International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 11, November 2014 **ISSN 2319 - 4847**

A REVIEW ON WEB MINING TECHNIQUES FOR ALIAS RECOMMENDATION

Ms. Mayuri S. Bawane[1], Prof. Monika Rajput[2]

[1]M.E. First Year CSE, P.R.Pote(Patil)COET, Amravati, Maharashtra, India

[2]Asst. Prof., Department of CSE, P.R.Pote (Patil) COET, AmravatiMaharashtra, India

ABSTRACT

Most of the celebrities and experts from different fields are generally referred by their personal names but they can also be referred by their aliases on the web. If we want to retrieve complete information about a particularperson then aliases plays an important role in information retrieval. Numbers of methods have been used to extract the information of the person. Suppose a person have the aliases or the nicknames then it is not easy to retrieve whole information. A large number of algorithms & methods have been developed in previous yearsfor extracting relations of persons, for detecting the groups of persons, and forobtaining keywords for a person. Recognizing the various aliases of an entity is a critical task for many applications, including Web search and e-discovery. Therefore we need to accurately identify entity aliases, especially the long tail one`s in the unstructured data. This paper includes the overview of various methods for detecting the aliases of an entity. To address this important need we would be using NLP ie Natural Language processor which would perform Parts of Speech Tagging (POS Tagging), and Chunking of data (extraction of only action words from the input sentence).

Keywords: entity aliases, web search, e-discovery, NLP, POS.

1.INTRODUCTION

Most of the information available on the web today is in an unstructured text format. Large amount of data such as map data, CAD, bimolecular, chemical molecules all are stored in the World Wide Web databases. There are some common graph structures enable virologists to make easy and effective design of corresponding drugs and vaccines.In the fields of Web mining and graph mining, many web documents and many chemical compounds are represented by ordered trees and outerplanar graphs, respectively. Therefore there is a structured discourse graph (SDG) model which can be used to improve the semantic integration, representation and interpretation of unstructured texts within the context of the Linked Data Web[5]. These Entities such as persons, organizations, etc. are commonly exist in various types of data, so identifying entity aliases is necessary in many applications[1]. For example, given an entity query “IBM” to a search engine, “International Business Machines” or “International Business Machines Corp.” may occur in the web search results. Accurately identifying these aliases could help get more applicable pages for the given search query.

The main objective of search engine is to provide the most applicable and required documents for a user’s query. Recognizing the various aliases of an entity is one of the difficult task for many applications, including Web search. Because of which we cannot accurately identify the entity aliases and also we are not able to see the aliases which have occurred. Many celebrities and experts from various fields may have been referred by not only their personal names but also by their aliases on web [4]. For example, people might use “Michel Jackson” as a query on search engine to know about him. The search engine might give the relevant documents met the information need of the user’s query. Apparently celebrities, experts and famous personalities might also be referred by their aliases on the web. Many of the web pages about person names might also be created by aliases. For example, a newspaper article might refer the particular persons using their original names, whereas a blogger might refer them using their nick names [9]. The user will not be able to retrieve all required information about a person if he only uses his personal name for searching. To retrieve complete information about a person name, one might know about all his aliases on the web. Various types of words are used as aliases on the web. Identifying aliases will be helpful in required information retrieval. Aliases are very important in information retrieval to retrieve complete information about a personal name from the web, as some of the web pages of the particular person may also be referred by his aliases. Natural Language Processing is the ability of a computer to understand what human is saying to it. NLP is a Natural Language Processor tool which is used for mining purposes. Currently, the most common technique for Natural Language parsing is done by using pattern matching through references to a database. But the huge variety of linguistical syntax and semantics are present, means

that accurate real time analysis is very difficult. Aliases arise from entities who are trying to hide their identities, from a person with multiple names, or from words which are unintentionally or even intentionally misspelled [11]. Therefore we need the visualization methods that give us the proper result. Given an entity and its alias candidate to ACS, it first splits them into a set of tokens respectively. And, the commonly used stop words such as "and", "for", and "of" are removed from them[1]. Next, we attempt to match the acronym tokens which are generated in the given entity and its candidate. After the acronym match, the remaining tokens in the entity are now required to find its matching tokens in the candidate. To compute the similarity of each pair of tokens respectively from the entity and its respective alias, we analyze the challenges in entity alias discovery for the low-redundant unstructured document, and propose a graph based solution called GRIAS by leveraging the entity relationships in both the structured and unstructured data.

2.RELATED WORK

The studies on entity resolution and record linkage are similar with this work. Also entity linking matches the strings in unstructured data with entities from Web sources. Some of the solutions can be referred as Web alias discovery, while alias discovery could be one of the core tasks in this work. As for the solutions, most of them extract these aliases assuming that the entity and its aliases co-occurrence frequently [6].One of the method for this is the combination of string match and graph-based match. Graph generators are developed to create graphs with a variety of sizes for simulations and experiments in many applications. Statistical graph generators have also attracted significant research interests to preserve important parameters [7]. Bollegala, Matsuo, and Ishizuka [4]proposed a method to extract aliases from the web for a given personal name. They have used lexical pattern approach to extract candidate aliases. The incorrect aliases have been removed by the page counts, anchor text co-occurrence frequency, and lexical pattern frequency respectively. However, this alias method considered only the first order co-occurrences on aliases to rank them but did not focus on the second order co-occurrences to improve recall and achieve a substantial MRR for the web search engine [10]. The problem of extracting aliases of a given name can be considered as a specialcase of the more general problem of extracting the words Y that have a given relation R with a word X. For example, extracting hyponyms, synonyms,metonyms are specific instances of this general problem of relation extraction. That is, hyponyms are the a words of more specific meaning than a general or superordinate term applicable to it and metonyms are the names or expressions used as a substitute for something else with which it is closely associated.Manually created or automatically extracted lexico-syntactic patterns have been successfully used to identify various relations between words. For example, patterns such as X is a Y and X such as Y are typically used to introduce hyponyms, whereas, X of a Y and X's Y are frequently used with the metonyms. NLP Interchange Format (NIF) [1] is a format which targets the interoperability among NLP tools, language resources and annotations [8].NIF consists of two vocabularies (the String Ontology and the Structured Sentence Ontology) which allow the annotation of documents and sentences. NIF concentrates on the documentation of the workflow of resources which are used on the analysis. Frequent patterns have been widely recognized as one of the most important graph characteristics in the graph data mining and analysis. Efficient graph mining algorithms for multiple graphs have been extensively studied.

3.OUTLINE OF PROPOSED SYSTEM

The main objective of search engine is to provide the most relevant documents for a user's query. Recognizing the various aliases of an entity is difficult task for many applications, we cannot accurately identify the entity aliases and also we are not able to see the aliases which have occurred. So, we have proposed a method which is outlined in Fig. 1. This system eliminates the challenges in entity alias discovery for the low-redundant unstructured document, and proposes a graph based solution called GRIAS (abbr. for a Graph based framework for discovering entItyAliaseS) [1]. We develop a Natural Language Processor so that all the text from the web would be processed via NLP. The NLP would be using a Stanford NLP API, which would perform Parts of Speech Tagging.

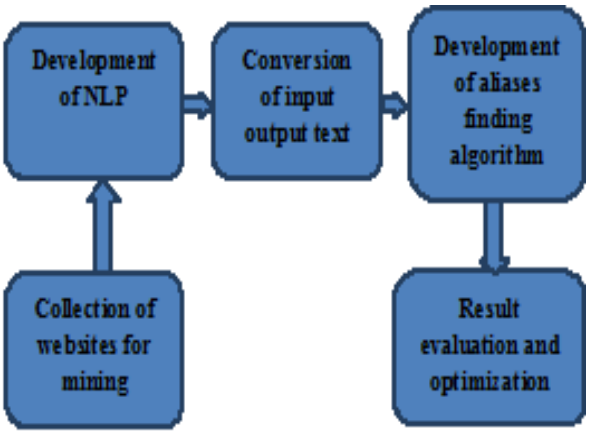

Figure 1 outline of proposed system

The proposed system can be divided into the following modules,

1. Collection of web sites for mining

In this module, we would be collecting various websites for mining of data; these websites would include sites which contain data about many individuals in proper format.

2. Development of Natural Language Processor

In this module, all the text from the web would be processed via a Natural Language processor. The NLP would be using a Stanford NLP API, which would perform Parts of Speech Tagging (POS Tagging), and Chunking of data (extraction of only action words from the input sentence). The output of this would be only the chunked or action words which are detected by the NLP chunker.

3. Conversion of input text into graphs

In this module, the output of NLP would be given for graph formation. In graph formation, the input text would be represented in the form of graphs, where the top node has the highest relevance and the bottom node has the lowest relevance to the sentence

4. Development of Aliases finding algorithm.

In this module, we would be developing an Alias finding algorithm, which would find the aliases from the input dataset, with the help of graph matching

5. Result evaluation and optimization.

In this module, results of our algorithm would be evaluated and the system would be optimized if required. In graph formation, the input text would be represented in the form of graphs, where the top node has the highest relevance and the bottom node has the lowest relevance to the sentence. Our GRIAS framework effectively combines the alias candidates into the graph model and refines them for each concerned entity. GRIAS explores the existing structured data and can be extended through adding more similarity functions based on the additional resources.

4.CONCLUSION

Entity alias discovery is a critical task in many real world applications on web. Hence finding an aliases has become increasingly important for a variety of applications in chemistry, virology, bioinformatics, social networking, etc. Present study reveals that various methods has been used. The list of references to provide more detailed understanding of the approaches described is enlisted. So, in this paper, we have proposed a graph-based approach, called GRIAS, to perform entity alias discovery in free documents. This proposed system would be developing NLP i.e. Natural language processor and converts the input text into graphs. The output will be displayed by developing Alias finding algorithm and the system would be optimized if required.

REFERENCE

[1] Lili Jiang, Ping Luo, Jianyong Wang, YuhongXiong, Bingduan Lin, Min Wang, Ning An, "GRIAS: an Entity-Relation Graph based Framework for Discovering Entity Aliases" IEEE 13th International Conference on Data Mining 2013.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: a generic approach to entity resolution," VLDB J., vol. 18, no. 1, pp. 255–276, 2009.

[3] D. G. Brizan and A. U. Tansel, "A survey of entity resolution and record linkage methodologies," Communications of IIMA, vol. 6, no. 3, 2006.

[4] D. Bollegala, T. Honma, Y. Matsuo, and M. Ishizuka, "Automatically extracting personal name aliases from the web," Proceedings of the International Conference on Natural Language Processing(GoTAL), vol. 5221, pp. 77–88, 2008.

[5] A. Freitas, D. S. Carvalho, J. C. P. da Silva, S. O'Riain, E. Curry, A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia. In Proc. of the 1st Workshop on the Web of Linked Entities, (ISWC), 2012.
[6] D. Milne and I. H. Witten, "Learning to link with wikipedia," in Proceedings of the 17th ACM conference on Information and knowledge management, ser. CIKM '08, 2008, pp. 509–518.
[7] Hong-Han Shuail, De-Nian Yang, Philip S. Yu, Chih-YaShen and Ming-Syan Chen, "On Pattern Preserving Graph Generation", IEEE 13th International Conference on Data Mining 2013.
[8] Andr´eFreitas, Se´anO'Riain, Edward Curry, Jo˜ao C. P. da Silva, Danilo S. Carvalho, "Representing Texts as Contextualized Entity-CentricLinked Data Graphs", 24th International Workshop on Database and Expert Systems Applications 2013.
[9] S. Chaudhuri, V. Ganti, and D. Xin, "Exploiting web search to generate synonyms for entities," in Proceedings of the International Conference on World Wide Web (WWW), 2009, pp. 151–160.
[10] E. Sapena, L. Padr´o, and J. Turmo, "Alias assignment in information extraction," ProcesamientodelLenguaje Natural, vol. 69, no. 39, pp. 105–112, 2007.

AUTHOR

Miss. Mayuri S. Bawane
is a student of M.E in Computer Science & Engineering from P.R.Pote COET,Amravati under SGBAU,India.

Prof.Monika Rajput,

AssitantantProfessor,Department of CSE, P.R.Pote(Patil) College of Engineering and Technology,Amravati, SantGadge Baba Amravati University,Amarvati,Maharashtra,India.