Webpage Keyword Extraction Using Term Frequency

Lamiaa Mostafa

***Abstract*—Search engines are used nowadays by all internet users. Keywords selection is a fast growing industry in which different tools are used by the companies to suggest their webpage's keywords. It is important to understand how different search engines would choose a webpage in a search results based on user's query. 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 adword tool. The accuracy of the proposed method is 82.4%, which consider being a promising result.**

***Index Terms*—Suggestion, terms frequency, webpage, search engine.**

I. Search Engine Introduction

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 .

Online adverting depends mainly on text adverting , since advertisers should aim to increase the volume of the bid phrases and also to chose the most relevant phrases to their products, otherwise online users would not click on the ad that transfer them to the purchase page of the product or the service being promoted .

The proposed approach is divided into four steps. First, loading and parsing the webpage into tokenizes (small pieces). Second, removing the Stopwords from the list of terms extracted from the parser, and then the third step is stemming the terms to return to the word's stem. Finally the list of terms and their frequency is extracted from the proposed system.

The rest of the paper is organized as follows. First, the author describes the problem. Then, previous work related to keyword selection is listed. After that, the proposed method is presented; the next part will describe the system snapshots. Finally the proposed method's performance is evaluated on the available dataset and the results were discussed. The paper conclusion includes the benefits of the proposed approach and the directions for future work.

Manuscript received June 17, 2012; revised July 11, 2012.

Lamiaa Mostafa is with Business Information System Department, Arab Academy for Science and Technology and Maritime Transport Alexandria, Egypt. (e-mail: Lamiaa.mostafa31@gmail.com).

II. Related Work

A. Keyword Suggestion Concept

The process of keyword suggestion is so important for different field e.g. semantic web, data mining, natural language process and advertising. Adverting 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, query-log mining, and meta-tag spidering [2].

Open Information Extraction (OIE) focuses on domain independent and scalable extraction of terms without requiring human input. [2] Proposes a model to OIE over search query logs, clusters generated from the query logs can be very effective in web search.Proximity search methods extract from the search engine's result pages terms that already exist in the search term.

Using the semantic meaning of the different words and extending keywords by semantically related phrases could be used by the advertiser's website [3], [4].

The Google Adwords Tool [5] and the Yahoo Search Marketing Tool are known examples for this method. The second type is the query log mining method, this method suggests past queries containing the search term[6]-[8].

The last method which is the Meta-tag spidering method uses the engine with the seed and extracts meta-tags from the best suggestion ranking. This method considered to be the lower quality than the other methods [9].

B. Keyword Suggestion Applications

The advertising companies use the research studies to reach more profitable level. Research studies investigate the suggestion keyword tools. Semantically related phrases can be extracted from any webpage and used to define the most important keyword to represent this webpage [1].

It is very important to extract keyword list that are less common but also most representative to a webpage, this is can be achieved by defining the most similar words semantically by using the statistical occurrence methods [3].

The list of the keywords that represent a specific webpage/document could be extracted not only by statistical methods but also by using the conceptual similarity knowledge [4].

The structure of the website can be improved by rearranging links between pages [6]; also the text content can be improved by identifying the most relevant keywords. Theses relevant keywords are extracted using Term Frequency Inverse Term Frequency (TFIDF) [6].

Search engine results depend on a query or keyword; sometimes users do not use the right or the specific keyword to reach the search goal [7].

DOI: 10.7763/IJCTE.2013.V5.672

Query recommendation systems are used to help the user rich his goal by suggesting some keywords that were used previously by the user or others (user log) .Another way to use help of an information source or a thesaurus [8], [10].

In [9], a novel approach was proposed which is called TermNet in which semantic relationships can be designed as a directed graph for defining the relevance terms extracted from a webpage.

C. Keyword Suggestion Tools:

Online advertisers bid on keywords through auctions, the winner of the auction can put his ad and links on the search result page of the search company when querying the Bidterm [2]. For the purpose of online advertising, different keyword suggestion tools also known as Bidterm suggestion tools [2] are used [5], [11], [12].

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], [2], [9]. However the drawback of this tool is the proximity based searches in which the keyword extracted must be listed as a search term.

One of the well known previous researchers who worked in online advertising proposed a method for phrase extraction for the advertising purpose using HTML tags and TFIDF and query logs [13].

Some systems solve the problem by adopting a strategy similar to proximity search, but using the result pages to build a bag of word vector for the seed, and then suggest keywords that have a similar vector [2] ,[9].

Worktracker is an online tool for keyword suggestion; the tool uses meta-tag spidering [9].

Query log is also used for keyword suggestion and enhancing the user keyword, however the query log failed to explore new words since it uses only the frequently listed query log data [9],[13].

III. Proposed Approach

The steps of the proposed method pass through three main modules which are loader, parser, Stopword 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.

A. Proposed Approach Phases

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 (tokenzs). After the webpage is fed to the system, the parser divides it into tokenzs. The output of the parser phase is the list of the words in the webpage. The second phase is the stopwords remover.

Stopwords are common words that carry less important meaning than keywords. Usually search engineers remove Stopwords from a keyword phrase to return the most relevant result. Examples of the stop words are: the, an, a, are different Stopwords removers can be used as PorterStemAnalyzer [14], [15], Weka3-6 [16]. The proposed approach uses Lucene English Stopwords removal [17] as it includes the most existing stopwords list.

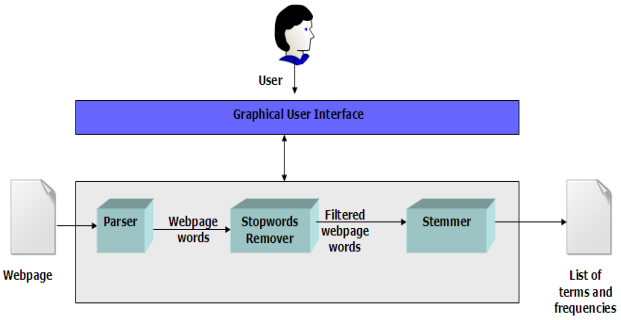

Fig. 1. Proposed approach three phases which include parser, stopwords remover and stemmer.

After removing the stopwords 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.

Different stemming algorithms can be used as Porter stemming [18], Lovins stemming algorithm [19] and Krovetz stemming algorithm [20]. One of the most used stemmers is the Snowball Stemmer [21]; this stemmer is used in the proposed framework.

The output of the stemmer phase is terms with their frequencies, it is important to count the frequency of the terms, so the terms with the highest frequency considered being one of the keywords of the webpage.

B. Approach Tools

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 stemming algorithm is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems, the stemmer used is Snowball Stemmer. Also the stopwords remover used is Lucene English stopwords removal.

IV. Proposed Approach System

The proposed framework is implemented using Java, a well known Object Oriented language. Different modules were added e.g. Snowball Stemmer. This section will describe the created system snapshots. The following figures will show the system's sequence.

The first step is feeding the webpage URL to the system, as described in the following Fig. 2.

The second step as explained in the proposed approach section is parsing, Stopword removal and stemming. This step is shown in Fig.3.

Fig.4. shows the last step which contains the list of the

terms and their frequencies.

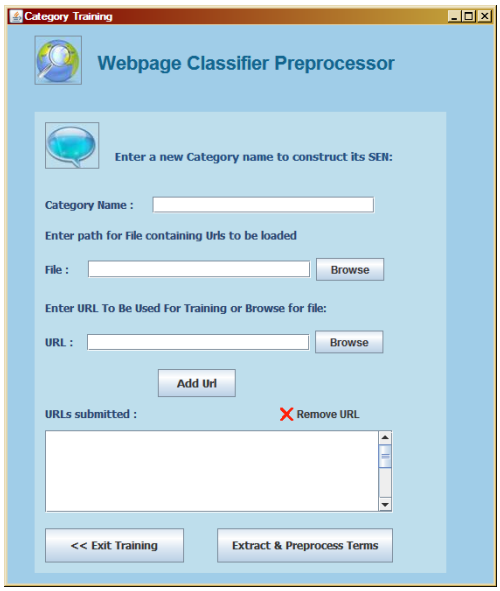

Fig. 2. The system screen in which the user can insert thewebpage URL.

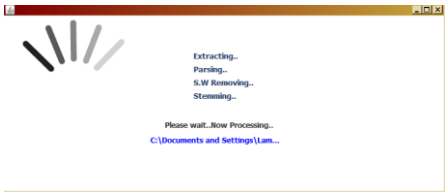

Fig. 3. The loading screen that represents steps of parsing,removal and stemming.

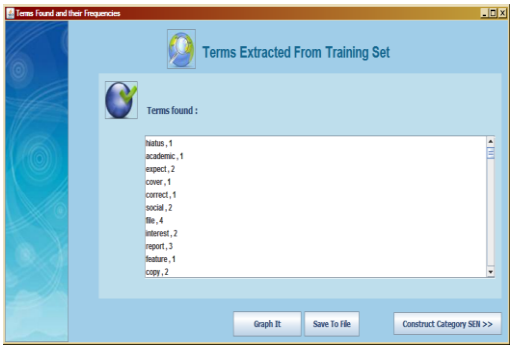

Fig. 4. The last step of the proposed systm that shows the output represented in list of Terms and their Frequencies

V. Limit Evaluation and Test Results

A. Experiment Design

The goal of the experiment is to validate the result of the proposed approach by measuring the F-measure and the accuracy level. The proposed method compares its keywords suggestion to Google Adwords keywords. The method suggests three keywords that can be used for webpage search query.

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[1] are used for this purpose.

$$\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items collection}} \quad (1)$$

$$\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}} \quad (2)$$

$$F_1(r, p) = \frac{2rp}{r + p} \quad (3)$$

To accomplish the experiment design, a data set for webpages should be used. DMOZ[2] 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 Webpages of shopping domain extracted from DMOZ dataset.

B. Experiment Results

For the purpose of defining the proposed approach validation F-measure is used. Based on the 50 Webpages in the dataset, the value of F-measure is 0.7 compared to the Google Adwords keywords as shown in Table I:

The experiment compares each keyword suggested by the proposed approach with the Google adword tool. Each correct suggestion will affect the value of precision and recall. After calculating the F-measure the accuracy level[3] is calculated for the proposed approach validation.

TABLE I:F-Measure and Accuracy Values

Precision	0.5467
Recall	0.8913
F-Measure	0.6777
Accuracy	82.467

VI. Conclusion and Future Work

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

1.http://wiki.uni.lu/secan-lab/docs/EugenStaab-E_and_F_measure.pdf [Retrieved Mar. 8, 2011]

2. http: //www.dmoz.org/ . [Retrieved Mar. 8, 2011]

3. http://www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf. [Retrieved Mar. 8, 2011]

suggested.

The process of suggestion in the proposed model depends on the combination of parser, Stopword 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 adword tool. The accuracy of the proposed approach is measured along with the precision and the recall values.

Future work could be started in creating keyword suggestion tool using different linguistic contents (e.g. Arabic webpages), using semantic relationships in choosing keywords and using spreading activation model along with the webpage preprocessing.

REFERENCES:

[1] S. Kiritchenko and M. Jiline, "Keyword optimization in sponsored search via feature selection," in *Proc. of JMLR: Workshop and Conference*, 2008, pp. 12-134.

[2] A. Joshi and R. Motwani, "Keyword generation for search engine advertising," in *Proc. of Sixth IEEE-ICDM*, 2006.

[3] V. Abhishek and K. Hosanagar, "Keyword generation for search engine advertising using semantic similarity between terms," *ICEC '07*, 2007, pp. 89-94.

[4] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and Bo Pang, "Automatic generation of bid phrases for online advertising," *WSDM '10: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 2010, pp. 341-350.

[5] Webpage. [Online]. Avaliable: https://www.adwords.google.com/o/Targeting/Explorer?__u=1000000000&__c=1000000000&ideaRequestType=KEYWORD_IDEAS#search.none.

[6] J. Vel ásquez, H. Yasuda, and T. Aoki, "Web Site Structure and Content Recommendations,". *IEEE/WIC Int. Conf. on Web Intelligence*, Beijing, China, September 2004.

[7] H. Zahera, G. Hady, and W. Abd El-Wahed, Query Recommendation for Improving Search Engine Results," in *Proc. of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010* San Francisco, USA, , October 2010.

[8] L. Li, S. Otsuka, and M. Kitsuregawa, "Query Recommendation Using Large-Scale Web Access Logs and Web Page Archive," LNCS 5181, *Springer-Verlag Berlin Heidelberg*, pp. 134–141, 2008.

[9] G. Chen and B. Choi, "Web page genre classification,"in *Proc.of ACM symposium on Applied computing, 2008*.

[10] J. Vel ásquez, S. Rıos, A. Bassi , H. Yasuda, and T. Aoki, "Identifying keywords to improve a web site text content," in *Proc. of 6th International Conference on Information Integration and Web-based Applications & Services* ,pp. 39-48, 2004.

[11] Freekeywords. [Online]. Avaliable: https://www.freekeywords.wordtracker.com/.

[12] Microsoft. [Online]. Avaliable: http://advertising.microsoft.com/support-center/adcenter-downloads/microsoft-advertising-intelligence.

[13] W. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on Web pages," in *Proc. of WWW*, 2006.

[14] L. Mostafa, M. Farouk, and M. Fakhry, "An Automated Approach for Webpage Classification," *ICCTA09 Proceedings of 19th International conference on computer theory and applications* ,Alexandria,Egypt, 2009.

[15] Koders. [Online]. Avaliable: http://www.koders.com/java/fid951FA1913F63BF9D69A987D01BDC123FB14B062.aspx?s=hibernate

[16] Sourceforge. [Online]. Avaliable: http://www.sourceforge.net/projects/weka/files/weka-3-6/.

[17] Lucen. [Online]. Avaliable: http://lucene.apache.org/java/2_3_2/api/org/apache/lucene/analysis/standard/StandardAnalyzer.html.

[18] Porter. [Online]. Avaliable: http://www.ils.unc.edu/~keyeg/java/porter/index.html.

[19] Stemmers [Online]. Avaliable: http://sourceforge.net/projects/stemmers/

[20] Lancs. [Online]. Avaliable: http://www.comp.lancs.ac.uk/computing/research/stemming/general/krovetz.htm.

[21] Snowball. [Online]. Avaliable: http://www.snowball.tartarus.org/