지오웹: 위치 정보를 포함한 웹의 중요성

박종세○ 김재홍 서상원 허재혁 맹승렬
한국과학기술원
{jspark, jaehong, swseo, maeng} @camars.kaist.ac.kr and {jhuh} @ cs.kaist.ac.kr
이문상
삼성전자
sang0627.lee@samsung.com

GeoWeb: The impacts of Geo-tagged Web Page

Jongse Park○ Jaehong Kim Sangwon Seo Jaehyuk Huh and Seungryoul Maeng
KAIST
Moonsang Lee
Samsung Electronics

1. 서 론

최근 인터넷 기술과 디지털 기기의 발달에 힘입어, 개인이 카메라를 이용하여 촬영한 사진들을 인터넷 사진 공유 사이트나 자신의 블로그(blog) 등에 올리고, 이를 공유하고 있다.

구글이나 마이크로소프트 등과 같은 인터넷 서비스 업체들이 사용자의 사진을 단순 인터넷 게시가 아니라, 촬영 위치와 관련된 정보를 입력하게끔 함으로써, 새로운 서비스 영역으로의 확장을 꾀하고 있는 것이 현 시점 이다. 이렇게 위치 정보를 입력하는 과정을 지오태깅(Geo Tagging)이라고 한다 [1, 2]

이렇게 위치 정보가 태깅된 정보를 활용하여, 지도를 보면서 실제 그 위치에서 다른 사람들에 의해서 촬영된 사진을 통해 미리 여행 정보를 얻는 등 다양한 정보를 획득하고 있다. 이와 같이 몇몇 대표적인 인터넷 서비스 업체가 위치와 관련한 사진 서비스를 보급함으로써, 지오태깅에 대한 관심이 증대 하고 있다. 예를 들면, Flickr [1] 라는 웹사이트에서는 사용자가 직접 위치 정보를 가진 사진을 업로드 할 수 있고 위치 정보에 의해서 지역별로 사진이 등록 된다. 그리고 Flickr는 사용자에게 지도 검색 등을 통해서 지역과 관련된 사진을 열람 할 수 있는 서비스를 제공 한다. 또한 2007년부터 서비스를 시작한 Google StreetView [2]를 활용하면 사용자는 지도 검색과 동시에 해당 지역의 주변 풍경 등을 담고 있는 이미지를 함께 열람 할 수 있다.

하지만, 기존의 Flickr나 Google StreetView는 위치 정보를 담고 있는 사진을 해당 서버에 직접 업로드 해야만 서비스가 가능하다는 한계가 있고, 이로 인해 다양한 정보를 제공하지 못하는 문제를 가진다. 만일 검색엔진에서 검색을 요청한 사용자의 현재 위치 정보를 검색에 활용할 수 있고 또한 지오태깅 된 사진을 포함한 웹 페이지들을 검색 결과에 이용할 수 있다면, 지오태깅 된 사진들을 능동적으로 검색해서 보다 다양하고 가치 있는 정보를 제공 할 수 있을 것이다.

지오태깅 정보는 앞서 설명 한 바와 같이 사용자가 수동으로 입력할 수 있지만, GPS를 활용하여 사진에 태깅하는 것도 가능하다. 특히 최근에 널리 보급하기 시작한 스마트폰 내부에 탑재되어 있는 GPS모듈을 이용하여 사진을 찍을 때마다 자동으로 지오태깅을 할 수 있다. 최근에 GPS모듈을 탑재한 디지털 기기 및 스마트폰의 광범위한 보급으로 일반 사용자들이 쉽게 지오태깅 된 사진을 찍고, 인터넷 블로그 혹은 웹 공간에 자신의 사진을 올리고 있다.

이렇게 지오태깅 된 사진을 포함하는 웹 페이지를 지오웹(GeoWeb) 페이지라고 한다. 웹 공간에 존재하는 지오웹 페이지들은 새로운 서비스에서 사용될 수 있는 유용한 자원 들이다. 예를 들어, 기존의 검색엔진의 경우 단순히 키워드 위주의 검색만이 가능했던 것에 반해 지오웹 페이지의 위치 정보를 이용하게 되면 자신이 원하는 지역과 키워드를 함께 이용한 검색이 가능하다. 또한, 여러 웹 페이지의 위치 정보를 파악하여 검색을 요청한 특정 사용자의 위치와 근접한 위치정보를 담고 있는 페이지를 검색결과로 제공해주는 검색서비스를 생각할 수 있다. 이 외에도 지오웹 페이지에 대한 수많은 응용을 생각할 수 있다는 점에서, 웹이 담고 있는 새로운 정보인 지오태깅 정보에 대한 가치를 높게 평가할 수 있다.

본 논문에서 우리는 지오웹 페이지의 증가 추세를 보여 주었으며, 그것의 의미가 점차 커지고 있음을 밝혀내었다. 또한 우리는 웹에서 지오웹 페이지를 크롤링하는 효율적이고 간단한 방법을 제시하였다.

2. 본 론

검색 서비스를 제공하기 위해 웹으로부터 페이지들을 끌어 모으는 작업을 크롤링이라고 한다. 하지만, 모든 페이지를 무작위로 크롤링하는 것이 아니라 지오웹 페이지와 같이 특별한 성격을 갖는 페이지를 효율적으로 크롤링하기 위해서는 종전의 방법과는 다른 기술적인 방법을 필요하다. 이를 위해 우리는 오픈 소스로 제공되고 있는 검색 엔진 Nutch [3]를 사용하여, Nutch에서 기본적으로 제공되고 있는 크롤러(crawler)에 세 가지 모듈을 추가하였다.

컨텐츠 필터(ContentsFilter), 태크 추출기(TagExtractor) 그리고 어댑티브 인젝터(AdaptiveInjector)가 그것이다. 컨텐츠 필터는 지오태깅 이미지가 아닐 가능성이 높은 이미지를 필터링하는 역할을 한다. 메타데이터에 GPS태그가 없는 포맷의 이미지, 작은 크기의 이미지, 아이콘이나 버튼 등의 이미지 등이 이에 해당한다. 태그 추출기는 컨텐츠 필터로부터 필터링 된 이미지에서 GPS태그를 추출해내는 역할을 한다. 마지막으로 어댑티브 인젝터는 지오웹 페이지의 지역성을 반영하기 위해서 크롤링의 시드(seed) URL을 동적으로 지오웹 페이지의 URL로 바꾸면서 반복적으로 크롤링을 수행할 수 있게 한다. 이를 통해 기존 Nutch에서 시드 URL을 고정한 상태로 크롤링을 하는 단점을 극복할 수 있다. 세 가지 모듈을 추가함으로 인해 Nutch의 기본 크롤러에 비해 40배 이상 많은 지오웹 페이지를 찾을 수 있었다.

전체 웹에서 지오웹이 차지하는 비중을 측정하기란 쉬운 일이 아니다. 위에서 제안한 효과적인 크롤러를 사용했지만 제한된 환경에서 지오웹 페이지의 증가 추세를 밝히는 것은 불가능했다. 그러나 현재 Flickr [1]에서 제공하는 자료에 따르면 지오태깅 된 사진의 숫자가 기하급수적으로 늘고 있다는 사실을 간접적으로 보여준다. 2006년 서비스가 시작된 Flickr는 2008년 당시 지오태깅 된 사진의 수가 4,000만개를 넘어섰고 2010년 현재 지오태깅 된 사진의 수는 1억개를 넘어서고 있다. 또한 GPS모듈이 탑재된 스마트폰의 판매가 급격히 증가하고 있다는 점에서 웹에서의 지오태깅 된 사진의 수는 현 시점보다도 더욱 많아질 것으로 추정된다.

지오웹은 향후에 여러 가지 목적을 위해 응용될 수 있다. 한 가지 예로 검색 엔진에서 사용자에게 좀 더 만족도 높은 결과를 돌려주기 위해 지오웹을 사용할 수 있다. 만약 어떤 사용자가 위치 정보와 관련된 정보를 검색하려 하고 검색 엔진이 이 사용자가 원하는 지오웹 페이지를 결과로 돌려줄 수 있다면 보다 개선된 서비스를 제공할 수 있을 것이다. 이를 위해서는 기존의 페이지랭크(PageRank)와는 다른 새로운 랭킹 알고리즘을 필요로 한다. 기존의 페이지랭크는 모든 페이지에 같은 초기 값을 부여하고 페이지 간의 인링크와 아웃링크를 분석하여 많은 페이지로부터 인링크를 갖고 있고 또 랭크가 높은 페이지로부터 많은 아웃링크를 받은 페이지일수록 랭크를 높게 준다. 그러나 지오웹 페이지는 일반적인 웹 페이지에 비해 위치 정보에 대한 보다 가치 있는 정보를 담고 있다. 지오웹 페이지의 랭크를 일반웹 페이지의 랭크보다 더 높여줄 수 있다면 사용자에게 만족도 높은 결과를 줄 수 있을 것이다. 이를 위해 페이지랭크에서 모든 페이지에게 일정하게 부여했던 초기 값을 지오웹 페이지의 경우에는 더 높게 주는 방법을 제안할 수 있다. 이 때 지오웹 페이지가 갖고 있는 위치 정보의 가치에 따라 다른 초기 값을 줄 수 있는데, 이러한 가치를 결정하는 요소에는 여러 가지가 있을 수 있다. 간단한 예를 들어, 지오웹 페이지 안에 지오태깅 된 사진의 개수가 많거나 지오태깅 된 사진의 크기가 큰 페이지일수록 초기 값을 더욱 많이 주는 방법을 생각해 볼 수 있을 것이다. 지오웹 페이지의 가치를 결정하는 방법을 연구하는 것은 향후 과제가 될 수 있을 것이다.

지오웹의 숫자가 기하급수적으로 늘어나고 있는 추세이기는 하지만 아직까지 지오웹이 전체 웹에서 차지하고 있는 비중이 그다지 크지 않다는 점을 보완하기 위하여 임시적으로 지오웹을 확장하는 방법론을 필요로 한다. 단순하게 생각할 수 있는 방법으로는 기존 지오웹 페이지가 가진 위치정보를 토대로 하여 지오웹 페이지가 아니면서 해당되는 위치의 정보를 갖고 있는 페이지에 지오웹 페이지의 위치정보를 전파(propagate)하는 방법을 생각할 수 있다. 이러한 방식은 근본적으로 정확한 결과를 만들어내기 어렵기 때문에 임시적인 방법으로 사용되어야 할 것이다.

3. 결 론

우리는 기존에 고려되지 않았던 지오태깅 된 사진을 포함하는 웹 페이지의 중요성을 인식했다. 그리고 그러한 특징을 갖는 웹을 통틀어 지오웹이라고 불렀다. 이러한 지오웹 페이지의 숫자는 증가 추세에 있지만 현재 전체 웹에서 지오웹이 차지하는 비중은 크지 않다. 우리는 여기서 지오웹 페이지를 효과적으로 크롤링하는 방법을 제시했다. 이를 바탕으로 전 세계에서의 지오웹 페이지 분포 및 비율을 통계적으로 제시하며 의미를 분석했다. 향후 과제로 지오웹 페이지를 실제로 검색 엔진에서 결과로 보여주기 위해 기존 페이지랭크 알고리즘을 수정하는 새로운 방식을 제시하였다. 마지막으로, 현재 절대적으로 많이 부족한 지오웹 페이지의 숫자를 보완하기 위한 방법으로 지오태깅 위치 정보를 지오웹 페이지가 아닌 페이지에 전파하는 방법을 제시하였다.

마지막으로 본 연구를 진행하는 데 있어서 여러 가지 도움을 주신 삼성전자 연구원 여러분들게 감사의 뜻을 표합니다.

[1]Flickr: http://www.flickr.com/map
[2]Street View: http://maps.google.com/intl/en_us/help/maps/streetview
[3]Nutch: http://nutch.apache.org