IDIS (2009) 2:155–169
DOI 10.1007/s12394-009-0026-2

Quantifying privacy in terms of entropy for context aware services

Athanasios S. Voulodimos · Charalampos Z. Patrikakis

Received: 12 January 2009 / Accepted: 2 September 2009 / Published online: 1 October 2009

Abstract In this paper, we address the issue of privacy protection in context aware services, through the use of entropy as a means of measuring the capability of locating a user's whereabouts and identifying personal selections. We present a framework for calculating levels of abstraction in location and personal preferences reporting in queries to a context aware services server. Finally, we propose a methodology for determining the levels of abstraction in location and preferences that should be applied in user data reporting during service provision, according to her personal privacy settings.

Keywords Context awareness · Entropy · Location based services · Personalization · Privacy

Abbreviation
API Application Programming Interface
GPS Global Positioning System
LBS Location Based Service

Introduction

Context aware services and primarily Location Based Services (LBS) are becoming increasingly popular, supported by the corresponding increase in the use of web

A. S. Voulodimos (✉) · C. Z. Patrikakis
School of Electrical and Computer Engineering, National Technical University of Athens,
9 Heroon Polytechneiou Str, Zographou, Athens 15773, Greece
e-mail: thanos@telecom.ntua.gr

C. Z. Patrikakis
e-mail: bpatr@telecom.ntua.gr

enabled, location and context aware mobile devices. The progress recorded in positioning techniques has been a supporting force. However, as location tracking capabilities are increasing, problems related to user privacy arise, since user's position and preferences constitute personal information and improper use of them violates user's privacy. Therefore, the need for protecting such personal information is eminent. On the other hand, the need for providing accurate answers to user requests regarding mapping information and directions based on personalized settings is also a requirement on mobile computing applications that can be considered as a means of measuring the quality of the offered services (Mokbel et al. 2006).

It is obvious that when it comes to personalized, location aware services over mobile architectures, the pin-pointing of the user's position, together with accurate information regarding user preferences and needs are instrumental in supplying the desired services. The server needs to know the user's whereabouts, as well as her personal preferences, so that it can provide her with the most precise, accurate, exhaustive personalized information; in other words, to be able to guarantee the quality of these services. On the other hand, communicating detailed data about the user's exact position and/or preferences undoubtedly raises privacy issues, which ought not to be neglected. Hence, there appear to be two divergent tendencies: quality and privacy, both of which are inarguably important for the end user.

Our goal is to provide a framework, in which the user will be the judge of how much each one of the two counterbalancing forces (quality and privacy) that drive the service provision process to an "equilibrium state" should contribute to the system, according to how much this weighs for her personally.

In order to go further in examining the relation between privacy and quality so as to achieve the best possible personalized result, we first need to find a way to systematically quantify privacy. According to Westin (1967), information privacy may be defined as "the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others". Up to now, a great deal of work has been carried out in the field of location privacy. Part of this work is summarized in the "related work" section. On the contrary, the issue of personal preferences privacy has not been addressed to the same extent.

In our approach, we attempt to take into consideration both privacy threats and let the user be the judge of the level of importance of each parameter.

Related work

An initial method that was proposed in order to protect user's location privacy is blurring. An indicative practical example of such techniques is given by Kido et al. (2005). They propose the transmission of several false position data (or "dummies") to the service provider, so that the latter returns an answer for each of the locations, without being able to distinguish the true one.

Hiding user's location information by mixing it with the same information provided by other users in order to achieve a certain level of "anonymity" (Sweeney 2002; Bayardo and Agrawal 2005) is another technique which has been widely proposed. Gruteser and Grunwald (2003) on the other hand, propose an algorithm

that deploys region quad-tree cloaking, in an endeavor to achieve k-anonymity in terms of spatial and temporal terms. The algorithm uses a recursive subdivision of location data around the user, up to the point that the selected quadrant includes a number of users below k-min, and then uses the previous level as the cloaking region. In this approach, the region required to achieve an adequate level of anonymity may be quite large. This is because the number of system users that have similar preferences in order to be included in a k-anonymous area may be scattered around a very large region. Beresford and Stajano (2003) introduce the concept of mix zones in order to model the spatiotemporal issue of anonymity. However, this approach again presupposes the existence of a sufficient user population, to whom the service is being supplied.

Gedik and Liu (2005) propose the CliqueCloak cloaking algorithm. The algorithm operates on a per-user basis, taking into account personal privacy settings and QoS requirements in terms of cloaking latency and cloaking region. All requests from users are anonymized through the use of an undirected graph consisting of user requests that have not been anonymized yet.

Mokbel et al (2006) follow the k-anonymity model, trying to provide a framework that meets the demands for privacy and quality in database queries for location based services. Their model is based on the use of a location anonymizer and a privacy-aware query processor, for hiding information about a user's location.

Other techniques use abstraction layers in reporting information about the user's location. This abstraction is equivalent to the increase of entropy with regard to positioning or representational accuracy (Jiang and Landay 2002). The higher the level of entropy is the better protection of user's privacy is achieved. Jiang and Landay (2002) however do not go further in proposing a way in which this measure could be useful in service provision.

Moreover, the concept of entropy is employed by Jiang et al. (2007). However, the framework proposed by Jiang et al is based on the use of a probability distribution calculated by an attacker for all users of a wireless LAN based on her observations about the users' locations.

Finally, Kapadia et al. (2007) present a model that allows users to deploy "virtual walls" which enable them to control the privacy of their digital footprints. It is obvious that all the aforementioned work cited in this section merely focuses on the issue of location privacy from various points of view. The work presented in this paper can therefore be considered as an extension to the currently existing location privacy protection research, as it introduces the personal preferences privacy aspect in context aware LBS.

Quantifying privacy

Using entropy to measure the level of privacy

Trying to find the appropriate method to uniformly address the level of privacy for each aspect of a user's personal profile, we use entropy (H) as the measurement of diversity, and therefore difficulty in identifying a user's personal preferences, parameters and whereabouts. Shannon (1948) developed the mathematical theory of

communication in an endeavor to quantify the uncertainty (or randomness) of an information source. Taking into consideration the following points:

- our aim is to quantify the uncertainty (of the user's geographical position and of her personal preferences, as well),
- some of the properties which Shannon sought in a measure of information uncertainty have a lot in common with the properties of a desired metric in our case: (i) with equally likely options there is more choice, or uncertainty, when there are more possible options (monotonicity) and (ii) if a set is defined as the combination of several disjoint subsets, H for the new set should be the weighted sum of the individual values of H for the subsets; that is, for a set S_c composed of two subsets, S_a and S_b: $H(S_c) = H(a, b) + aH(S_a) + bH(S_b)$, where $a=|S_a|/|S_c|$ and $b=|S_b|/|S_c|$, and $H(a,b)$ is the entropy of a system containing two types of elements with proportion (or probability) a and b respectively (recursiveness),
- information entropy is used in many related fields, such as: in ecology, in order to determine species' diversity (Lurie and Wagensberg 1980), in sociology, to examine societal evolution (Bailey 1990), in taxonomy, to evaluate classification methodologies (Jardine and Sibson 1971), in robotics, to measure multiagent system diversity (Balch 2000), it can be deduced that adopting Shannon's information entropy model can serve as the basis for our calculations. In the following sections, we will present the necessary mathematical framework on which our methodology is based.

The information entropy of a system X according to Shannon is:

$$H(X) = -K \sum_{j=1}^{M} p_j \log(p_j) \tag{1}$$

where:

- M is the number of possible values/states/subsets of the system under examination,
- p_j is the probability of a variable of the system taking the j^{th} value (being in the j^{th} state / belonging to the j^{th} subset, etc), where $\sum_{j=1}^{M} p_j = 1$, and
- K is a constant corresponding to a choice of measurement units (Shannon sets $K=1$ and we also adopt it in the paper).

We will apply this entropy model on both aspects which simultaneously constitute cornerstones of any personalized service for mobile devices and privacy threats for the user: geographical positioning and personal preferences. By doing so, these two seemingly different aspects will be placed on a common basis with regard to their influence on privacy of personal information. We will then be able to uniformly address the impacts of these two aspects on user's privacy, while letting her control the extent of privacy threat to which she is being exposed.

Location tracking entropy

Any mobile, location based services user is bound to be making use of a map system, such as Windows Live Local (a.k.a. Microsoft® Virtual Earth), Google

Maps™, Yahoo! Maps, etc. Most of these systems are based on the use of the Mercator projection. The most important characteristic of these systems, however, lies in their inherent support for different levels of abstraction and accuracy. In systems which are organized in an hierarchical structure using tiles, such as Virtual Earth and Google Maps™, there is a number of levels of detail (around 20—23). On each level, the world map is split into tiles of a specific pixel size, and every tile has an ID, e.g. on level 1, there are four tiles, named for example 0, 1, 2, 3. On the next level, the area of the map that corresponds to each tile is further split into a number of new tiles (again, usually four); e.g. level 1 tile 2 is split in four level 2 tiles: 20, 21, 22, and 23. This process continues for all levels, resulting in an hierarchical structure (in our case a quad-tree hierarchical structure), where positioning accuracy increases (and, consequently, privacy decreases) as the level of detail is enhanced. For instance, if a user is in an area represented by the tile 12202102331213 (14th level of detail), her location could also be reported as tile 1220210233121 (13th level of detail), and as tile 122021023312 (12th level of detail), with accuracy decreasing and vagueness increasing every time. It should be noted that in our framework, the selection of the geographical location system is by all means independent of the methodology presented for balancing privacy/quality in service provision, as long as the former uses a hierarchical data structure such as for example a region quad-tree (Samet 1984) to organize a geographical object space.

We will now apply Shannon's entropy model (1) to the case of a geographical location system based on an hierarchical structure. We will assume that the Global Positioning System (GPS) is used to track a user's location, since this is the most accurate positioning system that can be used globally. Let s_{GPS} be the area reported as the user's location (in an ideal situation this would be around one square meter, as this is the area that is occupied by a person (Neufville 2007), but since the issue of reporting accuracy of the GPS is also involved here, this is larger (the exact value will be discussed later). Let A be the actual area corresponding to the map region that the user reports to an LBS query server as her location, that is, in the above described case, the actual area corresponding to one tile of a particular level. In this case, M equals the number of the different possible distinct positions, which could correspond to the user's exact location on the area. Therefore, $M=\lceil A / s_{GPS} \rceil$. Consequently, p_j is the probability of each one of the reportable (by the GPS) locations on the current map being the exact user's position and will, thus, equal s_{GPS}/A. One could argue here that using the full area of a tile as A is not correct (i.e. in a city, a user is most probably located on the areas of a tile that correspond to streets). However, though in many cases this is true, there are often exceptions (e.g. users located near windows, on roofs, or in gardens). Furthermore, in rural areas, the possibility of a user being located anywhere in a tile increases. For this, as a general case we consider that a user has equal possibilities to be located anywhere in a tile area.

Therefore, the entropy regarding user positioning is:

$$H = -\sum_{j=1}^{M} p_j \log(p_j) = -\sum_{j=1}^{A/s_{GPS}} \frac{s_{GPS}}{A} \log\left(\frac{s_{GPS}}{A}\right) = -\log\left(\frac{s_{GPS}}{A}\right) \qquad (2)$$

The value of A depends on the detail level (in other words, the level of accuracy), and therefore, the value of H changes when tiles of different level of accuracy in the hierarchical structure (e.g. nodes of different depth in the quad-tree) are chosen by the user to report her location. A specific example of geographical entropy calculation is provided later in this paper.

Personal preferences entropy

Let us now move to the discussion of the second aspect of the privacy / quality equilibrium desired in personalized services: the user's personal preferences. The reporting of the user's personal preferences is obviously of vital importance for the efficient provision of the personalized service. These preferences may include numerous parameters, e.g. favorite film genres is essential information for a service that makes recommendations when the user wants to go to the cinema; preferred music styles and cuisine constitute two additional prominent examples. Nevertheless, preferences, again, can be expressed more specifically, leading to more accurate and better quality results, while at the same time sacrificing privacy, or they can be expressed in a more vague way, thus ensuring better privacy but also giving less accurate results. For instance, when it comes to movie types, a person could make an extremely specific statement of preference, such as "films about airplane disasters", or a much more general one, like "action films".

In order to apply our model to the preferences parameters (film types, music styles, cuisine, sports, etc), we need to group the possible preferences for each parameter in different levels of precision. There is not only one way to do this. It can be done based merely on logic, knowledge or experience. However, a more systematic method such as a clustering algorithm would probably constitute a better solution. Karamolegkos et al. (2007) proposed a framework so as to create a number of social groups among users, based on their profiles, which consist in n distinct keywords representing each user's preferences. K-means and spectral clustering algorithms are compared and the latter proves to be more efficient in the case. Given the request that every user choose n keywords for one specific parameter (for example, choose n different film types she enjoys watching) and use a spectral clustering algorithm to create groups, every step of the algorithm will give a number of clusters containing "related" film types based on real user's data. As the algorithm proceeds to a next step, the number of clusters will increase and the cardinality of the clusters will fall leading towards organization of user preferences in levels of different accuracy. In both described cases (organization based on knowledge and experience or based on keyword guided clustering), the goal is to organize the possible user preferences in levels of different accuracy, while it is out of the scope of the paper to discuss the most appropriate of these techniques. Applying the entropy model (1) now is quite simple by making use of the aforementioned recursive property:

Suppose a cluster C_j^l (the j^{th} cluster of level l) with cardinality N_j^l is broken down on the next level (l+1) and its N_j^l elements are "spread" into three distinct cluster of level l+1: C_x^{l+1}, C_y^{l+1}, and C_z^{l+1}. Specifically, n_x out of its N_j^l elements become elements of cluster C_x^{l+1}, n_y become elements of cluster C_y^{l+1}, and n_z become

elements of cluster C_z^{l+1} $\left(n_x + n_y + n_z = N_j^l \equiv n\right)$. The entropy H_j^l of cluster C_j^l is then calculated (Shannon 1948; Balch 2000) as:

$$H_j^l = H\left(\frac{n_x}{n}, \frac{n_y}{n}, \frac{n_z}{n}\right) + \frac{n_x}{n} H_x^{l+1} + \frac{n_y}{n} H_y^{l+1} + \frac{n_z}{n} H_z^{l+1} \quad (3)$$

It should be noted here that in the special case of a purely hierarchical structure, that is, every cluster on level l+1 is a subset of a cluster of level l (this would be the case if an hierarchical clustering algorithm was used), n_x equals the cardinality of cluster C_x^{l+1}, n_y equals the cardinality of cluster C_y^{l+1} and n_z equals the cardinality of cluster C_z^{l+1}.

Building the LBS query

Having now applied the entropy model in both geographical location and personal preferences, thus having quantified privacy related to both of these aspects, we are now able to handle the users' demands for privacy in a tangible way. Suppose that g_{max} levels of detail in the hierarchical mapping system are supported for a given service, with g_{max} being the level of greatest detail, and that a specific personal preferences parameter such as film types which is important to the specific service (e.g. "find a cinema in my area that shows a movie of a genre I like"), has undergone a grouping process (i.e. using a clustering algorithm), resulting in an organization of the possible preferences in p_{max} different levels, with p_{max} being the level of greatest accuracy.

The aim is to offer to the user personalized location based services in such a way that the level of abstraction/privacy of results is determined by the privacy settings profile of the user herself. On the other hand, accuracy in the provision of the service should not be compromised. Furthermore, the user should be given the possibility to select her privacy settings through a very simple user interface, hiding any complexity from her. The interface that we propose is the image of a quadrant where the horizontal axis represents location privacy and the vertical axis represents privacy in personal preferences/motivation. Both axes' values use a scale of 10, where 0 denotes minimum privacy (Fig. 1). The first time the user asks for a particular service, she is required to select a specific point on the aforementioned quadrant, which is representative of her privacy preferences as far as the specific service is concerned. Since the first time user will probably not be familiar with the mapping between the value on the axis and the actual level of detail at which the geographic frame will be reported, in the beginning some examples are presented to the user: e.g. a value of 2 with respect to location privacy concern corresponds to a reported area of a couple of blocks surrounding the user's exact location, whereas a value of 10 would mean that the user is really serious about protecting her location privacy and wishes to report a very abstract location such as the entire city. Respective examples could be provided to the user in reference to preferences privacy as well, so as to help her functionally operate the system.

Let (x_g, x_p) be the coordinates of the selected point of the quadrant. Transforming these user privacy preferences into usable values for our framework is rather simple:

$$H_g^* = H_g(g_1) \times x_g / 10 \quad (4)$$

$$H_p^* = H_p(p_1) \times x_p / 10 \quad (5)$$

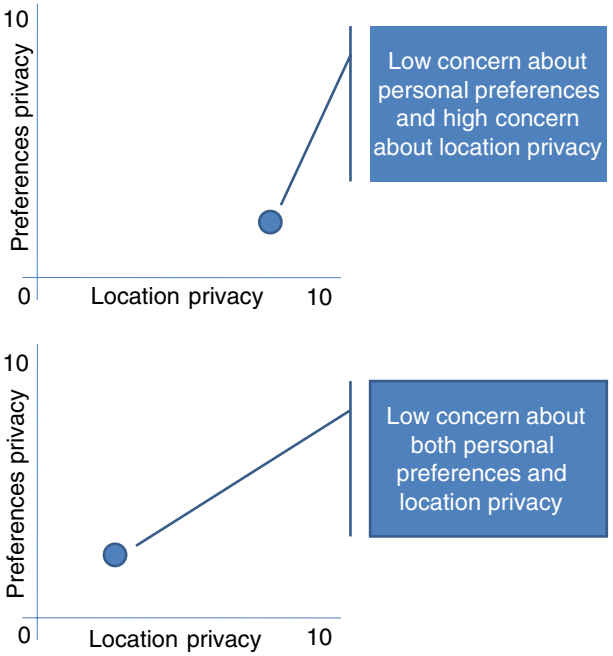

Fig. 1 User interface for privacy settings

where H_g^* and H_p^* are the entropy values that "ideally" represent the privacy settings of the user for the specific service as far as geographical location and personal preferences respectively are concerned. Since there is a certain number of distinct privacy/entropy levels both for location and for preferences too, it is highly unlikely that H_g^* or H_p^* equal any of the existing $H_g(g_i)$ or $H_p(p_j)$ respectively. We therefore need to determine the closest valid value. We are thus looking for the levels g_u and p_u for which:

$$\left|H_g^* - H_g(g_u)\right| = min_i \left|H_g^* - H_g(g_i)\right| \tag{6}$$

and

$$\left|H_p^* - H_p(p_u)\right| = min_j \left|H_p^* - H_p\left(p_j\right)\right|. \tag{7}$$

The levels g_u and p_u that satisfy the above conditions (6) and (7) are the privacy/accuracy levels that best describe the user's attitude towards the particular service and therefore constitute the privacy/accuracy levels at which the user will report her data to the service provider.

This approach could lead to a value $H_g(g_u)$ or $H_p(p_u)$ which could be smaller than the "ideal" H_g^* or H_p^* and some might argue that in this way the user is offered "less" privacy than she asked for. However, since privacy is not accurately quantified, there is no point in setting so strict limits. Nevertheless, demanding that $H_g(g_u)$ and $H_p(p_u)$ be greater than or equal to H_g^* and H_p^* respectively so as to ensure that the user's privacy settings are totally respected from a mathematically strict point of view would also constitute a perfectly adequate solution.

Explanation through a practical example

To assist the reader in understanding the proposed methodology, an actual example incorporating location awareness and personalization aspects is used. In the example, the geographical system used is Microsoft Virtual Earth, and the personalization aspect is the likes of the user regarding film types.

Virtual Earth splits the world map into square tiles with a size of 256×256 pixels. Currently 23 levels of detail are supported. At level 1, the world map is 512×512 pixels and consists of 4 tiles (tiles 0, 1, 2 and 3). Figure 2 graphically depicts the tile system in Virtual Earth.

We will start from the greatest level of detail, which corresponds to a level 23 tile. We will gradually "zoom out", that is, go to a lower level of detail in each step and calculate the entropy in every step.

The actual length u_i of the edge (in meters) of the area depicted by a tile at level i (i=1,2,..23) equals the product of ground resolution (the distance on the ground that is represented by a single pixel in the map) of that particular level, multiplied by 256 (pixels per tile). The lengths of the edge for tiles of two consecutive levels are correlated by the ratio (as can be easily deduced from the aforementioned description of the hierarchical mapping structure):

$$\frac{u_{i-1}}{u_i} = 2 \tag{8}$$

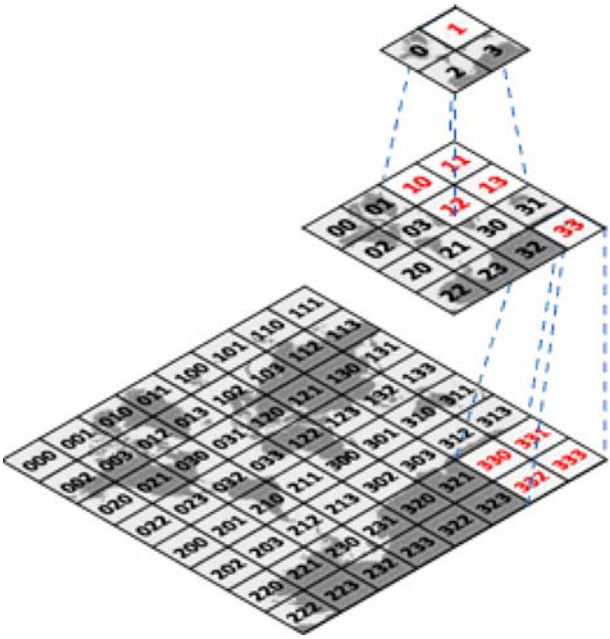

Fig. 2 Representation of tiles organization in Virtual Earth

As for the actual area α_i (in m^2) represented by one tile at level i, this equals u_i^2, and for two successive levels:

$$\frac{\alpha_{i-1}}{\alpha_i} = 4 \tag{9}$$

In order to calculate entropy, A (the actual area corresponding to the region that the user reports as her location) and s_{GPS} (the area reported as the user's location by the GPS device) need to be calculated. We have:

$$A_i = \alpha_i \tag{10}$$

Taking into account that α_{23} is the smallest possible area (level 23 is the level of greatest detail), (9) gives

$$\alpha_i = 4^{23-i} * \alpha_{23}. \tag{11}$$

Now (10) and (11) give:

$$A_i = 4^{23-i} * \alpha_{23}. \tag{12}$$

With regard to s_{GPS}, the average accuracy of GPS positioning needs to be determined. Since commercial devices advertise accuracy up to even 1 m, while practical tests indicate that in practice accuracy of devices is much less, we have tried to make an assumption on average accuracy that is as close to real life situations. In this, the results of Wing et al. (2005), in which they tested the accuracy and reliability of consumer-grade GPS receivers in a variety of landscape settings, have been used. According to these tests, the best accuracy, achieved in open sky conditions and in non urban areas was measured at an average of 5 m. As the area covered by a tile in the 23rd level (greatest detail level) of the Virtual Earth System is 22.9 m^2 (0.0187meters/pixel*256pixels)^2 (Virtual Earth Tile System), we have decided to use the area of the tile as the area of average GPS accuracy (being the equivalent of a near 5 m side square). We will therefore consider that

$$s_{GPS} = \alpha_{23}. \tag{13}$$

Taking (12) and (13) into account, (2) gives the following formula for the geographical entropy $H_g(i)$ on level l (l=1,2,..23):

$$H(g_i) \equiv H_g(i) = -\left(\log\left(\frac{a_{23}}{4^{23-i} * a_{23}}\right)\right) = (23 - i)\log 4 \tag{14}$$

Therefore, every time we enlarge (by going up one level at the Virtual Earth System tree) the area which defines the user's whereabouts, the additional entropy equals to log4, that is

$$\Delta H_g = H_g(i) - H_g(i+1) = log4 = 1.386 \tag{15}$$

At this point it should be noted that Shannon uses base2 logarithm in information entropy, because of dealing with binary data. For our study, we have chosen to use the natural (neperian) base e logarithm since it is the one appearing most often in physical processes (that is, log≡loge≡ln). The choice of base, nevertheless, does not influence our methodology.

In respect to the film preferences, we have created an organized (hierarchical) structure which includes a number of film genres and sub-genres, based on a specialized site for films (Dirks 2007). The structure is definitely not exhaustive and merely serves as an example of how the proposed methodology can be put to practice. As has been already pointed out, other methods for providing different levels of clustering for the film types such as spectral clustering may be used (Karamolegkos et al. 2007). Figure 3 depicts the hierarchical organization of film types used in our example.

Here $p_{max}=5$. We will calculate the entropy based on (3) for every node of the tree.

To begin with, $H=0$ for all the leaves since in these cases, the exact user preference is identified by the corresponding film type. For the rest, we have:

$H(\text{“Disaster”})=H(1/3,\ 1/3,\ 1/3)+1/3*H(\text{“Aircraft”})+1/3*H(\text{“Ship Wrecks”})+1/3*H(\text{“Natural Disasters”})=\log 3+0+0+0=1.099$

$H(\text{“Sports”})=H(1/4,\ 1/4,\ 1/4,\ 1/4)+1/4*H(\text{“Boxing”})+1/4*H(\text{“Martial Arts”})+1/4*H(\text{“Football”})+1/4*H(\text{“Motorcycle”})=1.386$

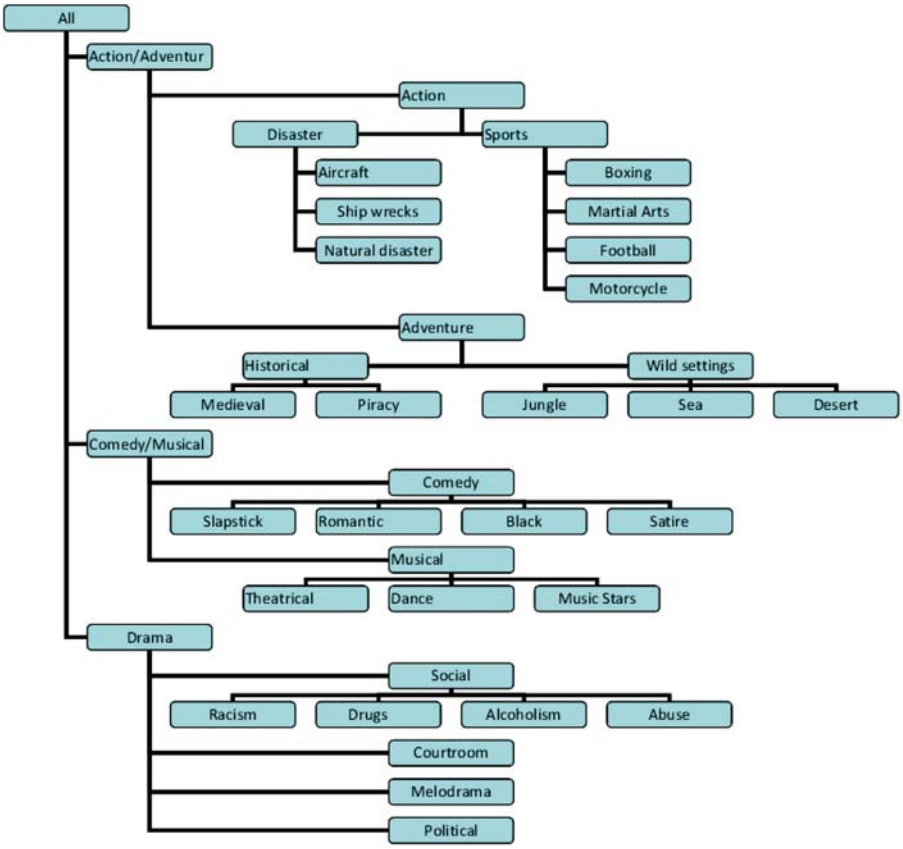

Fig. 3 Film genres hierarchical clustering

H("Historical Adventures")=0.693
H("Wild Settings")=1.099
H("Comedy")=1.386
H("Musical")=1.099
H("Social")=1.386
H("Action")=H(3/7, 4/7)+3/7* H("Disaster")+4/7*H("Sports")=1.946 (2)
H("Adventure")=H(2/5, 3/5)+2/5* H("Historical Adventures")+3/5*H("Wild Settings")=1.610
H("Action/Adventure")=H(7/12, 5/12)+7/12* H("Action")+5/12*H("Adventure")=2.485
H("Comedy/Musical")=H(4/7, 3/7)+4/7* H("Comedy")+3/7*H("Musical")=1.946
H("Drama")=H(4/7, 1/7, 1/7, 1/7)+4/7*H("Social")+1/7*H("Courtroom") + 1/7*H("Melodrama") + 1/7*H("Political")=1.513
H("All Movies")=H(12/26, 7/26, 7/26)+12/26*H("Action/Adventure")+ 7/26*H("Comedy/Musical") + 7/26*H("Drama")=3.142

In the above calculations, we consider the contribution ratios to equal the proportion of cardinalities of each subset to the cardinality of the new greater set.

Having calculated all the possible values of entropy regarding location and personal (film related) preferences reporting and supposing that the user's favorite film type is "natural disasters", we can determine the values of entropy H_p for all the clusters of different level which characterize the user:

$$H_p(p_1) = 3.142, H_p(p_2) = 2.485, H_p(p_3) = 1.946, H_p(p_4) = 1.099, H_p(p_5) = 0.$$

Regarding location hierarchy, the level of smallest possible detail g_1 will certainly not be the 1st level of Virtual Earth's system hierarchy, but will correspond to an area large enough to cover a demanding user's needs for location privacy, in which though the provision of the particular service makes sense. In our example, we consider that an appropriate area could be a relatively small arrondissement of Paris or a quarter of the area of a borough of New York City. This means that $H_g(g_1)$= H_g(level 15)=11.090, $H_g(g_2)$=9.704, $H_g(g_3)$=8.318, $H_g(g_4)$=6.932, $H_g(g_5)$=5.546, $H_g(g_6)$=4.159, $H_g(g_7)$=2.773, $H_g(g_8)$=1.386, and $H_g(g_9)$=0.

Let's now suppose that the user's concern about location privacy is rather high whereas her concern about preferences privacy is low. The user thus selects a point on the quadrant which corresponds to coordinates e.g. (7.2, 2.6).

(4) and (5) give:

$$H_g^* = 11.090 \times 7.2/10 = 7.985$$

$$H_p^* = 3.142 \times 2.6/10 = 0.817.$$

Applying now conditions (6) and (7) lead to: $g_u=g_3$ and $p_u=p_4$. The user's query to the LBS server will therefore consist in asking the server to find the cinemas that play movies about disasters in her area that will be reported at a detail of Virtual Earth's level 17 (out of 23).

If the number of returned results is not satisfactory, the user could be given the possibility to relax one or both of the parameters and repeat the query so that she receives a greater number of results.

Framework deployment—Future enhancements

In this paper, we have provided a methodology for addressing personal privacy and LBS queries accuracy through a simple, easy to present and utilize manner. The methodology has been based on the use of entropy for calculating the most appropriate level of privacy/accuracy according to user needs.

The methodology is independent from any server dependent architectures for providing privacy in user location and personal preferences reporting (i.e. anonymizers), while it uses a simple interface for getting the user's privacy settings and transforming them to the corresponding levels of abstraction. Furthermore, it does not require the correlation with any other user's reported data for offering privacy over the reported information, as this is achieved through a mixture of personal preferences abstraction and location blurring combined so as to provide the desired level of entropy.

Currently, the proposed methodology is used in practice for the provision of personalized, context aware mobile services through the PLASMA platform (2007). The platform is based on the use of PocketPC mobile devices with mobile telephone capabilities, and is using Microsoft Virtual Earth system for mapping and location tracking purposes. Figure 4 depicts the architecture of the platform.

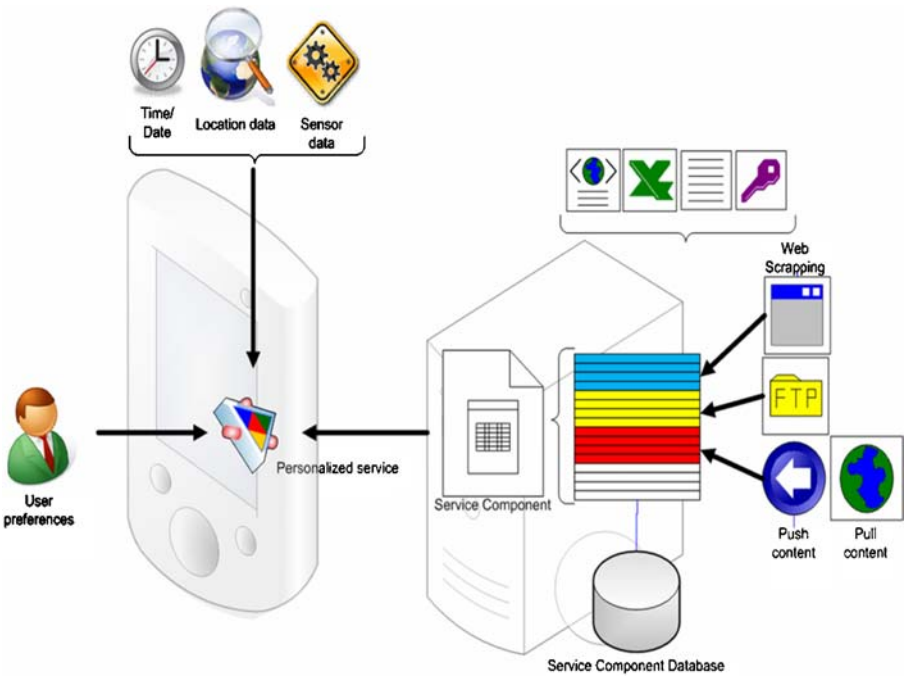

Fig. 4 PLASMA platform architecture

The platform supports the personalized provision of context aware services, using location, situation, time and personal settings related information to user initiated data queries to an LBS server. In order to be able to form and deploy queries, the mobile device application, using the service API, can identify all available information available to the LBS server. Once information about a specific point of interest needs to be transformed into a query, personalized user settings are used in order to determine the level of ambiguity the user desires to be used in the query to the system. Once this has been determined, then the query, incorporating personal information is addressed to the LBS server, while hiding the exact user's position and specific preference.

The platform is able to guarantee the protection of privacy of user data, since the reported information about the user location and personal preferences is reduced in accuracy using techniques deployed in the user's mobile device. Furthermore, classification of geographical information and preferences assists in decreasing the query search times, as the points of interest information is also characterized by the same classification fields, assisting in the indexing of information. An example of this is the storage of points of interest location information, which apart from the standard latitude and longitude fields, also includes the corresponding 23rd level tile quadnumber (the tile of highest accuracy), which is a 23 digit number. Once the user addresses a query to the system, the corresponding quadnumber of the tile reported as that including the user's position (having length equal or less than 23 digits), is used to produce the query results by matching the quadkey (or quadkey parts) of database entries to that of the query tile quadkey.

References

Bailey K. Social Entropy Theory. Albany: State University of New York; 1990.

Balch T. Hierarchical social entropy: an information theoretic measure of robot group diversity. Auton Robots. 2000;8(3):209–38. Springer Netherlands.

Bayardo R. J., & Agrawal, R. Data privacy through optimal k-anonymization. *Proceedings of ICDE 2005*, 217–228. 2005

Beresford, A. R., & Stajano, F. Location privacy in pervasive computing. *IEEE Pervasive Computing, 2, 1.* 2003

Dirks, T. Film Genres web site. Retrieved on December 20th, 2007 from the World Wide Web: 2007. http://www.filmsite.org/genres.html.

Gedik, B., & Liu, L. Location privacy in mobile systems: a personalized anonymization model. 2005. *Proceedings of ICDCS 2005,* 620–629.

Gruteser, M., & Grunwald, D. Anonymous usage of location-based services through spatial and temporal cloaking. *Proceedings of the 1st international conference on Mobile systems, applications and services*, 31–42. 2003

Jardine N, Sibson R. Mathematical Taxonomy. New York: Wiley; 1971.

Jiang, T., Wang, H., & Hu, Y. C. Preserving location privacy in wireless LANs. *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services,* 246–257. 2007

Jiang X, Landay JA. Modeling privacy control in context-aware systems. IEEE Pervasive Computing. 2002;1(3):59–63.

Kapadia A, Henderson T, Fielding J, Kotz D. Virtual walls: protecting digital privacy in pervasive environments. Pervasive Computing. 2007;4480:162–79. Springer Berlin/Heidelberg.

Karamolegkos, P. N., Patrikakis, Ch. Z., Doulamis, N. D., & Tragos, E. User-profile based communities assessment using clustering methods. Presented at the18th Annual IEEE International PIMRC 07. 2007

Kido, H., Yanagisawa, Y., & Satoh T. Protection of location privacy using dummies for location-based services. 2005. *Proceedings of the 21st International Conference on Data Engineering Workshops*.

Lurie, D., & Wagensberg, J. Information theory and ecological diversity. *Systems Far from Equilibrium*, Springer-Verlag, 290–303. 1980

Mokbel, M. F., Chow, C., & Aref, W.G. The new Casper: query processing for location services without compromising privacy. *Proceedings of the 32nd International Conference on VLDB*, 763–774. 2006

Neufville, R. Defining Capacity of Airport Passenger Buildings, Lecture notes for the course Airport Systems Planning, Design, and Management, Massachusetts Institute of Technology. Retrieved on December 1st, 2007 from the World Wide Web: 2007. http://ardent.mit.edu/airports/ASP_current_lectures/ASP%2004/Defining_Capacity04.pdf.

PLASMA Project web page. URL: http://www.telecom.ntua.gr/~bpatr/staticcontent/PLASMA.php

Samet H. The quadtree and related hierarchical data structures. ACM Comput Surv. 1984;16(2):187–260.

Shannon CE. The mathematical theory of communication. The Bell System Technical Journal. 1948;27 (379–423):623–56.

Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002;10(5):571–88.

Virtual Earth Tile System. Retrieved on March 15th, 2008 from the World Wide Web: http://msdn2.microsoft.com/en-us/library/bb259689.aspx.

Westin AF. Privacy and Freedom. New York: Atheneum; 1967.

Wing M, Eklund A, Kellogg L. Consumer-grade global positioning system (GPS) accuracy and reliability. J For. 2005;13(4):169–73.