Creating and Testing a Semantic Search Engine for the University of Sydney

KYE FELTON

Supervisor: Yash Shrivastava
Associate Supervisor: Joseph Davis
Associate Supervisor: Masoud Salehpour

A thesis submitted in fulfilment of
the requirements for the degree of
B.Eng (Hons)

School of Electrical and Information Engineering
Faculty of Engineering
The University of Sydney
Australia

20 April 2021

Abstract

Many users rely on search engines to retrieve information from a site. However, despite its huge importance, many organizations still face the challenge of developing effective search engines for their users. This is a problem the University of Sydney is yet to resolve. Experience with the University's current search engine shows that queries searched for often return irrelevant or unsatisfactory information, resulting in a poor user experience. Thus, there is reason to improve the University's current search implementation.

To solve this challenge faced by many organizations, a number of search tools have been developed that focus on understanding the meaning of the query instead of searching for literal matches of words and variants. These applications are known as semantic search engines. In the last decade, the application of ontologies and knowledge graphs as an instrument for semantic searching has become increasingly prominent within industry. This approach typically involves building a knowledge graph from various information sources, and then developing a search engine that can query over the graph to find relevant information. Knowledge graphs however introduce their own set of challenges as automating a process to create an extensive, truthful graph from various information sources proves to be a difficult task.

Acknowledgements

Thank everyone.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Search engines have played an integral role in information retrieval since the development of digital information. Many organizations provide search engines on their public Web sites to allow both external and internal users to retrieve public information within their domain. Site search engines have a much smaller subset of information than Web search engines, allowing them to be more customized to the user and updated more frequently. Information retrieval (IR) systems. Have you ever experienced better results from searching on a Google than searching on the site's search engine? This was the experience with the University of Sydney's search engine.

1.2 Scope

Several factors, including relevance, coverage, freshness, response time, and user interface, should be considered when evaluating the effectiveness of a search engine. Among these, relevance is the most critical but often hardest to get right. (Reference: Tefko 1, 2). Accordingly, this objective of this research paper is to design and implement a search engine that delivers relevant documents for a given query on the University's site.

The University currently provides several search engines:

- University
- Library

- Current Students

To design a search engine, it is important to note the key features that make that comprise the problem space.

- users: The University faces the challenge of accommodating both internal stakeholders (e.g. researchers, staff, students) and external stakeholders (e.g. future students, media). The search engine must be able to accommodate queries from students, researchers and staff, both current and future.
- data types: Documents types can range from contact info, images, videos, journals and unit outlines
- security: It is imperative that only users are only shown documents that they are permitted to have access to.
-

Once a the search engine has been developed, its effectiveness needs to be measured to evaluate if the new model has any improved results. Evaluation method will be discussed in future chapters.

1.3 Outline

CHAPTER 2

Literature review

2.1 History of Site Search

The first occurrences of site search had relied on literal matching of words determine relevance. This approach is referred to as lexical search. Documents would be broken down into words or variants (referred to as terms or bag-of-words), and then indexed based on the frequency of terms. Documents were matched and ranked to a query based on the degree of matching between terms. Although simple, this approach is fairly effective and is still prevalent in industry (Li and Xu 2013) [131, 52, 6].

2.2 Semantic Search

Instead of following a formula to match terms from queries to documents, semantic search aims to understand the meaning behind the query entered to deliver more relevant results. Meaning can involve understanding the user's intent and the contextual meaning of words. Consider the following example…

There are several approaches to developing a semantic search engine - semantic matching, knowledge graphs.

2.3 Knowledge Graphs

The development of knowledge graphs to aid semantic searching has increased in popularity. The goal of knowledge graphs is to reduce noisy Internet content to machine-readable facts about entities and their associations. Instead providing a link to a relevant document, once a search engine understands what the user is looking for, it can utilize the knowledge graph to return information that centers around the entity queried. For example...

2.3.1 Construction

Knowledge graphs consist of entities, attributes, classes, constraints (optional) and associations. Entities

2.3.2 Curation

2.3.3 Searching

2.4 Evaluation

2.4.1 Limitations

CHAPTER 3

Methods

Text.

3.1 Section

More text.

3.1.1 Subsection

Even more text.

CHAPTER 4

Results

Text.

4.1 Section

More text.

4.1.1 Subsection

Even more text.

CHAPTER 5

Conclusion

Something concluding.

Bibliography

Li, Hang and Jun Xu (2013). *Semantic matching in search*. Vol. 7. 5, pp. 343–469. ISBN: 1500000035. DOI: 10.1561/1500000035.