Privacy Protection in Context-Dependent Retrieval of Information and Multimedia

ELENA VILDJIOUNAITE, VESA KYLLÖNEN
Technical Research Centre of Finland
Kaitovayla 1, P.O.Box 1100, FIN-90571, Oulu
FINLAND

Abstract: - This work presents a method and experiments with privacy protection in multimedia and information retrieval system, currently being developed in Amigo project. Since Amigo environment is capable of recognition of user situations (contexts), the recommender system takes into account both long-term user interests and contexts when providing recommendations. We propose to utilize context recognition also for privacy protection, and suggest a method which either allows or suspends recommending of an item via non-personal UI, depending on a current context and retrieval history. This work studies how privacy protection affects precision and recall of recommendations. Two privacy-protection techniques were explored: protection based on the user's social context (other people around the user); and protection based on the user's location. Experimental results on data, collected via user interviews, show that social context-based protection works better than the location-based one; and that during normal family life privacy protection does not decrease system performance significantly. Instead, in some cases system performance has been improved.

Key-Words: - Privacy, Privacy Enhancing Technologies, Context Awareness, Information Retrieval, Recommender Systems

1 Introduction

As capabilities of computers to record, store and present to the users all kinds of information and multimedia constantly grow, personalization of their retrieval becomes an important research area. Indeed, it is more convenient for the users if their computers record and recommend web pages, TV programs and advertisements which the user is really interested in, instead of presenting same vast set of possibilities to everybody. However, recommender system can cause privacy problems to the user (can disclose the user's very private personal interests), if it recommends some special advertisements' topics or TV programs when the user is in a company of other people. Such risk is especially high in systems which learn user preferences by observations of user actions (instead of asking the users to manually enter their preferences and to update them each time when they change), because in these systems users have less control over system actions. On the other hand, these systems are more convenient for users because they reduce the need to enter preferences manually.

We suggest that privacy protection in recommender systems should be based on learning of user's behavior in different situations (contexts), because nowadays computers and personal devices become increasingly capable of reliable context detection. In Amigo project [1] the consortium is developing services for an intelligent home environment capable of recognizing such contexts as identities of people in the room (by means of speaker recognition [2] and by RFID positioning [3]); personal plans [4]; presence of visitors etc, which will be used in many applications, including personalized information and multimedia retrieval.

Current state–of–the-art in Privacy-Enhancing Technologies is mainly concerned with controlling what kind of information about the user is acquired, transmitted from him over networks and stored in the systems, and with its protection from malicious use. Research on protecting users' privacy in cases when information flow is directed towards users, as in recommender systems, is less active.

The main goal of privacy protection in network applications (such as Web browsing, e-commerce, ad-hoc communications between people in bars etc) is to provide anonymity or pseudonymity to the users, and to provide unobservability and unlinkability of their actions. Compare to network applications, achieving anonymity in an intelligent environments (especially home environment) is much more difficult, because users and applications are not spatially separated and number of users is fairly small. For example, if a recommender system includes unusual shopping advertisements or videos

in a list of "items of possible interest" when several family members are going to watch TV together, it would not be difficult to figure out who caused such system behavior. Thus, the main goal of privacy protection in smart environment is to check carefully whether each piece of information can be accessed by each user, or not. Without such access control it will be more difficult to buy surprising presents to somebody's birthday; and more difficult to save children from being aware that videos with violent and other "only for adult" scenes exist at home.

One way to solve this problem would be to let users to mark some information as "private", but it would require an extra effort, and users can forget to do it. There can be also certain predefined categories of private topics (e.g., adult videos should not be accessed by children). However, even if the topic of somebody's interest is not really a secret, anyway it is often beneficial to avoid public announcement of it. For example, every child knows that his/ her mother sometimes buys new clothes, and usually has nothing against – with the exception of cases when an advertisement of a nice dress comes just after the child was refused to buy a new computer game or a toy. Thus, we propose that recommender systems should consider all information as private by default.

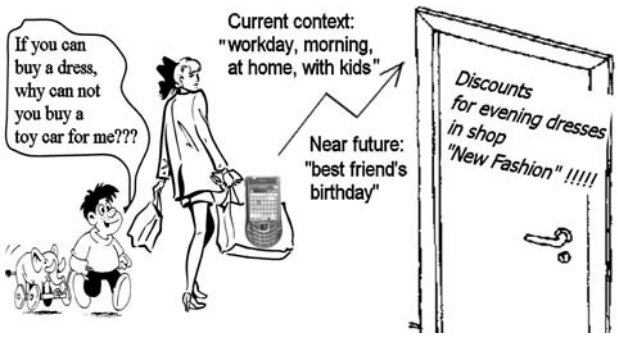

Figure 1 Scenario to avoid for a recommender system.

This work describes first experiments with protecting personal privacy in context-aware recommender systems by learning, in which contexts the user retrieves different items.

2 User Modeling in Context

Since personal information needs and multimedia preferences depend not only on the long-term user interests, but also on current situation (context) of a person (e.g., people are interested in weather forecast in their future destination, and people are interested only in short news digests when they are in a hurry), recommendation systems should take such dependency into account. Traditionally, user modeling in information and multimedia retrieval was aiming at learning only preferences of one person independently on context, whereas context-dependency of interests was not considered because context recognition is a fairly new research area.

Recently, since it is clear that watching videos and media generally is often a social activity, the researchers started to look for methods to take into account interests of multiple persons. The work of Masthoff [5] studies different strategies which humans apply to solve this problem. The work of Goren-Bar et al. [6] suggests to infer (based on time) which subset of family members is watching TV, and to personalize programs for this time. The work of Ardissono et al. [7] presents how learning dependency of user interests on time (day of week and time of day) improves recommendations. The work of Adomavicius et al. [8] suggests to model user interests in context as multidimensional spaces, and shows how this approach facilitates recommendations. However, these works do not pay special attention to privacy protection; while works devoted to privacy issues in recommender systems [9] discuss mainly risk of disclosure of stored data.

With the possibility to detect more contexts in Amigo home environment, we suggest to use following context variables in the user model:

- current and future social context (presence of people). Current context can be deduced from users' locations, detected by RFID tags and/ or speaker recognition; participants of future events can be acquired from the users' calendars
- current and future locations
- current and future events of the users and of their close relatives, such as trip, birthday etc, which can be acquired from the users' calendars
- day of the week and time of the day

We selected these context types after discussions with users regarding which factors affect their interests; and because these context types can be recognized by the Amigo system. We express each user context as a set of descriptors, for example:

Time of Day: morning; Day of week: Sunday; Location: home; Social Context: alone; Near Future Event: work trip to Brussels; its Social Context: alone; Near Future Event: child's birthday; its Social Context: family, relatives...

In the list above underlined items denote values which corresponding context types can take. Context values can be symbolic (e.g., "alone") or numeric (e.g., "19.00" for "time of day" context type). In this study we used only symbolic context values and converted numeric values into symbolic ones, e.g., converted hours into "morning", "afternoon" or "evening" symbolic values.

Context information is one side of user model. Another side is information and multimedia items the user is interested in (retrieves) in each context. User interests are expressed in a form similar to the context representation form, e.g.: *Shopping Pages: books: detective stories; Source: Amazon; News: Sports: skiing; Source: local newspaper; News: weather Brussels; Source: CNN....*

In these experiments we were interested to learn topics and genres of user interests, and used Case-Based Reasoning method for this purpose. CBR works as follows: the current user context is compared with all cases in interaction history, and most similar cases and corresponding interests are retrieved, along with their ranks. Ranking is based on similarity between current and stored contexts, which usually do not match exactly. The number of recommendations depends on whether the degree of similarity exceeds a threshold, calculated online (not predefined) in such a way that number of recommendations is not too small and not too large.

For calculating the similarity between contexts we apply Cosine measure, commonly used in Information Retrieval.

$$\textbf{\textit{sim}}\,(U,V)=\frac{\sum_{i=0}^{N} W_i * Sim(U_i,V_i)}{\sqrt{\sum_{i=0}^{N} U_i^2}\sqrt{\sum_{i=0}^{N} V_i^2}}$$

W_i (weight) is introduced to the formula in order to give application designer more control over the system, e.g., to assign higher importance to some context descriptors. However, in these initial experiments all weights were set equal.

3 Context – Based Privacy Protection

Diversity of personal interests and situations, that affect personal interests, makes it more comfortable for the users if the system learns their interests from the history of past interactions, instead of asking the users to set preferences manually and to update them each time when they change. However, such learning is dangerous, because it can disclose private interests of users when they are in a company of others. For example, recommending certain kind of videos during visit of one's parents or a girlfriend can cause problems; user's interest in this or that news in a stock market tells about his/ her financial situation; showing toy advertisements in presence of kids can destroy mystery of Christmas presents.

Naturally a system designer can predefine private categories of items, such as financial information and adult videos. However, people might prefer to keep as private also other, seemingly ordinary preferences regarding advertisements or news recommendations. For example, some people browse the web or read news when they need to relax at work, but many of them don't want their colleagues and bosses to know details about it. Same applies to family relations: parents might not want to inform their children about their purchasing plans in order to avoid discussions regarding e.g. why do parents restrict a child in buying computer games, if parents have enough money for expensive home appliances. Similarly, spouses might not want to inform each other about prices of hobby equipment or clothes which they are interested in, or how many TV programs they watch without the other one.

Apart from consumer media, personal recommender systems should be able to deal also with home media (home-made photos and videos), because people often show home media to friends and relatives, taking into account their interests. E.g., people show hobby videos to hobby friends and city views to those who are interested in sightseeing [10]. Since home media often contains private scenes (like chaos in a kitchen, not fully dressed people or unlucky facial expression), choice of home media to show depends also on a degree of intimacy between people. Thus, an attempt to predefine private and public categories in home media would fail because it is very individual.

Thus, we suggest that the recommender system should treat each information and multimedia item as private by default; and should create one list of public items, one list of private items and in some domains also a list of semi-public items. We suggest two approaches to the problem, whether an item can be recommended publicly or not: first, social context –based method, which considers an item as public only if interaction history contains the case when all currently present people (and possibly also some other people) were present when this item was retrieved. Second, location-based protection, which considers an item as public if it was previously retrieved in the same location (we mainly distinguish between work and home locations). Experiments with location–based protection were made because usually social situation depends on location, and because recognition of social context is not yet as common as location recognition. Despite active research on recognition of users' identities in smart spaces and in other context-aware applications (see e.g. [11] for acquiring social context with Bluetooth-enabled mobile phones), acquiring rough estimate of personal location is much easier: every mobile phone can distinguish between one's home and office

locations by means of cell-ID based positioning, with the exceptional case when one lives very near to his/ her workplace(in our city granularity of cell-ID based positioning in city area ranged from a few hundred meters up to one kilometer). Thus, a very simple option for users' privacy protection would be to have clear distinction between items retrieved at home and items retrieved at work, and to mark items as public if they were retrieved in the same environment earlier. This solution would allow users to deal with secrets from family members in a workplace, and to keep colleagues unaware of personal preferences in home environment.

Apart from lists of public and private items, we suggest to create also a list of semi-public items - items, similar to the public ones according to some domain-specific criteria. Home photos, shown to other guests earlier, can be marked as semi-public, because the probability that the photo can cause problems is smaller in this case. However, this probability is not zero. For example, if one's close friends are interested to see interiors of a hotel or a sauna, the user might decide to show them a photo with a piece of underwear in a background, whereas showing such photo in another context might be awkward. Since computers are not able to compute visual similarity of photos well enough to detect private scenes even when examples of private photos are presented, we suggest that with respect to privacy protection, degree of similarity between photos should be based on user actions only.

On the contrary, consumer videos usually are annotated in sufficient details, and computers can estimate similarity of videos better than that of home photos, although not perfectly. Thus, we suggest that movies, similar to the ones retrieved in a same social context, can be added to a list of semi-public items.

It is important to emphasis, that distinguishing between public, semi-public and private items makes more sense when some persons have higher priority in system control than others: e.g., an adult and children; family members and guests. In this case the system needs to create only one private list of recommendations and to let the higher priority user to check it. If co-located people are more or less "equal" (e.g., two spouses), creation of several recommendation lists is feasible only if everybody has personal device at hand. Such situation is not unrealistic (often all family members have own mobile phones), but in any case everybody would first look at the list of public recommendations presented via common UI (such as TV or computer screen in a living room). Consequently, it is needed to evaluate how privacy protection changes public recommendation lists when several users with similar priorities are present, e.g., when all family members are gathered together.

Next section describes the experiments regarding how privacy protection affects lists of public items.

4 Experiments

The data for the experiments was collected via user interviews: three persons reported their information retrieval cases and TV programs during certain time period (two users reported approximately two weeks data, and the third user reported approximately one month data). To these cases we added 20% noise (retrieval of similar and arbitrary information in arbitrary contexts), taking into account that some cases were not described correctly or were simply forgotten. This resulted in approximately 300 IR cases on 55 topics, among which were several favorite sets of topics for different sets of context descriptors. Examples of such favorite topics for different contexts include:

- choices of the whole family on Friday evening;
- choices of children with and without parents;
- workday morning favorites for father;
- mother's favorites which she watches alone.

Our data contained also four strongly event-related topics (retrieved only in relation with some event):

- interest in Brussels weather before a trip there;
- showing of hobby videos to guests;
- TV show for a particular public holiday;
- search for toys before a child's birthday.

Other cases were occasionally retrieved ones, such as programs of moderate interest which the users watch only if there is plenty of free time; or reading of accidentally encountered web page.

In this study we used altogether 48 context descriptors (that is, each context was represented as a list of 48 numbers). 44 descriptors were selected as most important after the user interviews and were used to express context types and values described in the Section 2. Some of these descriptors were actually adding noise, e.g., favorite topics of users on all workday mornings were the same. In order to increase noise even more (we decided so because data collected via self-reporting of users usually contains less noise than data collected in real life system operation), we added four extra meaningless context descriptors and set their values randomly.

During the tests we were interested to compare the system performance with and without privacy protection on two tasks:

- to recommend favorite set of topics for corresponding context;
- to recommend event-dependent topics for the corresponding event.

Typically in Information Retrieval the performances of the methods are evaluated by their precision and recall. Recall denotes percentage of provided recommendations, relevant for a current context (term "context" denotes a set of context descriptors), with respect to the overall number of relevant topics. Precision denotes percentage of relevant recommendation with respect to all recommended items. The goal of experiments was to compare, how two privacy protection methods affect precision and recall of public recommendations in comparison with the system without privacy protection. Performances of methods on favorite topics are presented in Figure 2; performances on event-dependent topics are presented in Figure 3.

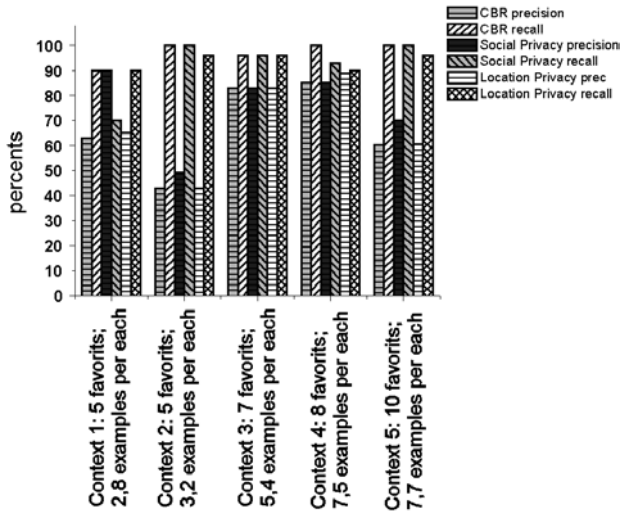

Figure 2 Precision and recall on favorite topics with and without privacy protection.

Since our data contained only a few positive examples per each favorite topic (average number of positive examples on topics considered as favorite in different contexts was six per topic, while average number of positive examples per event-dependent topic was three), it was impossible to split data for training and test sets, as usually. Thus, we generated test data per each set of favorite topics and per each event-dependent topic for corresponding relevant contexts additionally to the available training data. By "relevant contexts" here we mean a high-level description of situation when users are interested in a certain topic, for example, "workday morning" is a relevant context for reading sports news, but "child's birthday in near future" is irrelevant context for this topic. On the other hand, for retrieval of toys advertisements "child's birthday in near future" is the only relevant context, whereas "morning" or "afternoon" contexts were irrelevant in our data.

Accordingly, four test samples per each set of favorite topics (or the only event-dependent topic) were generated as follows: all context descriptors which were relevant for the topic retrieval were present, and irrelevant context descriptors were selected randomly. We did not add total strangers as a part of social context to the test data, because there would be no public suggestions in this case. Instead, we tested the affect of variations in presence of family members and friends. In these experiments we considered as potentially privacy-threatening all topics, because even a location for a suggested weather forecast can reveal personal plans.

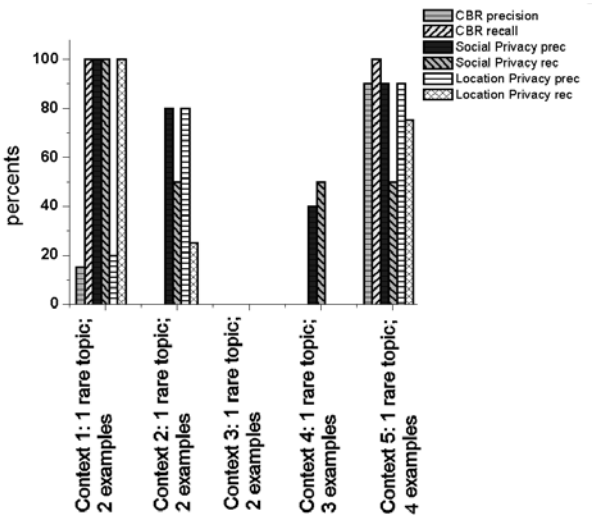

Figure 3 Precision and recall on event - dependent topics with and without privacy protection.

We would like to emphasis that in our experiments we were dealing with topics of user interest, such as genres of movies or topics of news, and our precision and recall would decrease when it comes e.g. to choice between particular movies of same genre. However, learning the topics of user's interests and distinction between public and private topics is an important first step for a recommender system. When more data is available, the recommendations will be refined.

4 Conclusion

This work proposes a method of privacy protection in context-aware recommender systems and experiments on information and multimedia retrieval data collected by self-reporting of users.

We propose that recommender systems should create lists of public, semi-public and private items, depending on current user situation (context) and interaction history. The most important for privacy

protection is social context (presence of other people around the user), but reliable recognition of it is not yet a common feature for most applications. Since social context depends on a user's location, which is easier to detect, we tested two different methods of privacy protection in recommender system: social context - based and location-based.

The tests were aiming at comparing precision and recall of public items. We compared three ways: no privacy protection; social context – based privacy protection and location–based privacy protection. In general, privacy protection can significantly decrease recall in public recommendations in case of unknown context, when most of items are marked as private or semi-public. However, we found that in case of normal family life privacy protection has not significantly decreased recall and precision. Instead, in some cases they have been improved, and in two cases of event-dependent retrieval social context – based privacy protection method has significantly improved the recall. This is due to the fact that the topics are recommended if their measure of similarity to previous cases exceeds some threshold, but this threshold is not predefined: it depends on overall number of candidates. We do so in order to control the total number of recommendations, which is especially important in cases when interactions with users happen via small screens of personal devices. In cases when private topics were removed from the list, extra topics (not recommended by a system without privacy protection) were added. For our data, where interests of family members differed from each other (which we believe to be a common situation), this turned out to be beneficial.

Although location-dependent privacy protection is less intelligent than social-context based privacy protection, nevertheless it has slightly improved precision for some favorite topics (whereas slightly decreased recall) and for one event-dependent topic. However, our previous experiences with context-aware systems have convinced us that users learn fast to benefit from such systems. Thus, if users know that privacy protection in recommender system is location-dependent, they might start intentionally utilize this feature.

It is worth noting that evaluation of effectiveness of privacy protection methods is very difficult due to limited number of life situations when disclosure of private data creates problems; and due to the fact that attempts to "catch" private data are against ethics. As it is generally the case with intelligent systems, privacy protection can lead to unpleasant surprises for users if they don't understand how the system works, e.g., the users might get confused when some items do not appear in a list of public suggestions, or when this list is empty due to presence of a stranger. We suggest that this problem can be solved by good design of user interface, which should present a prompt and an easy way to obtain also private suggestions, for example, by pressing an extra button. In a future we plan to study, how users understand our method of privacy protection, and how do they like it.

Long-term system use in real life and collection of users' opinions is needed for evaluation of how reliable and user-friendly are the proposed methods of privacy protection, but our initial experiments suggest their feasibility.

References:

[1] http://www.hitech-projects.com/euprojects/amigo/

[2] Haeb-Umbach, R., Kladis, B., Schmalenstroeer, J., Speech Processing in the Networked Home Environment - A View on the Amigo Project, *Interspeech 2005*, Lisboa, pp. 121-124

[3]Anne, M., Crowley, J., Devin, V., Privat, G., Localisation intra-batiments multi-technologies: RFID, WiFi et Vision, *UbiMob 2005*, Grenoble

[4] Lahti, J., Westermann, U., Palola, M., Peltola, J., Vildjiounaite, Context-aware mobile capture and sharing of video clips, *Handbook of Research on Mobile Multimedia,* Idea Group, 2006, pp.340-356

[5] Masthoff, J., Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers, *User Modeling and User-Adapted Interaction*, Vol.14, 2004, pp. 37-85

[6] Goren-Bar, D., Glinansky, O., FIT-recommending TV programs to family members, *Computers & Graphics*, Vol.28, 2004, pp. 149-156

[7] Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Chiarotto, A., Difino, A., Negro, B., User Modeling and Recommendation Techniques for Personalized electronic Program Guides, *Personalized Digital Television*, Vol. 6, 2004

[8] Adomavicius, G., Sankaranarayanan, R., Sen, Sh., Tughilin, A., Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach, *ACM Trans. Inf. Syst*., Vol.23, No.1, 2005, pp. 103-145

[9] Lam, Sh. K., Frankowski, D., Riedl, J., Do You Trust Your Recommendations? An Exploration of Security and Privacy Issues in Recommender Systems, *ETRICS* 2006, pp. 14-29

[10] Vildjiounaite, E., Sachinopoulou, A., Löthman, H., Lahti, J., Pietarila, P., Järvinen, S., Personalisation in content-based retrieval of home videos, *EuroIMSA 2005*, Grindenwald, 21 - 23 Feb. 2005. IASTED (2005)

[11] Mäntyjärvi, J., Gfeller, B., Social Cliques: Group Awareness for Mobile Terminals, *Enactive 05*