# 검색엔진 최적화를 위한 GAN 기반 웹사이트 메타데이터 자동 생성

안소정[1] · 이오준[1] · 이정현[1] · 정재은[1*] · 용환성[2]

[1]중앙대학교 · [2]리얼리티랩

# GAN-based Automated Generation of Web Page Metadata for Search Engine Optimization

Sojung An[1] · O-jun Lee[1] · Jung-Hyeon Lee[1] · Jason J. Jung[1*] · Hwan-Sung Yong[2]

[1]Department of Computer Engineering, Chung-Ang University · [2]Reality Lab

E-mail : {thwjd1258*, concerto34*, nmoonma7*, j3ung}@cau.ac.kr / zeus@realitylab.co.kr

## 요 약

본 논문에서는 검색엔진 최적화(SEO; Search Engine Optimization)에 인공지능 기법을 접목하여, 자동화된 SEO 도구 설계 및 구현을 목표로 한다. 기존의 SEO 온-페이지(On-page) 최적화 기법들은 웹페이지 관리자들의 경험적 지식에 의존하는 한계점을 보이고 있다. 이는 SEO 성능에 영향을 끼칠 뿐 아니라, 웹페이지 관리자들에게도 SEO 도입의 장벽으로 작용한다. 따라서, 위 문제를 해결하기 위하여 메타데이터의 효과적인 구성을 위해 다음과 같은 3단계의 접근법을 제안하고자 한다. i) 상위 랭킹 웹사이트들의 메타데이터를 추출한다. ii) 어텐션 메커니즘에 기반한 LSTM(Long Short Term Memory)을 이용하여 사용자 질의어와의 관련성 높은 메타데이터를 생성한다. iii) GAN(Generative Adversarial Network) 모델을 통하여 학습함으로써 전반적으로 성능을 높여주는 기법을 제안한다. 본 연구결과는 기업의 온라인 마케팅 프로세스를 평가하고 개선하기 위한 최적화 도구로서 유용하게 활용될 것으로 기대한다.

## ABSTRACT

This study aims to design and implement automated SEO tools that has applied the artificial intelligence techniques for search engine optimization (SEO; Search Engine Optimization). Traditional Search Engine Optimization (SEO) on-page optimization show limitations that rely only on knowledge of webpage administrators. Thereby, this paper proposes the metadata generation system. It introduces three approaches for recommending metadata; i) Downloading the metadata which is the top of webpage ii) Generating terms which is high relevance by using bi-directional Long Short Term Memory (LSTM) based on attention; iii) Learning through the Generative Adversarial Network (GAN) to enhance overall performance. It is expected to be useful as an optimizing tool that can be evaluated and improve the online marketing processes.

키워드

Search Engine Optmization, LSTM, GAN, Metadata, On-page Optmization

## Ⅰ. 서 론

웹페이지를 검색 상단에 위치시키는 것은 기업의 성과를 결정짓는 중요한 마케팅 전략이다. 웹 3.0의 등장으로 점점 더 많은 기업들이 인터넷 상에서 특히, 검색엔진을 고려한 온라인 마케팅에 노력을 기울일 필요성이 대두되고 있다. 전 세계 97%의 사람들이 온라인 쇼핑을 이용하며, 그 중 70% 이상의 소비자들은 검색 결과의 첫 번째 페이지만을 확인한다. 같은 맥락에 따라 배너 광고를

* corresponding author

이용하는 것보다 검색 상단에 위치하는 것이 온라인 마케팅의 성공요인으로 자리 잡게 되었다[1].

그러나 기존의 온라인 마케팅 방법들은 검색엔진을 고려하지 않고 웹페이지를 구성하는 데에만 치중하여 웹페이지 가시성이 떨어졌다. 기업 대부분은 메타데이터를 입력하지 않거나 경험적 지식에 의존하여 메타데이터를 입력하기 때문에 마케팅 투자 대비 효과는 점점 감소하게 된다. 웹상의 디지털 리소스의 가시성과 접근성을 향상시키는 것은 매우 중요하며, 최적화되지 않은 웹페이지는 온라인 마케팅에서 실패하는 요인으로 작용할 수 있다. 따라서 인공지능 기술을 접목하여 보다 더 객관적인 메타데이터 의사결정을 위한 지표가 필요성이 있다. 본 연구에서는 사용자 질의어를 고려한 GAN 모델링 기반 자동 메타데이터 자동 생성을 위한 접근법을 제안하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 SEO를 간단히 소개하고 기존 연구를 제공한다. 3장에서는 사용자 질의어에 대한 메타데이터 구성을 목표로 표준화된 모델을 제시하고, 이를 바탕으로 메타데이터 생성 및 추천을 위한 방법론을 제안한다.

## Ⅱ. 관련연구

검색엔진 최적화(SEO (Search Engine Optimization), white-hat of SEM (Search Engine Marketing), Decision-Making Fuzzy Rules, SEA (Search Engine Advertising), CA (Contextual Advertising))는 검색엔진이 웹사이트의 순위를 결정하는 여러 요소들을 고려하여 웹사이트를 설계함으로써 트래픽을 유도하는 마케팅 전략이다. 이를 통해 궁극적으로 보다 자신의 웹사이트가 검색엔진의 로봇에 잘 수집될 수 있도록 최적화 하는 방법이다. SEO는 크게 온페이지 기법과 오프 페이지 기법으로 분류할 수 있다. 온 페이지 최적화(on-page optimization, on-page SEO, on-site SEO, on-site optimization)는 검색엔진이 웹 사이트 내용을 읽을 수 있도록 최적화하는 방법을 의미한다. 대표적으로 키워드, 메타정보 최적화 기법이 있다. 오프 페이지 최적화(off-page optimization, off-page SEO, off-site SEO, off-site optimization)는 웹 사이트와 분리되어 사이트의 순위에 영향을 주는 외부 요인을 제어하는 기술을 의미한다. 자신의 웹 사이트 의 인기도, 관련성, 신뢰성 및 권한에 대한 검색 엔진 및 사용자 인식을 향상시키는 것을 목표로 한다.

연구에 따르면 메타데이터를 구현하는 것이 웹에서 사이트와 콘텐츠 가시성을 개선하기 위한 좋은 방법이다[3]. URL, 제목과 같은 다양한 태그에 적절한 단어를 선택하는 것으로 웹사이트를 검색엔진에게 더욱 노출시킬 수 있다[4]. (Park M, 2018)은 학술지에 메타데이터를 생성하여 검색엔진이 코리아사이언스에 관한 페이지를 쉽게 찾을 수 있도록 돕는 XML Sitemap을 설계하였다. 이를 통해 STEM 분야에서 접근을 용이하게 하였고 웹 사이트의 가시성 향상시킬 수 있었다. (Matošević G, 2016)는 백 링크(back link) 분석에 상위 순위의 앵커 텍스트(anchor text)를 바탕으로 기존 웹사이트의 제목을 개선하였고, 이를 통해 웹 트래픽을 증가시킬 수 있음을 밝혔다. 또한 메타데이터 정보가 누락된 웹 사이트에 대하여 SEO 전문가들의 수동 생성한 키워드를 기반으로 사용자 질의어에 따른 meta description을 생성 방법론을 제안하였다. (Luh C, Yang S, Huang T, 2016)는 웹사이트가 검색 상단에 위치하기 위하여 사용자 질의어와의 관련성 있는 웹사이트를 설계하는 것이 중요하며, 잠재 의미 분석(LSA; Latent semantic analysis) 기반으로 평가 지표를 설계하였다. (Armano G, Giulian A, Vargiu E, 2011)는 온라인 마케팅 전략으로 웹사이트의 구문(제목(T), 첫 번째 단락(FP), 등)을 추출하여 사용자 질의어에 따라 알맞게 텍스트 요약 기법을 제안했다. 본 연구에는 온-페이지 최적화를 위하여 문장과 키워드를 추천한다는 점에서 이전 연구와 공통점을 가진다. 그러나 문서의 내용을 고려하고, 특정 질의어에 따른 적절한 메타데이터의 패턴을 분석하여 각 상황에 따른 타켓 문장과 키워드를 생성한다는 점에서 차별성이 있다.

## Ⅲ. 검색엔진 최적화를 위한 메타데이터 자동 생성

적절한 메타데이터 생성을 목표로 본 논문에서는 기존 bi-directetional LSTM 모델의 확장을 기반으로 SEO 기술에 접목하고자 한다. 따라서 bi-directional LSTM 모델을 기반으로 각 특성을 분석하고, 그 결과를 바탕으로 발생 가능한 특성들의 분류와 패턴을 수립하고자 한다.

### 3.1 메타데이터 생성을 위한 분류 지표

표 1. 메타데이터의 분류

| No. | Attribute Value | description |
|---|---|---|
| 1 | title | 브라우저 상단에서 볼 수 있는 텍스트. 검색 엔진은 이 텍스트를 페이지의 "제목"으로 봄. |
| 2 | description | 그 <해당> 페이지에 대한 간략한 설명. |
| 3 | keyword | 문제의 페이지와 관련이 있다고 생각되는 일련의 키워드. |
| 4 | og:(title, description) | Open Graph(OG)는 어떠한 웹사이트도 소셜 네트워크와 연결될 수 있도록 페이지의 HTML <헤드> 섹션에 있는 추가 메타태그. |

사용자 질의어로부터 발생하는 각 상위 문서들을 추출하여 전처리 과정을 통해 단어 벡터를 생

성한다. 아래 [표 1]은 정형화된 각각의 태그들은 검색엔진에게 노출되는 웹사이트의 지시자 역할을 수행한다. 이는 학습을 위하여 전처리 과정에서 사용되며, 각각의 메타데이터들의 다른 특징점들을 추출하고, 그 평균 벡터를 바탕으로 메타데이터를 반복 복제함으로써 적절한 메타데이터를 생성 및 추출하고자 한다.

### 3.2 생성 모델(Generative Model)

기존의 bi-directional LSTM은 네트워크상의 의존 관계 정보를 포함하고 시스템의 성능 향상에 크게 기여해왔다. 본 논문에서는 (Linqing Liu et al, 2018) 방법론을 적용하여 문서의 내용을 적절히 고려한 메타데이터 생성 및 추천을 위한 방법론을 제안한다.

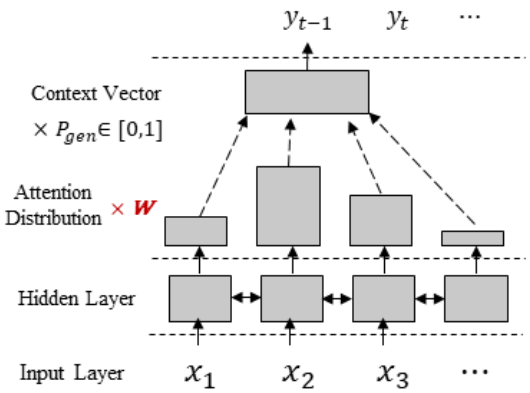


그림 1. 메타데이터 생성 모델링

검색 결과 첫 번째 웹 페이지의 사이트들을 상위 문서로서 정의한다. 임베딩 된 용어들은 입력벡터 $x$=($k_1, k_2, ..., k_i$)로서 표현되고, (See, Abigail, Peter, J. Liu, and C. D. Manning, 2017)의 bi-directional LSTM 모델을 기반으로 특징을 추출한다. 이는 은닉층 $h$=($h_1, h_2, ..., h_i$)을 거쳐, 벡터 $\hat{y}$=($y_1, y_2, ..., y_j$)가 출력 계층으로 출력된다. [그림 1]은 메타데이터 생성을 위한 모델링으로서 생성자(Generative) $G$ 의 파라미터들은 $\theta$에 의하여 표현된다. 상황 벡터(Context Vector) $c_t$은 다음과 같이 계산된다..

$$c_t^* = \sum_i a_i^t c_i w_i \quad (1)$$

이 때, $a_n^t$는 쌍곡탄젠트(hyperbolic tangent) 함수에 따라 학습시킨 파라미터 $\theta$를, $W \in w_i$는 사용자 질의어와의 감정적 연관성의 정도를 나타낸다. 이렇게 seq2seq 모델의 인코더와 디코더 사이에 어텐션 망(Attention Network)을 추가함으로써 입력값의 각 요소들이 출력 값의 각 요소들에 얼마나 영향을 미치는지 결정하는 역할을 수행할 수 있다. 따라서 어텐션 망을 통해 입력 단어 중 현재 출력할 내용에 필요한 단어를 선택적으로 집중하면 문서의 내용을 고려한 문장을 출력할 수 있을 뿐만 아니라 사용자 질의어와의 관련성이 높은 단어를 예측할 수 있다. 따라서 벡터의 유사성을 통하여 사용자 질의어와의 상관관계를 계산되며, 해당 점수 표본을 바탕으로 가중치 $W$가 부여된다. 이렇게 계산된 $c_t$를 기반으로 **Softmax** 층의 시간 스텝 $t$에 대하여 대상 단어들을 예측하는 확률의 계산법은 다음과 같다.

$$p_{vocab}(\hat{y_t}) = Softmax(V'(V[s_t, c_t] + b) + b') \quad (2)$$

$s_t$는 디코더의 상태, $V, V', b, b'$는 학습 파라미터를 의미한다. $p_{vocab}(\hat{y_t})$는 모든 단어들에 대한 확률 분포이며, GAN 모델 기반으로 학습되어 사용자 질의어에 따라 메타데이터를 동적 생성한다.

### 3.2 판별 모델(Discriminative Model)

생성 모델을 통하여 생성된 메타데이터는 판별자(Discriminator) $D(y)$를 통하여 메타데이터가 진짜인지 가짜인지를 분류 해낸다. 확률 $D(y)$가 1의 값을 갖는 것은 진짜와 가짜를 정확히 분별해냈다는 의미를 가진다.

### 3.3 파라미터 훈련

훈련을 위해 다층 은닉층을 둔 DCGAN을 기반으로 이루어진다[11]. $D(G(\theta))$는 생성자가 생성해낸 메타데이터를 판별자가 분류해낼 확률을 나타내며, $D(G(\theta))$가 1에 가까워지도록 학습된다. 학습을 위한 모델은 다음과 같이 표현될 수 있다.

$$\min_G \max_D V(D,G) = E_{y \sim P_{data}}[\log(D(y)] - E_{y \sim G_\theta}[\log(1 - D(G(y))]. \quad (3)$$

$G$를 고정한 채 판별자 $D$를 학습한다. 학습이 반복됨에 따라 원본과 생성자의 분포는 점점 비슷해지다가 완벽히 학습이 된 후에는 일치하게 되며 그 확률은 0.5로 수렴한다. 모델의 성취도, 성능을 평가하기 위하여 loss function을 이용하며 분류기가 실제 메타데이터를 구분케 한다.

## Ⅳ. 결 론

온라인 시장이 급격하게 성장함에 따라 온라인 마케팅은 기업과 고객이 양방향 소통을 위한 중요한 도구로 진화했다. 그러나 웹페이지 제작자들의 일반적이지 않은 키워드 선정은 결국 웹페이지의 SEO 품질을 떨어뜨렸고, 이는 온라인 마케팅을 실

패하는 요인으로 작용하였다.

이에 본 연구에서는 온라인 마케팅을 위해 동적 메타데이터 생성 및 추천을 위한 방법론을 제안하였다. 이는 온라인 마케팅 지표로서 확장될 수 있으며, 기업측면에서 온라인 마케팅 성공을 위한 해결책이 될 수 있음을 기대한다.

## Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 (20170001000031001).

## References

[1] Y. Yuniarthe, “Application of Articial Intelligence (AI) in Search Engine Optimization (SEO),” in Proceedings of the 2017 International Conference on Soft Computing Intelligent System and Information Technology (ICSIIT), Kuta, Indonesia, pp. 96-101, 2017.

[2] N. Sahu, R. Chhabra, “Review on Search Engine Optimization,” IEEE, Journal of Network Communications and Emerging Technologies (JNCET) Vol. 6, No. 6, pp. 19-21, 2016.

[3] J. Zhang, A. Dimitroff, “The impact of webpage content characteristics on webpage visibility in search engine Information results (part I),” Information Processing and Management, Vol. 41, No. 3, pp. 665-690, 2005.

[4] Zhu, Cen, W. Guixing, "Research and analysis of search engine optimization factors based on reverse engineeing," IEEE, in proceeding 3th International Conference on Multimedia Information Networking and Security, 2011.

[5] M. Park, “SEO for an Open Access scholarly information system to improve user experience,” Emerald Publishing Limited, Information Discovery and Delivery, Vol. 46 No. 2, pp. 77-82. 2018.

[6] Matosevic, Goran, "Using anchor text to improve web page title in process of search engine optimization," in Proceedings of the 26th Central European Conference on Information and Intelligent Systems, Varazdin, Croatia, 2015.

[7] Luh, Cheng-Jye, Y. Sheng-An, D.H. Ting-Li, "Estimating Google’s search engine ranking function from a search engine optimization perspective," Online Information Review Vol. 40, No. 2 pp. 239-255, 2016

[8] Armano, Giuliano, G. Alessandro, V. Eloisa, "Experimenting Text Summarization Techniques for Contextual Advertising," In IIR, 2011.

[9] L. Liu, Y. Lu, M Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in Proceeding of the 32th AAAI Conference on Artificial Intelligence, USA pp. 8109-8110, 2018.

[10] See, Abigail, Peter, J. Liu, C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.

[11] Radford, Alec, M. Luke, C. Soumith, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.