Unbiased Results Extraction Using Boolean Queries

P.SRINIVAS REDDY
M.TECH (CSE) STUDENT
AKULA SRIRAMULU COLLEGE OF ENGG
TANUKU, AP, INDIA

N.ANURAGAMAYI
ASST.PROFESSOR
AKULA SRIRAMULU COLLEGE OF ENGG
TANUKU, AP, INDIA

M.SATEESH KUMAR
HOD OF CSE
AKULA SRIRAMULU COLLEGE OF ENGG
TANUKU, AP, INDIA

Abstract

An ever-increasing amount of information on the Web today is available only through search interfaces: the users have to type in a set of keywords in a search form in order to access the pages from certain Web sites. These pages are often referred to as the *Hidden Web* or the *Deep Web*. Since there are no static links to the Hidden Web pages, search engines cannot discover and index such pages and thus do not return them in the results. However, according to recent studies, the content provided by many Hidden Web sites is often of very high quality and can be extremely valuable to many users. A naive approach would be to submit a disjunctive query with all query keywords, retrieve all the returned matching documents, and then re-rank them. Unfortunately, such an operation would be very expensive due to the large number of results returned by disjunctive queries. In this paper, we present algorithms that return the top results for a query, ranked according to an IR-style ranking function, while operating on top of a source with a Boolean query interface with no ranking capabilities (or a ranking capability of no interest to the end user). The algorithms generate a series of conjunctive queries that return only documents that are candidates for being highly ranked according to relevance metric. Our approach can also be applied to other settings where the ranking is monotonic on a set of factors

(query keywords in IR) and the source query interface is a Boolean expression of these factors. Our comprehensive experimental evaluation on the Pub Med database and a TREC data set show that we achieve order of magnitude improvement compared to the current baseline approaches

Keywords: Information retrieval, Boolean Query Interface, TREC dataset

1 Introduction

Recent studies show that a significant fraction of Web content cannot be reached by following links [7, 12]. In particular, a large part of the Web is "hidden" behind search forms and is reachable only when users type in a set of keywords, or *queries*, to the forms. These pages are often referred to as the *Hidden Web* [17] or the *Deep Web* [7], because search engines typically cannot index the pages and do not return them in their results (thus, the pages are essentially "hidden" from a typical Web user). According to many studies, the size of the Hidden Web increases rapidly as more organizations put their valuable content online through an easy-to-use Web interface [7]. In [12], Chang et al. estimate that well over 100,000 Hidden-Web sites currently exist on the Web. Moreover, the content provided by many Hidden-Web sites is often of very high quality and can be extremely valuable to many users [7]. For example, PubMed hosts many high-quality papers on medical research that were selected from careful peer-review processes, while the site of the US Patent and Trademarks Office 1 makes existing patent documents available, helping potential inventors examine "prior art." In this paper, we study how we can build a *Hidden-Web crawler*2 that can automatically download pages from the Hidden Web, so that search engines can index them. Conventional crawlers rely on the hyperlinks on the Web to discover pages, so current search engines cannot index the Hidden-Web pages (due to the lack of links). We believe that an effective Hidden-Web crawler can have a tremendous impact on how users search information on the Web:

RELATED WORK

A preliminary version of this work has been published as a short paper in

Top-k Queries

A significant amount of work has been devoted to the evaluation of top-k queries in databases. Inlays et al. provide a survey of the research on top-k queries on relational databases. This line of work typically handles the aggregation of attribute values of objects in the case where the attribute values lie in different sources or in a single source. Describe a framework for generating an approximate top-k answer, with some probabilistic guarantees. In our work, we use the same idea; the main and crucial difference is that we only have "random access" to the underlying database (i.e., through querying), and no "sorted access." Theo bald et al. assumed that at least one source provides "sorted access" to the underlying content.

3. Exploration versus Exploitation

The idea of the exploitation/exploration tradeoff (also called the "multi-armed bandit problem") is to determine a strategy of sequential execution of actions, each of which has a stochastic payoff. While executing an action we get back some (uncertain) payoff, and at the same time we get some information that allows us to decrease the uncertainty of the payoff of future actions. In our work, we are trying to maximize the payoff/exploitation of each query (which is the number of new, relevant top-k documents that the query retrieves) while minimizing the expense/exploration (number of queries sent, and documents retrieved).

3.1 Deep Web

Our work bears some similarities to the problem of searching and extracting data from the Deep Web databases. Examine the problem of estimating the number of useful documents in the database, assuming that the statistics about the frequency and the tf.idf weights of each word in the database is given. In our work, we estimate such statistics on-the-fly, as part of the explorative sampling process. Nodules et al attempt to download the

contents of a Deep Web database by issuing queries through a web form interface. The goal of Nodules et al. is to download and index the contents of databases with limited query capabilities, whereas in our case the focus is on achieving on-the-fly ranking of query results, on top of sources with no (or no useful) ranking capabilities. An alternative approach is to characterize databases by extracting a small sample of documents that is then used to describe the contents of the database. For example, it is possible to use query-based sampling to extract such a document sample, generate estimates for the distribution of each term, and then use the estimates to guide the choice of queries that should be submitted to the database. In the experimental section, we compare against this "static sampling" alternative and demonstrate the superiority of the dynamic sampling technique, which dynamically generates estimates tailored to the query at hand.

This section reviews the relevant literature by focusing on exploration and exploitation.

Learning Myopia Argument vs. Lock-in Argument As previously mentioned the literature consists of two seemingly contrasting views concerning our central question: the learning myopia and lock-in arguments. First, we turn to the learning myopia argument. In the face of radical technological change, the potential downfall of established companies has attracted a great deal of attention in research on technology management. These streams of research acknowledge that there are gains with respect to experience in a technology. Once a firm accumulates sufficient experience with one technology, it is natural for the firm to be trapped in this technology or to be blinded to alternative opportunities—this phenomenon is labeled "competency trap" by Levitt and March (1988) or "learning myopia" by Levinthal and March (1993). The literature, however, warns that too much exploitation of the existing technology may lead the firm to be locked out of opportunities in the long run. For example, March (1991, p. 73) noted: "Since long-run intelligence depends on

sustaining a reasonable level of exploration, these tendencies to increase exploitation and reduce exploration make adaptive processes potentially sell destructive." This is particularly true when an incremental gain in performance declines with the use of the existing technology. Indeed, research on learning curve has established that such diminishing performance gain over time is prevalent. With this assumption, Lee and Ryun (2002) numerically demonstrated that an undue focus on exploitation eventually leads to technological exhaustion in the market where firms compete for developing new products. The upshot is that firms exploiting the established technology to the exclusion of exploring a new technology are expected to underperform in the long run. On the other hand, the lock-in argument has focused on the difficulty of gaining a footing bya new, incompatible technology. When a given product is subject to network externalities, its value increases with the number of customers using similar ones or products compatible with it (Shapiro and Varian 1999). Studies have indicated that the presence of a dominant installed base can inhibit innovation due excess inertia in the buyers' adoption behavior (Because there are customers who appreciate compatibility, firms that achieve backward compatibility with existing products can ensure their immediate survival at least. Furthermore, when these firms grow faster and dominate the market, there maybe little room for their rivals exploring a new, incompatible technology.

3.2. Network Externalities

Technological Change

Our study also addresses a controversy in the literature on network effects. Despite the popularity of the lock-in argument, critics have argued that many new , incompatible technologies have been somehow successfully introduced (Katz and Shapiro 1994). In particular, and Margolis (1990, 1995) argued that cases of lock-in to inferior technologies are rare in the long history of technological change. These observations raise a question of how the market makes transition between incompatible technology.

In limited situations, the customer lock-in can be mitigated when "converters" are available (David and Bunn 1988). For example, Microsoft Word was introduced after WordPerfect dominated the market. But Microsoft Word was able to attract WordPerfect users because Microsoft Word was supplied with a converter that can translate a WordPerfect file into a Microsoft Word file. An economy could benefit from converters when their e costless and they can perfectly wipe out incompatibilities. Farrell and Salome (1992), however, argued that in many cases, the tasks to achieve compatibility through converters become more complex and remain costly. Furthermore, conversion often degrades performance.

Results and Discussion:

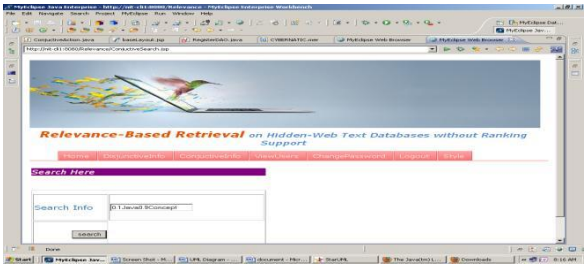

Fig 1: confidence based query extraction

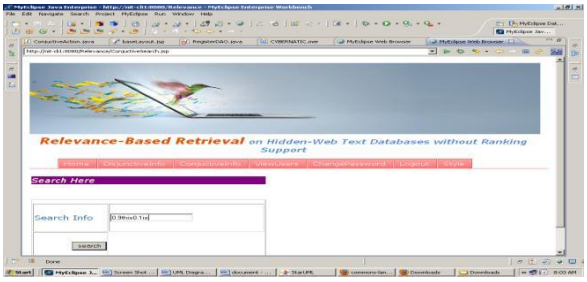

Fig 2: best edge cut results

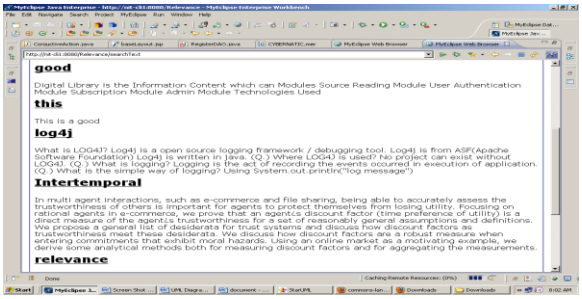

Fig3: Results display mechanism

CONCLUSIONS

We presented a framework and efficient algorithms to build a ranking wrapper on top of a documents data source that only serves Boolean keyword queries. Our algorithm submits a minimal sequence of conjunctive queries instead of a very expensive disjunctive one. Our comprehensive experimental evaluation on the Pub Med database shows that we achieve order of magnitude improvement compared to the baseline approach.

REFERENCES

[1] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A.Y. Halevy, "Google's Deep Web Crawl," Proc. VLDB, vol. 1, no. 2, pp. 1241-1252, 2008.

[2] A. Ntoulas, P. Zerfos, and J. Cho, "Downloading Textual Hidden Web Content by Keyword Queries," Proc. Fifth ACM and IEEE Joint Conf. Digital Libraries (JCDL '05), 2005.

[3] J.R. Herskovic and E.V. Bernstam, "Using Incomplete Citation Data for Medline Results Ranking," Proc. AMIA Ann. Symp., pp. 316-20, 2005.

[4] Z. Lu, W. Kim, and W.J. Wilbur, "Evaluating Relevance Ranking Strategies for Medline Retrieval," J. Am. Medical Informatics Assoc., vol. 16, no. 1, pp. 32-36, 2009.

[5] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, Inc., 1986.

[6] A. Singhal, "Modern Information Retrieval: A Brief Overview," Bull. IEEE CS Technical Committee on Data Eng., vol. 24, no. 4, pp. 35-42, http://singhal.info/ieee2001.pdf, 2001.

[7] S.E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at Trec-3," Proc. Text Retrieval Conf. (TREC), 1994.

[8] R.C. Geer et al., "Ncbi Advanced Workshop for Bioinformatics Information Specialists: Sample User Questions and Answers,"
http://www.ncbi.nlm.nih.gov/Class/NAWBIS/index.html, Aug. 2007.