Evaluating the Effectiveness of Keyword Search

William Webber
Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
wew@csse.unimelb.edu.au

Abstract

The prevalence of free text search in web search engines has inspired recent interest in keyword search on relational databases. Whereas relational queries formally specify matching tuples, keyword queries are imprecise expressions of the user's information need. The correctness of search results depends on the user's subjective assessment. As a result, the empirical evaluation of a keyword retrieval system's effectiveness is essential. In this paper, we examine the evolving practices and resources for effectiveness evaluation of keyword searches on relational databases. We compare practices with the longer-standing full-text evaluation methodologies in information retrieval. In the light of this comparison, we make some suggestions for the future development of the art in evaluating keyword search effectiveness.

1 Introduction

The rise of search engines as gateways to the Internet has made searching an everyday activity. The predominant mode of search is through the use of *keywords*, a small number of highly discriminating terms that the user anticipates will identify the web pages they are looking for. Keyword search offers a straightforward, intuitive, and flexible method of retrieving information. The success of keyword search on the web has generated interest in keyword search interfaces to relational databases and similar structured data sources. The traditional method of querying structured data stores is through formal query languages such as SQL. Such query languages, however, require much time to learn, and knowledge of a store's data schema to use. Keyword search interfaces offer a simple and flexible alternative, with (it is hoped) minimal loss of querying power.

Keyword search on unstructured text data has long been studied in the information retrieval community, where it goes under the name of free text search. Keyword searches provide only an approximate specification of the information items to be retrieved. Therefore, the correctness of the retrieval cannot be formally verified, as it can with query languages such as SQL. Instead, retrieval effectiveness is measured by user perception and experience. The empirical assessment of keyword-based retrieval systems is therefore imperative. Such assessment is the topic of this paper. We begin in Section 2 with a description of the characteristics of keyword search. Section 3 surveys the history of effectiveness evaluation on unstructured text in information retrieval. In Section 4, we examine the resources and methods used to date for keyword search evaluation over structured data. Finally, we consider in Section 5 some future directions for the evaluation of keyword-based retrieval on relational databases.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Characteristics of Keyword Search

Users perform searches to satisfy *information needs*. A keyword query is an expression of such an information need, and it is the task of the retrieval system to return information items that are *relevant* to that need. For unstructured text, the information items are discrete documents. For relational data, however, the information items are (possibly joined) tuples. The relational search system therefore has the additional responsibility of determining the candidate tuple joins. Additionally, the keyword query contains no schema information, so that each keyword potentially must be matched against each field of the joined tuple [1, 7].

A keyword query does not precisely define which information items are relevant to the user's information need. For instance, items may be relevant even though they do not contain all query terms, and conversely items containing all query terms may not be relevant. Instead, retrieval involves computing the *similarity* between the user's query and each information item. Similarity metrics generally take into account textual features such as the number of times a keyword occurs in an item, and the overall discrimination of the keyword in the collection [8]. With relational data, structural features can also be incorporated in the similarity metric. For instance, tuples that have fewer joins may be preferred as more coherent than tuples with many joins [1, 9]; or query keywords could be matched with schema terms [10]. The metric assigns a similarity score to each candidate item. The system then ranks the items by decreasing similarity, and returns some prefix of the ranking to the user. The goal of ranking by similarity is to bring more relevant items to the top, thus minimizing the amount of effort the user must expend to find the information they want.

In a structured query language like SQL, there is only one correct answer set. In contrast, there are many plausible similarity metrics, each with its own way of inferring a user's information need from a query, and of calculating the query's similarity to information items, to generate a ranking of answers. The *effectiveness* of a response to a keyword query, and hence of the similarity metric, is not something that can be formally proved; rather, it is determined by the user who realized the information need, formulated the query, and perused the response. This effectiveness must be empirically assessed.

3 Evaluation in Information Retrieval

The empirical approach to information retrieval began with the field itself, in the experiments conducted at the library of the Cranfield Aeronautical College, England, in the late 1950s and early 1960s, under the direction of the librarian, Cyril Cleverdon [4, 17]. The orthodoxy of the time in information science was that complex, hierarchical indexing schemes were essential to effective retrieval. The question the Cranfield experiments set out to answer was, which indexing scheme was best; and the answer the experiments arrived at was: none. It made no great difference which scheme was used; simply indexing documents by plain keywords was as good a method as any; what mattered was the process of retrieval. Cleverdon himself described these as "results which seem to offend against every canon on which we were trained as librarians" [4, p.252]. These findings turned attention away from information classification, and towards information retrieval.

The experimental method developed in the Cranfield tests has been highly influential. The first component of this method is the *corpus* of documents to index and retrieve. Against this corpus, user information requests must be processed; and a key insight of Cranfield was to divide these requests into two further components: first, the request statements themselves (what later became termed the *topics*); and second, assessments of which documents in the corpus were relevant to each request (called *qrels* in the jargon). These three components – corpus, topics, and qrels – together form a test collection; and the use of such a test collection in evaluation is often termed the "Cranfield methodology" or even the "Cranfield paradigm" [19].

The Cranfield experiments themselves were carried out entirely, and rather heroically, by manual means; the two hours required to process each of the 361 searches by hand was regarded as "relatively cheap compared to what would have been the cost for any form of machine searches" [4, p.91]. The test collection method is,

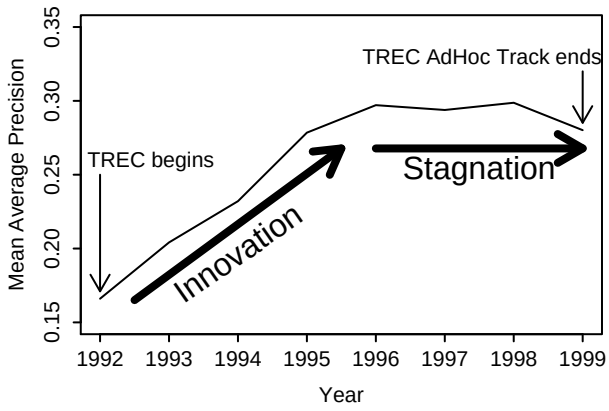

Figure 1: Retrieval effectiveness of the SMART versions from the first eight years of TREC, averaged across the first eight TREC collections [3].

though, ideally suited to automation and computerization. Relevance assessments are made in advance and are reusable, so experiments can be performed automatically and cheaply. The first such computerized retrieval systems were developed in the early 1960s, the most famous being that of the SMART project at Cornell University [14, 13]. Early progress was brisk, driven by the fast turnaround of collection-based evaluation, and the foundations of statistical information retrieval were laid down within a decade, including term weighting, query expansion, and result ranking. Over time, however, the field suffered from a lack of consolidation of results, due in part to the small and ageing test collections employed [16]. The largest collection used by SMART in 1990 had under 13,000 documents and was already 20 years old [15]. The credibility of experimental findings was undermined, impeding the adoption of research technologies in operational systems [13].

The second great impetus to empirical IR research came with the institution of the Text REtrieval Conferences (TREC) in 1992 [6, 20]. TREC produces large-scale, up-to-date test collections, encourages collaborative experiments upon them, and provides a venue for publishing and discussing results. The first collection used at TREC, known as TIPSTER, contained some 750,000 documents, a fifty-fold leap over what was available previously. The collaborative experiments run at TREC also provided a forum for the direct comparison, on equal terms, of many different research ideas. The first TREC experiment involved 22 research groups; by the fifth year, this has reached 38; and at its (apparent) peak in 2005, 117 research groups participated in TREC [18].

The impact that the TREC effort had upon the effectiveness of retrieval systems can be gauged from Figure 1. Even though SMART had been under active development for three decades, the first five years of TREC saw its retrieval effectiveness almost double, as measured by mean average precision, one common evaluation metric. And SMART's experience is typical of that of other participating groups. A number of important innovations were made during these early years of TREC, from new similarity metrics such as BM25 [12], now the standard retrieval formula, to smaller refinements that nevertheless had significant impacts, such as document length normalization. Few of these innovations were revolutionary in their nature; rather, the existence of a large and standard test environment allowed existing ideas to be extended, refined, and tuned. Figure 1 also suggests that improvements had plateaued by TREC's fifth year, leading to the retirement of the ad-hoc (plain text) task in favour of fresher tasks in 1998; and a recent survey has found little improvements in ad-hoc retrieval effectiveness during the following decade [2]. This underlines the impact that the standard, large-scale test collections produced by TREC had: in just a few years, they took retrieval technology from half of its potential, arrived at through three decades of piecemeal research, up to the effectiveness limits of current approaches.

4 Evaluation of Keyword Search

The earliest work in keyword search on relational databases was concerned with the practicality of performing it in a reasonably efficient manner. At this stage, any and only tuples that contained all of the query keywords were considered correct matches for the query. Results were ranked simply by the increasing number of joins (on the principle that the fewer joins, the more coherent the answer), and the evaluation of effectiveness was not considered [1, 7]. As technology developed, researchers became interested in not just the tractability of keyword retrieval, but the quality of the results. Proper metrics of similarity between query and answer tuple were introduced. Some of these metrics were specific to structured data, and were concerned with the conciseness or coherence of the retrieved answer [5]. Other similarity metrics were adopted from full-text information retrieval, treating either whole answer tuples or their individual fields as virtual documents [10, 8]. The most fruitful of the similarity metrics combined both a structural and a full-text component [9].

With the development of proper similarity metrics to process queries came the need to assess the effectiveness of the results. A number of different test datasets have been employed by different researchers. Two in particular have become widely used: the IMDB movie database, and the DBLP database of academic publications and citations. Such agreement is not to be found on query sets, though. Queries are generally formulated by the authors themselves, rather than taken from a query log, say, or written by independent third parties. Self-authored queries have a strong potential for bias: it is too easy to formulate queries that are favourable to your own over other algorithms. Query sets have not been re-used between experiments and experimenters, making comparison of results difficult, unless the researcher provides that comparison directly through re-implementing existing approaches as baselines – an important practice which, fortunately, is fairly common in the published keyword work. Query sets are frequently quite small, rarely more than 20 per dataset, and sometimes as few as 5. This is somewhat short of the 50 queries used in TREC collections, and even that figure of 50 is regarded by many as insufficient [21]. An exception is [10]: explicitly following TREC practice, they use a set of 50 queries; additionally, these queries are not created by the authors, but sampled from the log of a commercial search engine.

Just as queries are generally authored by the researchers, so too relevance assessment is generally performed by them or their colleagues. Like self-authoring, self-assessment has the potential for biasing results, with the same interpretation of relevance determining both algorithm design and relevance assessment – and query construction too, if self-authored. What should be a (correctly) subjective assessment of relevance can easily verge on an (incorrectly) objective verification of correctness. The suspicion that this has occurred is strongest where abnormally high effectiveness scores are achieved. For instance, the best fully automatic TREC participant systems achieve scores under the precision at depth 100 metric of around metric of around 0.25; but [9] report around 0.9. Similarly, top-end scores at TREC for mean reciprocal rank (MRR, an admittedly unstable metric) are around 0.8; but [11] achieve the rather astonishing, perfect MRR score of 1.

Given the disparity of query sets, corpora, and assessment methods, it is not straightforward to determine how retrieval effectiveness in keyword search has progressed. The only method is to follow the chain of baselines. So, [10] employ a variety of tuning mechanisms to improve the vanilla IR baseline of [8] from a MRR of 0.245 to 0.871. In turn, [11] give MRR scores of 0.243 and 0.333 respectively for baselines derived from the two aforementioned papers, and then report their impressive perfect score of 1. [22] count instead the proportion of top-ranked documents that are relevant (P@1); they report a mean P@1 score for their [8] baseline of 0, and for [11] baseline of 0.2; their own systems achieves a mean P@1 of 0.95. It is rather difficult to know what to make of this sequence of results, with each researcher tripling their predecessor's baseline to achieving perfect or near-perfect scores. There is persuasive *prima facie* evidence that substantial improvements in effectiveness are being made from a reasonable, standard IR beginning; [10] tantalizing report better effectiveness than Google (and their effectiveness has been tripled twice since!). But the reliable verification of such improvements would require a larger, more thorough, and more impartial, experimental environment.

5 Future Directions In Keyword Search Evaluation

The field of keyword search on structured data is well poised for growth towards maturity. The fundamental technical and formal problems of performing such search have been solved, and many important theoretical results have been achieved (in, for instance, graph theory). Concern is now turning to questions of the end-user effectiveness of such search systems. Traditional IR similarity metrics have been ported to the new domain, and combined with domain-specific structural features. There is also evidence of significant improvements in effectiveness, both through developing new methods and tuning existing ones.

Keyword search on structured data is therefore at roughly the same stage that information retrieval was pre-TREC; or, to be less sanguine, the stage that information retrieval had reached by the mid-seventies, and was not to clearly surpass for another two decades. There is much promise in the field, but more needs to be done to set it on a firm basis, to validate its results, and to inspire the confidence needed to convert this research technology into deployed tools. And most of these desiderata depend upon improvements in evaluation method.

What is needed is a standard method, combined with large-scale, independently curated test collections. Some straightforward quantitative points are immediately apparent – for instance, test sets of only 20 queries are scarcely adequate. Test collections also need to be more reusable; not merely document corpora should be made available, but the query sets and relevance judgments to go with them, even if the last of these are incomplete. And rather than unadorned keyword queries, test collections should have properly formulated topics; that is, fuller statements of information need. Such fuller topics form a number of useful functions. They guide additional relevance assessment, if such assessment should prove necessary. Moreover, they allow for different query formulations to be made for the one information need, with each formulation assessable by the same set of relevance judgments. The ability to reformulate queries is particularly important for a new and fluid field, as new retrieval methods may require different query methods to draw out their features.

The field of keyword search may, however, still be too young, and the technology too fluid, for a full TREC-style collaborative experiment to be achievable or even appropriate. Instead, the way forward would seem to be for individual research groups to create more thorough, credibly independent, and re-usable test collections, incorporating all three components – corpus, topics, and qrels. Such an undertaking requires a non-trivial amount of effort. But the experience of TREC demonstrates how much leverage standard collections and a standard methodology can achieve.

References

[1] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: a system for keyword-based search over relational databases. In *Proc. 18th International Conference on Data Engineering*, pages 5–16, San Jose, California, Feb. 2002.

[2] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proc. 18th ACM International Conference on Information and Knowledge Management*, pages 601–610, Hong Kong, China, Nov. 2009.

[3] C. Buckley and J. Walz. SMART in TREC 8. In E. Voorhees and D. Harman, editors, *Proc. 8th Text REtrieval Conference*, pages 577–582, Gaithersburg, Maryland, US, Nov. 1999. NIST Special Publication 500-246.

[4] C. Cleverdon, J. Mills, and E. Keen. *Factors determining the performance of indexing systems*. Aslib Cranfield Research Project, Cranfield, 1966.

[5] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: ranked keyword search over XML documents. In *Proc. 29th ACM SIGMOD International Conference on Management of Data*, pages 16–27, San Diego, California, 2003.

[6] D. Harman. Overview of the first text REtrieval conference (TREC-1). In D. Harman, editor, *Proc. 1st Text REtrieval Conference*, pages 1–30, Gaithersburg, Maryland, US, Nov. 1992. NIST Special Publication 500-207.

[7] V. Hristidis and Y. Papakonstantinou. DISCOVER: keyword search in relational databases. In *Proc. 28th International Conference on Very Large Data Bases*, pages 670–681, Hong Kong, China, Aug. 2002.

[8] V. Hristidis, L. Gravano, and Y. Papakonstantinou. Efficient IR-style keyword search over relational databases. In *Proc. 29th International Conference on Very Large Data Bases*, pages 850–861, Berlin, Germany, 2003.

[9] G. Li, B. Ooi, J. Feng, J. Wang, and L. Zhou. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proc. 34th ACM SIGMOD International Conference on Management of Data*, pages 903–914, Vancouver, Canada, June 2008.

[10] F. Liu, C. Yu, W. Meng, and A. Chowdhury. Effective keyword search in relational databases. In *Proc. 32nd ACM SIGMOD International Conference on Management of Data*, pages 563–574, Chicago, IL, USA, 2006.

[11] Y. Luo, X. Lin, W. Wang, and X. Zhou. SPARK: top-k keyword query in relational databases. In *Proc. 33rd ACM SIGMOD International Conference on Management of Data*, pages 115–126, Beijing, China, 2007.

[12] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. Harman, editor, *Proc. 3rd Text REtrieval Conference*, pages 109–126, Gaithersburg, Maryland, US, Nov. 1994. NIST Special Publication 500-225.

[13] G. Salton. The Smart environment for retrieval system evaluation–advantages and problem areas. In K. Sparck Jones, editor, *Information Retrieval Experiment*, chapter 15, pages 316–329. Butterworths, 1981.

[14] G. Salton, editor. *Information Storage and Retrieval*, volume Scientific Report No. ISR-11. Cornell University, Ithaca, New York, 1966.

[15] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 44(4):288–297, 1990.

[16] K. Sparck Jones. Retrieval system tests 1958–1978. In K. Sparck Jones, editor, *Information Retrieval Experiment*, chapter 12, pages 213–255. Butterworths, 1981a.

[17] K. Sparck Jones. The Cranfield tests. In K. Sparck Jones, editor, *Information Retrieval Experiment*, chapter 13, pages 256–284. Butterworths, 1981b.

[18] E. Voorhees. Overview of TREC 2007. In E. Voorhees and L. Buckland, editors, *Proc. 16th Text REtrieval Conference*, pages 1:1–16, Gaithersburg, Maryland, US, Nov. 2007. NIST Special Publication 500-274.

[19] E. Voorhees. The philosophy of information retrieval evaluation. In *Proc.2ndWorkshop of the Cross-Lingual Evaluation Forum*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370, Darmstadt, Germany, Sept. 2002.

[20] E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

[21] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proc. 17th ACM International Conference on Information and Knowledge Management*, pages 571–580, Napa, USA, Oct. 2008.

[22] Y. Xu, Y. Ishikawa, and J. Guan. Effective top-k keyword search in relational databases considering query semantics. In *Proc. APWeb-WAIM 2009 International Workshops*, volume 5731/2009 of *LNCS*, pages 172–184, Suzhou, China, Apr. 2009.