# Relevant Keyword Extraction for Website to Achieve Top Search Engine Results

**Ranjitha K S[1], Manjuprasad B[2]**

[1]*Computer Science and Engineering, GSSSIETW Mysore*
[2]*Computer Science and Engineering (Asst. Professor), GSSSIETW Mysore*

***Corresponding Email: ranjitha.k.swamy@gmail.com***

## ABSTRACT

**Technology is changing swiftly, similarly the way business and customers are interacting is also changing. Customers are gradually moving their priority to e-business when compared to traditional business. It has become inevitable for each and every company to appear online and become easily reachable to its customers. It is not just enough to have a website, to be online. One has to make sure that the website is visible to the searchers who are searching for the company's product or service. One of the important Search Engine Optimization (SEO) technique to improve visibility and therefore to get more visitors to a website is keyword optimization. This paper provides a novel approach to extract keywords for a website using the descriptions given about the products. Stuffing these keywords in the on-page SEO factors such as the meta descriptions, image alt tags, heading tags, and URL is expected to help the website grab search engine's attention. The identified keywords with high search rate and relevancy can also be used in off-page SEO factors like blogs and articles. Using the keyword optimization technique discussed in the work, the website will be shown in the top search results.**

***Keywords: keyword extraction, Search engine optimization, visibility.***

## INTRODUCTION

The way world is operating today has drastically changed when compared to previous years. Today with the help of internet we get everything at our finger tips within no time. One important area which is affected by the internet technology is business. People are rapidly shifting from traditional business to e-business as it is more advantageous when compared to old style of business. It becomes very important in this internet era for every company to change from traditional business to e-business. The basic need in e-business for any company is to have website so that it will be reachable to customers using internet technology. The very first question that pops in one's mind is that is it enough if a company has a website?. The answer is definitely a no; the company has to make sure that its website is visible to the customers at the top search results in any search engine results. For this to happen the website has to be optimized as per the rules or criteria of the search engine. Search engine optimization is the process of making some changes to the website so that the website appears at the top organic search result and also the site's user experience is also improved. There are many optimization techniques [6] like keyword optimization, content optimization, structure optimization and link optimization.

Among the optimization techniques mentioned above keyword optimization plays an important role for any website. It is very important for one to know and understand the keywords that will be used by the customer to search for any product or service. Once these targeted keywords are found out; using them in the title tags, meta descriptions, image alt tags and also in blogs, twits and articles [7] will help a website in getting more customers. If wrong keywords are used then targeted customer will not be drawn to the site as a result people quickly get out of your site and hence increase the bounce rate.

Using relevant keywords that are more often used by the customers will make search engine place the website on top positions. Therefore it becomes very essential for an organization to extract the keywords related to its products or services. Keyword extraction is the process of selecting the related key phrases from the document [3-4]. This paper proposes a method that can be used to extract related keywords given the text that describes what the product is.

## LITREATURE SURVEY

This section surveys two methods that are available to extract keywords form the given documents and also proposes a new method to extract keywords.

### Method1: Keyword extraction from conversation fragment

Here the author [1] considers the problem of identifying relevant keywords in a query and use these keywords to recommend documents. To identify the keywords, some prior data is needed hence a transcript of conversation fragment is used to select the relevant keywords. Using these identified keywords the documents will be recommended. The keywords chosen should cover as much as possible topics that are covered in the discussion and avoid selecting irrelevant keywords. There a three important steps in identifying the keywords that is proposed by the author. The first step is topic modeling which aims at finding out how each word w is distributed in each topic z represented as p (z|w). Topic modeling methods like Latent Dirichlet Allocation (LDA) or Probabilistic Latent Semantic Analysis (PLSA) can be used for this purpose. The author has worked using LDA as PLSA is prone to overfitting problem. The second step is to assign a weight for each of the topic in the conversation fragment considered. Finally a list of best keywords will be selected based on how much it contributes to the topic.

### Method2: Keyword extraction using n-gram approach

The keyword extraction algorithm proposed in this work [2] has three important steps. The list of bi-grams arranged in descending order based on the Fuzzy Bi-gram Index (FBI) is extracted first. A list of n-grams arranged in descending order with respect to the Fuzzy N-gram Index (FNI) is obtained in the second step. The third step makes a decision of whether to include the monogram keywords or not as in some cases monograms words provide more information about the document when compared to bi-gram and tri-gram words. By using the frequency measure approach these kind of logic will be implemented. Monograms with a frequency eight times of the bi-grams will be included by the algorithm. The final result will have five top monograms, ten top bi-grams and top ten n-grams.

Fuzzy Bi-gram Index (FBI) is defined as the membership values of the two fuzzy sets HWO and HWPO. Fuzzy N-gram index (FNI) is calculated only after the extraction of all bi-grams and their corresponding FBI. Suppose two bi-grams appear adjacently in the text then it is analyzed to form n-grams. First check if the first word of the second candidate and the last word of the first candidate are same if so further check if their respective FBI is greater than 0.5. Then if both the condition is true then combine the bi-grams to form tri-gram. By averaging the bi-grams of the constituents the fuzzy index of new tri-gram is calculated. Application of this procedure is recursively done to extract the n-gram keywords.

High Word Occurrence (HWO) is defined as a fuzzy set that corresponds to individual words also its membership values indicates that a word has high degree of appearance. High Word Pair Occurrence (HWPO) is a second fuzzy set that corresponds to a word pair and whose membership value represents that the word-pair in the text document has high degree of appearance.

## PROPOSED METHOD

The proposed method here also needs some prior text or data in order to extract the keywords. This data is nothing but the description of the product or service the company or organization is aiming to advertize in their site. The given document (relevant description about product or service) is first preprocessed to remove the

unwanted text it contains. To accomplish this data cleaning process, a database of stopwords is first created so that these words can be eliminated. The given document is compared with the stopwords in the database and the matched ones are eliminated from the original document. Once eliminated the remaining words of the document is again manually verified and the best related words are selected as keywords. The words that are rejected after the verification will be again added to the database so that the verification process in the further steps becomes easier. So that one need not reject those words that were rejected previously again and again. Figure1 shows the proposed keyword extraction diagrammatically.

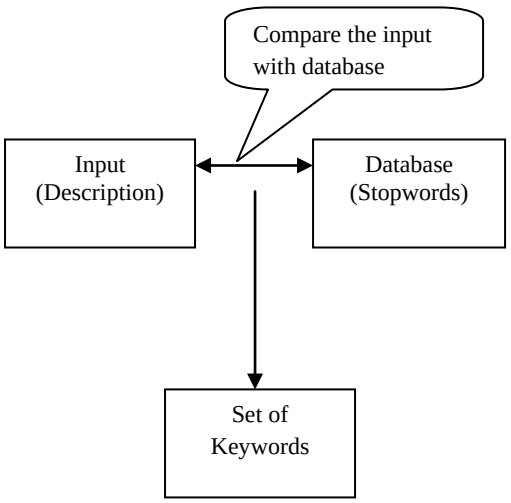


**Figure1: Proposed Method**

## CONCLUSIONS

Optimizing the website as per the norms of the search engine is very important for a company if it has a website. One such important optimization technique is keyword optimization. Identifying related and relevant keyword is an important task in keyword optimization. The work, has surveyed two existing approach that is used to extract keywords and has also proposed a new and simple that method that can be used to extract keywords from the given text document that has description about the product or service that the company offers.

## REFERENCES

[1] Habibi, Maryam, and Andrei Popescu-Belis. "Keyword extraction and clustering for document recommendation in conversations." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23.4 (2015): 746-759.

[2] Das, Bidyut, et al. "Automatic keyword extraction from any text document using N-gram rigid collocation." Int. J. Soft Comput. Eng.(IJSCE) 3.2 (2013): 238-242.

[3] Y. Matsuo and M. Ishizuka, ”Keyword extraction from a single document using word co-occurrence statistical information,” International Journal on Artificial Intelligence Tools, vol. 13, no. 1, pp. 157-169, 2004.

[4] Y. HaCohen-Kerner, ”Automatic extraction of keywords from abstracts,” in Proc. 7th Int. Conf. Knowledge-Based Intell. Inf. Eng. Syst., 2003, vol. 2773, pp. 843-849.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,”Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[6] Wang, Fuxue, Yi Li, and Yiwen Zhang. "An empirical study on the search engine optimization technique and its outcomes." Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on. IEEE, 2011.

[7] Gunjan, Vinit Kumar, et al. "Search Engine Optimization with Google." International Journal of Computer Science Issues 9.1: 206-214.