Standard Information Retrieval

Diagram:
- Intranet Web Server
- File Server
- Database Server
- Content Management System
- Search Engine Spider / Document Indexer
- Search Engine Index
- Search Engine and Application
- End User

Note: May reside on many physical machines.
Distributed Information Retrieval

Various Content Repositories

Vendor A
Search System 1
Index
Engine

Vendor B
Search System 2
Index
Engine

Vendor C
Search System 3
Index
Engine

Federated Data Silo
Silo's Search Engine

Search Federator

Search User
Motivations – Deep Web

Search results: 617 documents

Result 1

Little Tragedies
BOOK - 237 Pages

Show Table of Contents

Google

Your search - site:shop.ebrary.com Pushkin - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.
interactions and normalize the partial result rankings produced by potentially diverse models of probabilistic retrieval (cf. Callan, 2000a; Callan, Crestani, & Sanderson, 2004b). Automatically learned descriptions (Callan & Connell, 2001), cooperative protocols (Gravano, Chang, Garcia-Molina, & Paepcke, 1997), fusion heuristics (Callan, Crestani, Nottelmann, Pala, & Shou, 2003), and experimental testbeds (e.g., Callan, 2000b; French et al., 1999) have been proposed and extensively tested. Finally, a few research projects have carried out preliminary investigations into nontextual forms of content-based distributed retrieval (e.g., Nottelmann & Fuhr, 2003) and begun to explore the potential of cooperative and grid-enabled retrieval infrastructures.

More recently, core techniques developed for client/server retrieval have been repurposed in the more dynamic context of peer-to-peer retrieval, where queries may be posted, routed, and directly executed by any of a number of mutually, but intermittently connected peer engines (Callan, Fuhr, & Nejdl, 2004a).

Hybrid, double-tiered architectures, in particular, have offered ideal ground for bridging optimizations developed for client/server architectures with the advantages of fully decentralized control (i.e., increased potential for resource pooling, fault tolerance, dynamic self-configuration, and privacy) (Lu & Callan, 2003, 2005; Yang & Garcia-Molina, 2001).

Rather independently and over a longer period of time, the Digital Library (DL) community also has explored the potential of distributed retrieval in the practice of its information services. Here, retrieval has been mainly interpreted as a deterministic process defined against the explicit structure of descriptive and manually authored metadata records. Nonetheless, queries and results have still been exchanged within the client/server architecture described earlier, the Z39.50 protocol (Z39.50 Maintenance Agency, 2003), in particular, has standardized the syntax and semantics of such exchange. Recently, more lightweight, Web-based protocols for distributed retrieval also have been proposed (e.g., Paepcke, Brandriff, Janee, & Larson, 2003; Sanderson, 2003; Simon, Massart, Assche, Ternier, & Duvall, 2003).

Over the past 5 years, however, the DL community has progressively favored the complementary approach of iteratively and incrementally centralizing metadata as a precondition to the retrieval of the associated data: Metadata have been "pulled" toward the queries in advance of their execution, and the retrieval function has remained centralized. Figure 2 shows the data flow in metadata harvesting. In the offline phase, a Service Provider SP periodically and incrementally gathers Metadata M from a number of Data Providers DP1, DP2, & DPN and persistently stories it in a Metadata Repository MR. In the online phase, SP interfaces users and resolves their queries against the metadata in MR.

Standardized de facto by the Protocol for Metadata Harvesting of the Open Archive Initiative (OAI-PMH; Lagoze, de Sompel, Nelson, & Warner, 2002b), the strategy mentioned earlier has become known as the harvesting model of retrieval over distributed content. The model has proved particularly suitable to meet the technical and socio-logical requirements of retrieval—and, in fact, of many other metadata-based services—within large-scale Federated Digital Libraries (FDLs), most noticeably those built around Institutional Repositories (Crow, 2002), and the Open Access movement (Bailey, 2005). Among these are the cross-sectoral, nationally scoped initiatives which account for most of current development efforts within the DL field (e.g., Anan et al., 2002; Joint Information Systems Committee, 2001; Lagoze et al., 2002a; van der Kuil and Feijen, 2004). A principled analysis of such success is found in Simeoni (2004) and is summarized in the next section.
Resource Description

- **Cooperative Collections – STARTS, SDLIP**
- **Uncooperative – Query-Based Sampling**
  - Select an initial query term
  - Run a one-term query on a resource
  - Retrieve top N documents
  - Extract terms and frequencies and add them to the RD
  - Check for stopping criteria

- **Kullback-Leibler Divergence**

\[
KL(RD_a \parallel RD_e) = \sum_{t \in V} p(t \mid RD_a) \log \frac{p(t \mid RD_a)}{p(t \mid RD_e)}
\]

\[
p(t \mid RD_a) = \frac{n(t, RD_a)}{\sum_{t \in RD_a} n(t, RD_a)}
\]

\[
p(t \mid RD_e) = \frac{\sum_{d \in RD_e} n(t, d)}{\sum_{t \in V} \sum_{d \in RD_e} n(t, d)}
\]
Resource Selection

- **Super-document**
  - Collection Retrieval Inference Network (CORI net)
  - Vector Space Model - GLOSS family
  - Based on language model
- **Retain document boundaries**
  - Relevant Document Distribution Estimation
  - Unified Utility Maximization
  - others
- **Recall-oriented metrics**

\[ \hat{R}(n) = \frac{\sum_{i=1}^{n} R_i}{\sum_{i=1}^{N} R_i} \]
Results Fusion

- **Basic**
  - Round-Robin
  - CombSUM, CombMNZ, ...
- **DIR**
  - CORI
  - Semi-Supervised Learning (SSL)
- **Retrieval accuracy - precision at rank N**
  \[ P(n) = \frac{R_n}{n} \]
Updating Resource Representation

Ipeirotis et al.

- First paper on the problem
- 152 crawlable web-sites for 52 weeks
- **Metrics**
  
  recall:  \( \text{wr} = \frac{\sum_{w \in W_o \cap W_c} f_C(w)}{\sum_{w \in W_C} f_C(w)} \)  
  
  precision:  \( \text{wp} = \frac{\sum_{w \in W_o \cap W_c} f_O(w)}{\sum_{w \in W_o} f_O(w)} \)

- **KL Divergence:**  \( \text{KL} = \sum_{w \in W_o \cap W_c} p_C(w) \times \log \frac{p_C(w)}{p_O(w)} \)

- **Survival Analysis**
- **Individual updating policies**
Results

For some databases, we did not detect a change within the each content summary for different values of threshold as well as Definition 3, we computed the survival time of Computing Survival Times:

4.3. Using Cox Regression to Model Content Summaries

Next, we describe how we use the Cox regression model before using any survival analysis technique for our problem. An interesting variation of the Cox model that overcomes the PH assumption is the stratified Cox model allows for easier generalization of the results, since the applicability of Cox’s model is that the predictor variables, and contain only partial information, indicating that censored cases. (For more information, see [10].)

The Cox model effectively exploits incomplete or “censored” data, from cases that “survived” the whole study period. The Cox model can be generalized for the members of the population. The Cox model allows the extraction of a “normalized” hazard function that is not influenced by predictor variables. This results in reweighting of the remaining observations that adapt to the chosen value of 

\[
\sum_{i=0}^{0.60} = 0.62
\]

\[
0.64
\]

\[
0.66
\]

\[
0.68
\]

\[
0.70
\]

\[
0.72
\]

\[
0.74
\]

\[
g(\tau) = \log(h_0(t)) + \sum_i \beta_i x_i(t)
\]

\[
B_0 = \tau_0 \Rightarrow \beta_0 = \beta_1
\]

\[
K, size, tau
\]

\[
Naive
\]

\[
\pm 95\% \text{ Confidence Interval}
\]

\[
O, size, tau
\]

\[
K, size, tau
\]

\[
K, tau
\]

\[
Size, tau
\]

\[
\text{KL-based change definition improves not only the KL divergence but also precision and recall.}
\]

\[
D, t
\]

\[
\text{The only requirement for}
\]

\[
\text{the applicability of Cox’s model is that the predictor variables, and contain only partial information, indicating that censored cases. (For more information, see [10].)}
\]

\[
\text{Using the study of Section 3.2 for the definition of KL divergence), where}
\]

\[
\text{content summaries are different when KL}
\]

\[
\text{that two content summaries}
\]

\[
\text{change is unlikely to be of importance for database selection. Therefore, we relax this definition and say that two}
\]

\[
\text{between updates and averaged over each data–}
\]

\[
\text{because KL depends on the whole word-frequency distribution. As our later experiments show, an update policy derived from the}
\]

\[
\text{in Section 5, we show that we can define update schedules}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{that remain the same across the different strata, but each stratum}
\]

\[
\text{in our change model. Later,}
\]

\[
\text{longer survival times and the exact value of}
\]

\[
\text{higher values of}
\]

\[
\text{should be set}
\]

\[
\text{is a}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{definition for the PH assumption is that two individual groups}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{that remain the same across the different strata, but each stratum}
\]

\[
\text{in our change model. Later,}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]

\[
\text{portionality assumption. In this case, the variables that do}
\]

\[
\text{that overcomes the PH assumption is the stratified Cox model}
\]

\[
\text{the Cox model to incorporate}
\]

\[
\text{the strata. The}
\]

\[
\text{that adapt to the chosen value of}
\]

\[
\text{are}
\]

\[
\text{to represent changes in text database content summaries.}
\]
Shokouhi et al.

- Retrieval accuracy evaluation
- Methods
  - Constant updates (CU)
  - Based on query-logs (QL)
  - Based on collection size (SS)
- Fixed updating period
- 2 news web-sites for 8 weeks
- Same updating policy for each database

\[ n_C = n \times \frac{1}{N} \]

\[ n_C = n \times \frac{\text{Rank}_{est}(C)}{\sum_{i=1}^{N} \text{Rank}_{est}(C_i)} \]

\[ n_C = n \times \frac{\text{Size}_{est}(C)}{\sum_{i=1}^{N} \text{Size}_{est}(C_i)} \]
Out-of-Date RD Evaluations

Resource Description

Precision at N

Out-of-Date RD Evaluations

Resource Description

Precision at N

Out-of-Date RD Evaluations

Resource Description

Precision at N
One interesting observation during crawls seven and eight is that the KL divergence increased sharply for all three methods. This indicates that the representation sets were changing rapidly during these periods.

In comparison, the three updating policies display a consistent estimation close to the old-baseline. In the three updating methods, the QL updating method requires some data from the query logs, which can be used to determine the number of new documents that should be added to the collection representation set. The SS updating method is based on the size of the collection, with larger collections being updated more frequently. The CU updating method is based on the current relevance of the documents in the collection.

As expected, the QL updating method performed better than both the SS and CU updating methods. QL recorded higher precision values on average than SS, although both approaches showed relatively worse performance than CU. When comparing QL to both baselines, the policy was found to significantly improve the performance of the representation sets.

Overall, the SS updating policy provided the most accurate and consistent performance. It is not clear to what extent this is due to the size of the collection or the current relevance of the documents. Further research is needed to investigate what factors contribute to the performance of the updating policies.
Drawbacks

- **Data sets**
  - No statistics: size, content
  - Not unified
  - Pseudo relevance feedback instead of judgments
  - Results are averaged across data sets
- **Documents deletion**
- **No extensive experiments**
Goals

- Choose data sets, get statistics
- Evaluate evolution
  - weighted precision / recall
  - Kullback-Leibler Divergence
- Individual collection updating policies
  - Updating period
    - Survival Analysis
    - Other methods from web research
  - Evolving policies
- Reuse documents from user-query results
- Extensive experiments