검색어의 연관법칙

문상준
네이버 개발실
june69@nhncorp.com

최재걸°
네이버 개발실
jkchoi@nhncorp.com

ARMS : Association Rule for sMall Set

Sang-June Mun
Naver Development Department

Jonathan Choi °
Naver Development Department

요 약

검색엔진에 사용자가 입력한 검색어를 분석하면 상호 연관이 있는 검색어들을 찾아낼 수 있다. 검색어들 간의 상호 연관성을 찾기 위해서 데이타 마이닝 분야의 연관법칙을 위한 알고리즘을 적용하였다. 그러나 이 알고리즘들은 모두 일정 횟수 이상 검색된 검색어 간의 연관법칙에 집중되어 있어서 일정 횟수 이상 검색되지 않은 검색어들은 버려진다. 이 연구에서는 이런 검색어들을 스몰 셋(small set)이라고 정의하고 스몰 셋의 연관법칙을 찾기위한 방법을 제시한다. 실험결과는 이 연구에 제시한 방법이 효과적으로 동작하는 것을 입증해준다.

제 1 절 연구배경 및 문제점

1.1 연구 배경

사용자가 입력한 검색어를 바탕으로 연관된 검색어를 찾아서 사용자에게 제시하면 질 높은 검색서비스가 가능해진다. 연관된 검색어를 찾아내기 위하여 데이타 마이닝 분야 중 하나인 연관법칙을 이용할 수 있다. 연관법칙을 찾는 알고리즘은 현재까지 크게 두 가지로 분류할 수 있다. IBM 연구소에서 개발한 Apriori 스타일 알고리즘 [1] [2] [3] 과 캐나다의 Simon Fraser대학에서 개발한 PrefixSpan 스타일 알고리즘 [5] [6] 이다.

하지만 지금까지의 연관법칙은 시장 데이타(Market Basket)에 초점을 맞추어 둔 것으로서, 검색엔진에서 사용자가 입력한 검색어를 대상으로 할때 곧바로 이 알고리즘을 사용할 수는 없다. 사용자가 입력한 검색어를 모으는 특별한 방법이 요구되고, 또한 검색어에 알맞은 연관법칙을 찾는 알고리즘이 요구된다. 사용자가 입력한 검색어를를 수집하는 방법과 수집된 데이타를 처리하기 위해서 연관법칙 알고리즘에 가해지는 변형에 대해서 연구하게 되었다.

1.2 기존 알고리즘의 문제점

기존의 두 알고리즘에서 $t \Rightarrow c$ 라는 연관에서 지지도(support), 신뢰도(confidence)다음과 같다. [1]

$$Support = P[t \wedge c]$$

$$Confidence = \frac{P[t \wedge c]}{P[t]}$$

지지도(support)는 전체중에 차지하는 비율을 나타내는 값으로, 이 값이 큰 검색어들을 모아서 라지셋(Large set)이라고 분류하고 라지셋에 대해서만 연관법칙을 구한다. 그러나, 검색어에서 연관법칙을 찾을 경우 수회정도 검색된 검색어의 연관법칙도 중요한 의미를 지닌다. 따라서 지지도(support)의 값을 아주 작게 설정해야 한다.
지지도의 값을 작게 설정하면 신뢰도의 값에 문제가 발생한다. 예를 들어 신뢰도를 2%로 준 경우를 생각해보자. 10회 검색된 질의어가 있다고 할 경우, 이 횟수의 2%는 1회에도 못 미치는 숫자이므로 이 값을 신뢰도를 판정하는 기준으로 사용할 수 없다. 따라서 신뢰도의 값을 정할 때 감색어의 검색 횟수에 따른 함수로 주어져야 한다.

그리고, 단순히 많이 검색되었기 때문에 서로 신뢰도가 높은 검색어들이 있다. 이런 값들은 서로 실제로는 연관성이 없으므로, 제거해야할 필요가 있고 이를 위해서 다음과 같은 상호의존도(interest)값을 이용한다.[2]

$$Interest = \frac{P[t \wedge c]}{P[t] \cdot P[c]}$$

마지막으로, 검색어의 연관어가 구해지면, 이를 서비스하기 위해서는 연관이 있는 검색어중에 어떤 검색어가 가장 연관이 있는지 그 순서를 구할 필요가 발생한다. 따라서, 연관정도를 측정할 수 있는 기준이 필요해진다.

제 2 절 알고리즘의 변형

1.2절에서 제시한 문제를 해결하기 위해 필요한 변형은 세가지 이다. 검색엔진에서 데이타를 수집하기 위한 전처리 과정이 필요하고, 연관성을 판단하는 과정에서 신뢰도의 변형과 연관어를 추출하는 과정에서 연관정도를 계산하는 부분이 추가되어야 한다.

[1] t,c는 각각 검색어를 나타내고, P[t]는 검색어 t가 전체에서 나타날 확율을 뜻한다.

[2] 이에 대한 내용은 [7] 을 참조하기 바란다

전처리과정은 수집(collect), 정제(refinery), 매핑 (mapping)의 세가지 과정으로 구성된다.

첫째, 검색엔진에서 검색되는 데이타를 수집하는 과정이다. 검색엔진에서는 한 명의 사용자가 연속해서 검색하는 검색어들을 식별해서 수집하는 방법이 필요하다. 이를 위해서 웹세션(web-session)을 이용하였다. 사용자가 검색창을 처음 열면 세션을 생성하고 쿠키(cookie)를 기록하고 검색을 할 때마다 검색어와 함께 쿠키에 적힌 세션을 쌍으로 기록하여 세션이 같은 검색어는 함께 검색된 검색어 임을 인식하게 된다. 세션에는 시간제한(time-expired)를 두어서 일정 시간동안 사용자가 검색어를 입력하지 않을 경우 자동 소멸하는 방법을 택하여 양질의 데이타가 수집될 수 있도록 하였다.

둘째, 데이타 정제과정이 필요하다. 하나의 세션안에 일정 수 이상의 검색어가 존재할 경우 이 세션은 정상적이지 않은 것으로 판단하여 제거하도록 한다. 그런 세션은 사용자가 입력하기 보다는 공격성 검색어일 경우가 많기 때문이다. 또, 하나의 세션안에 동일한 검색어는 하나로 변경하여 주도록 한다.

셋째, 매핑(mapping)과정이다. 검색어와 세션은 모두 스트링데이타여서 그대로 연산과정에 사용하면 메모리와 처리속도에서 손해를 보게된다. 따라서 검색어와 세션을 숫자로 각각 매핑을 시키고, 매핑된 데이타를 가지고 연관성 알고리즘을 수행 할 수 있도록 한다.

2.2 신뢰도 값의 변형

이 연구에서 제시되는 방법은 아프리오리(Apriori) 스타일과 프리픽스(Prefix)스타일 알고리즘에 모두 적용될 수 있다. 앞으로 이 논문에서는 아프리오리-스타일(Apriori-style)의 경우를 예로 들어 설명하기로 하자. Apriori 알고리즘에서는 k+1번째 연관어를 찾기 위해서 k번째의 연관어로 부터 k+1번째 후보집합(candidate set)을 만들고 이로부터 k+1번째 연관어를 판단한다. k+1번째 연관어를 판단할때 지지도(support)와 신뢰도(confidence)를 사용하게 된다. 이때 기존 알고리즘에서는 신뢰도를 판정하는 기준을 일률적으로 결정하였으나, 검색어의 연관법칙에 있어서는 검색어의 횟수에 따라 상대적으로 결정해야 한다.

$$(Minimum \quad Confidence) = \frac{1}{\sqrt{N[t]}} \times 100\% \ ^{3}$$

이 수식을 따르면, 100회 검색된 검색어의 연관어가 되기위한 최저신뢰도는 10%, 10회 검색된 검색어의 최저신뢰도는 33.3%가 된다. 이처럼 검색된 회수의 함수로 최저신뢰도 값을 산출하면 연관성이 있는 결과물을 산출 할 수있다. 이는 실험결과에서 확인할 수 있다.
위의 수식은 실험적으로 구해진 값으로, 검색횟수가 작아질수록 높은 확률의 신뢰도 기준을 갖도록 설정되었다. 데이타

[3] N[t]는 검색어 t의 검색회수를 나타낸다.

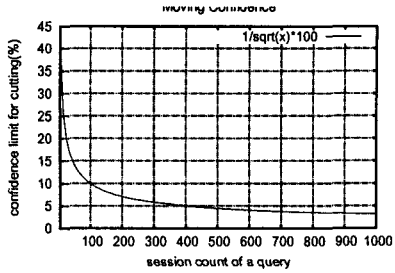


그림 1: 최저신뢰도 값의 변화

의 상태에 따라서 이 함수값은 변화할 수 있다. 데이타가 양질의 데이타일 경우, 최저신뢰도가 낮게 측정되도록 하는 등 변형이 가능하다.

2.3 연관성 정도의 판단

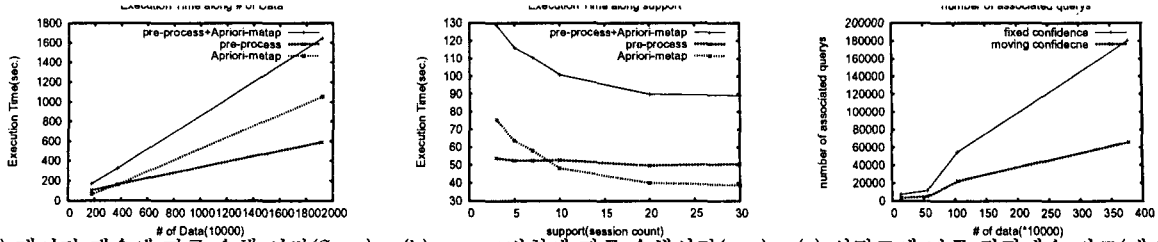
Apriori 알고리즘으로 연관된 검색어를 모두 찾은후, 연관법칙을 추출하게된다. 이 때, 연관성의 정도를 계산하여 연관성이 높은 것 순서대로 나열하도록 한다. 지금부터 연관성의 정도를 판단하기 위한 수치를 연관율 (Associated Rate) 이라고 부르기로 한다. 연관율을 구하기 위해서, 신뢰도(confidence), 상호의존도(interest), 클릭율(click rate)을 이용한다. 신뢰도 값이 크다는 것은 함께 검색된 횟수가 많다는 것이므로 연관율이 높음을 나타낸다. 상호의존도의 값이 크다는 것은 서로 독립성이 적다는 뜻이므로 이 또한 연관율이 높음을 나타내는 지표가 될 수 있다. 마지막으로 클릭율을 고려하기로 한다. 클릭율이란 검색어로 검색을 한뒤, 클릭을 한 횟수로서 사용자가 검색어에 대해 만족한 결과를 얻었다는 수치로 활용할 수 있다. 왜냐하면 만족한 결과가 많을 수록 클릭을 많이 하게 되기 때문이다. 클릭율을 이용하면 사용자에게 더욱 만족할 만한 연관어를 제공할 수 있게 된다. 다음 수식은 이 세가지 값을 이용한 연관율(Associated Rate)공식이다.

$$AssociatedRate = \alpha \cdot Confidence + \beta \cdot Interest + \gamma \cdot ClickRate$$

가중치를 조절 하므로서 각 값의 중요성을 조절 할 수 있다. 그런데, 지지도가 매우 낮을 경우 상호의존도의 값이 매우 커지므로, 상호의존도의 값에 bias되는 경우가 있다. 이를 해결하기 위해서 상호의존도의 값을 변형적으로 사용할 수 있다. 이것은 [7]를 참조하기 바란다.

제 3 절 성능 평가

모든 실험은 LINUX 운영체제에 메모리 1GigaByte, CPU Intel(R) Xeon(TM) CPU 3.00GHz인 Server장비에서 실행되었다.



(a) 데이타 개수에 따른 수행 시간(Secs.) (b)support변화에 따른 수행시간(sec.) (c) 신뢰도에 따른 결과개수 비교(개수)

그림 2: 실제 데이타 집합들에 대한 성능 평가

사용된 데이타는 국내 검색 서비스인 네이버에서 통합 검색에서 수집된 실제 데이타를 이용하였다.

실험에서는 변형된 Apriori알고리즘을 이용하였다.

- 전처리과정 (pre-process)
- Apriori 알고리즘의 변형(Apriori-metap)

그림 2 의 (a)는 세션의 개수(즉 data의 개수)에 따른 수행속도를 나타낸다. 최소지지도의 값은 5회를 이용하였다. 기존의 Apriori 알고리즘에서는 지지도 값을 전체 데이타개수의 %로 주지만, 이 경우에는 실험과 같이 5회, 10회 등과 같이 작은 수의 절대값으로 주고 실험을 한다. 작은 횟수가 검색되었다고 해도 의미를 가지기 때문이다. 선형적으로 수행속도가 증가하는 것을 볼 수 있다. (b)는 지지도의 변화에 따른 수행속도의 변화를 보여준다. 데이타의 개수는 100만개를 이용하여 실험하였다. 지지도가 감소함에 따라 수행속도가 느려지는 것을 확인 할 수 있다. (c)는 고정된 신뢰도를 사용한 경우와 이 논문에서 제시한 신뢰도를 사용한 경우에 추출되는 연관어의 개수를 비교한 그림이다. 모든 실험은 5회의 지지도를 공통적으로 이용하였다. 고정된 신뢰도를 사용한 경우 약 2 3배정도 많은 연관어가 추출되는 것을 볼 수 있는데 이 결과물에는 믿을만 하지 못한 데이타들이 많이 섞여 있다. 신뢰도가 고정되어 검색회수가 작은 검색어에 대해 잘못된 연관어가 구해진 것이다. 이 연구에서 제시한 신뢰도를 사용하면 매우 정제된 결과를 얻을 수 있음을 확인할 수 있다.

제 4 절 결론

본 연구에서는 검색엔진에서 수집된 검색어간의 연관법칙을 찾을 때 필요한 변형에 대해 다루었다. 검색어간의 연관법칙을 찾으면 보다 질높은 검색 서비스가 가능해지기 때문에 검색서비스를 제공하는 곳에서 유용하게 사용될 수 있다.

검색어의 연관법칙에서는 기존의 라지셋(*Large Set*)으로 분류되지 못하는 검색어도 중요한 의미를 갖게되므로 이를 처리하기 위하여 지지도(support)의 값을 매우 작게 설정하여야 하였다. 이로 인하여 신뢰도(confidence)값에 문제가 발생하여 유동적으로 이 값을 바꿈으로 해결하는 방법을 제시하였다. 또한, 연관어 간의 연관성을 수치화 함으로써 연관정도에 따라 연관어를 순서지을 수 있는 함수도 제시하였다.

참고 문헌

[1] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules," In *Proc. 1994 Int'l Conf. Very Large Data Bases (VLDB'94)*, pages 487-499, Santiago, Chile, September 1994.

[2] R. Agrawal and R. Srikant. "Mining sequential patterns," In *Proc. 1995 Int'l Conf. Data Engineering (ICDE'95)*, pages 3-14, Taipei, Taiwan, March 1995.

[3] R. Agrawal and R. Srikant. "Mining Generalized Association Rules," In *Proc. 21st Int'l Conf. Very Large Data Bases (VLDB'95)*, pages 3-16, Zurich, Swizerland, September 1995.

[4] M. Garofalakis, R. Rastogi, and K. Shim. "Spirit: Sequential pattern mining with regular expression constraints," In *Proc. 1999 Int'l Conf. Very Large Data Bases (VLDB'99)*, pages 223-234, Edinburgh, UK, September 1999.

[5] J. Han, J. Pei, and Y. Yin. "Mining frequent patterns without candidate generation," In *Proc. 2000 ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD'00)*, pages 1-12, Dallas, TX, May 2000.

[6] J.S. Park, M.S. Chen, and P.S. Yu. "An effective hash-based algorithm for mining association rules," In *Proc. 1995 ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD'95)*, pages 175-186, San Jose, CA, May 1995.

[7] Khalil M. Ahmed, Nagwa M. El-Makky, Yousry Taha. "A note on Beyond Market Basket: Generailizing Association Rules to Correaltions," *Proc. 2000 ACM SIGKDD Explorations. SIGKDD (SIGKDD'00)* , pages 46-48, Jan 2000.

[8] Mohammed J. Zaki, Ching-Jui Hsiao. "CHARM: An Efficient Algorithm for Closed Itemset Mining," *2002 SDM* , pages 02-27, 2002.