Information Retrieval of a Name by using its Aliases using Pattern Extraction Algorithm

Akhil M. Jaiswal
M.E. Scholar Department of
Computer Science & Engineering
H.V.P.M. C.O.E.T, Amravati, India

Anjali B. Raut, PhD
H.O.D Department of
Computer Science & Engineering
H.V.P.M. C.O.E.T, Amravati, India

ABSTRACT
A person may have multiple personal name aliases and that same thing might available be on the web. Identifying aliases of a name is useful in information retrieval, investigating about things, knowledge management, sentiment analysis, relation extraction and name disambiguation. Extracting aliases of an entity is important task in various jobs such as identification of relations among entities, web search and entity disambiguation. The objective of detecting aliases from the web is to retrieve all the information pertaining to a personal name whose content is described with different nick names called aliases in different documents of web. The Web contains aliases of popular personalities in various domains like sports, politics, music, cinema etc., and the problem is that it does not contain alias information or information related to its alias names. Recently, there are proven methods of extracting aliases through lexical pattern based retrieval tested using real-world name-alias pairs in Japanese and English as training data related to limited domains. In this paper, a method is proposed in which a lexical-pattern-based approach to extract aliases of a given name that is helpful in information retrieval. A set of names is used and their aliases to extract lexical patterns that describe numerous ways in which information related to aliases of a name is presented on the web the aliases which will be extracted using the proposed method can be successfully utilized in an information retrieval task and will improve in a relation detection task.

Keywords
Information Retrieval, Web mining, information extraction, web text analysis.

1. INTRODUCTION
One of the most common activities of the internet users is searching for information about people on the web. Around 40 percent of search engine queries include person names [1], [2]. However, Information retrieval about people from web search engines can become difficult when a person has nicknames or name aliases. For Example, A newspaper article on the Cricket player might use the real name, Saurav Gangully, whereas a blogger would use the alias, Dada, in a blog entry. We will not be able to retrieve all the information about this Cricket player, if we will only use his real name.

Particularly with keyword-based search engines, pages which use the real name to refer to the person about whom we are interested in finding out information will only be retrieved. In such cases, automatically extracted aliases of the name are useful to expand a query in a web search, thereby improving recall. Identification of entities on the web is difficult for two fundamental reasons:

1. Different entities can share the same name called lexical ambiguity.
2. A single entity can be designated by multiple names called referential ambiguity.

For example, the lexical ambiguity considers the tollywood actor name Shivaji Rao Gaikwad, is aside from the two most popular namesakes, Rajnikant and South Superstar. Referential ambiguity occurs because people use different names to refer to the same entity on the web. The problem of referential ambiguity of entities on the web has received much less attention.

We propose a fully automatic method to discover aliases of a given personal name from the web. Our contributions can be summarized as follows:

- Propose a lexical pattern-based approach to extract aliases of a given name using snippets returned by a web search engine. The lexical patterns are generated automatically using a set of real world name alias data. Evaluate the confidence of extracted lexical patterns and retain the patterns that can accurately discover aliases for various personal names. To select the best aliases among the extracted candidates, the authors propose numerous ranking scores based upon three approaches:
  - lexical pattern frequency
  - word co-occurrences in an anchor text graph
  - page counts on the web.
- The optimal combination of individual ranking scores to construct a robust alias extraction method. 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 [3] 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.

2. RELATED WORK
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 [4] 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. In personal name disambiguation the goal is to disambiguate various people that share the same name (namesakes) [5], [6]. However, the name disambiguation problem differs fundamentally from that of alias extraction. Because in name disambiguation the objective is to identify the different entities that are referred

by the same ambiguous name;–in alias extraction, the authors are interested in extracting all references to a single entity from the web. Approximate string matching algorithms have been used for extracting variants or abbreviations of personal names [7]. Bilenkoand Mooney [8] proposed a method to learn a string similarity measure to detect duplicates in bibliography databases. However, an inherent limitation of such string matching approaches is that they cannot identify aliases.

3. ANALYSIS OF PROBLEM.

Identifying aliases of a name are important in information retrieval [9]. In information retrieval, to improve recall of a web search on a person name, a search engine can automatically expand a query using aliases of the name [10]. For Example, user who searches for Amitabh Bacchan might also be interested in retrieving documents in which Bacchan is referred to as BigB. Consequently, we can expand a query on real name using his alias name BigB. 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 [3] 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.

4. METHOD

The proposed method is outlined in Fig.1 and comprises three main components: pattern extraction, and alias extraction and ranking. Using a seed list of name-alias pairs, first there will be extraction of 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. Various ranking scores can be defined using the hyperlink structure on the web and page counts retrieved from a search engine to identify the correct aliases among the extracted candidates.

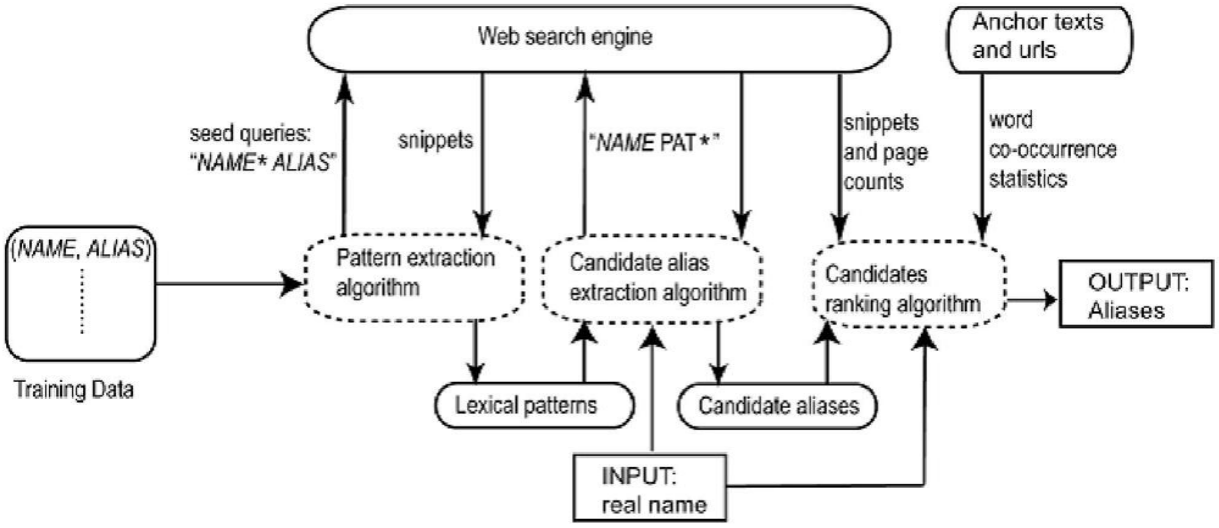

Fig.1: Proposed Method

4.1 Desired Implications

- It can be useful tool for analyzing the name of person from his alias name & extract the related information to it.
- In this it can be predictable step for detection of any kind of cyber crime.

5. CONCLUSION

The proposed lexical-pattern-based approach to extract aliases of a given name, uses 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. Next, the real name of the person is substituted that we are interested in finding aliases in the extracted lexical patterns, and download snippets from a web search engine. A set of candidate aliases from the snippets is extracted. The candidates are ranked using various ranking scores. Moreover, the extracted aliases significantly improved recall in a relation detection task and render useful in a web search task.

6. REFERENCES

[1]. R. Guha and A. Garg, "Disambiguating People in Search,"technical report, Stanford Univ., 2004.

[2]. J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570, 2005.

[3]. P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04),2004.

[4]. A. Baggaand B. Baldwin, "Entity-Based Cross-Document Coreferencing Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), pp. 79-85, 1998. (C : 240)

[5]. G. Mann and D. Yarowsky, "Unsupervised Personal Name Disambiguation," Proc. Conf. Computational Natural Language Learning (CoNLL'03), pp. 33-40, 2003. (C : 206)

[6]. R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of People in a Social Network," Proc. Int'l World Wide Web Conf. (WWW '05), pp. 463-470, 2005. (C : 166)

[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. (C : 26)

[8]. M. Bilenko and R. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. SIGKDD '03, 2003. (C : 418)

[9]. G. Salton and M. McGill, Introduction to Modern,Information Retreival. McGraw-Hill Inc., 1986.

[10]. M. Mitra, A. Singhal and C. Buckley, "Improving Automatic Query Expansion," Proc. SIGIR '98, pp.206-214, 1998.