Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection

P.G. Ipeirotis & L. Gravano
Computer Science Department, Columbia University

Amr El-Helw
CS856
University of Waterloo
aelhelw@cs.uwaterloo.ca

Outline

- Introduction
- Contribution
- Background
- Focused Probing for Content Summary Management.
- Exploiting Topic Hierarchies for Database Selection.
- Experiments
- Summary

Introduction

- From a searcher's perspective, the web can be classified into:

"Visible" Web	"Hidden" Web
❑Static Documents with links ❑Can be crawled ❑Indexed by search engines (e.g. Google)	❑Data hidden in databases, behind search interfaces, with no link structure. ❑Cannot be crawled. ❑Not indexed by search engines.

Introduction

- Information in databases can be accessed through *metasearchers*.
- A metasearcher performs the following tasks:
  - Database selection (based on content summaries)
  - Query translation (to each specific database)
  - Result merging

Contribution

- A *document sampling technique* for text databases that results in high quality content summaries.
- A technique to estimate the *absolute* document frequencies of the words in the content summaries.
- A *database selection algorithm* that proceeds hierarchically over a topical classification scheme.
- A thorough, extensive experimental evaluation of the new algorithms using both “controlled” databases and 50 real web-accessible databases.

Background

Database Selection

- Find best databases to evaluate a given query.
- Based on information about the database contents (e.g. document frequency for each word, and total number of documents)
- Example: **bGlOSS** algorithm [1].
- These algorithms assume that content summaries are accurate and up-to-date.

Background

Uniform Probing for Content Summary Construction

- Callan et al. 1999, 2001 [2, 3]
- Extract a document sample from the database and compute the frequency for each observed word.
- Variants of this algorithm: **RS-Ord** and **RS-Lrd**.
- They compute sample document frequency *SampleDF(w)* for each word *w* that appeared in the retrieved documents, not the actual frequency in the database.

Background

Focused Probing for Database Classification

- Ipeirotis et al., 2001 [4]
- Classify the database in a hierarchy of topics, according to its documents.
- Define rules associating query word(s) with categories

  e.g. : *jordan AND bulls → sports*

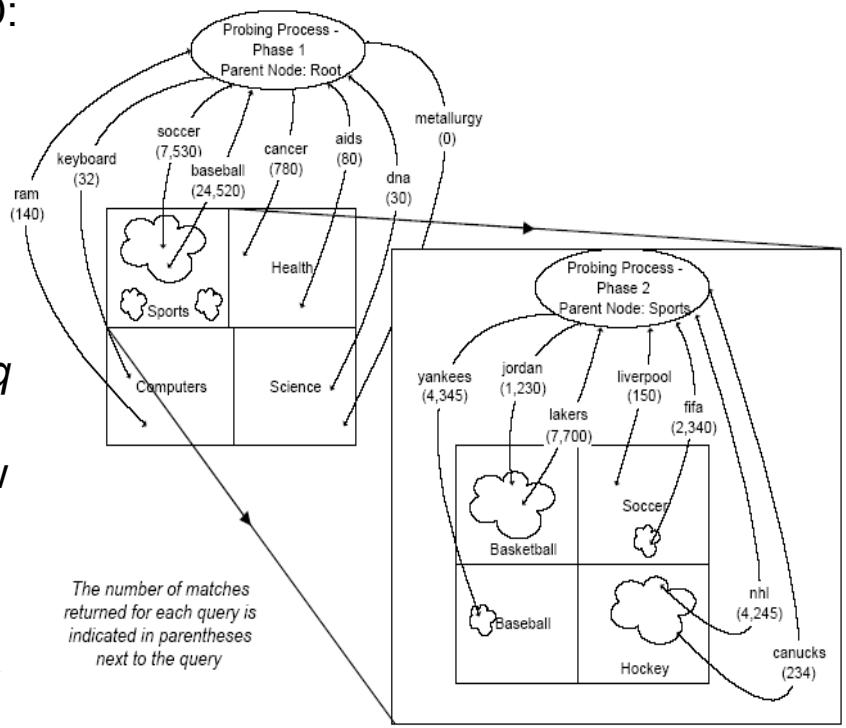
  hepatitis → health
- Rules can be learned automatically from a set of preclassified training documents.
- Categories can be divided into sub-categories.

Focused Probing for Content Summary Management

- Building Content Summaries from Extracted Documents

Starting with root category C, and database D:

- Probe database *D* with the query probes derived from the classifier for the subcategories of *C*
- For each probe *q:*
  - *retrieve top-k documents*
  - if *q is* a single word *w* then *ActualDF*(*w*) = #matches returned for *q*
- For each word *w* in the retrieved docs, SampleDF(w) = #documents that contain w
- For each subcategory C_i of C that satisifes coverage and specificity constraints:
  - Get content summary for C_i, and merge it with current content summary



Focused Probing for Content Summary Management

- Estimating Absolute Document Frequencies
  - Zipf (1949) and Mandelbrot (1988)
  - Mandelbrot's law: $f = P(r+p)^{-B}$

    r: rank of the document

    f: actual frequency of the document

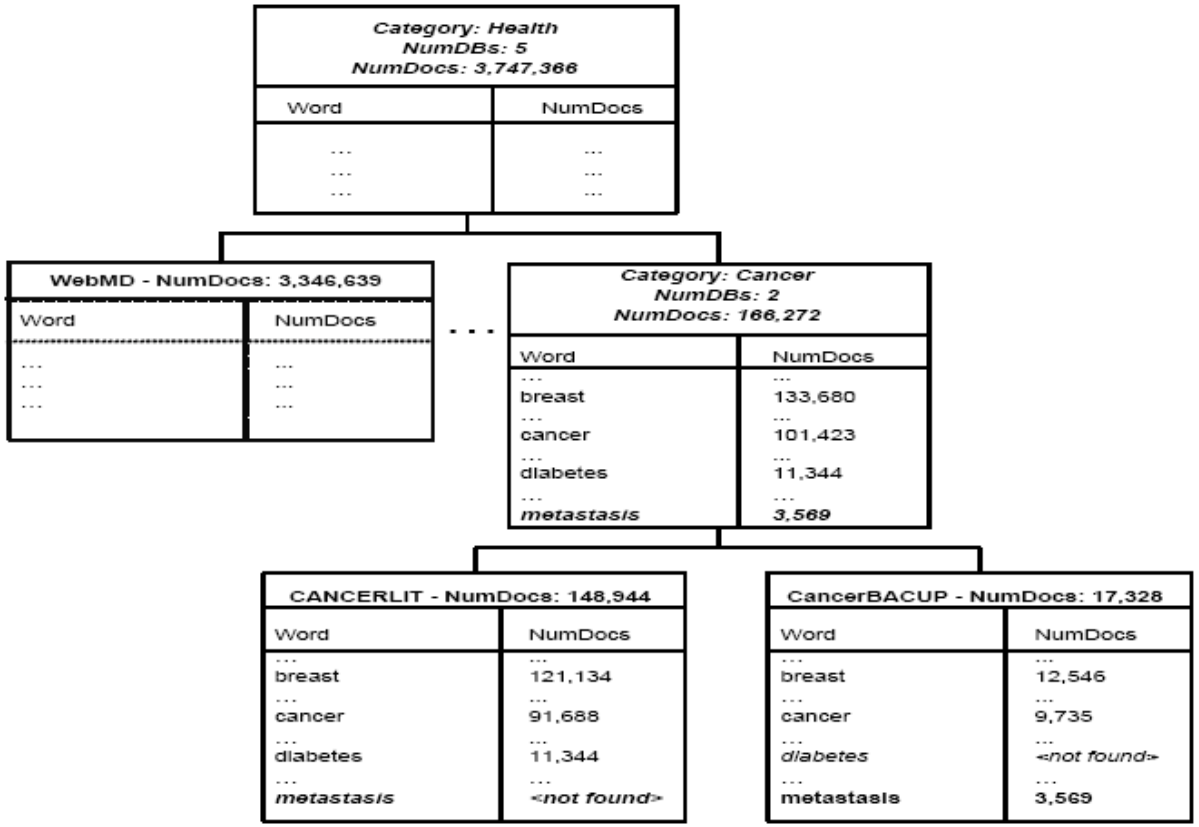
    P, ***B***, and ***p*** are parameters of the specific document collection.
  - The rank "r" can be computed from the sample frequencies obtained earlier.
  - The actual frequency can be estimated.

Exploiting Topic Hierarchies for Database Selection

- Database selection would suffer the most for queries with one or more words not present in content summaries.
- We can make use of the database categorization and content summaries to alleviate the negative effect of incomplete content summaries.
- This algorithm consists of two basic steps:
  - "Propagate" the database content summaries to the categories of the hierarchical classification scheme.
  - Use the content summaries of categories and databases to perform database selection hierarchically by zooming in on the most relevant portions of the topic hierarchy.

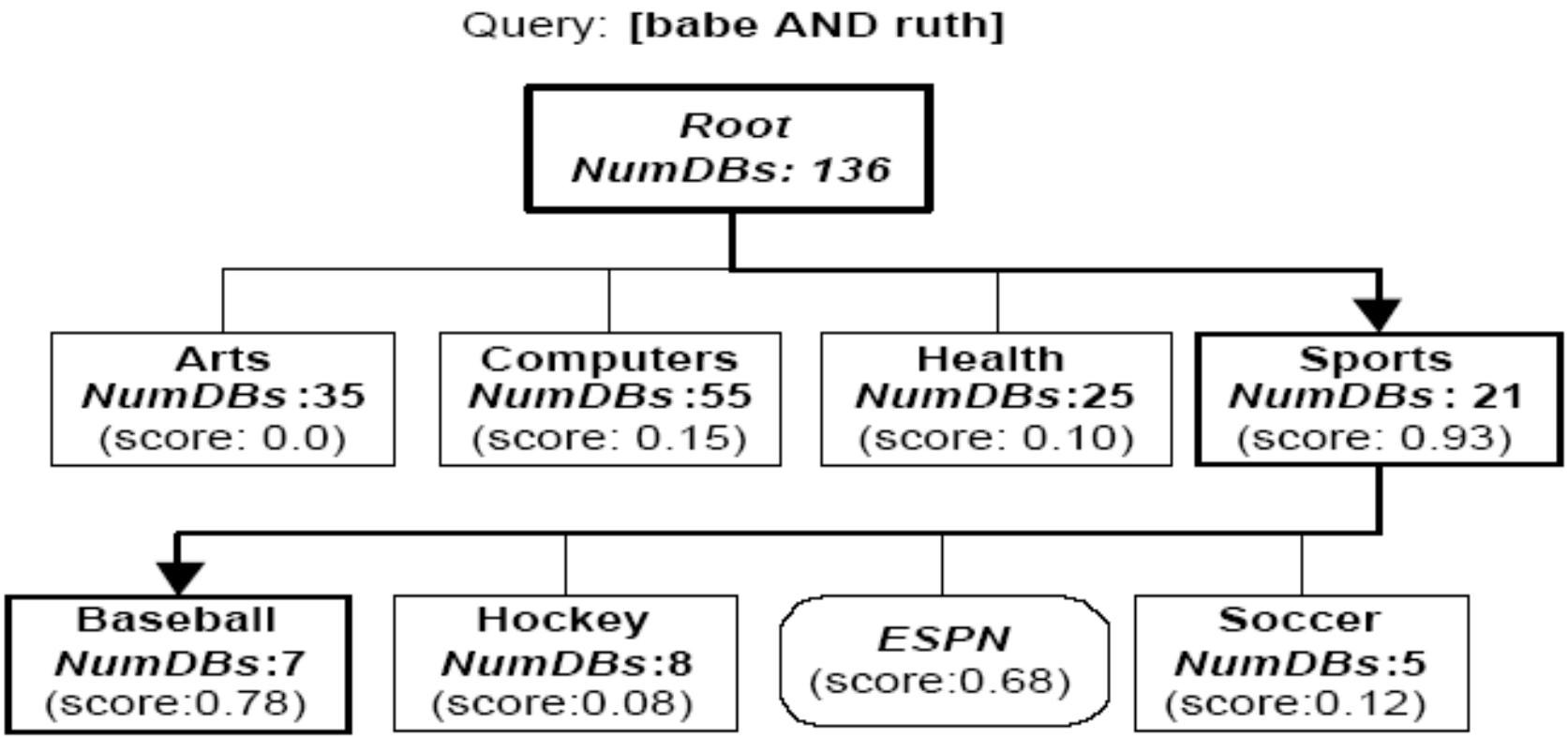
Exploiting Topic Hierarchies for Database Selection

- Creating Content Summaries for Topic Categories



Exploiting Topic Hierarchies for Database Selection

- Selecting Databases Hierarchically



Experiments

- Test Data:
  - **Controlled Database Set** (500,000 newsgroup articles from 54 newsgroups).
  - **Web Database Set** (50 real web-accessible databases).
- The experiments evaluate two sets of techniques:
  - Content-summary construction techniques.
  - Database selection techniques.

Experiments – Content Summary Construction

- The *Focused Probing* technique is tested against the two main variations of uniform probing (*RS-Ord* and *RS-Lrd)*.
- The following variations of the *Focused Probing* technique are considered (depending on the used classification technique):
  - **FP-RIPPER** (using RIPPER [5] as the base document classifier).
  - **FP-C4.5** (using C4.5RULES [6]).
  - **FP-Bayes** (using Naive-Bayes classifiers [7]).
  - **FP-SVM** (using Support Vector Machines with linear kernels [8]).

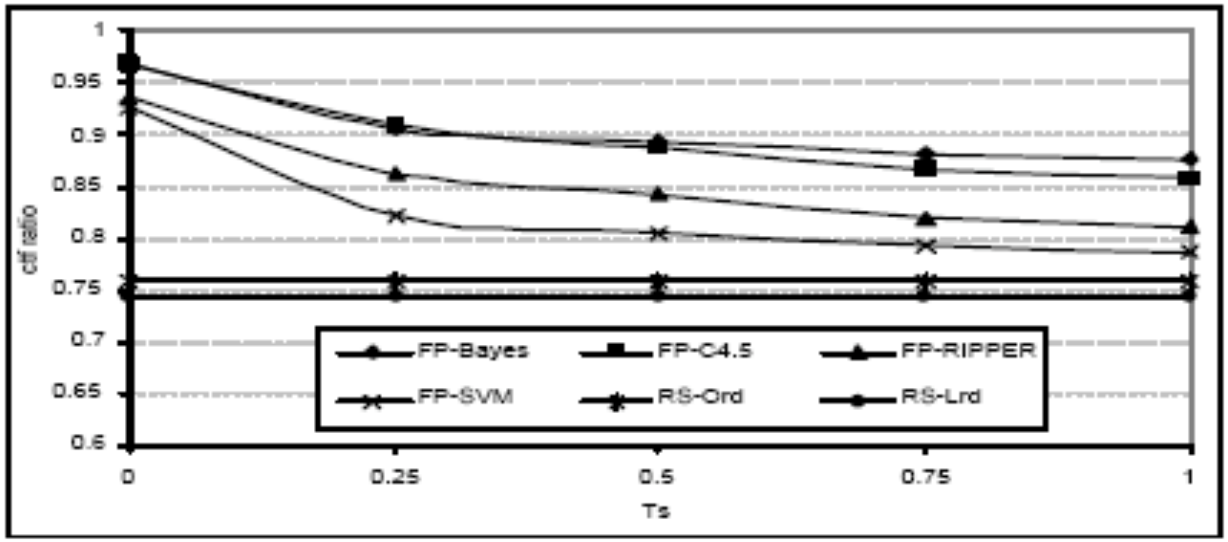
Experiments – Content Summary Construction

- **Coverage of the retrieved vocabulary**

$$ctf = \frac{\sum_{w \in T_r} ActualDF(w)}{\sum_{w \in T_d} ActualDF(w)}$$

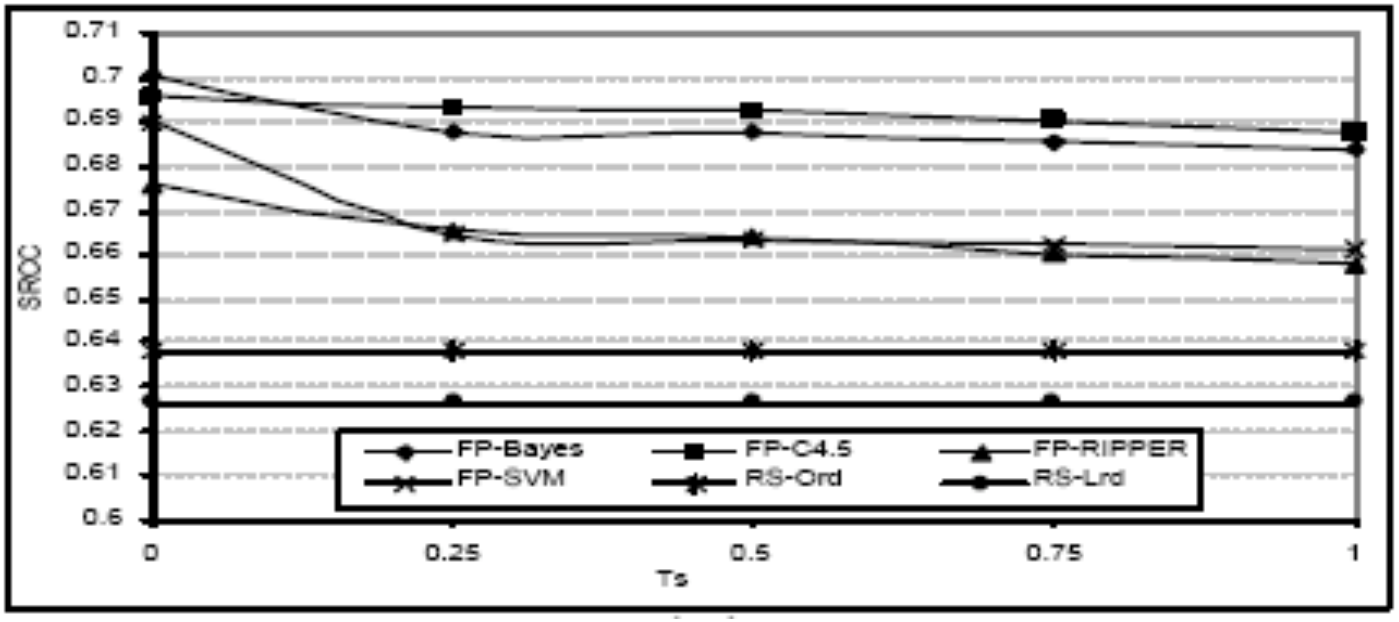
T_r: set of terms in a content summary

T_d: complete set of words in the corresponding database.



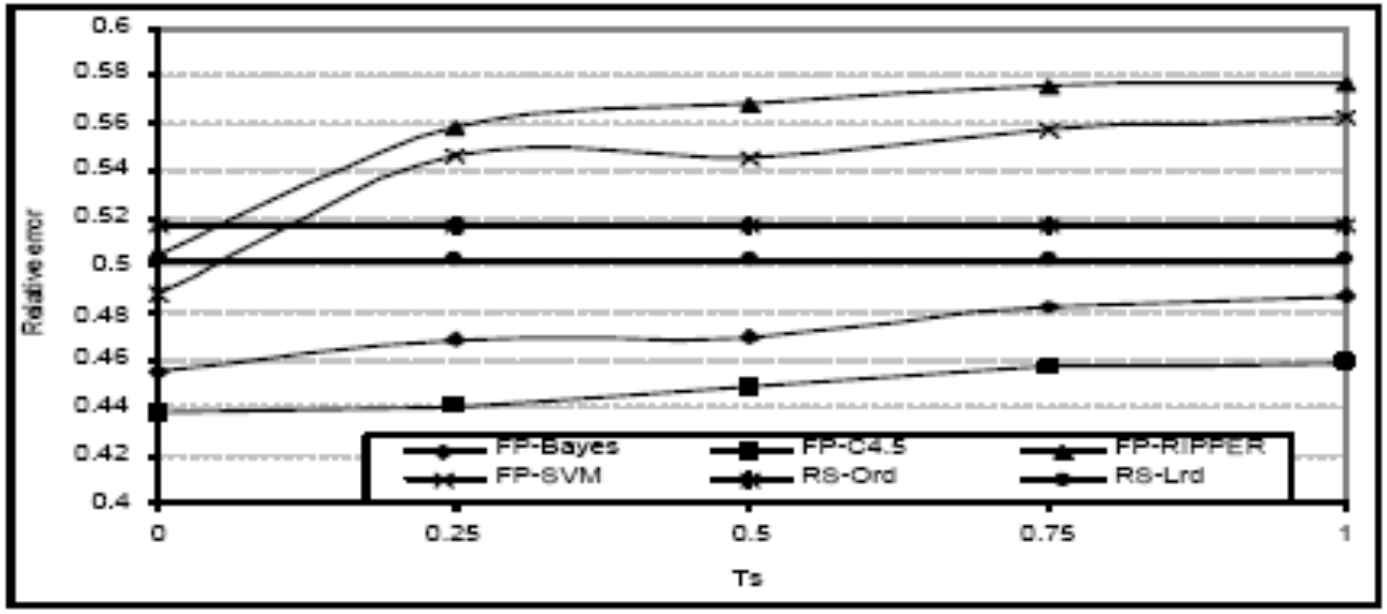
Experiments – Content Summary Construction

- **Correlation of word rankings**
  - *Spearman Rank Correlation Coefficient* (*SRCC*)



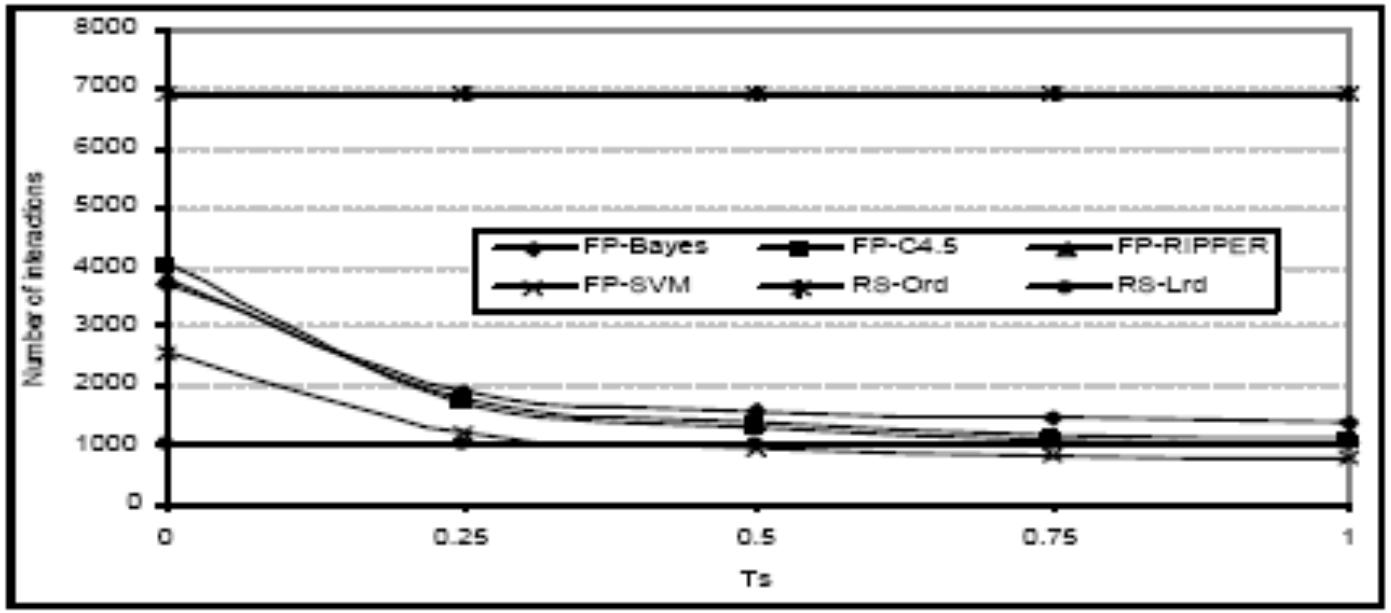
Experiments – Content Summary Construction

- **Accuracy of frequency estimations**
  - The average relative error for the *ActualDF* estimations for words with *ActualDF*> 3.



Experiments – Content Summary Construction

- **Efficiency**
  - *Number of interactions:* the sum of the number of queries sent to a database and the number of documents retrieved



Experiments – Database Selection

- Experiment procedure:
  - For each query pick 3 databases
  - Retrieve 5 documents from each database
  - Return 15 documents to user
  - Mark “relevant” and “irrelevant” documents
  - Precision (P_q) = |relevant documents in the answer| / |total number of documents in the answer|

Experiments – Database Selection

- Flat database selection algorithms used: CORI, bGlOSS
- Techniques compared: Focused Probing (FP-SVM), and Uniform Probing (RS-Ord and QPilot).

Technique	CORI		bGlOSS	
	Hierarchical	Flat	Hierarchical	Flat
FP-SVM	0.27	0.17	0.163	0.085
RS-Ord	–	0.177	–	0.085
QPilot	–	0.052	–	0.008

Summary

- This paper presents a novel and efficient method for the construction of content summaries of web-accessible text databases.
- The algorithm creates content summaries of higher quality than current approaches
- It categorizes databases in a classification scheme.
- The hierarchical database selection algorithm exploits the database content summaries and the generated classification to produce accurate results even for imperfect content summaries.
- Experiments showed that the proposed techniques improve the state of the art in content-summary construction and database selection.

References

[1] L. Gravano, H. Garcia-Molina, and A. Tomasic. *GlOSS*: Textsource discovery over the Internet. *ACM TODS*, 24(2), June 1999.

[2] J. P. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *SIGMOD'99*, 1999.

[3] J. Callan and M. Connell. Query-based sampling of text databases. *ACM TOIS*, 19(2), 2001.

[4] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: Categorizing hidden-web databases. In *SIGMOD 2001*, 2001.

[5] W. W. Cohen. Learning trees and rules with set-valued features. In AAAI-96,IAAI-96, 1996.

[6] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1992.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML-98*, 1998.

Comments…