Digital Library and Archiving for Qatar

Tarek Kanan
Virginia Tech, Department of Computer Science
Blacksburg, VA 24061 USA
tarekk@vt.edu

Sagnik Ray Choudhury
Penn State, College of Information Sciences and Technology
University Park, PA, USA
szr163@ist.psu.edu

C. Lee Giles
Penn State, College of Information Sciences and Technology
University Park, PA, USA
szr163@ist.psu.edu

Prashant Chandrasekar
Virginia Tech, Department of Computer Science
Blacksburg, VA 24061 USA
peecee@vt.edu

Edward A. Fox
Virginia Tech, Department of Computer Science
Blacksburg, VA 24061 USA
fox@vt.edu

ABSTRACT

Crawling and Indexing Qatari Scholarly Content- SeerQ

SeerSuite is a collection management system for digital libraries, developed at Penn State. It includes: 1) A Web crawler for scholarly articles; 2) A machine learning based automated system for metadata (title, abstract, author name/affiliation, citations) extraction; 3) A module for ingesting extracted information into a database and Solr; and 4) A JSP based front end for users. SeerQ reflects our modification of SeerSuite to address Qatari requirements. It uses both Heritrix and an in-house developed OAI-PMH based crawler, which accesses digital repositories in Qatar that expose their metadata and content, especially QScience, a publisher in Doha focusing on scholarly content produced in Qatar. Other seeds for crawling were provided by the Qatar National Library and cover websites such as QCRI, Qatar University, and varied research establishments. We have around 3300 documents ingested and around 4000 documents crawled. Metadata records with an author name, title, and citations are available through OAI-PMH.

Crawling and Searching- Lucidworks Fusion

Fusion is software by LucidWorks. It has the ability to collect and index documents using Apache Solr. It also provides utilities like pipelines, connectors, and logging routines. We devised our own interface to suit the needs in Qatar. We used this software to build many Qatari collections about online news resources, government activities, sports, etc. Fusion has the ability to handle multi-lingual content; our collections are in Arabic and English. We fed Fusion the seed URLs to the resources than it collected and indexed. Our Qatari Arabic news article collection has extra information, e.g., news article summaries we generated using machine learning based methods. Accordingly, we modified the Fusion configuration so there are extra fields in the schema file, and the interface to show both summary and article. Earlier we crawled 5200 PDF news files using Heritrix, each file having multiple news articles. We parsed the files, extracted the Arabic text, and ended up with roughly 120,000 news articles, for which we automatically generated summaries.

Crawling and Archiving in Qatar National Library (QNL)

QNL is crawling and archiving Web content related to Qatar. The challenges of this task include:

1. Seeds: Getting proper seeds for such a large-scale crawl is very hard, but ICT Qatar (the governing body for the Internet in Qatar) has agreed to provide a list of .qa domains.
2. Dynamic content: As the host of the upcoming 2022 FIFA World Cup, Qatar has been much in the news, with rapidly changing dynamic content. It is difficult for curators to apply tools like Heritrix, especially to adjust to large volumes of changes as well as sitemaps and RSS feeds.
3. Identifying "Qatar related" content: Many websites where Qatar is discussed (soccer or Middle East forums) contain text unrelated to Qatar. It is a challenge to quickly identify all and only the relevant URLs.

ACKNOWLEDGMENTS

We acknowledge QNRF for their support. This work was made possible by NPRP grant # 4-029-1-007 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] "ELISQ" *Electronic Library Institute SeerQ*. [Online]. Available: http:// http://elisq.qu.edu.qa/ . [Accessed: April-2015].

[2] "Heritrix" *Internet Archive*. [Online]. Available: https:// http://crawler.archive.org/index.html . [Accessed: April-2015].

[3] "SeerQ" *Seersuite*. [Online]. Available: http:// http://citeseerx.sourceforge.net/ . [Accessed: April-2015].

[4] "QNL" *Qatar National Library*. [Online]. Available: https:// http://www.qnl.qa/ . [Accessed: April-2015].

[5] "Fusion" *Lucidworks*. [Online]. Available: https:// http://lucidworks.com/fusion/ . [Accessed: April-2015].