Informatics Institute of Technology

In Collaboration with

University of Westminster, UK.

Duplicate Question Identification

A dissertation by

Mr. U. S. L. Perera

Supervised by

Mr. Saman Hettiarachchi

Submitted in partial fulfillment of the requirements for the

BSc (Hons) Software Engineering degree

Department of Computing

May 2016

Abstract

In the modern world, Internet has become a more popular place to seek information. And in Internet, Q&A websites are one of the most promising options available to get questions answered. Since its inception, number of users using these sites and accessing the information available in it has increased dramatically. The reason for this popularity is people trust Q&A sites due to the good content that is available. Nevertheless, due to the freedom that exists to ask questions, these sites have become much more difficult to manage because the same question is being asked again and again. Moreover, it has been a hard time for people looking for answers to get the right answer since the answers are scattered everywhere. However, some sites have introduced methods to close these types of questions and mark them as duplicates but requires it to be done manually by searching for questions. Since the same question can be interpreted in many ways this has not being the best solution so far. Also, searching through a large set of questions to identify duplicates will never be easy.

This project presents a portable software library that can be used to identify duplicate questions. The speciality of this library is that it can be incorporated to work as a search engine for any kind of software application. The library is developed with two major components to calculate the semantic similarity. First component uses distributional semantics, which uses a Tfidf matrix and a term co-occurrence matrix. The second component uses WordNet, a lexical ontology database for word sense disambiguation and similarity calculation. The results obtained from the first component is passed on to the second component to do further processing and get more accurate results. Also bigram, a character-based technique is used to manage misspelled words. The uniqueness of this project is that it uses a hybrid approach, both corpus-based and knowledge-based approaches to calculate similarity thus capable of finding documents with both semantically similar and related. The prototype is developed and evaluated with different techniques by using the Stack Overflow's data dump. Evaluation results show an improvement of accuracy by 15.9% and 16.38% when compared with Apache Solr for recall-rate@5 and recall-rate@10 respectively.

Subject Descriptors I.2: Artificial Intelligence, I.2.7: Natural Language Processing, H.3 Information Storage And Retrieval, H.3.3 Information Search and Retrieval

Keywords Tfidf, Ontology, Semantic, Word Sense Disambiguation, Corpus-based, Knowledge-base, Q&A, Duplicate