Identifying Users Across Social Tagging Systems

Tereza Iofciu[1], **Peter Fankhauser**[1], **Fabian Abel**[2], **Kerstin Bischoff**[1]
[1] L3S Research Center, Leibniz University Hannover
{*iofciu,fankhauser,bischoff*}*@L3S.de*
[2] Web Information Systems, TU Delft
f.abel@tudelft.nl

Abstract

How much do tagging activities tell about a user? Is it possible to identify people in Delicious based on the tags, which they use in Flickr? In this paper we study those questions and investigate whether users can be identified across social tagging systems. We combine two kinds of information: their user ids and their tags. We introduce and compare a variety of approaches to measure the distance between user profiles for identification. With the best performing combination we achieve, depending on the actual settings, accuracies of between 60% and 80%, which demonstrates that the traces of Web 2.0 users can reveal quite much about their identity.

Introduction

Today, people have online accounts on diverse Web portals where they leave plenty of multifaceted profile data. Users share their pictures, videos, or bookmarks at platforms such as Flickr, YouTube, and Delicious and annotate these resources using tags to facilitate retrieval of the resources, to express their opinion regarding some resource or merely to present themselves (cf. Marlow et al. 2006).

Aggregating profiles from different systems reveals more information about users and is beneficial for personalization and cross-domain recommendations – particularly for solving cold-start problems where systems suffer from sparse user profiles (Abel et al. 2010). However, for privacy reasons, people may not want their different online accounts to be connectable. Indeed, the interlinkage of profile information may be risky. Recently, *PleaseRobMe*[1] set an intimidating example and attracted public attention as they exploited *foursquare*[2] to detect the current location of Twitter users and identify – given the linkage to the address of these users – houses and apartments that were easy to burgle as the inhabitants were traveling at the time.

(Un)fortunately, automatically connecting the different Social Web identities of the users is difficult because they might (possibly on purpose) use varying usernames or have unequal profiles (e.g. fields such as homepage, birthday, etc.) on the different systems. Yet, the feasibility of exploiting individual tagging practices to identify a user and link her Social Web accounts has not been studied in detail.

In this paper, we close this gap and study the following research question: is it possible to identify users across systems based on their (tag-based) profiles? We analyze profiles of users from three collaborative tagging systems: Flickr, Delicious and StumbleUpon. While the latter two systems are for organizing public Web resources, Flickr is mainly for sharing personal pictures with friends and people rarely tag other people's photos.

Research Challenge User profiles can be constructed based on implicit and explicit user feedback. With explicit feedback, we refer to the data the user herself provides to the system directly, e.g. during the registration process. Usually such explicit data is structured as attribute-value pairs. In our approaches we experiment with the usernames as explicit profile information. With implicit feedback, we refer to the users' tagging activities within the folksonomy systems, i.e. the set of tag assignments performed by the user.

> **User Identification Challenge.** Given u_a, the tag-based profile and/or username of user X in system A, and U_B, the set of profiles from system B, the challenge of the user identification strategies is to rank the profiles from system B so that $u_b \in U_B$, the profile of X in system B, appears at the very top of the ranking.

Contributions Our main contributions can be summarized as follows:

- We propose different strategies that allow for the identification of users across systems.
- For tag-based profile mapping, we introduce a symmetric variant of BM25 using site specific statistics and compare it against measures like TF, TFIDF and conventional BM25. The results show that it is important to account for the specifics of a site.
- We evaluate the different matching approaches in experiments with public profiles from three different social tagging networks, Flickr, Delicious and StumbleUpon. We show how by combining implicit and explicit profiles we reach an accuracy of over 60%.
- Furthermore, we show how by aggregating the users' profiles from different sources, we can identify users with an accuracy of almost 80%.

[1] http://pleaserobme.com
[2] http://foursquare.com

Related Work

The issue of identifying users via their interaction over the web has been recently addressed in various application scenarios, such as personalization. Recent research mainly analyzed whether explicit profile information is sufficient to identify users across system boundaries. For example, Carmagnola and Cena (2009) introduce an approach that bases heuristics on profile attributes such as username, name, location or email address of a user. Vosecky, Hong, and Shen (2009) examine explicit user profile information from two similar social networking services to find which fields in the profiles are best suitable for user cross-system identification. Zafarani and Liu (2009) connect user accounts across 12 communities exploiting explicit profile information.

Szomszor, Cantador, and Alani (2008) focus on implicit tagging information. However, they do not aim to identify users across tagging platforms, but their goal is to align tag-based profiles users have in Flickr and Delicious and assume that linkage between different accounts of the same user is given. Nevertheless, the authors propose an approach for correlating tag clouds, which are filtered (to eliminate misspellings) and semantically enriched (e.g., via WordNet synonyms). We have implemented also their approach, but for our task it led to a significant decrease in performance.

More generally, the problem of identifying users can be regarded as an instance of duplicate detection – also known as record linkage or entity resolution – a long standing problem in computer science. For a comprehensive overview on the topic see Elmagarmid, Ipeirotis, and Verykios 2007. In a sense the experiments described in this paper return to the very root application domain of duplicate detection – identifying individuals – though under quite different circumstances. By tagging and other forms of interactions, Web 2.0 users provide a rich but fairly noisy trace, which as we will show can be readily exploited for identifying them.

Matching Users across Sites

Matching Users based on their Tags

For identifying users across social systems based on their tagging behavior, we experiment with standard techniques like TF, TFIDF and BM25 and compare it against a new symmetric variant of BM25 using site specific statistics.

Baselines One of the most straightforward approaches to match tag user profiles exploits tag frequencies. We evaluate this approach as one baseline (TF). However, this approach does not take into account the specificity of tags. Tags used by many users such as "web" contribute much less evidence for a match than more specific tags such as "NYC". To take into account tag specificity, frequency is typically combined with the inverse document frequency of a tag. Since together with the vector space model it is a standard method in Information Retrieval, we evaluate this approach as another baseline (TFIDF).

Each user profile u is modeled as a vector, where each dimension contains the TF or TFIDF value of tag $t \in T$. The matching score of two profiles u_1 and u_2 is then determined by the cosine distance between their weighted vectors.

BM25 A well known weakness of the TFIDF weighting scheme is that term frequency is not a very good indicator for the relevance of a term. If a document contains a term 20 times, it is not 20 times as relevant as occurring just once. Similarly, if a user assigns a tag 20 times, this is not 20 times as relevant as assigning a tag just once. Okapi BM25 (Spärck Jones, Walker, and Robertson 2000) addresses this weakness by tempering term frequency such that it quickly saturates with a maximum value.

Site specific IDF and BM25 Tagging behavior is influenced highly by the site's domain and design choices. People tag differently music items, images or Web resources (Bischoff et al. 2008). For example, in our experimental dataset "tools" is used for more than half of the resources in Delicious, but only few times in Flickr, conversely, "arts" is used very often in Flickr, and rarely in Delicious. Hence, "tools" is a very discriminative tag when matching against Flickr, while "arts" discriminates well against profiles in Delicious. As a consequence the document frequency of particular tags may differ substantially among the sites. For the same reason of dependency on system design choices like tagging rights, object type and ownership, etc. (Marlow et al. 2006), also profile lengths may be very different. Thus, we suggest to use BM25 together with a site specific IDF and site specific average profile lengths:

$$\begin{aligned} w(u,t,s) &= TF(u,t,s) * \sqrt{IDF(t,s)} \\ TF(u,t,s) &= \frac{c(t,u) * (k_1+1)}{c(t,u) + k_1 * (1 - b + b * \frac{|u|}{avgU(s)})} \\ IDF(t,s) &= max(0, log \frac{N(s) - n(t,s) + c}{n(t,s) + (1-c)}) \end{aligned} \quad (1)$$

Thereby, TF takes into account the site specific profile length $avgU(s)$ and IDF takes into account the site specific document frequency of a tag $n(t,s)$. As shown below, this approach leads to significantly improved matching.

Matching Users based on Explicit Usernames

Often, the username is the only explicit and publicly available user attribute common to various tagging systems. A straightforward approach for identifying users across systems is to analyze their usernames (Zafarani and Liu 2009).

We apply the following string similarity metrics: exact match, Jaccard similarity at character level, Levenshtein similarity (minimum number of editing operations required to transform one string into another string), Smith-Waterman similarity (the costs of aligning two strings by comparing segments of all possible lengths between two strings) and Longest Common Substring (LCS), a variation to Levenshtein distance allowing only addition and deletion.

Matching Users based on Combined Profiles

Now we present approaches to merge different sources of user information, first by combining implicit and explicit profile information and, second, by aggregating profiles from two systems to map against a third system.

Combining Username and Tags In order to combine the different types of profiles, tag- and username-based, we use a mixture model:

$$w(u_1,u_2) = \lambda * w_t(u_1,u_2) + (1-\lambda) * w_u(u_1,u_2) \quad (2)$$

where $w_t(u_1, u_2)$ is the normalized score obtained based on the tags the user assigned, as presented above; and $w_u(u_1, u_2)$ is the string similarity of the usernames of the two users. As the BM25 scores are not normalized between 0 and 1, we scale them to the same range as the scores on username similarity by dividing them with the maximum score of all compared user tag profiles between two systems. Thereby the choice of λ indeed reflects the relative importance of the two scores used for matching.

Aggregated Profiles Our evaluation will show that matching accuracy via tags depends heavily on the number of tags given by a user. Hence, we create an *aggregated tag-based user profile* by considering all the tags a user has assigned in two systems, on which the user is already identified (e.g. via explicit links on her profile page). Therefore, we accumulate the tag frequencies of the corresponding profiles and then apply the same comparison approaches for matching users between the aggregated profile and a third system as presented above. For creating an *aggregated username-based profile* from two systems, we consider the two usernames as matching candidates. When calculating the distance between the aggregated username profile and a third profile we select the highest matching username pair.

Evaluation

Method and Metrics

The user identification algorithms have to find for each user profile the corresponding profile that refers to the same user in another system, i.e. each algorithm is tested in different settings which are given by the different service constellations. For example, (i) given the Flickr profile of user u, the algorithm has to rank Delicious profiles so that u's Delicious profile appears at the very top of the ranking.

Dataset To investigate the questions above, we crawled public profiles of 421,188 distinct users via the Social Graph API[3], which makes information about connections between different user accounts of a user available. However, only a few users linked the profiles they have at social tagging platforms. Among these users, 1467 people had a Flickr *and* Delicious profile (*FD dataset*) and only 321 users had a tag-based profile at all three systems, i.e. Flickr *and* Delicious *and* StumbleUpon (*FDS dataset*).

A remarkable feature of the dataset is that only a few tags occur in more than one service: less than 20% of the distinct tags were used in more than one system. For each user and each pair of services we compute the overlap as the number of distinct tags that occur in both profiles. The Delicious and StumbleUpon profiles have the biggest overlap. However, the overlap is rather small: for more than 50% of the users the overlap of their Delicious and StumbleUpon profiles is less than 20% and there exist only 6 users for whom the overlap is slightly larger than 50%. It is interesting that the overlap is so small, as in Delicious and StumbleUpon the same type of resources are tagged, probably the tools are used for separate tasks. Flickr and StumbleUpon profiles offer the least overlap as for more than 40% the overlap is 0%. The small overlaps of individual profiles indicate that user identification based on tagging is not trivial.

[3] http://code.google.com/apis/socialgraph/

Metrics To measure the quality of the user profile rankings we use *MRR* and *S@k*. *MRR* (Mean Reciprocal Rank) indicates at which rank the correct profile occurs on average. The Success at rank k (*S@k*) stands for the mean probability that the correct profile occurs within the top k of the ranked results. In case of tied scores between the correct user profile pair and some other pairs, we penalized both metrics by dividing them by the number of tied scores.

Results

Matching Users based on Tags Table 1 compares the various approaches for identifying users. Regarding tag-based profiles (profile type: tag), BM25 clearly outperforms TFIDF, and BM25 with site specific IDF also clearly outperforms BM25 with global IDF. This suggests that accounting for site and domain specific characteristics in tag weighting is promising. All methods yield substantially better results than the baseline approach using TF with cosine similarity, BM25 with site specific IDF improves its *S@1* by even 2.5 times. By operating on a larger set of users (cf. FD column) the chance of a mismatch increases for all metrics. However, the relative ordering is consistent. BM25 with site specific IDF ($k_1 = 3.75$, $b = 1$, and $c = 1$) outperforms all other approaches and looking at MRR it is less influenced by the higher number of users. Evidently, there exists a strong correlation between profile size and matching accuracy (0.93).

Matching Users based on Username Regarding usernames (Table 1, username), Levenshtein and the Longest Common Subsequence (LCS) based distance perform fairly similar and outperform both Jaccard and Smith-Waterman distances as well as the ExactMatch baseline. Success rates (S@k) increase only slightly with increasing k, whereas they increase fairly substantially depending on k for matching based on users' tags. This is to be expected. User names tend to be much more unique than the tags assigned by users. For the best metric (LCS), string similarity works well for approximately 55% of the users but fails for the other 45%.

Matching Users based on Tags and Username When combining the best performing measures for the two types of profiles (Table 1, combined), i.e. BM25 with site specific IDF (tag-based) and LCS (username-based), we gain major improvements of 35% compared to the approaches that exploit just the tag-based profiles and of 8.9% compared to the username-based approaches. Furthermore, we analyzed how the user identification strategies perform for the different service settings: all approaches work best when comparing profiles from StumbleUpon and Delicious while they are less successful when Flickr is involved. Regarding matching based on tag-based profiles, this result is to be expected considering that the type of resources differ between these systems.However, a remarkable observation is that many users also tend to use similar usernames on StumbleUpon and Delicious as the success of the username-based

Table 1: Results based on user tags, username and mixture for Flickr, Delicious and StumbleUpon (FDS dataset); for FDS aggregated profiles; and for Flickr and Delicious (FD dataset). All improvements are significant ($p<0.05$, 2-tailed t-test)

		FDS				FDS-aggregation				FD			
Profile type	**Strategy**	**MRR**	**S@1**	**S@3**	**S@10**	**MRR**	**S@1**	**S@3**	**S@10**	**MRR**	**S@1**	**S@3**	**S@10**
tags	TF	0.181	0.126	0.180	0.278	-	-	-	-	0.108	0.070	0.110	0.178
	TFIDF	0.267	0.207	0.277	0.380	0.335	0.259	0.356	0.470	0.184	0.124	0.197	0.302
	BM25	0.301	0.242	0.317	0.405	0.391	0.326	0.414	0.505	0.259	0.204	0.274	0.370
	BM25 specific IDF	0.345	0.291	0.360	0.443	0.453	0.393	0.474	0.560	0.343	0.250	0.330	0.428
username	ExactMatch	0.387	0.372	0.375	0.375	0.555	0.542	0.547	0.547	0.555	0.542	0.547	0.547
	Jaccard	0.535	0.501	0.536	0.577	0.684	0.654	0.686	0.717	0.684	0.654	0.686	0.717
	SmithWaterman	0.462	0.357	0.475	0.607	0.437	0.217	0.476	0.747	0.437	0.217	0.476	0.747
	Levenshtein	0.574	0.552	0.572	0.591	0.721	0.701	0.722	0.735	0.571	0.459	0.496	0.524
	LCS	0.582	0.552	0.586	0.600	0.727	0.701	0.731	0.746	0.564	0.452	0.502	0.535
combined	Mixture	0.677	0.641	0.697	0.728	0.816	0.792	0.832	0.855	0.632	0.543	0.590	0.624

approach (LCS) is higher than 70%, whereas it is less successful ($S@1 < 50\%$) for Flickr profiles.

Using aggregated profiles Table 1 (FDS-aggregation) also compares the various approaches for matching aggregated profiles, i.e. the union of two individual profiles for a given user. We compare each aggregated profile with the remaining profile (e.g. Flickr-Delicious to StumbleUpon, Delicious-StumbleUpon to Flickr, etc.) and vice-versa. Regarding aggregated tag profiles, BM25 with site specific IDF leads again to a significant improvement over IDF and BM25 with global IDF. In summary, knowing more (tags) about a user improves user identification performance clearly. For example, S@1 improves from 0.291 to 0.393 for BM25 with site specific IDF.

Correspondingly, aggregated username-based profiles allow for improving the performance. For example, S@1 increases from 0.552 to 0.701 for the Levenshtein and LCS measures (Table 1, username, FDS-aggregation). Finally, for the mixture approach that combines the best tag- and username-based user identification strategies, we also see that having more user information increases the precision of the user identification challenge significantly. For the aggregated profiles, the mixture of the tag-based BM25 approach using site specific IDF and LCS for measuring similarity of usernames leads to the best performance of 0.816 and 0.792 regarding MRR and S@1 respectively. Moreover, it is interesting to see that for all settings where Flickr profiles are unified with Delicious or StumbleUpon profiles, the combination of tag- and username-based strategies achieves a success (S@1) of nearly 90%.

Conclusions and Future Work

In this paper we investigate whether users can be identified across Web platforms by analyzing their tagging practices. Therefore, we examine user profiles from three different social tagging services: Flickr, Delicious and StumbleUpon. We exploit implicit feedback (tagging behavior) as well as lightweight explicit profile information (usernames) to construct user profiles for identifying the users. In summary, we conclude that (1) it is possible to identify users across systems based on their tagging behavior even though the tagging behavior varies considerably between the analyzed systems, (2) for the user identification based on tag profiles our new approach of BM25 in combination with site specific IDF outperformed the other approaches significantly and (3) knowing more about the user (profile aggregation) and combining tag- and username-based approaches further improves the performance significantly to an accuracy of almost 80% and nearly 90% for specific settings.

While our user identification strategies can support cross-system personalization, they raise privacy concerns. For future work, we plan to study such privacy aspects in more detail. We will also investigate whether the consideration of network structure (such as friend links) in combination with tag-based profile features impacts user identification.

Acknowledgments The work was partially funded by the NTH (Niedersächsische Technische Hochschule) School for IT Ecosystems as well as the Crokodil project funded by the German Federal Ministry of Education and Research and the European Social Fund of the European Union (ESF).

References

Abel, F.; Henze, N.; Herder, E.; and Krause, D. 2010. Interweaving public user profiles on the web. In *Proc. UMAP*, 16–27.

Bischoff, K.; Firan, C. S.; Nejdl, W.; and Paiu, R. 2008. Can all tags be used for search? In *Proc. CIKM 2008*, 193–202.

Carmagnola, F., and Cena, F. 2009. User identification for cross-system personalisation. *Information Sciences: an International Journal* 179(1-2):16–32.

Elmagarmid, A. K.; Ipeirotis, P. G.; and Verykios, V. S. 2007. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on* 19(1):1–16.

Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M. 2006. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. Hypertext 2006*, 31–40. ACM.

Spärck Jones, K.; Walker, S.; and Robertson, S. E. 2000. A probabilistic model of information retrieval: development and comparative experiments. parts 1 and 2. *Information Processing and Management* 36:779–840.

Szomszor, M.; Cantador, I.; and Alani, H. 2008. Correlating user profiles from multiple folksonomies. In *Proc. Hypertext*, 33–42.

Vosecky, J.; Hong, D.; and Shen, V. Y. 2009. User identification across multiple social networks. In *Int. Conference on Networked Digital Technologies (NDT '09)*, 360 –365.

Zafarani, R., and Liu, H. 2009. Connecting corresponding identities across communities. In *Proc. ICWSM*, 354–357.