모바일 환경에서 다중 속성 검색을 위한 시그너처 기반의 인덱싱 기법

박성근[0], 정성원
서강대학교 컴퓨터학과

sgpark@sogang.ac.kr[0], jungsung@sogang.ac.kr

Signature-based Indexing Scheme for Multi-attribute Retrieval in Mobile Environments

Sunggeun Park, Sungwon Jung
Dept. of Computer Science, Sogang University

요 약

모바일 환경에서 효과적인 데이터 전송 방법인 브로드 캐스트 기법에서 중요한 문제 중의 하나가 데이터에 대한 인덱스 생성이다. 데이터에 대한 인덱스가 제공되면 클라이언트는 튜닝 타임과 엑세스 타임을 줄일 수 있고, 그와 함께 배터리 소모도 줄일 수 있다. 기존에 제시된 인덱스 생성 기법은 대부분 트리 구조를 기반으로 하고 있다. 트리 기반 인덱싱 기법은 튜닝 타임을 최소화하지만, 반면 멀티-어트리뷰트(multi-attribute)에 대한 엑세스나 다양한 종류의 멀티미디어 데이터들 혹은 클러스터링 된 데이터에 대한 인덱스 생성이 어렵다. 이러한 문제를 해결하기 위해 시그너처 기반의 인덱싱 기법이 제시되었다. 그러나 기존의 시그너처 기반 인덱싱 기법에서는 엑세스 타임이 전체 브로드 캐스트 타임으로 고정되는 문제가 있었다. 본 논문에서는 앞으로 브로드 캐스팅 될 데이터들에 대한 포괄적인 정보를 가지는 시그너처 집합을 인덱스로 제공해서 클라이언트의 엑세스 타임을 최소화시키는 시그너처 스킴을 제시한다.

1. 서론

무선 네트워크 기술과 컴퓨터 하드웨어의 비약적인 발전으로 인해서 모바일 컴퓨팅 분야는 빠르게 발전하고 있고, 그에 따라 많은 모바일 기기들이 널리 보급되고 있다. 모바일 단말기들은 상대적으로 적은 네트워크 대역폭, 연결의 잦은 단절, 저 전력 등으로 인해 어려움에 직면하게 되고, 기존의 유선 환경의 정적인 기기들에 적용되었던 기술과는 다른 새로운 기술이 요구된다[1,2,3].

이와 관련된 연구 중 하나로 서버가 다수의 클라이언트들이 필요로 하는 데이터를 무선 네트워크를 통해 브로드 캐스트(broadcast)하는 방법이 활발히 연구되고 있다. 이는 서버가 다수의 클라이언트들이 원하는 데이터를 구성하여 주기적으로 브로드 캐스팅하는 방법이다[4]. 이 경우 클라이언트의 수가 증가하더라도 서버에서는 추가적인 부하가 생기지 않음으로 확장성을 가지게 되고, 클라이언트는 상향 대역폭을 통해서 데이터를 요청함으로 인해서 발생되는 네트워크 대역폭과 배터리의 소모를 줄일 수 있다.

이러한 브로드 캐스트 기법은 인덱스와 결합할 때 효과적으로 사용될 수 있다. 인덱스를 통해 클라이언트가 원하는 데이터가 언제 브로드 캐스팅 되는 지 알 수 있으면 클라이언트는 그 사이에 대기 모드(doze mode)상태로 네트워크 대역폭과 배터리를 절약할 수 있다. 이를 위한 다양한 인덱스 생성 기법이 제시되었는데, 그것들 중 상당수는 트리를 기반으로 하고 있다. 이러한 트리 기반 인덱스는 튜닝 타임을 최소화하는 반면에 멀티-어트리뷰트에 대한 엑세스, 다양한 타입의 데이터, 클러스터링 된 데이터 등을 고려한 인덱스 생성이 어렵다. 시그너처를 이용한 인덱스 생성 기법은 이러한 문제를 한번에 해결해준다[8]. 하지만 기존의 시그너처 기반 기법은 클라이언트가 브로드캐스트 사이클 내의 모든 시그너처를 다 확인해야하고, 그에 따라 평균 엑세스 타임이 브로드 캐스트 사이클의 길이로 고정되는 문제가 발생하였다.

본 논문에서는 시그너처를 이용하여 위에서 언급한 문제들을 해결하면서 클라이언트의 엑세스 타임을 최소화시킬 수 있는 펄스펙티브 시그너처 스킴(perspective signature scheme)이라는 인덱싱 기법을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 인덱스 생성 기법의 문제점을 제시하고, 시그너처를 이용한 기법의 장점과 관련 기존 연구를 간략히 소개한다. 그리고 3장에서는 제안하는 기법에 대해서 기술하고, 4장에서는 제안한 기법의 성능을 분석하고, 마지막으로 5장에서는 결론을 기술한다.

2. 관련 연구

2.1. 트리 기반 인덱스의 단점

브로드 캐스트 데이터를 위해 기존에 제시된 인덱스 스킴들[5,6,7]의 대다수는 B+ 트리, 알파베티컬 호프만 트리(alphabetical hoffman tree)등의 트리 구조를 기반으로 하고 있다. 이러한 트리 기반의 인덱싱 기법은 디스크 기반에서는 노드와 노드 사이의 임의 접근이 가능하므로 좋은 성능을 보이지만, 브로드 캐스트 기반에서는 순차 접근만 가능하므로 그 장점을 잃어버린다. 또한 멀티-어트리뷰트를 엑세스 하는 클라이언트나 이미 클러스터링 된 데이터 혹은 이미지 데이터나 멀티미디어 데이터와 같은 다양한 데이터 타입을 고려하기 어렵다.

2.2. 기존의 시그너처 기반의 인덱스 생성 기법

이러한 이유로 모바일 환경에서는 시그너처를 이용한 인덱스 생성기법이 효과적일 수 있다[8]. 시그너처는 생성과 사용 방법이 매우 단순하다. 그래서 상대적으로 자원이 빈약한 클라이언트에서도 쉽게 사용할 수 있다. 그리고 데이터에서 비해서 상대적으로 아주 작은 크기이며, 텍스트 데이터, 이미지 데이터, 멀티미디어 데이터 등 다양한 데이터 타입에 적용될 수 있다. 뿐만 아니라 멀티-어트리뷰트에 대한 인덱스를 생성하는 것도 가능하다.

[8]에서는 시그너처를 이용한 기존의 세 가지 인덱스 생성 기법을 소개하고

있다. 첫 번째로 심플 시그너처 스킴(simple signature scheme)은 브로드 캐스트 사이클 내의 각각의 정보 프레임 앞에 해당 정보 프레임에 대한 시그너처를 두는 기법이다. 두 번째로 인티그레이티드 시그너처 스킴(integrated signature scheme)은 브로드 캐스트 사이클 내의 정보 프레임을 프레임 그룹으로 나누고, 각각의 그룹에 대한 통합 시그너처만을 제공하는 기법이다. 마지막으로 멀티-레벨 시그너처 스킴(multi-level signature scheme)은 위의 두 스킴의 혼합형으로 인티그레이티드 시그너처 스킴에서 정보 프레임에 대한 튜닝 타임 증가 문제를 개개 정보 프레임 앞에도 추가적인 시그너처를 할당함으로써 해결하는 기법이다.

위의 스킴 중 멀티-레벨 시그너처 스킴이 일반적인 상황에서 좋은 성능을 보인다. 하지만 세 스킴 모두 엑세스 타임이 전체 브로드 캐스트 타임으로 고정되어 있는 문제를 가진다. 본 논문에서 제시하는 스킴은 멀티-레벨 스킴과 같이 평균 튜닝 타임을 줄이면서, 평균 엑세스 타임도 최소화한다. 구체적인 스킴은 다음 장에서 기술한다.

3. 펄스펙티브 시그너처 스킴

펄스펙티브 시그너처 스킴은 위에서 언급한 것과 같이 클라이언트의 엑세스 타임을 최소화하는 인덱스를 시그너처 기반으로 생성한다.

3.1 데이터에 대한 시그너처 생성

시그너처는 데이터가 가진 정보에 대한 요약이고, 시그너처를 확인함으로써 데이터가 원하는 정보를 가지고 있는 지 평가할 수 있다. 레코드에 대한 시그너처는 일반적으로 해싱(hashing)을 통해서 생성한다. 하나의 레코드가 있을 때, 인덱싱 하고자 하는 어트리뷰트의 값을 각각 임의의 비트 스트링으로 해싱한 후, 생성된 비트 스트링들 전부를 OR 연산(∨)해서 S_R을 생성한다. 클라이언트는 원하는 질의에 대한 시그너처 S_Q를 같은 방법으로 생성한다. 실제 필터링 시에 $S_Q \wedge S_R = S_Q$ 이면, 레코드가 시그너처와 매치된 것으로 실제 데이터를 수신한다. 실제 데이터는 사용자가 원하는 레코드일 수도 있고, 아닐 수도 있다. 후자의 경우를 폴스 드랍(false drop)이라고 한다. $S_Q \wedge S_R \neq S_Q$ 이면, 레코드는 시그너처와 매치되지 않았으므로 데이터를 수신할 필요가 없다.

레코드	홍길동	24	서울

홍길동 → 1010 0000, 서울 → 0010 1100, 24 → 1110 1000
김철수 → 0011 0101, 대구 → 1110 0100, (→ = hashing)

레코드의 시그너처 = 1010 0000 ∨ 0010 1100 ∨ 1110 1000
= 1110 1100

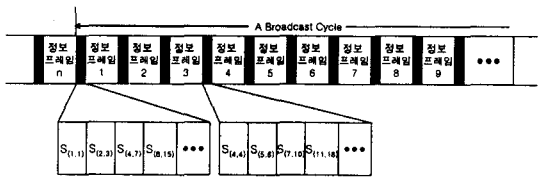
질의어 1) 서울 0010 1100 → match
2) 김철수 0011 0101 → not match
3) 대구 1110 0100 → false drop

그림 1: 시그너처의 생성과 필터링

3.2 시그너처를 이용한 인덱스 생성

클라이언트의 엑세스 타임을 감소시키도록 인덱스를 생성하는 기본 아이디어는 각각의 정보 프레임 앞에 향후에 브로드 캐스팅 될 데이터에 대한 포괄적인 정보를 제공하는 시그너처들의 그룹을 두는 것이다. 그때 그룹이 지나치게 커지는 것을 막기 위해 원근감 있게 정보를 두어서 시그너처들을 구성한다. 즉, 시그너처의 그룹이 들어갈 위치를 기준으로 가까이에 위치한 데이터일수록 자세한 정보를, 멀리 있는 데이터일수록 대략의 정보를 제공하도록 시그너처 그룹을 구성한다. 그러면 클라이언트는 인덱스를 받았을 때 앞으로 브로드 캐스팅 될 전체 데이터에 대한 정보를 알 수 있고, 원하는 데이터가 더 이상 나오지 않는다는 것이 확인되면 엑세스를 종료하고, 이는 엑세스 타임의 감소로 이어진다.

시그너처 그룹의 구성은 간단하다. 브로드 캐스팅 될 개별 정보 프레임에 대한 인덱스 역할을 하는 시그너처 그룹에 포함된 시그너처들을 차례로 $S_1, S_2, \cdots, S_{(N_S-1)}, S_{(N_S)}$라고 두자. 이 데이터를 포함하여 이후로 브로드 캐스팅 될 예정인 데이터를 $\epsilon^0, \epsilon^1, \epsilon^2, \ldots, \epsilon^{N_s-1}$개의 정보 프레임으로 그룹화한 후, 각각의 그룹에 대한 통합 시그너처 $S_1, S_2, \cdots, S_{(N_S-1)}, S_{(N_S)}$를 생성한다. 예를 들어 이렇게 하면 그림 2의 예와 같이 가까이 있는 정보 프레임에 대한 자세한 정보를 가지고, 멀리 있는 정보 프레임에 대한 대략의 정보를 가지는 원근감 있는 시그너처가 된다.



S_i = i번째 정보 프레임에 대한 시그너처

$S_{(i,j)} = S_i \vee S_{i+1} \vee \cdots \vee S_{j-1} \vee S_j$, $\epsilon = 2$

그림 2: 펄스펙티브 시그너처 스킴의 예

ϵ값이 클수록 하나의 시그너처 그룹은 많은 정보 프레임을 커버하는 반면, 시그너처가 제공하는 식별력은 떨어진다. 그림 3의 예와 같이 ϵ가 크면 상대적으로 넓은 범위를 커버하는 반면, 여러 정보 프레임에 대한 통합 시그너처를 생성해야하므로 식별력이 떨어진다. 즉, 폴스 드랍이 많아진다. 따라서 본 스킴이 효과적이기 위해서는 ϵ값과 시그너처의 식별력에 큰 영향을 미치게 되는 요인 중 하나인 시그너처의 비트수가 함께 고려되어야한다.

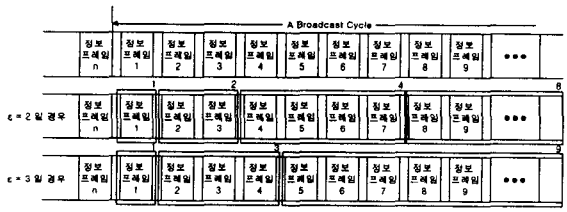


그림 3: ϵ값에 따른 정보 프레임의 그룹화 예

그리고 하나의 시그너처 그룹이 커버하는 최대 정보 프레임의 개수는 아래와 같다.

$$NI_S = \sum_{i=0}^{NS_S-1} \lceil \epsilon^i \rceil$$

NI_S : 하나의 시그너처 그룹이 커버하는 정보 프레임의 수

NS_S : 하나의 시그너처 그룹에 포함되는 시그너처의 수

ϵ : 정확성 팩터 ($\epsilon \geq 0$)

3.3 클라이언트의 엑세스 프로토콜

위와 같이 시그너처 그룹이 생성되어서 각 정보 프레임 앞에 위치해서 브로드 캐스팅 될 때 클라이언트의 엑세스 프로토콜은 다음과 같다. 하나의 브로드 캐스트 사이클에 포함되는 정보 프레임의 수를 N이라 둔다.

```
Q is 클라이언트의 질의에 대한 시그너처
A is {}
C is {1, 2, ··· , N-1, N}
for each i from 1 to N
  if i ∈ C
```

채널을 리스닝해서 시그너쳐 그룹 S_i를 수신
else
  다음 시그너쳐 그룹이 브로드 캐스팅 될 때까지 대기 모드 유지
if (S_{i1} ∧ Q) = Q
  이어지는 정보 프레임 수신 후 A에 저장
S_i is S_i - {S_{i1}}
for each S_{ij} in 시그너쳐 그룹 S_i
  if (S_{ij} ∧ Q) = Q
  C is C - {S_{ij}가 커버하는 영역의 정보 프레임들}
if C = {}
  최종 결과 집합 A를 반환 후 프로토콜 종료
else
  다음 시그너쳐 그룹이 브로드 캐스팅 될 때까지 대기 모드 유지

그림 4: 클라이언트의 수신 프로토콜

클라이언트는 하나의 브로드 캐스트 사이클에 포함되는 정보 프레임을 살펴서 사용자의 질의와 일치하는 모든 정보 프레임을 수신해야한다. 그림 2의 프로토콜은 검색할 대상 집합을 C로 두고 매 시그너쳐를 받을 때마다 시그너쳐를 통해서 검색하지도 않아도 될 정보 프레임 영역을 추출하고, 그 영역을 집합 C에서 뺀다. 그래서 집합 C가 공집합이 되거나 전체 브로드 캐스트 타임만큼 검색했다면 엑세스를 종료한다.

4. 성능 분석

본 스킴의 효율성에는 데이터의 사이즈와 시그너쳐의 비트 수가 중요한 변수로 작용한다. 먼저 데이터의 크기가 시그너쳐에 비해서 상대적으로 크면 클수록 효과적이다. 데이터의 크기가 클수록 시그너쳐에 의해서 발생하는 추가적인 비용은 영향이 줄어들기 때문이다. 시그너쳐의 크기 역시 중요한 요인이다. 시그너쳐가 크면 클수록 통합된 시그너쳐의 식별력이 좋아진다. 따라서 제안한 스킴이 엑세스 타임을 감소시킬 수 있는 여지가 생긴다. 반면 시그너쳐의 크기가 커지면 전체 브로드 캐스트 사이클이 길어지게 된다. 즉, 최악의 경우의 엑세스 타임은 나빠지게 되는 것이다.

그림 5는 시그너쳐의 비트 수 증가에 따른 제안한 스킴과 멀티-레벨 스킴의 평균 엑세스 타임을 비교한 것이다. 실험은 정보 프레임의 개수를 100, 정보 프레임의 크기를 100, 하나의 비트의 크기를 0.01로 두었다.

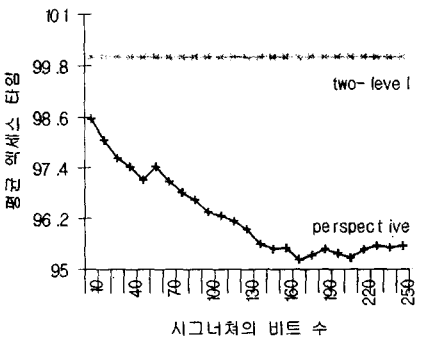


그림5: 시그너쳐의 비트 수에 따른 엑세스 타임의 변화

결과는 멀티-레벨 스킴의 경우 평균 엑세스 타임이 고정된 반면, 제안한 스킴은 시그너쳐의 비트 수가 증가할수록 평균 엑세스 타임이 감소하는 것을 보여준다. 실험에서는 데이터에게 임의의 비트 스트링을 할당했으나, 데이터들이 클러스터링 되어있고, 시그너쳐들이 클러스터링 된 데이터의 특성을 반영하여 생성된다면 더 좋은 성능을 보일 것으로 예상된다.

5. 결론

모바일 환경에서 브로드 캐스트 기법을 효과적으로 사용하기 위해서는 데이터가 언제 브로드 캐스팅 될 것인가에 대한 인덱스가 제공되어야 한다. 기존에 연구 되었던 트리 기반의 인덱스 생성 기법은 클라이언트의 평균 엑세스 타임과 튜닝 타임을 최소화하는 반면, 멀티-어트리뷰트에 대한 인덱싱 등 추가적인 고려 사항을 반영하기 어려웠다. 이러한 한계를 극복하기 위한 방법으로 시그너쳐를 이용한 인덱싱 기법이 제시되었지만, 클라이언트의 엑세스 타임이 고정된다는 단점을 가지고 있었다. 본 논문은 이러한 문제를 해결하여 평균 엑세스 타임을 최소화할 수 있는 필스펙티브 시그너쳐 스킴을 제시하였다.

제안한 스킴은 앞으로 브로드 캐스팅 될 데이터에 대한 포괄적인 정보를 제공하는 시그너쳐 집합을 인덱스로 제공한다. 따라서 클라이언트는 인덱스를 읽을 때마다 클라이언트가 질의한 데이터가 앞으로 브로드 캐스팅 될 것인지 혹은 그렇지 않을 것인지 판단할 수 있다. 그래서 원하는 데이터가 더 이상 브로드 캐스팅 되지 않는다는 것이 확인되면 수신을 종료하게 되고 따라서 평균 엑세스 타임을 감소시킨다. 그리고 이때 너무 많은 시그너쳐가 사용되지 않도록 정확성 팩터 ϵ를 두어서 가까이 있는 데이터에 대한 자세한 정보와 멀리 있는 데이터에 대한 대강의 정보를 가지는 원근감 있는 시그너쳐 그룹이 되도록 했다.

평균 엑세스 타임을 최소화하기 위해서는 데이터의 사이즈와 시그너쳐의 비트 수를 바탕으로 최적의 ϵ를 선택해야한다. 즉, 주어진 다양한 상황에 따라 최적의 ϵ값을 도출해야하는 데, 이는 남겨진 과제이다.

6. 참고 문헌

[1] G. Forman and J.Zahorjan, *The Challenges of Mobile Computing, in IEEE Computer*, 27(6), pp. 38-47. April 1994.

[2] T. Imielinski and B. Badrinath, *Wireless Mobile Computing : Challenges in Data Management*, Comm. of ACM, Vol 37, No. 10, pp. 18-28, 1994.

[3] D. Barbara, *Mobile Computing and Databases-A Survey*, IEEE Transaction on Knowledge and Data Engineering, Vol. 11, No. 1, pp. 108-117, 1999.

[4] S. Acharya, M. Franklin, S. Zdonik, and R. Alonso, *Broadcast Disks: Data Management for Asymmetric Communication Environments*, Proc. ACM SIGMOD Int. Conf. on Management of Data, pp. 199-210, 1995.

[5] T. Imielinski, S. Viswanathan, and B. Badrinath, *Data on Air: Organization and Access*, IEEE Transaction on Knowledge and Data Engineering, pp. 353-372, 1997.

[6] N. Shivakumar and S. Venkatasubramanian, *Efficient indexing for broadcast based wireless systems*, Mobile Networks and Applications, Vol. 1, pp. 433-446, 1996.

[7] M. Chen, K. Wu, and P. Yu, *Optimizing Index Allocation for Sequential Data Broadcasting in Wireless Mobile Computing*, IEEE Transactions on Knowledge and Data Engineering,, Vol. 15, No. 1, JAN/FEB, 2003.

[8] W. Lee and D. Lee, *Using Signature and Caching Technique for Information Filtering in Wireless and Mobile Environments, Special Issue on Databases and Mobile Computing*, Journal on Distributed and Parallel Databases, Vol. 4, No. 3, July 1996, 205-227.Conference on Communications (InfoCom'96), March 1996.

[9] Seung Joon Lee, Dan Keun Sung, *A New Fast Handoff Management Scheme in ATM-based Wireless Mobile Networks*, Proceedings of the Globecom '96 - Volume 2 ,11/18/96

[10] Sueng-Yong Park,Vaduvur Bharghavan, Sung-Mo Kang, *Data Ling Level Support for Handoff in Wireless ATM Network*, Proceeedings of the 1997 IEEE International Conference on Communications - Volume 2/3 , 5/08/97