Reverse Top-k 질의 처리 방법 비교 및 문제점 분석

임선영, 박영호*
숙명여자대학교 멀티미디어과학과
*교신저자
e-mail:{sunnyihm, yhpark}@sm.ac.kr

A Comparison and Study among Reverse Top-k Query Methods

Sun-Young Ihm, Young-Ho Park*
Dept. of Multimedia Science, Sookmyung Women's University
*Corresponding Author

요 약

Top-k 질의 처리가 사용자가 원하는 데이터를 검색하는 방법인 반면에, Reverse Top-k 질의 처리는 데이터의 관점에서 특정 데이터를 가장 선호할 만한 사용자를 검색하는 방법으로 생산자의 입장에서 매우 중요한 연구이다. 본 논문에서는 Reverse Top-k 질의 처리 방법들을 소개하고 비교 및 문제점을 분석한다.

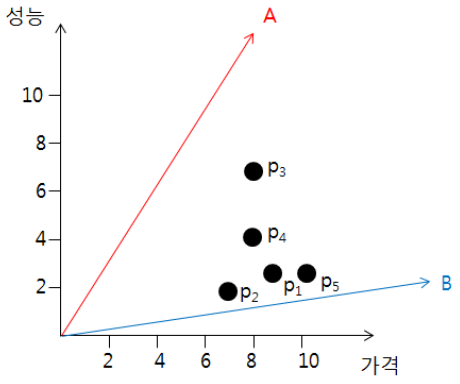
1. 서론

Top-k 질의 처리는 주어진 데이터에서 사용자가 원하는 k개의 데이터를 결과로 반환하는 방법을 말한다. 일반적으로 Top-k 질의 처리에서 데이터는 여러 속성을 가지게 되는데, 이 때 사용자는 속성 마다 가중치를 두어 질의 처리를 할 수 있다. 어떤 속성에 더 많은 가중치를 두느냐에 따라 결과는 달라진다. 예를 들어, 노트북을 산다고 가정해 보자. 이 때, 사용자들은 가격과 성능이라는 두 가지 속성에 가중치를 두고 검색을 하며, 2명의 사용자 A와 B가 있다. A는 가격에 0.3, 성능에 0.7의 가중치를 주고, B는 가격에 0.9, 성능에 0.1의 가중치를 주었다. 데이터베이스에는 총 5개의 노트북이 있으며, 노트북의 점수를 구하는 스코어링 함수는 f(p) = p[0]*w[0] + p[1]*w[1] 이다. 여기서 p는 데이터를 뜻하고, p[0]과 p[1]은 가격과 성능을 각각 점수로 나타낸 값이다. w[0]과 w[1]은 각각 가격과 성능의 가중치를 뜻한다. 사용자들은 2개의 결과를 원한다. 표 1은 데이터베이스에 저장된 데이터와 사용자 A와 B가 설정한 가중치에 의해 계산된 노트북의 점수를 나타내고 있다.

<표 1> 노트북 데이터베이스

id	가격	성능	점수-A	점수-B
p1	9	3	4.8	8.4
p2	7	2	3.5	6.5
p3	8	7	7.3	7.9
p4	8	4	5.2	7.6
p5	10	3	5.1	9.3

그림 1은 위의 예제를 2차원의 공간상에 표현하고 있다. 노트북 데이터베이스는 두 개의 속성을 가지므로 가격은 1차원을 뜻하는 x축으로, 성능은 2차원을 뜻하는 y축으로 하여 2차원의 공간에 표현이 가능하다. 그리고 사용자 A와 B의 질의를 벡터로 표현하였는데, 각각의 가중치가 반영되어 기울기가 다른 것을 볼 수 있다. A는 성능에 더 높은 가중치를 두어 성능 쪽으로 기울어져 있고, B는 가격에 더 높은 가중치를 두었기 때문에 가격 쪽으로 기울어져 있다. 사용자 A의 Top-2 결과는 p3과 p4이며 사용자 B의 Top-2 결과는 p1과 p3이다. 이렇게 사용자가 설정한 가중치에 따라 결과가 다르게 나타나는 것을 볼 수 있다.



(그림 1) Top-k 질의처리의 예

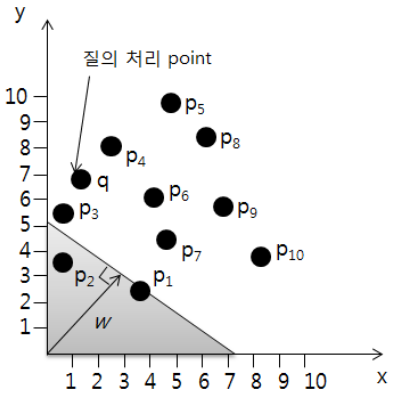
Top-k 질의처리는 이렇게 사용자의 관점에서 질의 처리를 수행한다. 이와 반대로 Reverse Top-k 질의 처리는

사용자의 관점이 아닌 제품, 즉 데이터의 관점에서 질의 처리를 수행하는 것으로 특정 제품을 가장 선호하는 가중치를 가지는 사용자를 검색하는 것이다. 예를 들어 위의 예제에서 노트북 p4는 사용자 A가 더 선호할 것으로 예측할 수 있다. 제품을 생산하는 생산자의 입장에서는 생산되는 제품을 선호할 것 같은 사용자를 예측이 가능하면, 모든 사용자가 아닌 특정 예측된 사용자에게만 보다 적합한 광고 또는 제품 추천이 가능하게 된다. 따라서 최근 Reverse Top-k 질의 처리에 대한 연구에 대한 관심이 높아지고 있다. 본 논문에서는 주요 Reverse Top-k 질의 처리 방법을 소개하고, 문제점을 분석한다.

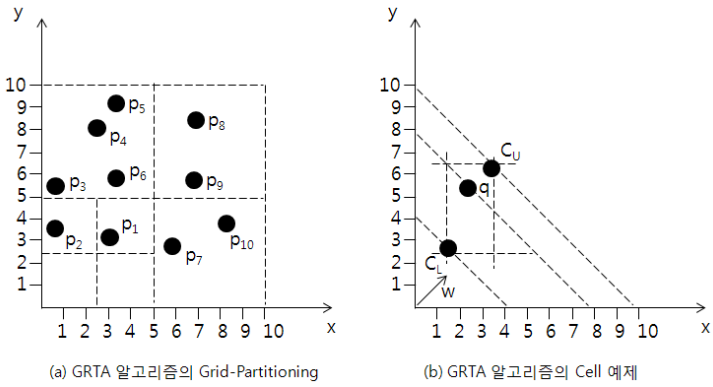
본 논문의 구성은 다음과 같다. 2장에서는 주요 Reverse Top-k 질의 처리 방법을 소개한다. 3장에서는 주요 연구들의 문제점을 분석하고 4장에서는 결론을 맺는다.

2. Reverse Top-k 질의 처리 방법

Reverse Top-k 질의 처리는 [4]에서 제안되었다. 질의 처리 포인트인 q와 정수 k가 주어졌고, 데이터 셋 S가 주어졌을 때, Reverse Top-k 질의 처리의 결과로 q를 선호할 가능성이 높은 가중치를 가지는 사용자 벡터를 가진다. Reverse top-k Threshold Algorithm (RTA) [4]은 top-k 질의처리의 계산 횟수를 줄이기 위하여 한계치를 사용하였다. 그림 2는 RTA 알고리즘의 예를 보여주고 있다 [4]. 점 q는 질의 처리 포인트를 뜻하고, 점 p1과 p2는 질의 처리 공간(회색 삼각형)에 의해 에워싸져 있다. Grid-based Reverse top-k Algorithm (GRTA) [4]은 격자 기반의 실체화된 뷰를 사용하여 Reverse Top-k 질의 처리를 수행한다. 그림 3은 GRTA의 예를 보여주고 있다 [4]. 그림 3(a)는 격자 기반으로 분할 된 것을 보여주고 있다. 그림 3(b)에는 질의 처리 포인트 q가 주어져있고, 점 q를 둘러싼 cell이 정해진다. cell의 한계치를 비교하여 질의 처리 결과를 탐색한다.



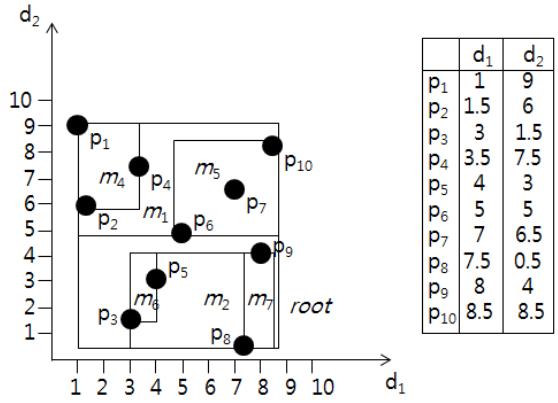
(그림 2) RTA 알고리즘의 예 [4]



(그림 3) GRTA 알고리즘의 예 [4]

Chester et al.은 [1]에서 2차원에서의 Reverse Top-k 질의 처리를 위한 인덱싱 기법을 제안하였다. Top-k 질의 처리를 선으로 표현한 후, 선에 기반 하여 인덱스를 생성하였다. [2]에서는 Top-k 질의 처리 방법들에 대한 평가와 모든 Top-k 질의 처리를 효율적으로 수행할 수 있는 방법을 제안하였다.

Branch-and-Bound Algorithm (BBR) [6]은 Reverse Top-k 질의 처리에 효율적으로 답하기 위하여 데이터는 R-tree 인덱스를 사용하여 인덱싱 한 후, 최소범위 사각형 (Minimum Bounded Rectangle, MBR)을 통해 질의 처리를 하는 방법을 제안하였다. 그림 4는 BBR 알고리즘의 예를 보여주고 있다 [6]. 먼저 데이터들은 2차원으로 이루어져 있으며, R-tree 인덱스를 사용하여 인덱싱 되어있다. 점들은 MBR 단위로 구성되어 있으며, MBR의 상계와 하계를 설정하여 비교한다. 그림 5는 BBR 알고리즘을 나타내고 있다 [6]. 행 3에서 R-tree의 root를 heep에 저장한 후 행 6에서는 INTOPk 알고리즘을 호출한다. INTOPk 알고리즘은 MBR의 데이터가 질의 처리의 결과인지 아닌지를 판별하는 알고리즘이다. 또한, 데이터에 대한 반복적인 접근을 줄이고자 결과를 공유하는 BBR* 알고리즘과, 통계 R-tree를 사용하는 BBRA 알고리즘도 함께 제안하였다.



	d_1	d_2
p_1	1	9
p_2	1.5	6
p_3	3	1.5
p_4	3.5	7.5
p_5	4	3
p_6	5	5
p_7	7	6.5
p_8	7.5	0.5
p_9	8	4
p_{10}	8.5	8.5

(그림 4) BBR 알고리즘의 예 [6]

Algorithm BBR

```
1:  Input: Point q, value k
2:  Output: Reverse top-k result set RTOP_k(q)
3:  heapW.enqueue(RtreeW.getRoot())
4:  while(!heapW.isEmpty()) do
5:    e ← heapW.dequeue()
6:    i ← INTOPk(e.m, q, k)
7:    if i=0 then
8:      heapW.enqueue(expand(e))
9:    else
10:     if i=1 then
11:       RTOP_k(q) ← RTOP_k(q) ∪ expandAll(e)
12: return RTOP_k(q)
```

(그림 5) BBR 알고리즘 [6]

또한 최근, Reverse Top-k 질의 처리를 사용하는 많은 어플리케이션들도 연구되고 있다. [3]에서는 모바일 기기를 사용하는 이동성을 가지는 사용자들의 위치에 기반한 선호도를 모니터링 하는 연구를 수행하였다. [5]에서는 가장 영향력이 큰 데이터 오브젝트를 판별하기 위한 알고리즘으로 SB와 BB를 제안하였다.

3. Reverse Top-k 질의 처리 방법의 문제점 분석

기존의 Reverse Top-k 질의 처리 방법들은 몇 가지 문제점들을 가지고 있다. RTA [4]는 Reverse Top-k 질의 처리에 답하기 위해 모든 사용자의 벡터를 검색하고, 이미 탐색한 사용자의 벡터에 대하여도 반복적인 Top-k 질의 처리를 수행해야 한다는 문제점이 있다.

BBR [6]은 RTA의 이러한 문제점을 해결하기 위하여 MBR을 사용하여 답이 될 가능성이 적은 사용자 벡터를 제거한다. 하지만 MBR의 크기를 설정한 후 질의와 데이터를 비교하기 위하여 상계와 하계를 계산하는데, 이 때 설정되는 상계와 하계의 범위의 크기에 따라 비교해야 할 사용자 벡터가 많아질 수 있다는 문제점이 있다. 따라서 상계와 하계의 범위를 데이터에 따라 동적으로 설정하는 방법이 필요하다.

4. 결론 및 향후연구

Reverse Top-k 질의 처리는 Top-k 질의 처리와 관점이 반대되는 개념으로 사용자보다는 생산자의 입장에서 매우 중요하다. 생산자들은 Reverse Top-k 질의 처리의 결과를 바탕으로 생산되는 제품을 가장 선호할 가능성이 큰 고객을 판별할 수 있으며, 예측된 고객들에게 더욱 알맞은 홍보를 할 수 있다.

Reverse Top-k 질의 처리의 대표적인 연구로는 RTA, GRTA, BBR 등이 있다. 향후 연구로는 기존 연구들의 문제점을 해결하기 위하여 MBR의 크기와, 상계, 하계의 크기를 동적으로 설정하는 방법에 대한 연구를 진행한다.

5. 사사문구

본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041854, 안전한 주거환경을 위한 실시간 위험요소 예측/방지용 스마트 홈 서비스 플랫폼 기술 개발]

참고문헌

[1] Chester S. Thomo A. Venkatesh S. Whitesides S. “Indexing Reverse Top-k Queries in Two Dimensions” In Proceedings of the 18th International Conference on Database Systems for Advanced Applications, pp.2013-208, 2013.

[2] Ge S. Hou U L. Mamoulis N. Cheung D.W. “Efficient All Top-k Computation－A Unified Solution for All Top-k, Reverse Top-k and Top-m Influential Queries” IEEE Transactions on Knowledge and Data Engineering, Vol.25 (5), pp.1015-1027, 2013.

[3] Vlachou A. Doulkeridis C. Nørvåg K. “Monitoring reverse top-k queries over mobile devices“ In Proceedings of International Workshop on Data Engineering for Wireless and Mobile Access (MobiDE), 2011.

[4] Vlachou A. Doulkeridis C. Kotidis Y. Nørvåg K. “Reverse Top-k Queries” In Proceedings of 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp.365-376, 2010.

[5] Vlachou A. Doulkeridis C. Nørvåg K. Kotidis Y. “Identifying the most influential data objects with reverse top-k queries” PVLDB, Vol.3 (1-2), pp.364-372, 2010.

[6] Vlachou A. Doulkeridis C. Nørvåg K. Kotidis Y. “Branch-and-Bound Algorithm for Reverse Top-k Queries” In Proceedings of ACM SIGMOD, 2013.