Linking Organizational Social Network Profiles

Jerome Cheng[1] Kazunari Sugiyama[1] Min-Yen Kan[1,2]
[1]School of Computing, National University of Singapore, Singapore
[2]Interactive and Digital Media Institute, National University of Singapore, Singapore
jerome@ayulin.net
{sugiyama,kanmy}@comp.nus.edu.sg

ABSTRACT

Many organizations possess social media accounts on different social networks, but these profiles are not always linked. End applications, users, as well as the organization themselves, can benefit when the profiles are appropriately identified and linked. Most existing works on social network entity linking focus on linking individuals, and do not model features specific for organizational linking. We address this gap not only to link official social media accounts but also to discover and solve the identification and linking of associated *affiliate* accounts – such as geographical divisions and brands – which are important to distinguish.

We instantiate our method for classifying profiles on social network services for Twitter and Facebook, which major organizations use. We classify profiles as to whether they belong to an organization or its affiliates. Our best classifier achieves an accuracy of 0.976 on average in both datasets, significantly improving baselines that exploit the features used in state-of-the-art comparable user linkage strategies.

CCS Concepts

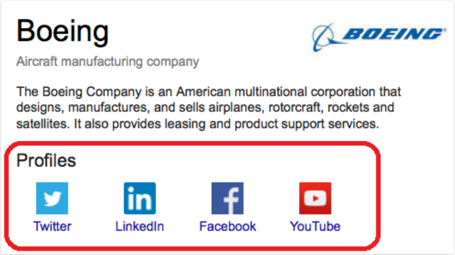
•**Information systems** → **Web and social media search;** *Entity resolution;* •**Human-centered computing** → *Social networking sites;*

Keywords

Organizational social profiles, Organization entity profiling, Record linkage, Social networks

1. INTRODUCTION

With the pervasiveness of social network services (SNSs), organizations also leverage them to reach out to and keep tabs on their fans and potential customers. However, unlike individual users, organizations usually maintain multiple profiles, even within a single SNS, to distribute targeted messages. This kind of intra-network profiles often represent *affiliates* – entities such as brands or geographical divisions which are part of the organization but is not the whole organization. An important task for both users and organizations is to automatically identify such profiles and link them to get a holistic social media profile of an organization, but to the best of our knowledge, up to now, this problem has been completely ignored by the community.



Boeing

Aircraft manufacturing company

The Boeing Company is an American multinational corporation that designs, manufactures, and sells airplanes, rotorcraft, rockets and satellites. It also provides leasing and product support services.

Profiles

Figure 1: Google's manual organization profile linkage for "Boeing" in search results (captured May 2015).

Individual users do follow organizations through these profiles to stay informed of their products, events and developments. However, they do not have ways to easily identify all such profiles that belong to these organizations. Search engines of most SNSs currently do not distinguish between individual user profiles and organizational ones. While some Web search engines do provide links to profiles in their search results, these links still rely on manual effort by the organization (*i.e.*, Google requires organizations to include specific markup in their Web page[1]; Figure 1). As such, one may think that the organization itself should be responsible to link their SNS account on their websites. Our experience shows that this is not always the case, possibly due to challenges in organizational knowledge management or ignorance. Finally, to the best of our knowledge, services that aggregate locations where we can find such comprehensive links do not currently exist: we can identify only an organization's main profile on the most popular SNS, but not identify its affiliate profiles.

Organizations would benefit from automatic profile linkage to help monitor their competitors' profiles and activities – a fact that has been proven by the existence of nascent start-ups[2]. In addition, organizations also face the problem of tracking and policing attempts by others to impersonate them or infringe upon their trademarks. For instance, *Jetstar Airways* fell victim to impersonation, with a "Jetstar Australia" profile created on Facebook. This is used to insult customers who mistakenly provided malicious feedback through it[3].

SIGIR '16, July 17–21, 2016, Pisa, Italy.
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2911451.2914698

[1] "Specify your social profiles to Google" in Google Developers. Accessed: Feb 12th, 2016, from https://developers.google.com/structured-data/customize/social-profiles

[2] e.g., http://nexgate.com

[3] http://bit.ly/1Eyruke

To solve these problems, it is important to identify an organization's social networking profiles on a specific network. We address this task by defining three categories of profiles related to a given organization:

Official. Profiles that represent the organization as a whole. For instance, Microsoft Corporation's official profile on Twitter is `@Microsoft`.

Affiliate. Profiles that represent one of the organization's brands, or a division. For example, Microsoft has several divisional profiles such as `@MicrosoftDesign`, `@MicrosoftAsia`, and so on.

Unrelated. These profiles are not run by the organization in question, belonging to an individual or other organization. We note that a profile unrelated to one organization may be an official or affiliate profile for another.

The contributions in our work are: 1) An implemented, automated system to perform organization profile linkage on social networks – given an organization's name as a query, it first searches for profile candidates from the target social networks, then constructs a model by using supervised learning to classify each candidate into **Official**, **Affiliate**, or **Unrelated** for the query; 2) The creation of a publicly available dataset that consists of 6,784 profiles from about 220 real-world organizations across Twitter and Facebook, where the profiles are manually classified into the above three categories.

2. RELATED WORK

The name fields of a profile have been shown to be of high importance in linking users. Two kinds of proposed linkage methods rely on usernames: either through generating candidate usernames to match with [7], or by examining behavioural patterns that influence the way usernames are created [8]. Another study found that the user and display names are the "most discriminative features" [3] when disambiguating profiles.

Several works have examined user activities to identify a profile's holder. These focused on using profile data and the content created by the user [1], or combining the information with other metadata (*e.g.,* location or time [2]). Their findings are helpful for our work as organizations often "cross-post" the same (or similar) material across different social networks.

Works have addressed social network profiles in the context of security – re-identification for anonymized social networks [4], spam bot detection [6], malicious profile detection [5] – but these are beyond the scope of our current review due to space considerations.

The aforementioned studies have identified profiles on social networks, to the best of our knowledge, they have all focused on profiles belonging to individual users and do not consider the specific properties of organization profiles, and are thus insufficient for linking organizations. For example, organizations often use an abbreviated form of their names online (*e.g.,* General Motors uses just "GM"), and their names often contain some common words such as "Company" or "Corporation." To accurately link organization profiles, it is necessary to take these properties into account.

3. FEATURES FOR CLASSIFIERS

We adopt supervised learning to classify organization names into the three categories previously defined. We extract 14 features that belong to the following four broad feature sets, annotating novel, specific features we introduce for this organization linkage problem with an asterisk (*):

Baseline Features (BL). These features examine the relationship between the target organization's name (hereafter, "query"), and the two name fields on a profile: the handle and the display names. The handle name is sometimes known as a username, and is a unique identifier that represents the profile on the social network. On the other hand, the display name is a free-text string that provides a more user-friendly name, but is not unique. For example, on *General Motors'* Twitter profile[4], the username is **GM** while the display name is **General Motors**. We adopt the following as baseline features:

- NED between the query and target handle name*
- NED between the query and target display name*

Normalized edit distance (NED) gives a relative measure of similarity accounting for their lengths. Let $ED(s_1, s_2)$ be edit distance between two strings s_1 and s_2. Then, normalized edit distance $NED(s_1, s_2)$ between them is defined as follows:

$$NED(s_1, s_2) = 1 - \frac{ED(s_1, s_2)}{\max(len(s_1), len(s_2))},$$

where $len(s_i)$ denotes the length of string s_i $(i = 1, 2)$.

In these features, we first process the query to generate a simple abbreviation and remove stop words (*e.g.,* "Company," "Corporation," and so on) defined for this task before calculating the edit distances between both the processed string and the original query. These features best simulate how a human would approach this problem.

Name-based Features (N). Name string lengths are also useful. We term these as name-based features:

- Length of the query
- Length of the target display name*
- Length of the target handle name*

Description-based Features (D). These features are based on the description field of the organization's profile. Besides the relationship between the description and query, we further searched for the organization's description from DuckDuckGo[5], a search engine that provides the results from sources such as Wikipedia.

- Cosine similarity between the target profile's description and the query
- Number of occurrences of the query in the target profile's description*
- Cosine similarity between the target profile's description and DuckDuckGo description*

The first two features give measures of the extent to which the profile's description is relevant to the organization's name. The third yields a similarity score between the profile's description and organization descriptions provided by the DuckDuckGo general search engine, which we deem as a more general, third-party description. Note that this feature requires an external API call.

Content-based Features (C). Posted content is also evidence for linkage. We construct two bigram language models for each category's training data (Official, Affiliate, or Unrelated): one from the profile descriptions in the target class, and the other from the textual content of recent 20 posts by the target profiles. We use add-one smoothing as a simple mechanism to address zero occurring bigrams in the query. This process yields six models in total that capture writing style differences:

- Probability that the query appears in bigram models constructed from official/affiliate/unrelated description*
- Probability that the query appears in bigram models constructed from official/affiliate/unrelated posted content*

[4] https://www.twitter.com/GM

[5] https://api.duckduckgo.com

Table 1: Statistics on our Twitter and Facebook dataset.

SNS	Organizations	Official	Affiliate	Unrelated	Total
Twitter	228	232	675	2,474	3,381
Facebook	216	145	491	2,767	3,403

4. EXPERIMENTS

We first construct the dataset for our experiments. We search for organization names which are listed on Freebase[6] as having at least one social network presence, and then submit the each organization name as a query to Twitter and Facebook. Finally, we manually label the profiles that Twitter and Facebook return. We obtain 3,381 profiles from 228 organizations and 3,403 profiles from 216 organizations on Twitter and Facebook, respectively (Table 1 shows relevant statistics). Note that the number of organizations included in each dataset differs from the number of official profiles that we found, as some organizations either lack an official profile or have more than one. To encourage the community to work on this important and understudied research work on social networks, we have made our entire dataset publicly available[7].

To study linkage performance across classification methods, we employ several different classifiers, namely: Bernoulli naïve Bayes (BNB), Gaussian naïve Bayes (GNB), decision tree (DT), logistic regression (LR), random forest (RF), support vector machine (SVM), and maximum entropy (ME), all which are implemented in *scikit-learn*[8]. We perform 10-fold cross validation, evaluating each classifier with their accuracy on the official and affiliate classes.

We compare the accuracies of classifiers constructed by combining "Baseline (BL)," "Name-based (N)," "Description-based (D)," and "Content-based (C)" features described earlier (Table 2) . In user identification works across SNSs, Iofciu *et al.* [1] reported that longest common subsequence-based distance is an effective feature in their "user maching based on username" experiments. In a related vein, Zafarani and Liu [8] disclosed their top 10 most important features, most of which are length-based ones. We also compare our classifiers with those constructed using the features from both [1] and [8].

5. DISCUSSION

In this section, we refer to "Twitter (Official)," "Twitter (Affiliate)," "Facebook (Official)," "Facebook (Affiliate)" as "TW-Off," "TW-Aff," "FB-Off," "FB-Aff," respectively.

Classification Accuracies. In both Twitter and Facebook datasets, DT, LR, RF, SVM, and ME achieve accuracies over 0.7 in almost all feature sets (Table 2). Naïve Bayes in both forms (BNB and GNB) yields lower accuracies (only about 0.2 to 0.7), especially in the Affiliate category, even when adding more features.

Among the classifiers, Random Forest (RF) achieves the best in both Twitter and Facebook datasets (0.983, 0.973, 0.972, 0.977 in TW-Off, TW-Aff, FB-Off, FB-Aff, respectively). RF is an ensemble approach, taking an average of multiple decision trees' decisions, which are separately trained on different parts of the same training set, to reduce overfitting. While this step slightly increases the bias of the forest, it decreases its variance. This framework of RF works well for our task as each of our features tends to have large variance.

With respect to features, we observe that using all features yields the best accuracy in each classifier, and also outperforms the accuracies obtained by the features used in [1] and [8] with statistical significance. This indicates that our ALL feature set is more effective than longest common subsequence-based distance in [1] and length-based features in [8].

With respect to individual feature contributions, we further study the improvement when adding N, D, or C feature set to the baseline (Table 3). We observe higher improvement rate on affiliate profiles in both Twitter and Facebook datasets (N: 15.02%, D: 7.57%, C:16.49% in TW-Aff and N: 7.61%, D: 4.54%, C:22.88% in FB-Aff). Among the feature sets, C is the most effective feature set to distinguish official and affiliate profiles (9.56%, 16.49%, 7.49%, 22.88% in TW-Off, TW-Aff, FB-Off, and FB-Aff, respectively). This indicates that content information extracted from profile description and recent posts do contain useful signal that can help distinguish organizational names. Although acquiring such data can be expensive, our experiments validates its significant impact on the linkage accuracy.

Linking Affiliates. The existence of affiliate profiles is a distinguishing characteristic of our problem, differentiating it from general linkage of individual profiles, or even linking official profiles alone. Both individuals and official profiles have names which are generally close to that of their holder, and as such linkage of affiliates would be expected to be more difficult. However, RF, ME, and SVM classifiers all prove to be good and equally adept at detecting both official and affiliate profiles, countering this intuition. Investigating this more fully reveals that this actually does not hold with certain subclasses of affiliate profiles – while geographic affiliates do tend to contain their parent company's name with a location as a suffix as their profile name (*e.g.,* "Microsoft Asia"), brand names often have little in common (*e.g.,* "Lenovo" and "ThinkPad"). Hence, the use of the parent company name for these brand profiles is insufficient to identify them as belonging to the organization; additional information is required. These classifiers, with the equipped features from our study, still significantly outperform the baseline and comparative methods.

Network Idiosyncrasies. On Twitter, all users are given the same type of profiles regardless of whether they are a person or an organization. On the other hand, Facebook has multiple profile types including *Pages*, which are designed for "artists, public figures, businesses, brands, organizations, and non-profits[9]."

We considered only Pages here, effectively resulting in a level of filtering provided by Facebook which Twitter does not have. Pages with handle and display names similar to those of an organization name are likely to belong to the organization.

However, this observation is not the case with affiliate profiles on Facebook. Facebook does not require pages that specify a unique handle name; in this scenario, we instead treat the page's numerical ID as its handle name. This substitution greatly impacts on the normalized edit distance scores in BL. It is sometimes the case that an affiliate profile does not have a specific handle name – for instance, *Netflix* uses "Netflix" as its official profile's handle name[10], but does not provide one for its Latin American affiliate[11].

In addition, Facebook also generates "topic pages" from Wikipedia articles, giving them a display name from the title of that article but no handle name. For example, a page exists for *Samsung* which has just "Samsung" as its display name[12]. We consider these pages to

[6] http://www.freebase.com

[7] http://wing.comp.nus.edu.sg/downloads/corpsearch/OrgSocialNetworkData.html

[8] http://www.scikit-learn.org (Version 0.15.0)

[9] "Products" in Facebook Newsroom. Accessed Feb 12th, 2016, from https://newsroom.fb.com/products/

[10] https://www.facebook.com/netflix

[11] https://www.facebook.com/pages/Netflix-Latinoam\%C3\%A9rica/553454298124413

[12] https://www.facebook.com/pages/Samsung/114938905184804

Table 2: Accuracies obtained by several classifiers constructed using various feature sets on Twitter ((a-1) Official, (a-2) Affiliate)) and Facebook ((b-1) Official, (b-2) Affiliate)) datasets. The best accuracy in each category is denoted with bold font. "" and "*" denote the difference between the best results obtained by the most effective feature set in [1] or [8] (underlined scores) and the best result obtained by each classifier (italic or bold scores) in our approach is significant for $p < 0.01$ and $p < 0.05$, respectively.**

(a-1) Twitter (Official)

Feature set	Classifier						
	BNB	GNB	DT	LR	RF	SVM	ME
Baseline (BL)	0.787	0.405	0.870	0.720	0.801	0.893	0.904
BL+N	0.911	0.410	0.919	0.842	0.923	0.921	0.943
BL+D	0.891	0.437	0.918	0.785	0.940	0.851	0.908
BL+C	0.894	0.461	0.929	0.820	0.961	0.944	0.914
BL+N+D	0.916	0.407	0.926	0.846	0.972	0.845	0.945
BL+N+C	0.931	0.578	0.942	0.951	0.976	0.956	0.949
BL+D+C	0.878	0.498	0.932	0.842	0.972	0.936	0.945
ALL (BL+N+D+C)	*0.934**	*0.645***	*0.952**	*0.955**	**0.983***	*0.963**	*0.954**
Iofcia et al. [1]	0.853	0.386	0.919	0.774	0.940	0.901	0.938
Zafarani and Liu [8]	0.846	0.414	0.929	0.863	0.967	0.869	0.926

(a-2) Twitter (Affiliate)

Feature set	Classifier						
	BNB	GNB	DT	LR	RF	SVM	ME
Baseline (BL)	0.235	0.413	0.788	0.681	0.740	0.828	0.878
BL+N	0.325	0.511	0.924	0.521	0.931	0.835	0.872
BL+D	0.140	0.501	0.922	0.531	0.935	0.798	0.846
BL+C	0.327	0.432	0.502	0.837	0.944	0.878	0.914
BL+N+D	0.324	0.468	0.917	0.528	0.932	0.805	0.870
BL+N+C	0.330	0.556	0.928	0.953	0.949	0.963	0.933
BL+D+C	0.322	0.515	0.919	0.831	0.937	0.848	0.898
ALL (BL+N+D+C)	*0.339**	*0.639**	*0.934**	*0.971**	**0.973***	*0.967**	*0.947**
Iofcia et al. [1]	0.321	0.499	0.927	0.535	0.931	0.824	0.873
Zafarani and Liu [8]	0.327	0.536	0.917	0.543	0.926	0.815	0.865

(b-1) Facebook (Official)

Feature set	Classifier						
	BNB	GNB	DT	LR	RF	SVM	ME
Baseline (BL)	0.802	0.502	0.827	0.816	0.794	0.816	0.959
BL+N	0.815	0.414	0.874	0.839	0.913	0.930	0.956
BL+D	0.738	0.401	0.905	0.739	0.937	0.939	0.963
BL+C	0.740	0.462	0.874	0.830	0.945	0.913	0.960
BL+N+D	0.755	0.344	0.901	0.860	0.951	0.948	0.955
BL+N+C	0.831	0.564	0.913	0.866	0.957	0.950	0.967
BL+D+C	0.699	0.412	0.897	0.821	0.946	0.921	0.958
ALL (BL+N+D+C)	*0.897**	*0.684***	*0.919**	*0.878**	**0.972***	*0.960**	*0.970**
Iofcia et al. [1]	0.815	0.460	0.881	0.797	0.917	0.948	0.955
Zafarani and Liu [8]	0.743	0.467	0.893	0.814	0.929	0.949	0.966

(b-2) Facebook (Affiliate)

Feature set	Classifier						
	BNB	GNB	DT	LR	RF	SVM	ME
Baseline (BL)	0.283	0.317	0.560	0.289	0.790	0.836	0.903
BL+N	0.309	0.327	0.685	0.385	0.859	0.847	0.911
BL+D	0.397	0.266	0.689	0.486	0.847	0.750	0.858
BL+C	0.359	0.308	0.699	0.703	0.962	0.796	0.942
BL+N+D	0.305	0.272	0.692	0.470	0.952	0.743	0.910
BL+N+C	0.404	0.360	0.704	0.712	0.966	0.927	0.953
BL+D+C	0.331	0.304	0.687	0.693	0.932	0.898	0.949
ALL (BL+N+D+C)	*0.443**	*0.399**	*0.727**	*0.734***	**0.977***	*0.964**	*0.968**
Iofcia et al. [1]	0.355	0.342	0.665	0.424	0.762	0.806	0.874
Zafarani and Liu [8]	0.359	0.314	0.684	0.439	0.743	0.802	0.869

Table 3: Improvement rate by adding N, D, or C feature set to baseline features (BL).

	Twitter (Official)	Twitter (Affiliate)	Facebook (Official)	Facebook (Affiliate)
N	6.56%	15.02%	6.72%	7.61%
D	2.39%	7.57%	0.46%	4.54%
C	**9.56%**	**16.49%**	**7.49%**	**22.88%**

be unrelated. However, they have a similar edit distance as legitimate affiliate profiles. This makes it difficult to distinguish an affiliate profile on Facebook using just the handle and display names.

As we go forward, it is likely that other social networks will exhibit their own unique set of quirks, which will similarly pose challenges to linking profiles on them. We recommend that the reader be mindful of these idiosyncrasies when adapting our work for linkage with other SNSs.

6. CONCLUSION

We have proposed an approach to organizational social profile linkage. To the best of our knowledge, this problem has not been tackled before in the literature, but is a problem of growing importance given the amount of influence that social media plays in lives today. A key difference in organization linkage is the presence of affiliates – both geographic and brand – that change the nature of the problem.

While previous efforts have done linkage through manual means, our automated approach finds that profile description and posted content are the most effective features for detecting linkage. In future work, we plan to leverage unique characteristics of the different types of affiliate accounts to achieve better classification accuracy. This will allow us to infer affiliate structure (with respect to the parent company), as well as to expand our work to include other social networks.

7. REFERENCES

[1] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying Users Across Social Tagging Systems. In *Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, pages 522–525, 2011.

[2] X. Kong, J. Zhang, and P. S. Yu. Inferring Anchor Links across Multiple Heterogeneous Social Networks. In *Proc. of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*, pages 179–188, 2013.

[3] A. Malhotra, L. C. Totti, W. Meira Jr, P. Kumaraguru, and V. Almeida. Studying User Footprints in Different Online Social Networks. In *Proc. of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1065–1070, 2012.

[4] A. Narayanan and V. Shamatikov. De-anonymizing Social Networks. In *Proc. of the 30th IEEE Symposium on Security and Privacy (SP 2009)*, pages 173–187, 2009.

[5] M. Singh, D. Bansal, and S. Sofat. Detecting Malicious Users in Twitter using Classifiers. In *Proc. of the 7th International Conference on Security of Information and Networks (SIN'14)*, pages 247–253, 2014.

[6] A. H. Wang. Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. In *Proc. of the 24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy (DBSec'10)*, pages 335–342, 2010.

[7] R. Zafarani and H. Liu. Connecting Corresponding Identities across Communities. In *Proc. of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM-09)*, pages 354–357, 2009.

[8] R. Zafarani and H. Liu. Connecting Users across Social Media Sites: A Behavioral-Modeling Approach. In *Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*, pages 41–49, 2013.