Linking Organizational Social Network Profiles

NUS National University of Singapore
School of Computing

Jerome Cheng, Kazunari Sugiyama, and Min-Yen Kan
jerome@ayulin.net
{sugiyama, kanmy}@comp.nus.edu.sg

1. Introduction

Social Network Entity Linking

(1) Individual Users

Most of existing works address "individual user profile" across social media

[Iofciu et al., ICWSM-11]

Flickr, Delicious

[Zafarani and Liu, KDD'13]

Flickr, Reddit, StumbleUpon, YouTube

- In both works, longest common subsequence between query (user name) and candidate works well.

(2) Organizations

- Search engines of most of SNSs do not distinguish between individual user profiles and organizational ones.
- Google requires organizations to include specific markup in their Web page.

Our Goal

- To identify an organizational social network profiles on a specific network.
- We define the following three categories, and then adopt supervised learning:

Official
e.g.) @Microsoft on

Affiliate
e.g.) @MicrosoftDesign on
@MicrosoftAsia on

Unrelated

Classifiers

- Bernoulli naïve Bayes
- Gaussian naïve Bayes
- Decision tree (DT)
- Logistic regression (LR)
- **Random forest (RF)**
- **Support vector machines (SVM)**
- **Maximum entropy (ME)**

Dataset

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

(We have released our dataset at http://wing.comp.nus.edu.sg/downloads/corpsearch/OrgSocialNetworkData.html)

2. Features for Classifiers

Baseline Features (BL)

Normalized edit distance between
- the query (organization's name) and handle name*
- query and display name*

* e.g.) "General Motors" { handle name: "GM" / display name: "General Motors"

Name-based Features (N)

Length of
- the query, - target handle name, and - target display name

Description-based Features (D)

We get organization's description from DuckDuckGo, a search engine that provides the results from 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.

Content-based Features (C)

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

3. Experimental Results (Selected)

Twitter (Affiliate)

	RF	SVM	ME
Baseline (BL)	0.740	0.828	0.878
BL+N	0.931	0.835	0.872
BL+D	0.935	0.798	0.846
BL+C	0.944	0.878	0.914
BL+N+D	0.932	0.805	0.870
BL+N+C	0.949	0.963	0.933
BL+D+C	0.937	0.848	0.898
ALL (BL+N+D+C)	**0.973***	*0.967**	*0.947**
Iofciu et al.	0.931	0.824	0.873
Zafarani and Liu	0.926	0.815	0.865

Facebook (Affiliate)

	RF	SVM	ME
Baseline (BL)	0.790	0.836	0.903
BL+N	0.859	0.847	0.911
BL+D	0.847	0.750	0.858
BL+C	0.962	0.796	0.942
BL+N+D	0.952	0.743	0.910
BL+N+C	0.966	0.927	0.953
BL+D+C	0.932	0.898	0.949
ALL (BL+N+D+C)	**0.977***	*0.964**	*0.968**
Iofciu et al.	0.762	0.806	0.874
Zafarani and Liu	0.743	0.802	0.869

Improvement Rate

	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%**