

Building a Test Collection for Investigating Contextual Information Retrieval

Diane Kelly
University of North Carolina
100 Manning Hall
Chapel Hill, NC USA 27599-3360
+00 919.962.8071
kelly@ils.unc.edu

The importance of building a test collection which can be used to explore contextual information retrieval (CIR) is great. Test collections such as those generated for TREC have proven invaluable to both TREC participants and others in the IR community. The creation of sharable test collections facilitates discovery and allows for more rapid progress since building a good test collection is such a difficult, laborious, and time-consuming task. Standard test collections also allow for multiple modes of inquiry including those that involve the comparison of various IR techniques, examination of alternative hypotheses and replication of previous findings.

The design and construction of a test collection for CIR introduces numerous challenges that are not present during the construction of more traditional IR test collections. For instance, a collection for CIR should contain more than just documents, topics and relevance judgments. Such a collection should contain information about users, their information needs and goals, their information-seeking context and their behaviors within this context. Collecting this type of information necessarily implies the construction new data collection tools since this type of information is neither obvious nor explicit in the interaction. Furthermore, these tools should produce valid, reliable and usable data. Constructing a collection whose data does not meet these fundamental criteria does little to further work in CIR. For such a collection to have the greatest ecological validity, and thus generalizability, it should be constructed within a natural use environment with real tasks and topics, rather than collected in a laboratory setting, with artificial tasks and topics. Moreover, such a collection should be collected over an extended period of time to allow for the investigation and modeling of more complex types of interactions such as successive searching and long- and short-term information needs. Finally, because the notion of context is so complex, building a test collection for CIR necessarily implies that some discussion about what context is, what elements of context matter in IR, and how these elements can be measured explicitly has to occur.

In this presentation, I will discuss the results of a naturalistic, longitudinal study that was designed to collect information about

users' information-seeking activities, context and behaviors in a natural setting over an extended period of time. One of the outcomes of this study was the development of a method for collecting data about users' information-seeking behaviors in natural search environments, with user-defined tasks and topics. Another outcome of this study was the development and evaluation of techniques for measuring aspects of information-seeking context. In this study, the entirety of users' on- and off-line interactions with their computers were unobtrusively monitored and recorded using both client- and proxy-side logging software. This included applications used, URLs visited, start, finished and elapsed times for all interactions, operating system commands, and all keystrokes such as queries and word processing text. In addition, a copy of every URL requested by each subject was saved on a local server. Throughout the study, subjects identified the various tasks and topics about which they were seeking information, and classified the documents that they viewed according to these tasks and topics. At weekly intervals, subjects updated each context measure and judged the usefulness of the documents that they viewed during the previous week. At the close of the study, subjects provided qualitative feedback about the study method including the various instruments and procedures that were used to measure context.

Although this study was not necessarily concerned with building a test collection for CIR, the resulting data makes for a potentially useful collection for exploring CIR. Furthermore, the study method can be viewed as a pilot for a larger data collection effort. I will discuss the method used to collect this data, including my attempt at measuring context, the results of the data collection effort, and the lessons learned from this effort. My purposes in sharing these results are to provide the audience with an overview of (1) how much effort is involved in planning and assembling such a collection; (2) how much data can be potentially collected; and (3) what are some potentially fruitful measures of information-seeking context. I will conclude by suggesting future directions for the construction of a new test collection that can be used to investigate CIR, and issues which need to be addressed before such an effort can commence.