C-rank: 웹 페이지 랭킹을 위한 기여도 기반 접근법

(C-rank: A Contribution-Based Approach for Web Page Ranking)

이 상 철 † 김 동 진 ††
(Sang-Chul Lee) (Dong-Jin Kim)

손 호 용 ††† 김 상 욱 ††††
(Ho-Yong Son) (Sang-Wook Kim)

이 재 범 †††††
(Jae Bum Lee)

요 약 수많은 웹 문서로부터 웹 서퍼가 원하는 정보를 찾기 위해 다양한 검색 엔진들이 개발되어왔다. 검색 엔진에서 가장 중요한 기능 중 하나는 사용자 질의에 대해서 웹 문서를 평가하고 랭킹을 부여하는 것이다. PageRank 등의 기존 하이퍼링크 정보를 이용한 웹 랭킹 알고리즘은 토픽 드리프트 현상을 발생시킨다. 이러한 문제를 해결하기 위하여 연관성 파급 모델이 제안되었지만, 기존의 연관성 파급 모델을 기반으로 하는 랭킹 알고리즘은 성능상의 이유로 실제 웹 검색 엔진에서 사용하기 어렵다. 본 논문에서는 이러한 토픽 드리프트 현상을 완화하면서 좋은 성능을 제공하는 새로운 랭킹 알고리즘을 제안한다. 다양한 실험을 통하여 기존 알고리즘들과 비교한 제안하는 알고리즘의 우수성을 검증한다.

키워드 : 웹 페이지 랭킹, 연관성 파급 모델

Abstract In the past decade, various search engines have been developed to retrieve web pages that web surfers want to find from world wide web. In search engines, one of the most important functions is to evaluate and rank web pages for a given web surfer query. The prior algorithms using hyperlink information like PageRank incur the problem of 'topic drift'. To solve the problem, relevance propagation models have been proposed. However, these models suffer from serious performance degradation, and thus cannot be employed in real search engines. In this paper, we propose a new ranking algorithm that alleviates the topic drift problem and also provides efficient performance. Through a variety of experiments, we verify the superiority of the proposed algorithm over prior ones.

Key words : Web Page Ranking, Relevance Propagation

1. 서 론

초기에는 각 웹 문서의 내용 분석을 수행함으로써 질의와 웹 문서와의 연관성 점수를 기준으로 랭크를 결정하는 방법이 사용되었다. 대표적인 초기 웹 문서 랭킹 기법으로는 TF-IDF[1]와 BM2500[2] 등이 존재한다. 그러나 이 기법은 상위에 랭크된 웹 문서가 다른 웹 문서 작성자들이 인정할 만큼 좋은 내용의 웹 문서인지 알 수 없다는 문제를 가지고 있다.

링크 기반의 웹 문서 랭킹 기법은 이러한 초기의 웹 문서 랭킹 기법의 문제를 해결하기 위하여 제안된 것이다[3]. 대표적인 기법으로는 PageRank[3]와 HITS[4] 등이 있다. 그러나 이 기법은 각 웹 문서에 대한 권위 점수와 연관성 점수를 독립적으로 계산하기 때문에 어떤 웹 문서가 사용자 질의와의 연관성이 낮음에도 불구하고 권위 점수가 높은 경우 검색 결과의 상위에 랭크되는 현상이 발생하게 된다. 이러한 현상을 토픽 드리프트(topic drift)라 부른다[5].

토픽 드리프트 문제를 해결하기 위하여 질의와 연관된 웹 문서만을 대상으로 권위 점수를 계산하는 연관성 파급 기법이 제안되었다. 대표적인 연관성 파급 기법에는 QD-PageRank[5-7] 등이 있다. 연관성 파급 기법은 사용자가 질의를 주는 시점에 (1)질의어와 연관된 웹 문서들을 찾고, (2)이 문서들 간의 하이퍼링크를 이용한 네트워크를 구성한 후, (3)이 네트워크를 대상으로 연관성 파급을 수행한다. 기존 연관성 파급 모델은 세 단계의 과정으로 인하여, 상당한 계산 오버헤드가 있다. 본

· 본 연구는 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원 회사의 입장을 대변하는 것은 아닙니다.
· 이 논문은 2009 한국컴퓨터종합학술대회에서 'C-rank: 기여도 기반의 웹문서 랭킹'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한양대학교 전자컴퓨터통신공학과
korly@hanyang.ac.kr
†† 비 회 원 : (주) NHN
dongjin.kim@nhncorp.com
††† 비 회 원 : 한양대학교 전자컴퓨터통신공학과
iso434@agape.hanyang.ac.kr
†††† 종신회원 : 한양대학교 정보통신학부 교수
wook@hanyang.ac.kr
††††† 종신회원 : (주) NHN
jblee@nhncorp.com

논문접수 : 2009년 8월 14일
심사완료 : 2009년 11월 3일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제1호(2010.1)

논문은 기존 연관성 파급 기법의 한계를 해결하기 위하여 제안된 C-rank[8]를 설명하고, 성능을 평가한다.

본 논문에서는 웹 문서의 내용을 대표하는 단어들을 핵심단어라 정의한다. 또한, 참조된 웹 문서는 참조한 웹 문서에 내용상으로 기여한다고 정의한다. C-rank는 각 웹 문서가 자신을 하이퍼링크로 직접 혹은 간접적으로 가리키는 다른 웹 문서들 중 자신과 핵심 단어를 공유하는 웹 문서들에게 내용적으로 얼마나 기여했는가를 계량화함으로써 기여 정도를 기반으로 해당 웹 문서의 랭크를 계산한다. C-rank는 핵심 단어에 대해서만 다른 웹 문서들에 대한 해당 웹 문서의 기여 정도를 계산하므로 질의어로 사용 가능한 전체 단어들에 대한 각 웹 문서의 랭크를 매우 효율적으로 구할 수 있다.

2. 제안하는 기법

2.1 기본 랭크 부여 방안

대부분의 웹 문서는 자신의 내용을 보안하기 위한 목적으로 그 내부에 하이퍼링크를 포함시킨다. 하이퍼링크를 받는 웹 문서가 질의어와 관련성이 클 때, 이 웹 문서는 해당 질의어에 대하여 자신을 하이퍼링크 하는 웹 문서에게 기여한다고 말할 수 있다. 따라서 주어진 검색 질의어와 관련된 좋은 내용을 가진 웹 문서(이후, 질의어 관련 좋은 웹 문서라 부름)는 다음과 같은 특징을 가지고 있다.

(1) 질의어와 연관성 점수가 높다.

(2) 질의어 관련 좋은 웹 문서들로부터 많은 하이퍼링크를 받는다.

이와 같은 좋은 문서의 특징을 기반으로 C-rank는 각 질의어에 대한 웹 문서의 랭크를 자신의 질의어에 대한 연관성 점수와 자신을 직접(혹은 간접)으로 하이퍼링크 하는 다른 웹 문서들에게 질의어에 대하여 기여한 점수들의 합으로 계산한다. 즉, 자신을 직접 하이퍼링크 하는 웹 문서 외에도 하이퍼링크로 d단계까지 떨어진 웹 문서들에 기여한 점수들도 모두 누적하여 반영한다.

그림 1은 C-rank를 이용하여 한 웹 문서가 자신에게 직접 하이퍼링크를 건 웹 문서들에 대하여 기여한 점수를 계산하는 방법을 나타낸다. 사각형은 웹 문서를 나타내고, 사각형 내의 숫자는 질의어에 대한 해당 웹 문서의 연관성 점수를 나타낸다. 웹 문서 q는 자신의 내용을 보완하기 위하여 웹 문서 p를 하이퍼링크로 연결하며, 이때 웹 문서 p는 웹 문서 q에게 내용 측면에서 기여한다. 웹 문서 q는 자신의 내용뿐만 아니라 자신이 하이퍼링크로 연결한 p의 내용까지 포함하는 셈이다. 따라서 웹 문서 q가 표현하고자 하는 전체 내용의 연관성 점수는 q 자신의 연관성 점수와 q가 하이퍼링크로 연결한 웹 문서들의 연관성 점수의 합이며, 이 중 웹 문서 p의

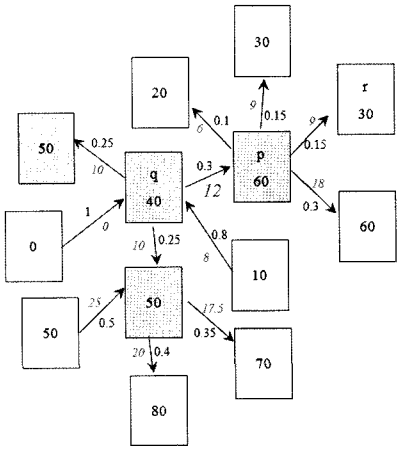


그림 1 C-rank 1단계 파급의 예

연관성 점수가 차지하는 비중을 p의 q에 대한 기여율이라 부른다. 식 (1)은 질의어 t에 대한 웹 문서 p의 웹 문서 q에 대한 기여율의 계산식을 나타낸 것이다. C-rank에서는 웹 문서 p가 웹문서 q에 기여하는 정도는 웹 문서 q의 연관성 점수와 p의 q에 대한 기여율의 곱으로 계산된다.

$$\alpha_t^1(q,p)\begin{cases}=\dfrac{R_t(p)}{R_t(q)+\sum_{r\in outlink(q)}R_t(r)} & ,\text{if } there\ is\ an\ edge\\ =0 & ,otherwise\end{cases} \quad (1)$$

여기서 $\alpha_t^1(q,p)$는 질의 단어 t에 대해 1단계의 기여율을 나타낸다. $R_t(p)$는 질의어 t에 대한 웹 문서 p의 연관성 점수를 나타낸다. $outlink(q)$는 웹 문서 q가 하이퍼링크로 연결한 웹 문서들의 집합을 나타낸다.

웹 문서 p는 웹 문서 q의 내용 보완에 있어서 자신이 기여한 정도를 자신에 랭크에 반영하게 된다. 이 기여한 정도는 내용을 보완한 웹 문서 q 자신의 내용의 가치가 높을수록 높은 점수를 부여받게 되는데, 이를 기여 가치라 한다. 따라서 기여 가치는 식 (2)에 나타난 바와 같이 기여율과 내용이 보완된 웹 문서의 연관성 점수의 곱으로 표현된다. 링크 위의 이텔릭으로 표시한 수는 각 웹 문서가 자신을 하이퍼링크로 연결한 다른 웹 문서에 대한 기여 가치를 표시한 것이다.

$$\text{기여가치} = \alpha_t^1(q,p)*R_t(q) \quad (2)$$

전술한 바와 같이, C-rank는 웹 문서의 전체 기여 가치를 계산할 때, 자신을 직접적인 하이퍼링크로 연결한 웹 문서 외에도 d단계까지 간접적인 하이퍼링크로 연결한 웹 문서들에 대한 기여 가치들을 모두 누적한다. 그림 1에서 웹 문서 p가 웹 문서 q에게 직접 기여한 것과 마찬가지로 웹 문서 r은 웹 문서 p에게 직접 기여한다. 이때, 웹 문서 r은 웹 문서 p에게 기여하기 때문에 웹 문서 r에 간접적으로 기여하는 셈이 된다. 웹 문서 r이

웹 문서 q에게 기여하는 정도는 웹 문서 r이 웹 문서 p에게 기여하는 정도 중 웹 문서 p가 웹 문서 r에게 기여하는 정도이므로 식 (3)과 같이 기여율 $\alpha_t^1(q,p)$과 기여율 $\alpha_t^1(p,r)$의 곱으로 표현한다. 여기서 $\alpha_t^2(q,r)$는 질의어 t에 대해서 웹 문서 r이 하이퍼링크로 2단계 떨어진 웹 문서 q에 대한 기여율을 의미한다. 이와 같은 방식으로 하이퍼링크로 d단계 떨어진 웹 문서에 대한 기여율을 계산하는 식을 식 (3)으로 일반화 할 수 있다.

$$\alpha_t^d(q,p) = \alpha_t^1(q,r_1) \times \prod_{i=1}^{d-2} \alpha_t^1(r_i, r_{i+1}) \times \alpha_t^1(r_{d-1}, p) \quad (3)$$

제안하는 C-rank 기법은 위의 과정을 d단계까지 수행하여 각 웹 문서가 자신을 하이퍼링크를 통하여 직간접적으로 연결한 웹 문서들의 내용을 보완해 준 정도를 반영하는 기여 가치를 저장한다. 해당 웹 문서의 최종 랭크는 이렇게 인정받은 전체 기여 가치와 더불어 이 웹 문서 자신의 연관성 점수를 가중합하여 결정된다. 기여 가치를 d단계 전달했을 때의 C-rank 모델을 수식으로 표현하면 식 (4)와 같다. 여기서 $D(p)$는 웹 문서 p로부터 d단계 떨어진 웹 문서 집합을 의미한다.

$$C_t(p) = \alpha R_t(p) + (1-\alpha) \sum_d \sum_{q \in D(p)} \alpha_t^d(q,p) R_t(q) \quad (4)$$

2.2 효율적인 랭크 계산 전략

C-rank도 질의어별로 웹 문서들의 연관성 점수들을 기반으로 기여 가치를 계산한다. 따라서 웹 문서들의 기여 가치는 질의어에 따라 달라진다. 그러나 검색이 요청된 시점에 온라인으로 기여 가치를 계산하게 되는 경우, 기존의 연관성 파급 기법과 마찬가지로 검색 사용자를 지나치게 기다리게 하므로 실제 검색 엔진에 적용이 어렵다. 또한, 오프라인으로 가능한 모든 질의어에 대하여 미리 모든 웹 문서들의 기여 가치와 연관성 점수를 계산하는 방법은 질의어를 구성하는 단어 조합 수의 무한함으로 인하여 적용이 불가능하다.

본 논문에서는 다음과 같은 관찰을 통하여 이러한 문제점을 해결할 수 있는 매우 효과적인 오프라인 계산 방법을 제안한다. 하나의 웹 문서는 주로 한 두 개의 주제들을 다루는 내용을 포함하며, 이 주제들을 표현하는 소수의 단어들로 대표될 수 있다. 이와 같이, 하나의 웹 문서가 표현하고자 하는 주제를 대표하는 소수의 단어들을 그 웹 문서의 핵심 단어라 정의한다. 핵심 단어는 그 웹 문서 내에 존재하는 단어들 중 해당 웹 문서와의 연관성 점수가 가장 높은 n개의 단어들을 의미한다. 이러한 핵심 단어의 개념과 관련하여 다음의 두 가지를 전제할 수 있다.

검색 사용자는 자신의 질의어를 핵심 단어로 가지고 있지 않는 웹 문서는 최종 검색 결과에 나타나는 것을 원하지 않는다.

웹 문서가 가지는 하이퍼링크는 그 웹 문서의 핵심 단어와 연관된 주제를 보완하기 위해서 그 주제를 다루는 다른 웹 문서를 가리키기 위하여 포함한 것이다. 두 문서는 동일한 핵심 단어를 공유하고 있을 가능성이 매우 높다.

이 두 가지 전제하에 본 연구에서는 C-rank를 오프라인 시간에 효과적으로 계산하기 위한 다음과 같은 두 가지 전략을 수립한다.

웹 문서에 대한 기여 가치는 그 웹 문서의 핵심 단어에 대해서만 계산한다. 이 웹 문서는 다른 단어들에 대해서는 검색 사용자를 만족시킬 수 있는 가능성이 매우 낮기 때문이다.

웹 문서의 자신을 하이퍼링크로 가리키는 다른 웹 문서에 대한 기여 가치는 두 웹 문서들에 대하여 공통된 핵심 단어가 존재할 때 그 존재하는 공통의 핵심 단어에 대해서만 인정된다. 이것은 공통의 핵심 단어를 가진 웹 문서가 자신을 하이퍼링크한 경우에 대해서만 기여 가치를 인정하고, 그렇지 않은 경우에는 기여 가치를 인정하지 않겠다는 것을 의미한다.

그림 2는 핵심 단어 집합을 이용한 C-rank 계산의 예를 보이고 있다. 이 예에서 핵심 단어의 수는 5이며 기여 가치 전달의 단계 수는 2단계이다. C-rank 계산을 위해 우선 하이퍼링크로 연결된 웹 문서 그래프를 구성한다. 그 다음, 핵심 단어를 결정한다. 그림 2에서는 핵심 단어의 개수를 5로 설정하였으며, 이때 웹 문서 q는 단어 A, B, C, D, E를 핵심 단어 집합으로 갖게 되며 각 단어에 대한 연관성 점수는 괄호 안에 표시되어있다. 그 다음, 모든 하이퍼링크에 각 핵심 단어에 대한 기여율을 부여한다. 하이퍼링크로 연결된 두 웹 문서가 공유하는 핵심 단어가 있다면 해당 단어에 대해 기여율을 계산한다. 예를 들어 하이퍼링크로 연결된 웹 문서 q와 p는 핵심 단어 B, C, D, E를 공유한다. 따라서 해당 핵심 단어들에 대한 기여율들을 하이퍼링크에 부여한다. 그 다음, 각 웹 문서로부터 하이퍼링크로 연결된 모든 웹 문서를 방문하며 기여 가치를 전달하는 것을 통하여 C-rank를 계산한다. 모든 문서에 대한 C-rank계산이 완료되면, 최종 랭킹을 위하여 각 웹 문서는 자신의 핵심 단어 별로 전달 받은 기여 가치를 합하여 유지한다. 표 1은 C-rank의 랭크 계산 알고리즘을 Pseudo 코드

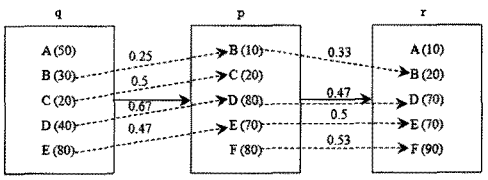


그림 2 핵심 공통 단어 집합을 이용한 예

표 1 C-rank의 랭크 계산 알고리즘

Algorithm C-rank
(1) 각 웹 문서에 대한 핵심 단어 추출.
(2) 링크 정보를 이용하여 웹 문서 네트워크 구축.
(3) 각 링크에 대해 공통 핵심 단어의 기여율 계산.
(4) 네트워크에서 웹 문서(노드) 초기화
(4-1) 각 노드는 자신의 핵심 단어 수만큼의 기여 가치 저장 공간 할당.
(4-2) 기여가치의 초기 값은 노드의 각 핵심 단어의 연관성 점수.
(5) 전 단계의 기여가치를 기여율의 비율로 아웃링크 노드로 전달.
(6) d 단계가 될 때 까지 단계 (5) 반복.

로 나타낸 것이다. 자세한 설명은 지면관계상 생략한다.

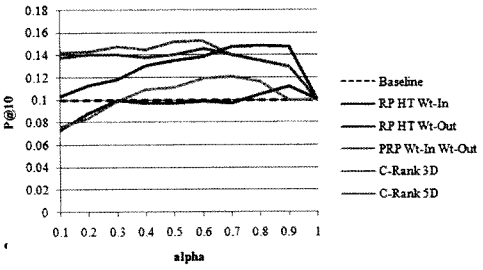
사용자 검색 요청 시, 미리 C-rank로 계산된 기여 가치를 사용하기 때문에 모든 사용자 질의에 대해 효과적인 랭킹 계산이 가능하다. 사용자의 검색이 요청되면, 우선 검색 엔진에서 이미 사용하고 있는 역 인덱스를 통하여 질의어와 연관된 웹 문서를 추출한다. 그 다음, 추출된 웹 문서들을 대상으로 사용자 질의어에 포함된 각 단어의 C-rank로 계산된 기여 가치 값과 연관성 점수를 로드(load)한다. 만약 해당 단어가 핵심 단어가 아닌 경우 기여 가치 값은 0이 되며, 그렇지 않은 경우 기존 저장된 값을 로드하면 된다. 그 다음, 로드한 두 값을 가중 합하는 방법으로 조합하여 각 단어에 대해 하나의 점수를 갖게 한다. 마지막으로 각 웹 문서에 있는 단어별 점수를 이용하여 각 웹 문서에 대해 하나의 점수를 계산한다. 이때, 여러 단어에 대한 점수를 하나의 점수로 계산하기 위하여 기존에 제안된 다중 단어 조합 방법을 사용한다[9].

3. 성능 평가

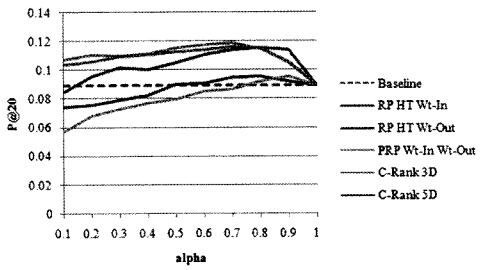
데이터 집합으로는 약 120만 개의 웹 문서와 1,120만 개의 하이퍼링크로 이루어진 .GOV 데이터 집합을 사용하여 실험을 진행하였다. 품질 측정의 기준이 될 질의 및 판정 기준으로는 2004년도 TREC의 Web Track[10] 중 TD(topic distillation) 질의 집합를 사용하였으며, 질의 수는 75개이다. 또한 웹 문서를 인덱싱 및 연관성 점수를 계산하는 방법으로는 Lucene에서 제공하는 역 인덱스와 TF-IDF기법을 사용하였다[9]. 정확도 평가를 위하여 기존에 널리 사용되는 P@10, P@20, 및 MAP[11]을 이용하여 측정하였다. 평가 대상이 되는 기존 기법으로는 연관성 파급 모델로 [6]과 [7]에서 제안한 방법을 사용하며, 본 논문에서는 각각 RP와 PRP로 부른다. 각 논문에서 추천하는 파라미터 설정을 사용하였다. 마지막으로 실험을 위한 환경으로는 1.6GHz의 Xeon 프로세스와 8GB의 메모리를 장착한 서버에서 OS로는 MS Windows 2003 R2 x64를 사용하였다.

3.1 정확도 평가

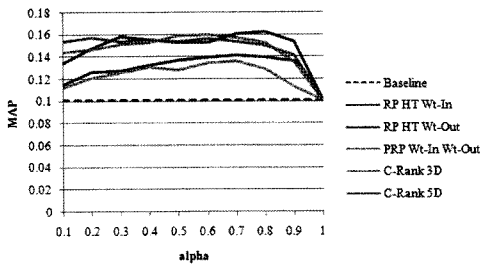
그림 3은 C-rank 및 기존 연관성 파급 기법의 정확도를 측정한 결과를 나타낸다. RP HT Wt-In와 RP HT Wt-Out은 RP모델에서 추천하는 파라미터 설정중 하나로 단어 빈발 레벨(term frequency level)에서 하이퍼링크로 가중치를 달리하여 파급을 시키는 두 가지 방법이다. 파급이 대상이 되는 집합은 전자는 인링크(inlink)이며, 후자는 아웃링크(outlink)이다. PRP Wt-In, Wt-out은 PRP 모델에서 추천하는 파라미터 설정중 하나이며, 연관성 점수를 가중치를 달리하여 인링크와 아웃링크 모두에게 파급하는 설정이다. 마지막으로 C-rank에 대해서는 5단계와 7단계로 파급을 설정하여 비교하였으며, 각각 C-rank 5D와 C-rank 7D로 명시하였다. 그림 3(a)은 각 기법의 P@10 결과를 나타낸 그래프이며, C-rank 3D기법이 가장 우수한 것으로 나타났다. 그림 3(b)는 각 기법의 P@20의 결과를 나타낸 그래프



(a) TD 2004년도 P@10 결과



(b) TD 2004년도 P@20 결과



(c) TD 2004년도 MAP 결과

그림 3 TD 2004년도에 대한 정확도 평가

이며, C-rank family가 가장 우수함을 볼 수 있다. 마지막으로 그림 3(c)는 각 기법의 MAP의 결과를 나타낸 그래프이며, 일정 구간에서 RP HT Wt-In이 우수하지만 전체적으로 C-rank family가 상위에 있으며 정확도 차이가 크게 나지 않음을 볼 수 있다.

3.2 질의 처리 성능 평가

P@10을 기준으로 가장 좋은 정확도를 보인 C-rank, RP, 그리고 PRP의 파라미터 설정을 이용하여 처리 성능을 비교하였다. 표 2는 하나의 단어에 대한 평균 질의 처리시간을 나타낸다. C-rank의 경우 질의 가능한 모든 단어가 인덱싱 되어 있기 때문에 실제 질의에 필요한 시간이 거의 없음을 알 수 있다. 그에 비해 RP와 PRP의 방법은 질의 처리 순간 질의어와 연관된 웹 문서들을 찾고, 이 문서들 간의 하이퍼링크를 이용한 네트워크를 구성한 후, 이 네트워크를 대상으로 연관성 파급을 수행하기 때문에 질의 처리 성능이 떨어진다. PRP의 경우 연관성 파급을 반복하면서 각 점수를 정규화 하는 등의 추가적인 작업이 요구되기 때문에 가장 낮은 성능을 보였다.

현실적으로 웹 서퍼는 하나의 질의에 대해 5~18초가량 기다리지 않는다. 따라서 C-rank와 같이 오프라인에서 단어별로 랭크를 구한 후, 사용자 검색이 요청되면 질의어에 포함된 각 단어에 대한 랭크 결과를 조합하는 방안을 고려해 볼 수 있다. 그림 4는 질의 가능한 단어 수가 증가함에 따라 C-rank와 기존 기법의 랭킹 계산 성능의 차이를 보이고 있다. 그림 4에서 보이는 바와 같이 질의 가능한 단어의 수가 증가함에 따라 C-rank는 영향을 받지 않는다. 왜냐하면 C-rank는 각 문서별로 n개의 핵심 단어에 대해서만 계산하기 때문에 C-rank를 계산하는데 질의 가능한 단어의 수는 무의미하다. 그에 비해, 기존의 연관성 파급 기법은 각 단어별로 랭킹을 계산해야 하기 때문에 질의 가능한 단어 수에 비례하여 랭킹 계산에 소요되는 시간은 선형적으로 증가한다. 따라서 질의 가능한 단어의 수가 매우 클 때, 기존의 연관성 파급 기법으로는 랭킹 계산이 어렵지만, C-rank를 통하여 랭킹 계산은 질의 가능한 수와 무관하기 때문에 가능하다.

표 2 단어 당 평균 질의 처리 시간

Model	C-rank	RP	PRP
시간(sec)	0.0	5.92	18.16

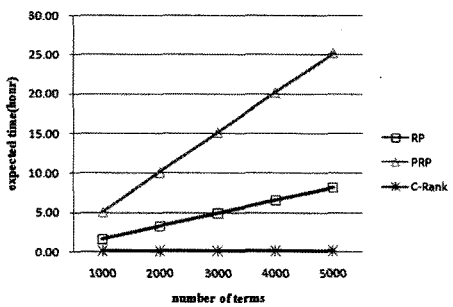


그림 4 질의 가능한 단어 수에 따른 랭킹 계산을 위한 예상 시간 비교

4. 결 론

본 논문에서는 기존의 연관성 파급 기법의 랭킹 계산 성능을 개선하는 새로운 기법인 C-rank 기법을 제안하였다. 제안하는 C-rank 기법은 각 웹 문서는 자신의 내용을 보완하기 위하여 하이퍼링크를 포함시킨다는 특징을 이용하여 각 웹 문서의 기여하는 정도를 계량화 하였다. 또한 각 웹 문서는 소수의 핵심 단어로 1~2가지의 주제를 다루고 있다는 특징을 이용하여 웹 문서간 공유하는 핵심 단어만을 계산함으로써 정확도와 랭킹 계산의 성능을 개선하였다. 제안하는 C-rank기법은 토픽 드리프트 문제를 완화할 뿐만 아니라 효과적인 웹 문서의 랭킹이 가능하기 때문에 실제 검색 엔진에 적용 가능할 것으로 기대한다.

참 고 문 헌

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addision-Wesley, 1999.
[2] S. E. Robertson, "Overview of the Okapi projects," *Journal of Documentation*, vol.53, no.1, pp.3-7, 1997.
[3] P. Lawrence et al., The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford University, 1998.
[4] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol.46, no.5, pp.604-632, 1999.
[5] M. Richardson and P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information In PageRank," In *Advances in Neural Information Processing Systems 14*, pp.1141-1448, 2002.
[6] T. Qin et al., "A Study of Relevance Propagation of Web Search," In *Proc. ACM Int'l. Conf. on Information Retrieval*, pp.408-415, 2005.
[7] A. Shakery and C. Zhai, "A Probabilistic Relevance Propagation Model for Hypertext Retrieval," In *Proc. ACM Int'l. Conf. on Information and Knowledge Management*, pp.550-558, 2006.
[8] Dong-Jin Kim, C-rank: A Contribution-Based Web Page Ranking Algorithm, NHN Internal Technical Report, TR-NHN-2007-158, 2007. (In Korean)
[9] Lucene, http://lucene.apache.org.
[10] TREC Web Track, http://es.cmis.csiro.au/TRECWeb.
[11] S. Chakrabarti, Mining The Web, Morgan Kaufmann, 2002.