JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 3, pp. 551-558, June. 2017

트렌드 지수를 반영한 블로그 랭킹 알고리즘

이 용 석 · 김 형 중
고려대학교 정보보호대학원 빅데이터 응용 및 보안학과

The Blog Ranking Algorithm Reflecting Trend Index

Yong-Suk Lee · Hyoung Joong Kim
Graduate School of Information Security, Korea University, Seoul 02841, Korea

[요 약]

블로그의 성장은 다양한 정보제공이라는 긍정적 측면과 마케팅적 활용이라는 부정적 수단으로 사용되고 있는 문제를 가지고 있다. 본 연구는 대형 포털의 블로그 포스트의 랭킹 결과를 OpenAPI를 이용하여 수집하였고, 탐색적 데이터 분석기법을 통해서 상위 랭크된 블로그의 특징들을 조사하였다. 분석 결과를 보면 상위 랭크에 영향을 주는 요소로는 블로거의 영향력과 포스트의 최근 생성일에 관련성이 높은 것을 알 수 있었다. 이런 평가 알고리즘의 약점으로 인해 파워 블로거의 포스트 중심으로 검색 결과를 편중되게 보여주는 문제가 있었다. 본 연구에서는 다양한 대중의 관심사를 나타내는 트렌드 지수를 통해 랭킹 점수 적용의 공정성을 확보하고, 전문가에 의해 검증된 신뢰 DB정보를 추가하여 컨텐츠 신뢰성을 높이는 알고리즘을 제안하였다. 개선된 알고리즘을 맛집 검색 결과가 실제 지역 학생들의 추천 맛집정보와의 유사도가 높은 것을 확인하였다. 개선된 알고리즘으로 좀 더 신뢰할 수 있는 정보제공이 가능해 졌으며, 방문자수 증가시키는 불법 앱에 의한 순위 조작이 어려워지는 부가적 개선 효과가 기대된다.

[Abstract]

The growth of blogs has two aspect of providing various information and marketing. This study collected the rankings of blog posts of large portal using OpenAPI and investigated the features of blogs ranked through the exploratory data analysis technique. As a result of the analysis, it was found that the influence of the blogger and the recent creation date of the post were highly influential factors in the top rank. Due to the weakness of these evaluation algorithms, there was a problem of showing the search results which is concentrated to the power blogger's post. In this study, we propose an algorithm that improves the reliability of content by adding the reliability DB information which is verified by the experts and reflects the fairness of the application of the ranking score through the trend index indicating various public interests. Improved algorithms have made it possible to provide more reliable information in the search results of the relevant field and have an effect of making it difficult to manipulate ranking by illegal applications that increase the number of visitors.

색인어 : 블로그, 랭킹 알고리즘, 블로그 검색, 정보 검색

Key word : Blog, Ranking algorithm, Blog retrieval, Information retrieval

http://dx.doi.org/10.9728/dcs.2017.18.3.551

Received 07 June 2017; **Revised** 15 June 2017
Accepted 25 June 2017

***Corresponding Author; Hyoung Joong Kim**

Tel: +82-02-3290-4895
E-mail: khj-@.korea.ac.kr

Ⅰ. 서 론

소셜 미디어는 이미 우리 생활의 일부가 되었으면, 여론 형성 정보취득, 의사결정 등 여러 분야에 영향력을 미치고 있다. 이에 따른 정확한 정보의 전달이 중요시 되고 있다. 이를 위해서 수많은 정보 중에 정확한 정보를 찾아주는 검색 엔진, 순위 알고리즘 등의 중요성이 많이 부각되고 있다. 단순히 정확한 정보가 아닌 좀 더 지능화된 정보 검색이 요구된다.

개인 중심의 소셜 미디어 발달과 정보의 공유로 개인이 생산하는 컨텐츠에 대한 정보신뢰가 커지고 있다. 특히 외식산업에서는 개인의 주관적 평판이 포함된 블로그 정보는 상당한 영향력을 미치고 있다. 하지만 이런 영향력의 증가로 인해, 오히려 마케팅의 수단으로 악용되는 사례가 적지 않다.

악의적 의도를 가진 블로거, 블로그 마케터, 불법 순위 조작 프로그래머 등의 등장을 볼 수 있다. 개인적 기록이라는 순순한 목적에서 벗어나, 마케팅과 특정인의 이익을 위한 정보 불평등을 초래하는 문제를 개선하기 위해, 랭킹 알고리즘의 개선, 컨텐츠 품질의 측정, 컨텐츠 저작권보호 등의 여러 가지 기술적, 관리적 조치들이 이루어지고 있다. 그럼에도 불구하고 각종 편법을 동원한 랭킹 결과에 대한 조작이 이루어지고 있다.

이에 블로그 검색, 특히 외식업 검색(이하 맛집검색)의 결과를 실제 수집하고 탐색적 데이터 분석 기법을 통해 랭킹 결과의 특징과 문제점을 조사하였다. 결과는 블로거 파워의 영향력과 컨텐츠의 최신성, 방문자 관심지수 등이 랭킹에 상당한 영향을 주는 것으로 확인 되었다.

본 연구는 2장에서 블로그의 기본적인 용어와 개념에 대해 정의하였고, 3장에서 선행연구에 대한 사례로서 랭킹알고리즘에 대해 조사하였으며, 4장에는 탐색적 데이터 분석을 통해 실제 포털 블로그 포스트 수집한 검색결과를 분석하였고, 5장에는 사용자 기반의 편중된 랭킹 알고리즘의 신뢰성 향상을 위해 전문가 의견 반영 랭킹 알고리즘을 구현하고 평가를 진행하였다. 6장에서는 본연구의 결론과 가능성에 대하여 기술하였다.

Ⅱ. 블로그 랭킹과 마케팅 영향

2-1 블로그 개념 및 특징

블로그(blog)는 인터넷을 의미하는 웹(web)과 자료를 뜻하는 로그(log)를 합친 웹로그(weblog)를 줄인 말이며, 웹을 통해 자신의 생각이나 주장 일상사 등을 자유롭게 표현하기 위해 빈번하게 갱신되는 개인적인 일지로서 탄생하였다. 스스로가 가진 느낌이나 품어오던 생각, 알리고 싶은 견해나 주장 같은 것을 웹일기(로그)처럼 차곡차곡 적어 올려서 다른 사람도 보고 읽을 수 있게 열어 놓은 글모음이다.

블로그는 약 2000년경부터 일반인들에 널리 활용되기 시작했다. 국내에서는 200년부터 인기가 많아지면서 포털사이트와 블로그 전문 사이트들의 대중화가 되었다. 블로그는 뉴스나 정보전달 위주의 컨텐츠 중심이었으나 점차 그 영역을 넓혀 여행, IT, 의료, 과학, 예술 등의 전문분야로 까지 확대되었다. 또한 전문 블로거들의 등장에 따라 생성하는 컨텐츠의 전문성도 매우 높은 수준이 되었다. 이런 경향은 일반인들로 하여금 블로거의 정보를 상당히 신뢰할 수 있는 인식을 심어주게 되었고, 해당 분야의 인기 블로거라고 할 수 있는 파워 블로거들이 등장하게 되었다.

블로그는 개인의 관심사를 일기, 칼럼, 기사 등 다양한 형태로 자유롭게 만들어 그림, 사진 등과 함께 올릴 수 있는 일인 미디어 웹사이트로서 만들기가 쉬우며 상호간의 커뮤니티를 형성하고, 온라인 기술을 통해 칼럼이나 일기, 기사나 소비자들의 제품 사용 후기, 제품 관련 정보, 개인출판, 개인방송 등에 대하여 자신의 생각, 가정이나 상태 등을 자유스럽게 표현할 수 있기 때문에 '일인 미디어' 또는 '풀뿌리 매체(grassroots media)'라고도 부른다[1].

2-2 블로그 관계성과 마케팅

블로그는 그 내용상의 특징을 갖고 있으며, 기존의 인터넷 게시물에 비해 편집이 적고 공개적이며 일대일, 다대다의 커뮤니케이션이 모두 가능하다. 또한 블로그가 갖는 가장 큰 기술적 특성은 포털 등 홈페이지에 들어가 회원가입을 하고 몇 가지 선택사항만 기입하면 자신만의 블로그가 생긴다는 운영의 편리성이다. 블로그는 개방형 네트워크의 형태를 띠고 있어 자신의 블로그에 등록한 글이나 이미지 등이 블로그 서비스를 제공하는 네트워크 포털에 시간 순 또는 카테고리별로 리스트 업 되기 때문에 불특정 타수의 사람들에게 노출될 수 있다. 또한 특정 블로그의 글이 다른 블로그로 옮겨져 포스팅 되거나 다른 사람이 관련 글을 트랙백(track-back)하는 경우가 많기 때문에 블로거들은 자신의 관심분야에 대한 정보의 장으로 활용할 수 있다.

블로그는 마케팅 수단으로서 여러 면에서 장점이 있다. 블로그를 통해 브랜드가 하나의 인격체로 인식될 수 있고, 특정집단에 대한 타겟 마케팅(target marketing)도 가능하며, 개인적 커뮤니케이션을 통한 관계 구축으로 브랜드에 대한 소비자의 충성도를 높일 수 있으며 소비자와의 장기적 관계 유지에도 효과적이다. 이는 최근 마케팅의 가장 중심적인 중추로 자리 잡고 있는 감성마케팅, 구전마케팅, 타겟마케팅에 있어서 블로그가 필수도구라는 인식이 확산이 되면서 기업마케팅에서도 새로운 트렌드로 자리잡아가고 있는 것이다.

블로그와 외신산업에 관한 선행연구는 이선령 등(2010)[2]이 외식고객의 블로그 이용 동기와 서비스 품질 속성 분석을 하였으며, 임성택과 조원섭(2011)[3]이 외식산업 분야에서 기업형 블로그의 구전정보특성이 온라인 구전효과에 미치는 영향 분석에 대해서 연구하였다. 송홍규(2014)[4]는 맛집 블로그의 신뢰성이 외식소비자의 지각혜택, 지각위험, 그리고 온라인 구전에 미치는 영향에 대해서 연구하였다.

2-3 블로그 검색 결과와 구매의사 결정의 영향

블로그의 검색 결과, 특히 상위 랭크된 결과는 맛집 홍보에 상당한 마케팅 효과를 보임을 알 수 있다. 이런 결과는 여러 연구에서 확인되었다.

외식 블로그의 정보 특성 지속성, 중립성, 시각성, 역사성 모두 신뢰와 구매의도에 영향을 미치는 것으로 나타났다[4]. 외식 블로그의 정보특성 중 중립성과 역사의 경우에만 신뢰가 부분 매개역할을 한다. 이상의 연구 결과를 토대로 외식 블로그의 정보 특성과 신뢰를 바탕으로 구매의도를 자극하기 위해서는 지속적으로 운영되고, 시각적으로 잘 정리된 블로그를 이용하여 무조건적인 홍보성의 블로그 포스팅보다는 객관적이고 정확한 정보를 게시하는 블로그를 마케팅에 적극적으로 활용할 필요가 있을 것이다.

홍보 효과를 위해 조회수를 올려주는 불법 소프트웨어를 사용한 악의적 순위 조작이 최근까지 발생하였다. 이런 시도를 하던 불법 앱 개발자, 판매자등이 사이버수사대에 구속되는 일까지 있었다. 이런 불법적 앱에 의한 순위 조작 시도는 2016년 9월, 2017년 3월에도 구속되었다. 최근에도 방송사의 뉴스에 의하면 파워 블로거에게 특정 맛집 홍보 요청을 하여 수 십 분만에 상위에 랭크되는 결과를 확인하였다[5]. 이와 같은 취재 사례를 보더라도 블로거 랭킹 순위 상승 요소는 많이 알려져 있고, 인위적으로 특정 포스트를 상위에 랭크 시킬 수 있다는 사실을 알 수 있다.

표 1. 블로거 평가 요소
Table. 1. Blogger Rating Factor

Classification	Data
Contents quality	Search accuracy, frequency, Content suitability, tag, Number of words, number of images, creation date
Blogger power	Expertise, total number of posts, duration of activity, reliability rating index, number of visitors per day, number of posts per day
Social relation index	Number of neighbors, number of sympathy, revisit rate, relationship rate of change

2-4 블로그 포털 랭킹 현황과 검색 트렌드의 변화

국내 대형 포털에서의 검색 노출도 상당히 민감한 요소이다. 요식업뿐만 아니라, 병원 등도 블로그에 의한 입소문 마케팅의 영향을 많이 받는 곳으로 블로그의 랭킹은 매우 중요한 요소이다. 대형 포털의 블로그 랭킹 알고리즘에 대한 평가 기준을 공개 게시판을 통해 확인한 내용을 아래와 같이 정리하였다. 네이버 블로그 검색 노출 순서는 여러 요소를 종합해서 만든 '관련도'이다. 관련도는 단지 글이 해당 단어를 포함하는지 뿐 아니라, 작성시간, 글의 품질, 인기도 등의 다양한 정보를 활용하여 계산된다. 이러한 다양한 정보들의 가치를 계산하여 수식에 적용한 결과가 검색결과로 나타나게 된다. 블로그 글의 수집과 검색은 자동화된 검색엔진이 수행하고 검색결과 노출 순서 역시 자동화된 계산 방식으로 이루어지기 때문에 개별적인 글에 대해 순위를 인위적으로 조정할 수 없다. 수시로 게시물이 등록/수정/삭제되고 검색결과도 업데이트되기 때문에 어떤 게시물이 상단에 노출될지는 예측하거나 보장할 수는 없다고 밝힌다.

알고리즘은 전체 글을 상대로 작동하는 만큼 정교한 인간의 판단 기준에 비춰 볼 때 일부 이용자에게 만족스러운 결과를 제공하지 못할 수도 있다. 위와 같이 블로그의 결과 순위는 매우 민감한 부분으로 정확한 알고리즘을 설명하고 있지는 않다. 순위에 반영되리라 예상 가능한 통계 지표를 표 1과 같이 본 연구자의 블로그 통계 정보 페이지를 통해 확인하였다[6].

그림 1. 맛집 검색 트렌드의 변화
Fig. 1. Changes in restaurant search trend
출처: 네이버 트렌드 검색 결과

블로그 포스트의 랭킹은 객관성과 정확성을 유지하기 위해서 여러 가지 지표를 수집하여 평가하게 된다. 블로그 평가 지표는 크게 3가지로 분류할 수 있다. 컨텐츠 품질, 블로그 파워, 관계성 지수이다[7].

2-5 맛집 검색에서의 전문가 정보의 트렌드 강화

일반인들이 맛집을 검색함에 있어서, 의존하는 매체에 대한 트렌드 지수를 조사한 결과는 그림 1과 같이 나타난다. 블로그 발생 초기에는 단순 키워드 기반 검색이었고, 블로그에 대한 결과 의존도가 높았으나, 최근에는 광고성 포스트의 범람과 신뢰도, 컨텐츠 중복 등의 문제로 인해, 신뢰성 높은 방송매체나 전문가에 의한 평가 정보를 맛집 검색의 키워드로 더 많이 사용되는 변화를 트렌드 변화를 보고 확인할 수 있다.

블로그 포스트의 랭킹에 대한 신뢰 저하로 인한 맛집 전문 방송 매체의 의존성이 확대 되고 있다.

Ⅲ. 선행연구

3-1 컨텐츠 품질 기반 랭킹 알고리즘

랭킹 알고리즘도 인터넷 서비스의 발전에 따라 같이 진화하고 있는데, 초기 알고리즘으로는 TF-IDF[8]와 BM25[9] 등이 존재한다. 이 알고리즘은 각 웹 문서의 내용 분석을 수행하여 질의된 내용과 웹 문서간의 상호 연관성에 대하여 스코어링을 하고 연관성 점수에 따라서 결정된 순위를 보여주는 방식이다. 간단하면서도 초기 웹 컨텐츠가 많지 않았을 때에는 사용가능한 알고리즘이었다. 그러나 이 기법은 단순히 정확도 위주의 랭킹 알고리즘으로써 컨텐츠 품질 위주의 단편적인 평가 기법이다.

이러한 문제를 개선한 대표적인 기법으로는 PageRank[10]와 HITS[11]등이 있다. 이 기법에서는 컨텐츠의 품질뿐만 아니라 제공자에 대한 권위 점수와 연관성 점수를 각각 계산한 결과

를 기반으로 검색 결과를 보여준다. 이와 같은 랭킹 알고리즘은 제공자의 신뢰도와 집단 평가를 기반으로 랭킹을 도출함으로써 검색결과에 적합한 컨텐츠 제공이 가능한 장점이 있다. 그러나 이런 경우 질의어와 직접적인 연관성이 줄게 되어 정확도 측면에서는 낮아질 수 있는 문제들이 있다. 또한 검색어의 결과와 상관없는 검색결과들을 보여 줄 수 있는데 이를 토픽 드리프트(topic drift)현상 이라고 부른다[12]. 토픽 드리프트 문제를 해결하기 위하여 질의와 연관된 웹 문서만을 대상으로 권위 점수를 계산하는 연관성파급 기법이 제안되었다. 대표적인 연관성 파급 기법에는 QD-PageRank등이 있다.

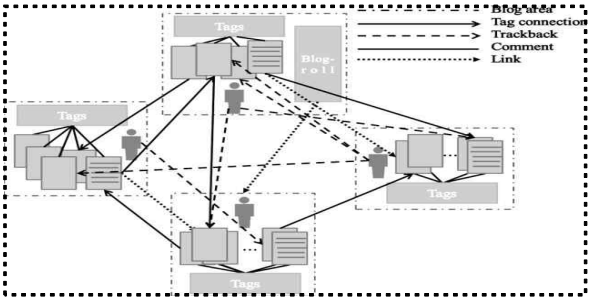

그림 2. 블로그와 포스트의 관계도
Fig. 2. Relation between blog and post
출처: A blog ranking algorithm using analysis of both blog influence and characteristics of blog posts [8]

연관성 파급 기법은 사용자가 질의를 주는 시점의 질의어와 연관된 웹 문서들을 찾고, 이 문서들 간의 하이퍼링크를 이용한 네트워크를 구성한 후, 이 네트워크를 대상으로 연관성 파급을 수행한다. 기존 연관성 파급 모델은 세 단계의 과정으로 인하여, 상당한 계산 오버헤드가 있다[13].

3-2 가중치 적용 알고리즘

답글과 역방향 연결고리 가중치 알고리즘(WCT, Weighted Comment and Trackback)은 2가지 평가 요소로 구성되어 있다. 컨텐츠 기반 평가 알고리즘과 연결 기반 평가 알고리즘이다. 연결기반 평가 알고리즘은 Blogsphere 내의 연관 고리에 대한 평가를 수행한다. 그림 2는 블로그간 네트워크를 보여준다. 블로그 간에는 다양한 관계가 있다. 블로그에는 하나의 블로거, 다중 포스트 페이지, 태그 세트 (즉, 선택된 포스트의 키워드), blogroll 등이 존재 한다. 게시물에는 제목, 내용, 태그, 설명, 트랙백, 등등. 또한, 몇 가지 종류의 연결이 있다.

WCT 알고리즘의 블로그 간의 상호 연결과 구조적 특징을 종합하여 랭킹을 출력한다. 특히, 코멘트와 트랙백에 가중치를 부여한다. 즉, WCT 알고리즘은 양질의 게시물은 더 많은 블로거가 관계성을 가지고 있고, 관심을 보여준다는 기본적인 아이디어를 기반으로 한다. WCT 랭킹 점수는 SPEAR(spamming-resistant expertise analysis and ranking)[14] 알고리즘을 적용한다.

3-3 기존 알고리즘의 한계와 개선

기존의 알고리즘은 포스트 컨텐츠와 관계성 중심으로 계산식에 의한 랭킹 결과를 보여준다. 이런 알고리즘은 몇 가지 문제점을 가지고 있다.

첫째, 블로거의 관계성 지수의 계산식 구조의 특성상 파워 블로거의 영향력에 의해 지속적으로 상위에 랭크되어 검색 결과가 특정 블로거들에게 쏠림현상이 발생하는 문제가 있다. 둘째, 파워 블로거들이 특정 음식점(상호)에 대해 유사한 포스트를 생산할 경우 상위에 중복되어서 노출되는 현상이 발생한다. 이는 사용자로 하여금 다양한 정보 제공을 못할 뿐더러 왜곡된 결과를 보여줄 수 있다. 셋째, 블로거 평가 알고리즘을 알고 있는 대다수의 블로거 마케터들은 이런 약점을 활용하여, 의도적으로 특정 포스트를 상위에 게재시킬 수 있다. 이런 한계 극복을 위하여 외부 객관적 평가 점수, 즉 검증된 전문가의 의견 반영을 통해 편중된 검색 랭킹 결과에 대한 조정이 필요하다.

본 연구에서는 평가 알고리즘의 객관성 확보를 위해, 다양한 외부 전문가의 평가 요소를 선정하였다. 대표적인 예로 전문 맛집 소개방송의 출연과 포털의 트렌드 지수를 고려한 평가 결과를 보여주었다. 각 평가요소의 적용비율은 트렌드 조사 결과 서비스를 활용하여 보다 객관적으로 랭킹 점수를 도출하기 위해 노력하였다[15].

Ⅳ. 포털에서의 블로그 랭킹 현황 조사

기존 알고리즘을 적용한 국내 대형 포털의 블로그 상위 랭킹에 대한 실증적 현황 조사를 위해 아래 표 2와 같이 블로그 데이터 수집 및 분석을 수행하였다. 포털에서 제공하는 OpenAPI를 활용하였으며, 일부 수집되지 않는 블로거 지수들에 대해서는 크롤링을 통해 추가 수집하였다.

표 2. 블로그 수집 데이터 개요
Table. 2. Blog Collection Data Overview

Item	Object	
Data collection targets	Naver blog	
Data Collection Period	2017.04.10~2017.04.16	
Data collection range	Search results for blogs	
Search Keywords	kangnam matzip, hongdae matzip, godae matzip	
Data Collecting method	Naver blog API Python BeautifulSoup crawling	
Data size	[1,000(100*10times query) * 3 area = 3,000] * 7days = 21,000	
Data information	Blogger Index	Total post, neighbor count, Activity
	post	Generation date, empathy, number of images, content length, keyword, Tag
	Time serial analysis	post ranking change

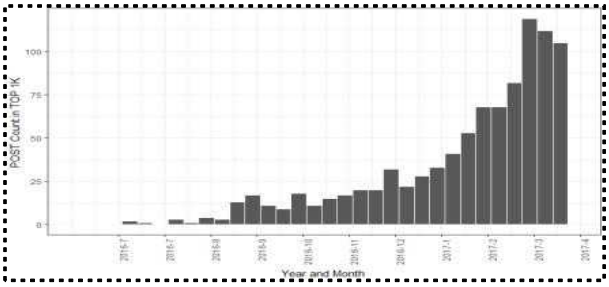

그림 3. 상위 1천건 블로그 포스트 생성일 히스토그램
Fig. 3. Histogram of top 1,000 blog posts' created date

4-2 블로거 상위 랭킹 기여도 분석

대상 검색어는 랭킹에 특히 민감한 요식업분야로 한정하였으며, 특히 음식점이 밀집한 강남맛집과 변화에 민감한 젊은 세대를 대변하는 홍대맛집, 마지막으로 주변 정보에 대한 경험 정보를 많이 가지고 있는 고대맛집으로 선정하였다. 최대한 랭킹에 영향을 주는 특징 수집을 위하여 기본 OpenAPI제공 정보 이외에도 크롤링을 통하여 activity와 이웃 수 등의 부가정보도 수집하였다. 수집된 자료는 R을 사용하여 데이터 특성을 분석하였다. 상위 1천 건 블로그 포스트의 히스토그램은 그림 3과 같다. 상위 1천 건 중 약 75%는 모두 최근 4개월 이내에 작성된 글임을 확인할 수 있었다. 그 중에서도 1개월 이내의 블로그 포스트들이 절반이상을 차지함을 알 수 있다. 상위에 랭크되는 블로그 포스트는 최신 글 일수록 높은 평가를 받고 있다.

상위 1천 건에 포함된 포스트의 블로거는 약 2,500명 이하의 이웃 블로거와 관계성을 가지고 있다. 5,000개 이상의 블로거 이웃과 관계를 가진 블로거도 있으며, 최대 62,914개의 블로그 이웃을 가진 블로거도 포함되어 있다. 블로거 이웃정보가 표기되지 않은 477개는 NA로 처리하였다. 상위 15위내에 있는 포스트의 블로거 데이터를 상세 조사한 결과를 보면 ID별 평균 등록 포스트수는 2,653개이며, 이웃은 12,071명, 일일 평균 방문자는 149,68명이다. 최근 1개월간의 포스트 작성수를 보면, 매일 2~3개정도의 포스트를 꾸준히 등록하고 있다. 블로그 포스트의 rank, 즉 표시되는 페이지 수를 10개의 페이지단위로 그룹핑하여 평균값을 계산하였다.

첫 번째 영향 요소로는 페이지와 제일 높은 상관 관계를 보여주는 값은 생성일이다. 최신 생성일일수로 상위 페이지에 표시되고 있는 현상은 3가지 검색에서 동일한 결과를 나타낸다.

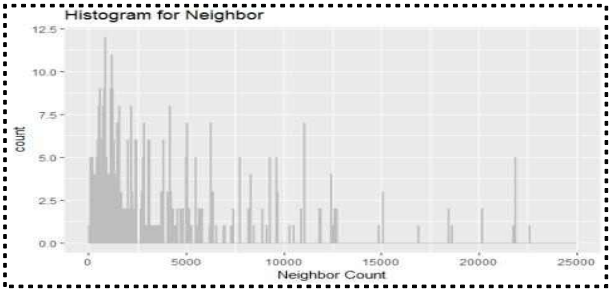

그림 4. 상위 1천건 블로거 이웃수 히스토그램
Fig. 4. Histogram of top 1 thousand blogger's neighbor

표 3. 상위 1K 포스트 특징 분석
Table. 3. Top 1,000 post features analysis

키워드	페이지 수	이웃 수	activity	전체 포스트	Avg. created date
강남	1P~	7,376	5,331	1,824	2017-03-29
	10P~	5,883	3,489	1,834	2017-02-25
	20P~	3,935	192	1,461	2017-02-19
	30P~	4,476	6,201	1,235	2017-02-02
	40P~	3,152	7,042	1,118	2017-02-11
	50P~	4,574	6,084	1,185	2017-01-14
	60P~	2,932	5,086	1,257	2017-01-08
	70P~	2,250	7,306	1,095	2016-12-20
	80P~	2,415	NA	1,272	2016-12-24
	90P~	6,029	1,469	1,153	2016-11-04
Avg.		4,357	4,658	1,193	2017-01-15

둘째, 이웃수와 activity, 전체 글과 어느 정도의 상관관계를 보여주나 강한 관련성을 보이지는 않는다. 특히 activity에서는 결측치들을 상당수 포함하고 있었다.

activity와 전체 글 수, 이웃 수는 다중공선성(multicollinearity) 문제를 가지고 있다. 다시 말하면 포스트에 대한 접속 빈도가 전체 글 때문일 수도 있고, 이웃 수가 많아서 일 수 있다. 상호간의 event에 대해 전후 관계를 설명하기 어렵다. 검색대상이 되는 전체 포스트의 수가 50만 이상이 되는 강남맛집에서는 4개월 전까지만 포함되어지지만, 5만정도 되는 고대맛집의 경우에는 2년 3개월 전의 포스트까지도 검색된다.

4-3 포털 블로그 랭킹 결과의 특징과 문제점

수집된 데이터 분석을 통해서 본 표 3에서와 같이 상위평가 글들은 대부분 이웃수 2~3천 이상의 파워 블로거가 많은 비율을 차지한다. 또한 1일 방문자수도 5,000명이상 정도로 일반 블로거와는 큰 격차를 보인다. 랭킹은 관계성과 방문자수에 많은 영향을 받음을 확인할 수 있었다.

또한, 최신 글들이 상위에 주로 랭크되며, 여러 확인되지 않은 이유로 상위의 글들이 삭제되는 현상도 보였다. 일반인 웹 검색시에 주로 참조하게 되는 30위 포스트(검색결과 3페이지 이내)의 컨텐츠를 조사해 보면, 중복된 컨텐츠가 전체 70%를 차지하고 있다.

표 4. 맛집 중복(고대맛집, 상위 30위)
Table. 4. Overlapping restaurants(kodae restaurant,top 30)

Restaurant	Duplicate number	Percentage
Murmur de gusto	11	52%
Chamchigongbang	4	19%
Lobster bada	4	19%
Samdeungsin	2	10%
Sum	21	70%

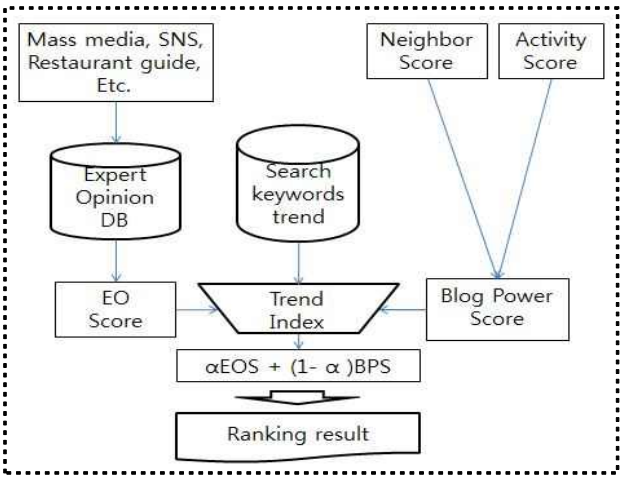

그림 5. TRR 알고리즘 개념도
Fig. 5. Diagram of TRR algorithm

특정 음식점이 상위 30개의 결과 중 50% 이상을 차지하고 있다. 실제 맛집일 수도 있고 블로그에서 홍보가 잘되는 음식점일 수도 있다. 만약, 마케팅 기법에 의해 중복이 된다면, 정보 다양성과 선택군 측면에서는 상당히 제한적이라는 문제가 있다. 믿을 수 있는 전문가의 평가 정보반영을 통해 랭킹 알고리즘 개선의 아이디어가 필요하다.

Ⅴ. 트렌드 지수를 반영한 블로그 랭킹 알고리즘

5-1 트렌드 지수 반영 랭킹 알고리즘(TRR, Trend Reflection Ranking algorithm)

기존 평가 방식은 포스트 생성 초기에 집단선택 우월성을 가진 파워 블로거가 유리한 평가를 받을 확률이 높다. 따라서 블로거 파워의 영향력을 줄이고, 전문가의 의견을 반영하여 신뢰도를 높힐 수 있는 알고리즘을 제안한다. 또한 검색어 기반 관심도 지수인 트렌드 지수를 사용하여, 블로거 파워와 전문가 견해(expert opinion)를 합리적으로 반영할 수 있도록 개선하였다.

본 논문에서는 파워 블로거의 영향력을 낮추고 집단 선택의 문제를 제한하는 개선된 평가모델을 트렌드 반영 랭킹 알고리즘(TRR, Trend Reflection Ranking algorithm)이라 명명한다. 그림 5는 트렌드 반영 전문가 랭킹 알고리즘의 절차를 제시한 개념도이다.

5-2 TRR 검색 알고리즘 구현

TRR 랭킹 알고리즘 구현의 평가 점수 산출식은 아래와 같이 산정 하였다.

$$TRR = \alpha EOS(\,) + (1-\alpha)BPS() \qquad (1)$$

EOS(Expert Opinion Score)는 전문가 점수로써 방송에서의 소개 유무에 따라 트렌드 비율의 총합이다. BPS(Blog Power Score)는 블로거의 파워에 대한 측정값으로서 본 연구에서는 이웃 수(N)와 활동성지수(A)를 사용하였다. 활동지수(A)는 수집 블로거 전체 중 최대 activity와 비율로 산정하였다.

5-3 전문가 평가 지수의 결정 절차

맛집 검색을 위한 사용자의 트렌드 기반으로 반영지수를 산정한다. 실제 검색어 트렌드 비교를 해보면 일반인들이 맛집 검색을 함에 있어서 방송 매체의 의존도가 매우 높음을 알 수 있다. 전문가 평가 점수는 맛집 검색어의 트렌드 지수를 기준으로 각각의 전문가 매체의 트렌드 지수로 나누어 총합으로 계산한다.

$$EOS = \sum_{i=1}^{n}(\omega_i \times \frac{(\text{전문가트렌드키워드})_i}{\text{맛집트렌드키워드}}) \qquad (2)$$
$$\text{if 해당맛집이 전문가}DB\text{ 존재},\omega = 1$$
$$\text{비존재},\omega = 0$$

본 연구에서는 국내 3대 주요 맛집 소개 방송프로그램의 트렌드 지수를 1년간 비교하여 각각의 비율로 지수를 산정하였다. EOS는 수요미식회 50%, 백종원 3대천왕 35%, 생생정보통 15%의 트렌드 값을 부여하였다. w는 방송 출연에 따라 1과 0을 입력하였다. 다음으로 블로거 파워 점수는 이웃 수와 활동지수를 임의로 5:5로 산정하여 아래 공식으로 산정한다. 실제 포털 계산 알고리즘이 공개되지 않아 공식 3으로 가정하였다.

$$BPS(Neighbor, Activity) = \qquad (3)$$
$$0.5 \times (\frac{Neighbor}{Neighbor_{MAX}}) + 0.5 \times (\frac{Activity}{Activity_{MAX}})$$

EOS와 BPS 점수를 공정하게 계산하기 위한 기준으로 포털의 검색어 트렌드 지수를 기준으로 사용하였다. 식 4에서처럼 TV 방송별 검색어 합계와 맛집 검색 키워드 비율에 의해 결정한다.

$$\alpha = \frac{\text{방송별트렌드지수의 합}}{\text{전체맛집트렌드지수}} \qquad (4)$$

표 5. 랭킹 비교표
Table. 5. Ranking comparison

No	Potal rank	BPS	TRR
1	Deulchangko	Youngcheol burger	Kodaedakbal
2	Chamchigongbang	Seorae galmaegi	Duki
3	Murmur de gusto	Chilbaekjip	Ujau
4	Lobster bada	Benares	911Onban
5	Samdeungsin	Bagueldokas	Eunhwasu
6	Chamchigongbang	Meonokameona	Youngcheol burger
7	Chamchigongbang	Maninnapoli	Seorae galmaegi
8	Murmur de gusto	Youngcheol burger	Chilbaekjip
9	Lobster bada	Chilbaekjip	Benares
10	Ilmiok	Meonokameona	Bagueldokas

5-4 랭킹 알고리즘의 비교 실험

표 5의 결과와 같이 포털의 10위 내의 랭킹결과를 참고하기 위해 10위까지 기록하였으며, 랭킹에서 중요한 처음 3페이지에 해당하는 30위권에 대하여 결과변화를 확인하였다.

블로그 파워 지수에 따른 랭킹 변화 대조군으로써 TRR을 사용한 결과를 비교하였다. 포털 랭킹은 1위에 들창코, 2위, 6위, 7위에 참치공방이 중복된 순위로 표시되었다. 3위 무르고무르드도 8위에 중복 랭크되었다.

기존 포털 랭킹 알고리즘과 BPS 순위 평가 방식은 TRR보다는 동일 맛집 정보가 중복되어 나타난다. 이는 다양한 맛집에 대한 정보를 얻는 사용자 입장에서는 정보획득의 걸림돌이 될 수 있다.

5-5 랭킹 알고리즘의 검증 결과 분석

본 연구에서 설계한 TRR 알고리즘의 검증을 위해서 고려대학생들의 포털 커뮤니티인 고파스(Koreapas.com)[16]을 사용하였다. 고파스 사이트는 고대 학생들의 정보 교환을 위한 목적으로 하고 있으며, 상호 질문에도 성실한 답변을 기대할 수 있는 커뮤니티이다. 고려대 학생들 사이에는 신뢰할 수 있는 정보를 교환하는 사이트여서 의견 수렴을 위해 사용하였다.

조사 방법으로는 고려대 주변 맛집에 대한 질문 글과 이에 대한 답변을 모두 조사하였다. 이 목록을 기존의 포털 블로그 검색 결과와 *BPS*, *TRR*의 랭킹 결과를 비교하여 얼마나 지역 학생들의 추천 목록과 일치하는지를 비교하였다. 상위 목록 10개에 대해서는 TRR이 40%로 실제 고대학생들의 추천 맛집과 일치함을 보여주었다. 포털 랭킹 순위와 비교하면 일미옥 1개만이 고대학생들이 추천하는 맛집에 포함되었다. BPS에서는 영철버거가 중복 2회로 20%로의 일치율을 나타내었다.

이런 결과에서 알 수 있듯이 고려대학교 주변을 생활 기반으로 하는 학생들의 추천 맛집과 포털의 랭킹 결과는 다른 우선순위를 나타내었다. 본 연구에서 구현한 TRR 알고리즘을 통해서 고대 주변 맛집 검색 결과에서는 현지 생활인이 추천하는 맛집에 근접한 결과를 확인 할 수 있었다.

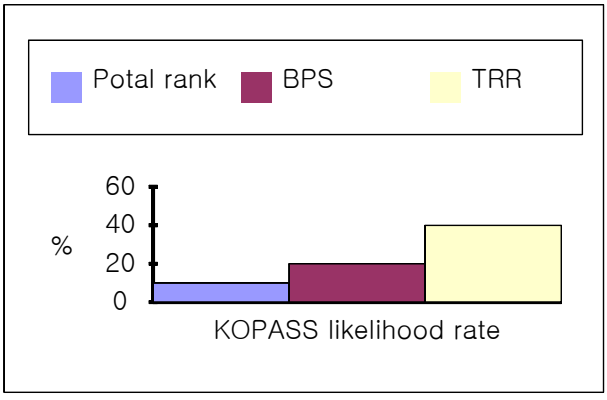

그림 6. 고대학생 추천(Koreapas) 맛집 유사율
Fig. 6. Likelihood ratio of the ranking results recommended by Korea university students.

Ⅵ. 결 론

본 연구에서는 첫째 성과로는 포털 블로그 데이터 수집과 분석을 통해서 상위에 랭킹되는 포스트의 특징에 대해서 탐색적 데이터 분석 기법을 통해 확인한 것과 랭킹 알고리즘의 문제점에 대한 발견이다. 상위 1,000개의 포스트는 블로거 파워에 의해 영향을 받고 있으며, 포스트의 생성일이 최신일수록 상위에 랭크되는 특징이 있었다.

이로 인해 파워 블로거들이 동일 맛집에 대한 포스트를 생성하게 되면, 맛집 검색결과에서 중복 컨텐츠 비율이 높음을 확인하였다. 이런 결과는 일반적으로 해당 지역에 대한 다양한 맛집 정보를 검색하고, 정보 비교를 통해 음식점을 선택을 원하는 사용자에게는 적합하지 않은 검색 결과이다. 또한 검색어 트렌드 결과를 보면 사용자는 지역 검색을 사용한 맛집 검색보다는 맛집 전문 방송의 출연 결과를 좀 더 많이 참고하는 경향을 보였다.

두번째의 성과로는 트렌드 지수를 적용한 전문가 평가 알고리즘을 사용하여 랭킹결과에 적용하는 방법을 제안하였고, 결과 개선에 대해 검증하였다. TRR알고리즘을 적용함으로써 고대맛집 검색 결과에서 실제 주변 맛집 정보를 많이 알고 있는 대학생들의 의견과 유사도가 높아졌다. 또한 전문가에게 검증되지 않은 인지도가 낮은 맛집이 상위에 중복 랭크되는 현상이 감소하는 효과도 있었다. 이로써 일반인이 블로그 검색을 통해 좀 더 다양한 맛집 정보를 획득할 수 있고 선택의 폭이 넓혀지는 효과가 있다.

검색 결과에 대한 검증을 위해 고려대학교 학생들의 커뮤니티인 고파스의 맛집 정보를 통해 TRR의 랭킹 결과와의 유사도를 확인하였으며, 기존 포탈의 검색 결과보다는 개선된 일치율을 보여주는 것을 확인 할 수 있었다.

추가적으로 예상되는 부가효과로는, 랭킹 조작에 대한 악의적 시도가 기존 알고리즘에 비해 어려운 장점이 있다. 기본적으로 전문가의 의견을 반영함으로써, 단순 클릭 수나 파워 블로거의 관계성 지수만으로는 상위에 랭크되기는 어렵다. 악의적 블로그 마케팅에 대한 개선이 가능할 것으로 예측된다[17].

하지만 본연구에서도 한계점은 여러 가지 존재 한다. 대형 포털의 C-Rank 기반의 알고리즘에 대한 구체적 정보 미공개로 실제 알고리즘에 대한 연구가 어려웠던 점과 클릭수와 같은 중요한 평가 정보는 OpenAPI기반으로 제공되지 않아서, 수집가능한 정보가 제한적이었다. 이는 평가 알고리즘에 독립변수로 사용되는 인자의 값이 불완전할 수 있다는 한계가 있었다.

그럼에도 불구하고 본 연구에서 사용한 전문가 정보의 반영과 트렌드를 지수를 반영하는 아이디어의 독창성을 활용하면, 순위조작의 개선방안으로 가능성이 있을 것으로 기대한다. 또한 전문가 의견요소는 다양한 외부 입력변수로 대체 가능하다. 전문가 의견 점수로써 방송 매체뿐만 아니라, SNS 언급도, 포털 일반 검색어 순위 등 다양한 정보순위를 트렌드 지수로 활용할 수 있다. 또한 본 실험에서는 맛집 만을 대상으로 하였지만 다양한 도메인에 대해서도 추가 연구가 기대된다.

참고문헌

[1] J.-E. Kim and Y.-Y. Kim, “How the characteristics of the food-blog marketing effect to purchasing intension with the mediation effect of trust,” *Korean Journal of Tourism Research,* vol. 30, no. 5, pp. 85-105, 2015.

[2] S.-L. Lee, H.-H. Yoon, and N. Young, "Blogs in the restaurant industry: Consumer usage motivation and service quality perception," *Korean Journal of Hotel Administration,* vol. 19, no. 6, pp. 273-287, 2010.

[3] S.-T. Lim, W.-S. Cho, "The effects of business blog Information characteristics influencing on electronic word-of-mouth in the food service industry: Emphasis on trust transference," *Korean Hospitality and Tourism Academe,* vol. 20, no. 5, pp. 165-180, 2011.

[4] H.-G. Song, "A study of relationship of gourmet blog’s reliability with the perceived benefits, perceived risk and online word of mouth of eating out consumer," *Culinary Science & Hospitality Research,* vol. 20, no. 6, pp. 275-291, 2014.

[5] The Kyunghyang Shinmun, news about blog rank hacking [Internet] Available: http://news.khan.co.kr/kh_news/ khan_art_view.html?artid=201609121242001&code=940202#csidx9f55f61f28a12039a949b1c036c78db

[6] NAVER, FAQ about ranking algorithm, [Internet]. Available: https://help.naver.com/support/contents/contents.nhn?serviceNo=606&categoryNo=15024

[7] J.-W. Kim, U.-I. Yun, G.-B. Pyun, H.-M. Ryang, G.-I. Lee, E.-C. Yoon, and K.-H. Ryu, "A blog ranking algorithm using analysis of both blog influence and characteristics of blog posts," *Cluster Computing,* vol. 18, no. 1, pp.100-104, 2015.

[8] Jiaul H. Paik, "A novel TF-IDF weighting scheme for effective ranking," *Proceedings of the International ACM Conference on Research and Development in Information Retrieval*, pp. 343–52, 2013.

[9] J.-Y. Lee, "A study on the pivoted inverse document frequency weighting method," *Journal of the Korean Society for Information Management*, vol. 20, no. 4, pp. 233–248, 2003.

[10] P. Lawrence, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Technical Report No. SIDL-WP-1999-0120, Stanford University, 1998.

[11] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM,* vol. 46, no. 5, pp. 604-632, 1999.

[12] M. Richardson and P. Domingos, "The intelligent surfer: Probabilistic combination of link and content information in pagerank," *Advances in Neural Information Processing Systems,* v*ol. 14,* pp.1141-1448, 2002

[13] S.-C. Lee, D.-J. Kim, H.-Y. Lee, S.-W. Kim, J.-B. Lee, "C-rank: A contribution-based approach for web page ranking," *,Journal of KIISE: Computing Practices and Letters,* vol. 16, no.1 , pp. 100-104, 2010.

[14] A. Yeung, G. Noll, N. Gibbins, C Meinel, and N, Shadbolt, “SPEAR: Spamming-resistant expertise analysis and ranking in collaborative tagging systems,“ *Computational Intelligence,* vol. 27, no. 3, pp. 458–488, 2011.

[15] M.-k. Seo, *R for Practical Data Analysis,* 1st ed. Gilbut, 2014.

[16] Koreapas.com, Community Bulletin board about restaurant recommended by Korea University students, [Internet]. Available: http://www.koreapas.com

[17] J.-H. Lee, W.-S. Lee, J.-W. Park, and J.-H. Choi, “The blog polarity classification technique using opinion mining,“ *Journal of Digital Contents Society,* vol. 15, no. 4, pp. 458–488, 2014.

이용석(Yong-Suk Lee)

1999년 : 경기대학교 전자공학과 학사
2015년~ 현재 : 고려대학교 빅데이터 응용 및 보안학과 (석사과정)

1999년～현 재: 한국후지쯔(주)
※관심분야 : 빅데이터분석, 도커, 오브젝트 스토리지, 그로스 해킹 등

김형중(Hyung-Joong Kim)

1978년 : 서울대학교 전기공학과 학사
1986년 : 서울대학교 제어계측공학과(공학석사)
1989년 : 서울대학교 제어계측공학과(공학박사)

1989년～2006년: 강원대학교 교수
2006년～현 재: 고려대학교 정보보호대학원 교수
관심분야 : 컴퓨터보안, 패턴인식, 가역정보은닉, 머신러닝, 빅데이터분석 등