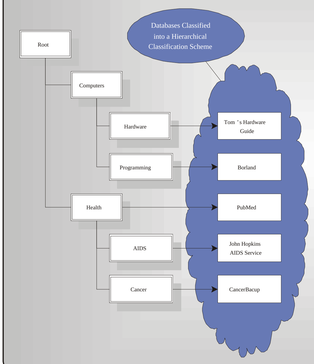
Abstract

The contents of many valuable web-accessible databases are only accessible through search interfaces and are hence invisible to traditional web "crawlers." Studies have estimated the size of this "hidden web" to be 500 billion pages, while the size of the "crawlable" web is only an estimated two billion pages. Recently, commercial web sites have started to manually organize web-accessible databases into Yahoo!-like hierarchical classification schemes.

We have developed QProber, a system that automates the classification of sites with hidden, but searchable, content. In fact our system works with any web-accessible document database, as long as its contents are searchable through a web-accessible form.

QProber automates this classification process by using a small number of query probes. The query probes are generated once, during the training phase. To classify a database, QProber uses these probes and sends them adaptively to the database in question. During the classification phase QProber does not retrieve or inspect any documents or pages from the database, but rather just exploits the number of matches that each query probe generates.

Frequently Asked Questions



Q: What is the Hidden-Web?
A: The content that is available through the web, but not indexed by search engines.

Q: Why do search engines not index these pages?
A: Because there are no static links pointing to these pages. The links are only generated as a response to a query submitted from a web-accessible form. Traditional crawlers cannot handle forms; hence they cannot find these pages.

Q: Why should we classify these document databases?
A: After classification, users can browse through the categories to find the databases of interest and then submit their queries there.

Q: How should we classify the databases?
A: Usually the documents in a database are about a specific topic. The database is assigned to the node(s) that best describe its contents (see figure).

Columbia University - PERSIVAL Project

QProber: Categorizing Hidden-Web Resources

http://qprober.cs.columbia.edu/

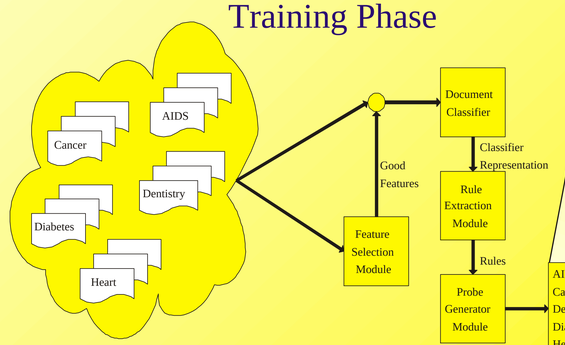
Panagiotis G. Ipeirotis, Luis Gravano, Mehran Sahami
[a] *Computer Science Department, Columbia University*
[b] *E.piphany Inc.*
Contact: Panagiotis G. Ipeirotis <pirot@cs.columbia.edu>

The **training phase** starts with a predefined topic hierarchy and an associated set of preclassified documents.

QProber then selects the best features (i.e., words) for classification by using an information-theoretic feature selection algorithm, and trains a document classifier to classify documents into the given classification scheme.

Based on the resulting classifier, QProber extracts classification rules like "***IF hepatitis AND transaminase THEN Health***". This rule classifies into the category "*Health*" all the documents that contain the words "*hepatitis*" and "*transaminase*".

Training Phase



All Query Probes for Categories
"AIDS," "Cancer," "Dentistry," "Diabetes," and "Heart."
(Total Number of Queries: 41, Maximum Query Length: 3)

```
Aids  IF Word ~ hiv Word ~ infection .
Aids  IF Word ~ aids Word ~ hiv .
Aids  IF Word ~ aids Word ~ hiv .
Aids  IF Word ~ aids Word ~ sexual .
Aids  IF Word ~ hiv .
Aids  IF Word ~ aids .
Cancer IF Word ~ cancer Word ~ chemo .
Cancer IF Word ~ cancer Word ~ orac Word ~ david .
Cancer IF Word ~ cancer Word ~ prostate .
Cancer IF Word ~ cancer Word ~ roda .
Cancer IF Word ~ physicians Word ~ roda .
Cancer IF Word ~ orac .
Cancer IF Word ~ radiation .
Cancer IF Word ~ pharma Word ~ death .
Cancer IF Word ~ soy .
Cancer IF Word ~ psa .
Cancer IF Word ~ rp .
Dentistry IF Word ~ dentist Word ~ tooth .
Dentistry IF Word ~ dental .
Dentistry IF Word ~ teeth Word ~ dentist .
Dentistry IF Word ~ tooth Word ~ crown .
Dentistry IF Word ~ teeth .
Dentistry IF Word ~ dentists .
Dentistry IF Word ~ dentist .
Dentistry IF Word ~ dentistry .
Dentistry IF Word ~ mercury .
Dentistry IF Word ~ tooth .
Diabetes IF Word ~ diabetes Word ~ mar .
Diabetes IF Word ~ diabetes Word ~ diabetic .
Diabetes IF Word ~ diabetes .
Diabetes IF Word ~ insulin .
Diabetes IF Word ~ diabetic .
Diabetes IF Word ~ bg .
Heart IF Word ~ heart Word ~ cardiologist .
Heart IF Word ~ heart Word ~ blood Word ~ surgery .
Heart IF Word ~ heart Word ~ cardiology .
Heart IF Word ~ heart Word ~ pressure .
Heart IF Word ~ heart Word ~ aspirin .
Heart IF Word ~ heart Word ~ cholesterol .
Heart IF Word ~ heart Word ~ chest .
Heart IF Word ~ cardiology Word ~ chung .
```

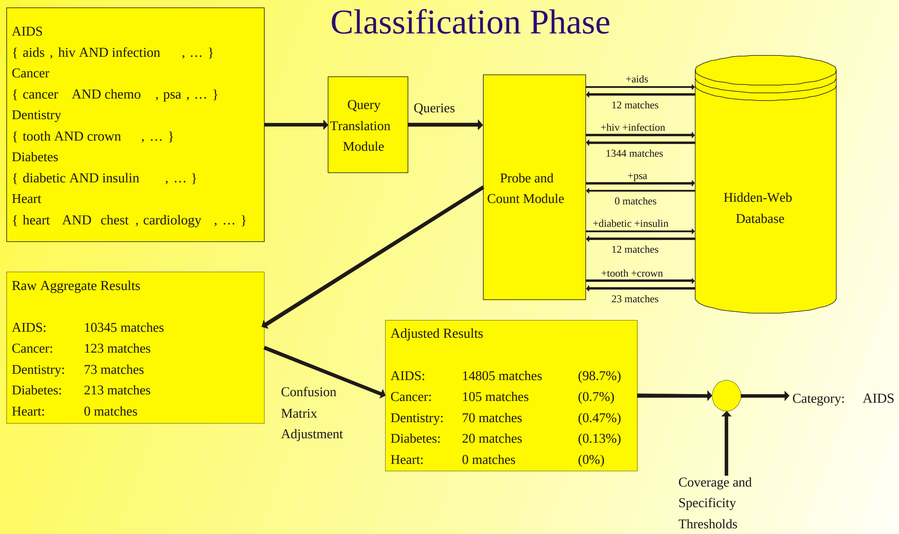
AIDS	{ aids , hiv AND infection , … }
Cancer	{ cancer AND chemo , psa , … }
Dentistry	{ tooth AND crown , … }
Diabetes	{ diabetic AND insulin , … }
Heart	{ heart AND chest , cardiology , … }

In the **classification phase**, QProber transforms each of these rules into a query, and issues the queries to the databases to classify.

Then QProber just monitors 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, QProber applies the Confusion Matrix Adjustment and, based on the resulting number of matches associated with each category, makes the proper classification decisions according to the given user preferences.

Classification Phase



Experimental Results

Training

- 72-node 4-level topic hierarchy from InvisibleWeb/Yahoo! (54 leaf nodes).
- Newsgroups assigned by hand to hierarchy nodes.
- 54,000 articles (1,000 articles per leaf) used to train RIPPER.
- 27,000 articles used to construct estimations of the confusion matrices.

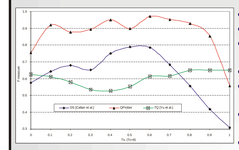
Data Sets

- 419,000 newsgroup articles used to build 500 **Controlled Databases**.
- 130 real **Web Databases** pre-classified from InvisibleWeb.

Alternatives for Comparison

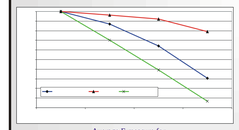
- DS: Random sampling of documents via query probes
  - Callan et al., SIGMOD'99
  - Different task: Gather vocabulary statistics
  - We adapted it for database classification
- TQ: Title-based Probing
  - Yu et al., WISE 2000
  - Query probes are simply the category names

Controlled Databases

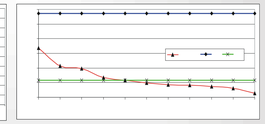


Average F-measure for
Different Values of the Ts Threshold (Tc=8)

- **Feature selection** helps.
- **Confusion-matrix adjustment** helps.
- **F measure above 0.8** - for most <*Tc, Ts*> combinations.
- Results **degrade gracefully** with hierarchy depth.
- Relatively **small number of probes** needed for most <Tc, Ts> combinations tried.
- Also, **probes are short**: 1.5 words on average; 4 words maximum.
- Both **better performance and lower cost** than DS [Callan et al. adaptation] and TQ [Yu et al.]

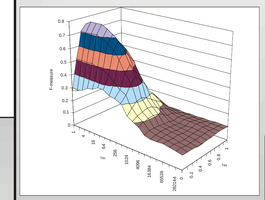


Average F-measure for
Different Hierarchy Depths



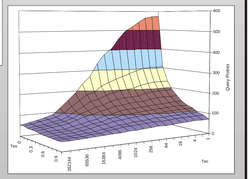
Average Cost for
Different Values of the Ts Threshold

Web Databases



Average F-measure for
Different Threshold Combinations

- **F-measure above 0.7** for best <Tc,Ts> combination learned.
- **185 query probes per database on average** needed for classification.
- Also, **probes are short**: 1.5 words on average; 4 words maximum.



Average Number of Query Probes for
Different Threshold Combinations

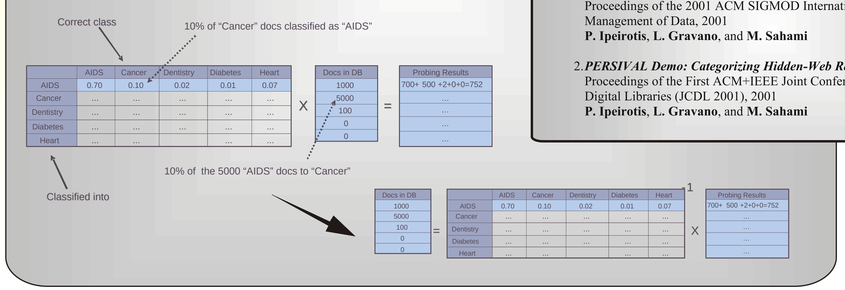
Confusion Matrix Adjustment

Probing is not perfect:

- Probes for one category might match documents of a different category.
- Documents might match multiple rules and be counted multiple times as separate documents.
- Some documents might not match any rule at all.

During training, QProber estimates the probing accuracy and keeps a "confusion matrix" with statistics about the errors.

Since QProber knows the usual errors, it applies a correction to the raw results and gets a better approximation of the real results.



References

1. ***Probe, Count, and Classify: Categorizing Hidden-Web Databases***
   Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 2001
   P. Ipeirotis, **L. Gravano**, and **M. Sahami**

2. ***PERSIVAL Demo: Categorizing Hidden-Web Resources***
   Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001), 2001
   P. Ipeirotis, **L. Gravano**, and **M. Sahami**