner as a digital rights management (DRM) system. The core of the idea is to force third parties to continuously request verifications from end user trusted party before performing working with the data. As a side effect, specific usage rights could be added to the lock inductive learning in the same way as DRM today hinders addition data to be added.

Another important part in keeping the privacy, we believe, is to distribute the information to a number of hosts. In this way the risk of someone misusing the information will decrease. If they would, the user could simply churn and the host will loose its market share.

How to treat the location information is still to be decided. Maybe the best solution is to offer to the user means to change their location by will. On the other hand, the fact that someone actively changes her location have been said to reveal more about the user than the actual location itself.

As far as the law concerns, there are still too few prejudice to state what should be done or not, but to conclude the best way to be on the safe side is to collect consent.

6. Conclusion and Future Work

Laws may restrict certain misuse of personal information, but we believe that it is equally important to create the necessary technology that offers users control of their digital footprints. When withdrawing their consent, the technique should hinder further usage, just as it should limit to whom the information may be spread to. Finally the personal data should preferably be distributed on hosts that value their good will, and therefore also how they treat the individual's information.

7. References

[1] The Data Protection Directive (officially Directive 95/46/EC)

[2] Thearling, K. An introduction to Data Mining, 2007 www.thearling.com (Accessed May 2010)

[3] EU directive 2002/58/EC article 6:3

[4] EU directive 2002/58/EC paragraph 35.

[5] Ericsson Labs http://labs.ericsson.com (Accessed May 2009)

[6] Netflix lawsuit 2009 on Wired http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/#ixzz0osDZLihW (Accessed May 2009)

[7] Predictions of Human mobility http://www.barabasilab.com/pubs/CCNR-ALB_Publications/201002-19_Science-Predictability/201002-19_Science-Predictability.pdf (Accessed May 2009)