IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Content Changes in Text Databases

Bhanu Shanker Prasad
P G Deptt of Statistics and Computer Application,TMBU,Bhagalpur,Bihar,India
Job Status:- M.D.N H/S SCHOOL,DUMRAMA ,AMARPUR,BANKA

**Abstract:-** Huge amounts of information are stored in web accessible text databases."Metasearchers" provide unified interfaces to query multiple such databases at once. For efficiency , metasearchers rely on succinct statistical summaries of the database contents to select the best databases for each query. So far , database selection research has largely assumed that database are static, so the associated statistical summaries do not need to change over time. However, databases are rarely static and the statistical summaries that describe their contents need to be update periodically to reflect content changes. In this paper , we first report the results of a study showing how the content summaries of 152 real web databases evolved over a period of 52 weeks. Then, we show how to use "survival analysis" techniques in general , and Cox's proportional hazards regression in particular, to model database changes over time and predict when we should update each content summary. Finally, we exploit our change model to devise update schedules that keep the summaries up to date by contacting database only when needed, and then we evaluate the quality of our schedules experimentally over real web databases.

## Introduction

A substantial amount of textual information on the web is stored in databases. While some databases are "crawlable" a significant fraction is not indexed by search engines. One way to provide one-stop access to the information in text databases (crawlable or not) is through *meta-searchers*, which can be used to query multiple databases simultaneously. The *database selection* step of the meta-searching process, in which the best databases to search for a given query are identified, is critical for efficiency, since a meta-searcher typically provides access to a large number of databases. The state-of-the-art database selection algorithms rely on aggregate statistics that characterize the database contents. These statistics, which are known as *content summaries* [Gravano et al. 1999] (or, alternatively, as *resource descriptions* [Callan 2000]), usually include the f*requency* of the words that appear in a database, plus perhaps other simple statistics such as the number of documents in the database. How to update these summaries, which provide sufficient information to decide which databases are the most promising for evaluating a given query, is the focus of this work.

So far, database selection research has largely assumed that databases are static. However, real-life databases are not always static and, accordingly, the statistical summaries that describe their contents need to be updated periodically to reflect database content changes. Defining schedules for updating the database content summaries is a challenging task, because the rate of change of the database contents might vary drastically from database to database. Furthermore, finding appropriate such schedules is important to keep content summaries up to date without overloading databases unnecessarily to regenerate summaries that are already (at least close to) up to date.

Now, we start by presenting an extensive study on how the content of 152 real web databases evolved over a period of 52 weeks. Given that small changes in the databases might not necessarily be reflected in the (relatively coarse) content summaries, we examined how these summaries change over time. Specifically, we analyzed the evolution of "complete" content summaries, which can be derived when we have full access to the database contents (e.g., via "crawlers" [Chakrabarti 2002]), as well as the evolution of "approximate" content summaries, which are used when database access is limited (e.g., as is the case for "hidden web" databases [Bergman 2001]). Our study shows that summaries indeed change and that old summaries eventually become obsolete, which then calls for a content summary update strategy.

In our approach for modeling content changes, we resort to the field of statistics named "survival analysis." Using the Cox proportional hazards regression model [Cox 1972], we show that database characteristics can be used to predict the pattern of change of the summaries. We then exploit our change models to develop summary update strategies that work well even under a resource constrained environment. Our strategies attempt to contact the databases only when needed, thus minimizing the communication with the databases. Our experimental evaluation, over 152 real web databases, shows the merits of our update strategies. Our experiments include a comparison with a technique from the literature developed for a different but related problem, namely how to decide when to recrawl (and update a search engine index of) crawlable web sites. We also develop and evaluate a machine learning approach for updating content summaries. Overall, our experiments show that our survival analysis approach significantly outperforms all the alternatives that we considered.

**Table I**

A fragment of the Content Summaries of Two Databases

| $D_1$ with $\|D_1\| = 51,500$ | |
|---|---|
| W | $f(w, D_1)$ |
| algorithm | 7,210 |
| cassini | 5 |
| Saturn | 2 |

| $D_2$with $\|D_2 = 5,73$ | |
|---|---|
| W | $f(w, D_2)$ |
| algorithm | 2 |
| cassini | 3,260 |
| Saturn | 3,73 |

## Background on Content Summary Construction

This section introduces the notation and necessary background .

We first define the notion of a *content summary* for a text database and briefly summarize how database selection algorithms exploit these summaries. Then, we review how to approximate database content summaries via querying.

***Definition 2.1. The content summary C(D) of a database D consists of:***

- *The actual number of documents in D, /D/, and*
- *For each word w, the number of D documents f (w, D) that include w.*

For efficiency, a metasearcher should evaluate a query only on a relatively small number of databases that are relevant to the query. The database selection component of a metasearcher typically makes the selection decisions using the information in the content summaries, as the following example illustrates:

***Example* 2.2.** Consider the query *[cassini saturn]* and two databases *D*1 and *D*2. Based on the content summaries of these databases (Table I), a database selection algorithm may infer that *D*2 is a promising database for the query, since each query word appears in many *D*2 documents. In contrast, *D*1 will probably be deemed not as relevant, since it contains only up to a handful of documents with each query word.

Database selection algorithms work best when the content summaries are accurate and up to date. The most desirable scenario is when each database either (1) is crawlable, so that we can (periodically) download its contents and generate content summaries, or (2) exports these content summaries directly and reliably (e.g., using a protocol such as STARTS [Gravano et al. 1997]). Unfortunately, the so-called *hidden-web* databases [Gravano et al. 2003], which abound on the web, are not crawlable and only provide access to their documents via querying; furthermore, no protocol is widely adopted for web-accessible databases to export metadata about their contents. Hence, it is generally not possible to extract the complete content summary of a hidden-web database. To characterize the

contents of a hidden-web database, an interesting observation is that we can easily extract document samples from the database via querying. In turn, we can approximate the content summary of the database using the documents in a sample. Here, we use the "hat" notation to refer to an *approximate, sample-based* content summary:

***Definition 2.3.*** An *approximate, sample-based content summary* $\hat{C}(D)$ of a database $D$ consists of:

- *An estimate* $|\hat{D}|$ *of the number of documents in D, and*
- *For each word w, an estimate* $\hat{f}(w, D)$ *of* $f(w, D)$.

The $\hat{C}(D)$ estimates are computed from a sample of the documents in $D$.

Here, we study two state-of-the-art strategies for constructing approximate, sample-based content summaries:

- *Query-Based Sampling (QBS), as presented by Callan and Connell [2001]: QBS starts by choosing words randomly from a dictionary and uses them to query a given database until at least one document is retrieved. Then, QBS continues to query the database using words that are randomly chosen from the retrieved documents. Each query retrieves up to k previously unseen documents (we set k = 4 in our implementation following the suggestions by Callan and Connell [2001], who experimented with other values of k as well). Sampling stops after retrieving sufficiently many documents (we stop after retrieving 300 documents, again following Callan and Connell [2001]). In our experiments, sampling also stops when 500 consecutive queries retrieve no new documents. (Getting no new results for 500 random queries is a signal that QBS might have retrieved the majority of the documents in the database.)*

- ***Focused Probing*** *(FPS), as presented by Ipeirotis and Gravano [2002]: Instead of sending (pseudo-) randomly picked words as queries, FPS derives queries from a classifier learned over a topic hierarchy. Thus, queries are associated with specific topics. For example, a query [breast cancer] is associated with the category "Health." We retrieve the top-k previously unseen documents for each query (we set k = 4 in our implementation, following the suggestions by Ipeirotis and Gravano [2002]) and, at the same time, keep track of the number of matches generated by each query. When the queries related to a category (e.g., "Health") generate a large number of matches, probing continues for its subcategories (e.g., "Diseases" and "Fitness"). The output of the algorithm is both an approximate content summary and the classification of the database in a hierarchical classification scheme. In our experiments, the queries are derived from an SVM classifier following the techniques described by Gravano et al. [2003], over the 72-node hierarchy also used by Ipeirotis and Gravano [2002] and Gravano et al. [2003]. In addition to the QBS and FPS approximate content summaries, we also study the evolution of the complete database content summaries (Definition 2.1), to which we will refer as complete (CMPL). To derive the complete content summary of a database, we retrieve all the documents from the database and compute the document frequency of each word. This technique requires that each database either allows direct access to its*

*documents or supports the functionality of a protocol such as STARTS [Gravano et al. 1997]. Next, we present the results of our study, which examined how CMPL, QBS, and FPS content summaries of 152 text databases changed over a period of 52 weeks.*

## STUDYING CONTENT CHANGES OF REAL TEXT DATABASES

Our One of the goals is to study how text database changes are reflected over time in the database content summaries. First, we discuss our data set in detail. Then, we report our study of the effect of database changes on the content summaries . The conclusions of this study will be critical , when we discuss how to model content summary change patterns.

### Data for our Study

Our study and experiments involved 152 searchable databases, whose contents were downloaded weekly from October 2002 through October 2003. These databases have previously been used in a study of the evolution of web pages [Ntoulas et al. 2004]. The databases in our study were–roughly–the five top-ranked web sites in a subset of the topical categories of the Google Directory, using the same topical categories as in Gravano et al. [2003]. Google Directory, in turn, reuses the hierarchical classification of web sites from the Open Directory Project. (Please refer to Ntoulas et al. [2004] for more details on the rationale behind the choice of these web sites.) From these web sites, we picked only those sites that provided a search interface over their contents, which are needed to generate samplebased content summaries. Also, since we wanted to study content changes, we only selected databases with crawlable content, so that every week we can retrieve the full database contents using a crawler. A complete list of the sites included in our experiments is available at http://webarchive.cs.ucla.edu/.

**Table II** shows the breakdown of web sites in the set by high-level DNS domain, where the *misc* category represents a variety of relatively small domains (e.g., *mil*, *uk*, *dk*, and *jp*). Similarly, Table III shows the breakdown of web sites by topical category, as assigned by the Google Directory. In this case, the *misc* category represents various small topical categories (e.g., world, shopping, and games).

We downloaded the contents of the 152 web sites every week for a period of one year. For each web site, we started our crawl from the root web page and continued to download pages–in breadth-first order–until either we exhausted all pages within the site or we downloaded 200,000 pages from the site.1 By following all the links recursively starting from the root page of each site we believe that we captured a relatively complete version of the contents of each site.2 Each weekly snapshot consisted of three to five million pages, or around 65 GB before compression, for a total over one year of almost 3.3 TB of history data. We treated each web site as a database, and created–each week–the complete content summary $C(D)$ of each database $D$ by downloading and processing all of

its documents. This data allowed us to study how the *complete* content summaries of the databases evolved over time. In addition, we also studied the evolution over

1. *Only four web sites were affected by this efficiency-motivated page-download limitation:* hti.umich.edu, eonline.com, pbs.org, *and* intelihealth.com.
2. *We are not aware of any site in our data set containing pages that are not reachable from the root pageof the site.*

**Table II**

Domain Distribution in Our Data Set

| **Domain** | **com** | **Edu** | **gov** | **misc** | **org** |
|---|---|---|---|---|---|
| % | 47.3% | 13.1% | 17.1% | 6.8% | 15.7% |

**Table III**

Category Distribution in Our Data Set

| Category | % | | Category | % |
|---|---|---|---|---|
| computer | 22.5% | | computer | 7.3 |
| science | 17.2% | | science | 53 |
| health | 9.9% | | health | 4.0% |
| arts | 8.6% | | arts | 4.0% |
| regional | 7.9% | | regional | 2.0% |
| society | 7.3% | | society | 4.0% |

time of an *approximate* content summary ˆ*C*(*D*) of each database *D*, computed weekly3 using either *QBS* or *FPS*. To reduce the effect of sampling randomness in our *QBS* experiments, we create five *QBS* content summaries of each database each week, in turn derived from five document samples, and report the various metrics in our study as averages over these five summaries.

**Measuring Content Summary Change**

We now turn to measuring how the database content summaries–both the complete and approximate versions–evolve over time. For this, we resort to a number of metrics of content summary similarity and quality from the literature. We discuss these metrics and the results for the 152 web databases next. For our discussion, we refer to the "current" and complete content summary of a database $D$ as $C(D)$, while $O(D, t)$ is the complete summary of $D$ as of $t$ weeks into the past. The $O(D, t)$ summary can be considered as an (old) approximation of the (current) $C(D)$ summary, simulating the realistic scenario where we extract a summary for a database $D$ and keep it unchanged for $t$ weeks. In the following definitions, $Wo$ is the set of words that appear in $O(D, t)$, while $Wc$ is the set of words that appear in $C(D)$. Values $fo(w, D)$ and $fc(w, D)$ denote the document frequency of word $w$ in $O(D, t)$ and $C(D)$, respectivel

***( a) Recall.***

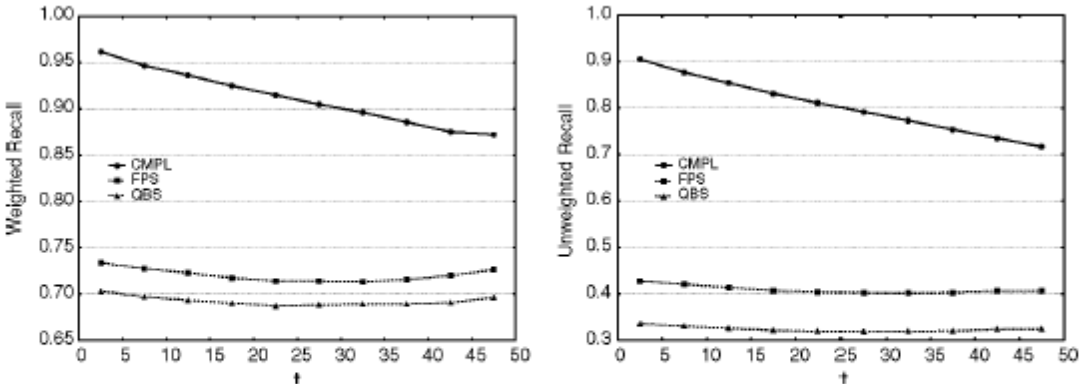
An important property of the content summary of a database is its coverage of the current database vocabulary. An up-to-date and complete content summary always has perfect recall, but an old summary might not, since it might not include, for example, words that appear only in new database documents. The *unweighted recall (ur)* of $O(D, t)$ with respect to $C(D)$ is the fraction of words in the current summary that are also present in the old summary:

$ur = |Wo \cap Wc| / |Wc|$ .

This metric gives equal weight to all words and takes values from 0 to 1, with a value of 1 meaning that the old content summary contains all the words that appear in the current content summary, and a value of 0 denoting no overlap between the summaries. An alternative recall metric, which gives higher weight to more frequent terms, is the *weighted recall (wr)* of $O(D, t)$

*To compute the approximate content summaries, we indexed and queried the data using ht://Dig (http://www.htdig.org/), an off-the-shelf indexing package.*

**Figure 1.**




*Weighted* *Unwaighted*

*The weighted and unweighted recall of content summary O(D,t) (CMPL) and of the approximate content summaries* $\hat{D}$ *(D,t) (QBS and FPS), with respect to the "current" content summary C(D), as a function of time t and averaged over each database D in the data set.*

with respect to *C*(*D*): *wr* $\frac{\sum_{wc} w_o \cap w_c \, f_c(w,D)}{\sum_{wc} w_c \, f_c(w,D)}$ We will use analogous definitions of unweighted and weighted recall for a sample-based content summary ˆ*O* (*D*, *t*) of database *D* obtained *t* weeks into the past with respect to the current content summary (*D*) for the same database.

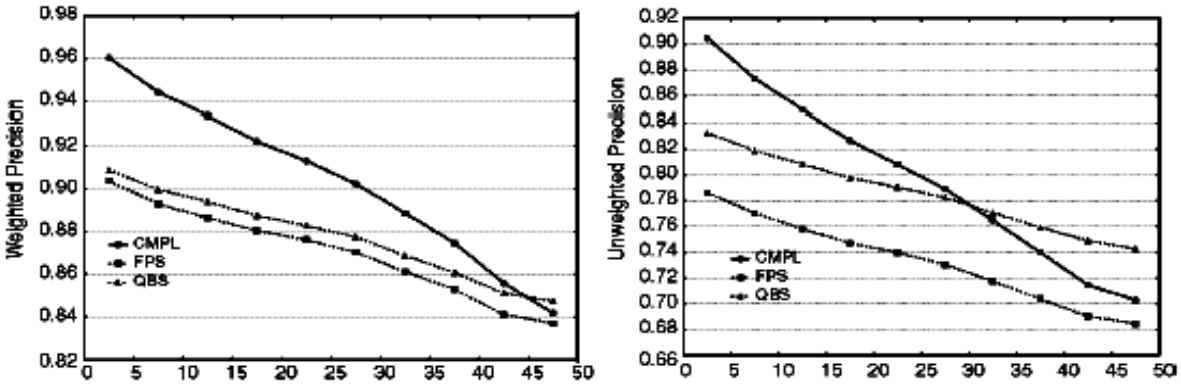
The *CMPL* lines in Figures 1(a) and 1(b) show the weighted and unweighted recall, respectively, for complete *t*-week-old summaries with respect to the "current" summary, as a function of *t* and averaged over every possible choice of "current" summary. Predictably, both the weighted and unweighted recall values decrease as *t* increases. For example, on average, 1-week-old summaries have unweighted recall of 91%, while older, 25-week-old summaries have unweighted recall of about 80%. The weighted recall figures are higher, as expected, but still significantly less than 1: this indicates that the newly introduced words have low frequencies, but constitute a substantial fraction of the database vocabulary as well.

The *QBS* and *FPS* lines in Figure 1 show the corresponding results for *QBS* and *FPS* content summaries. As expected, the values for all the approximate, sample-based summaries are substantially smaller than those for the complete summaries. Also, the recall values of the sample-based summaries do not change much over time, because the sample-based summaries are only marginally complete to start with and do not suffer a significant drop in recall over time. This shows that the inherent incompleteness of the sample-based summaries "prevails" over the incompleteness introduced by time.

Another interesting observation is that recall figures initially decrease (slightly) for approximately 20 weeks, then remain stable, and then, surprisingly, increase, so that a 50-week old content summary has higher recall than a 20-week old one, for example. This unexpected result is due to an interesting periodicity: some events (e.g.,

"Christmas," "Halloween") appear at the same time every year, allowing summaries that are close to being one year old to have higher recall than their younger counterparts. This effect is only visible in the sample-based summaries, which cover only a small fraction of the database vocabulary, and is not observed in the complete summaries, mainly because they are larger and are not substantially affected by a relatively small number of words.

**Figure 2.**




***(a) Weighted Precision*** ***(b) Unweighted Precision***

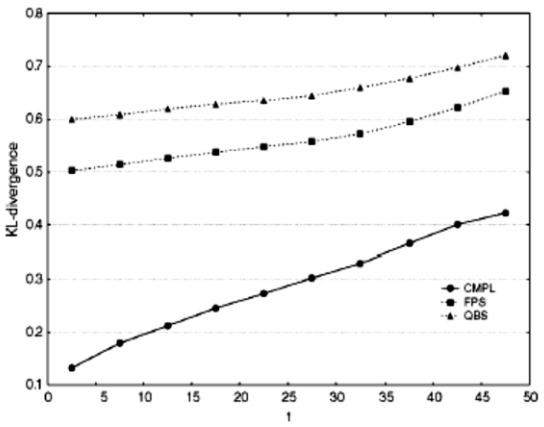
*The weighted and unweighted precision of content summary O(D,t) (CMPL) and of the approximate content summaries $\hat{D}$ (D,t) (QBS and FPS), with respect to the "current" content summary C(D), as a function of time t and averaged over each database D in the data set.*

***(b) Precision.***

Another important property of the content summary of a database is the precision of the summary vocabulary. Up-to-date content summaries contain only words that appear in the database, while older summaries might include obsolete words that appeared only in deleted documents. The *unweighted precision (up)* of $O(D, t)$ with respect to $C(D)$ is the fraction of words in the old content summary that still appear in the current summary $C(D)$: $up = \frac{|W_o \cap W_c|}{|W_o|}$. This metric, like *weighted recall*, gives equal weight to old content summary only contains words that are still in the current content summary, and a value of 0 denoting no overlap between the summaries. The alternative precision metric, which–just as in the *weighted recall* metric–gives higher weight to more frequent terms, is the *weighted precision (wp)* of $O(D, t)$ with respect to $C(D)$: $wp$ =

$\frac{\sum_{wc} w_o \cap w_c f_c(w,D)}{\sum_{wc} w_c f_c(w,D)}$ We use analogous definitions of unweighted and weighted precision for a sample-based content summary ˆ*O* (*D*, *t*) of a database *D* with respect to the correct content summary *C*(*D*). The *CMPL* lines in Figures 2(a) and 2(b) show the weighted and unweighted precision, respectively, for complete *t*-week-old summaries with respect to the "current" summary, as a function of *t* and averaged over every possible choice of "current" summary. Predictably, both the weighted and unweighted precision values decrease as *t* increases. For example, on average, a 48-week-old summary has unweighted precision of 70%, showing that 30% of the words in the old content summary do not appear in the database anymore. The *QBS* and *FPS* lines in Figure 2 show the corresponding results for *QBS* and *FPS* content summaries. As expected, precision decreases over time, and decreases much faster than recall. For example, almost 20% of the words in a 15- week-old *QBS* content summary are absent from the database. The periodicity that appeared in the recall figures is not visible for the precision results: the sample-based content summaries contain many more "obsolete" words that do not appear in the database anymore, so a small number of words that appear periodically cannot improve the results.

**Figure 3.**




***The KL divergence of the content summary O(D,t) (CMPL) and of the approximate content summaries $\hat{D}$ (D,t) (QBS and FPS), with respect to the "current" content summary C(D), as a function of time t and averaged over each database D in the data set.***

*(c) Kullback–Leibler*

*Divergence.* Precision and recall measure the accuracy and completeness of the content summaries, based *only* on the presence of words in the summaries. However, these metrics do not capture the accuracy of the frequency of each word as reported in the content summary. For this, the *Kullback–Leibler divergence* [Jelinek 1999] of *O*(*D*, *t*) with respect to *C*(*D*) (KL for short) calculates the "similarity" of the word frequencies in the old content summary *O*(*D*, *t*) against the "current" word frequencies in *C*(*D*): $KL = \sum_{w \in} W_o \cap W_c p_c(w|D).\log \frac{p_c(w|D)}{p_o(w|D)}$,

where $p_c(w|D)=\frac{f_c(w|D)}{\sum_{w'_c W_o \cap W_c} f_o(w',D)}$ *i*s the probability of observing *w* in *C*(*D*), and $p_o(w|D)=\frac{f_o(w|D)}{\sum_{w'_c W_o \cap W_c} f_o(w',D)}$ is the probability of observing *w* in *O*(*D*, *t*). The KL divergence metric takes values from 0 to infinity, with 0 indicating that the two content summaries being compared are equal. The *CMPL* line in Figure 3 shows that the KL divergence of old content summaries increases as *t* increases. This confirms the previously observed results and shows that the word frequency distribution changes substantially over time. Furthermore, the KL divergence of the old approximate summaries (lines *QBS* and *FPS*) also increases with time, indicating that approximate content summaries become obsolete just as their complete counterparts do.

### *Conclusion*

We studied how content summaries of text databases evolve over time. We observed that the quality of content summaries (both complete and sample-based) deteriorates as they become increasingly older. Therefore, it is imperative to have a policy for periodically updating the summaries to reflect the current contents of the databases. We now turn to this important issue and show how we can use "survival analysis" for this purpose.

## References


[1] BERGMAN, M. K. 2001. The deep web: Surfacing hidden value. J. Electron. Pub. 7, 1 (Aug.).

[2] BREWINGTON, B. E. AND CYBENKO, G. 2000a. How dynamic is the web? In Proceedings of the 9th International World Wide Web Conference (WWW9). 257–276.

[3] BREWINGTON, B. E. ANDCYBENKO,G. 2000b. Keeping up with the changing web. IEEE Comput. 33, 5 (May), 52–58.

[4] CALLAN, J. P. 2000. Distributed information retrieval. In Adv. Inf. Retriev. Kluwer Academic Publishers, 127–150.

[5] CALLAN, J. P. AND CONNELL, M. 2001. Query-based sampling of text databases. ACM Trans. Inf. Syst. 19, 2, 97–130.

[6] CHAKRABARTI, S. 2002. Mining the web. Morgan-Kaufmann, San Francisco, CA.

[7] CHO, J. ANDGARC´ıA-MOLINA,H. 2000. The evolution of the web and implications for an incremental crawler. In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000). 200–209.

[8] CHO, J. AND GARC´ıA-MOLINA, H. 2003. Estimating frequency of change. ACM Trans. Internet Tech. 3, 3 (Aug.), 256–290.

[9] CHO, J., GARC´ıA-MOLINA, H., AND PAGE, L. 2000. Synchronizing a database to improve freshness. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 2000). ACM, New York, 117–128.

[10] CHO, J. AND NTOULAS, A. 2002. Effective change detection using sampling. In Proceedings of the 28th International Conference on Very Large Databases (VLDB 2002). 514–525.

[11] COFFMAN, JR., E. G., LIU, Z., AND WEBER, R. R. 1998. Optimal robot scheduling for web search engines. J. Sched. 1, 1 (June), 15–29.

[12] COX, D. R. 1972. Regression models and life-tables (with discussion). J. Roy. Stat. Soc. B, 34, 187–220.

[13] DOUGLIS, F., FELDMANN, A., KRISHNAMURTHY, B., AND MOGUL, J. C. 1997. Rate of change and other metrics: A live study of the world wide web. In Proceedings of the 1st USENIX Symposium on Internet Technologies and Systems (USITS 1997). 16–31.

[14] DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. Pattern Classification, 2nd ed. Wiley, New York.

[15] EDWARDS, J., MCCURLEY, K. S., AND TOMLIN, J. A. 2001. An adaptive model for optimizing performance of an incremental web crawler. In Proceedings of the 10th International World Wide Web Conference (WWW10). 106–113.

[16] FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. 2003. A large-scale study of the evolution of web pages. In Proceedings of the 12th International World Wide Web Conference (WWW12). 669–678.

[17] GRAVANO, L., CHANG, K. C.-C., GARC´ıA-MOLINA, H., AND PAEPCKE, A. 1997. STARTS: Stanford proposal for Internet meta-searching. In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD'97). ACM, New York, 207–218.

[18] GRAVANO, L., GARC´ıA-MOLINA, H., AND TOMASIC, A. 1999. GlOSS: Text-source discovery over the Internet. ACM Trans. Data. Syst. 24, 2 (June), 229–264.

[19] GRAVANO, L., IPEIROTIS, P. G., AND SAHAMI, M. 2003. QProber: A system for automatic classification of hidden-web databases. ACM Trans. Inf. Syst. 21, 1 (Jan.), 1–41.

[20] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. H. 2001. The Elements of Statistical Learning. Springer-Verlag, New York.