A Data-Parallel Toolkit for Information Retrieval

Dennis Fetterly
Microsoft Research Silicon Valley
Mountain View, CA USA
fetterly@microsoft.com

Frank McSherry
Microsoft Research Silicon Valley
Mountain View, CA USA
mcsherry@microsoft.com

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Experimentation

1. EXTENDED ABSTRACT

Due to the explosive growth of the web that has occurred throughout its history, many researchers working on web corpora have begun to move toward distributed, data parallel computing. The size of the ClueWeb09 [2] corpus, at approximately one billion documents, is an indication of this. Even limiting the collection to only documents in the English language only halves the size of the collection.

In this work, we describe the collection of information retrieval algorithms we have implemented using DryadLINQ [8]. DryadLINQ is a data parallel processing system that allows programmers to write distributed programs without worrying about the implementation of a distributed system. DryadLINQ executes programs containing SQL-like Language Integrated Query statements (LINQ) by shipping the computation to nodes in the cluster for parallel execution. The ability to break a computation into many pieces that can be processed on individual machines means that even a small number of computers can be leveraged to reduce the time necessary to process large collections.

When researchers first obtain a collection of web documents, there is a substantial amount of preprocessing before analysis can commence. The toolkit assists with parsing, link extraction, associating discovered anchor text with the referenced document. Once the document content and links are in a standard format, then further processing can be performed. The toolkit provides implementations of text-based retrieval methods (BM25 [7] and BM25F [9]), query-independent link based scoring functions (PageRank, in-degree, and trans-domain indegree), query-dependent link-based scoring functions (SALSA-SETR [6]). Additionally, the toolkit provides an implementation of shingle based duplicate document detection [1], n-gram extraction, and a mechanism to build an inverted index.

The algorithms included in this toolkit include both traditional algorithms as well as recent research results. Elements of this toolkit formed the basis of the Microsoft Research entry in the TREC 2009 conference [3]. Given the implementation in a declarative, high-level language, these algorithms are easy to modify and extend making them a good basis for research into new algorithms.

In addition to discussing the use and implementation of this toolkit during the demonstration, we intend to release it [5] in source and binary form to others in the community to aid in large-scale information retrieval research. This, coupled with the public availability of the ClueWeb [2] dataset and the Dryad/DryadLINQ system [4] makes large-scale web information retrieval research substantially more accessible.

2. ACKNOWLEDGMENTS

We are very grateful to Nick Craswell, Marc Najork, and Emine Yilmaz for their assistance in writing the DryadLINQ implementations of the algorithms in this toolkit.

3. REFERENCES

[1] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic Clustering of the Web. In *Proc. of WWW6*, 1997.

[2] `http://boston.lti.cs.cmu.edu/Data/clueweb09/`

[3] N. Craswell, D. Fetterly, M. Najork, S. Robertson and E. Yilmaz. Microsoft Research at TREC 2009: Web and Relevance Feedback Tracks. In *Proc. of the 18th Text Retrieval Conference*, 2009.

[4] `http://research.microsoft.com/collaboration/tools/dryad.aspx`

[5] `http://research.microsoft.com/dryadlinqir`

[6] M. Najork, S. Gollapudi, and R. Panigrahy. Less is More: Sampling the neighborhood graph makes SALSA better and faster. In *Proc. of the 2nd ACM International Conference on Web Search and Data Mining*, pages 242–251, 2009.

[7] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of the 3rd Text REtrieval Conference*, 1994.

[8] Y. Yu, M. Isard, D. Fetterly, M. Budiu, Ú. Erlingsson, P. K. Gunda, J. Currey. DryadLINQ: a system for general-purpose distributed data-parallel computing using a high-level language. In *Proc. of the 8th USENIX Symposium on Operating Systems Design and Implementation*, pages 1–14, 2008.

[9] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft Cambridge at TREC–13: Web and HARD tracks. In *Proc. of the 13th Text Retrieval Conference*, 2004.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.