Automatic Discovery of Association Orders between Name and Aliases from the Web using Anchor Texts-based Co-occurrences

Rama Subbu Lakshmi B
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, India

Jayabhaduri R
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, 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 personal name from the web, as some of the web pages of the person may also be referred by his aliases. The aliases for a personal name are extracted by previously proposed alias extraction method. In information retrieval, the web search engine automatically expands the search query on a person name by tagging his aliases for complete information retrieval thereby improving recall in relation detection task and achieving a significant mean reciprocal rank (MRR) of search engine. 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.

General Terms
Information Retrieval, Relation Detection Task, Word Co-occurrences.

Keywords
Anchor Text mining, Graph Mining, Word Co-occurrence Graph.

1. INTRODUCTION

1.1 Information retrieval
This paper mainly deals with information retrieval system. Information retrieval is the area where users might search for documents, information within documents and metadata from documents on the web. Many users query might include retrieval of documents for personal names. Many celebrities and experts from various fields are referred by their original names on web. Most of the queries to web search engines include person names [1] [2]. 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 and experts might also be referred by their aliases on the web. Many web pages about person names might also be created by aliases. For example, a newspaper article might refer the persons using their original names, whereas a blogger might refer them using their nick names. The user will not be able to retrieve all information about a person if he only uses his personal name. To retrieve complete information about a person name, one might know about his aliases on the web. Various types of words are used as aliases on the web. Identifying aliases will be helpful in information retrieval. The aliases are extracted using previously proposed alias extraction method. The search engine expands the query on person names by tagging the extracted aliases to retrieve relevant web pages those are referred by original names as well as aliases thereby improving recall and MRR.

1.2 Outline of the proposed approach
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 [3]. 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

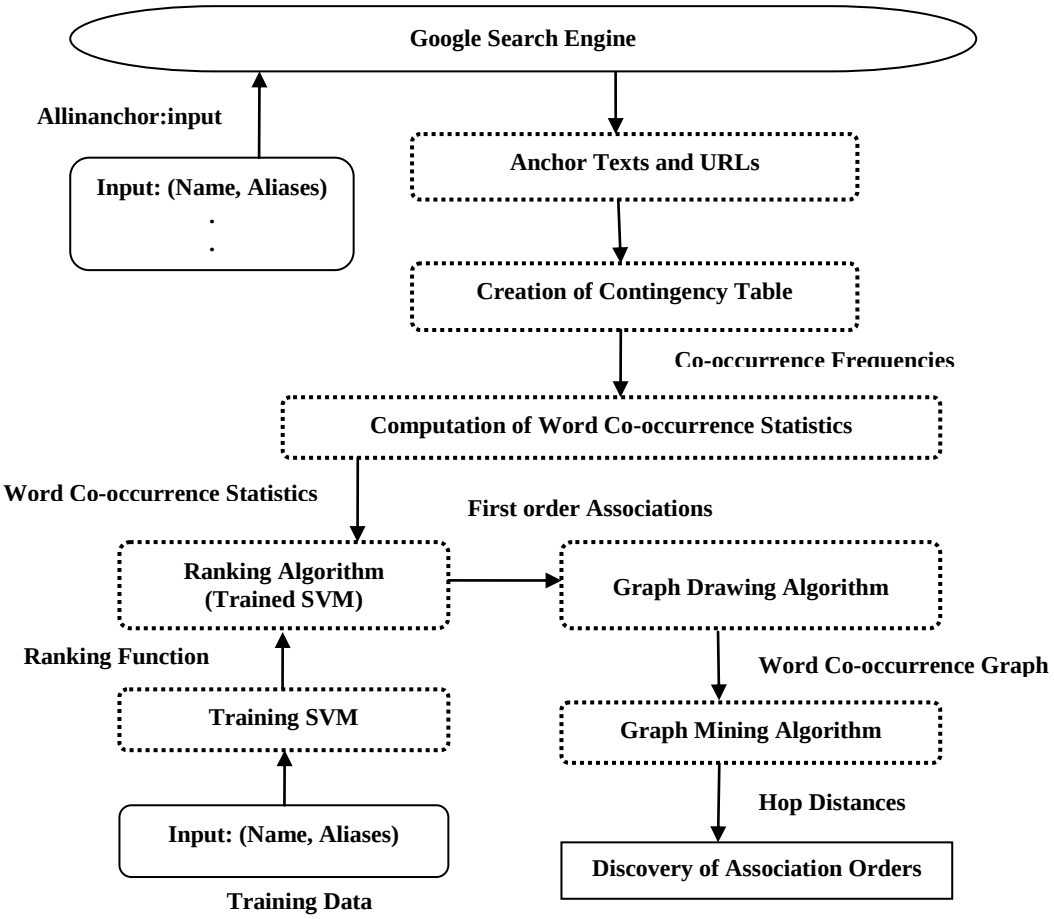

Fig 1: Outline of the proposed method

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.

2. RELATED WORK

2.1 Keyword Extraction Algorithm

Matsuo, Ishizuka [4] 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**.**

2.2 Transitive Translation Approach

Lu, Chien and Lee [5] proposed a transitive translation approach to find translation equivalents of query terms and constructing multilingual lexicons through the mining of web anchor texts and link structures. The translation equivalents of a query term can be extracted via its translation in an intermediate language. However this method did not associate anchor texts using the definition of co-occurrences.

2.3 Feature Selection Method

Liu, Yu, Deng, Wang, Bian [6] proposed a novel feature selection method based on part-of-speech and word co-occurrence. According to the components of Chinese document text, they utilized the words' part-of-speech attributes to filter lots of meaningless terms. Then they defined and used co-occurrence words by their part-of-speech to select features. The results showed that their method can select better features and get a more pleasant clustering performance. However, this method does not use anchor texts-based co-occurrences on words.

2.4 Data Treatment Strategy

Figueiredo et al. [7] proposed a data treatment strategy to generate new discriminative features, called compound-features for the sake of text classification. These c-features are composed by terms that co-occur in documents without any restrictions on order or distance between terms within a document. This strategy precedes the classification task, in order to enhance documents with discriminative c-features. This method extracts only a keyword from a

Table 1. Contingency Table for anchor Texts 'p' and 'x'.

Anchor Texts	x	C – {x}	C
p	k	n – k	n
V – {p}	K - k	N – n – K + k	N – n
V	K	N – K	N

document but not correlate any more documents using anchor texts.

2.5 Alias Extraction Method

Bollegala, Matsuo, and Ishizuka [3] 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 page counts, anchor text co-occurrence frequency, and lexical pattern frequency. However, this 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.

3. THE PROPOSED METHOD

The proposed method is outlined in Fig 1 and comprises four main components namely computation of word co-occurrence statistics, ranking anchor texts, creation of anchor text co-occurrence graph, and discovery of association orders. To compute anchor texts-based co-occurrence measures, there are nine co-occurrence statistics [3] used in anchor text mining to measure the associations between anchor texts: Co-occurrence Frequency (CF), term frequency – inverse document frequency (tfidf), Chi Square (CS), Log Likelihood Ratio (LLR), Pointwise Mutual Information (PMI), Hyper Geometric distribution (HG), Cosine, Overlap, and Dice. Ranking support vector machine (SVM) will be used to rank the anchor texts with respect to each anchor text to identify the highest ranking anchor text for making first order associations among anchor texts.

3.1 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 Inanchor 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.

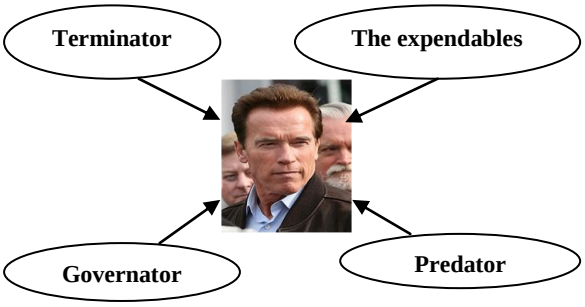

Fig 2: A picture of Arnold Schwarzenegger being linked by different anchor texts on the web

Next the contingency table will be created as described in Table 1 for each pair of anchor texts to measure their strength. Therein x and p are the two input anchor texts. C is the set of input anchor texts except p, V is the set of all words that appear in anchor texts, C-{x} and V-{p} are all the anchor texts except x and p respectively. Moreover, k is the co-occurrence frequency between p and x, whereas n is the sum of the co-occurrence frequencies between p and all anchor texts in C. K is the sum of co-occurrence frequencies between all words in V and x, whereas N is the sum of the co-occurrence frequencies between all words in V and all anchor texts in C.

3.1.1 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.

3.1.2 Anchor Texts Co-occurrence Frequency

The two anchor texts appearing in different web pages are called as inbound anchor texts [3] if they point to the same URL. Anchor texts co-occurrence frequency [3] 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.

3.1.3 Word Co-occurrence Statistics

To measure the association between anchor texts, nine popular measurements will be used and calculated from the Table 1.

3.1.3.1 CF

CF [3] is the simplest measurement among all and it denotes the value of k in the Table 1.

3.1.3.2 tfidf

The CF is biased towards highly frequent words. But tfidf [3] [8] resolves the bias by reducing the weight, that is, assigned to the words on anchor texts. The tfidf score for the anchor texts p and x is calculated from Table 1 as

$$tfidf\,(p,x) = k\log\left(\frac{N}{K+1}\right) \qquad (1)$$

3.1.3.3 CS

The Chi Square [3] is used to test the dependence between two words in natural language processing tasks. Given the contingency table in Table 1, the X^2 measure compares the observed frequency in Table 1 with the expected frequency for independence. Then it is likely that the anchor texts p and x are dependent if the difference between the observed and expected frequencies is large. The X^2 measure sums the difference between the observed and expected frequencies and is scaled by the expected values. The X^2 measure is given as

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

Where O_{ij} and E_{ij} are the observed and expected frequencies respectively. Using Equation (2), the X^2 score for anchor texts p and x from the Table 1 is as follows

$$CS(p,x) = \frac{N\{k(N-K-n+k)-(n-k)(K-k)\}^2}{nK(N-K)(N-n)} \quad (3)$$

3.1.3.4 LLR

LLR [3] [9] is the ratio between the likelihoods of two alternative hypotheses: that the texts p and x are independent or they are dependent. LLR is calculated using the Table 1 as follows

$$LLR(p,x) = k\log\frac{kN}{nK} + (n-k)\log\frac{(n-k)N}{n(N-K)} + (K-k)\log\frac{N(K-k)}{K(N-n)} + (N-K-n+k)\log\frac{N(N-K-n+k)}{(N-K)(N-n)} \quad (4)$$

3.1.3.5 PMI

PMI [3] [10] reflects the dependence between two probabilistic events. The PMI is defined for y and z events as

$$PMI(y,z) = \log_2\left(\frac{P(y,z)}{P(y)P(z)}\right) \quad (5)$$

Where P(y) and P(z), respectively, represent the probability of events y and z. Whereas P(y, z) is the joint probability of y and z. The PMI is calculated from Table 1 as

$$PMI(y,z) = \log_2\left(\frac{kN}{Kn}\right) \quad (6)$$

3.1.3.6 HG

Hyper Geometric distribution [3] [11] is a discrete probability distribution that represents the number of successes in a sequence of draws from a finite population without replacement. For example, the probability of the event that "k red balls are contained among n balls, which are arbitrarily selected from among N balls containing K red balls" is given by the hyper geometric distribution *hg*(N, K, n, k) as

$$hg(N,K,n,k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (7)$$

The hyper geometric distribution is applied to the values of Table 1 and the HG (p, x) is computed as the probability of observing more than k number of co-occurrences of p and x.

$$HG(p,x) = -\log_2\left(\sum_{i\geq k} hg(N,K,n,l)\right)$$
$$\max\{0, N+K-n\} \geq l \geq \min\{n,K\} \quad (8)$$

3.1.3.7 Cosine

Cosine [3] computes the association between anchor texts. The association between elements in two sets X and Y is computed as

$$\cos ine(X,Y) = \frac{|X \cap Y|}{\sqrt{|X|}\sqrt{|Y|}} \quad (9)$$

Where |X| represents the number of elements in set X. Considering X be the co-occurrences of anchor texts x and Y be the co-occurrences of anchor text p, then cosine measure from Table 1 is computed as

$$\cos ine(p,x) = \frac{k}{\sqrt{n}\sqrt{K}} \quad (10)$$

3.1.3.8 Overlap

The overlap [3] between two sets X and Y is defined as

$$overlap(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \quad (11)$$

Assuming that X and Y, respectively, represent occurrences of anchor texts p and x. The overlap of (p, x) to evaluate the appropriateness is defined as

$$overlap(p,x) = \frac{k}{\min(n,K)} \quad (12)$$

3.1.3.9 Dice

Dice [3] [12] retrieves collocations from large textual corpora. The Dice is defined over two sets X and Y as

$$Dice(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} \quad (13)$$

$$Dice(p,x) = \frac{2k}{n+K} \quad (14)$$

3.2 Ranking Anchor Texts

Ranking SVM [3] [13] 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.

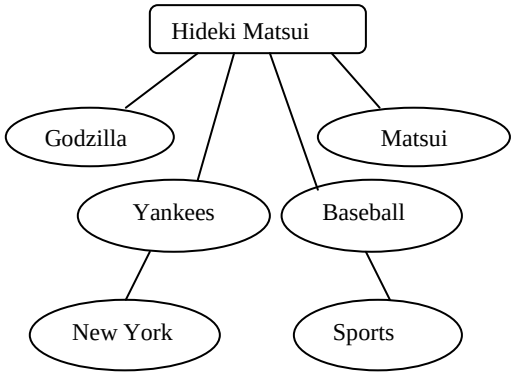

Fig 3: Word Co-occurrence graph for a personal name "*Hideki Matsui*"

3.3 Word Co-occurrence Graph

Word co-occurrence graph is an undirected graph where the nodes represent words that appear in anchor texts on the web. For each word in anchor text, a node will be created in the graph. According to the definition of co-occurrences if the two anchor texts co-occur in pointing to the same URL, then undirected edge will be drawn between them to denote their co-occurrences. A word co-occurrence graph like that shown in Fig 3 will be created for the name and aliases according to their first order associations among them. Each name and aliases will be represented by a node in the graph. The two nodes will be connected if they make first order associations between them. The edge between nodes will describe that the nodes bearing anchor texts co-occur according to the definition of anchor texts co-occurrences. Next the hop distance between nodes will be identified in order to have first, second, and higher order associations between name and aliases by graph mining algorithm.

3.4 Discovery of Association Orders

Using the graph mining algorithm [14] [15], 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.

4. DATA SET

To train and evaluate the proposed method, there are two data sets: the personal names data set and the place names data set. The personal names data set includes people from various fields of cinema, sports, politics, science, and mass media. The place names data set contains aliases for US states.

5. 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.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to the management of our institution for the valuable guidance and support.

7. REFERENCES

[1] J. Artiles, J. Gonzalo , and F. Verdejo, " A Testbed for People Searching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570, 2005.

[2] R. Guha and A. Garg, " Disambiguating People in Search," technical report, Stanford Univ., 2004.

[3] D.Bollegala, Y. Matsuo, and M. Ishizuka , "Automatic Discovery of Personal Name Aliases from the Web," IEEE Transactions on Knowledge and Data Engineering, vol. 23, No. 6, June 2011.

[4] Y. Matsuo, and M. Ishizuka," Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information", International Journal on Artificial Intelligence Tools, 2004.

[5] W. Lu, L. Chien and H. Lee, "Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach," ACM Transactions on Information Systems, Vol. 22, No. 2, Aprill 2004, Pages 242-269.

[6] Z. Liu, W. Yu, Y. Deng, Y. Wang, and Z. Bian," A Feature selection Method for Document Clustering based on Part-of-Speech and Word Co-occurrence," Proceedings of 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 10), pp. 2331-2334, Aug 2010.

[7] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M.A. Gonclaves, and W. Meira Jr, "Word Co-occurrence Features for Text Classification", Vol 36, Issues 5, Pages 843-858, July 2011.

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

[9] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, vol. 19, pp. 61-74, 1993.

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

[11] 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.

[12] F.Smadja, "Retrieveing Collocations from Text: Xtract," Computational Liguistics, Vol. 19, no 1, pp. 143-177, 1993.

[13] T. Joachims, ” Optimizing Search Engines using Clickthrough Data,” proc. ACM SIGKDD ’02, 2002.

[14] D. Chakrabarti and C. Faloutsos , “Graph Mining: Laws, Generators, and Algorithms,” ACM Computing Surveys, Vol. 38, March 2006, Article 2.

[15] C.C. Agarwal and H. Wang, “ Graph Data Management and Mining : A Survey of Algorithms and Applications,” DOI 10.1007/978-1-4419-6045-0_2,@ Springler Science+Business Media, LLC 2010.