HITS 알고리즘을 이용한 단어 연관 관계 링크 제어

문성천 ○ 이정훈 전서현
동국대학교 컴퓨터공학
Kornoon84@nate.com○, leeye123@naver.com, shcheon55@dgu.edu

A Link control of the word associated relation with using HITS Algorithm

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

요 약

많은 정보들을 인터넷을 통하여 접할 수 있게 됨에 따라 사용자가 만족하는 결과를 보여주는 것이 검색 엔진의 궁극적인 목표가 되었다. 하지만 방대한 양을 가진 다양한 정보에서 원하는 검색 결과를 검색하는 것은 과거와 현재까지 많은 연구를 통해 많은 시간과 노력이 필요하다는 것이 증명 되었다. 기존의 HITS 알고리즘을 개선하여 링크 제어를 이용한 페이지와 페이지간에 관련성을 높였다.

1. 서 론

전통적인 추천기법에는 컨텐츠 기반의 추천 방법과 상호 추천을 수행하는 협력적 추천 방법이 있다. 검색 엔진이 발전하면서 명시적인 방법과 묵시적인 방법이 등장하였고, 명시적인 방법은 개인정보나 관심정보를 이용하여 사용자의 관심정보를 빠르게 취득 할 수 있는 장점이 있는가 하면 동적으로 변화하는 사용자의 취향을 반영하기가 어렵다. 그리고 묵시적인 방법은 관심을 보인 행동을 기반으로 사용자의 관심을 추론하거나 일정 기간 사용자가 방문하여 이용해야만 사용자의 관심 정보가 분석 가능하다. 현재 상용화된 많은 검색 엔진이 있지만 개인에 맞는 검색 결과를 보여주는 것에는 한계가 있다. 하나의 단어가 여러 가지의 의미를 가질 수 있고 개인마다 의도하는 검색 결과가 다르기 때문이다. 사용자에게 더 정확한 검색 결과를 보여주기 위해 질의어 확장이나 사용자의 기호에 따라 검색 결과의 순위를 재조정하는 등의 연구가 진행 되고 있다.[1] 많은 웹 검색 기법들 중에서도 HITS(Hypertext induced topic Selection)[2] 알고리즘을 이용하여 검색 엔진 성능을 높이는 여러 연구가 진행 중이다. 하지만 이 알고리즘은 Authority값과 Hub값을 계산하는데 있어서 문제점이 있다. 첫 번째 문제는 새로운 문서일 경우 기존의 문서들과 링크가 연결되어야만 유용성을 측정할 수 있게 되다. 그런데 기존의 문서들과의 링크 생성에 시간이 많이 걸린다는 것이다. 두 번째 문제점은 처음 질의어와 맞지 않는 페이지가 링크로 연결될 수 있다는 것이다.

이 연구에서는 연관 단어 링크를 HITS를 이용하여 가중치를 측정한다. 또한 HITS의 문제점을 해결하기 위해 Clustring과 Kullback-Leibler(KL)을 사용하여 Authority 값과 Hub값을 연관 단어 별로 군집화하고 Kullback-Leibler(KL)을 이용하여 링크의 수를 제한하여 관련 없는 링크의 수를 줄여 HITS알고리즘의 성능을 높힌다.

2. 관련 연구

2.1 연관 단어 추출

단어의 연관 관계를 이용하여 문서에서 단어를 추출하는 연관 규칙을 이용한 방식과 단어가 문서를 다른 문서로부터 분류할 수 있는지를 계산하는 방식이 연구 되고 있다.

2.1.1 Apriori 알고리즘

Apriori알고리즘은 연관 규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발됐고 또 가장 많이 쓰인다. 이 알고리즘은 두 가지 단계로 구성된다. 우선 첫 번째 단계에서는 최소 지지도(Thresold) 설정 값에 따라 빈도수가 높은 항목의 집합들을 찾아내고 두 번째 단계에서는 이들 집합들로부터 신뢰도 설정 값을 모두 계산한다. 그리고 Apriori 알고리즘의 장점은 비교적 학습 속도가 빠르고 분류 규칙을 단순하고 이해하기 쉽게 변환 할 수 있다. 하지만 Apriori 알고리즘은 연관 단어를 추출하지만 최소 지지도 설정 값을 유동적으로 변화시키기 어렵다. 그에 따른 연관 단어 별 가중치 계산에 매우 취약하다.

2.1.2 TF-IDF 가중치 모델

TF-IDF(term frequency-inverse document frequency) 가중치 모델은 언어 자료 내의 특정 문서에서 어떤 단어의 중요도를 평가하기 위해 사용되는 통계적인 수치이다. TF-IDF는 정보검색과 텍스트 마이닝에 주로 사용되며, 검색 엔진에서 사용자 질의에 대한 문서의 유사도 순위를 정하는데 이 TF-IDF 가중치 모델의 변형들이 사용된다. 그 결과 TF-IDF 가중치 모델만으로도 원하는 검색 결과를 얻을 수 있다. 이 방법은 기계적인 학습 방법만큼의 좋은 성능을 보이지만 단어의 연관 관계를 전혀 고려하지 않는다.

2.1.3 단어 가중치 측정을 위한 HITS알고리즘

HITS(Hypertext Induced Topic Selection) 알고리즘은 링크 구조를 이용해서 Hub과 Authorities를 찾는 알고리즘이다. 연구자들에게 Hub라고 불리는 웹페이지들의 또 다른 중요한 카테고리들을 고려하도록 유도 하였다. Hub은 권한이 있는(Authorities)에 대한 링크의 집합을 제공하는 웹페이지 혹은 집합이다. 어떤 주제 에 관한 Authorities들을 많이 링크하고 있는 페이지를 중심축 역할을 한다는 의미에서 Hub이라고 한다. 그리 고 좋은 Authorites는 in-dgree가 높다는 점과 함께 많 은 Hub들로부터 링크 되어 있다는 공통점이 있다. <그림 1>과 같이 Hub 노드와 Authorites 노드는 링크 로 연결되어 있다. 좋은 Authorities는 좋은 Hub을 이용해서 찾아낼 수 있고 좋은 Hub은 좋은 Authorities를 통해서 찾아낼 수 있다.

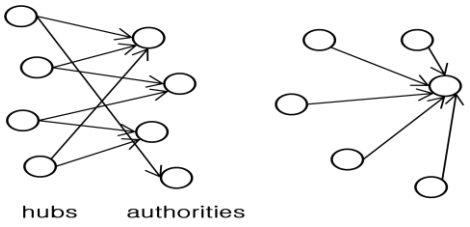

그림 1 Hubs과 Authorites의 링크 연결

예를 들어 연결 되어 있는 링크의 수가 많을수록 좋은 Hub와 좋은 Authorities이다. HITS 알고리즘은 문서 가중치 계산에 사용 되었지만 단어간의 관계를 이용하하여 가중치를 전파한다면 단어 가중치로도 쓸 수 있다.[2][3]

2.2 문서 분류

문서 분류를 이용하여 방대한 양의 온라인 문서의 레이블을 할당하며, 주제 디렉토리를 구성할 수 있다. 또한, 문서 집필 형태와 이후의 분석을 용이하게 하기 위해 문서들의 집합과 관련된 하이퍼링크들의 목적을 구분할 수 있다.[4]

2.2.1 클러스터를 이용한 문서 분류

클러스터는 대규모 데이터 집합을 유사성에 따라서 그룹들로 분할 한다. 문서 클러스터링은 높은 연관성을 갖는 문서들을 그룹화 시킨다. 그룹화 시킨 문서 별로 유사도를 계산하여 상호 관련성 여부를 확인한다.[5] 예를 들어 k-NN 알고리즘[4]처럼 벡터공간 모형에 유사한 문서 벡터를 공유하면 두 문서는 유사하다. 같은 클래스 레이블을 할당된 문서는 유사할 가능성이 높다고 볼 수 있다.

2.2.2 TF*IDF를 이용한 문서 분류

TF*IDF모델은 벡터 공간모델(Vector Space Model) 기반의 정보 검색을 위해서 문서를 표현하는 모델이다. 기본적으로 개별 문서에서 각 단어의 상대적 중요도를 표현 할 수 있어서 개별 문서에 존재하는 키워드 추출은 개별 문서에 존재하는 키워드를 추출하는데 활용하고 있다.

3. 연관 단어 그래프 추출

3.1.1 단어 그래프의 필요성

문서에서 추출된 단어는 서로 연관성을 가진다. 대부분의 논문에서는 질의와 문서간의 관계만 생각하지만 실질적으로 질의와 연관 단어와 문서의 관계를 따져야 한다.

여기서 문제는 연관 단어라는 것이 서로 연결되어 있으며, 하나의 단어가 하나의 단어를 가리키는 것이 아니고 단어가 서로를 가리키는 양방향성을 고려 했을 때 모든 단어가 서로 연결되어 있다면 가중치 분배는 적합하지 않게 된다.

그러므로 본 논문에서는 단어 그래프에서 서로간의 링크를 제거 했을 경우와 제거하지 않았을 경우에 대해서 분석 한다.

3.1.2 연관 단어 분류

연관 단어 추출을 위해 모든 문서 마다 연관 단어를 추출하지 않고 질의를 통해 검색된 문서를 클러스터로 군집화 시킨다. 연관 단어 추출은 하나의 군집을 대상으로 질의와 관련된 연관 단어를 추출한다.

3.1.3 연관 단어 가중치 추출

연관 단어의 추출에 가장 좋은 성능을 내는 것은 Apriori 알고리즘이다. 하지만 이 알고리즘은 연관성만

고려할 뿐 가중치 측정에는 적합하지 않다. 그러므로 연관 단어를 추출하여 단어의 연관성을 고려해 HITS 알고리즘을 이용해서 가중치를 계산한다. 하지만 모든 단어가 서로 링크되면 단어는 양방향성을 가지므로 모든 단어가 같은 가중치를 가지게 된다. 그러므로 논문과 같이 클러스터로 <그림 2>와 같이 분류한다.

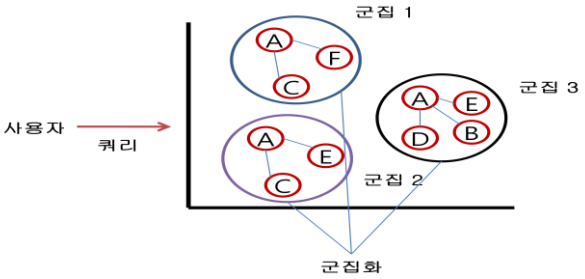

그림 2. 질의를 이용한 연관 단어 문서 분류 방식

군집에서 연관 단어를 추출하므로 일정 이상의 단어가 서로 연결되는 경우는 피할 수 있다. 다만 질의의 경우 모든 노드와 연결이 되므로 최상위 시작 노드가 될 수 있지만 질의의 경우 모든 노드와 연결된 것이 문제가 될 수 있다. 또한, 질의의 연관단어를 통해 다른 단어가 질의와 간접적으로 연관성을 가지게 될 수 있다. 이 경우 또한 결국 모든 단어가 연결되는 결과를 발생한다. 그러므로 일정 불필요한 링크를 끊어서 질의의 주제를 세분화할 수 있도록 링크를 제거하는 방법이 필요하다.

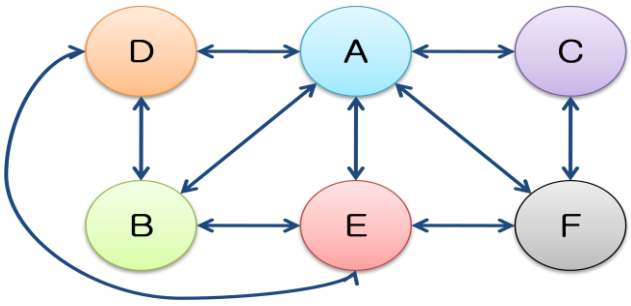

그림 3. 연관 단어 그래프

3.1.4 링크 제거

링크를 제거하기 위해서는 노드 사이의 관련성을 측정해야 한다. 노드와 직접적으로 연결된다거나 간접적으로 연결되는 것을 통해 최적화된 노드 상태를 찾아낸다. 노드를 최적화하기 위해 Kullback-Leibler(KL)을 사용한다. Kullback-Leibler(KL) 발산은 정보이론에서 사용되며 상대적 엔트로피를 측정하는 방법 중 하나이다.

표 1. 링크의 가중치 측정

단어	링크의 수	링크의 가중치
A	5	(A의 가중치)/5
B	3	(B의 가중치)/3
C	2	(C의 가중치)/2
D	3	(D의 가중치)/3
E	4	(E의 가중치)/4
F	3	(F의 가중치)/3

<표 1>은 <그림 3>의 그래프를 기준으로 각각의 단어가 연결된 링크의 가중치를 측정한 것이다. 예를 들어 단어 A는 연결된 링크가 5개 이므로 하나의 링크는 A의 가중치의 1/5만을 전파하게 된다.

4. 실 험

이 연구의 정당성을 증명하기 위해 단어의 연관성 그래프에 대한 실험을 실행한다. 실험을 위해 질의어를 입력하여 문서를 추출하고, 군집에서 연관 단어를 추출한다. 추출된 연관 단어를 링크 제거하지 않고 사용하였을 경우와 링크가 제거 되었을 때의 성능을 비교하는 것이다. 링크의 관계를 제외한 모든 실험 환경은 동일하게 구성 한다. 각 군집화 시킨 단어들의 가중치 순위를 <표 2>와 같이 적용하였다. 군집에서 랭킹이 높은 단어는 문서 집합에서 해당 군집을 가장 효율적으로 분류하는 단어의 순위이다. 즉, 단어의 랭킹이 높을수록 군집을 잘 분류한다.

표 2 군집에 따른 단어의 순위

랭킹 / 군집	1	2	3	4	5	6
군집1	F	C	A	D	E	B
군집2	C	E	A	D	F	B
군집3	B	D	E	A	F	C

<그림 4>는 <그림 3> 그래프에서 일정 링크가 제거된 그래프를 나타낸다. <그림 4>와 <그림 3>의 그래프를 HITS를 이용하여 단어의 유용성을 측정하고 <표 2>를 기준으로 NDCG 방법을 이용하여 평가한다. NDCG는 야후에서 만든 검색 퀄리티 측정 방법으로 검색엔진의 성능을 측정하는 한 방법이다.

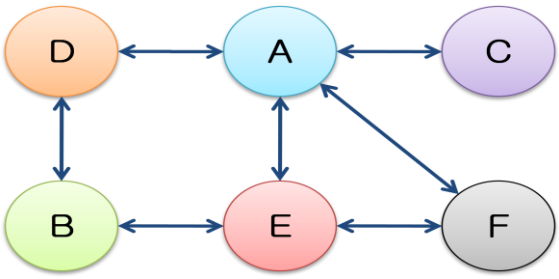

그림 4. 연관 단어 그래프의 링크 제어

실험에서는 단어 노드의 가중치를 측정하기 위해 HITS에서 Hubs과 Authorities 중에서 Authorities값을 이용한다. Hubs로써 가중치가 높은 단어는 다른 Authorities를 가리키는 노드로써 많은 노드를 가리키는 단어는 그만큼 많은 단어와 연관성을 가진다. 즉, 여러 주제와 연관된 단어라는 것이다. 이런 단어일 경우 하나의 질의에 대해 공통적인 관심사를 찾아 낼 경우 적합하다. 이 연구의 경우 단어의 모호성을 해결하기 위한 것으로 주제를 세분화할 수 있는 Authorities 값으로 노드의 가중치를 측정한다. Authorities는 다른 노드에서 연결 되어지는 노드로써 질의에 세분화된 연관 주제를 가리키는 노드이다.

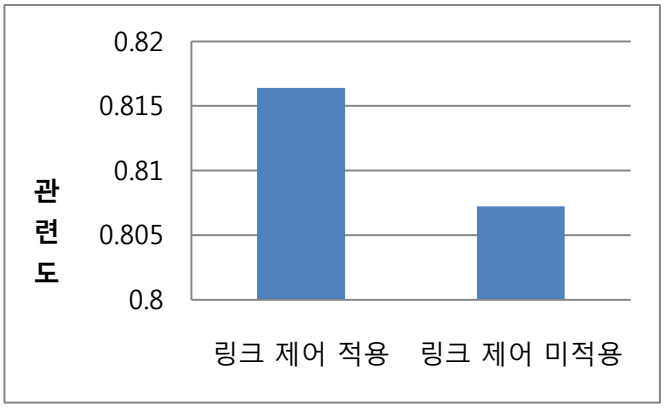

그림 5. NDCG 평균 성능 평가

<그림 5>는 <그림 3>과 <그림 4>의 그래프를 HITS로 가중치를 측정하고 단어의 순위를 NDCG를 측정 하였을 때의 평균 값이다. <그림 5>의 세로는 NDCG 값으로 0과 1사이의 값을 가진다. 가로의 링크 제어 적용은 <그림 4>를 뜻하며, 링크 제어 미적용은 <그림 3>을 뜻한다. 링크 연관 관계를 변화 시키며, HITS 알고리즘을 이용하여 단어의 가중치를 측정하고 <표 2>의 기준을 이용하여 NDCG값을 측정한 평균값이다. 성능 평가 결과 질의어에 대한 검색 결과가 관련 되어 있지 않은 링크를 제어를 했을 때 <그림 5>와 같이 본 논문에서 제시한 링크 제어 방법의 성능이 높은 것을 보인다.

5. 결 론

연관 단어 추출에서 단어의 연결 방향성을 고려하여 그래프를 만들고 모든 노드가 연결 되는 것을 방지하기 위해 군집을 이용하여 연관 단어를 추출 하였다. 또한 주제의 세분화를 위하여 KL 연산 식을 이용하여 정보량으로 단어간의 연관 관계를 고려하여 링크를 제거 하였다. 이를 통해 최적화된 그래프를 만들고 HITS를 이용해서 가중치를 계산하였다. 이것은 특정 군집을 효율적으로 분류하는 단어를 측정하는 방식인 TF-IDF와 유사하지만 단어의 연결성을 고려한다는 측면에서 뚜렷한 차이를 보인다. 추출한 연관 단어를 이용해서 질의를 확장할 수 있으며 검색 결과를 re-ranking 할 수 있게 된다.

6. 향후 연구 과제

연결성 제어를 이용한 HITS 알고리즘을 이용하여 군집에서 추출한 연관 단어의 가중치를 측정 하였다. 그 결과 검색 결과의 효율을 높일 수 있는 연관 단어를 추출할 수 있었다. 하지만, 실질적인 대규모 문서일 경우 단어의 연관성이 복잡해지며 그래프의 크기가 기하 급수적으로 늘어난다. 이러한 경우 링크의 연관성 측정이 전체 시스템의 처리 속도에 치명적인 문제가 될 수 있다. 또한 링크의 제거를 위해 일반적인 정보량 측정 이외에 링크의 제거 방법이 필요하다.

참고 문헌

[1] Sanghyun Ryu, Seunghwa Lee, Minchul Jung, Eunseok Lee, 'An Effective User-Profile Generation Method based on Identification of Informative Blocks in Web document', Sungkyunkwan University Dept. of Information and Communication Engineering, 2007.

[2] Jon M. Kleinberg,'Authoritative sources in a hyperlinked environment', Cornell Univ, In Proc.ACM SIAM Int. Conf, 1998.

[3] Yuanhua Lv, Le Sun, Junlin Zhang, Jian-Yun Nie, Wan Chen, Wei Zhang 2, 'An Iterative Implicit Feedback Approach to Personalized Search', Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp 585-592, 2006.

[4] Jiwei Han, Micheline Kamber, 'Data Mining : Concepts Techniques', pp.592-596, 2007.

[5] 김동희, 주길홍, 최진탁, '대용량 학습문서 관리를 위한 효율적인 문서 클러스터링 기법', 한국정보기술학회 하계학술대회, 2006.