Towards Alias Detection Without String Similarity: an Active Learning based Approach

Lili Jiang[1], Jianyong Wang[2], Ping Luo[3], Ning An[4], Min Wang[3]
[1]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]HP Labs China, Beijing, China
[4]Anhui Province Key Laboratory of Affective Computing And Advanced Intelligent Machine,
School of Computer and Information, Hefei University of Technology, Hefei, China
jianglili06@lzu.cn;
jianyong@tsinghua.edu.cn;{ping.luo, min.wang6}@hp.com;ning.an@hfut.edu.cn

ABSTRACT

Entity aliases commonly exist and accurately detecting these aliases plays a vital role in various applications. In this paper, we use an active-learning-based method to detect entity aliases without string similarity. To minimize the cost on pairwise comparison, a subset-based method restricts the alias selection within a small-scale entity set. Within each generated entity set, an active learning based logistic regression classifier is employed to predict whether a candidate is the alias of a given entity. The experimental results on three datasets clearly demonstrate that our proposed approach can effectively detect this kind of entity aliases.

Categories and Subject Descriptors: H.4[Information Systems Applications]:General

General Terms: Algorithms, Experiment, Measurement

Keywords: Alias Detection, Pairwise Comparison, Active Learning

1. INTRODUCTION

Solving the problem of alias detection is important for a large number of applications including entity identification, terrorist detection, and social network analysis. Some aliases can be detected through string similarity measures, such as "World Trade Organization" and its alias "WTO". While another aliases have a quite low string similarity with their original entities, such as "Bill Clinton" and its nickname "Slick Willie". Detecting the second type of entity aliases turns out to be the aim of this paper.

Some previous work [1][2][4][5] has investigated this problem. These studies hold the same goal with ours but focus on a special domain respectively. This paper tries to detect entity aliases regardless of domains. To solve this problem, we employ an active-learning-based method taking three issues into consideration: picking the potential alias candidates for each given entity, choosing features and labeled samples to train a high-quality classifier. The contributions of this paper include: 1) a subset-based method to reduce the cost of pairwise comparison when picking alias candidates; 2) an active learning method, which selects informative samples in training classifier; 3) experiments on three types of datasets and comparison with other four baseline methods.

2. THE OVERALL FRAMEWORK

Given a document corpus D and an entity e, the task of alias detection is to extract all the entities from D which denote the same real-world object with e. Firstly, since the given entity may have no/(quite low) string similarity with its aliases, and especially certain type of aliases (e.g., terrorist or ghostwriter) are intentionally hidden from their real identities, the commonly used rules[1] (e.g., "aka", "as well known as", and "also called") will not always work. Secondly, it is difficult to locate entity aliases, since the number of aliases of an entity is rather fewer compared to the millions of strings/entities in the document corpus, let alone the increasing volume of internet. Thirdly, data labeling for training a classifier is time-consuming. To address these challenges, we propose a solution depicted in Figure 1 and describe it in the following subsections.

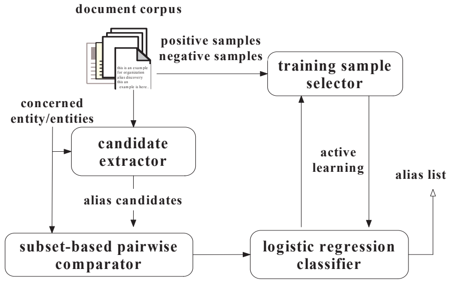

Figure 1: Entity Alias Detection Framework

2.1 The Subset-based Pairwise Comparison

Given a document corpus D and some entities to be resolved, the *candidate extractor* firstly employs a NLP tool (http://alias-i.com/lingpipe/) to extract all the named entities from D as candidates. Considering the first and second challenges mentioned above, the cost for pairwise comparison between entities and their alias candidates is non-trivial, which requires solutions to reduce the cost. The *subset-based pairwise comparator* in Figure 1 is designed to narrow the pairwise comparison scope for each given entity.

It is motivated by the observations that among the entities which are aliases with each other: 1) some ones frequently occur in documents, while others are mentioned in-

SIGIR August 12–16, 2012, Portland, Oregon, USA.
ACM ACM 978-1-4503-1472-5/12/08.

frequently; 2) the document union, where some of the most frequent entities appear, is often the superset of the documents union of the infrequent entities.

Based on these observations, the *subset-based pairwise comparator* is trying to keep each entity and its aliases in the same subset and it runs as follows: 1) builds an entity-document index and extracts the top $N\%$ most frequently used entities in the documents, and 2) for any two entities in the top $N\%$ ones and their corresponding occurring document sets, D_1 and D_2. If D_1 is the subset of D_2, D_1 will be removed. And when the intersection ratio between D_1 and D_2 is larger than a predefined threshold α, they are merged. After such iterations, we get a list of document sets. Finally, all the entities appearing in the documents of the same set are formed as an entity subset. Here, two parameters, α and N are trained using the 10 cross-validation, among which α denotes the entity overlapping ratio deciding whether two document sets are merged.

2.2 Active Learning for Alias Detection

For each given entity, the entities in the same entity subset are its alias candidates. We use a *logistic regression classifier* to output two values, among which p_1 denotes the probability that they are aliases and p_2 denotes the opposite. When p_1 is larger than p_2, we will conclude they may denote the same object. The classifier training is carried out in combination of active user learning, which tries to achieve higher performance with fewer training labels. The *training samples selector* randomly selects a subset of U and get labels for each sample, and then iteratively adds the labeled samples with high uncertainty[6] to the training set and trains a classifier until the trainers satisfy or U is null. The used features in the classifier training are presented as follows.

Co-occurrence relevance Pointwise mutual information(PMI) is used to measure the co-occurrence relevance in the corpus D between entities e_i and e_j, $PMI(e_i, e_j) = log$ $(p(e_i, e_j)/(p(e_i)p(e_j)))$. Herein, $p(e_i) = |D_{e_i}|/|D|$ where D_{e_i} is the occurring document set of e_i. $p(e_i, e_j)$ denotes the co-occurrence probability of e_i and e_j, calculated through dividing the co-occurring document count by the size of document union they appear respectively.

Social relevance Entities can be connected as a network based on their co-occurrence. For example, organization o and its aliases may co-occur with the persons affiliating with o. Thus, the social relevance between e_i and e_j is $CF(e_i, e_j) = (F(e_i) \cap F(e_j))/(F(e_i) \cup F(e_j))$ where $F(e_i)$ is the number of common nodes with e_i in the entity network.

Topic relevance We build an entity-document matrix M and each element in M is computed as $TFIDF(e_i, d_k) = (N^{e_i}_{d_k}/N^{*}_{d_k}) * \log(|D|/|D_{e_i}|)$. Herein, $N^{e_i}_{d_k}$ is the frequency of e_i appearing in document d_k, $N^{*}_{d_k}$ is the total entity count in document d_k, $|D|$ denotes the overall document count, and $|D_{e_i}|$ denotes the number of documents in which entity e_i appears. The similarity between any two entities can be calculated as $LSA(e_i, e_j) = cosine(V(e_i), V(e_j))$, where $V(e_i)$ is the entity-document frequency vector of e_i.

3. EXPERIMENTAL STUDY

We use three datasets to evaluate the proposed approach and compare it with other four methods in terms of F_1. The first dataset[1] contains 50 person names, 50 location names, and their aliases. To extend this dataset, we collected Web pages through issuing these names/aliases as queries to Google. The second dataset[3] contains terrorism data and spam emails. The third dataset was collected from an IT company including about 100k documents, 300k employee names and emails. As each email address can uniquely match one person, thus we regard an email as an alias of the person. We divided these datasets into training dataset and test dataset. The training data is used for parameter setting and classifier training, while the test dataset is used for evaluation. All ground-truth for evaluation is given in these datasets.

Table 1: Efficiency of subset-based method

	Pairwise Comparison Count	
Datasets/Methods	*no_subset*	*subset*
person/location	112, 020	20, 940
terrorism/spam	246, 030	56, 856
person/email	955, 631	187, 433

Table 2: Evaluation of the proposed approach

Datasets/Methods	*BPMI*	*BGraph*	*BLR*	*TLSA*	*Proposed*
person/location	0.21	0.20	0.25	0.19	**0.60**
terrorism/spam	0.43	0.39	0.41	0.35	**0.70**
person/email	0.40	0.24	0.32	0.28	**0.62**

Baselines Four state-of-the-art methods were implemented for comparison, namely, PMI based method(*BPMI*)[5], the graph based method(*BGraph*)[2], the logistic regression based method(*BLR*)[3], and a two-step LSA method(*TLSA*)[4].

Experimental Evaluation In Table 1, we present the optimization results in terms of pairwise comparison counts, which show that the subset-based method can reduce the comparison counts between all the given entities and their alias candidates. According to the experimental study, we found that the documents union in which only the top 50% entities appear can totally cover all the documents of the given corpus. Through parameter training, we choose merge threshold α as 0.2 and N as 50% for all test cases. Table 2 presents the performance of different methods in terms of F_1. We see that our method outperforms the four baseline methods significantly.

4. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program) under Grant No. 2011CB302206 and National High-Tech Research & Development Program of China (863 Program) under Grant No. 2012AA011103, National Natural Science Foundation of China under Grant No. 60833003 and Grant No. 90924025, and an HP Labs Innovation Research Program award.

5. REFERENCES

[1] Bollegalla et al. Identification of personal name aliases on the web. In *Proceedings of WWW* (2008).

[2] Hölzer et al. Email alias detection using social network analysis. In *Proceedings of LinkKDD* (2005).

[3] Hsiung, P. et al. Alias detection in link data sets. In *Proceedings of IA* (2005).

[4] Oates, T. et al. Using latent semantic analysis to find different names for the same entity in free text. In *Proceedings of the 4th international workshop on Web information and data management* (2002).

[5] Pantel, P. Alias detection in malicious environments. In *Proceedings of Proceedings of AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection* (2006).

[6] Lewis, D. D., and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR* (1994).