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 describes the possible ways of its enhancements. It describes one of the most popular search platform Solr, that will be used for building Citation Search Engine.

Chapter 2 Devoted to the exploring the structure of scientific papers and identifying future 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

In which we learn what have other done to address similar problems. For example, the work of Star [**?**]

2.1 A typical web search engine

2.2 Indexator - Solr

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.