Approach Research of Keyword Extraction Based on Web Pages Document

Yangxin Yu

ABSTRACT

Keywords selection is a fast growing industry in which different tools are used by the companies to suggest their webpage's keywords. The paper’s aim is to propose a method that suggests the keywords of a webpage based on the frequent terms. The method used in this paper is term frequency for defining the frequent terms. An Experiment is executed to validate the method results; and the result of the new method is compared to Google tool. The accuracy of the proposed method is 82.4%, which consider being a promising result. The experimental results show that this system in the judgment and the readability of web page content were superior to the general web page design of automatic summarization.

INTRODUCTION

With the rapid development of Internet, a variety of data on the Web has increased, the network has become a potential source of data warehouse and knowledge. Keyword selection is used in many applications as an example preprocessing, text classification, web mining, semantic web, sponsored search. Semantically keyword extension would allow improving the quality of the selected keywords.

The proposed approach is divided into four steps. First, loading and parsing the webpage into tokenizes (small pieces). Second, removing the Stop words from the list of terms extracted from the parser, and then the third step is stemming the terms

Yangxin Yu Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huai’an, China, 223003
to return to the word's stem. Finally the list of terms and their frequency is extracted from the proposed system.

RELATED WORK

The process of keyword suggestion is so important for different fields, e.g. semantic web, data mining, natural language process and advertising. Advertising companies use different tools that suggest keyword. This step is so important since the most appropriate keywords could maximize their profit using the click through rates [1].

Keyword suggestion can be classified into proximity search, querylog mining, and metatag spidering. Proximity search methods extract from the search engine's result pages terms that already exist in the search term.

The second type is the query log mining method, this method suggests past queries containing the search term. The Google Adwords Tool and the Yahoo Search Marketing Tool are known examples for this method[2].

The last method which is the Metatag spidering method uses the engine with the seed and extracts metatags from the best suggestion ranking. This method considered to be the lower quality than the other methods [3]. Using the semantic meaning of the different words and extending keywords by semantically related phrases could be used by the advertiser's website.

Also keyword selection tools are used by different search engines to provide relevant search results as a response to user query. Google Adwords tool can be used for keyword selection purpose[1]. However the drawback of this tool is the proximity based searches in which the keyword extracted must be listed as a search term.

PROPOSED APPROACH

This study aims at discovering how inspectors of the different object oriented background perform defect detection using Traceability-based reading technique, namely OORT (Object-oriented reading technique). The software inspector’s object oriented backgrounds include undergraduate degree earned, undergraduate object-oriented courses taken, undergraduate degree performance, and object-oriented work experience.

The steps of the proposed method pass through three main modules which are loader, parser, stop-words remover and stemmer as shown in Fig.1. The input of the proposed approach is the webpage and the output is list of words and their frequencies. The following figure shows the proposed approach framework.

The Parser module is the first step in which a parser is a program that breaks large units of data into smaller pieces which are called tokens. After the webpage is
fed to the system, the parser divides it into tokens. The output of the parser phase is the list of the words in the webpage. The second phase is the stopwords remover.

![Diagram](image)

Figure 1. Proposed approach three Phases which include Parser, stopwords remover and stemmer.

Stop words are common words that carry less important meaning than keywords. Usually search engineers remove Stop words from a keyword phrase to return the most relevant result. Examples of the Stop words are: the, an, a, are.

Different Stop words removers can be used as Porter Stem Analyzer[5]. The proposed approach uses Lucene English Stop words removal as it includes the most existing Stop words list.

After removing the Stop words from the list of keywords, the relevant words can be discovered. Stemmer can start working on these words.

Stemming is the process of reducing inflected or sometimes derived words to their stem, base or root form e.g. education to educate. The process of stemming is important for search engines, query expansion or indexing and natural language processing problems.

For preprocessing and extraction of keywords, an application was developed using Java language environment that includes Java Integrated Development Environment (IDE) which is the NetBeans and Java Virtual Machine.

The proposed framework is implemented using Java, a well-known Object Oriented language. Different modules were added e.g. Snowball Stemmer.

**EVALUATION AND TEST RESULT**

The experiment design was mainly divided into two parts. The first part examines F-measure (precision and recall); the second define the accuracy of the proposed approach.
F-measure is the combination of Precision (the percentage of positive predictions that are correct) and the recall (the percentage of positive labeled instances that were predicted as positive).

To calculate the F-measure it is required to compute the Precision and Recall values. The following equations are used for this purpose.

\[
\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}}
\]

\[
\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}}
\]

\[
F_i(r, p) = \frac{2rp}{r + p}
\]

To accomplish the experiment design, a data set for web pages should be used. DMOZ data set (Open Directory) is used for this purpose. DMOZ or The Open Directory Project is the largest, most comprehensive human edited directory of the Web. It is constructed and maintained by a global community of volunteer editors. The Open Directory was founded in the spirit of the Open Source movement, and is the only major directory that is 100% free. The data set used consists of 50 Web pages of shopping domain extracted from DMOZ dataset.

For the purpose of defining the proposed approach validation F-measure is used. Based on the 50 Web pages in the dataset, the value of F-measure is 0.7 compared to the Google Adwords keywords.

CONCLUSIONS

The paper provides a framework for web page keyword suggestion; the framework is ready for adding other modules in the future work since it is implemented using one of the Object Oriented Languages which is Java. The input of the approach is the webpage and the output is a list of keywords suggested.

The process of suggestion in the proposed model depends on the combination of parser, Stop words removal and stemmer. After finishing this preprocessing step, the words and their frequency is calculated. The experiment depends on the comparison between each suggested word by the proposed approach and the Google Adwords tool. The accuracy of the proposed approach is measured along with the precision and the recall values.
REFERENCES