Journal of Computer Science 7 (9): 1358-1362, 2011
ISSN 1549-3636

Optimizing Title and Meta Tags Based on Distribution of Keywords; Lexical and Semantic Approaches

Mohammad Farahmand, Abu Bakar M.D. Sultan,
Masrah Azrifah Azmi Murad and Fatimah Sidi
Department of Information System,
Faculty of Computer Science and Information Technology,
University Putra Malaysia, Malaysia

Abstract: Problem statement: To increase traffic on websites, Search Engine Optimization (SEO) has provided many costly and time-consuming options. One problem is the inadequate distribution of keywords especially those keywords that users use the title tag and Meta tags. **Approach:** This study described work on an initial model for handling some of the SEO factors to increase the distribution of keywords. Our purposed model provide users with the words and their values based on the key weights with initiated formula to provide a new title, keywords, or description in order to increase the relativity between content and HTML Meta tags and title tag. **Results:** The proposed model had been showed evidence of gaining the greater utilization of the distribution of keywords and prevents recognition of search engine spam. **Conclusion:** The result shows the significant enhancement of the proposed model on Title Weight by 51.69% of original Title Weight defined by user.

Key words: Keyword distribution, ranking, Search Engine Optimization (SEO), spam recognition, keyword generation, keyword extraction, semantic

INTRODUCTION

Ranking is the most important element in web search engines since searching specific terms through search engines requires proper ranking to obtain good results. Proper ranking is also important in online advertisements. In general, there are two types of online advertisements associated with internet search engines: paid placement and Search Engine Optimization (SEO). SEO is the process of improving the volume and quality of traffic to a web site from search engines via natural search results. Achieving the high rank in search engines depends on more than 200 parameters (Evans, 2007). Site owners or expert users will be able to customize and improve the rank if they manage all these parameters and use them in proper position and condition.

With respect to the mentioned parameters, there is an obvious and logical relation between Title tag, Keywords and Description Meta Tags (TKD) and web site content. Relativity between Title tag, Keywords and Description Meta tags, (especially Title tag) and body in the web pages are vital. When a search engine spider analyzes a web page, it determines keyword relevancy based on an algorithm, which is a rather large and complex formula that calculates how web pages are ranked (Thurow, 2008). Thus, as the TKD distribution rate in body gets higher, the ranking will improved, because the effect of this relation and distribution is a reason for achieving higher-ranking position in Search Engine Result Page (SERP). On the other hand, search engines may also penalize pages or exclude them from the index if they detect search engine "spamming". For instance, one word is repeated hundreds of times on a page to increase the frequency for propelling the page higher in the listing. Search engines watch for common spamming methods in a variety of ways, including complaints from users. For this reason, proper distribution of keywords is a crucial and noteworthy issue for a web page.

In the SEO field, many researches already done and many theories have been developed. Nowadays designers and site owners have understood what they want; they demand good rank in search result page. Therefore, many of specialists in SEO designed and developed different models to obtain a satisfying result. Ramos used Term Frequency-Inverse Document Frequency (TF-IDF) In order to find the term frequency in a document to determine word relevance in document queries (Ramos, 2001). A complete and

Corresponding Author: Mohammad Farahmand, Department of Information System, Faculty of Computer Science and Information Technology, University Putra Malaysia, Malaysia

persnickety work on keyphrase extraction in HTML page performed Humphreys (2002), He introduced a novel keyphrase extraction for web pages, which requires no training, but instead his work was based on the assumption that most well written WebPages "suggest" key phrases based on their internal structure. It is very fast, flexible and its results are state of the art in key phrase extraction (Humphreys, 2002).

Another significant work has been done to analyze some factors, which are used in search engine ranking. Their factors was based on word length elements such as number of bytes of the original document, average term length and it did not involve major factors that users can manipulate them (Bifet, 2005), So the method is not so practical. There is also a model for generating keywords for search engine advertisements based on semantic similarity between terms (Abhishek and Hosanagar, 2007). To find and test some factors on ranking in a specified search engine-Google-an analysis was provided through search engine optimization data (Evans, 2007). In 2008, Thurow analyzed and collected most factors which had effect on search engine ranking and worked on a marvellous topic called "do and don'ts in SEO" which seemed to be necessary (Thurow, 2008).

Another excellent experience on extract keywords from abstracts and titles in academic papers, which are useful for small sized text documents, had done by (Bhowmik, 2008) but the main purpose is different from SEO. Recently, another research on automatic keyphrase extraction has been performed, named KP-Miner system that works on two languages (English and Arabic) (El-Beltagy and Rafea, 2009). This system is same as (Humphreys, 2002) and does not need to be trained on a particular document set in order to achieve its goal. In order to improve the web advertisement (Xing and Lin, 2006) worked on some factors for helping managers to make informed advertising decisions.

Recently, Kumar *et al*. (2010) worked on an algorithm to improve the Google ranking algorithm, PageRank. Different algorithms for link analysis like PageRank (PR), Weighted PageRank (WPR) and Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared by (Kumar, 2010).

In this way, the proposed method figured out a model in both sides of semantic and lexical process in different ways and it is a kind of keyword suggestion for TKD enrichment.

The issue of proper distribution and optimizing keywords in TKD: One of the search engine parts, which is called spider (crawler), collects information such as content, Meta tags, title and so on, from websites and sends them to search engine database for calculating the ranking of each website. The spiders may come to the site any time, during day or night and the "return time" that they come again for checking the site depends on factors such as ranking, update period and number of visitors. The first visit of the spider is very important for a website because after the first visit, the spider determines when and after what period, it returns. Therefore a website owner should optimizes the TKD in order to make positive impact on the spider decision before uploading the web pages for the first time, for determining the return period and also obtaining the highest possible rank on the first visit.

A correspondent distribution of keywords and phrases in body is necessary for achieving the better rank as well as TKD. Meanwhile, changing the TKD is not necessarily going to help the page in the ranking position if the page has nothing to do with these parts. Keywords need to reflect in the page content too. Therefore, the problem is inappropriate influence of TKD in body and vice versa.

Proposed model: Three of the most important factors selected for this research (Thurow, 2008). The chosen factors are "Title", "Keyword" Meta tag and "Description" Meta tag, which are called TKD. One of the goals is to find a proper distribution of TKD in the body. This research assumes that the user is a semi-expert web designer or developer. Therefore, it supposed that the description tag is meaningful and related to the body content. Although after the processing, the results show which part needs to be change or modify but the proposed model preferred a minimum description and keywords.

The objective of this research is developing a model for optimizing TKD via body words to improve the preliminary ranking and making the good TKD with proposed model suggestions. This objective comprises of four smaller goals. The first is reducing the spider return time, the second, obtaining the uppermost rank, the third, checking TKD standards and the last, recognizing Spam page.

MATERIALS AND METHODS

The proposed model methodology contains six steps and every one gains a part of result for users. After importing a HTML file as an input, data pre-processing begins. Next, character-analyzing section analyzes the words and characters. After finishing the analyzing, keywords analysis and generation will extract the word from body via Semantic dictionary. This extraction is also performing for title tag. Moreover, at the end with an initiated formula, the model demonstrates the suggestions to user for further actions.

Data pre-processing: Extracting text from HTML format is the first move in this data pre-processing step.

Then, recognizing and removing the stop words, special and non-standard characters will perform. The next step is Tokenizing and counting the words and sentences. Calculation of word frequency is the last step.

Character analysis: This phase has three parts. At first, TKD are extracted. Second, extracted parts should be counted and compared with the standard search engines (Google, Yahoo and MSN) in number of characters. Third, stop words are removed from extracted TKD.

Keywords analysis and generation: Firstly, the model recognizes the keywords and descriptions in HTML format and extracts them. Then, it extracts the words from body and description tag, which are valuable as keywords via setting a repetition threshold by user. The extracted words from description are added automatically to the keywords since it is assumed that the users who have entered the descriptions are experts and have written something related to the document. Next, it creates a list of words for the user to choose some and add them to the keywords. This part enriches the keywords for title analysis. In this part, the model may encounter some words, which have a proper distribution in body, but the user didn't use of them in title or Meta tags. Therefore, the proposed model suggests these new keywords to user for adding them in proper places.

On the other hand, for improving the results, the model finds the synonyms of each word in current title tag via WordNet repository (Miller, 1995). The model has used dictionary based semantic words (i.e., semantic model) to find the synonyms for each keyword. The WordNet repository, which was first provided at the cognitive science laboratory at Princeton University in 2006 (sponsored by Google), was exploited for the dictionary. Users can check and add them if they are related to the content.

Title analysis and generation: After customizing keywords, the most important part of the research and the most significant factor in SEO, (i.e., Title tag) will be processed. First, Title tag is recognized and extracted. Then, for finding the title words' weights, the model should calculate the real values of the words. For this reason, a formula is initiated in order to calculating the words weight more accurate. Actually, the proposed model finds the title words' weights in the content to compare them with other words. This comparison helps to normalize the words' values in the title. This formula has three variables: word ratio (λ1), word contained sentences ratio (λ2) and average of word presence concerning word contained sentences (λ3). Therefore, the final formula multiplies these three factors.

The formula is given in details:

$$\lambda 1 = \frac{TF}{TC} \quad (1)$$

Where:
TF = Term frequency
TC = Total content words
λ1 = Word ratio:

$$\lambda 2 = \sum_{1}^{n} \text{Sign}\,(w \in L_i) / n \quad (2)$$

Where:
w = Number of title words
Li = Number of content lines
n = Number of lines
λ2 = word contained sentences ratio

$$\text{Sign}\,(w) = \begin{cases} 0, w, \notin L_i \\ 1, w, \in L_i \end{cases} \quad (3)$$

$$\lambda 3 = \sum_{1}^{n} \frac{\text{sign}\,(w \in L_i) * \left(\sum_{j=1}^{k=|Li|} \text{Sign}\,(w == Lij) / k \right)}{\sum_{1}^{n} \text{sign}\,(w \in L_i} \quad (4)$$

Where:
k = Total number of words of line Li
J = Line Li words
λ3 = Average presence of word (in word contained sentences)

The body weight (λ) is defined as:

$$\lambda = \lambda 1 * \lambda 2 * \lambda 3 \quad (5)$$

On the other hand, title words' synonyms are extracted from WordNet. Aforementioned formula is also calculated for them. This way, keyword weight is obtained. Therefore, the improvement of the title and keywords is more accurate.

Spam recognition: On the other hand, one of the main reasons that a webpage cannot obtain a suitable rank is that it has been recognized as a spam page. When a keyword is repeated more than usual (it depends on number of words in document), search engines mark them as a spam page in the sense that these pages try to gain a higher rank in listings illegally. Therefore, our model finds all mistakes in a page by calculating the repetition of keywords and alerts to resolve the issue.

Spam recognition in this model is based on threshold and percentage. Users can adjust the number

of repetitions or set a percentage in the developed application and find any suspicious word.

Methodology of the model: For evaluating the results, we need a dataset of random HTML file with standard structure. It means, the HTML files should have Title tag and Meta tags that created by developers or website owners. Unfortunately, we could not find any dataset in this format for our experiment and we made it by ourselves. To obtain more accuracy and precious, 100 pages randomly selected on the internet although all HTML files could be our sample and it is possible to import them into our model.

The process of selecting the random page was base on heuristic method. First, several random word have been generated by Random Word Generator (Watchout4snakes.com. Random Word Generator (Plus), 2007). Then, these words searched on the internet (by Google) and saved the content of first URL of first SERP page (excluding online dictionaries, movies and TV shows and Wikipedia websites). Although, the volume of the page is not our concern and any page can be process, but we preferred to extract more keyword to gain tangible results.

This model uses One-Group Pretest/Post-test design for experimental procedure. It means the results show with comparison between before and after using proposed model. This comparison performs over two measurement tools. The first one is the most famous and important measure tool "Term Frequency (TF)" and second one is the new weighting formula that is our proposed model. In addition, this statistics applied on generated dataset on 1st SERP page URL's. Finally, the narrow collations on the results of this dataset show the accuracy of the proposed model on Title weight and TF as well.

However, the sentences recognition method makes use of "." for recognizing the sentences. It means the model assumes the web designer or website owner has used the correct punctuation. In addition, the total words number is calculated after reducing the stop words in order to increase the worthiness. Because when stop words have not reduced, the numerical results are very small and unworthy.

RESULTS AND DISCUSSION

The contribution of proposed model is helping users to decide about the words that they are good enough to use in TKD or not. The model determines the value of a word according to whether it should add to TKD or it should remove. This suggestion will help users to manage the title and Meta tags since each place is valuable.

Table 1: Diversity and standard deviation comparison on 1st SERP dataset

		TF	TW
1st SERP	Diversity average	4.70	55.43
	Standard deviation	6.70	80.79

Table 2: Number of positive, negative and neutral answers

	1st dataset
Positive results	80
Neutral results	16
Negative results	4

For this reason, our comparison between the word values is based on the "biggest λ value" of title words. The model is nominating the words with nominal variable as Good (≥75% λ), partially Good (≥50 and <75% λ), Fair (≥25<50% λ) or Bad (<25% λ).

Table 1 shows the comparison of Diversity and standard deviation on TF and TW. The results show that the proposed model dramatically increased the TW diversity. Also, TF that is one of the most important factors in ranking (Thurow, 2008) is increased.

The created dataset has 100 HTML files and according to Table 2, the positive results are equal 80. It means 80% of the cases are improved. However, 16 cases or 16% of the files are improved but less than the model expectation. Neutral answers are the answers, which their enhancement is less than TF diversity average (this value according to Table 1 is equal to 4.70).

This way of comparison will help to choose a normalized title and keywords for documents that are dependent on own document because the proposed model suggests the words according to the result of comparison between word values in document itself.

In addition, the proposed model calculates the words weight in keywords Meta tag. It means the model calculates the percentage of the words in the keyword Meta tag which are either the same or have a synonym word among the title words and vice versa. In this case, Word Net repository has been used for finding the synonyms.

Furthermore, in order to check the standards in character count, our application checks the length of TKD separately and compares them with three famous search engine standards, (Google, Yahoo and MSN).

In addition, the percentage of total title weight toward body calculates before and after using the suggested words via proposed model. It can help the users for choosing the best combination of words to making the improved title that it is the proposed model core.

The parameters used to evaluate our proposed approach are Precision (P), Recall (R) and the weighted harmonic of these two, which is the F-measure (Rijsbergen, 1979).

In sum, in the 1st SERP page dataset we used three these parameters calculated as below:

C = 80, I = 16 and T = 100 then,
P = 0.83, R = 0.96 and F-measure = 0.89

Where:
C = Number of correct answers (positive results)
I = Number of neutral answers (less than TF diversity average)
T = Total number of documents

These measurements show the efficiency and reliability of the proposed model, which Precision is more than 83% and Recall is about 96%.

CONCLUSION

The results are shown by extracting some words from body via lexical and semantic approach and customizing the TKD with them, "Total Title Weight" is improved. Also with spotting the suggestion of proposed model, keyword and description Meta tags become more accurate and relative to body. On the other side, by observing the standard length, the primary parameters that are directly involved in SERP will be improved. On the other hand, with increasing this factor, the "spider return time" will be shorter.

For expanding the model, the study is in progress on more than three factors and on using a neural network solution instead of dictionary-based solution for semantic web purpose. One of the on-going process is creating the dataset with low quality HTML files and compare the results of the model between two datasets in order to finding the efficiency and accuracy of the proposed model.

In addition, there are many possibilities to synthesis some techniques to improving the results by making the complete and detailed sentences based on semantic approaches for Title and description as well.

REFERENCES

Evans, M.P., 2007. Analysing google rankings through search engine optimization data. Internet Res., 17: 21-37. DOI: 10.1108/10662240710730470

Thurow, S., 2007. Search Engine Visibility. 2nd Edn., New Riders, Indianapolis, ISBN: 0321503244, pp: 292.

Ramos, J., 2001. Using TF-IDF to Determine Word Relevance in Document Queries. Rutgers University.

Humphreys, J.B.K., 2002. Phrase rate: An HTML keyphrase extractor. University of California, Riverside.

Bifet, A., C. Castillo, P.A. Chirita and I. Weber, 2005. An analysis of factors used in search engine ranking. Technical University of Catalonia.

Abhishek, V. and K. Hosanagar, 2007. Keyword generation for search engine advertising using semantic similarity between terms. Proceedings of the 9th International Conference on Electronic Commerce, Aug. 19-22, Minneapolis, MN, USA., pp: 89-94. DOI: 10.1145/1282100.1282119

Bhowmik, R., 2008. Keyword extraction from abstracts and titles. IEEE Xplore Press. DOI: 10.1109/SECON.2008.4494366

El-Beltagy, S.R. and A. Rafea, 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. Inform. Syst., 34: 132-144. DOI: 10.1016/J.IS.2008.05.002

Xing, B. and Z. Lin. The impact of search engine optimization on online advertising market. Proceedings of the 8th international conference on Electronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet, ACM New York, NY, USA, pp: 519-529. DOI: 10.1145/1151454.1151531

Kumar, R.R. and A.K. Singh, 2010. Web structure mining: Exploring hyperlinks and algorithms for information retrieval. Am. J. Applied Sci.,. **7:** 840-845. DOI: 10.3844/AJASSP.2010.840.845

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38: 39-41. DOI: 10.1145/219717.219748

Watchout4snakes.com. 2007. Random Word Generator. Creativity Tool.

Van Rijsbergen, C.J., 1979. Information Retrieval. 2nd Edn., Butterworths, London, ISBN: 0408709294, pp: 208.