Explicit vs. Implicit Tagging for User Modeling

Enrique Frias-Martinez[1], Manuel Cebrian[1], J. Moises Pascual[2], Nuria Oliver[1]

[1] Data Mining and User Modeling Group, Telefonica Research
Emilio Vargas 6, 28043, Madrid, Spain
[2] Telefonica I+D, Emilio Vargas 6, 28043, Madrid, Spain
{efm, manuelc, mpascua, nuriao}@tid.es

Abstract. Tagging has been popularized by Web 2.0 sites as a way to describe resources. Typically, tagging has been done in an explicit way in which users directly describe with tags the resources in which they are interested. However in today's ubiquitous computing environments, it is possible to implicitly tag resources. This paper: (1) introduces the concept of explicit and implicit tagging for user models in two domains: Web 2.0 and mobile phone usage, respectively; and (2) compares the characteristics of both tagging mechanisms. Results indicate that the use of tags in both approaches is very similar, whereas the statistical characteristics of the common-interest networks are different.

Keywords: Tagging, User Modeling, Cell Phone Usage.

1 Introduction

Tagging has become a *de facto* method for assigning a set of descriptors (or keywords) to Internet digital content. The use of keywords for describing content was already in use before the Internet. However, with the advent of Web 2.0 technologies, a collaborative dimension was added. The main characteristics of a collaborative tagging system are [1]: (1) the free-nature of the tags, (2) tags are bottom-up non-hierarchical classifications and (3) there are no specific rules on how to annotate the resources and annotations are not necessarily done by experts. These characteristics are different from traditional classification hierarchies (taxonomies) where there are a limited number of tags which define a top-down hierarchy.

Most Web 2.0 tagging systems, such as Flickr (www.flickr.com) or Bibsonomy (www.bibsonomy.com), can be characterized as explicit or user-guided tagging, due to the involvement of the users in describing the resources. These systems, however, suffer from some drawbacks: (1) the semantics given to a tag is not necessarily the same for all users; (2) users may not be sure about what they are tagging; (3) users may tag different resources with the same set of tags to save time; etc. These limitations could be partially solved by *implicit* tagging systems. The concept of implicit or automatic tagging consists of assigning tags to a given resource *without* the intervention of a user. In today's ubiquitous computing environments, there are a lot of sources of information that can be used to automatically tag resources, including: geographical information, mobility and communication patterns, information about resources explicitly tagged, etc.

In this paper, we present two tagging systems, one explicit (Bibsonomy) and the other implicit (derived from mobile data usage) and compare their statistical properties. The goal is to test to which extent what it is known about explicit tagging environments can be applied to implicit tagging systems. The concept of explicit *vs.* implicit tagging is to some extent inherited from the adaptive *vs.* adaptable [9] or implicit *vs.* explicit [10] concepts used in user modeling.

2 User Model Generation from Explicit & Implicit Tagging

In this section, we present the two tagging environments used in our analysis. The explicit tagging system used is Bibsonomy. The implicit tagging system is constructed using the calling behavior of mobile phone users to businesses.

2.1 Bibsonomy

Bibsonomy [2] is a social bookmarking system in which users describe the resources added to their shared personal library by means of tags. The data considered for this study is freely available at [3]. We use the TAS (Tagging Association) file from [3] which contains 816,197 entries. Each entry consists of a user ID, one tag, and a resource (bookmark or publication) tagged by that user. The TAS file contains only non-spammers. No tag semantics were used in the experiments. Each user typically has multiple entries for a given resource, one for each tag introduced. The final dataset used in our experiments contains 2,467 unique users who assigned 69,902 unique tags to 268,692 resources. The set of tags explicitly introduced by a user are considered the user model that describes the interest of that particular user.

2.2 Cell Phone Usage (SME)

Data from a major cell phone carrier was obtained for a number of users close to 3,000. The original data set only contained the originating phone number and the destination phone number, both encrypted. In this context, the resource being tagged is each user and the tagging mechanism assigns tags that describe his/her interests. In order to automatically assign tags to users, an encrypted directory of businesses and services was considered. The directory contains a set of predefined tags for each encrypted business phone number: (1) the type of business (*e.g.* hotel, restaurant, car dealer, etc.); (2) a categorization of the business – only in the case of some businesses, (*e.g.* number of stars in the case of hotels, price range in the case of restaurants) and (3) an identifier of the town where the business is located.

The tagging interest model for each user was obtained by only considering the phone calls made by the user to the phone numbers included in the directory. A composite tag was generated for each one of these calls by concatenating the set of predefined tags for that particular business. In our study, we considered 57 business types, 3 of which also had a categorization (hotels, restaurants and academies), and 51 towns. From a total of 3366 possibilities after tag concatenation, 2044 unique tags were generated. We refer to this system as SME for "small and medium enterprises" because the directory of business used presented these types of companies.

3 Comparative Analysis

This section compares the statistical characteristics of both tagging systems. The comparative analysis focuses on: (1) tag frequency; (2) total number of tags per user and (3) characteristics of the common-interest social networks created.

3.1 Tag Frequency

The number of unique tags used in both systems is significantly very different: 69902 tags in Bibsonomy and 2044 tags in SME. Figure 1 (left) presents in a log-log representation the survival function of the tag frequency for Bibsonomy (bib, Bibsonomy, in circles) and SME (SME, in squares). The x-axis represents the probability of each tag, ordered by decreasing probability, while the y-axis represents the tag frequency. The software used was Clauset's et al. [11] algorithm for power law fitting and the MATLAB statistical toolbox for the lognormal and exponential fit. The best fit was identified in terms of root mean squared error. Both tagging systems show a similar power law distribution with similar slope values (1.75 –explicit-- and 1.8 --implicit--), which create almost two parallel distributions. The head of both distributions represents tags that are heavily used, such as "software" in the case of Bibsonomy or "Hotel4Stars28079" in the case of SME. Similar statistical behavior has been shown in other explicit tagging communities [4][5]. Preferential attachment behavior, which is typical in explicit systems, seems to hold true also in our implicit system, *i.e.* there are a few core businesses that receive the core of the calls.

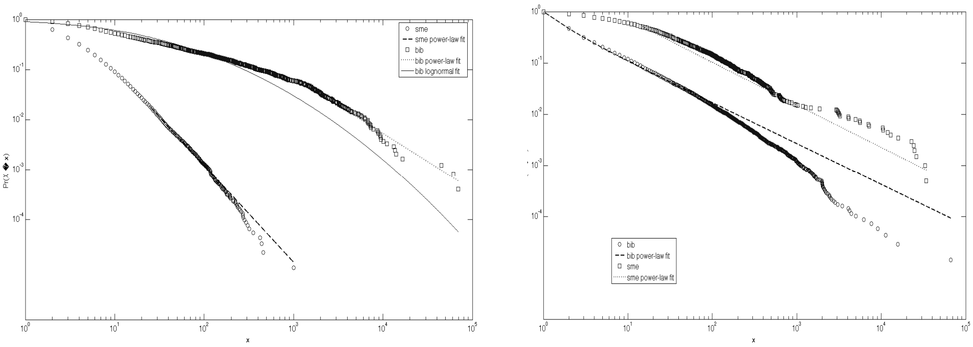

Fig. 1. Tag frequency (left) and number of total tags per user (right).

Intuitively, one of the biggest advantages of an explicit tagging system over an implicit one is the free nature of the vocabulary. Nevertheless, our statistical results indicate that the use of the tags is fairly similar in both cases. The results also highlight a well known behavior in explicit tagging systems: users may have an unlimited number of tags at their disposal. However, they end up using a reduced number of tags. Therefore and in the context of this experiment, the tagging behavior in the explicit system is similar to that of an implicit tagging system

3.2 Total Number of Tags per User

Figure 1 (right) presents in a log-log scale the survival function of the total number of tags used by each user in the explicit (squares) and implicit (circles) tagging systems. The total number of tags is given by the total number of tags introduced by a user (including repetitions) in Bibsonomy and by the total number of calls in SME. In this case, while SME has a straight forward power law distribution with α=2.93, the explicit system is better modeled by two distributions: a lognormal distribution (μ=2.87, σ=2.14) for its head, and a power-law distribution with α=2.13 for its tail. In the literature of explicit tagging systems, the tail of the distribution of tags per resource is usually approximated with a power law fitting [4][5]. This difference in the behavior of both tagging systems is probably not relevant, as the power law distribution of SME is somewhat defined by the size of the data collected.

3.3 Characteristics of the Interest-based Social Network

Next, we build a network of users that share common interests as measured by the number of common tags that are either introduced (Bibsonomy) or inferred (SME). The network is composed of two types of nodes: (1) user nodes and (2) interest nodes. The structure of the network is created by linking the user nodes with the interest nodes, as specified by each user model. Two elements can be studied to characterize the structure of each network: the number of unique tags per user and the number of unique users per tag. Figure 2 (left) depicts the number of unique tags per user for Bibsonomy (circles) and SME (squares). The x-axis represents the probability of the number of different tags per user while the y-axis presents the number of tags. Bibsonomy follows a power law distribution (α=3.44) while SME is better fitted by an exponential distribution with coefficient 1.67. Figure 2 (right) displays the survival function of the number of unique users per tag for Bibsonomy (circles) and SME (squares). In this case, both systems are better fitted by a power law distribution with α=2.37 (Bibsonomy) and α=1.83 (SME).

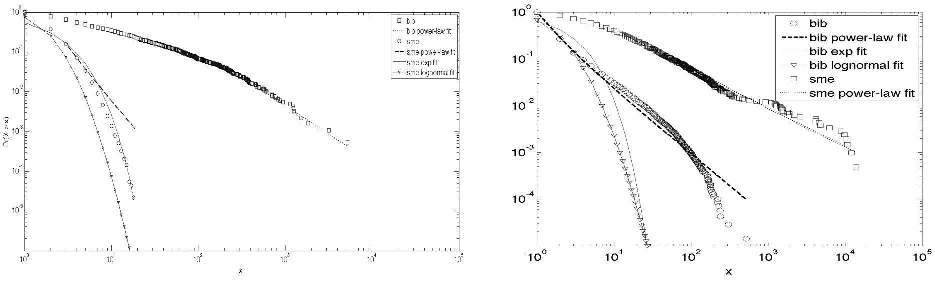

Fig. 2. (left) Unique tags per user and (right) distribution of different users per tag.

Power law distributions have been observed in a great variety of networks, including the WWW [7] and citation graphs [8]. These observations seem to hold true in a network based on common interests. Our results also indicate that the networks built from implicit and explicit tagging systems have a similar architecture, with the difference that the role of user nodes in the explicit social network is played by the interest nodes in the implicit network and *vice versa*. This inversion of roles is probably related to the different number of interests represented in each system.

4 Conclusions & Future Work

In this paper, we have compared two tagging mechanisms: an explicit system provided by Bibsonomy and an implicit one generated from cell phone usage behavior. Our results show that the statistical tag frequency of the explicit and implicit tagging systems is similar. While users in Bibsonomy may have the opportunity to use any tag, they end up using a reduced set of tags. This behavior is probably due to users having a limited set of interests that can be described by a reduced number of tags. The interest networks built from explicit and implicit tagging display different statistical behavior in the degree distribution of user and interest nodes. This difference might be caused by the limited number of interest tags used in the implicit case.

In future work, we plan to: (a) test if the statistical behavior of the tagging systems is better characterized by a Double Pareto log normal distribution [6]; (b) study how the size of the tagging vocabulary affects its statistical properties; (c) explore how well our results generalize to other domains and (d) develop mobile applications that include the user models learned from the implicit system.

References

1. Golder, S.: The structure of collaborative tagging systems. J. Inf. Science (2006).
2. http://www.kde.cs.uni-assel.de/ws/rsdc08/.
3. Jäschke, R., Hotho, A., Schmitz, C., Stumme, G.: Analysis of the Publication Sharing Behaviour in BibSonomy, *Proc. Knowledge Architectures for Smart Applications* (2006)
4. Sigurbjörnsson, B., van Zwol, R.: Flicker Tag Recommendation based on Collective Knowledge, *WWW 2008* (2008).
5. Angelova, R., Lipczak, M., Milios, E., Pralat, P.:Characterizing a social bookmarking and tagging network, *Mining Social Data Workshop, 18th Europ. Conf. AI* (2008).
6. Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., Leskovec, J.: Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions, 596-604, KDD 2008.
7. Broder, A.Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph Structure in the web: experiment and models, WWW 2000.
8. Redner, S.: How popular is your paper? An empirical study of citation distribution. European Physics Journal B (1998)
9. Brusilovsky P., Schwarz E. User as Student: Towards an Adaptive Interface for Advanced Web-Based Applications. *User Modeling: Proc.6th Int. Conf.* UM97, 177-188 (1997).
10. Quiroga, L.M., Mostafa, J. Empirical evaluation of explicit versus implicit acquisition of user profiles. In *Proc.fourth ACM conference on Digital libraries*, 238-239 (1999).
11. Cluset, A., Rohilla, C, in press. Power-law distributions in empirical data. SIAM Review.