International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 1 Issue: 10 795 – 799

# A new Approach for Automatic Discovery of Association Orders between Name and Aliases from the Web using Anchor Texts-based Co-occurrences

Sk. Salamuddin, Ch.Padmini, S.SureshBabu, U.Janardhan Reddy
Asst.Professor, Vignan University, Guntur, AndhraPradesh, India.

***Abstract*—** An individual is typically referred by numerous name aliases on the web. Accurate identification of aliases of a given person name is useful in various web related tasks such as information retrieval, sentiment analysis, personal name disambiguation, and relation extraction. We propose a method to extract aliases of a given personal name from the web. Given a personal name, t

He proposed method first extracts a set of candidate aliases. Second, we rank the extracted candidates according to the likelihood of a candidate being a correct alias of the given name. We propose a novel, automatically extracted lexical pattern-based approach to efficiently extract a large set of candidate aliases from snippets retrieved from a web search engine. We define numerous ranking scores to evaluate candidate aliases using three approaches: lexical pattern frequency, word co-occurrences in an anchor text graph, and page-counts on the web.

*****

## I. INTRODUCTION

To construct a robust alias detection system, we integrate the different ranking scores into a single ranking function using ranking support vector machines. We evaluate the proposed method on three datasets: an English personal names dataset, an English place names dataset, and a Japanese personal names dataset. The proposed method outperforms numerous baselines and previously proposed name alias extraction methods, achieving a statistically significant mean reciprocal rank of 0.67 Experiments carried out using location names and Japanese personal names suggest the possibility of extending the prop osed method to extract aliases for different types of named entities and for different languages .For the further substantial improvement on recall and MRR from the previously proposed methods, our proposed method will order the aliases based on their associations with the name using the definition of anchor texts-based co-occurrences between name and aliases in order to help the search engine tag the aliases according to the order of associations. The association orders will automatically be discovered by creating an anchor texts-based co-occurrence graph between name and aliases. Ranking support vector machine (SVM) will be used to create connections between name and aliases in the graph by performing ranking on anchor texts-based co-occurrence measures. The hop distances between nodes in the graph will lead to have the associations between name and aliases. The hop distances will be found by mining the graph. The proposed method will outperform previously proposed methods, achieving substantial growth on recall and MRR.

**Existing System**

The existing namesake disambiguation algorithm assumes the real name of a person to be given and does not attempt to disambiguate people who are referred only by aliases.

*Disadvantage:*

1) To low MRR and AP scores on all data sets.
2) To complex hub discounting measure.

## II. Proposed System

The proposed method will work on the aliases and get the association orders between name and aliases to help search engine tag those aliases according to the orders such as first order associations, second order associations etc so as to substantially increase the recall and MRR of the search engine while searching made on person names. The term recall is defined as the percentage of relevant documents that were in fact retrieved for a search query on search engine. The mean reciprocal rank of the search engine for a given sample of queries is that the average of the reciprocal ranks for each query. The term word co-occurrence refers to the temporal property of the two words occurring at the same web page or same document on the web. The anchor text is the clickable text on web pages, which points to a particular web document. Moreover the anchor texts are used by search engine algorithms to provide relevant documents for search results because they point to the web pages that are relevant to the user queries. So the anchor texts will be helpful to find the strength of association between two words on the web. The anchor texts-based co-occurrence means that the two anchor texts from the different web pages point to the same the URL on the web. The anchor texts which point to the same URL are called as inbound anchor texts. The proposed method will find the anchor texts-based co-occurrences between name and aliases using co-occurrence statistics and will rank the name and aliases by support vector machine according to the co-occurrence measures in order to get connections among name and aliases for drawing the word co-occurrence graph. Then a word co-occurrence graph will be created and mined by graph mining algorithm so as to get the hop distance between name and aliases that will lead to the association orders of aliases with the name. The search engine can now expand the search query on a name by tagging the aliases according to their association orders to retrieve all relevant pages which in turn will increase the recall and achieve a substantial MRR.

**Algorithm:**

Keyword Extraction Algorithm: Matsuo, Ishizuka proposed a method called keyword extraction algorithm that applies to a single document without using a corpus. Frequent terms are extracted first, and then a set of co-occurrences between each term and the frequent terms, i.e., occurrences in the same sentences, are generated. Co-occurrence distribution showed the importance of a term in the document**.** However, this method only extracts a keyword from a document but not correlate any more documents using anchor texts-based co-occurrence frequency**.**

## III. MODULE DESCRIPTIONS

**a) Co-occurrences in Anchor Texts:** The proposed method will first retrieve all corresponding URLs from search engine for all anchor texts in which name and aliases appear. Most of the search engines provide search operators to search in anchor texts on the web. For example, Google provides In anchor or Allinanchor search operator to retrieve URLs that are pointed by the anchor text given as a query. For example, query on "*Allinanchor:Hideki Matsui"* to the Google will provide all URLs pointed by Hideki Matsui anchor text on the web.

**b) *Role of Anchor Texts***

The main objective of search engine is to provide the most relevant documents for a user's query. Anchor texts play a vital role in search engine algorithm because it is clickable text which points to a particular relevant page on the web. Hence search engine considers anchor text as a main factor to retrieve relevant documents to the user's query. Anchor texts are used in synonym extraction, ranking and classification of web pages and query translation in cross language information retrieval system.

**c) *Anchor Texts Co-occurrence Frequency***

The two anchor texts appearing in different web pages are called as inbound anchor texts if they point to the same URL. Anchor texts co-occurrence frequency between anchor texts refers to the number of different URLs on which they co-occur. For example, if p and x that are two anchor texts are co-occurring, then p and x point to the same

URL. If the co-occurrence frequency between p and x is that say an example k, and then p and x co-occur in k number of different URLs. For example, the picture of Arnold Schwarzenegger is shown in Fig 2 which is being liked by four different anchor texts. According to the definition of co-occurrences on anchor texts, *Terminator* and *Predator* are co-occurring. As well, *The Expendables* and *Governator* are also co-occurring.

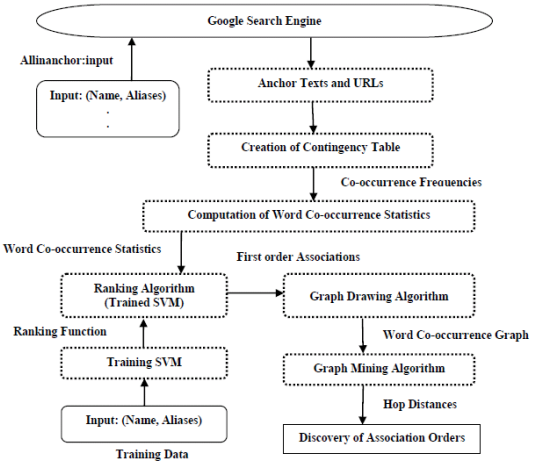
**d) Ranking Anchor Texts**

Ranking SVM will be used for ranking the aliases. The ranking SVM will be trained by training samples of name and aliases. All the co-occurrence measures for the anchor texts of the training samples will be found and will be normalized into the range of [0-1]. The normalized values termed as feature vectors will be used to train the SVM to get the ranking function to test the given anchor texts of name and aliases. Then for each anchor text, the trained SVM using the ranking function will rank the other anchor texts with respect to their co-occurrence measures with it. The highest ranking anchor text will be elected to make a first–order association with its corresponding anchor text for which ranking was performed. Next the word co-occurrence graph will be drawn for name and aliases according to the first order associations between them.

**e) Discovery of Association Orders**

Using the graph mining algorithm, the word co-occurrence graph will be mined to find the hop distances between nodes in graph. The hop distances between two nodes will be measured by counting the number of edges in-between the corresponding two nodes. The number of edges will yield the association orders between two nodes. According to the definition, a node that lies *n* hops away from *p* has an *n*-order co-occurrence with *p*. Hence the first, second and higher order associations between name and aliases will be identified by finding the hop distances between them. The search engine can now expand the query on person names by tagging the aliases according to the association orders with the name. Thereby the recall will be substantially improved by 40% in relation detection task. Moreover the search engine will get a substantial MRR for a sample of queries by giving relevant search results.

**Architecture**




## IV. CONCLUSION

The proposed method will compute anchor texts-based co-occurrences among the given personal name and aliases, and will create a word co-occurrence graph by making connections between nodes representing name and aliases in the graph based on their first order associations with each other. The graph mining algorithm to find out the hop distances between nodes will be used to identify the association orders between name and aliases. Ranking SVM will be used to rank the anchor texts according to the co-occurrence statistics in order to identify the anchor texts in the first order associations. The web search engine can expand the query on a personal name by tagging aliases in the order of their associations with name to retrieve all relevant results thereby improving recall and achieving a substantial MRR compared to that of previously proposed methods.

**REFERENCES:**

[1] Mann and D.Yarowsky, "Unsupervised personal name disam-
[2] biguation," in Proc. of CoNLL'03.
[3] R. Bekkerman and A. McCallum, "Disambiguating web appearances of people in a social network," in Proc. of WWW'05, 2005.
[4] G. Salton and M. McGill, Introduction to Modern Information Re-treival. New York, NY: McGraw-Hill Inc., M. Mitra, A. Singhal, and C. Buckley,

"Improving automatic query expansion," in Proc. of SIGIR'98

[5] P. Cimano, S. Handschuh, and S. Staab, "Towards the self annotating web," in Proc. of WWW'04.

[6] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An advanced social network extraction system," in Proc. of WWW'06.

[7] P. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in Proc. Of ACL'02

**About the Authors:**

Sk. Salamuddin worked as Asst.Professor at VIGNAN UNIVERSITY, Guntur. He has vast experience in Teaching field. He has author & co-author of different papers.

Ch.Padmini worked as Asst.Professor at VIGNAN UNIVERSITY, Guntur. She has vast experience in Teaching field.

S.SureshBabu worked as Asst.Professor at VIGNAN UNIVERSITY, Guntur. He has vast experience in Teaching field. He has author & co-author of different papers.

U.Janardhan Reddy worked as Asst.Professor at VIGNAN UNIVERSITY, Guntur. He has vast experience in Teaching field. He has author & co-author of different papers.