Specific Personal Alias Withdrawal from Web and Clustering of Similar Web Documents

Snehal S. Shinde[1], Prakash R. Devale[2]

[1]Bharati Vidyapeeth University, College of Engineering, Dhankwadi, Pune, India

[2]Professor, Bharati Vidyapeeth University, College of Engineering, Dhankwadi, Pune, India

Abstract: ***There are many names available for a person, place or an entity on the web. If accurate alias of a particular individual is identified it becomes very useful in numerous web related tasks like information extraction, relation extraction, biomedical fields, sentiment analysis, personal name disambiguation, etc. Here, one method is projected based on referential ambiguity to find the correct alias for a given name. After accepting real name as input lexical patterns are achieved from the web. Candidate aliases are extracted with the help of these patterns. The candidate aliases are ranked using various ranking scores like co occurrence frequency, web dice, hub discounting, and degree distribution. This method improves the recall and attains a statistically considerable mean reciprocal rank. Using candidate aliases and data files, related web documents are bunched or grouped. Grouping achieves high accuracy and reduces the complexity.***

Keywords: Web mining, ranking, clustering, web text analysis, co-occurrence frequency.

1. Introduction

Generally the information of an individual is searched on the internet using his/her name. This task becomes difficult if a person of the Interest has pet names or name aliases. Various personalities in various fields have nicknames; even the places have two names. Many times the personalities are referred by their profession, drama or the book they have published. For example, in many articles Narendra Modi is referred as PM or Prime Minister, A. P. J. Abdul Kalam is known as The Missile Man of India, Kareena Kapoor is named as Bebo, Ajay Deogan as Singham, Sachin Tendulkar as Master Blaster, etc. The alias may be of one word or more than two words as mentioned above.

Recognizing the entities like person, place, festival name on the web becomes difficult for two basic reasons. First: If persons are considered, two or more different persons can have same name. This is called as Lexical Ambiguity. For example Marathi Actress Sonali Kulkarni. Second: A single person can have different names. This is called as Referential Ambiguity. For example Indian captail Dhoni is known as mahi or msd. Lexical ambiguity has been surveyed broadly in the preceding learning of name disambiguation; much less attention was acknowledged to the problem of referential ambiguity of entities.

Here to find aliases of persons a fully automatic method is proposed. Given a name and alias pair as input, a lexical pattern based method identifies the aliases with the help of snippets returned by web search engine. To select the best aliases from the extracted candidate aliases various ranking scores are proposed. Semantically meaningful groups of web pages are presented to the users as clusters.

2. Related Work

Correct alias finding is important in information retrieval. In [1], Danushka Bollegala proposed a method which uses extraction techniques to automatically extract significant entities such as the names of other persons, organizations and locations on each webpage. In that method for given person name, it extract person name from the web by using lexical pattern matching method and anchor text analysis. They ranked the candidate alias from the list. For this they integrated various similarity measures scores and given to a single function to support vector machine.

In [2] Dmitri proposed a method in which automatic entity extraction techniques are explained. In addition, it extracts and parses HTML and Web related data on each web page, such as hyperlinks and email addresses. Then this information is presented in an Entity Relationship Graph. This method is used to find relative information of a particular person on the web.

In [3], A. Bagga proposed a method that summarizes the interested entities and ranks the similarity of summaries using various information metrics.

In [4], T. Hokama proposed a method, especially for Japanese language.

In [5], C. Galvez proposed a method for extraction of abbreviations of personal names that measures approximate string matching algorithms.

In [6], Christian Borgelt explained how text classification is done using graph mining and also explained different graph parameters.

3. Method

The planned method is outlined in Fig. 1. It includes 4 main components: pattern extraction, candidate alias extraction, candidate ranking and clustering.

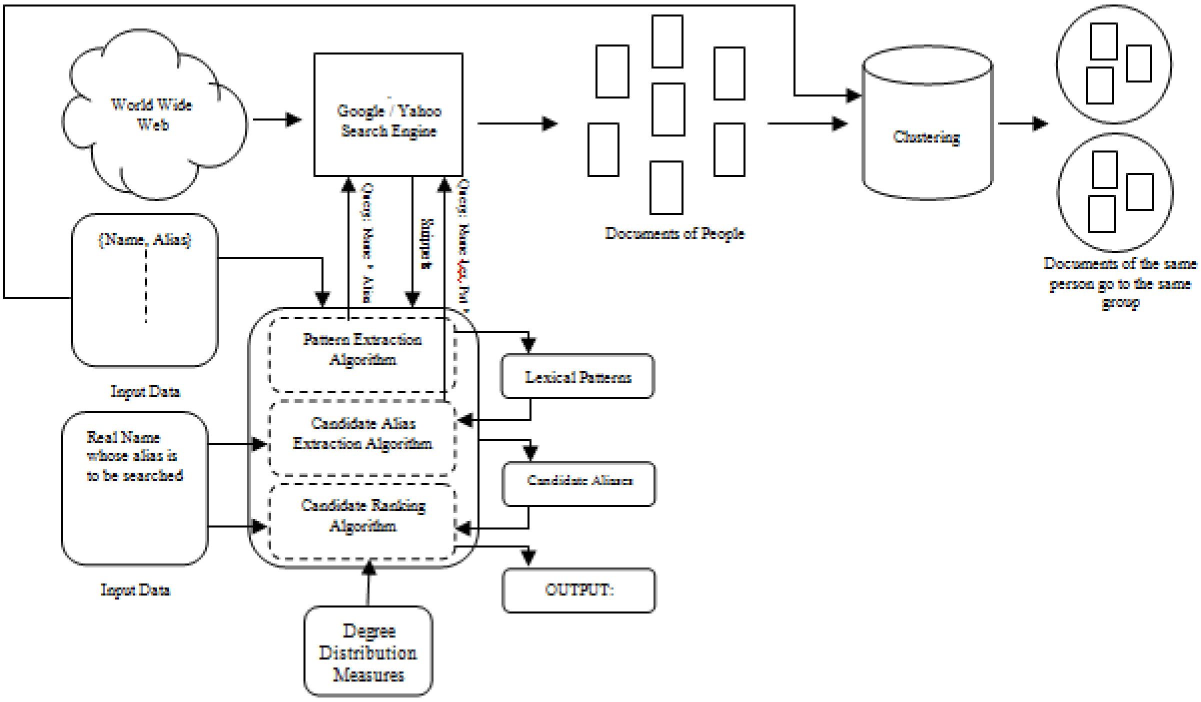
World Wide Web
Google / Yahoo Search Engine
Documents of People
Clustering
Documents of the same person go to the same group
{Name, Alias}
Input Data
Query: Name * Alias
Snippets
Query: Name aka, Pat *
Pattern Extraction Algorithm
Lexical Patterns
Real Name whose alias is to be searched
Candidate Alias Extraction Algorithm
Candidate Aliases
Candidate Ranking Algorithm
OUTPUT:
Input Data
Degree Distribution Measures

Figure 1: Outline of Planned System

After getting the input as name and alias pairs, Pattern Extraction Algorithm extracts lexical patterns which are supplied as input along with real name to Candidate Alias Extraction Algorithm. After applying degree distribution measures like lexical pattern occurrence, co-occurrence frequency and hub discounting on candidate aliases, correct aliases are identified. Clustering module takes name alias pair and html documents as input and semantically meaningful groups of web pages are presented to the users as clusters.

3.1 Take out Lexical Patterns from Snippets

Input to the system is training file which contains Name and Alias pair. For eg. Sachin Tendulkar * Master Blaster, Amitabh Bachhan * BigB, etc. The query in the form of Name * Alias is submitted to the search engine. Search engine returns the results in the form of snippets. In these snippets the given string is searched i.e. it is parsed. If such word is found, start and end of the query string is cropped and what remains is the lexical pattern.

For eg. Known as, aka, also known as, alias, famously known as, etc. these lexical patterns are saved in pattern file. The process is repeated for all the name alias pairs of input file and we get output as n number of patterns.

3.2 Find Candidate Aliases

Here we have to give the input as real name of the person whose alias is to be searched. One more input will be the pattern file which contains n number of patterns. First pattern from the file is taken and query is generated in the form of „Pattern RealName *‟. This query is given to the search engine. Search engine returns the results in the form of snippets. Again these snippets are parsed, cropped and what remains is the candidate alias. The process is repeated for all the n number of lexical patterns.

3.3 Ranking of Candidates

The candidate aliases extracted from previous module might have some invalid candidate aliases. From all these candidate aliases the most appropriate aliases of a given name must be identified. To model this problem various ranking scores are defined which calculate the connection between a name and a candidate alias using three diverse approaches: Lexical Pattern Occurrence, Co-Occurrence Frequency of Anchor Texts, Page-Count based Association Measure.

3.4 Lexical Pattern Occurrence

If a candidate alias and the real name of a person appear in many lexical patterns then that alias is a good alias for the real name. As a result, all candidate aliases are ranked in descending order of various lexical patterns in which they come into view with a real name.

3.5 Co-Occurrence Frequency of Anchor Texts

Anchor text is the highlighted hyperlinked text on the web page which is generally blue in color. If a real name *p* of interest and a candidate alias appear in different urls it is an indication that is really correct alias of name *p*. Fig. 2 shows a picture of Sachin Tendulkar being linked to by four diverse anchor texts.

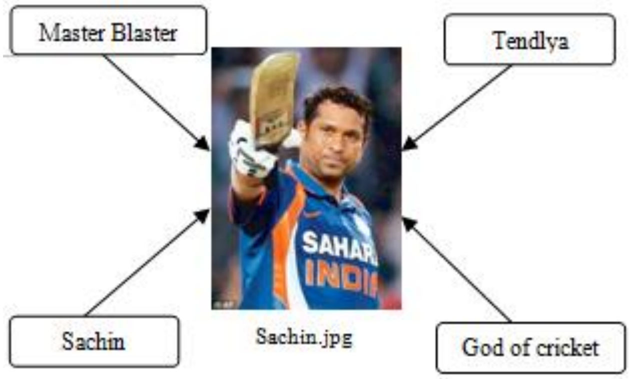

Figure 2: A picture of Sachin Tendulkar being linked by diverse anchor texts on the net.

3.6 Hub Discounting

A frequently observed phenomenon in case of web is that many web pages with different subject matters link to *hubs* like MSN, Yahoo, or Google. Two anchor texts might relate to a *hub* for totally diverse causes. So, co-occurrences coming from hubs are mostly noisy. Consider Fig. 3 mentioned below.

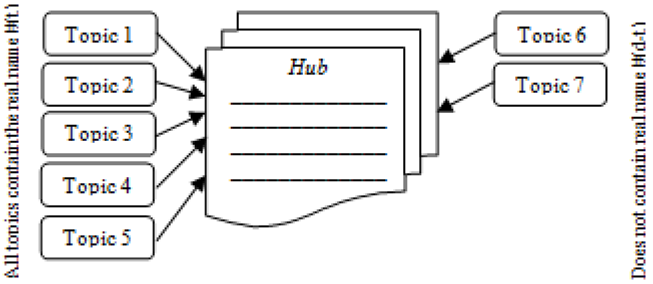
All topics contain the real name H(t)
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Hub
Topic 6
Topic 7
Does not contain real name H(d-t)

Figure 3: Discounting the co-occurrences in hubs.

Here the *hub* (web page) is linked to different sets of anchor texts. One set contains real name of interest and other does not contain real name but contains a variety of candidate alias. The confidence of a web page as a source of information rises if maximum anchor texts linked to it use the real name of interest.

While computing co-occurrence measures undesirable effect of a *hub h* can be overcome by

$$\alpha(h,p) = \frac{t}{k} \tag{1}$$

Where,
p: real name,
t: no. of inbound anchor texts of *h* that contain real name *p*,
k: total no. of inbound anchor texts of *h*.

Larger the value of *t*, more reliable the source of information, *h* for the real name *p*.
In contrast, if *k* is greater, then *h* can be considered as noisy and so gets discounted.

3.7 Page-Count based Association Measure: WebDice

When the query "*p and* x," for a name *p* and a candidate alias x is submitted to web search engine, it returns the page counts which can be viewed as an approximation of their co-occurrence in the web.

WebDice:
WebDice(p,x) between a name p and a candidate alias x using page counts can be computed as

$$WebDice(p,x) = \frac{2 * \text{hits}(\text{"}p \text{ AND } x\text{"})}{\text{hits}(p) + \text{hits}(x)} \tag{2}$$

4. Conclusion

For the extraction of aliases of a given name of interest a lexical pattern based approach is projected. To extract lexical patterns, as a training data, a set of names and aliases is used. Next, real name of the person of the interest is used with the lexical patterns to find the candidate aliases. To find accurate aliases among available candidate aliases, various ranking scores are applied. By finding the aliases, the referential ambiguity is removed. Using candidate aliases and data files, related web documents are grouped. Clustering achieves high accuracy and reduces the complexity.

References

[1] Danushka Bollegala, YutakaMatsuo and IitsuruIshizuka, Member , IEEE, "Automatic Discovery of Personal Name Aliases from the Web", *IEEE Transaction on knowledge and data engineering, vol. 23, no. 6,* June 2011.

[2] Dmitri V. Kalashnikov Zhaoqu Chen Rabia Nuray – Turan Sharad Mehrotra Zheng Zhang, "Web People Search via connection Analysis", *IEEE International Conference on Data Engineering,* 2009.

[3] A. Bagga and B. Baldwin, "Entity-Based Cross-Document Coreferencing using the vector space model", *Proc. Int's Conf. Computational linguistics (COLING '98), pp. 79-85,* 1998.

[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] C. Galvez and Fg. Moya-Anegon, "Approximate Personal Name Matching through Finite State Graphs", *J. Am. Soc. Fro Information Science and Technology, vol. 58, pp. 1-17,* 2007.

[6] Christian Borgelt, "Graph Mining: An Overview", *Proc, 19th GMA/GI Workshop Computational Intelligence, Germany,* 2009.

Author Profile

Mrs. Snehal S. Shinde is a student of M.Tech in Information Technology, Bharati Vidyapeeth Deemed University College of Engg, Pune-43.

Prof. Prakash Devale is Professor in Information Technology Dept., Bharati Vidyapeeth Deemed University College of Engineering, Pune-India. He did his B.E in Computer Science, M.E. from Bharati Vidyapeeth Deemed University College of Engineering, Pune in 2002 and Ph.D.Pursuing in BhartiVidyapeeth

Deemed University in the area of Machine Translation .He is having 21yrs of experience in teaching. His Publication in International Journals : 33, Publication in International Conference : 9,Publication in National Conference : 17. He is lifetime member of ISTE.