Summary of the SIGIR 2003 Workshop on Defining Evaluation Methodologies for Terabyte-Scale Test Collections

Ian Soboroff and Ellen Voorhees
National Institute of Standards and Technology
Gaithersburg, MD
ian.soboroff@nist.gov

Nick Craswell
CSIRO
Canberra, ACT, Australia
Nick.Craswell@csiro.au

1 Introduction

Early retrieval test collections were small, allowing relevance judgments to be based on an exhaustive examination of the documents, but limiting the general applicability of the findings. Karen Sparck Jones and Keith van Rijsbergen proposed a way of building significantly larger test collections by using pooling, a procedure adopted and subsequently validated by TREC. Now TREC-sized collections (several gigabytes of text and a few million documents) are small for some realistic tasks, but current pooling practices do not scale to substantially larger document sets.

This article summarizes a workshop held at SIGIR 2003 in Toronto, Canada, the goal of which was to develop an evaluation methodology for terabyte-scale document collections. The outcome of the workshop was a proposal for a new TREC track to investigate ad hoc retrieval on a collection of 100M web pages.

In particular, we began by assuming the existence of a collection of several hundred million web pages, and discussed methods for reliably evaluating retrieval tasks on such a collection. Search tasks currently evaluated using large web collections, such as known-item and high-precision searching, focus on the needs of the common web searcher but also arise from our inability to measure recall on very large collections. Good estimates of the total set of relevant documents are critical to the reliability and reusability of test collections as we now use them, but it would take hundreds of different systems, hundreds of relevance assessors, and years of effort to produce a terabyte-sized collection with completeness of judgments comparable to a typical TREC collection. Hence, new evaluation methodologies and ways of building test collections are needed to scale retrieval experiments to the next level.

2 Attendees

The following people attended the workshop: Ian Soboroff (NIST), Ellen Voorhees (NIST), Nick Craswell (CSIRO), Chris Buckley (Sabir Research), Lee Giles (Pennsylvania State University), Judy Johnson (NEC Laboratories), Kostas Tsioutsiouliklis (NEC Laboratories), Eric Jensen (Illinois Institute of Technology), David Carmel (IBM Research), Taher Haveliwala (Stanford University), Josh Coates (Internet Archive), Mark Sanderson (University of Sheffield), Justin Zobel (RMIT University), and Charlie Clarke (University of Waterloo)

Additionally, Padmini Srinivasan (University of Iowa) and William Pottenger (Lehigh University) had planned but were unable to attend.

3 Introductory Talks

Ellen Voorhees (NIST) began the workshop with a talk giving a brief history of test collections from Cranfield to TREC. The Cranfield II experiments laid the foundations of IR evaluation methodology which have largely been maintained through the years. Cranfield gave us many assumptions, some of which remain (such as a user's information need being modeled by relevance judgments) and some which have not (such as the need for complete relevance information). The TREC collections were the first to push retrieval experiments into the gigabyte range. They accomplished this using the pooling method proposed by Karen Sparck Jones and Keith van Rijsbergen. This brought us to the central questions of the workshop: can our existing methods scale to retrieval collections of a terabyte, or more? If not, what will have to change?

Nick Craswell (CSIRO) wrapped up the introduction with some comments on infrastructure issues. Previous large TREC collections, namely the web collections (VLC2, WT2g, WT10g, .GOV) were extracted whole or in part from web crawls done by the Internet Archive or by the University of Waterloo and Virginia Tech. The question of who would do a new crawl, and of what, would prove to occupy much of the morning discussion. Given the 15kb average page size of .GOV, a 1TB web crawl would comprise about 66 million pages. A 100 million page crawl would be about 1.5TB. Nick's examination of compression methods for web crawls has shown that by clever arrangement of pages, gzip can compress 500GB or more of web pages onto a 100 GB disk. Different audiences might want different forms of the collection more amenable to their research: while many would want the whole collection (likely shipped on one or more hard disks), others might rather have a subset of the text, or preprocessed term vectors, or a link graph. Interactive experiments also require images in order to render pages correctly, and this means multiple terabytes of data for those groups perhaps least interested in system engineering issues.

After the introduction, many participants took the opportunity to show a few slides on their thinking on these issues. Briefly summarizing,

- Justin Zobel proposed ad hoc retrieval experiments, evaluated using "precise queries", those queries which can be most completely judged. His proposal involves determining these topics from a larger topic set given to participants.

- Mark Sanderson discussed their investigation into the technique of Iterative Searching and Judging (ISJ) using a 94-million page, 1.2TB crawl done by Waterloo and Virginia Tech.

- David Carmel proposed the use of "T-rels", a term-vector-based approach to labeling relevant documents similar to the patterns developed in the TREC QA track.

- Eric Jensen mentioned the work at IIT on automatic evaluation of web search engines. He proposed a combination of manual and automatic approaches.

- Taher Haveliwala discussed the crawls at the Stanford Webbase project. Their crawls are about 100 million pages. Their group has worked on a number of web issues such as mirrors, URL canonicalization, frames, crawl coverage, and search UIs. Taher pointed out that rankings themselves can be compared without relevance judgments.

- Kostas Tsioutsiouliklis talked about NEC's experiences doing large crawls of the "adult" web. This web space has many unique characteristics. A product of this work is an impressive crawl architecture.

- Lee Giles talked about CiteSeer, which has nearly completely moved to PSU. Their search platform has also been applied in a number of other domains such as finding philanthropic organizations.

- Chris Buckley proposed that pooling should continue to work in larger scales. Incompleteness of judgments is fine, as long as the pooling and judging process is not biased. He suggested that effective methods for dealing with even less complete relevance judgments are a research goal for the track. He proposed a new measure which examines the relative rank of relevant and irrelevant documents.

The discussion preceding and following lunch diverged quite a bit from the original workshop plan. Rather than present that discussion as it happened, we summarize it below along the lines of collections and tasks.

4 Document Collections

The question of what kind of document collection could be used, and how it might be obtained, dominated the morning discussion. While there was an initial presupposition that we would use a large Web crawl, there are a lot of other terabyte-sized data sources in existence which would allow us to do different kinds of experiments. For example, Lee Giles was interested in the issues surrounding dynamic collections and the changing web, something which is hard to study within a single static crawl. Several people were also interested in structured and semistructured data, although others pointed out that this kind of data is hard to come by freely in large quantities. USENET postings, mailing lists, and web logs ("blogs") were also mentioned, some of which are already being crawled and/or archived by various groups.

On the subject of web crawls, Charlie Clarke indicated that they terminated the .GOV crawl at the desired size for that collection, leaving a very large queue of pages not yet collected. David Carmel said that, based on probing with Google, he thought there might be as many as 130 million pages in the entire ".gov" domain.

The idea of a terabyte .GOV collection was attractive for several reasons. Because U.S. government publications are not subject to copyright, it has generally been felt that distributing a crawl of government web sites is legally simpler. A web collection is amenable to ad hoc search tasks as well as web-specific tasks such as have been investigated in the Web Track; in contrast, mailing lists and the like imply different sorts of search tasks. A .GOV collection might be extended to include .GOV.AU, .GOV.UK, .GOV.NZ, etc. Finally, the estimates of those present at the workshop indicated that 100M pages might capture almost all of the current .GOV.

Several participants offered to help with the crawl, and in the end there will likely be a joint effort. Judy Johnson of NEC and Josh Coates of the Internet Archive both offered to do the actual crawl. Charlie Clarke is familiar with what was needed last year in crawling .GOV for a test collection. Finally, Nick Craswell in past years has done the page bundles and TREC formatting; additionally, he has a number of ideas for compressing crawls efficiently for distribution.

The participants spent some time discussing the issue of page duplicates and aliasing. Dealing with duplicates is a thorny problem, and is related to the question of how much the data should be cleaned before being released to participants. The current plan is to distribute a list of exact duplicates (using checksumming), and to not allow participants to return exact duplicate documents in a single run. The disadvantage of this plan is that duplicate documents subtract from the number of unique documents we can include in the collection. One or more lists of nearly-exact duplicates will be produced which would allow us to study the impact of duplicates on retrieval.

5 Tasks

The canonical TREC task, ad hoc search, was last done in TREC-9 using a web collection, WT10g. There was a lot of interest in the workshop of looking again at ad hoc search, for several reasons. The ad hoc search task is the best known and most reliable task for building reusable test collections. If we are expecting that fundamental retrieval algorithms will need to change at larger scales, we should test that hypothesis with the same search task. Lastly, since coping with incomplete relevance information is the real challenge, we should focus on tasks where recall is important.

Although Justin Zobel proposed starting with a large topic pool and selecting 50 evaluation topics for which we expected to find nearly-complete relevance information, it was decided that we should only have 50 topics total. This would force us to work with the incompleteness problem, and also encourage manual runs which would be impossible on a larger topic set.

The Web Track has focused on more Web-centric tasks, namely known-item search (navigational search) and topic distillation (informational search for homepages of key relevant sites). The Web Track could continue these investigations in the new terabyte collection alongside an Ad Hoc Track. Nick Craswell said that CSIRO might produce a known-item topic set themselves for the new collection and make it available.

A number of participants were interested in distributed retrieval, and also in crawling tasks. We decided not to make these the focus of the track, although participants might choose to make use of such techniques to reduce the problem size. Other suggested but tabled tasks included QA, time-based querying, spam detection, and pornography detection.

6 Conclusion

The workshop was successful, but in a rather different way than the organizers had planned. We had expected that choosing the right task and measure, when confronted by woefully incomplete relevance information, would dominate the discussion. In the end, the overwhelming sentiment was to carry on with a known task run in a straightforward fashion on the larger collection, and attempt along the way to measure the impact of incompleteness.

It was observed that at SIGIR, experiments are routinely presented which use the “classic” newswire collections from TREC 5-8. We hope that the efforts of this workshop, and the anticipated TREC track which we plan to propose this fall, will eventually produce a test collection equally useful and reliable.