Efficient Web Pages Identification for Entity Resolution

Jia Zhu
School of ITEE
University of Queensland
Brisbane, Australia
jiazhu@itee.uq.edu.au

Gabriel Fung
School of ITEE
University of Queensland
Brisbane, Australia
gfung@itee.uq.edu.au

Xiaofang Zhou
School of ITEE
University of Queensland
Brisbane, Australia
zxf@itee.uq.edu.au

ABSTRACT

Entity resolution (ER) is a problem that arises in many areas. In most of cases, it represents a task that multiple entities from different sources require to be identified if they refer to the same or different objects because there are not unique identifiers associated with them. In this paper, we propose a model using web pages identification to identify entities and merge those entities refer to one object together. We use a classical name disambiguation problem as case study and examine our model on a subset of digital library records as the first stage of our work. The favorable results indicated that our proposed approach is highly effective.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval, Clustering

General Terms: Algorithms

Keywords: Entity Resolution, Web Pages Identification, Name Disambiguation

1. INTRODUCTION

Entity resolution (ER) is a problem that has been widely discussed recently with the information explosion of the World Wide Web. Finding information about entities, e.g. people, and determine which information refer to one entity are extremely difficult because their attributes like person names are not unique identifiers. Web pages usually contain rich information that can help us identify entities but is difficult to gain [1]. In a typical search from search engine, the results for this search are a mix of pages about the entity we search for. Our task is to propose an approach to determine which pages are useful to identify entities and which pages are noisy information. At the first stage of our research, we use a classical name disambiguation problem in DBLP as case study and propose a model to identify web pages without going inside the pages to extract information in order to improve the performance.

In the rest of this paper, we give a formal problem formulation and describe our model in Section 2, then present our experiments in Section 3. We also conclude this study and ongoing work in Section 4.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. WEB PAGES IDENTIFICATION MODEL

There are many cases nowadays that multiple entities are difficult to be identified because they do not have unique identifiers. For example, it is very common that several authors share the same name in a digital library. In DBLP[1], there are at least 60 different authors who call "Wei Wang" and there are more than 400 entries under this name. Unfortunately, only limited information is provided in DBLP so that the entries are difficult to be identified. Therefore, we focus on how to gain more information from external, e.g. World Wide Web, to identify these entries.

Assume there are a list of entities E, and each entity has a list of web pages W associated with it which we collect from search engine. If the page is useful to identify the entity, we mark it as U. Two entities might refer to the same object if they have high overlapping U web pages. The key point is how we know the web page is useful. Most of existing approaches need to go to read the page contents in order to identify the page. However, it will produce very high cost when there are millions of entries need to be identified and each page associated with the entry might contain megabyte size information. In addition, it is also difficult to locate information like 'address' from a page because of different writing styles. Therefore, we propose a model to perform web pages identification without going inside each page.

2.1 Name Disambiguation

At the first stage of our research, we select the classical name disambiguation problem in DBLP as case study. In definition, name disambiguation in digital libraries refers to the task of attributing the publications to the proper authors. In our previous work [2], a taxonomy based clustering framework has been given to enhance the function to solve the name disambiguation issue with more information can be used because the relationship data has been extracted from attribute data. However, the experiments in this work showed that it is necessary to involve stronger resource, e.g. web pages, to provide evidence to identify authors.

Refer to the previous model description; the name disambiguation problem in DBLP can be described as:

Assume there are a list of entries E that have the same author name, and suppose there are existing k actual authors $a_1, a_2, ..., a_k$ having the name x. The goal is to assign each entry $e, e \in E$ to their real author a_i. Given a list of pages P associated to each entry e, a model should be proposed that can select those pages UP which are useful to identify the entry e from P. If two entries' UP have high overlapping,

[1]http://www.informatik.uni-trier.de/ ley/db/

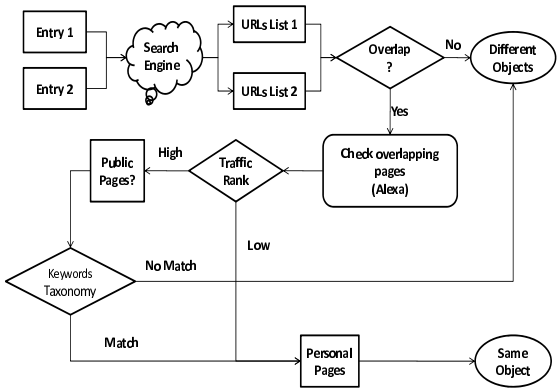

Figure 1: Proposed Model Flow Diagrams

then these two entries will be assigned to the same author a_i.

According to our observation, there are two types of web pages returning from search engine if we search the entry by using the author name plus the publication title as keyword, personal home pages and public pages, e.g. DBLP. Obviously, we are only interested in the former type of web pages because it is more sufficient that if two entries are found in a personal page than a public page. Therefore, we propose a model that can identify which pages are personal pages and details are given in the next section.

2.2 Proposed Model

Figure 1 shows the flow diagram of our proposed model. Assume there are two entries E_1 and E_2 need to be identified, and if they can be linked through their associated pages,then these two entries belong to the same author. Each entry has a list of URL returned from search engine. The process will continue only if there is any overlapping between two URL lists. Otherwise, these two entries refer to different authors.

For those overlapping pages, we query them in Alexa[2]. Alexa is a web information company which provides web site information like traffic rank and keywords for web pages. Because its web site information pages are very well organized, we can get the traffic rank and keywords for each page easily. We also set up a traffic rank range threshold β, which is generated according to the traffic rank of top 10 digital libraries, e.g. DBLP. If the page we query has the same or higher level of traffic rank to β, then our model takes this page as a public page, otherwise, the page will be treated a personal page.

However, some personal pages under the host of famous universities, e.g. Stanford University, have even higher traffic than public pages because Alexa calculates the traffic against the host. In order to solve this issue, we construct a taxonomy based on an existing education taxonomy[3] which contains terms like "University" by using the method in[2]. We propose a decision function $f(p) = \frac{N_t}{K}$ according to the taxonomy, where p is a page, N_t is the number of keywords associated with p and can be found in the taxonomy, and K is the total number of the keywords of p. If $f(p) > \theta$ where θ is a threshold value, then this page is a personal page.

[2]http://www.alexa.com/
[3]http://www.taxonomywarehouse.com/

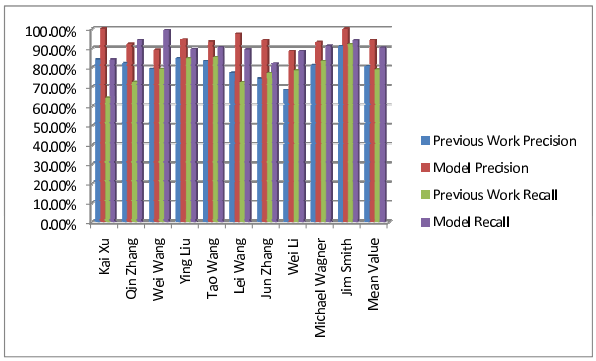

Figure 2: Comparisons of Precision and Recall

3. EXPERIMENT RESULTS

In our experiments, we use the same dataset previous work[2] and compare the web page identification model to it. Following the standard evaluation process, we use precision and recall value to evaluate our model and set the threshold $\beta = 100000, \theta = 0.6$:

$$Precision = \frac{PC}{PC+PIC}, \; Recall = \frac{PC}{PC+FIC}.$$

where PC is the number of pairs of entries being clustered correctly, PIC is the number of pairs being clustered incorrectly, and FIC is the number of pairs have not been clustered.

As shown in Figure 2, because of the strong evidence from personal pages, the proposesd model shows much higher precision and recall value than the previous model. Some mistakes are due to there is no shared page among entries or the web page information is not captured by Alexa.

4. CONCLUSIONS AND ONGOING WORK

In this paper, we have presented a web pages identification model for entity resolution by identify web pages without going inside the pages. We examine the model based on a classical name disambiguation problem in digital library. Experimental results indicate that the proposed model is effective.

Currently, we are conducting more experiments to refine our model by analyzing the web page description captured by search engine. The web page description contains many information can identify entities and we believe this work can improve the accuracy of name disambiguation and further extend to other entity resolution applications.

5. REFERENCES

[1] M. I. Lam and Z. Gong. Web information extraction. *Information Acquisition, IEEE International Conference*, vol. 27, 2005.

[2] J. Zhu, G. P. C. Fung and X. F. Zhou. A Term-based Driven Clustering Approach for Name Disambiguation. *Proc. Joint. APWeb/WAIM*, vol. 6, 2009.