캐쉬 관리를 위한 인기도 기반의 대체 기준치에 관한 연구*

홍진선 이상호
숭실대학교 컴퓨터학과
jshong@orion.soongsil.ac.kr shlee@computing.soongsil.ac.kr

Popularity-based Eviction Functions in Cache Managements

Jin-Seon Hong Sang-Ho Lee
School of Computing, Soongsil University

요 약

캐쉬 대체 알고리즘은 캐쉬 적재 공간의 한계성을 극복하는 방법 중에 하나이다. 기존의 많은 대체 알고리즘의 문제점인 대체 기준치의 부정확성 및 불충분성을 해결하기 위해 인기도를 제안하였다. 인기도는 인기 검색어의 순위를 정규화 한 값으로, 대량의 자료를 바탕으로 얻어진 통계치이다. 인기도 산출의 기반이 되는 인기 검색어는 시간적 흐름에 민감하고, 사회 전반적인 경향을 반영하며, 많은 중복을 가지고 있다. 인기도는 각 검색 엔진별로 단일 인기도와 누적 인기도를 산출한 후에, 이들을 모두 병합하여 산출된다. 이것을 병합 인기도라고 하며, 이는 임의의 검색어에 0에서 1사이의 소수값으로 부여된다. 인기도는 메타 검색 엔진에서 캐쉬 대체를 수행할 때 적용될 수 있으며, 다수의 자료 입력 경향에 관한 정보가 존재하는 문제 영역에 사용될 수 있다.

1. 서론

캐쉬는 전산분야에서 다양하게 사용되며, 많은 연구가 진행되어왔다. 특히, 항상 네트웍을 이용하여 타 검색 엔진에 접근하고 정보를 추출하는 메타 검색 엔진의 경우에는 검색 서비스 질(quality)의 향상에 많은 도움을 준다. 캐쉬를 관리할 때는 캐쉬 공간의 한계성을 반드시 고려해야 하며[1], 이에 대한 가장 쉬운 해결책은 데이터의 대체(replacement)를 수행하는 것이다.

캐쉬 대체를 위한 알고리즘은 다양하게 존재하며, LRU (Least Recently Used)나 LFU (Least Frequently Used)등이 가장 많이 사용되어 왔다. 하지만, 이러한 전통적인 캐쉬 대체 방법은 소량의 정보만을 이용하여 대체를 수행하기 때문에 좋은 성능을 나타내지 못한다[2]. 특히 LFU의 경우에는 미래에 참조 가능성이 높은 데이터를 퇴거(eviction) 시키는 단점이 존재한다[3]. 이러한 단점은 모두 대체 기준치의 정보가 부정확하거나, 불충분하기 때문에 발생한다.

본 논문은 캐쉬 대체 알고리즘의 성능을 높이기 위한 새로운 대체 기준(eviction function)으로 인기도를 제안한다. 인기도는 검색 엔진에서 제공하는 인기 검색어를 수집하여, 그들의 순위를 정규화하여 표현한 값이다.

*본 연구는 한국과학재단 목적기초연구(2000-2-51200-002-3) 지원으로 수행되었음.

본 논문은 다음과 같이 구성된다. 제 2장에서는 인기 검색어의 정의와 특징에 대하여 살펴보고, 제 3장에서는 인기도 산출법에 대하여 기술한다. 결론 및 향후 과제는 제 4장에서 언급한다.

2. 인기 검색어 (Popular Queries)

검색 엔진 사용자들이 질의한 단어들 중에 출현 빈도수가 높은 검색어들을 인기 검색어라고 한다. 인기 검색어는 각 검색 엔진의 추출 정책에 따라, 공시되는 인기 검색어 수, 필터링(filtering) 여부, 공시주기(posting cycle) 등이 상이하다. 국내의 대부분의 상용 검색 엔진들은 5개에서 100개의 인기 검색어를 매일 또는 매주 제공하고 있다.

인기 검색어를 살펴보면, 시대적 흐름과, 사회 전반적인 경향 및 유행(public trends/interests)을 잘 반영(reflection)함을 알 수 있다. [표 1]은 이러한 특성을 잘 나타낸다.

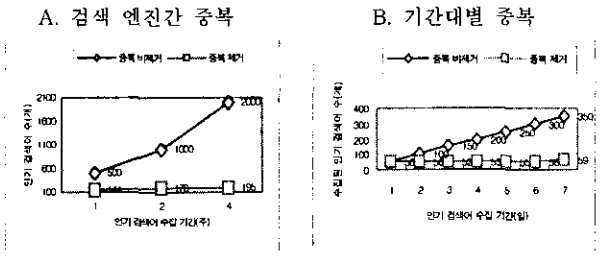
[표 1] 인기 검색어 입력 경향

사건	인기 검색어 목록
수학능력시험	수능, 대입 입시, 대학교
크리스마스	산타크로스, 캐롤, 선물
2월 14일	발렌타인데이, 선물, 쇼핑몰

표에서 볼 수 있듯이, 특정 사건에 따라 인기 검색어는 변경되며, 사건이 종료됨과 동시에 해당 검색어들은 인기 검색어 목록에서 사라지는 경향을 보인다.

한 상용 검색 엔진에서 추출한 질의 67만 건 중에 194개의 인기 검색어가 약 23%를 차지함을 나타낸 예에서, 인기 검색어가 사용자 입력 질의에 얼마나 높은 비율을 차지하는지를 알 수 있다. 다시 말하면, 인기 검색어는 사용자 질의 입력 경향을 잘 반영하는 지표라고 할 수 있다.

인기 검색어가 가지는 다른 특징은 많은 중복을 포함하고 있다는 것이다. 인기 검색어의 중복은 기간대별이나 검색 엔진간에 나타난다. 다음의 그림은 세 개의 검색 엔진에서 수집된 인기 검색어의 중복을 나타낸다.



[그림 1] 인기 검색어의 중복

[그림 1]의 A는 인기 검색어가 검색 엔진간에 많은 중복이 있음을 나타낸다. B의 경우에는 하나의 검색 엔진에서 기간대별로 많은 중복이 일어남을 나타낸다. 예를 들어, 7일째의 경우를 보면 350개의 인기 검색어가 수집되어야 하는데, 중복을 모두 제거하면 59개가 수집됨을 알 수 있다.

많은 중복이 일어난 검색어일수록 다수의 사용자에 의해 질의된 검색어로 여길 수 있다. 우리는 이러한 특성을 반영하는 인기도 산출법을 제안한다.

3. 인기도 (Popularity)

검색 엔진에서 제공하는 인기 검색어는 검색어간에 순위(ranking)를 가진다. 검색어간의 순위는 해당 검색어의 출현 빈도수에 의해 정해지며, 상위 순위를 갖는 몇 개만이 사용자에게 공시되고 있다.

우리는 사용자가 입력한 임의의 검색어에 0과 1사이의 소수값을 갖는 인기도를 부여한다. 인기도는 해당 검색어의 순위를 공시주기, 사전에 정의된 최대 인기 검색어 수, 외부 검색 엔진의 수로 나누어 표현된다. 인기도가 높을수록 많은 사람에 의해 질의된 검색어를 의미하며, 만일 임의의 검색어가 0의 인기도를 갖는다면, 해당 검색어는 비인기 검색어(non-popular queries)임을 의미한다. 산출되는 인기도의 특성에 따라 단일 인기도(unit popularity; UP), 누적 인기도(accumulated popularity; AP), 병합 인기도(fused popularity; FP)의 세 가지로 구분한다.

인기도 산출법을 기술하기 전에 W에 대하여 언급한다. W는 사전에 정의되는 값으로 인기도를 정규화 할 때 사용된다. W는 검색 엔진에서 발표하는 인기 검색어 수와는 무관하게 결정되며, 각 검색 엔진에서 발표하는 인기 검색어 수보다 항상 큰 값을 가진다. 또한 W는 한번 정의되며, 자속적으로 변하지 않는다.

3.1 단일 인기도 (Unit Popularity)

단일 인기도는 검색 엔진에서 제공되는 하나의 인기 검색어 목록에서 추출한 값으로, 검색 엔진의 공시주기마다 산출하며, 인기 검색어를 수집할 때 적용된다. 단일 인기도는 인기 검색어의 순위를 정규화하여 표현한다.

$R_i^j(X)$는 i번째 검색 엔진에서 j번째 공시되는 인기 검색어 X에 대한 순위를 나타내며, 검색 엔진에 따라 1 이상의 값을 가진다. i번째 검색 엔진에서 j번째 공시되는 인기 검색어 X의 단일 인기도를 나타내는 $UP_i^j(X)$는 다음과 같이 산출한다.

$$UP_i^j(X) = \frac{W - R_i^j(X) + 1}{W}$$

위의 수식을 전개하면, 인기 검색어는 각 검색 엔진별로 $0 < UP_i^j(X) \leq 1$의 범위를 갖는 단일 인기도를 부여 받는다. 특히, 서로 다른 검색 엔진에서 추출한 인기 검색어의 순위가 동일하다면, 두 인기 검색어는 동일한 단일 인기도를 갖는다. 이는 각 검색 엔진에서 제공되는 인기 검색어 개수와 상관없이 인기 검색어에 단일 인기도를 부여함을 의미한다. 단일 인기도 값이 0에 수렴할수록 낮은 순위를 갖는 인기 검색어이고, 1에 수렴할수록 높은 순위를 갖는 인기 검색어이다. 즉, 단일 인기도 값이 높으면 해당 검색 엔진의 대다수 사용자들에 의해 질의된 검색어라고 할 수 있다.

3.2 누적 인기도 (Accumulated Popularity)

누적 인기도는 일정 기간동안 동일한 검색 엔진의 단일 인기도를 누적하여 산출한다. 누적 인기도를 산출할 때에는 중복된 검색어의 단일 인기도를 더하여 각각의 인기 검색어에 하나의 누적 인기도가 부여될 수 있도록 한다.

누적 인기도는 사전에 산출된 단일 인기도의 누적 횟수를 의미하는 차수(order)를 가지고 있다. 차수는 인기도 값을 조정(adjustment)할 때 사용되며, 공시주기가 서로 다른 검색 엔진들간의 누적 인기도 병합을 용이하게 한다. n_i는 i번째 검색 엔진의 차수를 나타낸다. i번째 검색 엔진에서 제공하는 인기 검색어 X의 누적 인기도를 나타내는 $AP_i(X)$는 다음과 같이 산출한다.

$$AP_i(X) = \frac{1}{n_i}\sum_{j=1}^{n} UP_i^j(X)$$

누적 인기도는 단일 인기도의 합을 차수로 나누는 인기도 조정을 수행하여 생성하며, $0 < AP_i(X) \leq 1$의 범위를 가진다. 만일 임의의 검색어가 인기 검색어

목록에 많이 출현되면, 해당 검색어는 높은 누적 인기도를 갖는다. 하지만, 출현 빈도수가 적은 검색어는 누적 인기도의 차수가 높아질수록 인기도 값은 감소하게 된다. [그림 2]는 누적 인기도를 산출하는 알고리즘을 나타낸다.

// To compute AP of a query X for the i^{th} search engine
Let n_i be the order of AP of the i^{th} search engine
For each popular query X
j = 1, SUM = 0;
While (j <= n_i)

$$UP_i^j(X) = \frac{W - R_i^j(X) + 1}{W};$$

$SUM = SUM + UP_i^j(X);$
$j = j + 1;$
End of while-loop

$$AP_i(X) = \frac{SUM}{n_i};$$

End of for-loop

[그림 2] 누적 인기도 산출 알고리즘

위의 그림에서 보는 바와 같이, 인기 검색어의 단일 인기도는 누적 인기도를 해당 검색 엔진의 차수로 나눈 값이라 할 수 있다.

3.3 병합 인기도 (Fused Popularity)

병합 인기도는 이종의(heterogeneous) 검색 엔진에서 수집된 누적 인기도를 병합하여 생성되며, 검색 엔진의 수를 가지고 정규화 한다. 본 논문에서 병합 인기도는 모두 외부 검색 엔진에서 추출된 인기 검색어를 대상으로 산출한다.

N은 인기 검색어를 제공하는 검색 엔진의 수를 나타내는 값으로, 산출된 병합 인기도를 정규화 하는데 사용된다. 임의의 검색어 X에 대한 병합 인기도 FP(X)는 다음과 같이 산출한다.

$$FP(X) = \max\left(0, \frac{1}{N}\sum_{i=1}^{N} AP_i(X)\right)$$

병합 인기도는 단일 인기도 또는 누적 인기도와는 상이하게 $0 \le FP(X) \le 1$ 의 범위를 가지며, 모든 임의의 단어에 대하여 부여된다. 만일 병합 인기도가 0이면, 해당 단어는 인기 검색어 목록에 포함되어 있지 않은 값, 즉 비인기 검색어임을 의미한다.

본 논문에서, 임의의 검색어가 높은 인기도를 갖는 경우는 다음과 같다. (1) 인기 검색어 목록에서 높은 순위를 가질 경우, (2) 동일 기간에 다수의 검색 엔진에 출현될 경우, (3) 서로 다른 기간에 동일 검색 엔진에서 자주 출현될 경우이다. 본 논문에서 제안한 인기도 산출 방법은 공시되는 인기 검색어의 수, 외부 검색 엔진의 수, 각 검색 엔진의 공시 주기에 충분히 유연하게(flexible) 적용된다.

4. 결론 및 향후 과제

본 논문에서는 캐쉬 대체 알고리즘을 위한 새로운 대체 기준치인 인기도를 제안하였다. 인기도는 다수의 검색 엔진으로부터 추출된 인기 검색어의 순위를 정규화 한 값으로, 대량의 데이터를 바탕으로 산출된 통계치이다.

인기 검색어는 시대 흐름에 민감하며, 사회적 경향을 반영한다. 또한 기간대별 또는 검색 엔진간에 많은 중복을 포함하고 있다.

우리는 인기도는 단일 인기도, 누적 인기도, 병합 인기도로 구분하였다. 단일 인기도와 누적 인기도는 동일 검색 엔진에서 산출되는 인기도이다. 단일 인기도는 인기 검색어 수집 시에 산출되며, 순위를 정규화하여 표현한다. 누적 인기도는 차수를 가지고 있으며, 일정 기간(차수에 의해 부여된)동안 누적된 단일 인기도를 차수로 나누어 산출한다. 병합 인기도는 서로 다른 검색 엔진으로부터 산출한 누적 인기도를 병합하여 산출하며, 임의의 단어에 부여 된다.

인기도는 메타 검색 엔진을 위해 제안되었지만, 다수의 자료 입력 경향에 관한 정보가 존재하는 분제에 적용이 가능하다. 인기도를 적용한 대체 알고리즘은 기존보다 향상된 성능을 보일 것이라 사료된다.

메타 검색 엔진에서 인기도를 적용한 캐쉬 대체 알고리즘을 구현시에는 다음과 같은 사항을 고려해야 한다. (1) 수집된 인기 검색어의 수와 캐쉬 크기와의 관계, (2) 인기도 변경 주기(popularity update cycles), (3) 캐쉬에 저장된 결과의 진부성(staleness) 여부 등이다. 수집된 인기 검색어의 수에 따라 캐쉬 성능은 크게 달라지며, 이는 캐쉬 크기와도 밀접한 관련이 있다. 인기도 변경 주기는 시간적 흐름에 민감한 인기 검색어의 특성을 반영하기 위한 도구로 사용되며, 이 주기에 맞추어 현재 캐쉬 내에 적재되어 있는 값의 인기도도 변경된다. 각 검색 엔진들은 일정 기간마다 저장된 정보들의 재색인(re-indexing) 작업을 수행한다. 이는 이미 캐쉬에 저장된 자료와의 불일치를 유발하기 때문에, 반드시 고려되어야 할 사항이며, 이를 위해서는 캐쉬에 적재된 결과의 진부성 여부를 판단해야 한다.

References

[1] B. Chidlovskii, C. Roncancio, and M. Schneider, Semantic Cache Mechanism for Heterogeneous Web Querying, WWW8, 1999.

[2] E. O'Neil, P. O'Neil, G. Weikum, The LRU-K Page Replacement Algorithm for Database Disk Buffering, *Proc.* 1993 ACM SIGMOD, 1993.

[3] W. Stallings, Operating Systems 3rd Edition: Internals and Design Principles, Prentice-Hall, 479-482, 1998.