Received: 11 January 2019 | Revised: 1 May 2019 | Accepted: 17 May 2019
DOI: 10.1002/cpe.5407

WILEY

SPECIAL ISSUE PAPER

A heuristic approach on metadata recommendation for search engine optimization

Sojung An | **Jason J. Jung**

Department of Computer Engineering, Chung-Ang University, Seoul, South Korea

Correspondence
Jason J. Jung, Department of Computer Engineering, Chung-Ang University, Seoul, South Korea.
Email: j2jung@gmail.com

Funding information
National Research Foundation of Korea, Grant/Award Number: 2017R1A2B4010774

Summary

This study aims to recommend metadata for building a high ranking in Search Engine Result Page (SERP) by considering Search Engine Optimizations (SEO). For online marketing, it is important to place their websites on the top rank in a result of search engines. However, on-page techniques of traditional SEO do not have logical foundation to select metadata. Metadata is an important element to prioritize of websites when search engine indexing for user queries. Thereby, for online marketing, this study proposes a method for recommending metadata, which consists of two steps: i) combining keywords and metadata from high-ranked websites, and ii) evaluating the importance of terms based on semantic relevance. First, terms are selected with influential keywords and metadata by using their frequency and weight. Second, prioritize the terms according to semantic relevance based on a competitive learning model. We evaluated the validity of the proposed method by using three queries in Google. Experimental results demonstrate that it increases traffic of a website, by using terms, which are high-ranked websites and semantic relevance.

KEYWORDS

keyword, Hilltop algorithm, metadata, on-page optimization, search engine optimization

1 | INTRODUCTION

With the web 3.0, it is important to consider search engines (SEs) for successful online marketing. The reason is that SEs find and filter most relevant information matching a user query and display the information to users.[1] Customers may visit websites that already they know or advertise, but mostly they use SEs to effectively get information which they want. On the other hand, most companies have no marketing strategies and design their web pages by not considering a ranking system of SEs. Thousands of pages are created a day, and it is hard to find their website. According to HubSpot, more than 70% marketers fail with conducting online marketing. For the reason, it is significant to appear as a top search result, SEO is the technology for placing website at the top of search results. Marketing using ICTs, especially SEO, has completely transformed the marketing strategies for business. SEO is regarded as one of effective marketing method to make web pages be located to the top of the result. Regarding BCAMA's marketing survey about SEO proficiency in 2017, 96% of SEO experts think that SEO affects their benefits strategies.[2,3] Not surprisingly, researchers in the field of information retrieval have actively studied to solve complex challenge related to SEO.

Hilltop is one of the popular ranking algorithms in SEs, it is ranked by the match between the query and terms in metadata. Thereby, it is crucial issue to optimize metadata, which is an SEO on-page technique, since SE stores meta information.[4,5] Thousands of web pages are generated and published in SE every day, it is difficult to make a visible and accessible website for users. Furthermore, general administrator, especially managers of individuals web pages, does not usually understand marketing strategy about SEO. In this regard, to make higher visibility for successful online marketing is not an easy work. To achieve a high ranking of websites, existing SEO techniques mainly use selecting appropriate keywords, optimizing title, enriching content, and so on.[6] In particular, leveraging proper metadata is one of the representative techniques that can optimize web pages.[7] The effectiveness of metadata for high visibility of websites has already been proved.[8,9] However, utilization of metadata that is not fit to target content comes at the expense of quality. As a simple example, for a query that is ‘‘women shirt’’, metadata is written ‘‘Check

This paper is significantly revised from an earlier version presented at the 2019 International Conference on Information and Communication Technology and Digital Convergence Business (ICIDB 2019) held in Seoul, the Republic of Korea on January 2019.

out our trendy collection of casual shirts" on a website. In another website, metadata is written "Browse women"s tops perfect for lounging around, running errands, or work. In this case, you can recognize that the former site is more related to the query than the latter.[10] Likewise, using appropriate metadata can attract public attention, and the effectiveness of the metadata can change with queries.[11,12] For this reason, we need to consider which terms are useful in online marketing.

Thereby, we propose a recommender system for optimizing metadata by using the metadata and keywords. For an effective decision of metadata of websites, it consists of two methods as follows; (i) extracting keywords from metadata and content of websites ranked at high in SEs, and (ii) classifying the keywords with semantic relevance generated through Self Organizing Map) (SEO).

To verify the efficacy of the proposed methods, we have conducted experiments with the following research questions.

- RQ 1. Is it important to use terms that are semantic relevance with of a query?
- RQ 2. Is it effective to improve website ranking by using keywords and metadata?
- RQ 3. Is it helpful for placing the first SERP by selecting terms to belong in top ranking?

The outline of this paper is as follows. In Section 2, we take a briefly about the background and related works of this research area. In Section 3, we propose the solution of support system for the metadata recommendation problem, and Section 4 shows the data is how valuable in marketing, briefly compares all of the models in terms of both statistical performance and total profit. Finally, Section 5 draws a conclusion of this study and provides further research direction.

2 | BACKGROUND AND RELATED WORK

In this section, we introduce the fundamental background and definitions related to the SEO, especially on-page optimization. The on-page optimization is a technique belonging to the SEO, and the proposed heuristic method in this paper is a technique for on-page optimization.

2.1 | Search engine optimization

SEO refers to the task of making a websites appear above the search results by designing a website according to rank algorithms of SE. SEO improves rankings of web sites by optimizing SE.[13] SEOs are well-known and influential online marketing techniques. According to Econsultancy's report conducted with 1500 advertisers in 2010, approximately 90% of participants have used SEO techniques for their online marketing.[8] SEO technology has grown into a multi-billion dollar business over the last few years, and it has been obviously shown for years that SEO affects the decision-making of consumers.[9] It has been obviously shown for years that SEO affects consumers' decision making.[9] SEs are generally composed of two algorithms. One is Page Rank (PR) that ranks web sites based on quantity and quality of in- and out-coming links. Another one is Hilltop algorithm that ranked according to relevance with search query or keywords. They judge whether web pages are proper target or not by analyzing topics of the pages.[13,14] Google's SE is mostly used and combines PR and Hilltop to rank websites. Based on this, SEO can be largely divided into on-page and off-page techniques.

The on-page optimization (SEO (Search Engine Optimization), white-hat of SEM (Search Engine Marketing), Decision-Making Fuzzy Rules, SEA (Search Engine Advertising), and CA (Contextual Advertising)) mean a method of optimizing the search engine to read the contents of the website. In general, on-page optimization is limited to the selection of terms. The specific examples of this technique include selecting terms of title, metadata, title of image, and anchor text. It translates influential factors (e.g., keywords appearing in title tag, meta tag, headlines, URL tag, and body text) into a score for SEO.[14] For the hilltop algorithm, the association with queries is evaluated under the expert score, which consists of a levelscore and a fullnessfactor. According to Krishna, a key phrase is a piece of text that qualifies one or more URLs in the page. Levelscore checks whether the terms of the query belong to the key phase. If a term contains in the phase, it is given 16 scores for title phrases, 6 for headings, and 1 for anchor text. Fullnessfactor is computed how many query belongs among the terms in key pieces and multiplied by levelscope.[15] Hence, the topics and keywords that raise ranking of websites are important to increase high visibility of web pages.[16] In addition, to be contained a reliable link in anchor text and URL tag is an effective improving the ranking.

On the other hand, the off-page optimization is to control an external factor that influences the ranking of the site apart from the website.[13] This is an algorithm that evaluates reliability of websites based on links.[17-20] Many studies have been done in off-page optimization for the public, it is a difficult technology by considering the reliability of one's website (i.e., PR score). On the other hand, on-page optimization technology is easy but powerful.

2.2 | On-page optimization

As shown in Table 1, many studies have shown that metadata optimization is a good way to improve site and content visibility on the web. Hoque et al (2018) proved that ranking is advanced 17.27% with On-page SEO, especially, search result is improved 20.45% by length, unique, and terms.[21] It is proved that the use of metadata improves the search traffic.[22-24] Zhang and Dimitroff[22] found that visibility of website can

TABLE 1 Basic features to on-page optimization area

No.	Ref. num.	Summary
1	26	Creating an XML Sitemap to improve visibility for a national scholarly open access information website in the field of STEM.
2	27	Build meta description for websites which are missing it.
3	28	Using top anchor text from backlink analysis to improve page titles of existing web pages.
4	29	Research and analysis of search engine optimization factors based on reverse engineeing.
5	30	Text summarization techniques applied to Contextual Advertising (CA hereinafter).
6	31	Estimate Google search engine's ranking function from a SEO perspective.

be improved by increasing the frequency of terms in the title and in the body text. The website can be further exposed to the SE by selecting the appropriate words for a variety of tags such as URLs and titles. Park[23] designed XML Sitemap to create metadata in academic journals so that SEs can easily find pages about Korea Science. It showed that applying metadata can improve accessibility of content in an open access scholarly information system. Matosevic[24,25] has improved the title of the existing website based on the higher-priority anchor text in the back link analysis, which could increase web traffic. It also proposed a methodology for generating meta descriptions based on user queries based on manually generated terms from SEO experts for Web sites with missing metadata information. Luh et al[26] showed the efficacy of metadata optimization and proposed to generate metadata according to the query by using Latent Semantic Analysis (LSA). Armano et al[27] proposed text summary techniques to suit the user query language by extracting phrases (Titles (T), First Paragraph (FP), and so on) from the website as an online marketing strategy. Using metadata is relevant with user queries play an important role in increasing the visibility of websites. To expand into a decision-making system for online marketing, metadata need to find optimal standard for increasing SEO quality.

Moreover, effectiveness of utilizing tags to improve visibility of websites in Google's SE. Recently, many marketing companies have sprung up to support on-page optimization such as Straight North, Ignite Visibility, Boostability, and Twinword. The companies emphasize the importance of selecting terms for titles, description, and so on. Each tag is a significant factor for optimizing SEs. The closer the term is to the beginning of the title tag, the more weight it has with SEs. The selection of terms has a positive impact on online marketing, in particular, there has many tools that recommend terms as a way to improve SEO quality.[28-30] SEO company, Straight North, claims that metadata are extremely useful in getting people to click on your links.

This study has something in common with previous research in that it recommends metadata and terms for on-page optimization. However, there are limitations to recommend terms only evaluated as competitiveness or search volume. In general, when the companies or researchers recommend terms to optimize websites, they rank based on how many they are searched and how competitive they are based on the user query (target).

3 | AUTOMATIC METADATA GENERATION

In this section, we propose a method for recommendation system to select the metadata of websites. Metadata is important to represent information of websites to SE and it can be applied for online marketing (SEO). In this regard, by analyzing contents on the web pages, this study is focusing on two kinds of heuristics (i) terms located the first SERP and (ii) semantic relevance with a user query. Figure 1 shows the conceptual design of the proposed method, which consists of five steps as follows.

- **Keywords and metadata extraction (Steps 1, 2, and 3)**: First, we extract terms that consist of keywords and metadata from HTML of websites. The process is involved in transforming the HTML documents by processes such as stop-word removal and stemming and generating body text and metadata. Next, the keywords are selected through KeyGraph[31] by reason such as an enormous amount of body text. For metadata recommendation, we combine keywords and metadata according to the third curriculum about metadata recommendation. The combination based on the weight of keywords and metadata is mixed, normalizing the weights of terms.
- **Semantic relevance measurement (Step 4)**: In this process, the terms are evaluated by semantic relevance with normalized frequency. To find association of terms, we use Bag of Words (BoW) model and Term Frequency-Inverse Document Frequency (TF-IDF) and generate vectors from the terms. On this basis, we use the correlation coefficient to compare how well fit user queries With distributions.
- **Metadata recommendation (Step 5)**: To achieve top ranking in the websites, the terms are chosen as metadata candidates by evaluation under two course. First, each term is prioritized by a ranking of the website a term appears. Second, they are evaluated with semantic relevance to the user query. By using the above two options, we recommend proper terms for selecting metadata.

3.1 | Selecting terms by using keywords and metadata

According to Hubspot,* 75% of online shoppers never scroll past the first SERP. Therefore, it is effective marketing strategies that a website occurs in the first SERP. According to HillTop algorithm used on SEs, to increase a ranking of websites, it is important to greatly boost an organization's

*https://www.hubspot.com/marketing-statistics#SEO

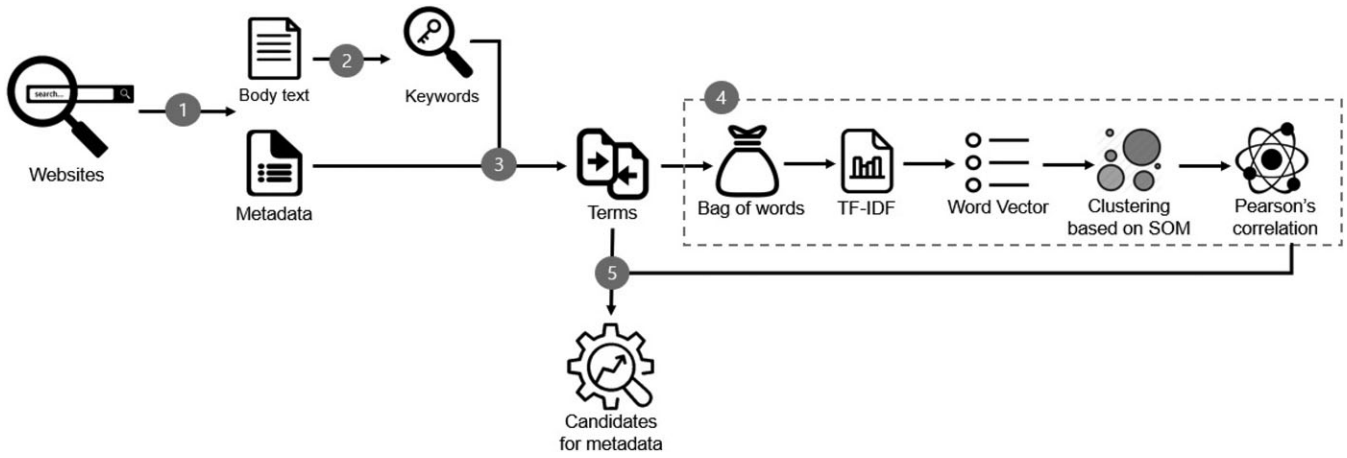

FIGURE 1 Conceptual design of the proposed for metadata recommendation. From a user query, an algorithm generates metadata candidates in the websites for optimizing a website. The structure is as follows: (1) Parsing from Websites to HTML, (2) Extracting keywords, (3) Combining between keywords and metadata into terms, (4) measuring semantic relevance of terms on the basis of SERP, (5) Computing importance of terms by SERP and semantic relevance. The method can be increased website traffic by eliciting users visit that site

TABLE 2 A example of metadata extraction according to meta tags

No.	Tag	Attribute Value	Example
1	meta	title	Women's Tops - Shirts & Blouses for Women
2	meta	description	Shop our wide selection of women's tops for every occasion, including sweaters, women's t shirts and blouses in dressy or casual styles!
3	meta	keyword	t-shirt, women, sleeve

visibility. Especially, the algorithm focuses on semantic relevance by matching the metadata and body text. For this reason, we use all the terms that are consisted of metadata and keywords. In this paper, the keywords mean topic terms of websites that define what your body text is about.[32] The keywords are terms representing the websites, it can be a criterion how it interrelated with the query. The metadata describes a resource for purposes such as discovery and identification of websites, including elements like title tag, description tag, and keyword tag. Metadata, first of all, need to be relevant user's query and are selected terms that are used frequently by users. Hence, it is significant to use combine between the keywords and the metadata; in this section, we consider a method that keywords and metadata coalesce into terms for selecting novel metadata.

To select the keywords, we exploit KeyGraph that extracts relevant keywords using the graph-based approach. The technique is sampling main point in a document when processing large noisy data collections based on the segmentation of a graph, representing the co-occurrence between terms in the documents from websites, and it is working as described below. Firstly, with body text collected in web pages, keywords are selected based on their frequency and calculated. A predetermined amount of terms are selected based on their frequency (high-frequency set, HF) and added as the initial nodes of the graph. The association strength between each of these terms is calculated using co-occurrence frequency, and then each of these new terms is then linked to every cluster using the strongest scoring edge amongst the possible ones called columns. it is associations between foundations and roofs that are used for extracting keywords. As the following setting, it will be made that this node displays the contents of the clusters after the pruning terms.

Next, we collect the metadata belong in tags as noted above. Metadata is written by administrators of the websites with their own ideas; it is used when SE matches with user query searched. In this regard, we collect it by not KeyGraph, but term frequency. Table 2 shows an example of extracting each metadata elements. Metadata are performed by pre-processing methods such as stemming, stop-words filtering, and term weight about frequency.

As in the content, three meta tags are meaningful among the meta tags that can match with SEs. Title tag specifies the title of a web page, description tag provides concise summaries of a website, and keyword tag consists of topic or terms that directly pertain to the content of website. All the tags is invisible to visitors but visible to SEs, it is significant for SEO.

For combining, the weight of terms is normalized their weights from 0 to 1, the reason is that weight of terms between keywords and metadata was measured in different ways. Those weight of terms can be described as the importance of the terms, and metadata candidates are recommended to largest weight. It is an integration method as:

$$\mathcal{T}_l = \mathcal{K}_i \mathcal{W}_t + \mathcal{M}_j(1 - \mathcal{W}_t), \tag{1}$$

where $\mathcal{T}_l$ denotes the terms, $\mathcal{W}_t$ indicates the weight of the combination, $\mathcal{K}_i$ refers to centrality of keywords, and $\mathcal{M}_j$ indicates frequency of metadata.

3.2 | Selecting terms based on semantic relevance

In this work, we measure the semantic relevance. With regard to the search algorithm, for locating first SERP, it is another issue to select terms that are related to query. Only Raising a high SERP ranking, it can lead to better brand development. Even with the same with a query, semantic relevance of each website can be different. However, no matter how a website use terms better, if the terms are little relevance to the query, a ranking of the website has no choice but to fall. As a term can have multiple meanings, and it can be used for a variety of reasons, it is a significant component in determining the ranking of websites. This is why we should consider semantic relevance, hence, we propose the following methodology for measuring semantic relevance.

In this regard, to evaluate semantic relevance between terms and a user query, terms are clustered based on SOM with their each feature vector. For the first step to cluster the terms, we composed a feature vector of all the terms by applying the BoW model and TF-IDF. The feature vector is a measured term whether existing or not in all websites with calculating the TF-IDF about the stem of terms. A term vector is presented in the following form:

$$t_l = (\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \ldots, \mathcal{D}_\varepsilon), \tag{2}$$

where t_n refers to term vectors, and $\mathcal{D}_\varepsilon$ indicates websites. For the evaluation of semantic relevance between terms and queries, if each term belongs to the websites, the TF-IDF value of the terms is put into the $\mathcal{D}_\varepsilon$. If not, it is set to 0.

Specifically, for clustering the terms, we apply SOM algorithm, an architecture suggested for artificial neural networks, to reduce input data to a low dimensional space (usually two dimensional).[33] This algorithm can cluster of the input grid without loss of useful information. Moreover, it is able to quickly learn a high dimensional vector.[30,34] To train the vector, Kohenen's SOM has been proposed for generally used classification. In this regard, it can make a challenging task considerably easier. The following simple procedure was implemented to train with SOM. At first, the algorithm first determines the size of a map as follows: $\mathcal{N}_i = 3\sqrt{\mathcal{N}_m}$, which $\mathcal{N}_m$ is the number of input vectors. Second, the node which is the nearest to the input vector is determined as a winning node. The winning node is calculated by:

$$\mu = \text{argmin}_{\alpha \in [0,I]} \|\mathcal{V}_\alpha - \mathcal{V}_t\|, \tag{3}$$

where μ denotes the winning node, $\mathcal{V}_t$ refers to the input data vector which is targeted by random, and *alpha* indicates a number of input vectors. Finally, weights of the winning node and neighboring nodes are updated, until sequence motifs result in negligible. Each node is associated with a weight vector, which is a position in the input space, the weights are updated as:

$$\mathcal{W}_\beta(\delta + 1) = \mathcal{W}_\beta(\delta) + h_{\mu,\beta,\delta} \cdot \gamma(\delta) \cdot (\mathcal{V}_t - \mathcal{W}_\beta(\delta)), \tag{4}$$

where β indicates the index of the node in map and μ denotes the index of the best matching unit (BMU). The $h_{\mu,\beta,\delta}$ is to classify neighborhoods of the BMU by:

$$h_{\mu,\beta,\delta} = exp\left(\frac{-\|r_\beta - r_\mu\|^2}{2\mathcal{V}_\delta^2}\right). \tag{5}$$

According to the proposed approach, each clustered group has a semantic relationship. If a lot of aspect of semantic relationships changes depending on the distribution of websites, it is a meaningful term. In particular, if the ranking of a website decreases and the frequency terms are less used, then the terms are likely to be often used to users using that query. On the other hand, terms that have a negative correlation between rankings of websites and frequency of the terms would be unfavorable terms for raising the ranking of websites. Hence, terms are given weight, terms that have high weight are recommended for metadata.

3.3 | Recommending metadata based on heuristics

In this section, we propose a metadata recommendation system based on two features above. The terms are weighted on each important scale, and, at the top, any number of the weights is recommended as metadata. There are two main types of scale assessments. The first is the ranking of the website. The higher the web site's ranking is, the more frequently it is used by users, so it can be ranked by increasing the website's traffic. In this regard, terms combined with metadata and keywords are weighted according to the ranking of the website. When the ranking of a term is high, a weight for the term ($\mathcal{W}_r(l)$) is large, and vice versa. When the number of sites is $\mathcal{C}$, and a set of ranking is $\mathcal{R}$, weights of the term are multiplied as follows:

$$\mathcal{W}_r(l) = \frac{\mathcal{C} - (\mathcal{R}(l) - 1)}{\mathcal{C}}. \tag{6}$$

TABLE 3 Features of dataset

	Sites			Keywords			Metadata		
Query	Top	Middle	Low	Top	Middle	Low	Top	Middle	Low
Women shirt	35	41	54	69	77	99	69	61	88
New bag	28	33	59	99	69	99	51	77	77
London restaurant	49	49	67	129	139	96	129	89	63

All numbers are counted according to each option. One web page is a collection of websites, on average, there are 14 sites on a web page. Data consist of Top, Middle, and Low based on ranking. Top means data extracted from the 1st page to the 3rd page, Middle is from 4th to 6th, and Low is from 7th to 10th.

Second, terms are evaluated through how much semantic relevance it is. To use Pearson's correlation coefficient, a process was explained as causal association to indicate between the two variables. If hit rate and ranking are exactly the same, the weight of relevance is set to 1, or the opposite direction, weight of relevance is set to 0. If it is exactly the same in the opposite direction, it is set to −1. If semantic relevance of a term is high, a weight of term will be high. A correlation coefficients between the normalized frequency of terms and SERP is calculated as:

$$r_{\mathcal{HP}}(l) = \frac{\sum(\mathcal{H}_l - \overline{\mathcal{H}})\sum(\mathcal{P}_l - \overline{\mathcal{P}})}{S_{\mathcal{H}} S_{\mathcal{P}}}, \tag{7}$$

where $S_{\mathcal{H}}$ denotes standard deviation for hit rates, $S_{\mathcal{P}}$ indicates standard deviation for pages, and $r_{\mathcal{HP}}(l)$ refers to the correlation coefficient between the normalized frequency of terms and SERP. It is a term that should be targeted if semantic relevance grows less and less. Conversely, if a term is a positive correlation, it is better not to use it. For this reason, the correlation coefficients are changed to their sign and normalized between 0 and 1 for evaluation as follows:

$$r_{\tilde{\mathcal{H}}\mathcal{P}}(l) = \frac{-r_{\mathcal{HP}}(l) - \min(-r_{\mathcal{HP}})}{\max(-r_{\mathcal{HP}}) - \min(-r_{\mathcal{HP}})}. \tag{8}$$

To consider the two option, for metadata, we proposed the terms which are placed to the upper SERP and are semantic relevance. The terms are measured by:

$$\mathcal{T}_l := \mathcal{T}_l \cdot \mathcal{W}_r(l) \cdot r_{\tilde{\mathcal{H}}\mathcal{P}}(l). \tag{9}$$

Based on these calculated results, terms can be chosen in high order as desired values. The terms are applicable to tags such as title, description, and keyword for optimizing websites. If terms can be predicted to user queries through the two additional conditions, the expert score can be high. If a website uses the user query term, it can be weighted from 1 point to 16 points, depending on its location. It means that to select terms is important, and user queries and the terms recommended under the two assessments proposed in this section (which are in the top ranking and semantic relevance higher to the expecting user queries) are same together in some way.

4 | EVALUATION

In this section, we validate the 3 research questions. First, we compare hit rates of the combination between keywords and metadata to find out which combining weights are most efficient. Next, we compare hit rates[35] between top ranking terms and low ranking terms. Finally, we measure semantic relevance through hit rates of groups clustered based on SOM.

The data are organized to 3 queries which are described as follows: i) Women shirt, ii) New bag, and iii) London restaurant. It is searched by an average of 2.2 terms and is combined with common terms used in marketing statistically. Data is downloaded through the Google SE that is the most used, and composed of 17860 rates of hit (415 websites, 131 terms) from 1 and 10 pages. Table 3 shows the number of data for each query.

4.1 | Efficiency of combining keywords and metadata

As we mentioned, the performance is evaluated by hit rate of web pages, with comparing terms belonging to groups between top, middle, and low ranking. In this study, if the term exists on the website, the hit is set to 1, if not, hit is set 0. For example, if it exists on seven out of 10 sites, the hit rate is 0.7. To find the optimal $\mathcal{W}_t$, we evaluated the hit rate of terms adjusting $\mathcal{W}_t$ from 0 to 1. For the experiment, we select the top 10 terms in combining Keywords and Metadata. Figure 2 depicts result each query by 5 conditions. The x-axis and y-axis represent pages and hit rates, respectively.

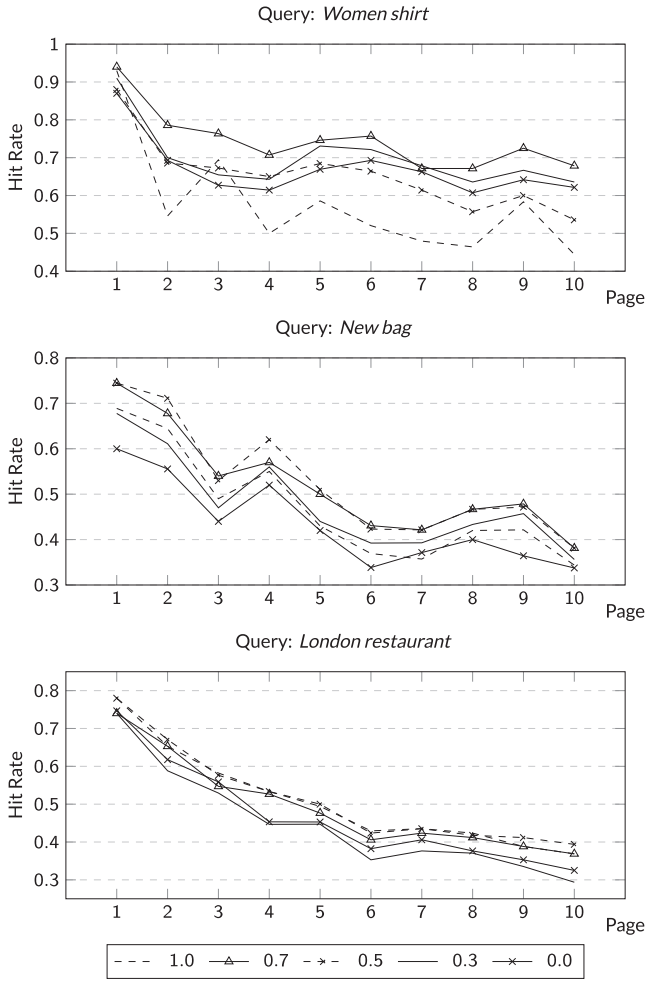

FIGURE 2 Hit Rates of Terms' combination for Websites on each Page according to Weighting Factor ($\mathcal{W}$)

$\mathcal{W}_t$	Top	Middle	Low
0	0.66	0.47	0.41
0.3	**0.71**	**0.57**	**0.51**
0.5	0.69	0.56	0.48
0.7	0.65	0.53	0.47
1	0.64	0.52	0.46

TABLE 4 Average hit rates in the three queries according to combination weight

Figure 2 shows average of terms' hit rate by adjusting $\mathcal{W}_t$. As a result, when query is "Women shirt", hit rate in $\mathcal{W}_t$ is 0.3 was the highest. It means terms are outperformed when combining weight is 0.3. In the other hand, hit rate of the other queries when $\mathcal{W}_t$ is 0.5 are the highest. The results also show that the lower the ranking, the lower the hit rate. For more details, Table 4 demonstrates the result by averaging based on ranking such as Top, Middle, and Low to identify the most efficient combination. According to the table, an approach that $\mathcal{W}_t$ is 0.3 was the highest in all the ranking. Combining between keywords and metadata is a key to increase web traffic, with the result of outperforming when the weight was 0.3 and 0.5.

4.2 | Efficiency of concentrating on high-ranked pages

We compare hit rates between the top ranking and the low ranking terms for validation whether terms selection is important to rank of websites. The hit rate is the same as the previous experimental method. Figure 3 depicts the hit rates for all the terms.

Figure 3 shows that the terms in the top ranking have significantly better performance than those in the middle and low ranking. We verified that the terms of the top ranking are performed well. However, for the "women shirt" query, the hit rate of the top-ranked terms was high from pages 1 to 10, but for the other two queries, the hit rate of the low-ranked terms tends higher than the hit rate of the top-ranked after page 5.

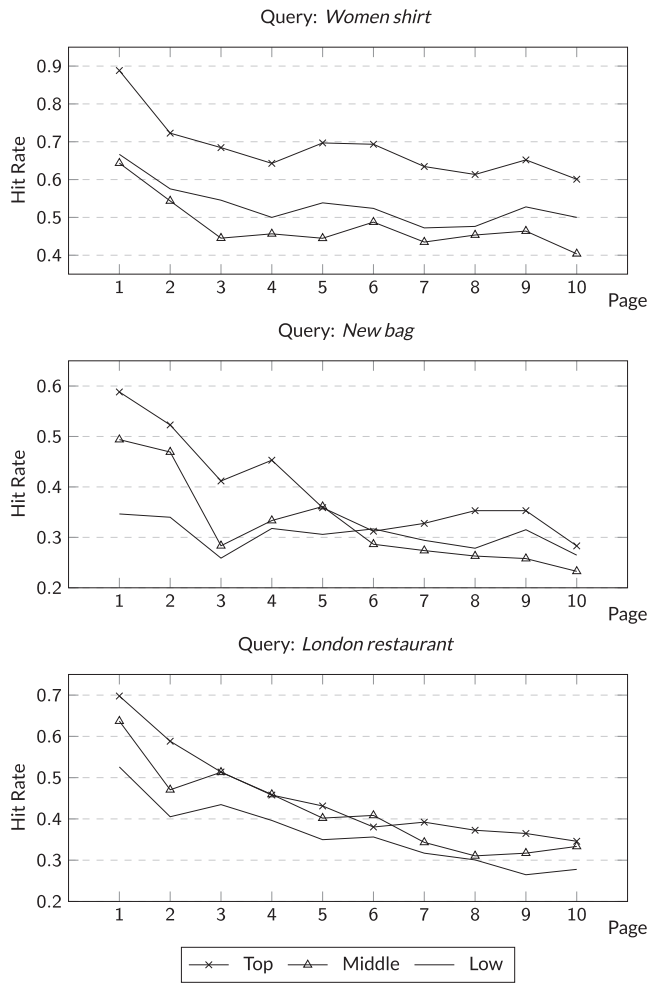

FIGURE 3 Hit Rates on Each Page according to Ranking where the Terms Belong to

TABLE 5 Average hit rates in the three queries based on ranking groups

Ranking of Terms	Top	Middle	Low
Top	**0.62**	**0.49**	**0.44**
Middle	0.49	0.38	0.31
Low	0.46	0.40	0.36

The reason for this result can be explained by semantic associations. The terms used on the first page and the terms used on 10th pages are different in meaning, and we will discuss it more specifically in the next experiment. Table 5 represents the average of hit rate grouping every three pages. According to Table 5, the terms that are used in the top three pages outperformed the other groups with a hit rate of 0.62.

4.3 | Efficiency of semantic relevance of terms

In this section, to validate the patterns of semantic relevance change, we cluster terms according to semantic relevance by using SOM. The vectors of the terms are input vectors of SOM, and it is calculated by the SOM algorithm. In this paper, the training cycle is set to 1000 times and an initial value of learning parameter learning rate is set to 1.0. The terms are grouped into 6 term vectors based on the appropriate map size. The group of semantic relevance is clearly different on each page, and to evaluate definite variation, we normalized. Figure 4 depicts the result of changing aspects about semantic relevance normalized on each page. The $\mathcal{SG}$ group represents a list of groups clustered based on SOM with word embedded values, which can be defined as semantic groups.

With regard to Figure 4, on average, each standard deviation is about 6.47, 5.94, and 9.12. As expected, some groups of hit rate was a gradual rose or fell, however, hit rates of others plummet. It shows that there are semantic changes and have associations depending on the page. The result brings about too rapidly decreased, increased, or plummeted in case of clustering less such as 15 terms. Moreover, the results show that the terms of the $\mathcal{SG}0$ group are highly distributed with a negative correlation in case of "Women shirt". In this regard, the terms of $\mathcal{SG}0$ group are most importance and are measured the highest.

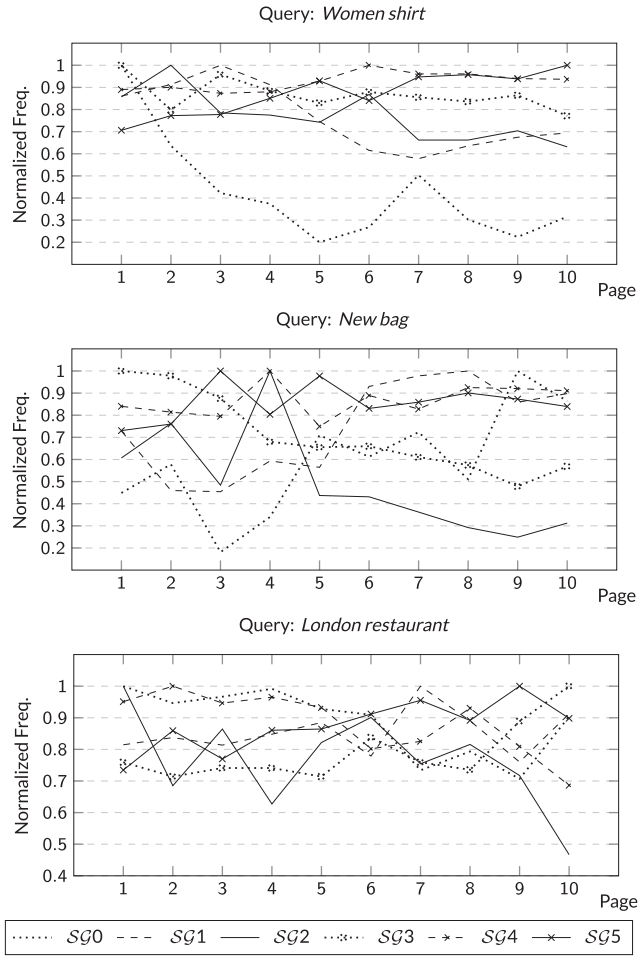

FIGURE 4 Results of the experiment for evaluating semantic relevance on each page

5 | CONCLUSIONS

In this study, we proposed metadata recommender system to optimize websites, the process is to be at the top in SERP. As the online market grows rapidly, online marketing has evolved as an important tool in business for communication with customers. However, the inappropriate selection terms reduce the quality of the SEO effect. In particular, in this paper, we take into account selecting the terms which are important to increase visibility the web page in Google. Hence, two main methods were adopted to enhance visibility. Therefore, we suggest the approach that can choose suitable terms by using artificial intelligence techniques. Through three experiments, we found out that the keywords and metadata that belong to the top ranking will be used, and terms have to relate semantic relevance with the query. Additionally, when metadata and keywords are combined, it has received the increasing amount of its traffic. Future research needs to prove how the traffic increase by using extracted terms based on this system. Furthermore, in this study, there was a limitation to crawl the image and video, since we only focus on HTML. In future studies, this issue will be considered how to download non-text data.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A2B4010774).

ORCID

Jason J. Jung https://orcid.org/0000-0003-0050-7445

REFERENCES

1. Patil SP, Pawar BV, Patil AS. Search engine optimization: a study. *Res J Comput Inf Tech Sci*. 2013;1(1):10-13.
2. VanBoskirk S, Overby CS, Takvorian S. US interactive marketing forecast, 2011 to 2016. Cambridge, MA: Forrester Research; 2011.
3. Yuniarthe Y. Application of artificial intelligence (AI) in search engine optimization (SEO) . In: Proceedings of the 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT); 2017; Denpasar, Indonesia.
4. Gudivada VN, Rao D, Paris J. Understanding search-engine optimization. *Computer*. 2015;48(10):43-52. https://doi.org/10.1109/MC.2015.297

5. Chianese A, Marulli F, Piccialli F, Benedusi P, Jung JE. An associative engines based approach supporting collaborative analytics in the Internet of cultural things. *Futur Gener Comput Syst*. 2017;66:187-198. https://doi.org/10.1016/j.future.2016.04.015
6. Lee O, Jung JJ. Modeling affective character network for story analytics. *Futur Gener Comput Syst*. 2019;92:458-478. https://doi.org/10.1016/j.future.2018.01.030
7. Killoran JB. How to use search engine optimization techniques to increase website visibility. *IEEE Trans Prof Commun*. 2013;56(1):50-66. https://doi.org/10.1109/TPC.2012.2237255
8. Econsultancy State. *Search Engine Marketing Report 2010 (in association With SEMPO)*. New York, NY. Econsultancy; 2011.
9. Berman R, Katona Z. The role of search engine optimization in search marketing. *Marketing Science*. 2013;32(4):644-651. https://doi.org/10.1287/mksc.2013.0783
10. Cuomo S, Michele PD, Piccialli F, Galletti A, Jung JE. IoT-based collaborative reputation system for associating visitors and artworks in a cultural scenario. *Expert Syst Appl*. 2017;79:101-111. https://doi.org/10.1016/j.eswa.2017.02.034
11. Bui KHN, Jung JJ. Computational negotiation-based edge analytics for smart objects. *Information Sciences*. 2019;480:222-236. https://doi.org/10.1016/j.ins.2018.12.046
12. Hong M, Jung JJ. Multi-sided recommendation based on social tensor factorization. *Information Sciences*. 2018;447:140-156. https://doi.org/10.1016/j.ins.2018.03.019
13. Sahu N, Chhabra R. Review on search engine optimization. *J Netw Commun Emerg Technol*. 2016;6(6):19-21.
14. Van Hulle MM. Self-organizing maps. In: *Handbook of Natural Computing*. Berlin, Germany: Springer-Verlag Berlin Heidelberg; 2012:585-622. https://doi.org/10.1007/978-3-540-92910-9
15. Bharat K, Mihaila A. A search engine based on expert documents. In: Proceedings of the 9th International WWW Conference; 2000; Amsterdam, The Netherlands.
16. Zhu C, Wu G. Research and analysis of search engine optimization factors based on reverse engineering. Paper presented at: 2011 Third International Conference on Multimedia Information Networking and Security; 2011; Shanghai, China. https://doi.org/10.1109/MINES.2011.99
17. Ohsaka N, Sonobe T, Kakimura N, Fukunaga T, Fujita S, Kawarabayashi K. Boosting PageRank scores by optimizing internal link structure. In: *Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3-6, 2018, Proceedings, Part I*. Cham, Switzerland: Springer Nature Switzerland AG; 2018:424-439. https://doi.org/10.1007/978-3-319-98809-2_26
18. Datta P, Vaidhehi V. Influencing the PageRank using link analysis in SEO. *Int J Appl Eng Res*. 2017;12(24):15122-15128.
19. Dye K. Website abuse for search engine optimization. *Network Security*. 2008;3:4-6. https://doi.org/10.1016/S1353-4858(08)70028-X
20. Multazam M, Purnama B. Influence of classified ad on Google page rank and number of visitors. *J Theor Appl Inf Technol*. 2015;81:174-181.
21. Hoque M, Alsadoon A, Maag A, Prasad PWC, Elchouemi A. Comprehensive search engine optimization model for commercial websites: surgeon's website in Sydney. *J Softw*. 2018;13(1):43-56.
22. Zhang J, Dimitroff A. The impact of webpage content characteristics on webpage visibility in search engine results (part I). *Inf Process Manag*. 2005;41(3):665-690. https://doi.org/10.1016/j.ipm.2003.12.001
23. Park M. SEO for an open access scholarly information system to improve user experience. *Inf Discov Deliv*. 2018;46(2):77-82.
24. Matošević G. Text summarization techniques for meta description generation in process of search engine optimization. In: *Artificial Intelligence and Algorithms in Intelligent Systems Proceedings of 7th Computer Science On-line Conference 2018, Volume 2*. Cham, Switzerland: Springer International Publishing AG; 2018:165-173. https://doi.org/10.1007/978-3-319-91189-2_17
25. Matošević G. Using anchor text to improve web page title in process of search engine optimization. Paper presented at: Central European Conference on Information and Intelligent Systems; 2015; Varaždin, Croatia.
26. Luh CJ, Yang SA, Huang TLD. Estimating Google's search engine ranking function from a search engine optimization perspective. *Online Inf Rev*. 2016;40(2):239-255. https://doi.org/10.1108/OIR-04-2015-0112
27. Armano G, Giuliani A, Vargiu E. Experimenting text summarization techniques for contextual advertising. Paper presented at: Second Italian Workshop on Information Retrieval (IIR); 2011; Milan, Italy.
28. Yalçin N, Köse K. What is search engine optimization: SEO? *Procedia Soc Behav Sci*. 2010;9:487-493. https://doi.org/10.1016/j.sbspro.2010.12.185
29. Cui M, Hu S. Search engine optimization research for website promotion. In: Proceedings of the 2011 International Conference of Information Technology, Computer Engineering and Management Sciences - Volume 4; 2011; Nanjing, China. https://doi.org/10.1109/ICM.2011.308
30. Kent P. *Search Engine Optimization For Dummies*. Hoboken, NJ: John Wiley and Sons; 2012.
31. Ohsawa Y, Benson N, Yachida M. KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings IEEE International Forum on Research and Technology Advances in Digital libraries-ADL; 1998; Santa Barbara, CA. https://doi.org/10.1109/ADL.1998.670375
32. Hoffman MD, Blei DM, Bach FR. Online learning for latent Dirichlet allocation. In: Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010, Advances in Neural Information Processing Systems; 2010; Vancouver, Canada.
33. Arribas-Bel D, Nijkamp P, Scholten H. Multidimensional urban sprawl in Europe: a self-organizing map approach. *Comput Environ Urban Syst*. 2011;35(4):263-275.
34. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. *IEEE Trans Neural Netw*. 2000;11(3):586-600. https://doi.org/10.1109/72.846731
35. Arlitsch K, O'Brien PS. Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar. *Libr Hi Tech*. 2012;30(1):60-81. https://doi.org/10.1108/07378831211213210

How to cite this article: An S, Jung JJ. A heuristic approach on metadata recommendation for search engine optimization. *Concurrency Computat Pract Exper*. 2019;e5407. https://doi.org/10.1002/cpe.5407