Research Issues for Large Scale Digital Library Search Engines in the Cloud: CiteSeerX

Principal Investigators

Dr. C. Lee Giles, Pradeep Teregowda, Dr. Bhuvan Urgaonkar
Information Sciences and Technology
Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802, USA
giles@ist.psu.edu

Cloud Research Interests:

- Data Portability, Consistency and Management
- Cloud Test-beds

Current Cloud Research activities: Information retrieval tools such as digital libraries and search engines are invaluable resources for finding authoritative and relevant resources on the web. However, building large scale digital libraries from crawling the web present several challenges, which include processing, ingesting, and indexing documents, and then presenting such resources to the user in a effective and accessible manner. SeerSuite, an open source framework for digital libraries, enables users to build digital library search engines such as CiteSeerX. In addition SeerSuite enables users to crawl the web in search of the most relevant documents, process these documents and access them through a web based user interface. It includes a user portal myCiteSeerX to support personalization and interaction. SeerSuite thus includes several components common to other information retrieval systems and web applications. Our interest is to extend the SeerSuite framework into the cloud using CiteSeerX as a testbed in order to take advantage of a cloud computing infrastructure both in house and for researchers and other users who wish to expand features already available in SeerSuite.

In many ways a digital library such as CiteSeerX is very attractive for cloud research and design. We'll list some of these advantages:

- The data structures and files are on the order of several terabytes making the system large but manageable.
- The data has few proprietary and privacy issues making it easy to use and share and is available under a creative commons attribute license.
- Use is substantial with several million hits a day and nearly a million unique users.
- The data is constantly growing and must be managed.
- New features are implemented on a regular basis.
- There are many unique information extraction services.

Thus, in many ways CiteSeerX can be seen as an instantiation of an enterprise search service and results from its cloud design will have broad implications in enterprise search.

Cloud Computing for Enterprise Search and Academic Digital Libraries

The main goals of our research are to deploy SeerSuite instances and components in a public cloud infrastructure, with an extended feature set. In particular we propose to investigate and research the following:

- Develop and deploy a scalable cloud based mechanism for crawling and processing documents from the web and improving the metadata extracted.
- Make available a cloud infrastructure for repository and repository access.
- Expose metadata available in the cloud through the index, repository and database by extending both the interfaces already available and add new interfaces, explore data access issues.
- Deploy myCiteSeerX in the cloud, examining various privacy, confidentiality, integrity and security issues with user data.
- Develop an optimization model for evaluation of the various components and their processes for SeerSuite cloud deployment.

Technical Description

The SeerSuite architecture is built around loosely coupled modules that exploit commercial grade open source applications and state of the art machine learning techniques. The components can be grouped into broad groups based on the tasks performed: Web application, data storage, document ingestion, document conversion, metadata extraction, and focused crawling. Components interact with each other through SOA/REST based interfaces and data access objects [1,2,3].

Expected Outcomes and Results

This research with SeerSuite and CiteSeerX will produce the following:

- A cloud infrastructure design for specialty and enterprise search engine and digital libraries.
- An initial deployment of such an engine in a private cloud.
- An investigation of the engine process, services and security needs.
- A study of document processing services in the context of Hadoop/MapReduce based deployments.
- An optimization model and tuning of services and software for cloud deployment.
- An investigation of large scale deployment of the above in a commercial cloud.

As a result, we intend to provide a clear guideline for deployment of enterprise search and digital libraries in the cloud, In addition all algorithms and software developed as part of this project will be made public and open source, to enable widespread community of users, researchers and developers to take advantage of algorithms and software.

References

1. BP.B. Teregowda, B. Urgaonkar, C.L. Giles, "Cloud Computing: A Digital Libraries Perspective", *3rd IEEE International Conference on Cloud Computing*, 2010.
2. P.B Teregowda, B. Urgaonkar, C.L. Giles, "Cost Implications Of Moving To The Cloud: A Digital Libraries Perspective", *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud '10)*, 2010.
3. P.B. Teregowda, I.G. Councill, J.P Fernández R., M. Kasbha, S. Zheng, C.L. Giles, "SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web", *1st USENIX Conference on Web Application Development*, 2010.