Information Sciences 261 (2014) 89–100

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Toward detection of aliases without string similarity

Ning An [a,b,*], Lili Jiang [c], Jianyong Wang [d], Ping Luo [e], Min Wang [e], Bing Nan Li [a,f,*]

[a] *Gerontechnology Lab, Hefei University of Technology, Hefei, China*
[b] *School of Computer and Information, Hefei University of Technology, Hefei, China*
[c] *Max Planck Institute for Informatics, Saarbrucken, Germany*
[d] *Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[e] *HP Labs China, Beijing, China*
[f] *Department of Biomedical Engineering, Hefei University of Technology, Hefei, China*

ARTICLE INFO

Article history:
Received 20 June 2012
Received in revised form 24 September 2013
Accepted 8 November 2013
Available online 18 November 2013

Keywords:
Alias detection
Entity subset
Active learning
Supervised classification

ABSTRACT

Entity aliases commonly exist. Accurately detecting these aliases plays a vital role in various applications. In particular, it is critical to detect the aliases that are intentionally hidden from the real identities, such as those of terrorists and frauds. Most existing work does not pay close attention to the aliases that have low/no string similarity to the given entities. In this paper, we propose a classifier that is based on active learning for detecting this type of aliasing. To minimize the cost of pair-wise comparison, a subset-based method is designed to restrict the selection within entity subsets. An active learning classifier is then employed in each entity subset to find the probability of whether a candidate is the alias of a given entity within the subset. After all of the results from the classifier are integrated, a list of aliases is returned for each given entity. For evaluation, we implemented four state-of-the-art methods and compared them with our proposed approach on three datasets. The results clearly demonstrate that this new active learning classifier is superior to those existing methods.

1. Introduction

Solving the problem of entity alias detection [20,22] is important for a large number of applications. Typically, there are two types of aliases: one type of alias can be roughly detected through string similarity, for example, “John Smith” and “J.M. Smith”; the second type of alias, such as nicknames, fraud, and terrorist aliases, has a low string similarity, and certain number of these aliases are intentionally hidden from their real identities. It is obvious that a pure string similarity search could fail to handle semantically identical entities. For example, “Wisconsin state” has the nickname “dairy state”, and “Abu Abdallah” was used as an alias of “Osama bin Laden”. It is appropriate to call this type of alias a semantic entity alias. Detecting semantic entity aliases is very useful in the real world, and the aim of this paper is to detect such semantic entity aliases.

Researchers have investigated this issue with respect to various domains, including people alias extraction [5,14], fraud detection [6,23], medical alias extraction [10,18], and terrorist recognition [16,22]. Their solutions focus on a special domain (e.g., peoples’ names, terrorism or fraud detection) but fail to correctly obtain the true aliases under a broad range of circumstances. Compared to detecting an entity alias with string similarity, it is more challenging to discover a semantic entity alias. First, there is no or quite low string similarity between a given entity and its semantic aliases. In particular, certain types of

[*] Corresponding authors. Address: Hefei University of Technology, P.O. Box 112, Tunxi Road 193, Hefei 230009, China. Tel.: +86 139 65089390; fax: +86 551 62901760.

E-mail addresses: ning.g.an@acm.org (N. An), bingoon@ieee.org (B.N. Li).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.11.010

entities (e.g., terrorists or frauds) are intentionally hidden from their real identities; hence, the commonly used rules (e.g., ‘‘aka’’, ‘‘as well known as’’, and ‘‘also called’’) [5] do not work. Second, with the increasingly growing volume of information and data, the number of an entity's aliases is rather small compared to billions of strings/entities. Consequently, it is difficult to accurately detect the true aliases for a concerned entity.

To achieve this goal, we decided to employ a supervised method that is based on active learning to solve this problem. Then, the following two issues arise: first, because there is low/no string similarity, it is expensive to choose potential alias candidates for each given entity from a large-scale document corpus; second, classification features and training samples play important roles in the supervised learning methods [19]. To address these two issues, we propose a probabilistic classification method that is based on active learning. Initially, to reduce the cost of pair-wise comparisons for selected alias candidates, a subset-based method is introduced to divide the extracted entities into subsets. Next, three informative features are employed to train a probabilistic classifier. In particular, the strategy of active learning is used to choose high-quality training samples. Finally, the classifier assigns a probability to each pair of a concerned entity and its corresponding extracted entity (alias candidate) in the same subset.

The contributions of this study include:

- Proposing a subset-based method to decrease the cost of pair-wise comparisons and to improve the overall precision of alias detection.
- Designing a classifier that is based on active learning, for which the training samples are selected for high-quality classification.
- Conducting extensive experiments on three types of datasets.

It is noteworthy that a preliminary version of this paper was published as a poster [17]. We have since then made significant enhancements to this paper, as follows:

- Explaining the proposed method in more detail, with additional background knowledge.
- Enhancing the proposed method by adding graph-based features and user-selected samples during the training.
- Adding state-of-the-art indexing measures to make a better evaluation of the proposed method.
- Presenting more experiments and related analyses.

The remainder of this paper is organized as follows. After Section 2 introduces the related work, Section 3 formulates the problem and describes the overall framework. We then present the proposed methods in Section 4 and show the experiment results in Section 5. Conclusions and future work are summarized in Section 6.

2. Related work

The problem of entity alias detection has a close connection with the data matching problem [8], including deduplication [12,24], record linkage [3] and entity resolution [4]. Here, deduplication is the process of removing duplicate records, i.e., records that refer to the same entity, in one data set; record linkage is the process of finding related records and creating a linkage between them [7]; and entity resolution is the process of finding duplicate records and merging them. There has also been some work that addresses the data-matching issue by using active learning [1,24,28]. Most of these studies are aimed at resolving the entities and have string matching as the initial stage. There are two major strategies, namely grouping and indexing, to reduce the cost of performing pair-wise comparisons [9,13,15]. The former strategy is often adopted in database-based applications, where each entity has various attributes, and the entities are grouped according to the similarity between the attribute values. The latter strategy depends on string similarity, which does not work for semantic alias detection. In this paper, we propose a new subset-based method for effectively reducing the scope of the pair-wise comparisons for each concerned entity, thus making the subsequent processing more efficient.

While many existing studies have focused on detecting the general aliases, such as a person's name, an organizational name, and a place name [5,10,11,14], several other studies have paid close attention to detecting deceptive and unknown identities, including falsified identifiers created by frauds [6,23] and terrorists [16,22]. The aim of [5] was to extract person aliases in the Web; this approach constituted three steps, namely pattern extraction, model training, and candidate extraction. The authors employed a few string patterns, such as ‘‘aka’’ and ‘‘better known as’’, and evaluated the method on person's names and location names. Because they focused on the prominent entity aliases, the common rules (e.g., ‘‘aka’’, ‘‘well known as’’, and ‘‘also called’’) were effective. The authors of [23] employed point-wise mutual information (PMI) to measure the importance of various features and, then, used the cosine measure to compute the similarity between each given entity and its alias candidates. Additionally, researchers [11] attempted to detect entity aliases by learning a set of rules and using the tree augmented naive Bayesian networks (TAN) to calculate the probability for a pair of entity and alias candidates. The study in [14] constructed a social network S from collections of email addresses and identified all of the aliases for a given email address that also appears in S. One of the disadvantages of this method is its specificity on an email alias instead of a general setting. The authors of [16] used a training model to predict terrorist and spammer aliases. The authors of [22] proposed a two-stage latent semantic analysis (LSA) for terrorist detection. It is noteworthy that this study used only the test