ISSN:0975-9646

A Review on Identification of Aliases from Web

Snehal R. Kamble, Prof. S.S.Dhande, Prof. H.R.Vyawahare

Sipna College of Engineering and Technology,
Amravati, Maharashtra, India.

***Abstract*— Many celebrities and experts from various fields may have been referred by not only their personal names but also by their aliases on web. Aliases are very important in information retrieval to retrieve complete information about a person from the web. Various methods have been used before to extract the information of the person. If the person have the aliases or the nicknames then it is difficult to retrieve whole information. A large number algorithms & methods have been developed in last decades to extract relations of persons, to detect groups of persons, and to obtain keywords for a person. This includes cross-document co-reference resolution algorithm, a social network extraction system called *POLYPHONET*, unsupervised frameworks, WePS corpus method. This paper describes all these methods.**

***Keywords*— Text Mining, Information extraction, Web Mining, Web Text Analysis.**

I. Introduction

Social networks play important roles in our daily lives. People conduct communications and share information through social relations with others such as friends, family, colleagues, collaborators, and business partners. Our lives are profoundly influenced by social networks without our knowledge of the implications. Social studies have been conducted from the 1930s, observing and modeling the network structure, its influence and dynamics, and information flow from tribal and village societies on to global corporate and industrial societies [1].

Searching people on the web is a common activity of Internet users. Around 30% of search engine queries include personal names. However, retrieving information about a person merely using his or her real names is insufficient when that person has nicknames. Particularly with keyword-based search engines, we will only retrieve pages which use the real name to refer to the person about whom we are interested in finding information. In such cases, automatically extracted aliases of the name are useful to expand a query in a web search, thereby improving recall [2], [3].

Identification of entities on the web is difficult for two fundamental reasons: first, different entities can share the same name. Second, a single entity can be designated by multiple names. For example, the lexical ambiguity, consider the name Jim Clark 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 [2].

However, referential ambiguity occurs because people use different names to refer to the same entity on the web. Although lexical ambiguity, particularly ambiguity related to personal names has been explored extensively in the previous studies of name disambiguation. The problem of referential ambiguity of entities on the web has received much less attention. We will specifically examine on the problem of automatically extracting the various references on the web of a particular entity.

For the set A, *e* is an entity, the set A of its aliases to be the set of all words or multiword expressions that are used to refer to *e* on the web. For example, Godzilla is a one-word alias for Hideki Matsui, whereas alias The Fresh Prince contains three words and refers to Will Smith. Various types of terms are used as aliases on the web. For instance, in the case of an actor, the name of a role or the title of a drama (or a movie) can later become an alias for the person (e.g., Fresh Prince, Knight Rider). Titles or professions such as president, doctor, professor, etc. are also frequently used as aliases. Variants or abbreviations of names such as Bill for William, and acronyms such as JFK for John Fitzgerald Kennedy are also types of name aliases that are observed frequently on the web.

II. Literature Review And Related Work

D. Bollegala *et al.* proposed the lexical pattern based approach to extract the candidate aliases from the web automatically providing the real name of the person then they gave the ranking to the aliases being the good aliases of the person [2].

Hokama and Kitagawa (HK) proposed an alias extraction method, that is, specific to the Japanese language. For a given name p, they search for the query "* koto p" and extract the context that matches the asterisk. The Japanese word koto, roughly corresponds to *also known as* in English. However, koto is a highly ambiguous word in Japanese that can also mean incident, thing, matter, experience and task. As reported many noisy and incorrect aliases are extracted using this pattern, which requires post-processing heuristics that are specific to Japanese language to filter out the incorrect aliases [4].

Duplicate detection is an important problem in data cleaning, and an adaptive approach that learns to identify duplicate records for a specific domain has clear advantages over static methods. Experimental results demonstrate that trainable similarity measures are capable of learning the specific notion of similarity that is appropriate for a specific domain.

Bilenko and Mooney presented two learnable distance measures that improve over character-based and vector-space based metrics and allow specializing them for specific datasets using labeled examples. They have also shown that support vector machines can be effectively utilized for some datasets both for string similarity and record similarity computations, outperforming traditional methods. Their overall framework for duplicate detection integrates previous work on adaptive methods with learnable similarity measures, leading to improved results [5].

J. Artiles, J. Gonzalo, F. Verdejo described the creation of a testbed to evaluate strategies addressing this people searching task on web documents. They propose a method called the WePS corpus which an initial testbed to test people search strategies over the web. They provide (i) a corpus of web pages retrieved using person names as queries to web search engines;(ii) a classification of pages according to the different people(with the same name) they refer to; (iii) manual annotations of relevant information found in the web pages describing them (e-mail, image, profession, phone number). (iv) the results of applying a general purpose clustering algorithm to that annotated data, which serve as a baseline for the ambiguity resolution problem[3].

G. Mann and D. Yarowsky presented a set of algorithms for finding the real referents for ambiguous personal names in text using unsupervised clustering and feature extraction methods. In particular, they have shown how to learn and use automatically extracted biographic information to improve clustering results, and have demonstrated this improvement by evaluating on pseudonames. They had presented initial results on learning these patterns to extract biographic information for multiple languages, and intend to use these techniques for large-scale multilingual polysemous name clustering.

The results they presented support the automatic clustering of polysemous personal name referents and visualization of these induced clusters and their motivating features. These distinct referents can be verified by inspection both of extracted features and of the high weighted terms for each document. Clusterings may be useful in two ways. First as a useful visualization tool themselves, and second as seeds for disambiguating further entities [6].

S. Sekine and J. Artiles describe the Web People Search 2 attribute extraction task (WePS2-AE). It was conducted in September-December 2008 along with the WePS2 clustering task. Six groups participated in the Attribute Extraction task. They describe the motivation, task definition, evaluation set up, participating systems, and evaluation results and discuss the problems and future directions [6].

C. Galvez and F. Moya-Anegon presented how finite-state methods can be employed in a new and different task: the conflation of personal name variants in standard forms. They classify the personal name variants as non-valid and valid forms. In establishing an equivalence relation between valid variants and the standard form of its equivalence class, they defend the application of finite-state transducers. The process of variant identification requires the elaboration of (a) binary matrices and (b) finite-state graphs [7].

R. Bekkerman and A. McCallum, presents two unsupervised frameworks one based on link structure of the Web pages, another using Agglomerativ/Conglomerative Double Clustering (A/CDC) an application of a recently introduced multi-way distributional clustering method[8].

Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka proposed a social network extraction system called *POLYPHONET*, which employs several advanced techniques to extract relations of persons, to detect groups of persons, and to obtain keywords for a person. Search engines, especially Google, are used to measure co-occurrence of information and obtain Web documents [9].

Bagga and Baldwin 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 [10].

III. Analysis Of Problem

Baldwin proposed a cross-document co-reference resolution algorithm. He was the first to perform within document co-reference resolution for each individual document to extract co-reference chains. Then the chains are clustered 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.

Approximate String matching algorithms have been used for extracting variants or abbreviations of personal names. Rules in the form of regular expressions and edit-distance-based methods have been used to compare names. Bilenko and Mooney proposed a method to learn a string similarity measure to detect duplicates in bibliography databases. However, limitation of such string matching approaches is that they can not identify aliases, which share no words or letters with the real name.

Hokama and Kitagawa (HK) propose an alias extraction method, that is, specific to the Japanese language. For a given name p, they search for the query "* koto p" and extract the context that matches the asterisk. The Japanese word koto, roughly corresponds to also known as in English. However, koto is a highly ambiguous word in Japanese that can also mean incident, thing, matter, experience, and task. As reported, many noisy and incorrect aliases are extracted using this pattern, which requires various post processing heuristics that are specific to Japanese language to filter out the incorrect aliases. Moreover, manually crafted patterns do not cover various ways that convey information about name aliases.

IV. Proposed Work

An individual is typically referred by numerous name aliases on the web. Accurate identification of aliases of a given person name is useful for various tasks. It is used for information retrieval, personal name disambiguation, and relation extraction. A method will be implemented to extract aliases of a given personal name from the web.

Given a personal name, the method first extracts a set of candidate aliases. Second, it will rank the extracted candidates according to the likelihood of a candidate being a correct alias of the given name. Automatically extracted lexical pattern-based approach to efficiently extract a large set of candidate aliases from snippets retrieved from a web search engine shall be developed. The numerous ranking is done to extract the correct aliases of the person. Three methods are used: lexical pattern frequency, word co-occurrences in an anchor text graph, and page counts on the web. To construct a robust alias detection system, integrate the different ranking scores into a single ranking function using ranking support vector machines.

The proposed method comprises two main components: pattern extraction, 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. Extracting Lexical Patterns from Snippets 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.

V. CONCLUSION

This paper has attempted to review a significant number of papers to cover the recent development in the field of personal name disambiguation & to identify the correct person we want to search. Present study reveals that various methods has been used. It includes a cross-document co-reference resolution algorithm, a social network extraction system called *POLYPHONET,* unsupervised frameworks, and many other. The list of references to provide more detailed understanding of the approaches described is enlisted. We apologize to researchers whose important contributions may have been overlooked. In the proposed method, the lexical pattern are used to retrieve the aliases of the person for the complete information extraction about the person.

REFERENCES

[1] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida,and M. Ishizuka, “Polyphonet: An Advanced Social Network Extraction System,”Proc. WWW ’06,2006.

[2] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, ”Automatic Discovery of Personal Name Aliases from the Web” IEEE Tractions on knowledge and data Engineering, Vol. 23, NO. 6, June 2011.

[3] J. Artiles, J. Gonzalo, and F. Verdejo, “A Testbed for PeopleSearching Strategies in the WWW,” Proc. SIGIR ’05, pp. 569-570,2005.

[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] M. Bilenko and R. Mooney, “Adaptive Duplicate Detection Using Learnable String Similarity Measures,” Proc. SIGKDD ’03, 2003.

[6] G. Mann and D. Yarowsky, “Unsupervised Personal Name Disambiguation,” Proc. Conf. Computational Natural Language Learning (CoNLL ’03), pp. 33-40, 2003.

[7] S. Sekine and J. Artiles, “Weps 2 Evaluation Campaign Overview of the Web People Search Attribute Extraction Task,” Proc. Second Web People Search Evaluation Workshop (WePS ’09) at 18th Int’l World Wide WebConf. 2009.

[8] 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.

[9] R. Bekkerman and A. McCallum, “Disambiguating Web Appearances of People in a Social Network,” Proc. Int’l World Wide WebConf. (WWW ’05), pp. 463-470, 2005.

[10]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.