Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Grouping of Personalized Web Documents by Naming Aliases

S. S. Shinde
Department of Information Technology
Bharati Vidyapeeth Deemed University, Pune , India

P. R. Devale
Department of Information Technology
Bharati Vidyapeeth Deemed University, Pune , India

Abstract: In lots of works such as information retrieval, sentiment analysis, person name disambiguation as well as in biomedical fields it is required to identify the accurate references to an entity among a list of references. More previous work had been done on solving lexical ambiguity. Here we proposed a method that is based on referential ambiguity. In this paper we proposed a method which is based on referential ambiguity to extract correct alias for a given name. Given a person name and / or with context data such as location, organization retrieves top K snippets and depth up to level two from a web search engine. With the help of lexical pattern extract candidate aliases. To find correct alias from a list of aliases we used n-depth crawling method. This method is useful to improve the precision and minimize the recall than the previous baseline method. Using these candidate aliases related personalized web documents are clustered or grouped. Grouping attains high accuracy and reduces the complexity.

Keywords: Web mining, information retrieval, n-depth crawling, clustering, and web text analysis.

I. INTRODUCTION

Finding a relevant, information of a particular entity on the web is very important task as it is helpful in information retrieval process. Retrieving information of a person simply by using his or her name is quite insufficient if the person has nick names. Now a day celebrities are known by 2 or more name in the web. Entities may be a person, an organization, a location, a festival name, etc. Identification of entities on the web is difficult for two basic reasons. First: different entities may share the same name (Lexical ambiguity). Second: One entity is known by different names (Referential ambiguity). 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, we are interested in extracting all references to a single entity from the web.

For example: Diwali is also known as Deepavali as a one word alias or Festival of Lights as a three word alias. The cricketer, Mahendra Singh Dhoni, is also known as Dhoni or Mahi. Similarly, entities are also referenced by drama, profession, etc. Grouping of web pages identifies semantically meaningful groups of web pages and presents these to the users as clusters. The clusters provide an overview of the contents of the result set and when cluster is selected the result set is refined to just the relevant pages in that cluster.

II. RELATED WORK

Correct alias finding is important in information retrieval. In [1], Danushka Bollegala proposed a method which uses extraction techniques to automatically extract significant entities such as the names of other persons, organizations and locations on each webpage. In that method for given person name, it extract person name from the web by using lexical pattern matching method and anchor text analysis. They ranked the candidate alias from the list. For this they integrated various similarity measures scores and given to a single function to support vector machine.

In [2] Dmitri proposed a method in which automatic entity extraction techniques are explained. In addition, it extracts and parses HTML and Web related data on each web page, such as hyperlinks and email addresses. Then this information is presented in an Entity Relationship Graph. This method is used to find relative information of a particular person on the web. In [3], A. Bagga proposed a method that summarizes the interested entities and ranks the similarity of summaries using various information metrics. 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 [4], T. Hokama proposed a method, especially for Japanese language.

In [5], C. Galvez proposed a method for extraction of abbreviations of personal names that measures approximate string matching algorithms. But by using such string matching approaches we cannot identify aliases, which do not share any words or letters with the real name of person i.e. approximate string matching methods would not identify Mahendra singh Dhoni as Mahi, an alias for Mahendra singh Dhoni.

In [6], Christian Borgelt explained how text classification is done using graph mining and also explained different graph parameters.

III. GENERAL IDEA OF PROPOSED METHOD

To overcome the limitations of previous information retrieval techniques, the goal is to group all the entity descriptions that refer to the same real world entities. In this paper we use algorithms such as extract patterns, extract candidate aliases and ranking of candidates aliases.

The proposed work involves the design and implementation of cluster based web search system. The following are the steps of the overall approach, in the context of middleware architecture. A user submits a query to the middleware via a specialized web – based interface. The middleware queries a search engine with this query via the search engine API and retrieves a fixed number of relevant web pages. Then these web pages are processed and first we find out the different aliases of personal name. Then evaluate candidate aliases by using different ranking scores and finally the documents of the same people go to the same group. These aliases are used to find all documents of one person having different nicknames. Finally, we get several groups of documents; each group contains documents related to the same person.

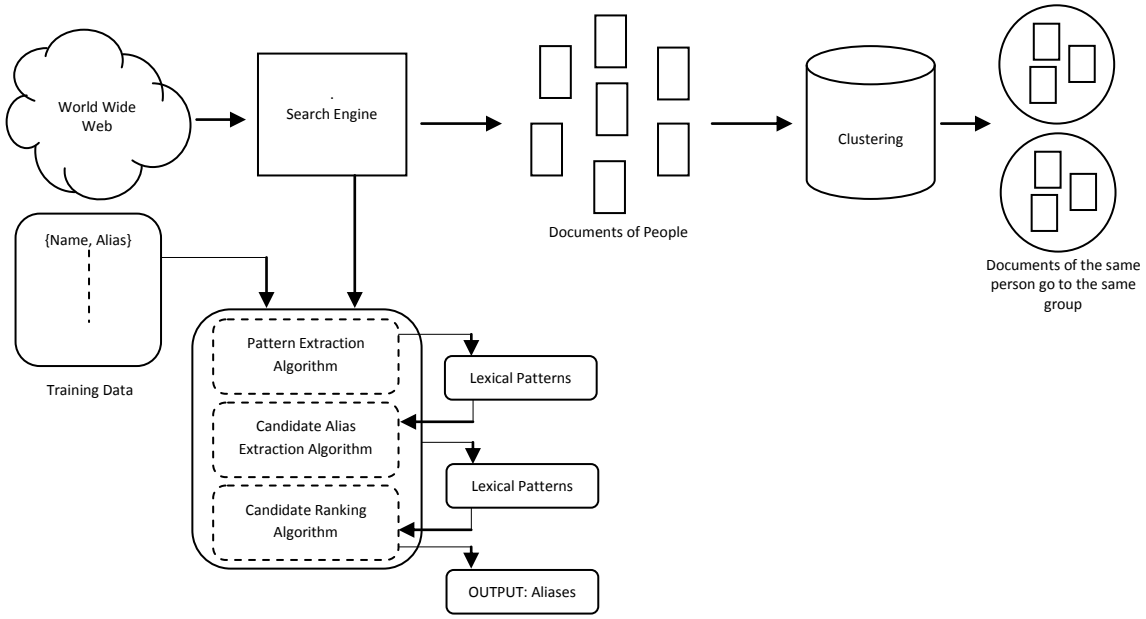

Fig. 1 Overview of the Proposed Method

IV. CONCLUSION

In this paper we propose a cluster based web search approach that is based on personal name aliases in order to get better disambiguation quality. We use a lexical-pattern-based approach to extract aliases of a given name. Referential ambiguity is also removed by finding the aliases.

REFERENCES

[1] Danushka Bollegala, YutakaMatsuo and IitsuruIshizuka, Member , IEEE, Automatic Discovery of Personal Name Aliases from the Web, *IEEE Transaction on knowledge and data engineering, vol. 23, no. 6,* June 2011.

[2] Dmitri V. Kalashnikov Zhaoqu Chen Rabia Nuray – Turan Sharad Mehrotra Zheng Zhang, Web People Search via connection Analysis, *IEEE International Conference on Data Engineering,* 2009.

[3] A. Bagga and B. Baldwin, Entity-Based Cross-Document Coreferencing using the vector space model, *Proc. Int's Conf. Computational linguistics (COLING '98), pp. 79-85*, 1998.

[4] T. Hokama and H. Kitagawa, Extracting Mnemonic Names of People from the Web*, Proc. Ninth Int'l Conf. Asian Digital Libraries (ICADL ' 06), pp. 121-130*, 2006.

[5] C. Galvez and Fg. Moya-Anegon, Approximate Personal Name Matching through Finite State Graphs, *J. Am. Soc. Fro Information Science and Technology, vol. 58, pp. 1-17*, 2007.

[6] Christian Borgelt, Graph Mining: An Overview, *Proc, 19th GMA/GI Workshop Computational Intelligence, Germany,* 2009.

