위치 정보 블로그 문서의 공간 통합과 정확도 분석에 관한 연구[1)]

A research on the spatial integration and the trust analysis of location information blog documents

백성하 · 김진아 · 김경배 · 배해영
Sungha Baek · Jinah Kim · Gyoungbae Kim · Haeyoung Bae
인하대학교 정보공학과 · 서원대 컴퓨터교육과
shbaek@dblab.inha.ac.kr · gbkim@seowon.ac.kr · {jakim · hybae}@inha.ac.kr

요 약

본 논문은 위치 정보 검색 서비스에서 POI에 등록된 업소의 정보 열람 시 연관된 블로그 문서를 보다 정확히 제공하고 해당 문서의 신뢰도를 분석하기 위한 기법을 제시한다. 이 기법은 업소명위주로 문서를 통합하여 연관되지 않은 문서를 제공하는 문제를 해결하기 위해 POI에 등록된 업소명뿐 아니라 주소 정보를 추가적으로 이용한다. 또한 시간에 지남에 따라 정보가 변질되는 위치 정보 블로그 문서의 신뢰도를 제공하지 않아 문서의 정확성을 판단 할 수 없는 문제를 해결하기 위해 문서가 포스팅된 시간 정보를 이용하여 문서의 신뢰도를 분석한다.

1. 서론

웹 2.0은 정보의 소유자나 독점자 없이 누구나 손쉽게 정보를 생산하고 인터넷에서 공유할 수 있도록 한 사용자 참여 중심의 인터넷 환경이다. 블로그는 웹 2.0의 하나로 사용자들이 다양한 분야의 정보를 생산하고 공유하고 있고, 최근에는 디지털 카메라, 휴대폰 카메라의 보급으로 맛집 정보와 같이 특정 장소를 소개하는 "위치 정보 블로그" 문서가 포스팅되고 있다.

최근 웹 2.0이 위치 정보 검색 서비스에 도입되고 있다. 맵을 기반으로 한 위치 정보검색 서비스는 지도상의 주요위치(POI: Point of Interest)를 설정하고, 이 위치에 해당 업소(음식점, 병원등)의 정보를 제공하고 있다. 그러나 이 정보는 업소에 대한 기본적인 설명만 포함하고 있어, 정보 이용자가 업소의 구체적인 정보를 획득할 수 없다. 따라서 해당 POI의 업소 정보에 사용자가 직접 정보를 등록하거나 포스팅된 위치 정보 블로그를 연결하는 방법을 사용하고 있다.

네이버 윙버스, 네이버 지도, 다음 플레이스, 구글맵 모두 유사한 방식으로 서비스를 제공하고 있다[1,2,3,4]. 네이버 윙버스와 다음 플레이스는 설정된 POI에 업소 평점과 리뷰를 등록할 수 있는 인터페이스를 제공하고, 자신이 생성한 위치 정보 블로그의 URL이나 RSS를 이용하여 이를 연결 할 수 있다. 그러나 이 방법은 사용자가 직접 블로그를 연결하지 않으면 기존의 방대한 양의 위치 정보 블로그를 검색에 이용 할 수 없다. 네이버 지도는 이와는 다르게 해당 POI의 업소명을 이용하여 연관된 블로그를 자동으로 검색하여 검색된 블로그 문서를 연결한다. 그러나 이 방법은 업소명이 "아파치"나 "장충

1) 본 연구는 건설교통부 첨단도시기술개발사업 - 지능형국토정보기술혁신 사업과제의 연구비지원(07국토정보C05)에 의해 수행되었습니다.

동 왕족발” 처럼 일반적인 용어이거나 지역명을 포함하는 경우 해당 업소와 관련이 없는 문서들이 검색될 수 있다.

또한 이 방식은 모두 저장된 업소가 영업을 중단한 경우 이를 즉시 반영하지 못하여 서비스 이용자에게 잘못된 정보를 제공 할 수 있다. 이 문제를 해결하기 위해서 지도 서비스는 사용자가 직접 오류를 신고 할 수 있다[5]. 그러나 사용자가 직접 신고를 해야 하고 신고를 하여도 지도에 저장된 정보에만 반영이 되고, 연관된 모든 블로그에는 반영되지 않는다.

정보의 신뢰도 분석을 위해 영향력 있는 파워유저를 파악하는 연구가 진행 되었다[6,7,8,9]. 파워유저는 신뢰도 높은 질 좋은 정보를 생산하는 검증된 사용자로, 정보의 신뢰도를 판단하는 기준으로 사용될 수 있다. 그러나 이 방법은 용어의 정의와 같이 변하지 않는 정보의 신뢰성 판단에는 적합하나, 정보가 변질될 수 있는 장소와 연관된 정보의 판단에는 부적합하다. 해당 업소가 사라지는 경우는 파워유저도 알 수 없다. 따라서 위치와 연관된 정보의 신뢰도의 분석은 새로운 판단 기준이 필요하다.

본 논문에서는 위치 정보 블로그 문서를 POI를 기반으로 자동으로 통합하고, 통합된 문서들의 포스팅 시간과 관련 글 및 오류 신고 정보를 이용하여 정확도를 분석하는 방법을 제안한다. 이 방법은 블로그 연결망에 존재하는 문서들에 포함된 위치와 관련된 정보(제목, 주소, 좌표, 전화번호등)를 이용하여 위치 정보 블로그 문서만 선별하여 해당 POI에 해당하는 블로그들을 그룹화 하여 통합한다. 그리고 통합된 그룹의 정확도 분석을 위해서 블로그 문서가 포스팅된 마지막 시간과 포스팅 되는 문서들의 주기를 이용한다. 일반적으로 마지막으로 포스팅된 시간이 최근이 경우 정확한 정보일 가능성이 높다. 또한 평균 1달 간격으로 포스팅되던 문서가 최근 몇 개월간 포스팅되지 않았다면 부정확한 정보일 가능성이 높다. 그밖에 관련 글에서 해당 업소의 영업 중지 정보를 얻은 경우 이를 이용할 수 있다. 본 제안 기법은 다음과 같은 장점을 갖는다.

• POI와 연관된 블로그 문서의 통합

기존에 존재하는 위치 정보 블로그를 활용하여 지도의 POI와 연결을 지원한다. 통합 시 공간 인덱스를 사용하여 업소명을 통한 검색 보다 빠르게 검색이 가능하다.

• 위치 정보를 이용한 보다 정확한 통합

업소명과 위치 정보를 모두 이용하여 정확한 연관 문서들의 통합이 가능하다.

• 지역정보의 정확성의 판별

시간에 따라 변하는 지역 정보에 대한 정확성이 분석 가능하다.

• 오류신고의 연쇄적인 반영

그룹의 특정 문서에서 오류 신고를 통해 그룹의 모든 문서에 적용 가능하다.

2. 관련연구

2.1 연관 블로그 문서의 통합 기법

현재 위치정보 검색 서비스는 POI를 지도 상에 표시하고 이 POI의 기본적인 정보를 제공한다. 또한 해당 POI의 업소명과 같은 기본 정보를 이용하여 연관된 블로그 문서를 검색하여 제공한다.

그림 1. 위치정보 검색 서비스

[그림1]은 아파치라는 업소명에 해당하는 연관된 문서를 검색하여 제공한다. 그러나 아파치는 헬기이름으로 더욱 많이

사용되기 때문에 해당 업소와 관계없는 문서가 제공되는 문제가 있다.

다른 방법으로 RSS나 블로그의 URL을 이용하여 사용자가 직접 연관 블로그 문서를 해당 업소와 연결하는 방법이 있다. 그러나 이 방법은 사용자가 직접 문서를 등록해야 하는 어려움과 등록되지 않은 수많은 문서들을 이용하지 못하는 한계가 있다.

또 다른 통합 기법으로 블로그 문서에서 인용한 문장의 존재 유무를 이용하여 중복된 문서를 통합하는 기법이 있다[10,11,12]. 그러나 위치 정보 문서는 개인적인 의견과 사진을 중심으로 구성되기 때문에 이 방법은 적합하지 않다.

2.2 블로그 문서의 신뢰도 판단 기법

블로그는 사용자들이 자유롭게 컨텐츠를 생산하기 때문에 정보의 신뢰도가 중요하다. 신뢰도를 판단하기 위한 대부분의 연구는 파워 유저의 파악 방법이다[9]. 파워유저는 신뢰성 있는 질 좋은 정보를 생산하는 검증된 사용자로, 블로그에 포스팅된 정보의 신뢰성을 판단하는 기준으로 사용될 수 있다.

파워 유저를 파악하는 기존 연구는 사회 연결망에서 해당 사용자가 중앙에 위치한 정도를 측정 하는 방법들이 제안되었다. 즉 많은 이웃 관계를 갖고 있는 사용자가 일반적으로 영향력 있다는 파워유저라고 판단하는 방법이다[6,7]. 또한 컨텐츠로 인한 영향력을 분석하여 파워유저를 판단하는 방법도 제안되었다[8].

그러나 이 기법은 시간에 따라 변할 수 있는 지역정보의 정확성을 판단하기 위한 방법으로는 어렵다. 질 좋은 컨텐츠를 등록하는 파워 유저도 해당 업소가 사라지는 것을 알 수 없기 때문이다.

또한 사용자로부터 정보의 오류를 신고받는 방법이 있다. 블로그 스피어는 사용자의 신고를 통해 오류를 정정하는 방법을 제안하였고, 현재 다음 플레이스와 네이버는 우리 동네지도 오차제로 프로젝트를 서비스하고 있다[5]. 그러나 문서의 신뢰도가 제공되지 않고 사용자의 신고 이전에는 문서의 오류를 판단할 수 없고, 신고를 받은 문서에만 오류가 적용이 되고 관련된 다른 문서에는 적용되지 않는다.

3. 제안기법

본 장에서는 제안기법인 위치 정보 블로그 문서를 POI를 중심으로 통합하고 문서의 신뢰도를 관리하는 기법을 설명한다.

3.1 위치 정보 블로그 통합 기법

위치 정보 블로그 문서의 통합기법은 두 단계로 구성된다. [그림2]은 이 과정을 설명한다.

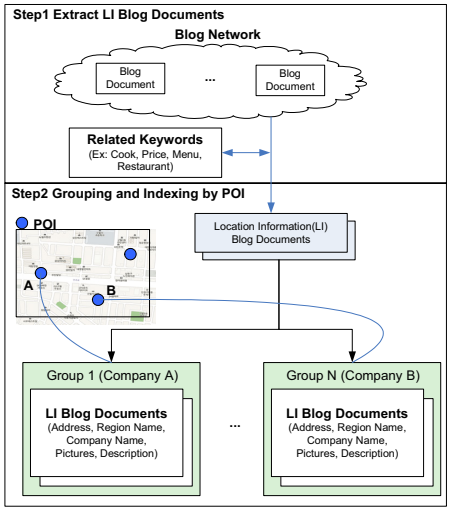

그림 2. 위치 정보 블로그의 통합 과정

1단계에서는 블로그 연결망에 존재하는 다양한 문서들에서 위치 정보 블로그 문서를 추출한다. 추출방법은 위치 정보 블로그 문서는 일반적으로 주소와 연관된 단어(Region Keyword)와 업소의 특성에 관련된 단어(Related Keyword)를 포함하기 때문에, 이 단어들을 포함하는 문서들을 추출한다.

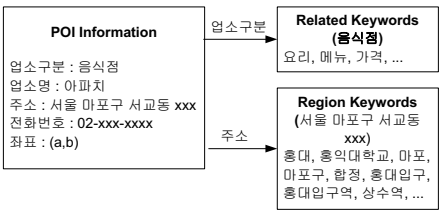

그림 3. POI Information과 Keyword 생성

주소와 연관된 단어는 주로 제목에 포함된 지역명(홍대)이나 업소명과 문서에 포함된 업소의 주소, 전화번호, 지역명, 업소명을 이용할 수 있다. 업소의 특성에 관련된 단어는 동종 업소 소개에 주로 사용되는 표현을 이용할 수 있다. 예를들면 식당은 음식, 메뉴, 가격과 같은 표현이 주로 사용된다.

2단계에서는 1단계에서 추출된 지역정보 블로그 문서에 포함된 주소와 연관된 단어(주소, 업소명, 지역명)를 이용하여 지도상의 POI를 중심으로 이에 해당하는 문서들을 그룹화 하고 빠른 검색을 위해 인덱스를 구성한다. 구축방법은 추출한 위치 정보 블로그 문서가 포함하고 있는 주소정보를 POI의 업소 정보(POI InformatioN)와 비교하여 해당되는 문서를 그룹화 한다. 비교는 제목, 주요 지역 정보(주소, 전화번호), 기타 정보를 가중치를 두어 비교한다.

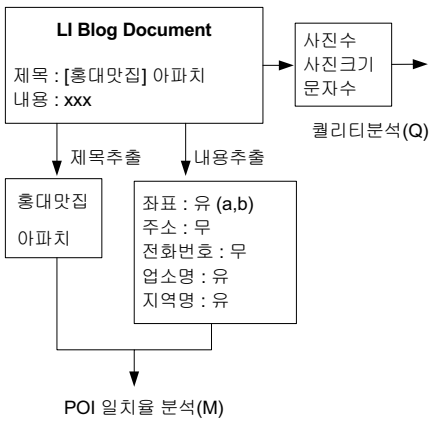

그림 4. POI 중심의 그룹화 과정

[그림4]는 이방법을 설명하고 있다. 추출된 위치 정보 블로그 문서의 제목에서 주소와 연관된 단어(홍대맛집)와 업소명(아파치)가 추출되어서 POI Information에서 아파치라는 업소와 유사함을 1차적으로 판단하다. 그리고 내용을 추출하여 해당 내용에서 좌표, 주소, 전화번호, 업소명, 지역명의 포함여부를 파악하여 해당 POI와의 일치율을 분석한다. 이 일치율은 검색 결과를 제공할 때 우선순위를 반영하여 상위에 가장 일치하는 블로그를 제공하기 위해서 사용한다. 또한 이 분석과정에서 등록된 사진수, 사진크기와 문자수를 분석하여 퀄리티를 분석한다. 이는 보다 자세히 설명한 블로그 문서를 최상위에 제공하기 위해 분석한다.

3.2 정확도 분석 기법

정확도 분석은 그룹화된 문서의 포스팅 시간과 포스팅 주기 변화, 답글, 오류 신고정보를 이용한다. 그룹화된 문서의 마지막 포스팅 시간은 해당 지역 정보가 유효한 마지막 시점이다. 또한 포스팅 주기는 문서가 포스팅되는 평균 주기와 최대 주기를 이용해서 포스팅의 중지를 판단한다. 답글에서 발견되는 오류신고와 오류신고는 그룹화된 문서에서 발생한 경우 오류의 사실을 판단하고 그룹화된 모든 문서에 적용한다.

[그림5]은 하나의 그룹에 대한 정보이다. N 개의 문서가 있고, L_d,M_d,α 3개의 상수가 주어진다. L_d은 가장 마지막에 포스팅된 문서의 시간이다. M_d는 시간 순서로 포스팅된 문서간의 가장 큰 시간차이다. α는 포스팅된 시간 차이를 정확도에 반영하기 위해 사용자가 임의로 지정한 가중치 상수이다.

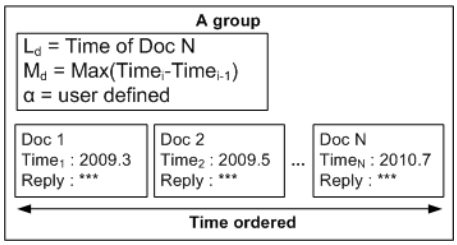

그림 5. 그룹 정보

$$acc = 100 - (C_d - L_d) \times \alpha \;,\; M_d \le (C_d - L_d)$$
$$acc = 100 - \frac{(C_d - L_d) \times \alpha}{(M_d - (C_d - L_d))} \;,\; otherwise$$

식 1. 지역정보 블로그의 정확도

[식1]은 지역정보 블로그의 정확도(acc)를 임의의 그룹 내 블로그 정보를 열람하는 시점(C_d)에 제공한다. 이 식은 "시간차 $(C_d - L_d)$"를 이용하여 마지막으로 포스팅된 시점이 열람 시점보다 오래 될수록 정확도를 낮게 계산한다. 그러나 포스팅이 잦은 지역과 아닌 지역 모두 시간차만 이용하는 방법은 적절하지 않다. 왜냐하면 6개월에 한번 포스팅 되는 문서가 3개월이 지난 경우의 정확도와, 1개월에 한번 꾸준히 포스팅되는 문서가 3개월이 지난 경우는 엄연히 다르다. 따라서 보다 정확한 정확도 분석을 위해 포스팅 간격(M_d)을 이용하여 더욱 정확한 정확도를 얻을 수 있다.

그 이외에 정확도를 위한 방법은 문서들의 답글(Reply)이 등록되면 답글의 문장에서 지역정보의 오류에 대한 단어를 포함하는지 확인하고 오류 가능 문서로 설정한다. 그리고 관리자가 최종 오류로 판명이 되면 해당 그룹의 모든 문서에 오류 문서임을 반영한다.

4. 성능분석

본 장에서는 POI에 위치 정보 블로그를 연결하는 정확성을 평가한다. 기존 자동 블로그 검색은 업소명만 이용해서 연관 블로그를 검색하기 때문에 검색어의 특징에 따라 다른 결과가 발생한다. 이를 비교 분석하기 위해 검색어는 업소명의 특징에 따라 3가지로 분류하여 실험한다. 첫 번째는 업소명이 일반적으로 사용되는 다른 의미의 용어를 사용하는 경우이다(일반적 업소명). 예를 들면 아파치라는 주점은 아파치 헬기라는 고유명사와 중복된다. 두 번째는 업소명의 혼하거나 일부가 지역명을 포함하는 경우이다(지역적 업소명). 예를 들면 마포 숯불갈비는 마포 이외의 지역에도 두루 존재한다. 세 번째는 업소명이 일반적으로 사용하지 않는 비교적 고유한 용어를 사용하는 경우이다(고유한 업소명).

위 실험을 위해 포털사이트에서 3가지 특징에 따라 검색을 하여 나오는 결과 블로그 중 해당 POI에 해당하는 문서의 비율을 계산하고, 제안기법을 이용하영 경우의 비율을 계산한다. 실험 결과는 다음과 같다. 검색어는 각 특징별로 3가지를 사용하였다. 일반적 용어는 아파치, 솔로몬, 포도나무를 사용하였고, 지역적 용어는 마포 숯불갈비, 장충동 왕족발, 춘천 닭갈비를 사용하였다. 고유한 용어는 서야 왕족발, 조폭 떡볶이, 겐지를 사용하였다. 검색결과는 상위 10개의 결과를 이용한다.

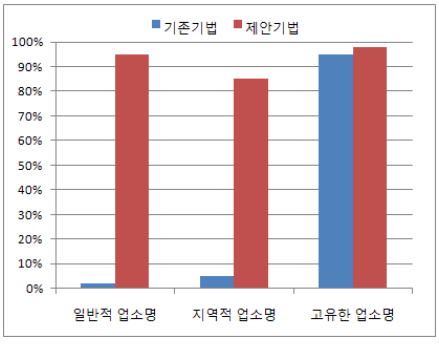

그림 6. 업소명 특징별 검색 정확성

[그림6]은 업소명을 특징별로 검색한 실험 결과이다. 기존 기법은 고유한 업소명 이외에는 해당하는 POI의 업소 블로그를

정상적으로 검색하지 못하지만, 제안기법은 높은 정확도로 검색이 가능하다. 이는 검색 방법이 업소명에만 의존하지 않고 다른 지역에 관한 정보를 검색에 추가로 이용하기 때문이다.

5. 결론 및 향후 연구

본 논문에서는 POI의 정보와 블로그 문서에 포함된 주소 정보를 이용하여 POI 업소와 정확히 연관된 블로그 문서를 추출하여 그룹화 하는 방법과 이 문서들의 정확도를 분석하는 방법을 제안하였다. 이 기법은 업소명만 이용하는 기존 기법보다 정확하게 연관된 블로그를 추출할 있다. 또한 문서의 신뢰도를 제공하여 위치 정보 검색 이용자들이 보다 질 좋은 검색 서비스를 제공 받을 수 있다.

향후 연구로는 POI를 중심으로 하는 통합은 POI에 포함되지 못한 수많은 업소 정보를 제공하지 못하는 문제가 있다. 그래서 그룹화 과정에서 POI에 등록되지 않은 업소에 대한 그룹을 생성하여 POI를 추가하는 방법에 대한 연구가 필요하다.

참고문헌

[1] (주)NHN, 윙버스 서울 맛집, http://r.wingbus.com/seoul

[2] (주)NHN, 네이버 지도, http://map.naver.com

[3] (주)Daum, 다음 플레이스, http://place.daum.net

[4] (주)Google, 구글 지도 http://maps.google.co.kr

[5] (주)NHN, 우리 동네 지도 오차 제로 프로젝트, http://maps.google.co.kr

[6] X. Song et al., “Mining in Social Networks Information Flow Modeling based on Diffusion Rate for Prediction and Ranking,“ In Proc. Int’l. Conf. on World Wide Web, pp.191-200, 2007

[7] J. Iribarren and E. Moro, “Information Diusion Epidemics in Social Networks,“ Arxiv, 0706.0641, 2007.

[8] 임승환 외., “블로그 연결망 활성화를 위한 컨텐츠 파워 유저의 파악 방안”, 한국정보과학회논문지 데이터베이스 제36권 제6호, 2009

[9] A. Nitin et al., “Blogosphere: Research Issues, Tools, and Applications”, SIGKDD, Vol10, Issue 1, 2008

[10] M. Henzinger, “Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms,“ In Proc. ACM Int’l. Conf. on Information Retrieval, SIGIR, pp.284-291, 2006

[11] Jong Wook Kim, K. Selcuk Candan, and Junichi Tatemura, “Efficient Overlap and Content Reuse Detection in Blogs and Online News Articles,“ In Proc. Int’l. World Wide Web Conference, WWW, pp.81-90, 2009

[12] 이상철 외., “블로그 서비스 시스템을 위한 효과적인 중복문서의 검출 기법”, 정보과학회논문지 데이터베이스 제37권 제1호, 2010