Topic-Specific Off-line Web Search

Rodger Benham
RMIT University
Melbourne, Australia

ABSTRACT

Web search-engine users around the world rely upon information shared via the Internet to deliver value to society. Internet search-engines exist that allow users to issue queries and retrieve a set of online uniform resource locators (URLs) likely to satisfy an information need for a user. If the hosting, search-engine and Internet infrastructures are operating correctly, the user will receive the data that is likely to satisfy their information need. Conversely, the user will not receive the data they need and fail to perform the information seeking activity. This paper argues for the investigation of a retrieval architecture that enables search over a local cache of web pages pertinent to a set of topics, in the event of temporary or long-term Internet disconnection.

IR techniques have been applied offline for the purposes of file system search [5]. Büttcher and Clarke [2] show that searching documents in an offline context can present indexing and query efficiency challenges, as users frequently delete files in the index. In related work to storing information generated online for offline usage, Le et al. [3] survey digital preservation techniques employed by libraries to cache and store important data, while maintaining the utility of obsolete digital resources. NASA collated media to share cultural aspects of humanity to extraterrestrial life as a contingency for human extinction on a gold-plated copper disk [1]. However, the utility of the web caches we discuss in this paper is likely to be maximal closest to the most recent update of the local cache, as this will yield the most temporally relevant answer set to user queries.

As crawls of the Internet are intractably large to store on consumer computing devices, finding reduced partitions of documents relevant to the interests of the user might be a useful surrogate for Internet access. Space constraints are highly variable for consumers, therefore, retrieving and storing the cached set of documents for offline retrieval purposes is size-bound in bytes, rather than the well-studied top-k document retrieval paradigm. We name this novel retrieval paradigm top-B retrieval, where B is the maximum size in bytes of all documents retrieved for a given query. Figure 1 describes an architecture using top-B retrieval to enable topic-specific offline web search. Initially, the user supplies a daemon process with a set of topics they are interested in having an offline cache for, and the maximum size in bytes of the offline cache to be formed. An online top-B search-engine is periodically queried with respect to the set of user interests and returns an answer set. The daemon walks the answer set, downloading and updating the local cache and index. Upon Internet disconnection, the user can issue top-k queries on the local index as a surrogate for Internet access.

Selective search has been shown to be a fast retrieval approach [4]. Returning topical shards may seem like a natural fit for selecting documents to exist in an offline cluster, but unfortunately, the typical number of topical shards found by k-means clustering ($k \approx 10^2$) will provide an answer set too large for consumer hardware to cache. Implicit feedback of offline search user satisfaction might be used to form smaller, more precise topical shards for search-engines; further improving selective search efficiency.

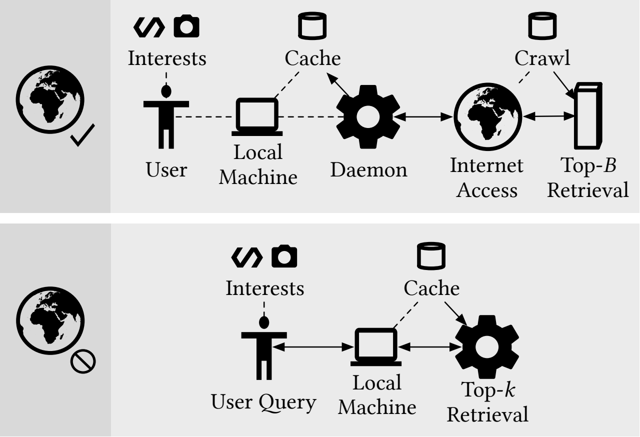

Figure 1: A possible solution architecture. Directed solid lines indicate the primary flow of data for each of the cases of Internet connectivity and disconnection. Dashed lines represent an intrinsic relationship between entities as a visual aid.

Varying B in a top-B retrieval is likely to alter the user satisfaction for the lengths of documents returned. For smaller B, textbooks may be the most effective way of ensuring coverage of a topic. For larger B, smaller documents might be easier for a user to navigate different facets of their information need. TREC collections may be unhelpful in evaluating ΔB due to pooling bias towards shorter documents, and so a collection with judgments for documents of diverse lengths may be needed to observe ΔB and its effect on user satisfaction in the proposed offline search scenario.

Automatically forming offline web caches for a prescribed set of topics on consumer devices may increase the productivity of users without Internet access. This poses new research questions in IR efficiency, effectiveness and evaluation.

Acknowledgements. Thanks to J. S. Culpepper, A. Moffat and J. Mackenzie and anonymous reviewers for their feedback.

REFERENCES

[1] The golden record. URL https://voyager.jpl.nasa.gov/golden-record.
[2] S. Büttcher and C. L. A. Clarke. Indexing time vs. query time: Trade-offs in dynamic information retrieval systems. In *Proc. CIKM*, pages 317–318, 2005.
[3] K.-H. Le, O. Slattery, R. Lu, X. Tang, and V. McCrary. The state of the art and practice in digital preservation. *J. Res. Natl. Inst. Stand. Technol.*, 107(1):93, 2002.
[4] H. R. Mohammad, K. Xu, J. Callan, and J. S. Culpepper. Dynamic shard cutoff prediction for selective search. In *Proc. SIGIR*, pages 85–94, 2018.
[5] K. Peltonen. Adding full text indexing to the operating system. In *Proc. ICDE*, pages 386–390, 1997.

DESIRES 2018, August 28-31, 2018, Bertinoro, Italy. Copyright held by the author(s).