User data distributed on the social web: how to identify users on different social systems and collecting data about them

Francesca Carmagnola
Dipartimento di Informatica
Corso Svizzera, 184
Torino, Italia
carmagnola@di.unito.it

Francesco Osborne
Dipartimento di Informatica
Corso Svizzera, 184
Torino, Italia
francesco.osborne@gmail.com

Ilaria Torre
Dipartimento di Informatica,
Sistemistica e Telematica
Via all'Opera Pia, 13
Genova, Italia
Ilaria.torre@unige.it

Abstract

This paper presents an approach to uniquely identify users and to retrieve their data distributed in profiles stored in different systems. The objective is exploiting the public user data available in the Web and especially in social networks. The approach does not require the implementation of specific protocols and the provision of authentication data. The evaluation provides good results that encourage us in carrying on the extension of the project. The extension we are working on is aimed at aggregating, using heuristic techniques, the data stored in the retrieved profiles and at inferring new data about the user.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces And Presentation**]: Hypertext/ Hypermedia – *User issues.*

General Terms

Algorithms, Human Factors.

Keywords

User Model Interoperability, Social Web, User Identification.

1. Introduction

In the current Web, most of the websites collect data about users to provide different services. This implies that a lot of data on a specific user (e.g. his/her preferences, interests, activities, etc.) are scattered over many systems on the Web and the user profile is inherently distributed. This phenomenon grew with the diffusion of Social Web and social systems, which store lots of data, often public, about users.[1]

Given this context, an interesting opportunity is to develop environments that effectively enable systems to benefit from the distributed knowledge about users, favoring the exchange and reuse of user data for adaptation purposes [1,7]. This is known as "cross-system personalization". Many researches have explored cross-system personalization, focusing on issues such as the communication among distributed systems [2], the heterogeneity management of distributed user and domain data [7, 1], the protection of user's privacy during the interoperability process [9] and the identification of the user whose data are exchanged among systems [3, 5, 13]. This last issue is a starting requirement to enable cross-system personalization since it means, for the different systems partaking into the user data exchange process, to *discover if* they are *referring to the same user*. The problem is particularly critical in the context of the Social Web, where users are often identified by nicknames chosen by themselves.

This paper presents an approach and an algorithm to support adaptive systems to uniquely identify users in different social systems. The approach we present does not require the provision of authentication data and user identification is performed by using the public data available on the Web.

In a cross-system personalization perspective, the contribution of this approach is creating new opportunities for gathering more user data to reach better adaptation results.

To show how the algorithm can be exploited, let us consider the scenario of an adaptive system that knows a small set of data about a user, like the nickname (s)he uses in that system, for example *billsmith*, the gender, for example male, and the country, for example US, California. Collecting further data about *billsmith* would allow the adaptive system to extend the profile of the user and consequently to improve the adaptation result. But how can the system be aware of other systems that collect data about *billsmith* and how can the user be identified in these other systems?

The algorithm we developed can support the adaptive systems in discovering whether *billsmith* has a profile in other systems on the Web. The bigger the set of crawled systems, and the number of users of these systems, the higher the probability of discovering

HetRec '10, September 26, 2010, Barcelona, Spain.

[1] In the so called *Social Web*, people interact one with each other, sharing knowledge and interests. In this context, s*ocial system* allow to create relations among users.

billsmith somewhere. In the public Web, crawling the data of big social networks offers the highest probability of finding *billsmith*. Moreover, social networks offer also the advantage that lots of profile data are public and up-to-date.

The algorithm receives in input the small set of known data about *billsmith* and returns a set of profiles associated to an *identification probability* that represents the chance they belong to the searched user. Moreover, it returns a set of *probable attributes*, like *billsmith*'s age, city, interests, profession, etc. obtained from the *billsmith*'s data included in the retrieved profiles. Since each attribute is associated to an identification probability, the adaptive system can decide, for each one, to acquire it or not, according to its polices.

To test our algorithm, we ran it on a set of real-world social systems. The evaluation showed good ability of the algorithm to identify users but showed also some limits. Currently we are working on the revision of the algorithm and its extension. In particular, we are working on the aggregation of user profiles data using heuristics rules and using ontologies.

The paper is structured as follows: Section 2 offers an overview of the related works, Section 3 describes the algorithm we propose for user identification, Sec. 4 provides the results obtained by running the algorithm on a set of real world social systems and Sec. 5 concludes and points at future research directions.

2. Related works

Cross-system personalization is growing in importance and diffusion. Already in 2001, Kobsa et al [10] observed that adaptive systems could use cross-system personalization to speed-up the process of user model creation. This approach is also useful to the user, avoiding him/her to repeat the boring process of filling in similar forms for different services [14].

Systems partaking in the process of user data exchange can enrich their own profiles or can enrich repositories of user profiles [7]. In [1] the authors propose a *framework* to import and integrate user data from other recommender systems. Even though several obstacles exist to user data integration, such as different representation formats, different contexts of acquisition, privacy risks, etc., user models mediation can nevertheless be useful as a support to the personalization service.

The use of standards such as SKOS, and in general standards related to the Semantic Web, can make easier sharing, exchanging and integrating data coming from different systems, as in Morpho framework [11]. In our approach, this could be useful especially in the last part of the process, when results are returned to be used by different applications. Instead, using it in the phase of search and comparison between profiles reduces the performances of the algorithm.

Regarding the user identification issue, currently, only a few solutions have been suggested to support user-adaptive systems in discovering if they are referring to the same user. Sometimes, this issue has been considered as a starting assumption, without proposing specific solutions to face it; other times, identification was not a problem since the systems used a common identification mechanism. For example, in [12] user identification is ensured by making the user hold a passport with his/her data to be provided to the personalization systems he/she interacts with.

In [15], the issue of user identification is managed by using OpenID (openid.net), an identification system developed in the spirit of the Web 2.0, which provides an authentication method to allow users to log on different services with the same digital identity. Other recent initiatives, like OpenSocial (opensocial.org), Connect[2] and MySpaceID[3], add also the possibility of profile portability across systems on the Web. All the solutions mentioned above require the application to join a framework or support proper protocols to allow the cross-systems user identification. Identifying the users across systems independently of the protocols supported by each system and independently of the authentication data supplied by the user is, so far, a challenge.

The approach presented in this paper aims at uniquely identifying users and retrieving their data distributed on the profiles stored on different systems. It does not require the implementation of specific protocols and the provision of authentication data, even if, compared to the solutions above, it can use only the public data available on the Web. From this point of view, the work of Szomszor et al. [13] is very close to our project. They perform a cross-folksonomies profiling based on collecting all the tags used by a user on different Social Systems. For the automatic identification of users on different systems, they use the Google Social Graph API[4], which includes a matching technique for cross-profiling based on the user homepage. In our framework we compare sets of attributes relative to the user as found in different social systems. The choice of using a set of user attributes and not only the homepage is aimed to identify also users who do not have a personal homepage or do not want to publicly display this information. Moreover, in our approach, user attributes are used not only for user identification, but also for obtaining an aggregated user profile.

An interesting project regarding the aggregation of data from social networks is SONAR [6], an API for gathering and sharing social network information. In particular it is focused on identifying and exploiting relationships between individuals, who may be linked in several ways, as co-authors of papers, file co-sharers, blog comments, etc. In this project, however, the issue of user identification cross-system is not specifically addressed.

3. Cross-Systems Identity Discovery

In the Social Web, users typically interact with different social systems (in the following SSs), having different accounts and thus different identities and profiles. Some of the data are private, but a big amount of these data are public (users make them public to be searched) and can be accessed by other people and systems which can reuse them (with the limitation that some data cannot be stored on third systems). Moreover, some of the data in these profiles are *replicated*, while other data do not overlap and provide *new information* about the user. As we will describe in this section, our approach exploits *replicated public data* to perform user identification and the other data to *enrich* the user profile. Obviously, if the crawled SSs are popular and numerous, the probability of finding one or more profiles of the searched user increases.

As in other related works (e.g. [11]), to have a SS crawled, a specific parser has to be developed.

[2] http://developers.facebook.com/connect.php

[3] http://wiki.developer.myspace.com/index.php?title=Category:MySpaceID

[4] http://code.google.com/intl/it/apis/socialgraph/

Given a set of *input attributes* about the user to be searched, the SSs' parsers search for user profiles that match the request. The result is passed to the *User Identification process* (Sec. 3.1), which calculates a *score* for each profile. Subsequently, for each retrieved profile, the *Identification Probability technique* calculates, starting from this score, a percentage value that we call IdP (Identification Probability) and then calculates the probability that the new discovered user attributes actually belong to the searched user (Sec. 3.2).

3.1 User Identification process

As mentioned above, the algorithm requires some initial attributes about the user to set up the search. We refer to this set of initial attributes as *input profile* `Pi`, while we will refer to the *retrieved profiles* as `Pr`. The search process starts by looking for the given nickname included in the input profile and its possible variations, or looking for the full name, if included in the input profile. Once the user profiles, with matching nickname/full name, have been retrieved, the other initial attributes are used to compute a score of this match.

In assessing this score, we consider the user identity as a collection of attributes [16]. The notion of identity is extremely important to disambiguate individuals. In the object-oriented programming, "identity" is defined as the set of properties of an object that allows it to be distinguished from the others. Referring to this definition, we consider the user identity as a collection of properties, or attributes, that uniquely represents a user.

Among these identity attributes, we distinguish between attributes that are strong indicators of a positive match between the profiles and those that are strong indicators of a negative match. For example, the user's homepage is a strong indicator of positive match; in fact, if its value is the same on two profiles, the probability that the two profiles belong to the same user is high. Other attributes, especially persistent user attributes, work as strong indicators to trigger profiles that cannot belong to the same user. For example, the match of gender in two profiles is not very significant, since the probability for each value is 0.5; conversely, a negative match is a strong indicator to exclude a match between such profiles. We apply this mechanism by defining a table of weights for positive and negative matches and then performing a semi-combinatorial weighed match between the attributes in `Pi` and the corresponding attributes included in the retrieved profiles `Pr`. Moreover, the final score takes into account also the *specificity* of the nickname, namely how uncommon and rare it is. We define the specificity score as a function of the *length* of the nickname (long nicknames are presumably more specific than shorter ones) and of its *rareness* in terms of combination of letters, numbers and special characters, in a sample of the system's population. For a more detailed description, refer to [4]. Notice that, given the results of the evaluation that will be presented in Section 4, we are working on slightly modifying these algorithms in order to solve some problems and in particular we are working on the formula for combining the scores from the nickname match and scores from the other attributes.

As a result of the User Identification process, for each crawled social system, the algorithm returns: **i)** the list of discovered user profiles $\mathtt{Pr}_{\mathtt{j}}$, **ii)** their score and **iii)** the set of user attributes discovered, included in the retrieved profiles.

As an example, let us consider a system that needs to identify a user given the following initial user attributes: ***nickname*:** billsmith*;* ***gender*:** *male;* ***country*:** *California*.

To simplify the example, let us assume that the algorithm has been run on two social systems only, SS.1 and SS.2, and that, based on such a query, it returns the user profiles `Pr` reported in Table 1: two profiles out of SS.1 and four profiles out of SS.2. For each profile the algorithm returns a score of the match with the input profile and the set of attributes stored in the retrieved profiles. These attributes include the user attributes provided to the algorithm as input profile and a set of other attributes, previously unknown.

Table 1. Example of results of the User Identification process, given an input profile `Pi` with the following initial user attributes: ***nickname*:** billsmith*;* ***gender*:** *male;* ***country*:** *California*

	Profiles `Pr`	**Score**	**Discovered user attributes**
SS.1	1	0.67	Nickname: billsmith; gender: male; age: 22; city: San Diego; country: California
	2	0.50	Nickname: billsmithers44; city: Sacramento; country: California
SS.2	1	0.55	Nickname: billsmith_1999; gender: male; country: California
	2	0.55	Nickname: billsmit77; gender: male; country: California
	3	0.66	Nickname: billsmith2009; age: 22; country: California
	4	0.44	Nickname: bill_s; city: Sacramento; country: California

Analyzing such attributes, we can see that some of the profiles seem more similar than others. We are working on finding relations between profiles in order to discover if couples of attributes on two different social systems could belong to the same user.

Consider the example above. In such an example, the user profile 1 in SS.1 and the user profile 3 in SS.2 share, together with all the correct values of the initial attributes, the value of the user attribute *age*, which is a discovered new attribute, since it was not provided as initial attribute. At the same time, the user profile 2 in SS.1 and the user profile 4 in SS.2 share the value of the discovered new user attribute *city*. We observe that the match of one or more new discovered attributes on profiles of two different social systems can be seen as a sort of connection between the profiles. It links the two profiles, increasing the probability they belong to the same user.

The problem is how to exploit this information coming from the relationship between similar profiles in different social systems. We intend to use this relationship to adjust and possibly increase, the score of a pair of similar profiles.

The issue originates from the consideration that the score of a retrieved profile highly depends on the number of attributes that match with those in the input profile. A profile with few matching attributes, and thus a low score, but linked to another profile, which has been granted a high score, can receive a boost from this correlation, increasing its original score to a value at most equal to that of the correlated richer profile. We are working on tuning the procedure to obtain this result.

3.2 Identification Probability calculation

In the previous step, the identification process returned the score of the retrieved profiles Pr and the new attributes discovered.

The next step of the identification process is converting the score of Pr into a percentage value (IdP) defined as a sort of probability, conditioned by the initial attributes in the input profile Pi and, slightly, by the probability of the other retrieved profiles Pr: $IdP(Pr_i|Pi,Pr)$.

Converting the score into a percentage value has the advantage of using a self-explaining measure of correctness of the retrieved profile, avoiding a user profile to be bound to a certain score threshold. The procedure we used to convert the score of Pr into a percentage is mapping the score of Pr to a percentage value on a scale 0%-100%, where 100% corresponds to the maximum score the retrieved profile Pr can obtain, given Pi, with the condition that such a maximum value is over a specified threshold. The maximum score that Pr can obtain, given Pi, is calculated by matching Pi with itself (using the algorithm in Sec. 3.1).

Moreover, given that the IdP of Pr slightly depends also on the IdP of the other profiles retrieved on the same SS, we use a variation of the conditional probability formula $P(x|y)=P(x\land y)/P(y)$, namely (favorable Pr)/(total Pr), to calculate this value, combining it with the probability as above defined.

According to the Identification Probability calculation, the scores in the example in **Table 1** are converted as reported in **Table 2**.

Table 2. Identification Probability (IdP) estimation for each profile.

	Profiles Pr	**Score**	**IdP**
SS.1	1	0.67	36.8%
	2	0.5	5.9%
SS.2	1	0.55	7.2%
	2	0.55	7.2%
	3	0.66	23,6%
	4	0.44	2.2%

Considering the **discovered attributes**, as displayed in **Table 3**, each attribute value is coupled with the list of the SSs where the attributes come from, each with a probability percentage inherited from the IdP of the profiles to which they belong (column 3). Moreover, an *average probability percentage* (column 4) for that value is provided. Currently, it is simply calculated as an average of the percentages in column 3 for each attribute. This is a rough estimation, aimed at showing the possibility to expand and enrich a given user profile, even starting from a relatively low number of initial data.

We plan to refine and improve this calculation, by introducing heuristic rules to combine the attribute values and ontologies to generalize or specialize the value of the discovered attributes.

Table 3. Discovered Attributes

Attr. discov.	**Value**	**Source (SSs)**	**Attr. Prob.**
Age	22	SS.1.1 (36.8%)	30.2%
	22	SS.2.3 (23.6%)	
City	San Diego	SS.1.1 (36.8%)	36.8%
	Sacramento	SS.1.2 (5.9%)	4%
	Sacramento	SS.2.4 (2.2%)	

It is important to notice that *the attributes that can be discovered are numerous and various*. Depending on the crawled social system, the algorithm can retrieve the user's bookmarks (e.g. Delicious, StumbleUpon, etc.), his/her interests (e.g. MySpace, Facebook, Diig, etc.), photos (e.g. Flickr), friends (e.g. Facebook), people followed (e.g. Twitter) , profession, education, connected people (e.g. Linkedin), etc.

However, the preliminary need is that the searched user is *properly identified*.

4 Experimental Evaluation

The definition and set up of the algorithm described in the previous sections were performed using a recurrent process of definition-evaluation-tuning of the algorithms. This section focuses on the final cycle of evaluation, structured in two steps:

1. given a set of input profiles, calculating the probability of identifying other profiles belonging to the same user on other social systems (Sec. 4.1)
2. measuring the accuracy (precision and recall) of the distribution probability calculated in the previous step (Sec. 4.2).

To perform the evaluation, we ran the algorithm on two popular SSs, MySpace and Flickr. This required the parsers for such systems to be built.

However, to check the accuracy of the retrieved profiles, this was not enough: we needed a real world dataset, made of user profiles linked to profiles of the same user on MySpace and Flickr. A dataset like that is available on Profilactic (*profilactic.com*), an aggregator where users can specify data about themselves and link their Profilactic profile to their own profiles on other social systems. For our evaluation, we selected the Profilactic profiles with links both to MySpace and to Flickr profiles.

The core of the evaluation was providing the algorithms with Profilactic profiles, *hiding the links to MySpace and Flickr profiles* and *letting the service to retrieve them* on the two SSs. To this aim, beside the parsers for MySpace and Flickr, mentioned above, we developed a parser for extracting the dataset of the Profilactic profiles.

Sampling. On January 2010, we extracted the Profilactic profiles containing links both to MySpace and to Flickr profiles. On a total of 511 profiles extracted from Profilactic, we selected the full profiles, namely the profiles with all the following types of attributes filled in: nick or full name, age, gender, city, province, country (besides the links to MySpace and to Flickr profiles). We obtained 333 profiles, and we extracted a sample of 300 profiles, according to a random sampling strategy.

In the following we provide details about the first and second step of the evaluation.

4.1 Calculating the Probability Distribution of User Identification

As a first step of the experimentation, we ran the algorithms using as input attributes the profiles of our sample, hiding information about MySpace and Flickr profiles. The objective was to calculate the User Identification Probability $\mathrm{IdP}(\mathrm{Pr}_j|\mathrm{Pi}_i,\mathrm{Pr}_i)$, being Pi_i the set of input profiles (with $0<i<=300$), Pr_j the profiles retrieved on MySpace and Flickr with an associated probability of belonging to the same user and Pr_i the retrieved profiles that condition the final IdP, as defined in Sec. 3.2..

In order to perform a realistic simulation of a querying system searching for a user profile, we *subdivided the evaluation in 8 tests*, by varying the number and type of the users' attributes provided to the algorithm as input profile Pi. In this way Pi can be made of $k=8$ different combinations of attributes (a), so that we evaluate 8 different probabilities $\mathrm{IdP}(\mathrm{Pr}_j|\mathrm{Pi}_i=a_k)$.

As said before, the complete Profilactic profiles include information about the user nick/full name, age, gender, city, province and country. We group together full names and nicks since, in the evaluated social networks, full name is often filled in by users as a nickname. Moreover, since province and country can be mostly derived by city, we just consider the attribute city.

Running the algorithm, for each test we obtain (see Sec. 3.2) a list of profiles retrieved on Flickr and on Myspace, each one associated to a *probability of identification IdP* and a list of attributes, with an associated probability, coming from the retrieved profiles.

Results for the IdP(Pr_j | Pi_i) on each test and for the average IdP(Pr_j | Pi_i)

Considering the probability of identification of each Pr profile retrieved in each test, we obtain a distribution of probability of identification for each combination a_k of attributes. **Fig. 1**, displays the probability distribution obtained in each of the 8 tests. Notice that the retrieved profiles Pr_j, on the x-axis, include all the profiles retrieved in all tests. Such profiles are numbered in descending order (Pr_1 is the profile with x=1 and with y=the highest probability of identification). Notice, moreover, that we refer to the attributes of each test with their initial letter.

Fig. 1 displays also the *average result* for the eight tests (in the figure, it is the dotted red line). Indeed, the main reason for splitting the evaluation in 8 tests is the possibility to obtain a realistic *average result*, by combining the probabilities of identification estimated for each different combination of input attributes. Moreover, this splitting allows us to analyze if the probability of identification significantly changes by changing the type and number of attributes provided to the algorithm.

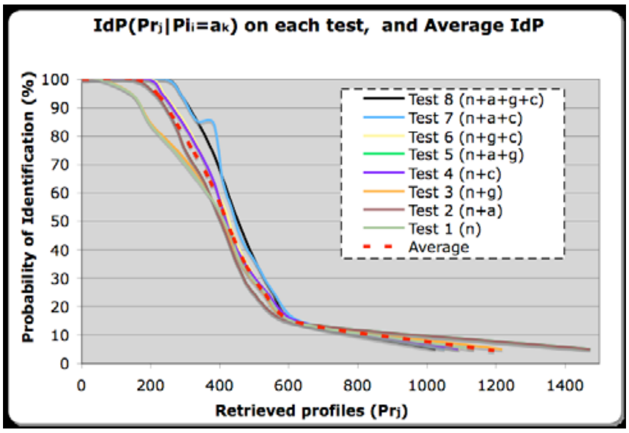

Fig. 1. Probability of identifying MySpace and Flickr profiles belonging to the Profilactic users.

Changes can depend both on the algorithm (e.g. the weights used for attributes) and on the real data available in the social networks considered. As it can be seen, the distribution is very polarized in all the tests so that, considering the average result, all the profiles under the third decile have a probability of identification over 86%, which means that more than 290 profiles are identified with a very high probability (given that each one of the 300 Profilactic profiles has a link on both the social networks, the maximum number of correct profiles that can be retrieved is 600).

4.2 Evaluating the accuracy of the probabilities associated to the retrieved profiles

In this section we introduce the second step of the evaluation, which consists in *evaluating if the probabilities of identification obtained are correct*, that is evaluating the accuracy of the identification of profiles.

To this purpose, we exploited the dataset of Profilactic profiles. The availability of this dataset allowed us to use the same set of data as experimental set, also called test set, and as control set (namely the set used as a standard of comparison in a control experiment). In the previous step of the evaluation we provided the Profilactic profiles to our algorithm, hiding information about the links to MySpace and Flickr profiles. Then we compared the profiles retrieved by the algorithm with the true profiles on MySpace and Skype associated to each Profilactic profile.

Notice that, as regard to the accuracy of the probability of the discovered attributes, in the following we will not display specific results, being such results highly correlated to the IdP of the retrieved profiles.

To evaluate the accuracy of the identification, we used a technique similar to those frequently used in Information Retrieval: the Precision and Recall metrics [8]. In measuring Precision and Recall, we considered as true retrieved profile those profiles which have been correctly identified by applying the User Identification Algorithm.

In **Table 4** we provide the results of Precision and Recall of the average identification probabilities $\mathrm{IdP}(\overline{\mathrm{Pr}} \mid \overline{\mathrm{Pi}})$ computed in the previous step as a mean between the 8 tests. The identification probabilities of the retrieved profiles are grouped in ten ranges: 0<IdP<10%, 10<=IdP<20%, etc. and for each Precision and Recall result we specify, in brackets, the numerator and denominator of the metrics. Finally, **Table 5** displays precision and recall for progressive IdP thresholds.

Table 4. Precision and Recall of the average identification probabilities IdP(Pr | Pi)

Range %	Precision of average IdP(Pr|Pi)		Recall %
	IdP of retrieved profiles %	Precision % (true $\mathtt{Pr_j}$ / $\mathtt{Pr_j}$)[1]	% (over 600)
IdP 0 -9	4,5	5 (28.5/572.4)	4.7
Idp 10 -20	14,5	14.3 (13/91.1)	2.2
IdP 20-29	24,5	28.6 (16/55.9)	2.2
IdP 30-39	34,5	38.9 (14.6/37.6)	2.4
IdP 40-49	44,5	48.2 (15.1/31.3)	2.5
IdP 50-59	54,5	65.0 (23.5/36.1)	3.9
IdP 60-69	64,5	75.5 (34.6/45.9)	5.8
IdP 70-79	74,5	81.3 (40.7/50.1)	6.8
IdP 80-89	84,5	88.1 (46.5/52.7)	7.7
IdP 90-99	94,5	89.9.1 (57/63.4)	9.5
IdP 100	100	94.9 (150.5/158.6)	25.3
			Tot 73%

[1] We report decimal values since they are mean values.

Table 5. Recall and Precision for progressive IdP thresholds

Range %	IdP Thresholds %
IdP 0 -9	IdP 100 --> R. 25.3% - P. 94.9%
Idp 10 -20	IdP over 90% --> R. 34.8% - P. 93.4%
IdP 20-29	IdP over 80% --> R. 42.5% - P. 92.4%
IdP 30-39	IdP over 70% --> R. 49.3% - P. 90.7%
IdP 40-49	IdP over 60% --> R. 55.1% - P. 88.8%
IdP 50-59	IdP over 50% --> R. 59% - P. 86.7%
IdP 60-69	IdP over 40% --> R. 61.5% - P. 84%
IdP 70-79	IdP over 30% --> R. 63.9% - P. 80.4%
IdP 80-89	IdP over 20% --> R. 66.1% - P. 75%
IdP 90-99	IdP over 10% --> R. 68.3% - P. 66.1%
IdP 100	IdP 0% --> R. 73% - P. 88.8%

Notice that Precision values have to be read by comparing them with the estimated IdP (second column in Table 4) since the evaluation aims at testing if the probabilities provided to a requiring system are reliable or not. In this way, the choice whether to accept or not profiles with probabilities below some threshold depends on the querying system itself. As it is clear from the table, the distance of the estimated IdP from the true probabilities (precision), for a given range of probabilities, is very low, and thus the result is good.

The distance between IdP and precision can be also seen in **Fig. 2**.

Considering **Table 4,** we observe that the overall Recall is 73%, but, as typical, it has an inverse relationship with Precision, hence it is possible to increase the former only at the cost of reducing the latter.

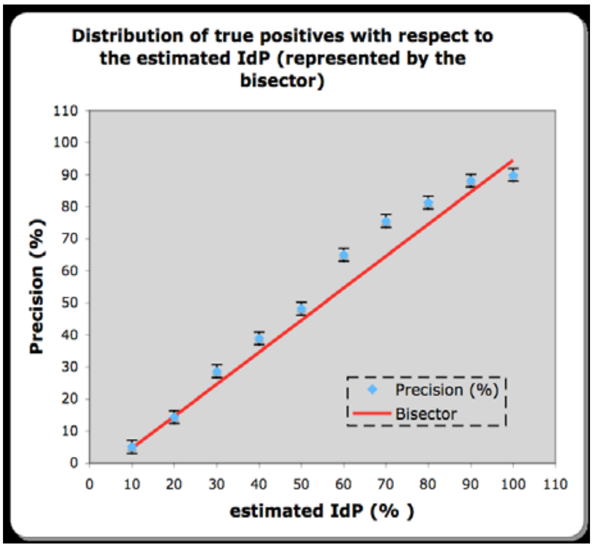

Fig. 2 Distance between prevision and true values.

Table. 5 shows the relation of the two metrics for progressive IdP thresholds. We can observe that even with IdP thresholds over 60%, the Recall does not decrease too much, being 55.1%, and the mean Precision is 88.8%.

Notice, moreover, that Recall includes all those profiles that cannot be intrinsically retrieved, given the definition of the algorithm (e.g. profiles with username completely different from the input profile and no full name available). We estimated that at least 20% of profiles cannot be identified, given the features of our algorithm, mainly non consistent profiles belonging to the same user, or with errors and linguistic variations.

Besides providing an average evaluation of the discovery probability, as explained in the previous section, the objective of performing 8 different tests was also to analyze the impact of providing the algorithm with different types and number of attributes. From this point of view, it is interesting to note the relevance of the attribute city, which entails the increase of recall values with respect to tests where it is not indicated. This is due to the fact that our algorithm gives city a weight higher than other attributes such as gender and age in case of positive match (depending on the a priori probability of their values).

As a final remark, let us consider that using Profilactic as a dataset we wanted to test the validity of the algorithm, not the probability of identification of a user in a real context. In fact, people that use an aggregator of profiles, as Profilactic subjects, are users probably more careful than other users in filling in the attributes of their profile. Moreover, Profilactic sample was very useful, providing a test set and a control set, but with limited dimension. The size of our sample should be increased in order to be more significative.

5 Conclusions

In a cross-systems personalization scenario, user identification, which means for different user-adapted systems partaking into the user data exchange process to discover if they are referring to the same user, is so far a challenge.

This paper presents an approach to support adaptive systems to uniquely identify users on different social systems and to retrieve the user data distributed over the profiles stored in such systems. The approach we present does not require the provision of authentication data and user identification, since it is performed by using the public data available on the Web.

To test such algorithm we ran it on two social networks: MySpace and Flickr.

One of the main advantage of our approach is that it does not require the implementation of specific protocols and the provision of authentication data. Moreover, notice that the algorithm we implemented and the scenario above allows to exploit the public data collected from social systems.

Anyway, notice that the approach is quite flexible, since it can be implemented as well within a specific framework of co-operating systems, institutions, organizations. In this case, the distributed profiles that can be crawled can include not only public data, but also all data our service is authorized to access to. This possibility extends the applicability of the approach, which, otherwise, could be limited with respect to the possible needs of a personalization system. Depending on the kind of crawled systems the data about users can be very specific and domain dependent. This requires particular attention in matching the attributes since their meaning could be different. The phases of parsers' building have always to be realized with a careful analysis of the attributes that will be matched. Anyway, given that these phase are manual and not automatic, they do not determine relevant problems.

Regarding possible limitations, we should consider that the user identification and attributes discovery can be limited by the fact that the profiles scattered across social system are not always correct and updated. However we notice that social systems, and social networks in particular, are used to have relationships with friends, for profession aims, etc. thus users are motivated to provide true data and to update them. A further consideration regards the Identification Probability IdP (Sect. 3.2) and the choice of a third system can do whether to accept or not the retrieved profiles, given the initial known attributes for the search. The possibility of knowing the Precision and Recall for different IdP thresholds and for different combinations of attributes can support this decision. Considering the average results of different combinations of attributes, we saw, in Table 4, that for a threshold IdP over 50%, the Precision and Recall are good enough: respectively over 86,7% and 59%, thus the advice to a searching system could be to accept the results over such a threshold.

Varying the combination of initial attributes provided, Precision and Recall change for each threshold.

4. REFERENCES

[1] Berkovsky, S., Kuflik, T., Ricci, F. :Mediation of User Models for Enhanced Personalization in Recommender Systems. Journal of User Modeling and User-Adapted Interaction 18(3), 245–286 (2008)

[2] Brooks, C., Winter, M., Greer, J., McCalla, G.: The Massive User Modelling System (MUMS). In: Intelligent Tutoring Systems. pp. 635–645, (2004)

[3] Carmagnola, F., Cena, F.: User Identification for Cross-System Personalisation. Information Sciences 179(1-2), pp. 16–32, (2009).

[4] Carmagnola, F., Osborne, F., Torre, I.: Cross-Systems Identification of Users in the Social Web. In: 8th IADIS Int. Conf. WWW/INTERNET (2009), Rome, Italy, (2009).

[5] Dolog, P., Schäfer, M.: A framework for browsing, manipulating and maintaining interoperable learner profiles. In: 10th International Conference on User Modeling, pp. 397-401, Edinburgh, Scotland, UK, (2005).

[6] Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., and Farrell, S.: Harvesting with SONAR: the value of aggregating social network information. Proc. CHI '08, 1017-1026 (2008)

[7] Heckmann, D.: Ubiquitous User Modeling. Ph.D. thesis, Department of Computer Science Saarbrucken, University of Saarlandes, (2005).

[8] Herlocker, J., Konstan, J., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1):5–53, (2004).

[9] Kobsa, A., Fink, J.: An LDAP-Based User Modeling Server and its Evaluation. Journal of User Modeling and User-Adapted Interaction 16(2), 129–169, (2006).

[10] Kobsa, A., Koenemann, J., Pohl, W.: Personalized hypermedia presentation techniques for improving online customer relationships. Knowledge Eng. Rev. 16(2), 111–155 (2001)

[11] Leonardi, E., Houben G-J, van der Sluijs, K,, Hidders, J., Herder E., Abel F, Krause D, Heckmann, D: User Profile Elicitation and Conversion in a Mashup Environment. In Int. Workshop on Lightweight Integration on the Web, in conjunction with ICWE 09, San Sebastian, Spain, June, (2009).

[12] Mehta, B., Niederee, C., Stewart, A., Degemmis, M., Lops, P., Semeraro, G.: Ontologically-Enriched Unified User Modeling for Cross-System Personalization. In: 10th Int. Conf. on User Modeling, pp.119-123, Edinburgh, Scotland, UK, (2005).

[13] Szomszor, M., Alani, H., Cantador, I., O'Hara, K. and Shadbolt, N.: Semantic modelling of user interests based on cross-folksonomy analysis. In: 7th Int Semantic Web Conf. ISWC, Karlsruhe, Germany, pp. 632–648, (2008).

[14] Vassileva J, Distributed user modelling for universal information access, International Journal of Human–Computer Interaction, 122–126 (2001).

[15] Wang, Y., Cena, F., Carmagnola, F., Cortassa, O., Gena, C., Stash, N., Aroyo, L.: RSS-based Interoperability for User Adaptive Systems. In: 5th Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 353-356, Hannover, Germany, (2008).

[16] Windley, P., *Digital Identity*. O'Reilly Media, Inc., Sebastopol, CA (2005)