Automatic Name Disambiguation of Bibliographic Data

Ting Chen
Department of Computer Sciences
University of Wisconsin, Madison

ABSTRACT

Person name is a commonly used but often ambiguous identifer. Bibliographic services like DBLP[3] or Citeseer [2] usually group all the publications with a common author name together for the ease of search. However, because different authors may happen to have the same name, publication records of non-identical researchers may mix up. In this project, we propose a novel application of disambiguating publication author names based on the information from a researcher's own homepage. We use a variation of the classic Naive Bayesian text classification method[4] to test if a publication belongs to the owner of a homepage. We envision that researchers who wants to disambiguate their publications on sites like DBLP can use our program on their homepages instead of have a hyperlink pointing directly to those bibliographic sites. Our experiment results shows the effectiveness of our approach.

1. INTRODUCTION

Person name is a commonly used identifier for people. However, it is well known that different people may have the same name and this causes problems in many services which rely on person names to differentiate people. As an example, many current bibliography services group publications under a common name for the ease of keyword search. However, due to the increasing number of oriental names appearing in research literature, it is quite common that several researchers may have the same common name and thus have their publication list mixed together. As an example, there are 65 papers listed under the name "Ting Chen" in the popular DBLP computer science biliographic service. However, as a graduate student, I (who happens to have such a popular name) contribute only 8 of them.

We observe that many researchers today have their own research homepages and they often want to provide a link to a complete and accurate list of their publications on their homepages. Currently they often do so by putting a hyperlink on their homepages to the sites like DBLP. As pointed out earlier, this may result in an inflated list of paper if they happen to have common name with other researchers. On the other hand, their homepages contain valuable information such as research fields and topics, co-authors which can be utilized during the disambiguation process. In this project, we tried to use researchers' homepage text information to disambiguate their publication lists on sites like DBLP. We envision that researchers who wants to disambiguate their publications on sites like DBLP can use our program on their homepages instead of have a hyperlink pointing directly to those bibliographic sites.

This report is organized as follows: Section 2 will explain the problem we try to solve. Section 3 presents a machine learning framework to solve the problem based on a variation of the classic Bayesian method. Section 4 shows our experiment results. Section 5 concludes the paper.

2. PROBLEM DEFINITION

We consider the following problem setting: suppose there are two webpages W_1 and W_2. W_1 is a webpage containing a list L of publications written by authors with name N. Note that these authors may not be the same person but they have the same name. W_2 is a webpage with information about a researcher R named N. Our task is to obtain a sublist L' of L in W_1 which contains each and every publication written by R.

An example of W_1 is a DBLP webpage listing all publications under an author name. An example of W_2 may be a researcher's homepage.

We assume that the structure of the publication list page W_1 is known in advance so that we could extract the publication list. On the other hand, we assume no prior knowledge of W_2 except the fact that it is an HTML page. We envision that our application could be deployed in such a manner that a researcher could use the information in his/her own webpage (i.e W_2) and a well known bibliography page (W_1)with all publications under his/her name to automatically obtain a list of his/her own publications. Thus it is reasonable to assume that we have prior knowledge about the structure of W_1 because there are only a limited number of bibliographic service websites. On the other hand, W_2 is different from researcher to researcher.

2.1 A machine learning formulation

The name disambiguation task can be modelled as a classic machine learning task naturally. The list of publications in a bibliographic webpage W_1 is the test data and researcher's own page W_2 provides training examples. We will discuss in detail how to obtain training examples from W_2 in the next section. Our task then is to classify each publication in the test data based on the training examples in W_2.

3. POSITIVE NATIVE BAYESIAN BASED AUTOMATIC NAME DISAMBIGUATION

Our name disambiguation task is a text classification problem for which the Bayesian framework is proven to be effective. In this project we use a variation of the classic Naive Bayesian method. Before we get into the details of our Bayesian solution, we first describe how to obtain training data from a researcher's homepage.

3.1 Extraction of training examples

Given a HTML document, we extract a set of training examples based on the idea of local proximity. More specifically, we assume words close to the name of the page owner are more likely to hold important personal information. For example, the self introduction section of a researcher's (say Dr. Tom Johnson) homepage may contain the following sentence: "Dr. Tom Johnson is an expert in machine learning". If we consider the 10 words before and after the owner's name, we could obtain the information that Dr. Tom Johnson is working on the machine learning field. This piece of information could be very useful in classifying publications because another author who happens to have the same name but works on a different field may have a quite different set of words in his/her publication records.

We use the following procedure to extract training examples from a webpage:

1. Find all occurrences of the page owner's name in the given webpage
2. For each name occurrence, get N words immediately preceding and following the name. The two $2N$ words then form a training example.

Notice that using the above approach, we could only get positive examples because we assume all close-by words are more likely to contain information related to the page owener.

3.2 Positive Native Bayesian based Name Disambiguation

Given a publication D for which one of its authors has name N and a set of positive training examples written by a researcher R named N, our task now is to find out if D is written by the researcher R. This is a typical binary classification problem [5] where the output class is either YES or NO. We use a variation of techique called Positive Naive Bayesian (PNB)[1] in our implemenation: the main difference in our approach is the treatment of negative examples.

Under the Bayesian learning framework, the most likely class of a given publication D is determined by the value

$$c = argmax_{v \in V} P(v|D) \tag{1}$$

$$= argmax_{v \in V} \frac{P(D|v)P(v)}{P(D)} \tag{2}$$

$$= argmax_{v \in V} P(D|v)P(v) \tag{3}$$

where $V = \{Yes, No\}$.

The value of $P(Yes)$ is the ratio of the number of publications the page owner has (usually the page owner has a good estimation) and the total number of publications listed in the bibliographic service. The value of $P(No)$ is $1 - P(Yes)$. In case the above ratio is not available, we just use $P(Yes) = P(No) = 0.5$.

The value of $P(D|Yes)$ is calculated by

$$P(D|Yes) = \prod_{i=1}^{n} P(w_i|Yes) \tag{4}$$

where w_i where $i \in \{1 \ldots$ are the words appearing in the given document D. The value $P(w_i|Yes)$ is estimated by

$$P(w_i|Yes) = \frac{1 + N(w_i, PD)}{Card(Voc) + N(PD)} \tag{5}$$

where PD is the set of positive examples; $N(PD)$ is the total number of words in PD and $N(w_i, PD)$ is the total number of word occurences of w_i in PD; $Card(Voc)$ is the size of vocabulary and it is used to smooth the probability value in case w_i does not occur in the training examples.

Because there is no negative example, we estimate the value of $P(D|No)$ by

$$P(D|Yes) = \prod_{i=1}^{n} P(w_i|Yes) \tag{6}$$

where

$$P(w_i|Yes) = \frac{1}{Card(Voc)} \tag{7}$$

We use this estimation because we do not have any knowledge about negative cases. An alternative estimation may be the word frequency in English text for each w_i but we didnot explore it further in our project due to lack of time.

4. EXPERIMENT

We conducted an experiment to disambiguate the DBLP publication list of the author of this paper (i.e. Ting Chen). Thus, the publication webpage W_1 used in the experiment is Ting Chen's DBLP publication page ($http://www.informatik.uni-trier.de/\tilde{l}ey/db/indices/a-tree/c/Chen:Ting.html$) and the researcher's homepage used is the current author's own page ($http://www.cs.wisc.edu/\tilde{t}chen$). The DBLP publication page has 65 listed paper under the name Ting Chen out of which only about 8 of them are contributed by this paper's author (a graduate student in University of Wisconsin, Madison).

4.1 Content of a sample author homepage

The content of the homepage at $http://www.cs.wisc.edu/\tilde{t}chen$ is shown below. There are four occurences of the name "Ting Chen": one appears as the page title and the rest in the section of selected publications. As shown, the surrounding texts contain such useful information about the page owner as occupation, field of study, research topics and co-author names. These information then will be utilized in the disambiguation process.

```
Ting Chen

Graduate Student,  Department of Computer Science,
                   University of Wisconsin,Madison.

email: Office:CS4354

-----------------------------------------------

I am a Graduate Student in Computer Science in
University of Wisconsin, Madison since 08/2005.
I obtained my B.Comp (1998-2002) and Master of
Science in Computer Sciences from School of
Computing, National University of Singapore.

My bio info can be found here [ps].

[My DBLP Entry] Note: Very unfortunately, my
publication records are mixed a Professor with
identical name (he is doing Computational
Biology).

Selected Publications
1. Eric Chu, A. Baid, Ting Chen, Anhai Doan,
Jeff Naughton " A Relational Approach to
Incrementally Extracting and Querying Structure
in Unstructured Data" to appear in VLDB 2007
2. Jiaheng Lu , Tok Wang Ling, Chee-Yong Chan,
Ting Chen "From Region Encoding To Extended
Dewey: On Efficient Processing of XML Twig
Pattern Matching " VLDB 2005
3. Ting Chen, Jiaheng Lu, Tok Wang Ling "On
Boosting Holism in XML Twig Pattern Matching
using Structural Indexing Techniques " SIGMOD
2005 [pdf ]

Last modification: 27 July 2007
```

4.2 Experiment Results

We implement the automatic name disambiguation program in Perl. The program consists of three main modules:

1. DBLP extractor, which extracts the list of publications under a certain name
2. Homepage extractor, which extracts the set of positive training examples about the page owner.
3. Positive Naive Bayesian Learner, which classifies the list of publications based on the training data obtained from a researcher's homepage.

The 10 words just before and after the homepage owner's name will form a positive example. Because the homepage is a HTML file, we exclude HTML tags from the 20 words. We set the value $Pr(Yes) = 0.15$ because I have 8 publications out of 65 in total.

Table 4.2 lists the disambiguation results of our program. The column No. Pubs shows the total number of publications classified as my publications whereas the column No. right Pubs shows the total number of publications really belonging to me.

The only parameter that we vary in the experiments is $Card(Voc)$ the size of vocabulary set.

$Card(Voc)$	No. Pubs	No. right Pubs	recall	precision
75	5	5	0.625	1.0
100	6	5	0.625	0.83
200	11	6	0.75	0.55
300	17	7	0.875	0.41
400	19	7	0.875	0.37
500	24	8	1.0	0.33

Table 1: Disambiguation results on Ting Chen's homepage and DBLP publication list.

As the results show, our Positive Naive Bayesian method significantly reduces the number of irrelevant publications (from 65 to 24 at most). At the same time, the methods achieve good recall rate ($> 60\%$ in all tests) while maintaining a acceptable level of precision ($> 33\%$ in all tests). We found setting $Card(Voc)$ as 200 achieves both good recall and precision.

5. CONCLUSION

In this project, we propose a novel application of disambiguating publication author names based on the information from a researcher's own homepage. We use a variation of Naive Bayesian text classification methods to test if a publication belongs to the owner of a homepage. Our experiment results shows the effectiveness of our approach: we dramatically reduce the number of irrelevant publications while maintaining a good comprise between recall and precision rates. We envision that researchers who wants to disambiguate their publications on sites like DBLP can use our program on their homepages instead of pointing directly to those bibliographic sites.

6. REFERENCES

[1] F. Denis, R. Gilleron, and M. Tommasi. Text classification from positive and unlabeled examples, 2002.

[2] Citeseer http://citeseer.ist.psu.edu/. Citeseer http://citeseer.ist.psu.edu/.

[3] Michael Ley. DBLP http://www.informatik.uni-trier.de/ ley/db/.

[4] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[5] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.