인터넷 사용 패턴 분석을 통한 인터넷 LBS상에서의 고성능 위치 검색 기법 설계 및 구현

김민경° 조민정* 류옥현**
KT컨버전스본부 유비쿼터스개발담당*, 한국산업기술대학교 e-비즈니스학과**
{kimminky°, mini}@kt.co.kr*, {ok-ryou}@kpu.ac.kr**

High Performance Location Query Method based on Access Pattern Analysis

MinKyung Kim°*, MinJeung Cho*, OkHyun Ryou**
KT Convergence Business Unit*, Korea Polytechnic University, e-Business Department**

요 약

웹을 이용하는 사용자들은 통상적으로 짧은 시간 내에 포탈의 여러 페이지를 방문하는 현상이 있는데, 이 때 방문하는 각 페이지 상에 위치 정보를 필요로 하는 콘텐츠(위치 기반 광고 배너, 위치 기반 날씨 등)가 존재한다면 웹의 특성상 짧은 시간 내에 동일 IP에 대한 위치 정보 검색을 반복하게 된다. 본 논문에서는 이러한 웹 사용 패턴을 캐시를 통해 반영한 고성능 인터넷 위치 검색을 제안, 구현하고 그 성능을 검증한 것이다. 이는 초당 2-3천 건 이상의 대용량 위치 정보를 검색을 수행하데 특히 적합한 방법으로, 적은 비용으로 위치 검색 성능을 획기적으로 높일 수 있었다.

1. 서론

접속 단말의 위치를 활용하는 서비스, 즉 LBS는 이동 통신 영역에서 시작되어 현재에는 인터넷 영역에서도 서비스가 이루어지기 시작하였다. 특히, 접속 단말의 유동 IP와 접속 위치를 실시간으로 매핑하는 기술이 최근 등장함에 따라[1] 서비스 활성화가 가능하였는데, Yahoo의 '거기'나 KT의 bizmeka Near(위치기반 배너, 검색), Naver의 '근처검색' 등이 바로 그것이다.

하지만, 이동 통신 LBS가 'GPS를 이용한 차량 안내' 등과 같은 서비스에서 볼 수 있듯이, 개별 app를 통하여 단독으로 사용되는 것과는 달리, 현재의 인터넷 LBS는 주로 배너 광고 또는 위치 기반 검색 등의 형태로 웹 페이지에 포함되어 서비스되고 있다. 그 이유는 인터넷 LBS는 단말의 위치가 거의 변하지 않기 때문에, 단말의 이동을 지속적/주기적으로 추적하여 정보를 제공하는 동적인 서비스 보다는, 미리 단말의 위치를 파악하거나 접속 당시 실시간으로 그 위치를 파악한 후, 그 위치에 맞는 여러 정보를 생성하여 제공하는 정적인 서비스가 적합하기 때문이며, 또한, 이러한 정적인 정보는 현재 포탈이 가장 잘 서비스 할 수 있기 때문이다.

그렇지만, 포탈에 서비스되는 만큼, 포탈 동시 사용자가 많으면 많을수록, 하나의 웹 페이지에 위치 콘텐츠가 많으면 많을수록, 위치 콘텐츠를 포함하는 여러 페이지를 사용자가 방문하면 방문할수록 위치 정보 검색의 수가 많아지고, 따라서 이를 처리하는 LBS 위치 검색 시스템은 높은 성능을 가져야 한다는 문제가 발생한다. 특히, 대형 포탈의 루트 페이지에 위치 정보 콘텐츠가 있는 경우에는 하루 수천/수억 건의 Page가 노출되므로 이 페이지에 대해 위치정보를 제공할 수 있는 대용량 위치 정보 검색 기법 및 시스템이 반드시 필요하다.

본 논문은 대용량의 위치 정보에 대한 고속 검색 기법을 제안한 것이다. 본 논문의 2장에서는 위치 정보 콘텐츠를 포함한 웹 페이지에 대한 사용자 접근 패턴을 실제 예를 통하여 기술하였으며, 3장은 웹 접근 패턴을 캐시 기법을 통해 반영한 검색 시스템의 구현을 기술하였다. 또한, 4장에서는 모의실험을 통해 검색 시스템의 성능을 검증하였고, 5장에서는 결론을 기술하였다.

2. 사용자의 웹 접근 패턴 분석

웹 페이지에 대한 사용자 접근 패턴은 사용자 PC에 감시 클라이언트를 설치하기 이전에는 방문하는 모든 페이지를 감시하기가 매우 어렵다. 하지만, 위치 정보 콘텐츠를 사용하는 패턴은 위치 정보 콘텐츠가 노출되기 전에 반드시 인터넷 LBS에 위치를 검색하기 때문에 단위 시간동안 특정 IP에서 얼마만큼의 위치 정보 콘텐츠를 사용하는지를 접속 로그를 통해 쉽게 알 수 있다.

다음 [표1]과 [표2]는 국내 대표적인 미디어 렙 중 하나인 A사의 배너 광고 송출에 대한 로그와 네이버 '근처 검색'의 로그를 실제 예로 사용하여 사용자 웹 접근 패턴을 추출한 것이다.

단위시간	총 방문IP수	Unique IP수	중복 IP 비율
30초	3,127	2,970	5.02%
1분	6,206	5,870	5.41%
2분	12,505	10,102	19.22%
3분	18,808	14,981	20.35%
4분	25,109	20,103	19.94%
5분	31,426	25,435	19.06%

[표 1] A사의 배너 송출로그(2006/03/21 16시-16시05분)

단위시간	총 방문 IP수	Unique IP수	중복 IP 비율
30초	22,740	12,680	44.24%
1분	45,264	18,148	59.91%
2분	91,074	25,002	72.55%
3분	137,227	30,852	77.52%
4분	183,295	36,175	80.26%
5분	229,078	41,051	82.08%

[표 2]네이버 '근처검색' 위치 정보 접근 로그(2006/3/21 15시)

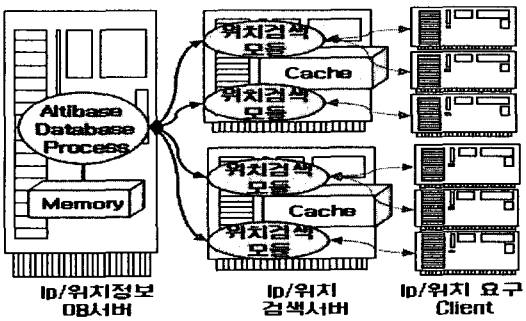
[표1]과 [표2]에서 볼 수 있듯이 단위 시간(30초에서 5분)동안 동일 IP에서 동일 위치 정보를 요구하는 비율이 배너의 경우 44~82%내외이고, 검색의 경우 5~20% 정도임을 알 수 있다.

3. 웹 사용 패턴을 반영한 위치 정보 검색 시스템 구현

이 장에서는 2장에서 제시한 사용자의 웹 접근 패턴을 캐시를 통해 반영한 위치 검색 시스템을 구현한 것으로, 인터넷 LBS 플랫폼 전체 구성을 나타낸 3.1절과, 캐시 구조를 나타낸 3.2절, 3.3절에서는 캐시의 정확성을 유지하면서도 성능을 유지할 수 있는 캐시 지속 시간 및 캐시 전체 크기에 대한 분석을 기술하였다.

3.1 인터넷 LBS 플랫폼 구성

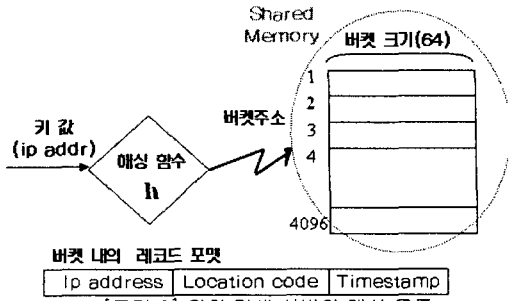
인터넷 LBS위치 검색 시스템은 위치 검색시스템과 위치 저장 시스템으로 나뉘어져 있으며, 대용량/고속 위치정보 검색을 위해 메모리 DB를 사용한다[2]. 그림[1]은 위치 정보 플랫폼을 나타낸 것으로 IP/위치 검색 서버에 사용자의 웹 접속 패턴을 반영하는 Cache를 사용하여 검색 성능을 높이고, 메모리 DB의 접근을 줄여 비용을 절약하는 효과를 얻도록 설계되었다[2].



[그림 1] 인터넷 LBS 플랫폼 구성

3.2 캐시 구조

웹 접근 패턴을 반영하여 검색 속도를 높이기 위한 캐시는 [그림 2]와 같은 구조를 가진다.



[그림 2] 위치 검색 서버의 캐시 구조

[그림 2]에서 볼 수 있듯이 캐시는 IP에 대한 버켓 해싱을 사용하는데, 전체 버켓은 4096개로 구성되고, 64개로 이루어진 각 버켓에는 과거 접근한 IP와 위치 코드, 접근 시간으로 이루어진 레코드들이 저장된다.

한편, 캐시의 동일 버켓에 있는 레코드들은 IP를 key값으로 정렬되어 있으며, 데이터 입력이나 검색을 위해서는 binary search를 사용한다. 또한 버켓이 overflow가 생겼을 경우에는 시스템에서 미리 정한 캐시 지속 시간보다 큰 것은 모두 지워버림으로써 overflow를 해결하고, 만일 캐시 지속 시간 이후의 것이 없는 경우 가장 오래된 레코드를 지우게 된다.

3.3 최적 캐시 지속 시간 및 캐시 크기

인터넷 LBS위치 검색 시스템에서 캐시의 사용은 높은 성능을 낼 수 있는 반면, 사용자의 위치에 대한 부정확한 정보를 제공할 수 있는 단점을 가진다. 따라서 성능은 높으면서도, 부정확한 정보를 제공하지 않도록 캐시의 정보를 원래 위치 정보 DB와 일치시키려는 노력이 필요하다.

캐시의 일관성을 위해서는 LRU 등 여러 가지 알고리즘을 이용되어 왔다[3]. 하지만 본 논문에서 기존 알고리즘들과는 달리 단순히 캐시 지속 시간만을 사용하는데, 그 이유는 다음과 같다.

1) 인터넷 LBS의 위치 정보는 유동 IP사용자의 네트워크 접속 기록을 실시간 수집하여 주소 정보와 결합하여 생성하는데 [1], 유동 IP 사용자가 접속을 한 후 1-5분 이내에 네트워크 접속을 끊을 확률이 매우 적다[표3 참조].
2) 캐시의 데이터가 잘못될 경우는 특정 IP에 대한 사용자가 위치 정보 콘텐츠를 사용하다가, 네트워크를 끊고, 그 IP가 다시 다른 사람에게 할당되어, 그 사람이 위치 정보 콘텐츠가 있는 site를 1-5분 이내에 방문해야 하는데, 그 확률이 매우 적다[표3 참조].

단위시간	총IP수	Uniq IP수	중복IP비율	보정계수	가능성
30초	1294	1278	1.24%	0.05	0.06%
1분	1357	1330	1.99%	0.1	0.20%
2분	2651	2589	2.34%	0.2	0.47%
3분	3959	3855	2.63%	0.3	0.79%
4분	5240	5073	3.19%	0.5	1.59%
5분	6534	6108	6.52%	0.6	3.91%

[표 3]유동IP 가입자 네트워크 인증 기록(2006/03/21 16시~)

[표 3]은 메가패스 유동 가입자의 네트워크 인증 기록을 바탕으로 1), 2)번의 이유를 설명한 것이다. 보정 계수는 2)번 내용을 기반으로 설정한 것으로, 특정 시간에 대하여 중복 IP에 기록에 대해 분석한 후, 2번의 내용만 추출한 비율을 적용한 것이다.

따라서 [표 3]에서 알 수 있듯이 캐시의 지속 시간은 정확도 99%를 만족하기 위해선 3분 이내가 가장 좋은 것을 알 수 있으며, [표1]과 같이 동일한 IP가 중복 되어 검색된다면 충분히 성능을 증가시킬 가능성이 있음을 알 수 있다.

또한, 캐시 데이터의 양은 3분 내에 중복된 IP가 하나도 없는 경우의 양에 대해, 그 정보의 50%를 캐시 할 수 있는 정도(4096*64)로 정했는데, 그 이유는 위치 검색 단위 서버 당 1초에 약 2,000건 이상의 요청을 처리하는 것으로 설계되었기 때문이다.(4096*64 ≒2000*180초 * 50%)

4. 성능 분석

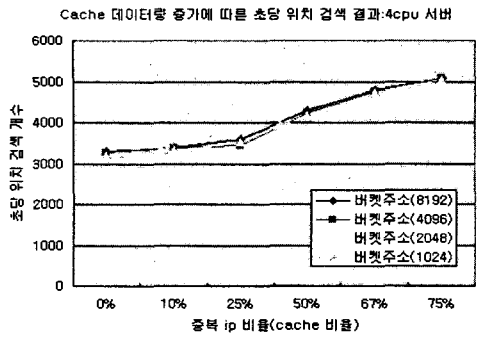
이 장은 사용자의 웹 사용 패턴을 cache로 반영한 인터넷 위치 검색 서버의 성능 분석을 기술한 것이다.

4.1 성능 분석 방법

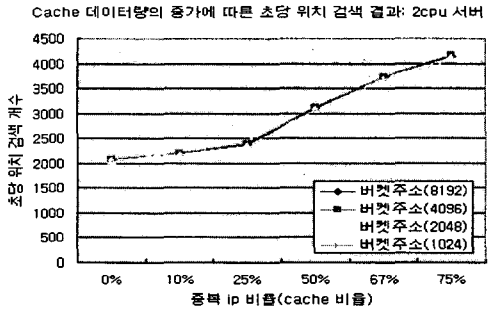
성능 분석은 50만개의 IP에 대해 각각 0%, 10%, 25%, 50%, 67%, 75%의 중복이 되도록 파일을 만든 다음, 이를 여러 대의 서버에 나눈 후, 'Microsoft Application Center Test Tool[4]을 사용하여 가능한 최대 속도로 위치 정보 Query를 요청하도록 하여 각각의 결과를 살펴보았다. 또한, 위치 검색 서버는 각각 2cpu와 4cpu가 장착된 리눅스 서버를 사용하였으며, 캐시 레코드의 총 량(4096*64)은 그대로 둔 채 버켓 주소의 크기를 1024, 2048, 4096, 8192로 각각 나누어 실험을 수행하였다.

4.2 성능 분석 결과

[그림 3]과 [그림4]는 2cpu와 4cpu의 위치 검색 서버에서 중복 IP 비율(캐시 비율)에 따라 위치 검색을 수행한 결과이다. [그림3]과 [그림4]에서 볼 수 있듯이 캐시 지속 시간(3분) 이내에 위치 검색 IP가 중복되면 버켓 주소의 크기가 4096인 경우, 캐시를 적용하지 않았을 때보다 2cpu 서버에서는, 각각 중복 IP 비율에 따라 6%, 16%, 51%, 89, 100%가 증가하였으며, 4cpu 서버 에서는 3%, 5%, 29%, 45%, 54%의 성능 증가가 이루어지는 것을 알 수 있었다. 이 결과는 초기 설계치인 초당 2,000개보다 높은 결과를 나타낸 것으로, cpu가 적은 저 사양의 서버일수록 캐시 사용에 대한 효과가 높은 것을 알 수 있었다. 또한, 버켓 주소의 크기를 1024에서 8192까지 변화시키며 실험을 했지만 의미 있는 성능의 변화를 발견할 수 없었다. 이는 버켓 크기가 매우 큰 경우(256개 이상)를 제외하고는 버켓 검색에 사용되는 binary search는 한두 번 더해도 별 문제가 없는 것으로 파악되었다.



[그림 3] cache 데이터양 증가에 따른 위치 검색결과:2cpu



[그림 4] cache 데이터양 증가에 따른 위치 검색결과:4cpu

5. 결론

본 논문에서는 사용자의 웹 접근 패턴을 파악하여 이를 캐싱함으로써, 대용량의 고속 위치 검색을 수행할 수 있는 인터넷 LBS 상에서의 위치 검색 서버의 설계와 구현을 기술하였다. 위치 기반 광고 및 위치 기반 검색과 같이 포탈의 웹 페이지에 포함되어 서비스되는 위치 콘텐츠들은 단위 시간동안 매우 중복되게 위치 검색을 요구하므로, 동일 IP로부터의 중복 위치 요구 결과를 캐시에 저장하고, 이를 검색에 활용함으로써 높은 검색 성능을 얻을 수 있었다. 또한, 고 비용의 메모리 DB에 대한 직접 검색을 줄임으로써 시스템 구축 및 운영 비용 절감을 꾀할 수 있었다.

추후, 위치 검색 요구량에 따라 능동적으로 캐시의 양과 시간을 변경시킬 수 있는 기법에 대한 연구가 진행될 것이다.

[참고문헌]

[1] 김민경, 백규태, "초고속 인터넷상에서 위치기반 서비스를 위한 실시간 IP/위치 매핑 시스템 구현", 한국정보통신설비학회 학술대회 논문집, pp. 10-15, 2005년 8월

[2] 김민경, 변성원, 김석우, "초고속 인터넷 IP 기반 LBS 플랫폼 구현", KT R&Dzine, Vol, 2, 2005년 9월

[3] 반효경, "CDN 서비스를 위한 웹 캐싱 기법", 정보과학회지 제20권 제9호, 2002년 9월

[4] Microsoft Application Center Test 1.0, Visual Studio .NET Edition "http://msdn.microsoft.com/library/default.asp?url=/library/en-us/act/htm/actml_main.asp"