연관 관계와 TF*IDF를 이용한 검색 결과 Re-Ranking

이정훈○ · 전서현
동국대학교 컴퓨터공학
leeye123@naver.com○, shcheon55@dgu.edu

Re-ranking for Search result using association relationship and TF*IDF

Jung-Hun Lee○ Suh H. Cheon
Department of Computer Science and Engineering, Dongguk University

요 약

질의를 이용한 정보 검색 기술에서 단어 의미의 모호성에 의해 사용자가 검색 하고자 하는 주제 이외의 문서 까지 검색되고 있다. 이러한 문제는 모바일기기의 검색 환경에서 두드러진다. 모바일에서의 검색은 문서의 로딩속도가 느리며 작은 화면에 의해 스크롤이 잦다. 그러므로 원하는 검색 결과가 검색 첫 페이지 이외에 위치하거나, 또는 페이지 하단에 위치할 경우 검색 결과를 확인하는 대에 많은 시간과 노력이 필요하다. 이러한 문제를 해결하기위해선 단어 의미의 모호성을 해결하고 사용자가 검색하고자하는 주제의 검색결과를 검색 상위에 위치시킬 수 있는 방법을 필요로 한다. 이 연구에서는 연관 단어 추출과 TF*IDF를 이용하여, 검색결과를 re-ranking하는 방법을 제시한다.

1. 서 론

정보전달의 기술이 발달하면서 여러 매체를 통한 정보 전달 방법이 제공되고 있다. 하지만 검색 기술의 경우 아직 텍스트에 의한 검색이 대다수를 이룬다. 이러한 텍스트 기반 검색 환경은 기존의 PC를 이용한 검색환경을 기반으로 한 것이다. 하지만 모바일 기기의 발달에 의해 언제 어디서든 사용자가 원하는 곳에서 검색이 가능한 시대가 열리면서 기존의 검색 환경을 모바일 기기에서 적용하는 것에 문제가 발생되었다. 모바일 기기는 기존의 PC환경보다 문서의 로딩 속도가 느리며, 많은 데이터를 짧은 시간에 수신하기 어렵다. 또한 작은 화면으로 인해 원하는 검색 결과가 하단에 위치할 경우 잦은 스크롤로 인해 사용자의 불편을 야기한다. 그러므로 정확한 정보를 신속하게 검색하고 re-ranking을 이용하여 사용자가 원하는 정보를 우선 적으로 제공하는 방법이 필요하게 되었다. 또한 이러한 기술은 기존의 PC 검색 환경에서도 기하급수적인 정보의 증가에 의해 사용자가 필요로 하지 않는 정보까지 검색 결과 상위에 출현하여 사용자가 원하는 정보를 찾는데 더 많은 시간과 노력이 필요하게 되는 문제점의 해결방법이 될 수 있다.

re-ranking은 검색된 결과에서 사용자가 입력한 질의(query)와 연관성 있는 단어를 이용하여 문서의 순위를 재 정렬하는 것이다. 현재의 일반 적인 정보 검색 시스템은 수억 건 이상의 웹 문서를 데이터베이스화하여 사용자가 입력한 질의에 대해 유사도가 높은 문서를 보여주는 방식 이므로 하나의 질의에 의해 다양한 주제의 문서가 검색될 수 있기 때문에 연관 단어를 이용하여 검색결과를 재 정렬할 필요성이 있는 것이다.

이 연구에서는 질의와 질의어의 연관 단어의 관계를 파악하고 연관단어와 문서의 관계를 파악하여 질의와 관련된 문서의 순위를 측정하는 방식을 제시한다.

2. 관련 연구

re-ranking을 위한 방법은 연관 단어를 이용하여 질의를 확장하는 방법, 질의와 문서의 관계를 이용하는 방법 그리고 문서와 문서간의 관계를 이용하여 문서의 순위를 정하는 방법이 있다.

2.1 연관 단어 추출 이용한 re-ranking

검색된 문서에서 질의어와 연관성을 가진 단어를 이용하여, 질의를 확장하고 확장된 질의를 이용하여 re-ranking하는 것이다.

2.1.1 연관 규칙 알고리즘

연관 규칙 알고리즘 단어의 연관성을 이용하여, 문서에서 연관 단어를 추출 하는 방법이다. 연관규칙 마이닝에서 용어간의 연관관계(association relationship)는 항목들 사이에 존재하는 유사성 또는 패턴을 의미하는 것으로 한 단락이나 한 문장을 하나의 트랜잭션으로 하여 용어간의 연관성을 측정하는 것이다.

Aporiori 알고리즘은 Agrawal and Srikant[1]이 제안한 것으로 후보항목 집합을 생성하고, 발생 빈도를 계산한 후 사용자가 정의한 최소지지도를 가진 빈발 항목집합을 결정하는 것이다. Aporiori 알고리즘은 각 패스에서 빈발 항목집합들의 후보 항목집합을 구성한 후에 각후보 항목집합의 발생 빈도를 계산하고, 사용자가 정의한 최소지지도를 기준으로 하여 빈발 항목집합들을 결정한다. 다음 단계에서는 이들 빈발 항목집합들로부터 최소신뢰도 임계치를 만족하는 연관규칙을 모두 찾는다. 최소신뢰도 임계치는 발생된 항목집합에서 사용자가 지정한 최소한의 빈도수를 뜻한다. 최소한의 빈도수를 넘지 못하는 경우 항목집단에서 제외된다. 하지만 문서마다 단어의 발생 빈도 및 텍스트 정보의 양이 다르므로 최소신뢰도 임

계치를 고정 값으로 사용하는 것은 단어 추출의 한계를 보인다. 또한 빈번하게 같은 문장 또는 문서에서 발생하였다 하여 연관단어 라고 단정 지을 수 없다.

2.1.2 TF-IDF

TF-IDF(term frequency-inverse document frequency)는 여러 문서로 이루어진 문서집단이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다. 기계적 학습 방식이 아니지만 유사한 성능을 보인다[2]. 하지만 단어 사이의 연관성을 고려하지 않았을 경우 또는 문서의 양이 기하급수적으로 많아질 경우 핵심어가 아닌 단지 문서를 구분 지을 수 있는 용어만이 추출 된다.

2.2 단어와 문서간의 연관 관계를 이용한 re-ranking

질의어를 이용하여 유사도를 측정하고 유사도를 가중치로한 re-ranking 방식이다. 질의와 문서간의 직접적인 관계를 측정한다.

2.2.1 HITS (Hub-Authority model)

HITS 알고리즘[3]은 문서간의 링크(link)를 통해 문서의 가중치 값을 전파하여 Hub 문서와 Authority 문서를 구분 지어 좋은 Hub과 Authority문서를 찾는 것에서 시작된다. 이 방식을 Yuanhua Lv[4]는 단어와 문서간의 관계 측정에 이용하였다. 단어를 Hub으로 이용하고 문서를 Authority로 하여, 링크에 가중치를 부여하여 Hub의 가중치 값을 조절하는 방식이다. 이것은 Pagerank와 유사한 방법이다.

단어와 문서간의 관계는 단어가 문서에 속하였는지를 이용하는 것으로 오로지 단어에서 문서를 가리키는 단방향 형태이다. 이러한 형태는 오로지 단어에서 문서로만 가중치를 전파하므로 단어가 Authority가 되거나 문서가 Hub가 되는 경우는 발생되지 않는다. 오로지 단어는 Hub의 기능만을 가지고 문서는 Authority의 기능만을 가진다. 이러한 단순한 관계로 인해 단어와 문서 간의 직접적인 관계를 파악할 수 있지만 일방적인 가중치 전파로인해 Hub과 Authority값은 무한으로 상승한다. 이것은 기존의 HITS 처럼 Loop에 의한 가중치의 변화가 안정기에 들어서서 크게 변동되지 않는 현상이 발생되지 않는 것이다. 또한 문서와 단어 사이에 링크의 가중치에 의해 특정 문서만 전파되는 값을 독식하게 되므로 새로 발생한 문서 또는 단어 등은 기존의 문서 또는 단어만큼의 가중치 값이 쌓일 때까지 오랜 시간이 허비된다.

2.2.2 TF*IDF

기존의 TF-IDF는 오로지 문서 집합에서 특정문서를 구분지을 수 있는 핵심어를 추출하였다. 이 방식은 같은 단어가 모든 문서에서 서로 다른 값을 가진다. 즉, 단어의 가중치를 이용하여 문서를 분류할 수 있는 것이다. 하지만 단지 문서만을 구분 지을 뿐 단어와 문서의 유사도를 측정할 수는 없다. TF*IDF는 질이어 벡터(vector)와 문서의 단어 벡터를 이용하여 코사인(Cosine)유사도 값을 측정하는 방식이다. 실제 문서에서 발생된 질의어의 가중치는 TF-IDF 방식으로 측정한다. 이 방식 또한 기계적 학습방법과 유사한 성능을 보인다. 또한 문서와 질의어 간의 직접적인 관계를 파악하여 유사성을 측정하므로 문서의 re-ranking에 적합하다. 하지만 질의어의 단어 수가 적거나 질의어로 사용된 단어가 모든 문서에서 분포되지 않을 경우 유사도 측정에 의미가 없다. 즉, 질의어 수가 적을 경우 TF-IDF와 유사해 진다는 것이다.

2.3 문서와 문서간의 관계를 이용한 re-ranking

사용자가 방문한 문서 또는 검색 결과로 제시된 문서간의 링크 또는 연관성을 고려하여 가중치를 측정하고 re-ranking하는 방식이다.

2.3.1 PageRank

Sergey Brin[5]이 제시한 방식으로 문서의 가중치를 링크를 통하여 전파하고 Loop를 통해 가중치를 재정이 하는 방식으로 Loop를 실행할 때마다 값이 안정기에 들어서게 된다. 이것은 HITS 알고리즘을 기반으로 한 것으로 현재까지 문서와 문서간의 관계 측정에 좋은 성능을 보였다. 하지만 PageRank 역시 새로운 문서 또는 유용하지만 링크가 많이 연결되지 못한 문서의 경우 가중치 분배가 정상적으로 이루어 지지 못한다. 또한 seed page를 설정하는 것 또한 문제가 될수 있다.

2.3.2 클러스터(cluster)

클러스터는 검색 결과를 좌표로 분류하고 좌표와 좌표 사이의 거리를 측정하여, 거리가 가까운 좌표를 묶어 나가는 방식을 사용한다. 좌표의 군집 중 질의와 유사도가 높은 군집을 검색 결과 상위로 제시하는 방식이다. 하지만 클러스터 자체의 성능이 기계학습을 이용한 방식보다 성능이 낮으며, 클러스터의 차원(Dimension)이 늘어남에 따라 성능의 문제가 발생된다.

3. 연관 단어 확장을 이용한 TF*IDF

이 연구에서는 질의어와 검색결과 간의 직접적인 관계를 파악하지 않고 질의어와 단어들 간의 연관성을 파악하고 연관 단어와 문서간의 연관 관계를 파악하여 질의어와 문서간의 연관 관계를 파악한다. 또한 연관 관계 파악에서 단어와 문서 같은 객체의 방향성을 고려하도록 한다. 질의어와 연관 단어 사이의 관계성 파악을 위하여 HITS를 이용하고, 연관단어와 문서 사이의 관계성 파악을 위하여 TF*IDF를 이용한다.

단어와 단어는 서로 양방향성의 성향을 가진다. 즉, 단어와 단어 사이의 모든 관계는 서로 양방향을 이루는 그래프(Graph)라는 것이다. 2.2.1의 방식처럼 HITS를 서로

다른 객체의 관계를 파악할 경우 단방향성에의해 가중치가 쌓이기만 하는 결과를 가진다. 하지만 단어의 유용성 측정처럼 동일한 객체에 대한 유용성 측정에는 서로의 연관 관계에 의해 그래프를 이루어서, 가중치의 전파가 Loop를 통하여 자신에게 다시 부여되는 방식으로 가중치가 안정기를 가지게 되는 것이다. 그림 1은 이러한 가중치 변화를 그래프로 나타낸 것이다.

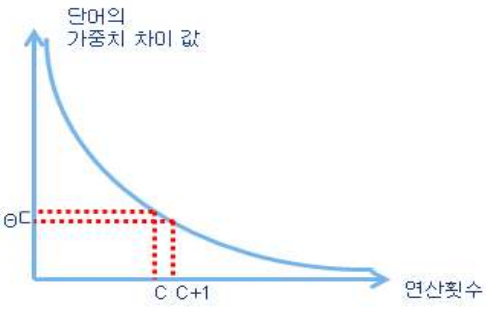

그림 1 연산 횟수에 따른 가중치 변화

그림 1의 연산횟수는 Loop의 횟수를 뜻하며, Θ는 연산횟수가 C일 때 단어 가중치 값과 C+1 에서의 가중치 값의 차이 값이다. 이 Θ가 일정 기중치 값(Threshold)보다 낮으면 Loop를 멈춘다. 다만 모든 단어는 양방향성을 가지므로 모든 단어가 Hub이며 또한 Authority가 된다. 그러므로 연결된 단어의 Hub 과 Authority값을 링크마다 똑같이 전파하지 않고 그림 2의 PageRank와 같이 1/(단어가 연결된 모든 링크 수)를 이용하도록 한다.

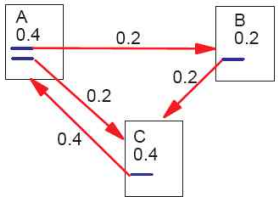

그림 2 가중치 전파

문서와 단어 사이의 연관성 측정은 연관단어를 질의어 벡터로 이용한 TF*IDF를 이용한다. 질의어의 연관 단어를 이용하므로 충분한 질의어 벡터를 보유 하게 되며, 기계적 학습 없이 유사한 성능을 보인다.

TF*IDF 의 단어 가중치 계산 TF-IDF에서 단어 또는 문서의 발생 빈도를 이용하지 않고 HITS 알고리즘의 유용성 값을 이용한다.

$$TF-IDF(d,w) = r(d,w)*weight(w)$$
$$weight(w) = \log(1+N/f(w)) \qquad \text{수식 1}$$
$$r(d,w) = 1+\log(f(d,w))$$

N: 총 문서들의 weight값의 합
w: 단어
d: 문서
f(d, w): d에서 키워드 w의 HITS 유용성 값
r(d, w): d와 w 사이의 연관성을 수치화.
f(w): 단어 w가 등장하는 문서들의 weight 값
weight(w): 단어 w의 가중치

수식 1은 HITS 알고리즘의 유용성 값을 이용한 TF-IDF 연산식을 나타낸 것이다. 질의어 연관 단어의 벡터를 Q={w_1,w_2,w_3}라고 하고 문서에서 연관 단어들의 벡터 값을 d_1, d_2, d_3, ... , d_j 라고 했을 경우 이것을 좌표로 나타내면 그림 3과 같다. 단 연관 단어 중에서 문서에 출현되지 않는 단어의 가중치 값은 0이된다.

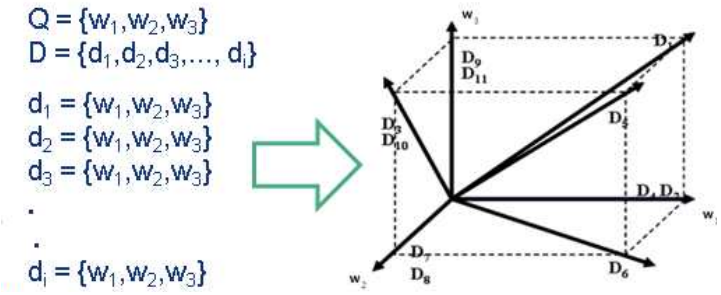

그림 3 단어를 이용한 문서 좌표 표현

그림 3의 내용을 기반으로 수식 2의 코사인 유사도를 측정하면 그림 4와 같은 결과를 얻게 된다.

$$sim(Q,D_i) = \sum_{j=1}^{t} w_{qj} * w_{d_{ij}} \qquad \text{수식 2}$$

유사도 값은 그림 4의 α를 이용하여 질의 벡터와의 유사도를 측정하는 것이다.

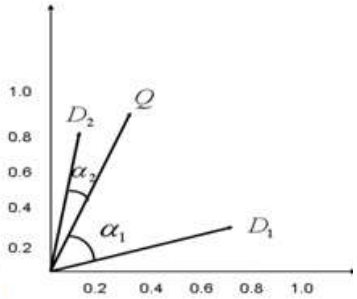

그림 4 질의 벡터와의 유사도 측정

측정된 유사도 값을 이용하여 문서를 re-ranking 한다.

4. 실 험

실험을 위하여 인위적인 문서를 이용하여 HITS, TF*IDF, HTID(연관단어 확장을 이용한 TF*IDF)를 NDCG 성능 측정 방식을 이용하여 평가한다. 기존의 정확도(precision) 그래프는 검색 결과의 랭킹과 관련 없이 상위 N개의 문서에 관련 문서가 포함되었는지를 측정하는 방식 이다. 하지만 검색 엔진의 결과라는 것은 사용자의 입장에서 상위에 관련 문서가 다수 위치하는 것이 가장 이상적이다. 즉, 정확도 및 랭킹 까지도 성능 실험에 포함 되어야 한다. NDCG는 정확도 및 문서의 랭킹을 반영하여 성능을 평가함으로 NDCG를 이용하여 성능을 평가한다. NDCG 평가 값의 최대값은 1이다. IDCG를 위한 문서의 평가 가중치는 실험 때마다 랜덤하게 가중치를 뽑

아 사용하며, TF-IDF 계산을 위한 발생빈도 수는 랜덤 값을 이용하였다. HITS에서 단어의 초기 가중치는 모두 1에서 시작한다.

그림 5는 HTID의 성능을 평가한 것이다. 가로축은 평가 값이 큰 알고리즘을 나타내고, 세로축은 NDCG의 차이 값을 나타낸 것이다. 세로축의 최대값은 1이다. HITS>HTID는 HITS의 NDCG 평가 값이 클 경우에 HTID와 HITS의 평가 값 차이의 평균이며, HITS<HTID는 HTID가 HITS보다 평가 값이 클 경우의 평가 값 차이의 평균이다. HTID가 HITS에 비하여 NDCG성능이 뛰어난 것을 보인다.

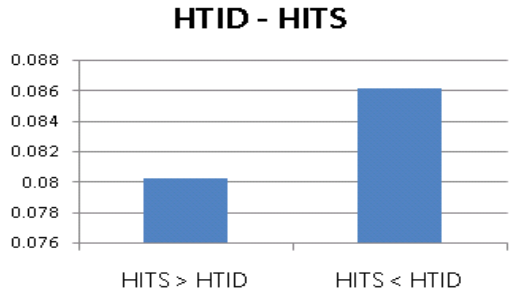

그림 5 HTID와 HITS의 성능 비교

그림 6은 그림 5와 같은 비교방식으로 HTID와 TF*IDF의 성능을 비교한 것이다. 그림 6과 같이 HTID가 TF*IDF보다 평균적으로 높은 성능을 보인다.

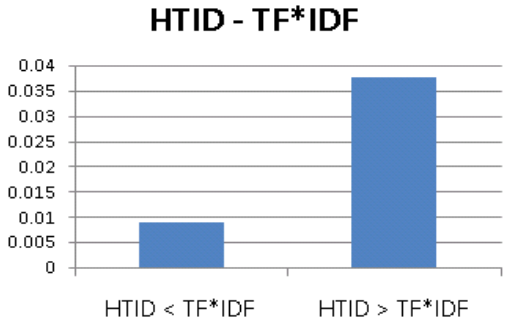

그림 6 HTID와 TF*IDF의 성능 비교

그림 7은 각 알고리즘의 성능의 평균을 나타낸 것이다. HTID가 우수한 성능을 보인다. 가로축은 각 알고리즘을 나타내며, 세로축은 각 알고리즘의 평균 NDCG 성능 평가 값이며, 최대값은 1이 된다.

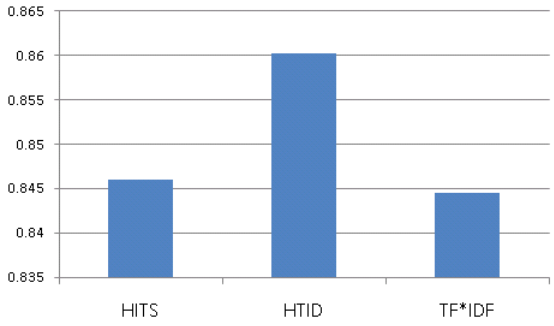

그림 7 NDCG 성능 평가 평균

5. 결 과

본 논문에서 제시한 HTID의 방식이 TF*IDF 또는 HITS보다 높은 성능을 나타내었다. 이것은 두 개의 알고리즘을 객체의 방향성과 관계에 특화된 방식을 사용하여 기존의 알고리즘을 단독으로 사용한 것 보다 높은 성능을 보인 것이다. 또한 수식에 의한 단순 계산 값을 이용하여 TF*IDF로 통계 값을 추출하는 것이므로 알고리즘을 단독으로 사용하는 것과 같은 처리속도를 가진다. 이러한 re-ranking 방식으로 사용자가 필요로하는 정보를 효과적으로 검색 결과 상위에 제시 하여 검색의 성능을 높이게 되는 것이다.

6. 향후과제

실험을 통해 기존의 알고리즘보다 높은 성능을 나타내는 것을 증명하였다. 하지만 HTID가 항상 높은 성능을 보이는 것은 아니다. 그림 5 또는 그림 6과같이 기존의 알고리즘이 더 높은 성능을 보이는 경우도 발생 한다. 더 높은 성능을 보이는 경우의 횟수를 줄이기 위하여 사용자의 프로파일을 사용하거나, 가중치 조절하는 방식이 필요할 것이다.

참 고 문 헌

[1] Agrawal, R., and Srikant, R. Fast Algorithms for Mining Association Rules. Proceeding of the 20th International Conference on Very Large Databases, pp.487-499, 1994.

[2] Salton, Michael J. McGill. Introduction to modern information retrieval. McGraw-Hill. separate volume. 1983.

[3] Kleinberg, Jon. Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5): 604-632. 1999.

[4] Y. Lv, L. Sun, J.-Y. Nie, and W. Z. Wan Chen. An iterative implicit feedback approach to personalized search. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 585-592. 2006.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report. Stanford University, 1998.