



POLİTEKNİK DERGİSİ

*JOURNAL of POLYTECHNIC*

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)

URL: <http://dergipark.org.tr/politeknik>



# Keyword extraction for search engine optimization using latent semantic analysis

*Gizli anlamsal analiz ile arama motorları için anahtar kelime çıkarma*

Yazar(lar) (Author(s)): Fahrettin HORASAN

ORCID: 0000-0003-4554-9083

**Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article):** Horasan F., “Keyword extraction for search engine optimization using latent semantic analysis”, *Politeknik Dergisi*, 24(2): 473-479, (2021).

**Erişim linki (To link to this article):** <http://dergipark.org.tr/politeknik/archive>

**DOI:** 10.2339/politeknik.684377

# Keyword Extraction for Search Engine Optimization Using Latent Semantic Analysis

## Highlights

- ❖ *SEO processes were performed automatically*
- ❖ *The study has been tested with a well-known data set.*
- ❖ *It has become easier to create content suitable for SEO*

## Graphical Abstract

The following figure shows the process of obtaining keywords from textual data used as web content, in stages.

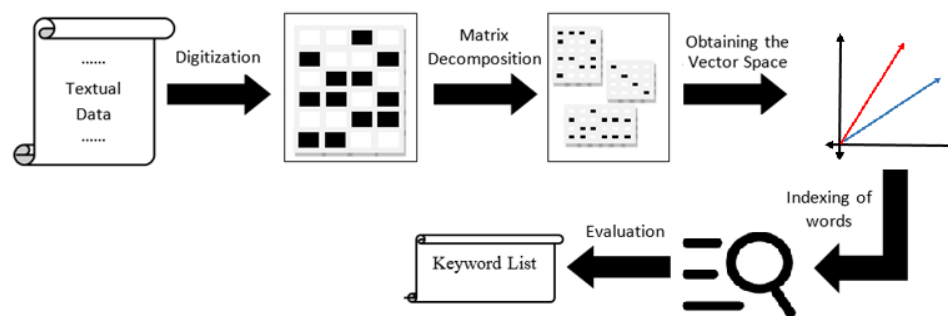


Figure. 1

## Aim

In this study, the keywords that best represent the content were obtained using the LSA in order to ensure that the content on the web pages comply with the SEO.

## Design & Methodology

Latent semantic analysis known as vector space based is used in this study.

## Originality

Keyword lists used in SEO were examined according to different dimension reduction parameters.

## Findings

Better results were obtained with the method based on the proposed term and document similarity.

## Conclusion

Keyword lists that best represent textual web content were obtained by considering their similarity with other terms/sentences.

## Declaration of Ethical Standards

The author of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

# Keyword Extraction for Search Engine Optimization Using Latent Semantic Analysis

*Araştırma Makalesi / Research Article*

**Fahrettin HORASAN\***

Engineering Faculty, Computer Engineering Department, Kırıkkale University, Kırıkkale, Turkey

(Geliş/Received : 04.02.2020 ; Kabul/Accepted : 11.04.2020)

## ABSTRACT

It is now difficult to access desired information in the Internet world. Search engines are always trying to overcome this difficulty. However, web pages that cannot reach their target audience in search engines cannot become popular. For this reason, search engine optimization is done to increase the visibility in search engines. In this process, a few keywords are selected from the textual content added to the web page. A responsible person who is knowledgeable about the content and search engine optimization is required to determine these words. Otherwise, an effective optimization study cannot be obtained. In this study, the keyword extraction from textual data with latent semantic analysis technique was performed. The latent semantic analysis technique models the relations between documents/sentences and terms in the text using linear algebra. According to the similarity values of the terms in the resulting vector space, the words that best represent the text are listed. This allows people without knowledge of the SEO process and content to add content that complies with the SEO criteria. Thus, with this method, both financial expenses are reduced and the opportunity to reach the target audience of web pages is provided.

**Keywords:** Search engine optimization, keyword extraction, latent semantic analysis, text mining.

## Gizli Anlamsal Analiz ile Arama Motorları için Anahtar Kelime Çıkarma

### ÖZ

Devasa bilgi yığınının bulunduğu internet dünyasında artık istenilen bilgiye erişmek zor hale geldi. Arama motorları bu zorluğun altından kalkmak için çaba sarf etmektedirler. Ancak arama motorlarında hedef kitlesine ulaşamayan bir web sayfası popüler hale gelememektedir. Arama motorlarındaki görünürlüğün artırılması için arama motoru optimizasyonu yapılır. Bu süreçte web sayfasına eklenen metinsel içeriklerden anahtar kelimeler seçilir. Bu kelimelerin belirlenmesi için hem içerik hakkında ve hem de arama motoru optimizasyonu konusunda bilgili bir sorumlu kişi gereklidir. Böyle olmadığı durumlarda etkili bir optimizasyon çalışması elde edilemez. Bu çalışmada gizli anlamsal analiz tekniği ile metinsel verilerden anahtar kelime çıkarma işlemi gerçekleştirilmiştir. Gizli anlamsal analiz yöntemi, metin içerisindeki doküman/cümle ve terimler arasındaki ilişkileri lineer cebir yönüyle modellemektedir. Elde edilen vektör uzayındaki terimlerin benzerlik değerlerine göre metni en iyi temsil edilen kelimeler listelenir. Bu işlem SEO süreci ve içerik hakkında bilgisi olmayan insanların da SEO kriterlerine uygun içerik eklemelerine imkan tanıyacaktır. Dolayısıyla, bu yöntemle hem maddi gider azaltılmış hem de web sayfalarının hedef kitlesine ulaşma fırsatı sağlanmıştır.

**Anahtar Kelimeler:** Arama motoru optimizasyonu, anahtar kelime çıkarımı, gizli anlamsal analiz, metin madenciliği.

### 1. INTRODUCTION

The increase in the usage of the internet and the facilitation of the publishing of web sites lead rapid increase in the number of web pages published on the internet [1]. Besides, the most important purpose of publishing a website is to ensure that the data is delivered to target users on time. Most of the time, under standard circumstances, it is time consuming for users to access the information they want on the internet since it requires the users to know all the content on the web and where the specific content is located. This issue is largely overcome by web sites called web search engines. According to World Internet Usage and Population Statistics data, 58.8 percent of the world's population use

internet technologies in mid-2019 [2]. Figure 1 shows the increase in the number of hostnames and active websites in the World. Difficulties are encountered in accessing correct information in the internet environment where there are too many internet users and websites [3]. The search engines are an information-navigation-tool that enables users to access the information they want on the internet [4]. On the other hand, in parallel with the development of web technologies, the importance of the search engines is increasing day by day [5].

The search engines scan contents of the web pages at certain time intervals via a software called crawlers or spiders and add the necessary information about the contents of the scanned web pages to their databases. Thus, they make it possible for internet users to access the desired content or web pages through themselves [6]. In order for the crawlers to find and scan the web pages, the

\*Sorumlu Yazar (Corresponding Author)  
e-posta : fhorasan@kku.edu.tr

URL details of the web sites need to be recorded in the database of the search engines [6]. For this reason, it is important for web developers to implement this addition process and regularly update the site contents in order for

the websites to get more visits [6,7]. The whole set of operations performed to ensure that websites are indexed in the search engines in the best way is called Search engine optimization (SEO)[8,9].

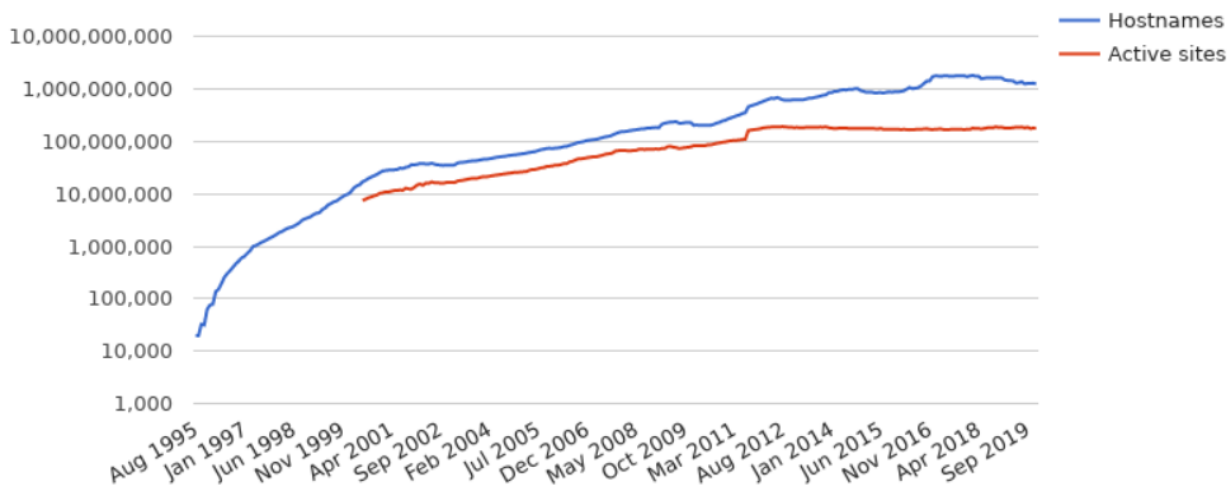


Figure 1. The increase in the number of hostnames and active websites in the World [2]

During the SEO process, some additional edits are made in the required places of the web sites. These edits may look like simple operations, but they have a significant impact on the site's search results. They may improve performance, but it should also be taken into account that they may reduce performance so, the edits need to be done carefully [6,7]. The web developers need to do a separate SEO work for each site since content and content presentation are different on each site. Moreover, the person performing the SEO process must be knowledgeable about the content posted on the web page, otherwise, the SEO process may cause the website to become a site that does not appeal to the target audience [7].

The SEO operations can be facilitated by designing the necessary processes automatically. There are many studies in this subject. First of all, the SEO operations should be carried out for the visitors, not the website publisher. The SEO directing users to the correct web pages are called White hat SEO [10]. Therefore, the main principle in SEO studies is to be user oriented. There are many studies in the literature that automatically perform the SEO process as user-oriented or facilitate this process. For example, a content management system that performs white hat SEO has been developed for an active web page called Fragformet [11]. This system, in which content is added and managed on the web page, realizes SEO automatically. One of the platforms that attract attention in the internet world is electronic market sites. A product-content optimization was developed in a study for the electronic market site. The basis of this study is multi-criteria optimization model such as discount time, visual presentation and product relations [12]. In another study, researchers have done a new heuristic scanning process with additional learning techniques that learn by

looking at the data that affects their web site's ranking on the search engine. They have proved that a system that automatically combines intuitive scans based on the data coming from users makes a system that gets better rank in search engines [13].

The search engines make recommendations according to the query sentence consisting of keywords or keyphrases [14]. Studies under the name of keyword extraction [15,16,17], query suggestion [18] and query classification [19] are based on this.

Latent semantic analysis (LSA) can be useful in finding keywords or keyphrases that best represent the semantic structure in the text [20,21]. In this study, the keywords used in the seo process of the contents added to the web pages were determined with the LSA model. The LSA, which is used in many fields such as text mining, image processing, data mining, signal processing, voice analysis, is a dimension reduction based approach [21,22,23]. The tests performed according to the parameters in the technique called rank-k in the dimension reduction stage were examined. Thus, a more efficient and faster recommendation process was obtained. The contributions of the developed model can be listed as follows.

- SEO processes are performed automatically for each content added to the web pages.
- Someone who has nothing to do with the content or seo stages can also add content that complies with the SEO criteria.
- The study has been tried with a well-known data set. It is comparable to future studies

- It was examined according to different dimension reduction parameters.
- The study can shed light on studies such as topic extraction, query classification.

In the next part of the paper, the realization of the seo process with the latent semantic analysis is explained. In Section 3, experiments and test parameters are explained. The last part is the discussion and conclusion section.

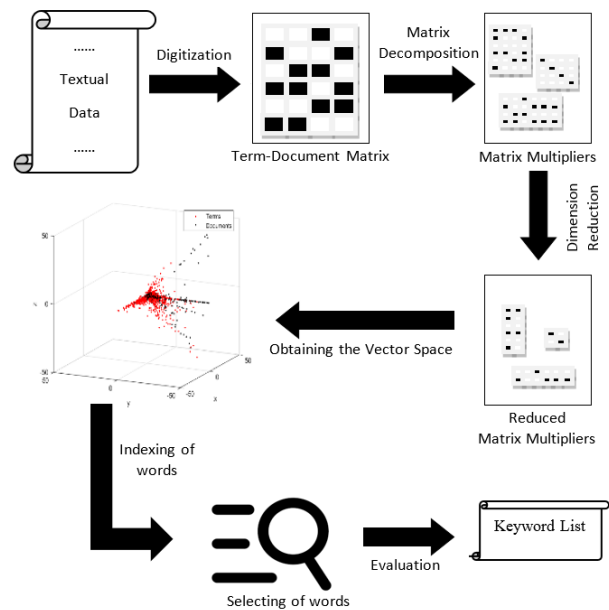
**2. SEO VIA LATENT SEMANTIC ANALYSIS**

The Latent semantic analysis is a statistical / mathematical technique that reveals latent relations between term-term, term-document and document-document [22,23]. The LSA, which is a dimension reduction based approach, aims to consider only the important data groups in the dataset. Data included in the data stack that does not contribute to the meaning or negatively affect the meaning are not included in this process. For this, the low-rank approach of the term-document matrix is used to find the latent semantic structure between the term and the documents [23].

Terms and documents are represented by elements of the row and column of the matrix, respectively. This matrix is called as term-document matrix. i-th row and j-th column of the term-document matrix contains the mathematical value of i-th term in j-th document. This mathematical value is known as the weight. The calculation of this mathematical value is known as weighting. The weight of the document for each term is calculated according to some methods [21,23].

Usually, the textual dataset is processed in the LSA. By passing through this dataset through the pre-process, stop words and punctuation marks are removed. Stemming is applied for each term. Then the weights in the term and document matrix are obtained. In this study, the most used TF \* IDF method was chosen among the weighting methods [21,22,24]. Thus, the term-document matrix ( ) is obtained by using the weight of the term in each document. The SVD is applied to the obtained matrix. The rank k approach is applied to the matrices obtained as a result of the matrix decomposition in order to reduce the dimension. After applying the Rank k approach, the term matrix and the document matrix are obtained by multiplying the left and right orthogonal matrices by the singular value matrix, respectively [25,26]. Each row in the term matrix represents the vector of the same indexed term in the term-document matrix. Each column in the document matrix represents the vector of the same indexed document in the term-document matrix. Thus, the term and document vectors are represented in the same vector space. After obtaining the vector space, documents/terms are listed according to the query from most similar to least similar. Ultimately, the documents / terms associated with the query (This can be a document, term or sentence) are discovered [23,27,28].

In this study, a single text is considered as a data stack and each sentence in this text is used as a document. The term document (sentence) matrix is obtained for terms in each sentence. As mentioned in the previous paragraph, the SVD is applied to this obtained matrix. The vectors of terms and documents are determined by the value of k in the rank k approach. The terms are listed according to their proximity to each other, taking into account their similarities to all terms and phrases mentioned in the text. Thus, word lists that have a very similar resemblance to terms and documents can be accessed in this way. These are the words that have the most discrimination in the text and can represent the text. Figure 2 shows the flow chart of the study.



**Figure 2.** Flow chart of keyword extraction technique with LSA

**2.1. Singular Value Decomposition**

The SVD of the  $A \in \mathbb{R}^{m \times n}$  is found by the formula

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T \tag{1}$$

Here  $m > n$ ,  $U^T U = U U^T = I_m$  and  $V^T V = V V^T = I_n$ . In addition,  $\Sigma$  matrix of diagonal and containing singular values of A is in the format

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq \sigma_{k+1} = \dots = \sigma_n = 0. \tag{2}$$

**2.2. Rank-k Approach**

The rank k approach is applied to reduce the cost of calculation and increase efficiency in Formula 2. When the rank of A matrix is k

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > \delta \geq \sigma_{k+1} \geq \dots = \sigma_n. \tag{3}$$

Here,  $\delta$  represents the threshold value. In order for  $\delta$  to be the most optimal, the difference between  $\sigma_k$  and  $\sigma_{k+1}$  should be significantly higher.

If the rank approach of the matrix A ( $A_k$ ) is used instead of the matrix A in LSA,  $A_k$  is represented by the equation  $A_k = U_k \sum_k V_k^T$ .  $U_k$  and  $V_k^T$  in the formula represent the first k columns of  $U$  and  $V^T$  matrices, respectively.  $\sum_k$  is the diagonal matrix  $\sum_k = \text{diag}(\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_k)$ .

### 2.3. Obtaining Vector Space

Representatives of terms and documents in vector space are obtained as  $T_k$  ve  $D_k$ , respectively, with equations

$$T_k = U_k \sum_k \tag{5}$$

$$D_k = \sum_k V_k^T \tag{6}$$

Herein, the i'th row of the  $T_k$  matrix is the vector symbolizing the term i'th, and the j'th column of the  $D_k$  matrix is the vector symbolizing the j'th document.

### 2.4. Listing Words

At this stage, words are listed in three ways. According to the first, only other terms are taken into account when calculating the similarity of the terms. According to the similarity value, the terms are listed from the least value to the most value. This listing method is Term Similarity Based listing (TSBL). The second is that the words are listed according to the similarity of the documents. The name of this method is Document Similarity Based Listing (DSBL). Another method of listing is the method in which the similarity of terms and documents are calculated together. The name of this method is Term and Document Similarity Based Listing (TDSBL).

In calculating each of these methods, the cosine similarity technique was used. The cosine similarity technique was preferred because it makes an angular similarity measurement. The cosine similarity technique takes into account the cosine value of the angles at which two vectors intersect with each other in the same vector

space [21,24]. The similarity between the two vectors is calculated by the formula

$$\text{Cos\_Similarity}(X, Y) = \frac{X \cdot Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \tag{7}$$

In the formula,  $X$  and  $Y$  represent  $m \times 1$  dimensional vectors.

### 2.5. Choosing Words and Evaluating

It is the selection of the best  $N$  of the words listed in the previous stage.  $N$  keywords are determined by selecting the best  $N$  of the terms listed in order from least similar to most similar. The keyword lists obtained were examined according to the TSBL, DSBL and TDSBL techniques. In addition, according to rank k approach, their performances according to different k values were examined. As an example, Table 1 shows the 20 most similar word lists according to the different rank-k values of a text according to the TDBL technique.

## 3. EXPERIMENTAL ANALYSIS

In this study, BBC news collection was used as a data set. In this collection, there are 1313 documents and 15393 words under 5 classes in total.

The performances of the keyword lists obtained according to the TSBL, DSBL and TDSBL techniques were examined according to the rank k approach. In the word groups listed according to two different k values, the number of similar words ( $\mathfrak{n}$ ) and the similarity ratio ( $sr$ ) were examined.  $sr$  is calculated according to the equation

$$sr = \frac{\text{The number of similar terms}}{N} \tag{8}$$

Here, N is the number of the best N terms.

The algorithm complexity of the keyword extraction technique using the  $m \times n$  dimensional term document matrix is  $O(mn^2)$ . In this study, where the Rank k approach is used, the algorithm complexity is  $O(mnk)$ .

It is also considered to be  $k < \min(m, n)$ . Thus, a less costly system was developed.

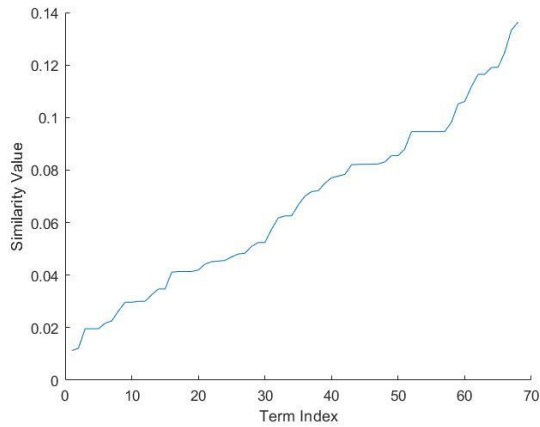
During the analysis process, keyword extraction was performed for all documents in the dataset. Firstly, figures 3.a., figures 4.a. and figures 5.a., which show

**Table 1.** Words listed according to different k values

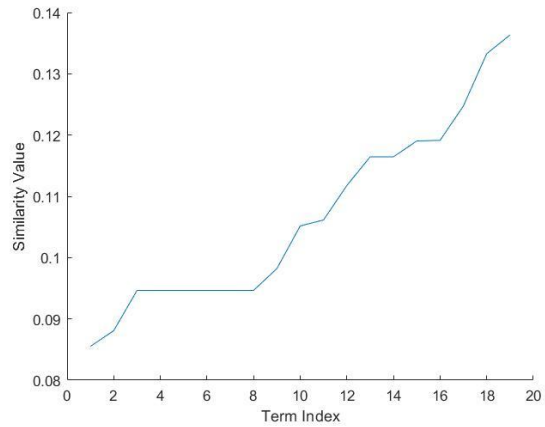
k	Ordered Word List
20	boost, profit, timewarn, year, earlier, high, speed, internet, aol, revenu, catwoman, contrast, third, final, lord, ring, trilog, full, chief, execut
15	boost, profit, timewarn, year, earlier, internet, aol, revenu, help, box, offic, alexand, catwoman, sharp, contrast, final, ring, trilog, full, post
<b>Similar Words</b>	boost, profit, timewarn, year, earlier, internet, aol, revenu, catwoman, contrast, final, ring, trilog, full

similarity changes according to the TSBL, DSBL and TDSBL techniques mentioned in Section 2.4, should be examined. In these figures, the terms are sorted according to their similarity values to all documents. As can be seen, there is an increasing similarity change. In figure

3b, figures 4b, and 5b, which show the performances of the 20 words that resemble the best, the similarity change of the term groups that can represent the document better is seen.

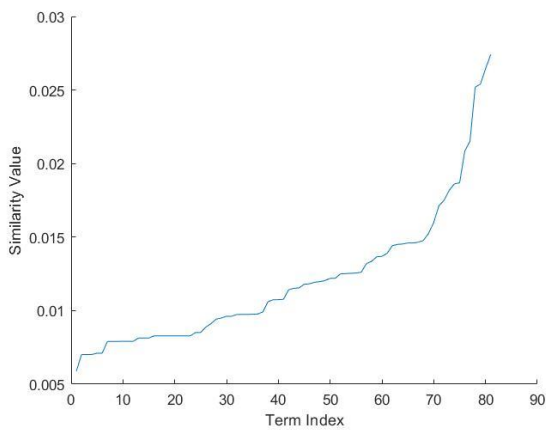


a. All of the listed terms

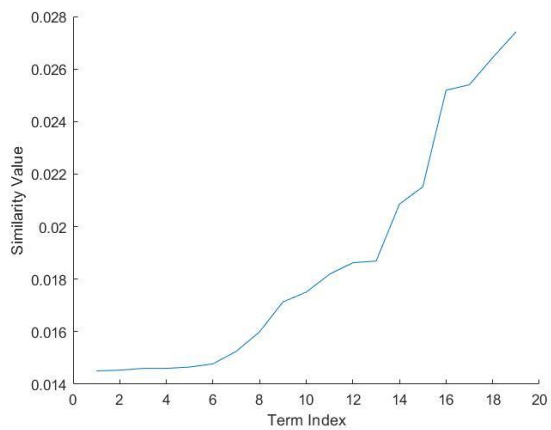


b. Top 20 of the listed terms

**Figure 3.** Similarity changes according to the TSBL technique

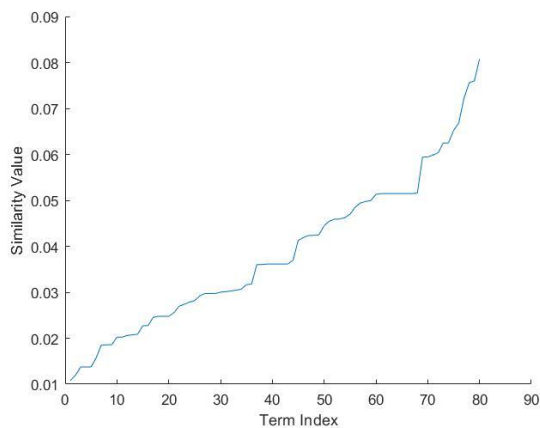


a. All of the listed terms

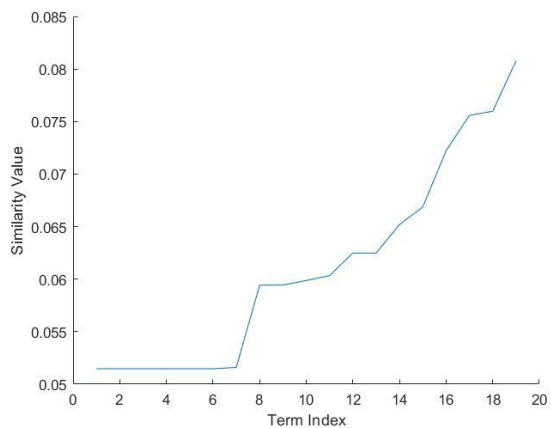


b. Top 20 of the listed terms

**Figure 4.** Similarity changes according to the DSBL technique



a. All of the listed terms



b. Top 20 of the listed terms

**Figure 5.** Similarity changes according to the TDSBL technique

**Table 2.** Performances of the TSBL technique

rank k	rank k	The number of similar terms	sr
rank 2	rank 10	1,3	6,5
rank 2	rank 15	0,5	2,5
rank 2	rank 20	0,3	1,5
rank 2	rank 25	0,1	0,5
rank 10	rank 15	8,1	40,5
rank 10	rank 20	7	35
rank 10	rank 25	8,6	43
rank 15	rank 20	14,2	71
rank 15	rank 25	11	55
rank 20	rank 25	13,1	65,5

**Table 3.** Performances of the DSBL technique

rank k	rank k	The number of similar terms	sr
rank 2	rank 10	6,1	30,5
rank 2	rank 15	5,4	27
rank 2	rank 20	3,7	18,5
rank 2	rank 25	3,3	16,5
rank 10	rank 15	12,1	60,5
rank 10	rank 20	12,1	60,5
rank 10	rank 25	11	55
rank 15	rank 20	12,2	61
rank 15	rank 25	10,5	52,5
rank 20	rank 25	16	80

**Table 4.** Performances of the TDSBL technique

rank k	rank k	The number of similar terms	sr
rank 2	rank 10	2	10
rank 2	rank 15	1,1	5,5
rank 2	rank 20	0	0
rank 2	rank 25	1,1	5,5
rank 10	rank 15	10,2	51
rank 10	rank 20	7,3	36,5
rank 10	rank 25	8,4	42
rank 15	rank 20	15,2	76
rank 15	rank 25	14,1	70,5
rank 20	rank 25	16,4	82

In Table 2, Table 3, and Table 4, the similarities of the keywords according to the TSBL, DSBL and TDSBL techniques were examined according to different rank k values. In the TSBL technique, the best result was observed when k 15 and 20. According to the DSBL technique, good results are observed when k is in the range of 20-25. In the TDSBL technique, which uses both techniques, good results were obtained when k value is between 15 to 20.

**4. CONCLUSION**

Firms/ people have to work with Search Engine Optimization consultants or companies for their web pages. They need this to reach their target audience or increase their audience in the e-commerce environment. As a result, SEO process of websites causes both labor and financial expenses. With this study, in order to eliminate/ reduce these losses, keyword determination processes in the seo transaction were performed automatically. The Latent semantic analysis and keyword

extraction method used in the study will shed light on future studies, especially in areas such as question answering, topic detection, and text classifying.

**DECLARATION OF ETHICAL STANDARDS**

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

**AUTHORS' CONTRIBUTIONS**

**Fahrettin HORASAN:** performed the design and implementation of the research, analysis of the results and writing the article.

**CONFLICT OF INTEREST**

There is no conflict of interest in this study.



## REFERENCES

- [1] Leavitt, Neal, "Network-usage changes push internet traffic to the edge.", *Computer*, 43.10: 13-15 (2010).
- [2] Internet World Stats Internet users of the world: World Internet Usage And Population Statistics, "2019 Mid-Year Estimates", [www.internetworldstats.com](http://www.internetworldstats.com) (2019).
- [3] Wood, Steve, "Web of Deception: Misinformation on the Internet". *New Library World*, (2003).
- [4] Yan, L., Gui, Z., Du, W., & Guo, Q., "An improved PageRank method based on genetic algorithm for web search.", *Procedia Engineering*, 15: 2983-2987, (2011).
- [5] Cui, M., & Hu, S., "Search engine optimization research for website promotion". In 2011 International Conference of Information Technology, *Computer Engineering and Management Sciences*, 4: 100-103, (2011).
- [6] Killoran, J. B., "How to use search engine optimization techniques to increase website visibility.", *IEEE Transactions on professional communication*, 56(1): 50-66, (2013).
- [7] Yalçın, N., & Köse, U., "What is search engine optimization: SEO?.", *Procedia-Social and Behavioral Sciences*, 9: 487-493, (2010).
- [8] Malaga, R. A. , "Worst practices in search engine optimization. *Communications of the ACM*, 51(12): 147-150, (2008).
- [9] "Google's Search Engine Optimization Starter Guide", (2013).
- [10] Mittal, M. K., Kirar, N., & Meena, J. "Implementation of Search Engine Optimization: Through White Hat Techniques.", In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 674-678, (2018).
- [11] Gandour, A., & Regolini, A., "Web site search engine optimization: a case study of Fragfor.net.", *Library Hi Tech News*, (2011).
- [12] Asllani, A., & Lari, A., "Using genetic algorithm for dynamic and multiple criteria web-site optimizations.", *European journal of operational research*, 176(3): 1767-1777, (2007).
- [13] Boyan, J., Freitag, D., & Joachims, T., "A machine learning architecture for optimizing web search engines.", In *AAAI Workshop on Internet Based Information Systems*, (1996).
- [14] Kiritchenko, S., & Jiline, M., Keyword optimization in sponsored search via feature selection. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 122-134, (2008).
- [15] Zimmewicz, M., Kurowski, K., & Węglarz, J., "Scheduling aspects in keyword extraction problem.", *International Transactions in Operational Research*, 25(2): 507-522, (2018).
- [16] Joshi, A., & Motwani, R., "Keyword generation for search engine advertising.", In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, 490-496, (2006).
- [17] Abhishek, V., & Hosanagar, K., "Keyword generation for search engine advertising using semantic similarity between terms.", In *Proceedings of the ninth international conference on Electronic commerce*, 89-94, (2007).
- [18] Sordani, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., & Nie, J. Y., "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion." In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 553-562, (2015).
- [19] Hong, Y., Vaidya, J., Lu, H., & Liu, W. M. Accurate and efficient query clustering via top ranked search results. In *Web Intelligence*, 14(2): 119-138, IOS Press, (2016).
- [20] Süzek, T. Ö. Using latent semantic analysis for automated keyword extraction from large document corpora. *Turkish Journal of Electrical Engineering & Computer Sciences*, 25(3): 1784-1794, (2017).
- [21] Varçın, F., Erbay, H., & Horasan, F., "Latent semantic analysis via truncated ULV decomposition.", In *2016 24th Signal Processing and Communication Application Conference (SIU)*, 1333-1336. IEEE, (2016).
- [22] Horasan, F., Erbay, H., Varçın, F., & Deniz, E. "Alternate Low-Rank Matrix Approximation in Latent Semantic Analysis.", *Scientific Programming*, (2019).
- [23] Martin, D. I., & Berry, M. W., "Mathematical foundations behind latent semantic analysis.", *Handbook of latent semantic analysis*, 35-55, (2007).
- [24] Berry, M. W., & Fierro, R. D., "Low-rank Orthogonal Decompositions for Information Retrieval Applications.", *Numerical linear algebra with applications*, 3(4): 301-327, (1996).
- [25] Duman E., Erbay H., "Latent semantic analysis approach for automatic classification of web pages contents.", *Master Thesis*, (2013).
- [26] Shima, K., Todoriki, M., & Suzuki, A., "SVM-based feature selection of latent semantic features.", *Pattern Recognition Letters*, 25(9): 1051-1057, (2004).
- [27] Uysal, A. K., & Gunal, S. Text classification using genetic algorithm oriented latent semantic features. *Expert Systems with Applications*, 41(13): 5938-5947, (2014).
- [28] Jessup, E. R., & Martin, J. H., "Taking a new look at the latent semantic analysis approach to information retrieval.", *Computational information retrieval*, 121-144, (2001)