Connecting Corresponding Identities across Communities

Reza Zafarani and Huan Liu
Department of Computer Science and Engineering, Arizona State University, USA
{reza@asu.edu, huanliu@asu.edu}

Motivation

Problem: a well known barrier in cross-community (multiple website) analysis is the disconnectedness of the websites.

- **No Real Identity:** communities preserve the anonymity of users by allowing them to freely select usernames instead of their real identities .
- **Different Authentication Systems:** different websites employ different username and authentication systems.
- **No Single-Sign-On:** communities rarely share Single-Sign-On procedures, where users can logon to different communities using a single username (e.g., as in Orkut and YouTube).

Definitions and Experiment Setting

Cross-Community Corresponding Username Elicitation: given a username-community pair $< u_1, c_1 >$, called base-username and base-community, and a community c_2 (target community), a solution to the cross community corresponding username elicitation problem is a username $u_2 \in \prod_{c_2}$, called the target-username, such that

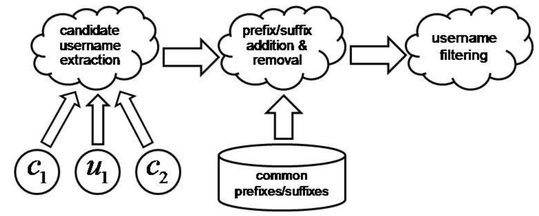
$$U^{-1}(u_1, c_1) = U^{-1}(u_2, c_2)$$

- BlogCatalog was used in our experiment. Users in BlogCatalog are provided with a feature called “My Communities”. This feature enables users to list their usernames in other communities.
- Overall, 38,093 username-username pairs were gathered. Each pair consists of the username in the BlogCatalog community and the corresponding username in another community
- Besides BlogCatalog, the dataset contains usernames from 36 different communities.

Hypotheses

1. **(100% accurate) There exist URLs containing the usernames (profiles)**
   i.e., http://www.myspace.com/**test**
2. **(100% accurate) Given a target community, you can easily find the community’s domain.**
   i.e., query: “communityName”
3. **(81% accurate) You can easily find profile WebPages using web search.**
   i.e., query: “site:communityNameDomain username”
4. **(68% accurate) For any two usernames of the same person, there is a high chance of co-occurrence of these two in search engine results.**
   i.e., query: “username1 username2”
5. **(38% accurate) For any two usernames of the same person, there is a high chance of occurrence of one them on the profile webpage of the other.**
   i.e., query: “baseUsername site:targetCommunityDomain targetUsername”
6. **(66% accurate) Users use the same usernames.**
   i.e., John.smith on Flickr, then john.smith on MySpace
7. **(83% accurate) Users use one of their many usernames.**
   i.e., either john.smith on j.smith on Flickr.

Approach



- **Candidate Username Extraction:** Query “base-username”, i.e., u_1 and then Keyword extraction on the URLs retrieved.
- **Prefixes** {the, i, b, iam, my, free, happy, dr, x, mister, coach,…}
- **Suffixes** {f1, 2, s, dotcom, b, blog, 7, 07, 77, 13, a, z, 66, 0, 50, 08,..}
- **Add/Removal**: John.smith -> thejohn.smith, theone -> one.
- **Filter candidate set.**

Evaluation and Conclusions

	Del.icio.us	Digg	Flickr	Furl	Last.fm	Multiply	MyBlogLog	MySpace	Reddit	StumbleUpon	Technorati	Twitter	YouTube
Del.icio.us	1	0.68	**0.66**	0.84	0.76	0.62	**0.73**	0.47	0.9	0.72	0.78	0.76	0.58
Digg	0.7	1	0.57	0.78	0.82	0.54	0.63	0.4	0.84	0.62	0.68	0.64	0.54
Flickr	0.66	0.64	1	0.66	0.71	0.45	0.51	**0.58**	0.56	0.63	0.59	0.65	0.6
Furl	0.78	0.76	0.63	1	**0.88**	0.74	**0.73**	0.45	**0.92**	0.78	**0.82**	0.76	0.6
Last.fm	0.74	0.78	0.6	0.82	1	0.64	0.64	0.53	0.72	0.64	0.72	0.64	0.54
MyBlogLog	0.71	0.67	0.47	0.63	0.66	0.46	1	0.35	0.71	0.6	0.67	0.67	0.47
MySpace	0.57	0.56	0.54	0.61	0.57	0.49	0.56	1	0.57	0.52	0.53	0.53	0.58
Reddit	**0.84**	**0.8**	0.54	**0.86**	0.68	**0.78**	0.67	0.43	1	**0.8**	0.76	**0.77**	**0.62**
StumbleUpon	0.74	0.68	0.5	0.78	0.68	0.6	0.62	0.38	0.86	1	0.66	0.6	0.58
Technorati	0.74	0.66	0.5	0.8	0.72	0.48	0.65	0.4	0.78	0.64	1	0.66	0.58
Twitter	0.64	0.64	0.53	0.68	0.7	0.53	0.65	0.33	0.81	0.58	0.62	1	0.52
YouTube	0.58	0.6	0.58	0.6	0.68	0.52	0.55	0.56	0.67	0.6	0.68	0.62	1

- On average, our method predicted the correct target-username in more than 66% of the cases and is up to 92% accurate in the best case scenario.
- Usernames can be used quite successfully to identify corresponding usernames in various communities.

Problem → Hypothesize and Validate → Approach → Evaluation

This work is, in part, sponsored by AFOSR Grant FA95500810132.

Data Mining and Machine Learning Lab

ASU Arizona State University