Citation Search Engine

Academic paper search engine

Master Thesis

Aliya Ibragimova
from
University of Fribourg

Faculty of Natural Sciences
University of Bern

January 2015

Prof. Dr. Oscar Nierstrasz
Mr. Haidar Osman, Mr. Boris Spasojevic

Software Composition Group
Institut für Informatik und angewandte Mathematik
University of Bern, Switzerland

Abstract

Nowadays the amount of documents in World Wide Web grows exponentially. Tools that can facilitate information retrieval present a particular interest in the modern world. A typical web search engine that search for information in World Wide Web is a software system that performs full-text indexing without considering meta information. This paper is devoted to the design of the academic paper search engine that takes advantage of meta-information, specifically citations. It is believed that citation is a very concise statement describing the source it refers to. Retrieving such statements can be particularly useful in writing scientific papers, for example, to build up a good argument.

This paper describes implementation of Citation Search Engine, a system that makes an attempt to automatically extract, index and aggregate citations from a set of scientific articles in PDF format. Besides it analyses the results of the deployment of the system on the collection of scientific papers provided by Software Composition Group.

Contents

1 Introduction

1.1 Thesis statement

We believe that considering meta information helps to build enhanced search systems that can facilitate information retrieval. Particularly, we target information retrieval for scientific papers. We consider citations as important text blocks summarising or judging previous scientific findings assisting in creating a new scientific work. We propose Citation Search Engine a software system that extracts citations from scientific papers, aggregates citations based on the referred source, then indexes extracted content. It provides a practical web interface that allows users to search for citations.

1.2 Goals

We set following goals:

- Introduce the state of the art techniques in information search.
- Explore the structure of scientific articles, reveal common patterns
- Design and implement the academic search engine.
- Deploy the system on the given collection of scientific papers
- Analyse results, define future work

1.3 Outline

The rest of the paper structured as follows:

Chapter 1 The chapter gives an overview of a typical web search engine and shortly reviews popular academic search engines.

Chapter 2 Devoted to the exploring the structure of scientific papers and identifying parsing challenges.

Chapter 3 Describes the desig and implementation of Citation Search Engine.

Chapter 4 The chapter describes the deployment process and analysis the result of setting up the system on the given collection of scientific articles.

Chapter 5 Contains conclusion and possible future work.

2
Related Work

2.1 Typical web search engine

Figure 2.1 illustrates a high level architecture of a standard web engine. It consist of three main components:

- Web Crawler
- Data indexer
- Search interface

Figure 2.1: A high-level architecture of a typical web search engine

Web Crawler is a program that browses the World Wide Web reading the content of web pages in order to provide up-to-date data to Data Indexer. Data Indexer decides how a page

content should be stored in an index database. Index helps to quickly query information. Users can search and view query results through Search Interface. When a user makes a query a search engine analysis its index and returns best matched web pages according to specific to indexer criterias.
Web crawlers that fetch web pages with the content in the same domain are called focused or topical crawlers. An example of focused crawlers are academic-focused crawlers that crawls academical documents. Such crawlers become components of the ”focused” search engines. Next chapter reviews some of popular academical search engines.

2.2 Popular academic search engines

CiteSeer^x CiteSeer^x is an autonomous citation search engine [3, 5]. It automatically parses and index publicly available scientific articles found on the World Wide Web. It uses the impact of citations to rank documents. CiteSeer^x is built on the open source infrastructure SeerSuite [6] and uses Apache Solr [2] search platform for indexing documents. CiteSeer^x can extract meta information from papers such as title, authors, abstract, citations. The extraction methods are based on machine learning approaches such ParseCit [1]. CiteSeer^x one of the world's top repositories and was rated number 1 in July 2010 [4]. It currently has over 4 million documents with nearly 4 million unique authors and 80 million citations.
At first glance it may seem that both CiteSeer^x and Citation Search Engine index citations in the same manner, however the approaches are different. By indexing citations in CiteSeer on should imply indexing bibliography links, while in Citation Search Engine we intend to index text of the citation in the body of the document.

Google Scholar
Microsoft Academic Search

3
The Problem

In which we understand what the problem is in detail.

4
Citation Search Engine

In which you describe your solution.

4.1 Overview of Architecture

4.2 Components

4.3 Parser -Challenges

5
The Validation

In which you show how well the solution works.

6
Conclusion and Future Work

In which we step back, have a critical look at the entire work, then conclude, and learn what lays beyond this thesis.

Bibliography

[1] Isaac G. Councill, C. Lee Giles, and Min yen Kan. Parscit: An open-source crf reference string parsing package. In *INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION*. European Language Resources Association, 2008.

[2] The Apache Software Foundation. Apache Solr. http://lucene.apache.org/solr/.

[3] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, DL '98, pages 89–98, New York, NY, USA, 1998. ACM.

[4] Cybermetrics Lab. Ranking Web of Repositories. http://repositories.webometrics.info/.

[5] The Pennsylvania State University. Citeseer. http://citeseerx.ist.psu.edu/.

[6] The Pennsylvania State University. SeerSuite. http://citeseerx.sourceforge.net/.