July 2015

PERSONAL NAME ALIASES ON AUTOMATIC DISCOVERY FROM THE WEB

Y. SARATH KUMAR

St.Anns College of Engineering & Technology, Chirala, AP, India, ballak999@gmail.com

ESWAR KODALI

St.Anns College of Engineering & Technology, Chirala, AP, India, kodali_eswar77@gmail.com

P. HARINI

St.Anns College of Engineering & Technology, Chirala, AP, India, hpogadadanda@yahoo.co.in

Follow this and additional works at: <https://www.interscience.in/ijcct>

Recommended Citation

KUMAR, Y. SARATH; KODALI, ESWAR; and HARINI, P. (2015) "PERSONAL NAME ALIASES ON AUTOMATIC DISCOVERY FROM THE WEB," *International Journal of Computer and Communication Technology*: Vol. 6 : Iss. 3 , Article 3.

DOI: 10.47893/IJCCT.2015.1296

Available at: <https://www.interscience.in/ijcct/vol6/iss3/3>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

PERSONAL NAME ALIASES ON AUTOMATIC DISCOVERY FROM THE WEB

¹Y. SARATH KUMAR, ²ESWAR KODALI & ³P. HARINI

St.Anns College of Engineering & Technology, Chirala, AP, India
ballak999@gmail.com, kodali_eswar77@gmail.com, hpogadadanda@yahoo.co.in

Abstract: In this paper we proposed a lexical-pattern-based approach to extract aliases of a given name. We use a set of names and their aliases as training data to extract lexical patterns that describe numerous ways in which information related to aliases of a name is presented on the web. An individual is typically referred by numerous name aliases on the web. Accurate identification of aliases of a given person name is useful in various web related tasks such as information retrieval, sentiment analysis, personal name disambiguation, and relation extraction. We propose a method to extract aliases of a given personal name from the web. Given a personal name, the proposed method first extracts a set of candidate aliases. Second, we rank the extracted candidates according to the likelihood of a candidate being a correct alias of the given name. We evaluate the proposed method on three data sets: an English personal names data set, an English place names data set, and a Japanese personal names data set. The proposed method outperforms numerous baselines and previously proposed name alias extraction methods, achieving a statistically significant mean reciprocal rank (MRR) of 0.67.

Keywords: Automated Discovery, Aliases, lexical patterns.

1. INTRODUCTION

Searching for information about people in the web is one of the most common activities of internet users. Around 30 percent of search engine queries include person names. However, retrieving information about people from web search engines can become difficult when a person has nicknames or name aliases. For example, the famous Japanese major league baseball player Hideki Matsui is often called as Godzilla on the web. A newspaper article on the baseball player might use the real name, Hideki Matsui, whereas a blogger would use the alias, Godzilla, in a blog entry. We will not be able to retrieve all the information about the baseball player, if we only use his real name. Identification of entities on the web is difficult for two fundamental reasons: first, different entities can share the same name (i.e., lexical ambiguity); second, a single entity can be designated by multiple names (i.e., referential ambiguity). For example, the lexical ambiguity consider the name Jim Clark. Aside from the two most popular namesakes, the formula-one racing champion and the founder of Netscape, at least 10 different people are listed among the top 100 results returned by Google for the name. On the other hand, referential ambiguity occurs because people use different names to refer to the same entity on the web. For example, the American movie star Will Smith is often called the Fresh Prince in web contents. Although lexical ambiguity, particularly ambiguity related to personal names has been explored extensively in the previous studies of name disambiguation [3], [4], the problem of referential ambiguity of entities on the web has received much less attention. In this paper, we specifically examine on the problem of automatically extracting the various references on the web of a particular entity.

The semantic web is intended to solve the entity disambiguation problem by providing a mechanism to add semantic metadata for entities.

However, an issue that the semantic web currently faces is that insufficient semantically annotated web contents are available. Automatic extraction of metadata [4] can accelerate the process of semantic annotation. For named entities, automatically extracted aliases can serve as a useful source of metadata, thereby providing a means to disambiguate an entity. Identifying aliases of a name are important for extracting relations among entities. For example, Matsuo et al. [8] propose a social network extraction algorithm in which they compute the strength of the relation between two individuals A and B by the web hits for the conjunctive query, “A” and “B”. However, both persons A and B might also appear in their alias names in web contents. Consequently, by expanding the conjunctive query using aliases for the names, a social network extraction algorithm can accurately compute the strength of a relationship between two persons.

Figure 1 Proposed Method

Alias identification is closely related to the problem of cross-document co reference resolution in which the objective is to determine whether two mentions of a name in different documents refer to the same entity. Bagga and Baldwin [10] proposed a cross-document co reference resolution algorithm by first performing within document co reference resolution for each individual document to extract co reference chains, and then, clustering the co reference chains under a vector space model to identify all mentions of a name in the document set. However, the vastly

numerous documents on the web render it impractical to perform within document co reference resolution to each document separately, and then, cluster the documents to find aliases.

In personal name disambiguation the goal is to disambiguate various people that share the same name (namesakes) [3], [4]. Given an ambiguous name, most name disambiguation algorithms have modeled the problem as one of document clustering in which all documents that discuss a particular individual of the given ambiguous name are grouped into a single cluster.

The web people search task (WePS)¹ provided an evaluation data set and compared various name disambiguation systems.

2. RELATED WORK

The proposed method is outlined in Fig. 1 and comprises two main components: pattern extraction, and alias extraction and ranking. Using a seed list of name-alias pairs, we first extract lexical patterns that are frequently used to convey information related to aliases on the web. The extracted patterns are then used to find candidate aliases for a given name. We define various ranking scores using the hyperlink structure on the web and page counts retrieved from a search engine to identify the correct aliases among the extracted candidates.

Many modern search engines provide a brief text snippet for each search result by selecting the text that appears in the web page in the proximity of the query. Such snippets provide valuable information related to the local context of the query. For names and aliases, snippets convey useful semantic clues that can be used to extract lexical patterns that are frequently used to express aliases of a name. For example, consider the snippet returned by Google2 for the query “Will Smith The Fresh Prince.”

Here, we use the wildcard operator to perform a NEAR query and it matches with one or more words in a snippet. In Fig. 2 the snippet contains aka (i.e., also known as), which indicates the fact that fresh prince is an alias for Will Smith.

...Rock the House, the duo's debut album of 1987, demonstrated that **Will Smith**, aka **the Fresh Prince**, was an entertaining and amusing storyteller...

Figure 2 Query Return Example

In addition to aka, numerous clues exist such as nicknamed, alias, real name is, which are used on the web to represent aliases of a name. Consequently, we propose the shallow pattern extraction method illustrated in Fig. 3 to capture the various ways in which information about aliases of names is expressed on the web.

Lexico-syntactic patterns have been used in numerous related tasks such as extracting hypernyms [4] and meronyms [5].

Algorithm 3.1: EXTRACTPATTERNS(*S*)

comment: *S* is a set of (NAME, ALIAS) pairs

P ← null

for each (NAME, ALIAS) ∈ *S*

do { *D* ← GetSnippets(“NAME * ALIAS”)
for each snippet *d* ∈ *D*
do *P* ← *P* + CreatePattern(*d*)

return (*P*)

Figure 3 Pattern Extraction model

3. PROPOSED METHOD

The proposed pattern extraction algorithm can extract a large number of lexical patterns. If the personal name under consideration and a candidate alias occur in many lexical patterns, then it can be considered as a good alias for the personal name. Consequently, we rank a set of candidate aliases in the descending order of the number of different lexical patterns in which they appear with a name. The lexical pattern frequency of an alias is analogous to the document frequency (DF) popularly used in information retrieval.

A frequently observed phenomenon related to the web is that many pages with diverse topics link to so-called hubs such as Google, Yahoo, or MSN. Two anchor texts might link to a hub for entirely different reasons. Therefore, cooccurrences coming from hubs are prone to noise. Consider the situation shown in Fig. 4 where a certain web page is linked to by two sets of anchor texts. One set of anchor texts contains the real name for which we must find aliases, whereas the other set of anchor texts contains various candidate aliases. If the majority of anchor texts linked to a

particular web site use the real name to do so, then the confidence of that page as a source of information regarding the person whom we are interested in extracting aliases increases. We use this intuition to compute a simple discounting measure for co-occurrences in hubs as follows.

To overcome the adverse effects of a hub *h* when computing co-occurrence measures, we multiply the number of co-occurrences of words linked to *h* by a factor $\alpha(h, p)$, where

$$\alpha(h, p) = \frac{t}{d}.$$

Here, *t* is the number of inbound anchor texts of *h* that contain the real name *p*, and *d* is the total number of inbound anchor texts of *h*. If many anchor

texts that link to h contain p (i.e., larger t value), then the reliability of h as a source of information about p increases. On the other hand if h has many inbound links (i.e., larger d value), then it is likely to be a noisy hub and gets discounted when multiplied by that factor. Intuitively, Formula 14 boosts hubs that are likely to contain information related to p , while penalizing those that contain various other topics. 3.6 Page-Count-Based Association Measures. We defined various ranking scores using anchor texts. However, not all names and aliases are equally well represented in anchor texts. Consequently, in this we define word association measures that consider co-occurrences not only in anchor texts but in the web overall. Page counts retrieved from a web search engine for the conjunctive query, “ p and x ,” for a name p and a candidate alias x can be regarded as an approximation of their co occurrences in the web. We compute popular word association measures using page counts returned by a search engine.

4. RESULTS & CONCLUSIONS

Fig. 4 shows The Activity diagram of the Project. The other diagrams are also shown in fig 5-7. We utilized both lexical patterns extracted from snippets retrieved from a web search engine as well as anchor texts and links in a web crawl. Lexical patterns can only be matched within the same document. In contrast, anchor texts can be used to identify aliases of names across documents. The use of lexical patterns and anchor texts, respectively, can be considered as an approximation of within document and cross-document alias references. Experimentally, we showed that by combining both lexical patterns based features and anchor text-based features; we can achieve better performance in alias extraction.

An alias might not always uniquely identify a person. For example, the alias Bill is used to refer many individuals who has the first name William.

The namesake disambiguation problem focuses on identifying the different individuals who have the same name. The existing namesake disambiguation algorithms assume the real name of a person to be given and does not attempt to disambiguate people who are referred only by aliases. we showed experimentally that the knowledge of aliases is helpful to identify a particular person from his or her namesakes on the web. Aliases are one of the many attributes of a person that can be useful to identify that person on the web. Extracting common attributes such as date of birth, affiliation, occupation, and nationality have been shown to be useful for namesake disambiguation on the web [8].

Figure 4 Activity Diagram

Figure 5 Sequence Diagram

Figure 6 Screen Shot of the Current Model

Figure 7 Creating Own Access Screen Shot

We proposed a lexical-pattern-based approach to extract aliases of a given name.

We use a set of names and their aliases as

training data to extract lexical patterns that describe numerous ways in which information related to aliases of a name is presented on the web.

We evaluate the proposed method using three data sets: an English personal names data set, an English location names data set, and a Japanese personal names data set. The proposed method reported high MRR and AP scores on all three data sets and outperformed numerous baselines and a previously proposed alias extraction algorithm. Discounting co-occurrences from hubs is important to filter the noise in co-occurrences in anchor texts. For this purpose, we proposed a simple and effective hub discounting measure.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their comments which were very helpful in improving the quality and presentation of this paper.

REFERENCES:

- [1] G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill Inc., 1986.
- [2] M. Mitra, A. Singhal, and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, pp. 206-214, 1998.
- [3] P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04), 2004.
- [4] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An Advanced Social Network Extraction System," Proc. WWW '06, 2006.
- [5] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. Assoc. for Computational Linguistics (ACL '02), pp. 417-424, 2002.
- [6] A. Bagga and B. Baldwin, "Entity-Based Cross-Document Coreferencing Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), pp. 79-85, 1998.
- [7] C. Galvez and F. Moya-Anegon, "Approximate Personal Name- Matching through Finite-State Graphs," J. Am. Soc. for Information Science and Technology, vol. 58, pp. 1-17, 2007.
- [8] M. Bilenko and R. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. SIGKDD '03, 2003.
- [9] T. Hokama and H. Kitagawa, "Extracting Mnemonic Names of People from the Web," Proc. Ninth Int'l Conf. Asian Digital Libraries