장소에 내재된 토픽 기반 기사 추천

노윤석○, 손정우, 박성배, 박세영, 이상조
경북대학교 컴퓨터학부
{ysnoh○, jwson, sbpark, sypark, sjlee}@sejong.knu.ac.kr

Article Recommendation based on Latent Place Topic

Yunseok Noh○, Jung-Woo Son, Seong-Bae Park, Se-Young Park, Sang-Jo Lee
School of Computer Science and Engineering
Kyungpook National University

요 약

스마트폰의 대중화와 함께 그에 내장된 GPS를 활용하여 컨텐츠를 제공하는 서비스들이 점차 늘어나고 있다. 그러나 이런 컨텐츠를 단지 위도, 경도 좌표 정보만을 기초로 구성하게 되면 실제 그 위치가 가지는 의미적 특성을 제대로 반영하지 못하게 된다. 사용자의 위치를 기반으로 그에 맞는 서비스를 제공하기 위해서는 장소의 토픽을 고려해야한다. 본 논문은 장소에 내재된 토픽을 바탕으로 한 기사 추천 방법을 제안한다. 장소와 관련된 문서로부터 장소의 토픽을 표현하고 그 토픽을 기사 추천에 이용한다. 제안한 방법이 실제로 장소에 내재된 토픽을 잘 반영함을 보이고 또한 이를 바탕으로 장소와 관련된 적합한 기사를 추천하는 것을 보여준다.

주제어: 장소의 토픽, Explicit Semantic Analysis, 토픽 모델, 기사 추천

1. 서론

GPS를 탑재한 휴대 기기가 대중적으로 널리 보급되면서, 사용자의 '현재 위치'가 중요한 정보로 부각되고 있다. 실제로 스마트폰을 염두에 둔 다양한 소셜 네트워크 서비스(SNS)들이 사용자의 위치 정보를 적극적으로 활용하고 있다. Foursquare[1]는 사용자가 방문한 POI (Point of Interest)에 대한 로그를 남기고 지인과 그 정보를 공유하는 서비스를 제공하고 있고, Google Earth[2]와 Flickr[3]를 비롯한 몇 가지 서비스들은 위치 정보를 이용해 컨텐츠를 지도상에 보여주는 기능을 제공한다. 사용자 위치 정보의 활용은 비단 SNS 뿐만 아니라 연구자들 사이에서도 활발하게 연구의 소재로 사용되고 있다. 사용자의 GPS 궤적 및 이력을 협업 필터링에 사용하여 추천을 한다거나 [1, 2] 역시 사용자의 위치 이력에 기초하여 그들 간의 유사성을 비교하고 이를 추천에 이용하는 연구 [3] 등이 그 사례이다.

많은 모바일 서비스들과 연구들이 사용자의 위치 정보에 주목하고 있지만, 이들은 사용자의 위치 그 자체만을 고려할 뿐 사용자가 현재 머물고 있는 장소가 어떤 곳인가에 대해서는 간과하고 있다. 특정 장소에 방문한 사람의 관심사는 대체로 그 곳의 토픽과 연관이 있다. 따라서 장소에 내재된 토픽을 고려함으로써 사용자가 현재 위치에 무엇 때문에 머무르고 있는지 짐작할 수 있고, 그에 맞춰 사용자의 기대에 부합하는 적절한 서비스가 가능하다. 예를 들어 세종문화회관에 오페라를 관람하러 온 관객의 GPS 정보를 활용하여 기사를 추천하는 경우를 생각해보자. 장소의 토픽을 반영하게 되면 세종문화회관 및 그 주변과 직접적으로 관련된 뉴스나 정보를 추천하는 것 외에 비슷한 토픽을 가진 예술의전당에 관련된 기사나 그 곳에서 공연되는 발레에 대한 기사를 추천하는 것이 가능하다. 이런 기사들은 문화행사, 공연 등에 관심 있는 세종문화회관 관객에게 충분한 흥미를 불러일으키는 유용한 기사라 할 수 있다. 또는 사용자가 예전에 방문했던 장소를 고려하여 유사한 토픽을 가지는 다른 장소를 추천하는 서비스도 좋은 예가 될 수 있다. 사용자가 가보지 않아 모르는 곳이지만 평소 사용자가 즐겨 찾는 곳과 유사한 장소에 대한 정보는 좋은 컨텐츠가 될 것이다.

본 논문에서는 내재된 장소 토픽을 이용하여 기사를 추천하고자 한다. 이를 위해 먼저, 사용자가 현재 위치한 장소의 토픽을 발견하고, 그 토픽과 비슷한 토픽을 가지는 기사를 추천한다. 장소의 토픽을 기사 추천에 고려함으로써 장소의 물리적 정보 활용 즉, 위도, 경도 좌표 정보만을 활용하는 기사 추천을 벗어날 수 있다. 실험에서 실제 장소의 예를 통해 주어진 장소에 대한 토픽이 정확히 표현됨을 보인다. 또한 사용자의 현재 위치와는 멀리 떨어졌지만 유사한 토픽을 가지는 곳에 대한 기사 또는 장소의 토픽을 전반적으로 반영하는 기사가 추천되는 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자의 위치 정보를 활용한 기존 연구와 지역의 토픽 발견에 관한 관련연구를 살펴본다. 3장에서는 본 연구에서 제안하는 장소의 토픽 발견과 이를 통한 기사 추천에 대해 자세히 설명한다. 4장에서는 실험 및 그 결과에 대해 논의하고 마지막으로 5장에서는 결론 및 향후 계획을 보인다.

1) http://foursquare.com
2) http://earth.google.com
3) http://www.flickr.com

2. 관련연구

모바일 기기의 GPS 정보를 사용하는 SNS가 점점 늘어나고 사용자의 실시간 위치 정보를 활용한 여러 연구들이 진행되고 있다. Takeuchi et al. [1]과 Zheng et al. [2]은 여러 사용자의 위치 궤적 정보를 모으고, 이 데이터를 바탕으로 하여 협업 필터링을 사용한 추천 시스템을 제안했다. Li et al. [3]은 사용자의 위치 이력을 시간 흐름에 따라 계층적으로 적용하여 사용자간의 유사성을 측정하였다. 이와 같은 사람들의 위치 정보 비교에 근거한 연구들은 같은 장소를 방문한 사람들은 비슷한 사용자 선호도 성향을 보일 것이라는 것을 기본 가정으로 하고 있다. 즉, 이런 연구들은 사실상 장소의 토픽을 암묵적으로 이용한다고 볼 수 있다. 그러나 명시적으로 장소에 내재된 토픽을 고려함으로써 위치 이력이 상이한 사용자간에도 유사성을 발견할 수 있고, 또한 장소의 토픽을 추천의 중요한 근거로 활용할 수 있을 것이다.

이런 점에 대한 착안으로, 그리고 위치와 연관된 데이터의 양이 풍부해지면서, 장소 또는 지역의 토픽을 발견하기 위한 연구도 진행되고 있다. 이들 연구들은 주로 GPS 정보와 명시적으로 연결된 문서를 활용하여 지역에 내재된 토픽을 분석한다. Sizov [4]가 제안한 GeoFolk 모델은 Flickr tag 데이터로부터 지역의 의미 정보를 파악하고 이에 따라 지역을 분류하고 군집화 하였다. Yin et al. [5]도 Flickr tag 데이터를 사용하였으며, 위치/문서에 기반한 모델을 합친 Latent Geographical Topic Analysis (LGTA) 모델을 통하여 지역의 토픽을 발견하고 비교하였다. 이런 연구들을 통해 웹 상의 위치 관련 데이터들로부터 지역의 토픽을 발견할 수 있고 그 토픽들이 실제로 해당 지역을 잘 반영한다는 것을 확인할 수 있다. 그러나 언급한 방법들은 일정 지역의 GPS 연관 문서 데이터들로부터 토픽을 찾는데 그 결과 산, 해안 등과 같은 넓은 지역의 토픽을 나타내거나 지역별로 주로 먹는 음식 종류, 시 또는 주 단위의 축제 등과 같은 광역에 대응하는 토픽 일반을 주로 발견한다. 이런 토픽들은 사용자의 GPS에 맞춰 개인에 대한 서비스를 제공하기에는 포괄하는 범위가 너무 넓다.

본 논문에서는 현재 사용자가 위치한 장소로 발견할 토픽의 범위를 좁힌다. 장소의 토픽을 발견하기 위해 웹 상에 이미 존재하는 많은 관련 문서를 이용할 수 있다. 이들로부터 발견한 장소의 토픽을 기사 추천에 이용한다. 이를 통해 각 장소에 내재된 토픽이 그 곳을 방문한 사용자에게 개인화된 서비스를 제공하는데 큰 기여를 할 수 있음을 보인다.

3. 장소의 잠재적 토픽에 기반을 둔 기사 추천

그림 1은 사용자의 현재 위치로부터 장소의 토픽을 발견하고, 그와 관련된 기사를 추천하는 일련의 과정을 나타낸 것이다. 장소와 관련된 문서를 얻기 위해 구글 검색엔진을 사용하였으며, 검색 결과 중 스니펫을 사용하여 문서를 구성하였다. 얻어진 문서에서 토픽 모델을 이용하여 토픽을 찾고, 추천할 대상이 되는 기사들은 토픽 모델을 통해 미리 그 토픽들을 뽑아놓는다. 본 연구에서는 문서에서 토픽을 발견하기 위한 토픽 모델로 Explicit Semantic Analysis (ESA) [6]를 사용하였다. 장소의 토픽이 모델링 되면 장소와 기사들의 토픽 유사성을 측정하고 그 결과 장소의 토픽과 가장 유사도가 높게 나타난 기사들을 최종적으로 추천한다.

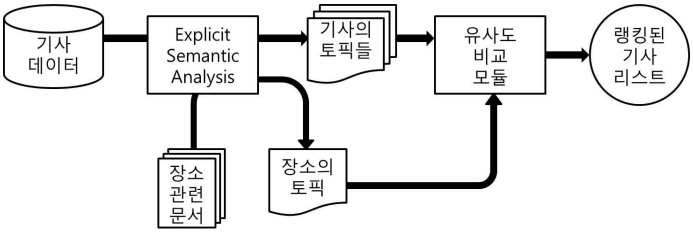

그림 1 장소의 토픽에 기반을 둔 기사 추천

3.1 장소에 관한 문서 구성

본 논문에서는 장소의 토픽을 얻기 위해 먼저 장소와 관련된 문서를 구성하고 그 문서의 토픽을 ESA를 통해 찾는다. 토픽을 알아내기 위해 사용될, 장소를 기술하는 문서를 위한 특별한 제약 조건은 없다. 웹 상에서 쉽게 장소와 관련된 문서를 구할 수 있는 방법에는 다음 두 가지를 생각할 수 있다.

- 위키피디아를 비롯한 백과사전은 장소에 대한 명시적인 정보를 체계적으로 기술하고 있다. 그러나 기술하고 있는 장소의 양이 부족하다.
- 장소를 키워드로, 검색엔진 검색을 통해 상대적으로 장소와 관련이 높은 문서를 얻을 수 있다. 대단히 많은 장소를 아우를 수 있지만, 검색엔진의 성능에 문서의 질이 결정된다.

본 논문에서는 장소를 기술하는 문서를 구글 검색 결과로 나오는 스니펫으로 구성한다. 검색엔진의 검색 결과를 사용함으로써 본래 장소가 가지고 있는 고유의 토픽 외에 시의성 있는 토픽을 추가로 기대할 수 있다. 구글은 문서 요약 기술을 사용하여 검색 결과 문서로부터 후보 요약문을 구성하고 그 중 최적의 후보 요약문을 스니펫으로 보여준다. 따라서 상위 검색 결과로 보이는 스니펫은 장소에 대한 핵심 키워드들을 포함하고 있으며, 이는 ESA로부터 발견될 토픽의 질을 높이는데 기여한다.

3.2 Explicit Semantic Analysis

ESA는 위키피디아 같은 대규모의 백과사전으로부터 생성된 고차원의 표제어 벡터로 토픽을 표현한다. 위키피디아에 기반한 ESA를 사용하는 경우, 문서는 위키피디아의 표제어 벡터로 표현된다. 벡터의 값은 문서에 나타나는 단어들과 위키피디아 표제어 사이의 연관성 정도를 나타낸다. 표제어와 단어 사이의 연관성 정도는 tf-idf 값을 사용한다.

그림 2는 문서의 토픽이 위키피디아에 기반한 ESA를

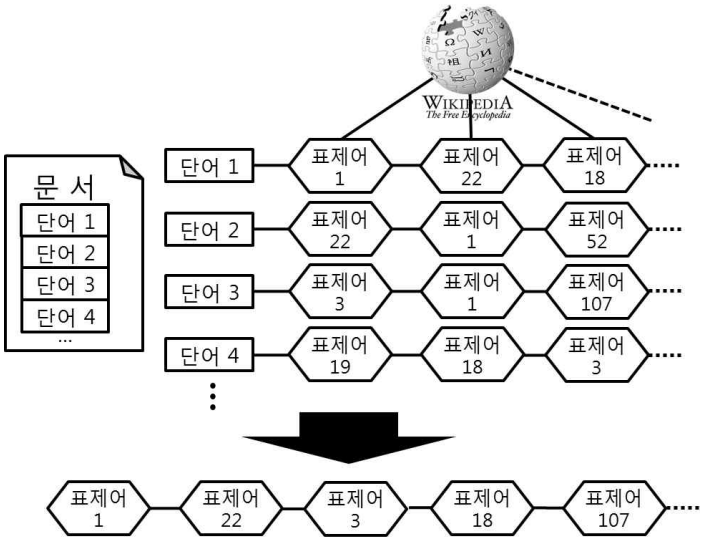

그림 2 문서로부터 ESA 토픽 모델 생성

통해 표제어 벡터로 표현되는 과정을 나타낸 것이다. 문서를 구성하는 각 단어들과 연관되는 표제어들을 모아 표제어 벡터를 만든다. 각 단어의 표제어 벡터들을 모두 더하면 문서의 표제어 벡터가 된다. 문서 $T=\{w_i\}$는 각 단어 w_i의 tf-idf 값을 가지는 벡터 $\langle v_i \rangle$로 표현할 수 있다. 그리고 위키피디아의 표제어 $e_j, \{e_j, \in e_1, ..., e_N\}$ (N은 위키피디아 전제 표제어 수)와 단어 w_i와의 연관성 정도를 k_j로 나타내면, 문서 T는 각각의 위키피디아 표제어 e_j가 $\sum_{w_i \in T} v_i \cdot k_j$의 가중치를 가지는 N 크기의 벡터 E_T로 표현된다. ESA는 토픽을 찾을 대상이 주어졌을 때 벡터를 실시간으로 구성할 수 있도록 위키피디아에 나타나는 단어들과 그 단어를 포함하는 표제어들 사이에 역색인을 구축해놓는다.

토픽 모델에는 ESA외에도 통계적 방법을 사용하는 PLSA [7]와 LDA [8]가 있다. 특히 LDA는 다양한 분야에서 토픽을 분석하는데 널리 응용되고 있다. 그럼에도 불구하고 본 연구에서 ESA를 사용한 이유는 다음과 같다.

- 사용자의 이동성을 고려해 장소의 토픽을 즉각적으로 찾아내야 한다.
- 미리 구축한 기사들의 토픽과 현재 장소의 토픽 간에 유사도 비교가 가능해야 한다.

위의 두 가지 제약 사항에 대해 ESA가 더 좋은 대처를 할 수 있다.

3.3 토픽 유사도에 기반한 기사 추천

본 논문에서는 장소를 기술하는 문서들을 수집하고 ESA를 통해 문서의 토픽을 위키피디아 표제어 벡터로 나타낸다. 이렇게 모델링된 장소의 토픽 벡터를 마찬가지 방법으로 미리 모델링된 기사들의 토픽 벡터들과 유사도 비교를 하고, 그 결과를 이용하여 추천될 기사를 선정한다.

먼저 장소 P에 관한 문서들을 $P=\{T_1, ... T_k\}$로 놓자. 여기서 T_k는 스니펫이 된다. 그러면 문서 T_k는 ESA 벡터 $E_{T_k}=\langle e_j \rangle$로 모델링 된다. 각각의 문서에 대해 토픽 벡터를 만들고, 이들을 합하면 장소 P의 토픽 E_P가 만들어진다.

$$E_P = \sum_{i=1}^{k} E_{T_i}$$

추천할 대상이 되는 전체 m개의 기사들은 각각 $E_{A_l}, \{A_l \in A_1, ..., A_m\}$로 토픽이 표현된다. 장소와 각 기사간의 토픽 유사도는 코사인 유사도를 사용하여 간단히 계산한다.

$$score(P, A_l) = \frac{E_P \cdot E_{A_l}}{\| E_P \| \| E_{A_l} \|} = \frac{\sum_{j=1}^{N} E_{P_j} \times E_{A_{l_j}}}{\sqrt{\sum_{j=1}^{N} (E_{P_j})^2} \times \sqrt{\sum_{j=1}^{N} (E_{A_{l_j}})^2}}$$

최종적으로 전체 기사들 중 $score(P, A_l)$가 높은 것들을 추천 기사를 선정한다.

4. 실험

4.1 실험 설정

본 논문에서는 추천할 대상이 되는 기사 데이터를 교육, 스포츠 등 7개 카테고리에서 총 39,033개의 웹 문서를 수집하여 사용하였다. 장소를 기술하는 문서와 기사 데이터의 토픽 모델링을 위한 ESA 구축에는 한국어 위키피디아 6월 8일자 덤프를 사용하였다. 한국어 위키피디아는 155,480개의 표제어로 이루어져 있는데, 이 중 50개 이하의 단어(명사)로 이루어진 짧은 표제어들을 추려내고 최종 65,031개의 표제어를 토픽 표현에 사용한다.

기사 추천 실험에서는 장소의 토픽을 고려치 않은 두 가지 비교 모델을 설정하였다. 첫 번째 비교 모델은 키워드 기반(keyword-based) 방법으로, 장소를 키워드로 하여 기사 검색을 하고 tf-idf 값 순으로 기사 추천 순위를 결정한다. 두 번째 비교 모델은 장소를 키워드로 먼저 구글 검색 결과의 스니펫을 얻고, 그로부터 추가로 키워드를 확장하여 추천할 기사를 검색하는 질의확장 기반(Query-expansion based) 방법 [9]이다.

기사 추천의 성능 평가는 정보 검색 분야에서 랭크된 리스트의 품질 평가를 위해 사용하는 Normalized Discounted Cumulated Gain (NDCG) [10]을 사용하였다. DCG는 질의와 높은 연관성을 가지는 문서일수록 더 유용하며 높은 연관성을 가지는 문서가 더 높은 순위의 검색결과로 나타나는 것이 좋다는 가정에서 출발하는 평가 방법이며 수식 (1)로 표현된다.

$$DCG_P = \sum_{i=1}^{P} \frac{2^{rel_i} - 1}{\log_2(1+i)} \quad (1)$$

표 1 장소의 스니펫 문서와 ESA 토픽 표현

세종문화회관	예술의전당
• 대극장 및 연간공연 일정, 관련 기사, 월간지 문화공간, 교향악단 등 예술단 소개. • 2011.08.23 ~ 2011.09.10 | 세종문화회관 대극장, |종료임박. 출연 : 박은태, 조정은, 양희경, ... 사미인곡. 2011.09.15 ~ 2011.09.16 | 세종문화회관 대극장, |개막예정 • 세종문화회관(世宗文化會館, Sejong Center for the Performing Arts)은 서울특별시 종로구 세종대로 175(세종로 81-3)에 위치한 53202m²크기의 종합예술시설이다. • 세종홀 일요일 예식 특전, 2011-08-30. 돌잔치 상차림(BLUE), 2011-08-24. • 세종문화회관 대극장 :3022석 세종문화회관 대극장 배치도 보기; 세종문화회관 ... 세종문화회관 체임버홀 :443석 세종문화회관 체임버홀 배치도 보기; 세종문화회관...	• 공연 및 전시일정, 예매일정, 공연장, 월간 정보지 등 소개. 회원할인, 예술의전당 유료회원 2천원 할인(골드 1인4매/블루 1인2매) / 20명 이 ... • 예술의전당 Seoul Arts Center ... 2011년 예술의전당 프로그램. 월간일정 ... 신 ... • 주 최 : 예술의전당, 프랑스국립오르세미술관 주관/협찬/후원 :주관 : 지엔씨미디 ... • 의정부예술의전당. 예술의전당소개. 공연/전시. 회원서비스. 문화예술아카데미. 대관안내. 열린광장. 이용안내. 의정부예술의전당소개; 인사말; 조직과기구; 윤리 경영 ... • 예술의전당은 클래식음악, 오페라, 연극, 무용 등 공연예술 전용극장을 갖추고 있으며 미술관, 서예박물관, 디자인 미술관 등 전시장까지 겸비한 복합문화예술 공간 ...
서울시 유스 오케스트라 마리오네트 (공연) 클럽 (공연) 원기범 (방송인) 부민관 진보라 서울시민회관 화재 사고 코리안 심포니 오케스트라 동방신기의 아카펠라 공연 신데렐라 (발레)	서애운수 부천 필하모닉 오케스트라 만년동 진보라 클럽 (공연) 마리오네트 (공연) 신데렐라 (발레) 동방신기의 아카펠라 공연 대전문화예술의전당 김지언 (피아니스트)

표 2. 장소의 토픽 간 유사도

	세종문화회관	예술의전당	청와대	국회의사당	경북대학교	서울대학교
세종문화회관		**0.7968**	0.1145	0.2968	0.0090	0.0234
예술의전당	**0.7968**		0.0968	0.0327	0.0099	0.1253
청와대	0.1145	0.0968		**0.4485**	0.0133	0.2615
국회의사당	0.2968	0.0327	**0.4485**		0.1788	0.3681
경북대학교	0.0090	0.0099	0.0133	0.1788		**0.6119**
서울대학교	0.0234	0.1253	0.2615	0.3681	**0.6119**	

여기서 rel_i은 검색 시스템이 도출한 i번째 문서와 질의와의 연관성 정도를 나타낸다. 본 연구에서는 DCG 평가를 위해 장소와 기사의 관련성 정도에 대한 기준을 다음과 같이 정하였다.

- 3점 – 현재 장소와 정확히 부합하면서 장소의 토픽과 관련된 기사
- 2점 - 현재 장소에 정확히 맞지는 않더라도, 장소의 토픽과 관련성이 높은 기사.
- 1점 - 장소의 토픽과 어느 정도의 관련성이 있는 기사
- 0점 - 장소의 토픽과는 관련이 없는 기사

각 문서들은 낮은 검색 순위에 위치할수록 연관성 점수에 페널티를 받게 된다. 랭크 위치 P에 대해 이상적인 DCG_P를 나타낸 것이 $IDCG_P$이다. 예를 들어, 검색엔진의 결과로 나온 문서 상위 5개가 {3, 0, 1, 3, 2}의 연관성 점수를 받았다면 $IDCG_P$는 {3, 3, 2, 1, 0}을 기준으로 DCG_P를 계산한 것이다. $NDCG_P$는 수식 (2)와 같이 $IDCG_P$로 DCG_P를 정규화한 것으로 최고값은 1이 된다.

$$NDCG_p = \frac{DCG_p}{IDCG_p} \quad (2)$$

4.2 실험 결과

기사 추천에 대한 성능 평가 실험에 앞서 ESA를 통해 표현된 장소 토픽의 특성을 살펴보았다. 문서로부터 찾아낸 토픽이 실제 장소의 토픽을 반영하는 지를 확인하기 위한 실험이다. 이를 위해 추출된 토픽의 형태를 살펴보고 유사한 장소와 그렇지 않은 장소 간의 토픽 유사도를 비교하였다.

표 1은 구글을 통해 얻은 세종문화회관과 예술의전당에 관한 스니펫 문서들과 그 문서의 토픽을 ESA를 통해

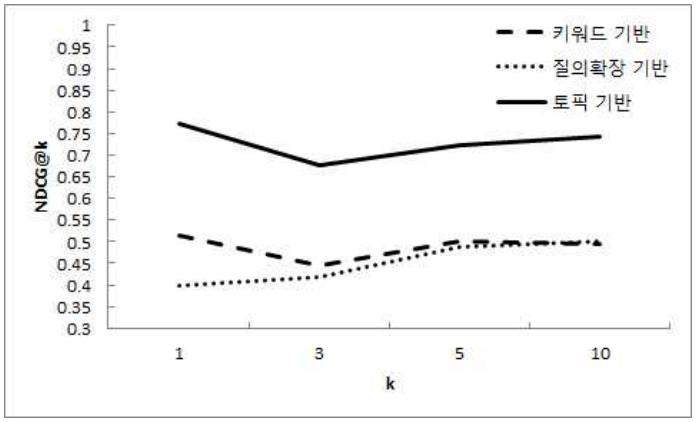

그림 3 모델 간의 기사 추천 성능 비교

위키피디아 표제어 벡터로 나타낸 것이다. 스니펫은 검색 결과 중 상위 5개, 표제어 벡터는 전체 벡터 중 상위 10개의 표제어를 보인 것이다. 두 장소의 스니펫들은 주소 정보와 같은 장소의 위치 정보는 물론 가수, 오페라, 클래식, 문화공간 등 장소에 내재된 토픽과 깊은 연관이 있는 단어들로 문서가 이루어진 것을 확인할 수 있다. 한편, 예술의전당의 표제어 벡터를 보면 실제 토픽과 상관없어 보이는 표제어들이 포함되어있다. 그러나 이 표제어들을 살펴보면, 서애운수-예술의전당으로 가는 버스 노선 운행, 만년동-대전문화예술의전당이 위치한 곳으로 어느 정도 예술의전당과 연관된 표제어임을 확인할 수 있다. 결국 두 벡터 모두 해당 장소의 토픽에 부합하는 표제어들로 채워졌으며, 두 장소의 공통된 표제어들을 통해 공연/문화 등과 연관된 토픽을 두 장소가 공유하는 것을 확인할 수 있다.

표 2는 유사한 두 장소를 한 쌍으로 하여 3쌍의 장소를 선정하고, 이들 간의 ESA 토픽 유사도를 비교한 것이다. 역시 사람들에게 비슷하게 인식되는 장소 간에 토픽 유사도가 높은 것을 알 수 있다. 그리고 이런 특성은 장소의 물리적 위치에 제약받지 않음을 경북대학교와 서울대학교 및 다른 서울 지역의 장소 유사도 간의 차이로 확인할 수 있다. 이에 따라 경북대학교와 서울대학교에 있는 사용자들에게 각 대학교와 관련된 기사뿐만 아니라 등록금 문제 등과 같은 대학교에 전반에 관련된 기사를 추천하는 것이 가능하다.

그림 3은 앞서 언급한 두 가지 비교 모델과 장소의 토픽을 기반으로 한 모델의 기사 추천 성능을 비교한 결과이다. 8개의 장소에 대해 각 모델들이 추천한 기사의 상위 1, 3, 5, 10개의 기사에 대한 NDCG 측정 결과를 평균낸 것이다. 토픽 기반 모델과 키워드 기반 모델은 평균적으로 약 0.24의 NDCG 값 차이를 보이고 있고, 토픽 기반 모델과 질의확장 기반 모델 간에는 평균 약 0.28 정도의 차이를 보인다. 토픽 기반 추천의 경우 모델이 높은 순위로 결과를 낸 기사들이 관련성 정도에서 고루 2점 이상의 점수를 받은 기사들로 채워졌다. 반면에 키워드와 질의확장 기반 모델의 경우 기사에 포함된 키워드가 장소에 정확히 일치하는 기사(3점)를 찾는데 결정적인 역할을 하지만, 또한 기사에 장소가 언급되지만 장소의 토픽과 실질적으로 상관관계가 없는 기사(0점)도 많아 전체적으로 점수가 떨어졌다. 이런 결과는 제안한 방법이 다른 두 모델보다 추가적인 분석을 거쳐 뽑은 토픽을 사용하기 때문에 가능한 것이다.

표 3. 잠실야구장에 대한 기사 추천

키워드 기반	질의확장 기반	토픽 기반
양궁 대표팀, AG 대비 잠실야구장서 소음 적응훈련	양궁 대표팀, AG 대비 잠실야구장서 소음 적응훈련	서울열전 17탄 잠실야구장
서울열전 17탄 잠실야구장	서울열전 17탄 잠실야구장	준PO를 앞두고 살펴본 롯데 가을야구의 역사
양궁대표팀, 2010년 광저우 AG 대비 특별훈련	KBO 야구발전실행위원회, '야구장 건립 매뉴얼' 발행	KBO, 2011년 프로야구 경기일정 발표
1982년 제27회 세계야구선수권 한·일 결승전	야구의 계절이 찾아오기도 전에 맛보는 야구, WBC	만화 같았던 롯데, LG 연장 11회 대혈전
야구의 계절이 찾아오기도 전에 맛보는 야구, WBC	이대호, 세계야구史를 바꿨다	오빠, 4번타자 등번호가 왜 4번이 아니야?

마지막으로 표 3은 잠실야구장에 대해 세 모델이 추천한 기사의 상위 5개 목록이다. 여기서 눈 여겨 볼 점은 키워드 기반 추천과 질의확장 기반 추천의 경우 잠실야구장에서 소음 적응 훈련을 하는 양궁 국가 대표에 관한 기사가 상위에 올라있다는 것이다. 이 기사는 잠실야구장을 언급하지만 실질적으로는 양궁에 관련된 기사로, 야구장을 찾은 야구팬에게는 비교적 관심사가 되지 못하는 기사이다. 반면 토픽 기반의 경우에는 양궁 관련 기사를 배제하고, 잠실야구장과 직접적으로 연관된 기사는 물론 야구에 대한 전반적인 토픽을 아우르는 기사들로 추천이 되었음을 확인할 수 있다.

5. 결론

본 논문에서는 사용자가 머물고 있는 장소의 토픽을 고려하여 기사 추천하는 방법을 제안하였다. 장소와 관련된 문서를 검색 엔진을 통해 구성하고 ESA 토픽 모델을 통해 그 문서의 토픽을 표현하였다. 이렇게 발견된 장소에 내재된 토픽과 기사들의 토픽 비교를 통해 기사를 추천하였다. 제안한 방법으로부터 토픽을 구성하는 표제어들이 실제 장소의 토픽과 관련되어있고 비슷한 장소 간에 토픽 유사도가 높게 나오는 결과를 통해 토픽이 실제 각 장소가 가지는 특성을 잘 반영하고 있음을 확인하였다. 또한 기사 추천 성능을 비교해본 결과, 토픽을 고려치 않은 방법과 의미 있는 성능 차이를 보임으로써 장소에 내재된 토픽이 유용하게 활용될 수 있음을 보였다.

향후에는 트위터[4]와 같이 일반 웹 문서와는 관점이 다른 문서로부터 발견한 장소의 토픽을 기준으로 하여

4) http://twitter.com

본 연구에서 제안한 방법이 추천한 기사의 순위를 재배치(re-ranking)하는 방법을 적용하고자 한다. 트위터에서는 그 장소의 지역적 특색이 강한 토픽 또는 글을 남기는 사람 개인의 토픽이 이야기되며 때로는 전 사회적 토픽이 이슈가 되기도 한다. 이런 것들을 현재 그 장소를 대변하는 토픽으로 보고, 이를 기준으로 장소 고유의 토픽으로 추천된 기사들을 재평가하는 것은 또 다른 의미 있는 결과를 보일 것이다.

Acknowledgement

본 논문은 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발)의 지원으로 수행되었음.

참고문헌

[1] Y. Takeuchi, and M. Sugimoto. Cityvoyager: an outdoor recommendation system based on user location history. In *Proceedings of UIC*, pp. 625-636, 2006.

[2] V. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang. Collaborative Filtering Meets Mobile Recommendation: A User-centered Approach. In *Proceedings of AAAI*, pp. 236-241, 2010.

[3] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining User Similarity Based on Location History. In *Proceedings of GIS*, pp. 298-307, 2008.

[4] S. Sizov. GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media. In *Proceedings of WSDM*, pp. 281-290, 2010.

[5] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical Topic Discovery and Comparison. In *Proceedings of WWW*, pp. 247-256, 2011.

[6] E. Gabrilovich, and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of IJCAI*, pp. 1606-1611, 2007.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*, pp. 50-57, 1999.

[8] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3. pp. 993-1022, 2003.

[9] G. Salton, and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, Vol. 41(4). pp. 299-297, 1990.

[10] K. Jarvelin, and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, Vol. 20(4). pp. 422-446, 2002.