Information Sciences 276 (2014) 123–145

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Namesake alias mining on the Web and its role towards suspect tracking

Tarique Anwar [a,1], Muhammad Abulaish [a,b,*]

[a] Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia
[b] Department of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India

ARTICLE INFO

*Article history:*
Received 24 September 2012
Received in revised form 19 November 2013
Accepted 9 February 2014
Available online 20 February 2014

*Keywords:*
Web content mining
Web people search
Alias mining
Namesake disambiguation
Clustering

ABSTRACT

With the proliferation of social media, the number of active web-users is rapidly increasing these days. They create and maintain their personal web-profiles, and use them to interact with others in the cyber-space. Currently two major problems are being faced to automatically identify these web-users and correlate their web-profiles. First is the presence of *namesakes* on the Web, and the second is the use of *alias names*. In this paper, we propose a context-based text mining approach to discover alias names for all the namesakes sharing a common name on the Web, and leave the task of selecting the namesake of interest on part of the user. The proposed method employs a search-engine API to retrieve relevant webpages for a given name. The retrieved webpages are modeled into a graph, and a clustering algorithm is applied to disambiguate the webpages. Thereafter each obtained cluster standing for a namesake is mined for alias identification following a text pattern based statistical technique. The existing research works do not consider the presence of namesakes on the Web to mine aliases, which is impractical. The novelty of the proposed approach lies in discovering this drawback of existing works. Additionally the contribution includes the disambiguation technique that does not need to have a pre-determined number of clusters to be generated and the light-weight text pattern based alias mining technique. The number of clusters in the proposed method is rather determined dynamically by the inflation parameter, the pre-determination of which is comparatively much easier. Experimental results on different components demonstrate the robustness of the proposed alias mining approach. This paper also brings forth the significance of alias mining to the problem of suspect monitoring and tracking on the Web.



## 1. Introduction

With the widespread digitization of printed materials and the cheap and easy accessibility of Internet, the Web is growing rapidly both in scope as well as in depth [25]. Since last few years, the newer generation people are undergoing through a great revolution in their lives adopting several recent trends [33,27]. They find the Web as helpful, interesting and entertaining to interact with others whom they know well in their real life, and very often they even do not know. Sometimes these

* Corresponding author at: Department of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India. Tel./fax: +91 11 26980014.
*E-mail addresses:* tAnwar@swin.edu.au (T. Anwar), abulaish@ieee.org (M. Abulaish).

[1] Currently Tarique Anwar is a PhD Candidate in the Web and Data Engineering research group at Swinburne University of Technology, and a member of the Victoria Research Lab (VRL), National Information and Communication Technologies Australia (NICTA), Melbourne, Australia.

http://dx.doi.org/10.1016/j.ins.2014.02.050
0020-0255/

interactions are intended to perform some personal tasks (e.g. *online-shopping* to buy something needed). Quite often they are just to have some entertainment by exchanging thoughts with different people. Social media (e.g. *Facebook*, *Twitter*, *Youtube*, various weblogs, and so on) have become an important part of their lives. Creating and maintaining personal webpages are among the other activities which keep them intact with WWW. According to the statistics of *DoubleClick Ad Planner*,[2] social media sites are amongst the top ranking websites with largest number of visitors as well as largest number of page views. During the month of April in 2011, *Facebook* has remained on top with 880,000,000 unique visitors and 910,000,000,000 page views. Table 1 presents a list of top 10 websites visited in this month, along with the number of unique visitors and number of pages navigated. These numbers show the importance of their move towards social media. Thus, in addition to the real world, they have started being in a virtual world of WWW, which can be said as a reflection of their real world. The user-generated contents resulting from their frequent interactions with the Web has made it ever since the largest repository of electronically accessible data with a potential to reveal a lot of undiscovered crucial information about users' online behaviors and activities [3,28,42,47]. These online activities could further be used to infer their real life activities and involvements. However, due to the unstructured and unorganized nature of the available data, it is a challenging task to retrieve and integrate all these information collected from diversified source. Search engines, like *Google* and *Yahoo!*, make use of keywords of the given search query to find matches on the Web, and return a ranked set of webpages. For people search, this simple method is ineffective. Machine learning and data mining techniques need to be applied further on the returned results of search engines to analyze the information, and deduce some meaningful and relevant information about web-user profiles and activities.

As said in previous works [11,26], Web people search has become very common. Around 30% of Web search queries account for person names. However, due to the unstructured nature of Web contents, many ambiguities exist in the returned results. For the task of Web people search, there are two major problems currently being faced. (*i*) The first one is that, as the Web is a common unit of global access, just like the real world, multiple persons sharing the same name called *namesakes*, exist on the Web as well. It becomes difficult to consider them as different individuals. For example, on passing a search query on Google for the text "*tarique anwar*" to find exact matches, the top 20 returned results consist of pages referring to 12 different individuals with this name and it has no such functionality to mark them as of different persons. The Web is also having a dominating nature for famous personalities. As there is no such person with the name *tarique anwar* so popular, the results are highly varied. At the same time, a search for "*azim premji*", the famous Indian business tycoon and the chairman of Wipro Technologies, returned webpages with 18 out of top 20 referring to this individual directly. The remaining two referred to *azim premji university* and *azim premji foundation* which indirectly refer to the same person. (*ii*) The second problem is that, again like the real world, quite often a single person is known by multiple names on the Web called *alias names* or *mnemonic names*, and it becomes difficult to relate all the pages referring to the same individual by different alias names. For example, *Albert Einstein* on the Web is also known as the *father of modern science*, *Albert* and *Alby*. Sometimes these alternate names are used just because of their simplicity (e.g. *Alby*), or sometimes to highlight any specific aspect (*father of modern science*). Quite often they are also used by the person to represent himself or herself to a specific group of people who know the original identity, while it remains hidden to the rest of the public (e.g. using a nickname while discussing with others through any social media platform or chat server). In addition to these intentionally used alias names, it's quite common for the person to misspell a name which produces a different lexical structure than the actual name. Whatever be the reason of use, these aliases are of great importance to gather facts for a specific person. They can increase its scope of search by expanding the query after personalizing it [12], which in turn will increase its recall. Some other applications of aliases are using them in the form of metadata for Web entities to annotate them [13,48], for disambiguating Web entities [14] by annotating them by aliases, identifying relationships among entities in social media [32], and analyzing sentiments from comments on social media [18,34] by including alias names to identify the person.

In this paper, we propose a context-based approach to mine alias names of persons from the vast electronically accessible data on the WWW, taking into account the issue of namesakes. The system starts working with retrieving target webpages using Google API. A graph-based clustering technique is then applied on the retrieved pages to group them into different clusters, where each cluster is expected to correspond to a specific namesake. Thereafter, webpages from each of the clusters are processed by a light-weight alias mining algorithm to extract aliases for each namesake. The novelty of the proposed approach lies in application of the clustering technique for namesake disambiguation, the alias mining algorithm, and their integration to sort out person name ambiguities on the Web.

## 2. Related work

In this section, we present a brief review of web-mining and its applications to identify aliases on the Web. According to Kosala and Blockeel [25], web mining tasks can be categorized into three major areas – *web content mining*, *web usage mining*, and *web structure mining*. In addition to its huge size, the Web is also characterized by its dynamic and diverse nature, which calls for a continuous treatment in the due course of time. Research on this area has gained adequate attention of researchers bringing forth solutions to various kinds of novel problems, which range from recommender systems [15,52] and personalized search [43] to problems like spam filtering [51,41] and the question–answering systems [29,31]. The ambiguous nature of the Web is found a major hindrance in them. Very often different entities are designated by the same name and also vice

[2] http://www.google.com/adplanner/.