PERSIVAL Demo: Categorizing Hidden-Web Resources

Panagiotis G. Ipeirotis
Computer Science Dept.
Columbia University
pirot@cs.columbia.edu

Luis Gravano
Computer Science Dept.
Columbia University
gravano@cs.columbia.edu

Mehran Sahami
E.piphany, Inc.
sahami@epiphany.com

1. INTRODUCTION

The information available in electronic form continues to grow at an exponential rate and this trend is expected to continue. Although traditional search engines like AltaVista can address common information needs, they ignore the often valuable information that is "hidden" behind search interfaces, the so-called "hidden web."

Automating the classification of "hidden web" resources is challenging, since the contents of these collections are available only by querying, not by traditional crawling. For example, consider the PubMed medical database from the National Library of Medicine, which stores medical bibliographic information and links to full-text journals accessible through the web. This database is accessible through a query interface[1]. A query to PubMed with keyword "cancer" returns 1,313,266 matches, which are high-quality citations to medical articles, stored locally at the PubMed site. The contents of PubMed are not "crawlable" by traditional search engines. Thus, a query on AltaVista for all the pages in the PubMed site with keyword "cancer"[2] returns only 16,380 matches. Hence, techniques that need to have the documents available for inspection are not applicable to analyze and classify the "hidden web" resources.

The ability to access these resources and organize them for subsequent use is a central component of the Digital Libraries Initiative – Phase 2 (DLI2) project at Columbia University. The project is named PERSIVAL and its main goal is to provide personalized access to a distributed patient care digital library with all kinds of collections. The manual inspection and classification of these resources is a non-scalable solution, so we developed a novel technique to automate this task.

2. SYSTEM IMPLEMENTATION

In [2], we present a novel technique to automate the classification of searchable text databases. Our technique has two steps: a training step, followed by the actual database classification step.

In the training step, we start with a comprehensive, predefined topic hierarchy with an associated training set of preclassified documents. We then select the best features (i.e., words) for classification by using a feature selection algorithm that eliminates the words that have the least impact on the class distribution of documents [3]. Then, we train a rule-based document classifier [1] to produce rules like "IF transaminase AND hepatitis THEN Health". According to this rule, a document having the words "transaminase" and "hepatitis" will be classified into category "Health".

In the classification step, we transform each of these rules into a query probe (a query containing all the words in the antecedent of a given rule), and issue the queries to the databases that we want to classify, *extracting only the number of matches for each query*, without retrieving or inspecting any documents. The number of documents that match a specific query (e.g.,"transaminase AND hepatitis") at a database represents the number of documents that would match the corresponding classifier rule if we could run it over every document in the collection. After the probing phase, we classify the database based on the query-result statistics. As a result, our strategy efficiently produces an accurate collection classification using a small number of query probes.

In our demonstration, the classification process is completely interactive and controllable through an easy-to-use GUI. The user can pick a database to be classified, select among different classification schemes and change the parameters that affect the classification decisions. Finally, during probing an illustrative diagram displays the number of documents that are estimated to belong to each category in the database in question.

[1] `http://www.ncbi.nlm.nih.gov/PubMed/`
[2] The query is `cancer host:www.ncbi.nlm.nih.gov`.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
ACM 1-58113-345-6/01/0006.

3. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-98-17434. Panagiotis G. Ipeirotis is also partially supported by Empeirikeio Foundation.

4. REFERENCES

[1] W. W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of AAAI'96, IAAI'96*, pages 709–716, 1996.
[2] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, Count, and Classify: Categorizing Hidden-Web Databases. In *Proceedings of ACM SIGMOD 2001*, 2001.
[3] D. Koller and M. Sahami. Hierarchically Classifying Documents Using Very Few Words. In *Proceedings of ICML-97*, pages 170–178, 1997.