Context dependent keyword extraction for short text matching

M.Tech Project Report

Phase I

Submitted in partial fulfillment of the requirements
for the degree of

Master of Technology

by

Jayaprakash S

Roll No: 123050045

under the guidance of

Prof. Dr. Pushpak Battacharya

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Mumbai

Acknowledgement

I would like to thank my guide, Prof. Dr. Pushpak Battacharya for the consistent support and guidance he provided throughout the semester. I also would like to thank the team Mr. Arjun and Mr. Swapnil for valuable guidance and discussions.

TODO:Write something more

Abstract

Information is represented in different for like text, images, audio, video etc. The aim of this work is to understand and improve the

Contents

Chapter 1

Introduction

Query represents the information need of user, by which retrieval systems are expected to fetch relevant documents from the collection. Traditional information retrieval focuses on the frequency of word appearance, co-occurrence statistics to uncover the relationships among query and documents.

1.1 Applications

The aim of this work is to explore various methods related to improving information retrieval utilizing the Semantic Web, ranking the relationships among concepts in the structured data and combining the Semantic Web with full text search.

1.2 Challenges

1.3 Roadmap

The remaining part of this report organized as follows: Section 2 introduces the semantic web. Section 3 explains similarity measurements between concepts from ontology and using the similarity measures in the ranking system. Section 4 describes essence of integrated semantic full text search and Broccoli system. Section 5 is about ranking methods for ranking relationships between two concepts.

Chapter 2

A closer look at the problem

2.1 Classification fo Images

Chapter 3

Literature Study

3.1 Unsupervised Keyword Extraction Methods

3.1.1 RAKE

3.1.2 TextRank

3.2 Supervised Keyword Extraction Methods

3.2.1 KEA

3.2.2 Naive Bayes

Chapter 4

Experiments

4.1 Unsupervised Approaches

4.1.1 Boosting based on occurrence and co-reference

4.1.2 TextRank - Modified

4.2 Supervised Approaches

4.2.1 Naive Bayes

Chapter 5

Conclusion

In this work, we have studied various ways of mapping between large text and short text through extracting keyword from large text.

In future, we are planning to using the external resources to better understand the relation between the large text and short text. For example, using the click-through logs will be helpful.

Chapter 6

Future work

In this work, we have studied various ways of mapping between large text and short text through extracting keyword from large text.

In future, we are planning to using the external resources to better understand the relation between the large text and short text. For example, using the *click-through logs* will be helpful.

6.1 Click-through logs