링크기반 문서구조를 고려한 검색어 가중치 함수

Term Weighting Function for Link-based Document Structures

박재휘(Jaehui Park)*, 양병주(ByoungJu Yang)**, 이상구(Sang-goo Lee)***
jaehui@europa.snu.ac.kr, bjyang@europa.snu.ac.kr, sglee@europa.snu.ac.kr

초 록

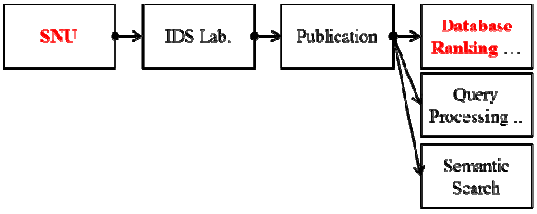
최근, 웹을 통한 상품 거래에 있어서 단일 문서가 아닌, 연관 상품 카탈로그, 사용자 거래 명세 등을 비롯한 연결 관계를 가진 문서의 집합이 의미 있는 정보로 활용되고 있다. 하지만, 기존 검색어 기반 가중치 함수는 검색 결과 단위를 하나의 문서로만 가정하고 있기 때문에 하이퍼링크와 같은 링크 구조에 기반하여 연결된 문서 집합에 대해서는 효과적인 가중치 값을 구하는데 어려움이 있다. 이 논문은 링크기반 문서구조를 고려한 새로운 검색어 가중치 함수를 제안한다. 먼저, 검색어가 포함된 문서 집합 내 링크 구조와 문서의 중복을 고려하여 새로운 정규화 지수를 정의한다. 그리고 정의된 정규화 지수를 사용하여 기존의 TF-IDF 함수를 확장하여 링크기반 문서구조에 적합한 새로운 검색어 가중치 함수를 제안한다.

1. 서론

최근 웹을 통한 전자 상거래가 활발해지면서, 하이퍼링크와 같은 문서간 연결 구조를 활용하는 다양한 검색 응용 (예를 들어, 연관 상품 검색, 소셜 네트워크 기반 검색 등) 프로그램이 늘어나고 있다. PageRank [1]와 같이 그래프 탐색 모델에 기반한 랭킹 알고리즘을 비롯하여, 연결된 문서간 근접성 [2]을 고려한 랭킹 알고리즘 등 다양한 접근 방식들이 문서간 연결 구조를 고려한 검색을 시도하여 왔다. 이 논문에서는 검색 결과의 단위가 하나의 문서가 아닌 연결된 문서들의 집합인 경우를 고려하고자 한다. 이 문제는 위에서 제시한 검색 응용들과 상호보완적인 관계에 있다. 검색어로 제시된 단어들이 하나의 문서에 모두 포함되지 않고, 다수의 연결된 문서에 분산되어 나타나는 경우, 연결된 해당 문서들의 집합을 검색 결과로 간주하는 방식은 앞선 응용 프로그램들이 해결하지 못한 다른 차원의 문제를 해결 가능하다. 예를 들어, 서울대학교 대학원 연구실 중, 데이터베이스 랭킹 관련 논문을 제출한 연구실 홈페이지를 찾는 사람이 있다고 가정하자. 검색어 “SNU, Database, Ranking” 를 모두 포함하는 단

본 연구는 서울시 산학연 협력사업(WR080951)의 연구결과로 수행되었습니다.
* 서울대학교 컴퓨터공학과 박사과정
** 서울대학교 컴퓨터공학과 석사과정
*** 서울대학교 컴퓨터공학과 교수

일 페이지는 존재하지 않으나, 서울대학교의 연구실 홈페이지 (예를 들어, IDS Lab.)를 비롯하여 논문 목록 페이지 (Publication) 까지 연결된 페이지 집합 내에서 Database Ranking 논문 관련 정보를 검색 결과로 간주할 수 있다. 그림 1은 링크기반 문서에서의 검색 결과로 문서 집합이 유용하게 활용되는 예시를 표현한다.



<그림 1> 링크 기반 문서 구조에서의 검색 결과 예시

기존의 정보 검색에서 주로 사용하는 랭킹 함수 [3]는 검색어 K가 주어졌을 때, 검색어 t가 문서 d 에 대해 가지는 가중치 w를 모두 합하여 해당 문서의 관련도 (혹은 중요도)를 계산할 수 있다. (식 1)

$$Score(K,d) = \sum_{t \in K} w(t,d) \qquad (1)$$

이와 동일한 방식으로, 다양한 크기와 형태를 가진 문서 집합이 검색의 결과로 생성되었을 때, 검색어에 대한 해당 문서 집합의 관련도는 검색어가 문서집합에 대해 가지는 가중치에 의해 결정된다.

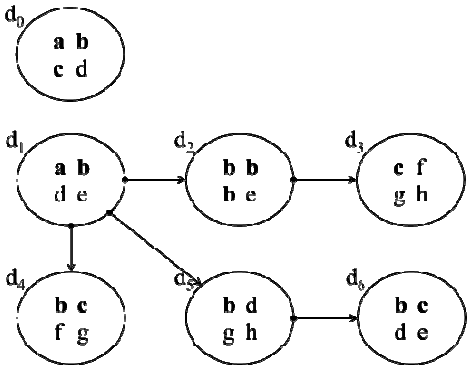
연결된 문서 집합을 고려한 최근 연구들 [4, 5]은 그래프 기반의 문서 연결 폐쇄 (closure)를 구하는 최적화 문제에 주로 연구의 초점이 맞추어져 있어서, 랭킹 알고리즘을 위한 검색어 가중치 계산 기법은 기존의 단일 문서 기반 가중치 함수 (예를 들어, Term Frequency – Inverse Document Frequency (TF-IDF)) 를 그대로 적용하고 있는 수준이다. 간단히 말해, 문서 집합 내에서 검색어를 포함한 문서에 대해 TF-IDF 함수값을 구하여 더하는 방식으로 문서 집합에 대한 가중치를 결정한다. 본 논문에서는 이런 방식이 가져오는 문제점을 지적하고 이를 문제를 해결하기 위한 새로운 검색어 가중치 함수를 제안하고자 한다.

2. 링크기반 문서 구조를 고려한 새로운 가중치 함수

기존의 검색어 가중치 함수는 단어 집합 (즉, bag-of-words) 기반의 문서를 검색의 단위로 고려하였기 때문에 검색어가 속한 문서간의 관계성 (inter-relationship)을 고려한 가중치를 효과적으로 계산할 수 없다. TF-IDF 함수를 예로 들면, 아래의 식 (2)으로부터 검색어 t가 문서 d에 대해서 가지는 가중치를 계산할 수 있다 (tf(t,d): 단어 t가 문서 d에서 등장하는 횟수, df(t): 단어 t가 등장하는 문서 개수, N: 전체 문서 개수).

$$w(t,d) = (1 + log(tf(t,d)) \, log \frac{N}{df(t)} \qquad (2)$$

기존의 단일 문서에서 사용하던 방식을 문서 집합에 적용하기 위해 TF-IDF 값을 계산하여 단일 문서에 각각 적용하고 단순 합을 구하는 방식은 여러 가지 문제를 야기한다. 예를 들어, 검색어 $a, b, c \in K$ 에 대한 검색 결과 (그림 1)을 생각해 보자.



<그림 2> 검색된 문서 집합

불필요하게 연결된 문서들은 제거하고 주어진 검색어만을 모두 포함하는 연결된 문서 집합을 구하면, $\{d_1,d_2,d_3\}$, $\{d_1,d_5,d_6\}$, $\{d_1,d_4\}$을 검색 결과로 생각할 수 있다. $\{d_1,d_2\}$와 같은 문서집합은 검색어 c 를 포함하지 못하기 때문에 검색결과로 적합하지 않다. 문서 d_0는 단일문서로써 원소가 하나인 집합 $\{d_0\}$으로 간주하여 이 논문에서 제안하는 문서 집합에 기반한 방식의 특별한 예로 간주할 수 있다. 이 방식의 문제점은 중첩되는 집합의 원소로 인한 문서의 중복이 발생하여 가중치 값 계산에 적용된다는 것이다. 그리고 조밀하게 연결된 전체 문서에 대해서 N 값이 2^N 에 가깝게 증가한다는 점이다.

기존 TF-IDF 방식을 그대로 적용하여 문서 집합에 대한 검색어 가중치를 구하면 표 1 과 같다 (N=7).

<표 1> 문서집합에 대한 가중치 값

문서집합	TF-IDF
$D_0 = \{d_0\}$	0.7291
$D_1 = \{d_1,d_2,d_3\}$	0.8754
$D_2 = \{d_1,d_5,d_6\}$	0.845
$D_3 = \{d_1,d_4\}$	0.8022

$$\mathrm{tf}(\mathrm{a}, D_0) = 1, \mathrm{tf}(\mathrm{b}, D_0) = 1, \mathrm{tf}(\mathrm{c}, D_0) = 1$$

$$\mathrm{tf}(\mathrm{a}, D_1) = 1, \mathrm{tf}(\mathrm{b}, D_1) = 4, \mathrm{tf}(\mathrm{c}, D_1) = 1$$

$$\mathrm{tf}(\mathrm{a}, D_2) = 1, \mathrm{tf}(\mathrm{b}, D_2) = 3, \mathrm{tf}(\mathrm{c}, D_2) = 1$$

$$\mathrm{tf}(\mathrm{a}, D_3) = 1, \mathrm{tf}(\mathrm{b}, D_3) = 2, \mathrm{tf}(\mathrm{c}, D_3) = 1$$

$$\mathrm{df}(\mathrm{a}) = 4, \mathrm{df}(\mathrm{b}) = 4, \mathrm{df}(\mathrm{c}) = 4$$

검색어 a는 실제로 문서 d_1에 한번 등장하기 때문에 기존 방식으로는 Document Frequency (DF) 값이 1이지만, 문서 집합이 하나의 문서로 간주되는 상황에서는 4로 계산된다. 문서 d_1이 3개의 문서집합에서 중복되어 나타나기 때문이다. 또한, 문서 d_1에 연결된 문서가 많아질수록 d_1에 속하는 단어들의DF값이 상승하게 된다. 검색어 b의 DF 값도 마찬가지로, 실제 b가 등장하는 문서는 6개임에도 불구하고, df(b)는 4로 계산된다. B가 등장하는 여러 문서가 하나의 집합의 원소에 해당하기 때문이다. 단순히 각 문서에 해당하는 TF-IDF 값을 구해 합하는 기존 방식은 실제 존재하는 문서와 문서간 연결 구조에 따라 값이 왜곡되어 나타나게 되는 문제를 발생시킨다. 이와 같은 문제를 해결하기 위해 문서간 연결의 개수 (e(t,D): 임의의 문서 집합 D 내에서 검색어 t를 포함하는 문서로부터 연결 가능한 (reachable) 문서들의 개수)를 고려한 정규화를 수행하여 새로운 TF-IDF 값을 다음과 같이 정의하였다.

$$w(t,D) = \frac{1+\log\left(1+\log\left(tf^s(t,D)\right)\right)}{n(D)\times lD} \cdot \log\left(idf^s(t)\right) \quad (3)$$

$$tf^s(t,D) = \sum_{d\in D} tf(t,d) \quad (4)$$

$$idf^s(t) = \frac{N^s}{\sum_{\forall D, t\in D} e(t,D)} \quad (5)$$

N^s: 문서집합의 총 개수 (본 예제에서는 N^s을 전부 구할 수 없으므로 매우 큰 값으로 가정)

lD: 문서집합 D에 속하는 문서들의 평균 길이

n(D): 문서집합 D에 속하는 문서 개수

idf^s(t)값이 정의되는 문서는 예제에 등장한 문서가 전부라고 가정한다.

식 3 에서 제시된 새로운 TF-IDF 값은 문서 집합을 구성하는 문서들간의 연결 상태, 중복성을 고려하여 기존 방식의 문제점을 해결하였다. 표 2 는 새로운 정규화를 수행한 문서 가중치 값을 다시 계산한 값으로, 기존 방식이 가진 a값의 중복 계산과 링크 구조로 인한 값의 왜곡이 제거되었음을 확인 할 수 있다. 예를 들어, 문서 집합에 대한 검색어 관련도를 $D_0 > D_3 > D_1 > D_2$ 의 순서로 확인 가능하다. 이는 검색어를 전부 포함한 단일 문서가 가장 관련도가 높고, 다수의 링크를 가진 문서는 관련도가 낮아지는 현상을 정량화된 값으로 표현하는 직관적인 예시라고 볼 수 있다. 본 연구는 여전히 진행 상태에 있기 때문에 해당 값에 대한 검증은 추후 연구에서 수행할 예정에 있다.

<표 2> 문서집합에 대한 새로운 가중치 값

문서집합	TF-IDF
$D_0 = \{d_0\}$	1.1505
$D_1 = \{d_1,d_2,d_3\}$	0.4176
$D_2 = \{d_1,d_5,d_6\}$	0.4117
$D_3 = \{d_1,d_4\}$	0.6038

$$tf^s(\mathrm{a}, D_0) = 1, tf^s(\mathrm{b}, D_0) = 1, tf^s(\mathrm{c}, D_0) = 1$$
$$tf^s(\mathrm{a}, D_1) = 1, tf^s(\mathrm{b}, D_1) = 4, tf^s(\mathrm{c}, D_1) = 1$$
$$tf^s(\mathrm{a}, D_2) = 1, tf^s(\mathrm{b}, D_2) = 3, tf^s(\mathrm{c}, D_2) = 1$$
$$tf^s(\mathrm{a}, D_3) = 1, tf^s(\mathrm{b}, D_3) = 2, tf^s(\mathrm{c}, D_3) = 1$$
$$idf^s(\mathrm{a}) = 20, idf^s(\mathrm{b}) = 100, idf^s(\mathrm{c}) = 20$$

3. 결론

본 논문에서는 TF-IDF 함수를 문서 집합에 적용한 기존의 연구들에 대한 문제점을 지적하고 검색어가 포함된 문서의 링크의 개수와 문서의 중복을 고려한 새로운 정규화 지수를 적용한 새로운 TF-IDF 함수를 제안하였다. 본 연구는 추후에 문서 연결 구조상의 방향성과 연결선에 대한 가중치 값을 계산하여 보다 포괄적인 TF-IDF 함수를 문서 집합에 대하여 정의하고자 한다. 현재, TREC 문서 집합을 생성하여 제안된 가중치 함수를 검증하는 실험을 진행 중이다. 추후 연구에서는 다양한 실험 결과를 통해 제시한 가중치 함수가 실제 랭킹 알고리즘에 활용되어 좋은 품질의 검색을 수행하는지를 증명해보고자 한다.

참고문헌

[1] Lawrence Page and Sergey Brim. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World-Wide Web Conference*, pages 107-117, Brisbane, Queensland, Australia, April 1998

[2] J. Glen and J. Widom. SimRank: a measure of structural-context similarity, SIGKDD, pp:538-543, 2002

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge Univ Press, 2008

[4] W.S. Li, K. S. Candan, Q. Vu and D. Agrawal. Retrieving and Organizing Web Pages by "Information Unit". WWW, 230-244, 2001

[5] R. Varadarajan, V. Hristidis, T. Li. Searching the Web Using Composed

Pages. SIGIR, August 6-11, 2006