Clustering of Personalized Documents from the Web by Personal Name Aliases

Sucheta Kokate, D. G. Chougule, Manjiri Kokate

Abstract— The web is a huge resource for people who use engines to search documents related to specific person. The traditional approach is to organize search results into groups, one for each meaning of the query. According to the topical similarity of the retrieved documents, these groups are usually constructed but it is impossible for documents to be totally dissimilar and still correspond to the same person. To overcome this problem, in this paper we will implement a rigorous technique to find out all the documents regarding personalized information within short period of time. In this novel approach we propose a technique in which we cluster personalized documents from the web by personal name aliases. Given a personal name, the proposed method first extracts a set of candidate aliases and then clusters the documents by these aliases to achieve high accuracy and reduce the complexity.

Keywords - Web mining, information retrieval, web text analysis, searching, surfing.

I. INTRODUCTION

Web search is difficult because it is hard for users to construct queries that are both sufficiently descriptive and sufficiently discriminating to find the web pages that are relevant to the user's goal. Queries are often ambiguous: words and phrases are frequently polysemantic and user search goals are often narrower in scope than the queries used to express them. Search result sets contain distinct page groups due to ambiguity that meet different user search goals. Web page clustering is one approach for assisting users to both comprehend the result set and to refine the query. Web page clustering identifies groups of web pages related to the query by personal name aliases and represents these to user as clusters. In this paper we apply clustering is mechanism in which we have find out the aliases with the help of generic algorithms such as extract patterns, extract candidates and Ranking of Candidates to give the results on web within short period of time. We will apply a lexical pattern-based approach and extract aliases of a given name using snippets returned by a web search engine. This work presents a novel approach to find structured information from the vast repository of unstructured text which is to be placed on web server. We propose a fully automatic method to discover all the documents of a given personal name from the web by aliases. Web page clustering identifies semantically meaningful groups of web pages and presents these to the users as clusters. The clusters provide an overview of the contents of the result set and when cluster is selected the result set is refined to just the relevant pages in that cluster.

Manuscript received July 05, 2013

Sucheta Kokate, Department of Computer Science & Engineering, BSCOER, Pune, India,

D.G.Chougule, Department of Computer Science & Engineering, TKIET, Warananagar, India,

Manjiri Kokate, Department of Computer Science & Engineering, JSPM, Pune, India.

II. RELATED WORK

There has been huge boy of work on disambiguation, entity resolution. The goal of personal name disambiguation is to disambiguate various people that share the same name (namesakes). Given an ambiguous name, most name ambiguity algorithms have modeled the problem as one of document clustering in which all documents that discuss a particular individual of the given ambiguous name are grouped into a single cluster. In Web people search via connection analysis [1], it is proposed that if there are errors in clusters (multiple people into same cluster) the advantage of cluster based approach is not obvious. A novel algorithm is developed for disambiguating people that have same name. In web people search application, the main techniques used are unambiguation and entity resolution. The authors have pointing out that they rely primarily on analyzing object features for making their conference decisions by over viewing several existing entity resolution approaches. A. Bagga and B. Baldwin[2], proposed a cross-document co-reference resolution algorithm 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 by first performing within document co-reference resolution for each individual document, i.e. The problem of cross-document co-reference resolution is used to determine whether two mentions of a name in different documents refer to the same entity which is closely related to alias identification. But it is impractical to perform within document co-reference resolution to each document separately due to vastly numerous documents on the web, and to find aliases by clustering the documents. In Approximate string matching algorithms [3], variants or abbreviations of personal names (e.g., matching Will Smith with the first name initialized variant W. Smith) are extracted. For Comparing names, Rules in the form of regular expressions and edit-distance-based methods have been used. Many commercial proposed a method to learn a string similarity measure to detect duplicates in bibliography databases. So by using such string matching approaches we cannot identify aliases, which don't sharing any words or letters with the real

name of person i.e. approximate string matching methods would not identify Sachin Tendulkar as Master blaster an alias for Sachin Tendulakr. Hokama and Kitagawa [4], proposed an alias extraction method for Japanese language. They search for the query "*koto p" for a given name p, and extract the context that matches the asterisk. The meaning of Japanese word koto is also known as in English. However, in Japanese "koto" is a highly ambiguous word that has different meanings like incident, thing, matter, experience, and task. Many noisy and incorrect aliases are extracted using this pattern and so to filter out the incorrect aliases requires various post processing heuristics that are specific to Japanese language.

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka [7], proposed a method of alias extraction in which given a personal name, first extracts a set of candidate aliases and then extracted candidates are ranked according to the likelihood of a candidate being a correct alias of the given personal name.

III. OVERVIEW OF PROPOSED APPROACH

To overcome the limitations of previous information retrieval techniques, the goal is to group all the entity descriptions that refer to the same real world entities. In this paper we define accuracy in technically belongs algorithm such as extract patterns, extract candidates and Ranking of Candidates with referenced. The proposed work involves the design and implementation of cluster based web search system. The following are the steps of the overall approach, in the context of middleware architecture. A user submits a query to the middleware via a specialized web-based interface. The middleware queries a search engine with this query via the search engine API and retrieves a fixed no of relevant web pages. Then these web pages are processed and first we find out the different aliases of personal name. Then evaluate candidate aliases by using different ranking scores and finally the documents of the same people go to the same group. These aliases are used to find all documents of one person having different nicknames. Finally, we get several groups of documents; each group contains documents related to the same person.

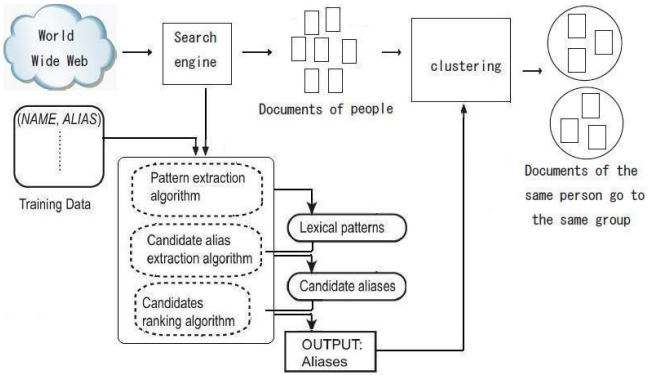

Figure 1. Outline of the proposed method

IV. CONCLUSION

In this paper we propose a cluster based web search approach that is based on personal name aliases in order to get better disambiguation quality. We use a lexical-pattern-based approach to extract aliases of a given name. Referential ambiguity is also removed by finding the aliases.

REFERENCES

1. D.V Kalashnikov, s.Mehrotra, R.N,Turen and Z.Chen,"Web People Search via connction analysis," IEEE transactions on knowledge and data engineering, vol. 20, no. 11, june 2008.
2. A. Bagga and B. Baldwin, "Entity-Based Cross-Document Co-referencing Using the Vector Space Model," Proc. Int'l Conf. Computational Linguistics (COLING '98), pp. 79-85, 1998.
3. C. Galvez and F. Moya-Anegon, "Approximate Personal Name-Matching through Finite-State Graphs," J. Am. Soc. for InformationScience and Technology, vol. 58, pp. 1-17, 2007.
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. J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570, 2005.
6. G. Mann and D. Yarowsky, "Unsupervised Personal Name Disambiguation," Proc. Conf. Computational Natural Language Learning (CoNLL '03), pp. 33-40, 2003.
7. Danushka Bollegala, Yutaka Matsuo, an d Mitsuru Ishizuka" Automatic Discovery of Personal Name Aliases from the Web", IEEE transactions on knowledge and data engineering, vol. 23, no. 6, june 2011