Survey: Detection of Personal Name Aliases from the Web

Abha Jain[1], Mily Lal[2], Akanksha Goel[3]
[1, 2, 3] Department of Computer Engineering
[1, 2, 3]DYPIEMR,Pune

Abstract- *A human being is known by more than one name in person as well as on the web. Identification of the aliases is very much related to the predicament of cross- document co reference resolution. Here, purpose is to determine whether two mentions of a name in different documents refer to the same entity. Lexical ambiguity, problem is solved by the traditional system but it lagged in identification of referential ambiguity. This paper is survey of discovering of name aliases on the web.*

Keywords- Web mining, information extraction, web text analysis.

I. INTRODUCTION

Internet users are now-a-days more interested in searching for information about individuals on the web. Most of the queries on the search engine includes name of persons. It is very tedious to retrieve data about people from web search engine when a person uses a nickname or alias name. For example, the famous cricket player Sachin Tendulkar is also known as Little Master as a two word alias. Mumbai city is known as Bombay or an Indian festival like Deepavali is known as the festival of light or Diwali. Many times various types of terms are used as alias.

In information retrieval, a search engine can automatically expand the query using aliases of the name. Here a scheme is proposed as an alias identification method that is based on two main things such as links extraction and association measures used.

Here the scheme is a fully automatic identification system for name alias. The task is followed in threefold ways as below.

1. Lexical pattern extraction algorithm is used to retrieve pattern with the help of name alias dataset. This lexical pattern is useful for candidate alias extraction and which are independent of languages.

2. To extract candidate aliases, consider a set of patterns and real name taken as an input to the system. By considering all possible combination of name and pattern, for a given query, we get top-k URLs. After pre-processing we get final list of candidate aliases.

3. Four approaches are proposed to correct aliases from list of candidate aliases. They are listed as below: a) Lexical pattern frequency b) Word co-occurrence in an anchor text c) Page count on the web d) Graph mining method.

4. Last but not least one needs to integrate word score from all approaches and consider normalized weight for each candidate aliases.

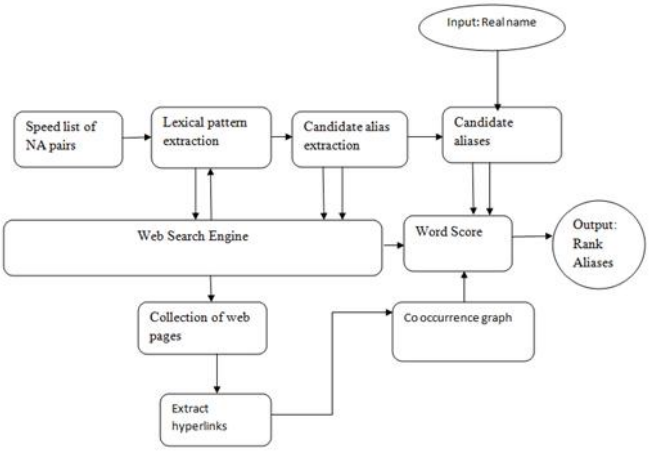

Fig 1: System Architecture

II. METHOD

The projected scheme is outlined in Fig. 1 and has two main mechanisms: pattern extraction, and alias extraction and ranking. Firstly we extract lexical patterns by using a seed list of name-alias pairs. Extracted patterns are used to find candidate aliases for a given name. Defining of various ranking scores is done by using the hyperlink structures on the web. Page counts are then retrieved from a search engine so that correct aliases can be identified from the extracted candidates.

2.1 Extraction of Lexical Patterns using Snippets

Modern search engines provide a brief text snippet for each search result by selecting the text that appears in the web page in the immediacy of the query. These snippets then 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 "Sachin Tendulkar _Little Master."

2.2 Candidates Ranking

Taking into consideration the noise in web snippets, candidates extracted by the shallow lexical patterns may include a number of invalid aliases. From among these candidates, we must identify those, which are most likely to be correct aliases of a given name. We model this problem of alias recognition as one of ranking candidates with respect to a given name such that the candidates, who are most likely to be correct aliases are assigned a higher rank.

III. COMPARATIVE STUDY

Paper Title	Description	Advantages	Disadvantages
Adaptive Duplicate Detection using Learnable String Similarity Measures	A universal framework Used for learning string similarity measures for duplicate detection Two algorithms are used character based & vector space based text distance.	This approach can be extended as required.	Has over fitting issues
Automatic Acquisition of Hyponyms from Large Text Corpora	1. Proposed automatic acquisition of semantic lexical relations from unrestricted text. 2. 3. Is meant to provide an incremental step toward large goals of natural processing	1. Low cost approach 2. Is complimentary to statistically based approach. 3. Also useful as a critiquing component for existing knowledge based & lexicons.	Noun Phrases can't be identified.
Finding parts in Very Large Corpora	1. A method for extracting parts of objects from whole.	1. Suitable for cars data 2. Large corpus and use of more refined statistical measures for ranking the output	Sparse data is the source of most of the noise.
Word Association Norms, Mutual Information, Lexicography	1. The term word association is used in a very particular sense in the psycholinguistic literature	1. It provides precise statistical calculation that could be applied to a very large corpus of text 2. Helps lexicographer in organizing concordance.	Large corpus data and lexical patterns knowledge needed.

Topic Word Selection Based on Combinatorial Probability	1. Developed a term weighting measure for selecting words characterizing a document set. 2. Measure is based on combinatorial probability 3. HGS method	1. Effective in the task of word selection 2. HGS can be applied to dimension reduction in text categorization/clustering.	Knowledge of combinatorial probability is must.
Measuring Semantic Similarity between words using Web Search Engine	1. The proposed method exploits page counts and text snippets returned by a web search engine.	1. Automatically extracted lexico syntactic patterns 2. Robustly captures semantic similarity between named entities	1. No query suggestion 2. Automatic synonyms not present. 3. Can't identify name alias.
Optimizing Search Engines using Clickthrough Data	1. Presents an approach to automatically optimizing the retrieval quality of search engine using clickthrough data. 2. Taking a support of SVM approach this paper presents a method for learning retrieval functions.	1. Click through data is available in abundance and can be recorded at very low cost. 2. Opens a series of question regarding the use of machine learning in search engines.	1. Single user 2. Optimization problem

IV. CONCLUSION

We have just seen some methods to detect personal name aliases from the web. This helps in information retrieval and improves recalling of a web search on a person name. A search engine can automatically expand a query using aliases of the name. The researchers continue to find the best solutions supporting to this system. In this paper, we have just presented some existing methods and comparative study of the seven papers and collected information for further study.

REFERENCES

[1] M. Bilenko and R. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. SIGKDD '03, 2003

[2] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. Int'l Conf. Computational Linguistics (COLING '92), pp. 539-545, 1992.

[3] M. Berland and E. Charniak, "Finding Parts in Very Large Corpora," Proc. Ann. Meeting of the Assoc. for Computational Linguistics (ACL '99), pp. 57-64, 1999.

[4] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, vol. 24, pp. 513-523, 1988.

[5] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, pp. 22-29, 1991.

[6] T. Hisamitsu and Y. Niwa, "Topic-Word Selection Based on Combinatorial Probability," Proc. Natural Language Processing Pacific-Rim Symp. (NLPRS '01), pp. 289-296, 2001.

[7] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines," Proc. Int'l World Wide Web Conf. (WWW '07), pp. 757-766, 2007.

[8] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD '02, 2002.