Extracting Key Phrases to Disambiguate Personal Names on the Web

Danushka Bollegala[1], Yutaka Matsuo[2], and Mitsuru Ishizuka[1]

[1] University of Tokyo
{danushka,ishizuka}@miv.t.u-tokyo.ac.jp
[2] AIST
y.matsuo@carc.aist.go.jp

Abstract. When you search for information regarding a particular person on the web, a search engine returns many pages. Some of these pages may be for people with the same name. How can we disambiguate these different people with the same name? This paper presents an unsupervised algorithm which produces key phrases for the different people with the same name. These key phrases could be used to further narrow down the search, leading to more person specific unambiguous information. The algorithm we propose does not require any biographical or social information regarding the person. Although there are some previous work in personal name disambiguation on the web, to our knowledge, this is the first attempt to extract key phrases to disambiguate the different persons with the same name. To evaluate our algorithm, we collected and hand labeled a dataset of over 1000 Web pages retrieved from Google using personal name queries. Our experimental results shows an improvement over the existing methods for namesake disambiguation.

1 Introduction

The Internet has grown into a collection of billions of web pages. One of the most important interfaces to this vast information are web search engines. We send simple text queries to search engines and retrieve web pages. However, due to the ambiguities in the queries and the documents, search engines return lots of irrelevant pages. In the case of personal names, we may receive web pages to other people with the same name (*namesakes*). However,the the different namesakes appear in quite different contexts. For example if we search for *Michael Jackson* in Google, among the top hundred hits we get a beer expert and a gun dealer along with the famous singer. However, the context in which the singer appears is quite different from his namesakes. However, context associated with a personal name is difficult to identify. In cases where the entire web page is about the person under consideration, the context could be the complete page. On the other hand the context could be few sentences having the specified name. In this paper we explore a method which uses terms extracted from web pages to represent the context of namesakes. For example, in the case of Michael Jackson, terms such as *music, album, trial* associate with the famous singer, whereas we