웹 사이트 구조를 이용한 토픽 검색 연구

An Experimental Study on Topic Distillation Using Web Site Structure

이지숙(Jee-Suk Lee)*, 정영미(Young-Mee Chung)**

초 록

이 연구에서는 TREC이 제시한 토픽 검색의 정의에 따라 질의에 적합한 웹 사이트를 검색하는 효과적인 토픽 검색 알고리즘을 제안하고 실험을 통해 그 성능을 평가하였다. 이 연구의 토픽 검색 알고리즘은 먼저 질의에 대한 웹 페이지 검색 결과로부터 적합한 웹 사이트를 선정한 다음, 선정된 사이트의 구조를 이용하여 질의에 대한 적합성 점수를 산출한다. TREC의 .GOV 실험 문헌 집단과 TREC-2004 실험의 질의 및 적합문헌 리스트를 이용한 검색 실험 결과 이 토픽 검색 알고리즘은 상위 10위 안에 최소 2개 이상의 적합 사이트를 검색하여 비교적 높은 수준의 성능을 보였다. 또한 TREC-2004의 적합문헌 리스트 분석을 통해 적합문헌 선정에 토픽 검색의 정의가 엄격하게 적용되지 않은 경우가 있음을 확인하고, 수정된 적합문헌 리스트를 이용하여 토픽 검색 성능을 재평가한 결과 이 연구에서 제안한 토픽 검색 알고리즘의 성능이 월등히 향상되었다.

ABSTRACT

This study proposes a topic distillation algorithm that ranks the relevant sites selected from retrieved web pages, and evaluates the performance of the algorithm. The algorithm calculates the topic score of a site using its hierarchical structure. The TREC .GOV test collection and a set of TREC-2004 queries for topic distillation task are used for the experiment. The experimental results showed the algorithm returned at least 2 relevant sites in top ten retrieval results. We performed an in-depth analysis of the relevant sites list provided by TREC-2004 to find out that the definition of topic distillation was not strictly applied in selecting relevant sites. When we re-evaluated the retrieved sites/sub-sites using the revised list of relevant sites, the performance of the proposed algorithm was improved significantly.

키워드 : 웹 검색, 토픽 검색, 사이트 검색, 웹 사이트 구조, 하이퍼링크
web search, topic distillation, site searching, web site structure, hyperlink

* NHN(주) 기술연구센터 (jeesuklee@nhncorp.com)
** 연세대학교 문헌정보학과 교수 (ymchung@yonsei.ac.kr)
■ 논문접수일자 : 2007년 8월 15일
■ 게재확정일자 : 2007년 9월 10일

1. 서 론

웹을 통해 접근할 수 있는 정보의 양과 종류가 증가하면서 웹은 이용자들에게 중요한 정보원이 되고 있다. 방대한 양의 웹 자원 중에서 이용자가 필요로 하는 정보를 찾아서 제공하기 위한 웹 검색 엔진은 웹 자원이 증가함에 따라 이용자들에게 더욱 필수적인 도구가 되고 있다. 이에 따라 웹 검색 엔진의 성능을 향상시키고자 하는 연구가 계속되고 있으며, 웹 검색의 성능 향상을 위한 최신 연구 과제의 하나가 TREC(Text Retrieval Conference)의 토픽 검색(topic distillation)이다.

토픽 검색은 적합성만을 평가 기준으로 삼아 웹 페이지를 검색하는 일반 웹 검색(ad-hoc retrieval)과는 달리 웹 자원의 적합성과 신뢰성을 평가 기준으로 한다. 토픽 검색은 1998년 Bahrat and Henzinger(1998)가 처음 제안하였으며, 이들은 질의와 관련된 가장 권위있는 웹 페이지를 이용자에게 제시하는 것을 검색의 목적으로 보았다. 이후 토픽 검색은 2002년 TREC의 Web Track 과제의 하나로 선정되어 여러 연구자들이 관심을 갖게 되었으며, Craswell and Hawking(2003)은 토픽 검색을 "주어진 넓은 주제의 질의에 대해, 가장 적합한 사이트의 홈페이지를 검색하는 것"으로 정의한 바 있다.

토픽 검색은 넓은 주제의 질의의 경우 내용의 적합성만으로는 이용자가 만족할 만한 검색 결과를 제공할 수 없다는 가정에서 출발하였다. 넓은 주제를 표현하는 질의에서는 주제에 관련된 페이지 수가 너무 많기 때문에 이용자가 검색 결과를 소화하기 어렵고(Kleinberg 1999), 이용자는 많은 검색 결과를 모두 확인하지도 않으며(Bahrat and Mihaila 2002), 검색된 모든 웹 페이지가 믿을 만한 정보를 가지고 있다고 보장하기 어렵다. 따라서 검색 결과에서 신뢰성 있는 가장 중요한 몇 개의 페이지를 선택하여 이용자에게 제시할 수 있는 필터링 과정이 필요하다고 보는 것이다.

이 연구에서는 질의에 대해 적합한 소수의 웹 사이트를 검색하는 것을 목표로 한 효과적인 토픽 검색 알고리즘을 제안하였다. 이 알고리즘은 일반 웹 검색 기법을 이용하여 검색한 웹 페이지들로부터 적합 사이트 후보들을 선정한 후 각 사이트의 구조를 이용하여 산출한 적합성 점수에 의해 사이트를 순위화한다. 이 토픽 검색 알고리즘은 TREC의 .GOV 실험 문헌 집단과 TREC-2004 실험의 질의 집합을 사용하여 실험하였다.

2. 토픽 검색 관련 연구

초기 토픽 검색 연구자들은 신뢰성 측정에 중요한 요소는 웹 페이지가 갖는 링크 정보라고 보았기 때문에 토픽 검색은 링크 분석 알고리즘과 함께 연구되어 왔다. 저널의 인용에서처럼 링크를 많이 받는 웹 페이지는 상대적으로 중요한 정보를 가지고 있는 페이지로 간주하고 신뢰성 여부를 링크의 수를 통해 계량적으로 측정하였다. 초기 토픽 검색 연구의 중심이 되었던 Kleinberg(1999)의 HITS(Hyperlink Induced Topic Search) 알고리즘도 내용에 기반한 검색 결과 집합 안에서 링크를 통해 웹 페이지의 신뢰성을 측정한다. "topic

distillation"이라는 용어를 처음 사용한 Bahrat and Henzinger(1998)도 분석 대상 웹 페이지들과 질의와의 유사성 정도를 측정하고 이를 링크 가중치에 반영하여 HITS 알고리즘의 성능을 향상시키고자 하였다. 이후 Chakrabarti et al.(1999), 박기림 등(2003)의 연구에서도 HITS 알고리즘을 통한 링크 구조 분석에 내용 적합성을 추가하여 신뢰성 있는 웹 페이지를 찾아내고자 하였다.

초기의 토픽 검색 연구는 질의의 주제에 대해 다루고 있는 웹 페이지로 연결 가능한 페이지를 검색하는 것을 목표로 하였다고 볼 수 있다. 토픽 검색의 질의가 광범위한 주제를 나타내고 있어, 이용자의 정보 요구에 정확히 일치하는 검색 결과를 제공하기 어렵기 때문에 이용자의 질의에 관련된 내용으로 연결해 줄 수 있는 링크 페이지를 검색하여 이용자가 자신에게 적합한 내용을 찾을 수 있도록 유도하는 것이다.

이후 토픽 검색이 TREC Web Track의 과제의 하나로 채택되면서, 토픽 검색을 위한 검색 모형과 신뢰성 측정 방법에 대한 다양한 연구가 진행되었다. 이를 위해서는 토픽 검색의 정의와 목적에 대한 구체적인 서술이 필요하기 때문에 TREC에서는 토픽 검색을 위한 새로운 정의를 내리고, 이 정의에 따라 질의와 적합문헌을 선정하였다.

특히 TREC-2003과 TREC-2004의 토픽 검색에서는 주제에 대한 사이트나 하위사이트를 검색하는 데에 더 집중할 수 있도록 웹 자원의 신뢰성에 대한 정의를 명확히 하였다. 주제에 관한 사이트가 신뢰성 있는 자원, 또는 이용자에게 제시할 소수의 중요 자원이 되는 이유는 이 사이트가 하나의 기관 혹은 한 저자가 단일 주제에 대해 작성한 체계적인 정보일 것이기 때문이다. 반대로 주제에 대한 많은 페이지를 링크하는 경우에는 다수의 단편적인 정보를 제공하게 될 가능성이 있다. 따라서 특정한 주제에 관한 몇 개의 사이트/하위 사이트를 검색하여 이용자에게 제시하는 것이 이용자에게는 보다 만족스러운 검색 결과가 될 수 있을 것이다.

즉 최근 연구에서는 소수의 중요한 자원을 검색한다는 토픽 검색의 기본적인 정의는 변하지 않았으나, 중요한 자원을 웹 사이트 단위로 제한한 것으로 보인다. 웹 사이트 단위로 제한한 것은 특정 주제에 대해서 다루고 있는 사이트를 이용자에게 추천하기 위한 목적인 것으로 해석할 수 있다.

토픽 검색에서 검색해야 할 사이트는 다음과 같은 조건을 갖는다(Craswell and Hawking 2003). 첫째, 주로 해당 주제에 대해 다루고 있는 사이트, 둘째, 주제에 대한 믿을만한 정보를 제공하는 사이트, 셋째, 해당 주제에 대해 다루고 있는 더 큰 사이트의 일부분이 아닌 사이트 등이다. 이 중 세 번째 조건 때문에 토픽 검색에서 검색해야 할 사이트는 일반적으로 말하는 웹 사이트와는 다르다. 동일 도메인 안에서, 특정 주제에 대해서 다루고 있으면서 계층 구조를 이루고 있는 웹 페이지의 집합이 토픽 검색에서 말하는 사이트가 된다. 따라서 토픽 검색은 사이트로 정의된 웹 페이지 집합으로의 접근점이 되는 엔트리 페이지(entry page)를 검색하는 것이다.

TREC의 토픽 검색 연구에서는 주로 웹 사이트 내에서 웹 페이지가 갖는 계층적 구조를

이용하여, 하위에 적합한 페이지들을 가지고 있을 것으로 추측되는 사이트의 엔트리 페이지를 검색한다. 이를 위한 구체적인 방법으로는 URL의 길이를 이용하여 짧은 길이의 URL을 가진 웹 페이지를 검색하거나(Zaragoza et al. 2004), URL의 유형을 나누어 엔트리 페이지의 URL을 가진 웹 페이지를 검색하는 방법(Plachouras et al. 2003; Tomlinson 2003) 등이 있다. 일부 연구(Zhang et al. 2003; Song et al. 2004; Qin et al. 2007)에서는 웹 사이트의 계층 구조 안에서 페이지 간의 상대적인 위치 관계를 고려하여 토픽 검색을 수행하였다.

3. 토픽 검색 실험 설계

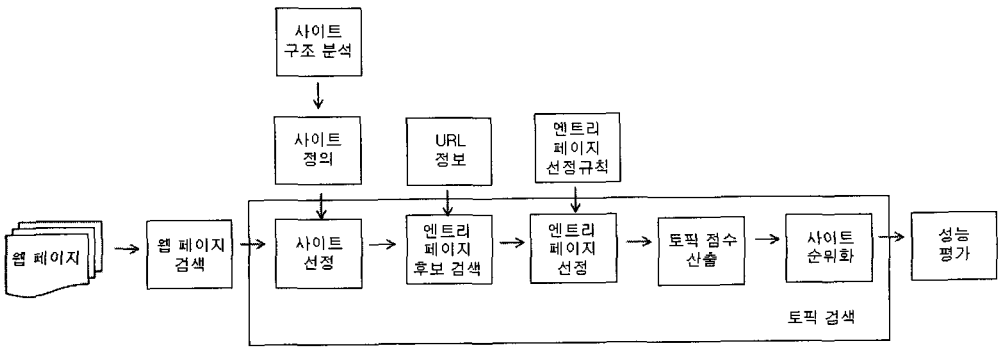
3.1 실험 개요

이 연구의 토픽 검색 실험은 웹 페이지 검색, 사이트 선정, 엔트리 페이지 선정, 토픽 점수 산출, 검색 사이트 순위화의 순서로 진행된다. 질의에 대한 웹 페이지 검색에는 오카피(Okapi) 시스템의 BM25 함수를 이용하였고, 이 검색 결과 집합에 대해 토픽 검색 알고리즘을 적용하였다.

각 질의에 대한 웹 페이지 검색 결과 상위 1,000개의 웹 페이지 집합 안에서 사이트 선정 규칙에 따라 토픽 검색 대상이 되는 사이트와 하위 사이트를 선정하고, 각 사이트와 하위 사이트의 엔트리 페이지를 선정한다. 선정된 사이트/하위 사이트의 토픽 점수를 계산하고, 이 점수에 따라 각 엔트리 페이지를 정렬한 것이 토픽 검색 결과가 된다. 검색된 사이트는 TREC 적합문헌 리스트와 비교하여 알고리즘의 성능을 평가한다. 전체 실험 과정은 〈그림 1〉과 같다.

3.2 실험 집단

토픽 검색 실험에 사용한 TREC의 .GOV 실험 문헌 집단은 수집 당시인 2002년 초 .gov 도메인 내의 모든 웹 페이지를 수집한 것으로서 전체 18.1GB, 1,247,753개 웹 페이지를 포함하고 있다. 토픽 검색 연구를 위한 질의와 적합문헌 집합을 가지고 있는 웹 실험 문헌 집



〈그림 1〉 실험 개요

〈표 1〉 실험용 질의 집합

질의 번호	질의	적합문헌 수	질의 번호	질의	적합문헌 수
Q5	American music	27	Q111	species identification	31
Q12	oil petroleum resources	33	Q115	information security	20
Q15	welfare reform	17	Q136	career information	114
Q19	toxic waste	40	Q137	maps, mapping, cartography	40
Q21	substance abuse	21	Q148	low-income housing	25
Q30	HIV/AIDS	51	Q153	technology transfer	75
Q48	federal and state statistics	86	Q157	federal grant programs	147
Q53	telecommuting	16	Q165	national laboratories	49
Q58	automobile emissions vehicle pollution	34	Q169	how to save energy	26
Q73	solar flares	20	Q177	minority business	30
Q74	skin cancer	19	Q185	estuary conservation	24
Q99	salmon	29	Q189	homeopathic medicine	18
Q101	migrant farm workers	26	Q194	cigarettes, nicotine, health	32
Q103	vanity license plates	19	Q200	maritime trade	34
Q104	space exploration	85	Q204	endangered species	16

단으로는 현재 .GOV가 유일하다.

실험용 질의 집합으로는 TREC-2004의 질의 집합을 구성하는 총 75개의 토픽 검색 질의 중 적합문헌의 수가 많은 30개 질의를 선정하였다. 〈표 1〉은 실험에 사용한 질의와 각 질의에 대한 적합문헌 수를 보여 준다.

3.3 웹 페이지 검색 기법

웹 페이지 검색 이전에 실험 문헌 집단의 모든 페이지를 대상으로 Porter's Stemmer를 이용하여 텍스트를 색인하였다. 일부 페이지는 추출할 텍스트가 존재하지 않았으므로 전체 1,247,753개 웹 페이지 중에서 1,094,936개의 페이지가 색인되었다.

30개의 토픽 검색 질의에 대한 웹 검색은 오카피 시스템의 BM25 함수를 이용하였다. BM25는 그간 TREC 실험을 통해 성능을 인정받아 왔으며, 웹 검색에서도 우수한 성능을 보여주었다(Robertson et al. 1994; Qin et al. 2007). 뿐만 아니라 TREC의 토픽 검색에서도 BM25는 상대적으로 좋은 결과를 낸 것으로 보고되었다(Kamps et al. 2003; MacFarlane 2002; Zhang et al. 2002).

문헌 D_i의 질의 Q에 대한 적합성 점수는 다음과 같이 계산한다.

$$SC(Q, D_i) = \sum_{j=1}^{t} w \times \frac{(k_1+1)\, tf_{ij}}{K + tf_{ij}} \times \frac{(k_3+1)\, qtf_j}{k_3 + qtf_j}$$

$$K = k_1((1-b) + b \times \frac{dl_i}{avdl})$$

위 공식에서 tf_{ij}는 문헌 D_i에서 색인어 j의 출현 빈도를, qtf_j는 질의어의 빈도를 말한다. dl_i는 문헌 D_i의 길이를, $avdl$은 문헌 길이의 평균을 의미한다. t는 전체 색인어의 수를 나타내는데, 따라서 문헌 D_i의 질의에 대한 적합성 값은 문헌 내 각 색인어의 적합성 값의 합이 된다. 상수 k_1과 k_3, b는 각각 tf, qtf, dl을 조절하며 상수 값은 이전의 토픽 검색 연구에서와 같이 k_1=1.5, k_3=8, b=0.8로 하였다. w는 Robertson and Sparck Jones(1976)의 적합성 가중치 공식으로 $w(4)$의 초기값 공식을 사용하였다. 전체 문헌 집단의 수를 N으로, 색인어가 출현한 문헌의 수를 n으로 표현할 때, 초기값 공식은 다음과 같다.

$$w = \log \frac{N+0.5}{n+0.5}$$

3.4 토픽 검색 알고리즘

3.4.1 사이트 및 엔트리 페이지 선정

일차 검색된 웹 페이지를 포함하는 웹 사이트 중에서 토픽 검색 대상 사이트/하위 사이트를 선정한다. 검색된 웹 페이지 집합으로부터 적합한 모든 하위 사이트를 선정하고자 하였기 때문에, 웹 페이지의 디렉토리 경로를 분석하여, 이들을 적절히 포함하는 모든 가능한 하위 사이트를 토픽 검색 대상 사이트로 선정하였다. 검색된 웹 페이지를 적절히 포함하는 사이트에 대한 두 가지 선정 규칙을 정의하였다.

규칙-1) 검색된 웹 페이지가 포함된 URL 디렉토리 경로는 하나의 하위 사이트가 된다.

규칙-2) 규칙-1로 찾은 2개 이상의 하위 사이트의 상위 디렉토리 경로가 같다면, 이 상위 경로도 하나의 사이트가 된다.

예를 들어 질의에 대하여 두 개의 웹 페이지 www.nih.gov/health/food.html과 www.nih.gov/health/nutrition.html이 검색되었다면 사이트 선정 규칙-1에 따라 하위 사이트인 www.nih.gov/health/는 토픽 검색 대상 사이트로 선정된다. 또한 이 하위 사이트와 함께 www.nih.gov/drugs/가 하위 사이트로 선정되었다면 사이트 선정 규칙-2에 따라 이들의 상위 사이트인 www.nih.gov/도 토픽 검색 대상 사이트로 선정된다.

사이트 선정 규칙-2에 의하면, 동일 수준의 하위 사이트를 둘 이상 가진 디렉토리 URL은 검색된 웹 페이지가 없다고 해도 사이트가 된다. 질의에 대한 내용을 담고 있는 두 하위 사이트를 가지고 있는 사이트라면 이 상위 사이트 전체가 질의에 대한 것일 수 있다고 생각하였기 때문이다. 토픽 검색의 정의에 따르면 사이트는 질의에 대해서 다루고 있는 더 큰 사이트의 일부여서는 안 된다. 따라서 둘 이상의 하위 사이트가 질의에 관한 것이라면, 이 둘을 연결하는 상위 디렉토리 URL이 토픽 검색 결과에 포함되고 두 개의 하위 사이트는 제외되어야 한다. 그러나 이들 또한 토픽 검색을 위한 사이트로 선정하였는데, 규칙-2에 따라 선

정된 사이트가 갖는 모든 하위 사이트가 질의에 대한 것인가를 확신할 수 없기 때문이다. 상위 사이트가 질의에 관한 것인가는 사이트의 적합성 점수를 계산함으로써 결정될 수 있을 것이다.

토픽 검색 대상 사이트를 선정한 뒤에는, 각 사이트를 대표할 엔트리 페이지를 선정한다. 우선 엔트리 페이지가 될 수 있는 후보 URL을 검색한 후, 이 중에서 엔트리 페이지를 선정한다. 엔트리 페이지 후보 리스트는 전체 실험 문헌 집단 안에서 선정된 사이트/하위 사이트의 웹 페이지의 URL로 구성된다. 선정된 사이트/하위 사이트의 URL을 입력하여 이와 동일한 디렉토리 경로를 가진 웹 페이지의 URL을 검색하는 것이다.

후보 URL 리스트에서 엔트리 페이지를 선정할 때에는 URL의 파일명 정보를 이용하는데, 파일명의 확장자를 제외한 부분만을 이용한다. 본 연구에서는 선행 연구(Tomlinson 2002; Lim et al. 2005; Sun and Lim 2003)에서 사용한 index, default, main, home, main_default, welcome, homepage 등을 엔트리 페이지가 될 수 있는 파일명의 조건으로 정하였는데, 사용한 7개의 단어가 포함된 URL이 2개 이상 있을 경우를 위하여 7개 파일명의 사용에 우선순위를 정하였다. 사용 순서는 index, main, default, home, welcome, homepage로 이는 각각의 파일명이 전체 실험 문헌 집단 내에서 출현한 수에 근거한 것이다. main_default는 실험 문헌 집단 내 웹 페이지의 URL에 포함되지 않아 제외하였다. 이들 파일명과 함께 마지막 디렉토리명이 파일명에 포함되는 URL을 엔트리 페이지로 정하였다.

위의 방법으로 엔트리 페이지를 선정할 수 없는 경우에 한하여 질의어가 포함되는 파일명을 가진 웹 페이지도 엔트리 페이지가 될 수 있다고 보았다. 이는 엔트리 페이지로 보이는 웹 페이지가 없다면 이 사이트의 주제어 즉 질의어가 파일명에 포함되었을 것으로 여겨지기 때문이다. 엔트리 페이지 선정을 위한 질의어로는 원래의 질의어와 함께 확장 질의어를 이용하였다. 확장 질의어는 초기 웹 검색 결과 상위 10개 문헌에 출현한 색인어에 대해 Robertson 등(Robertson et al. 2000)이 사용하였던 *TSV*(term selection value)에 따라 질의당 10개 색인어를 선택하였다.

$$TSV = w \times r/R$$

$$w = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-R+0.5)}$$

위 공식에서 N은 전체 문헌 집단의 크기, n은 해당 색인어가 출현한 문헌의 수를 의미하며 R과 r은 각각 전체 적합문헌의 수와 해당 색인어가 출현한 적합문헌의 수를 말한다. 확장 질의어 선정을 위해 이 실험에서는 검색된 상위 10개의 웹 페이지를 이용하였으므로 모든 경우에 R=10이 된다.

사용한 네 가지 규칙을 적절히 사용하기 위하여 각 방법의 사용에 우선 순위를 두고, 정해진 순위에 따라 첫 번째 규칙으로 엔트리 페이지를 찾고, 찾을 수 없는 경우 그 다음 규칙에 따라 엔트리 페이지를 검색하도록 하였다. 〈표 2〉는 엔트리 페이지 검색 규칙의 우선순위와, 해당 규칙이 검색하는 URL의 예시이다. 질의어가 포함되는 파일명을 가진 URL은

〈표 2〉 엔트리 페이지 선정 규칙

우선순위	파일명 규칙	예시
1	파일명 없음	http://pwrc.nbs.gov/resource/
2	파일명 정보	http://pwrc.nbs.gov/resource/index.htm
3	마지막 디렉토리명이 포함	http://pwrc.nbs.gov/resource/resource.htm
4	질의어가 포함	http://pwrc.nbs.gov/resource/oil.htm

다른 모든 규칙으로 엔트리 페이지를 찾을 수 없는 경우에만 사용하도록 하였다.

3.4.2 토픽 점수 산출

검색될 사이트/하위 사이트와 각 사이트의 엔트리 페이지를 선정한 뒤에는 각 사이트/하위 사이트의 토픽 검색 점수를 계산한다. 토픽 점수 $TScore(S_k)$는 사이트의 질의에 대한 적합성 점수 $Rel(S_k)$와 인링크의 수$(Num-Inlink)$를 정규화한 링크 점수를 결합한 값이다.

임의의 하위 사이트 S_k의 적합성 점수 $Rel(S_k)$는 엔트리 페이지 P_E의 적합성 값 $Rel(P_E)$과 하위 페이지 P_i및 하위 사이트 S_{kj}의 적합성 값 $Rel(P_i)$과 $Rel(S_{kj})$의 평균을 합하여 구한다. 이 때 S_k내 하위 페이지와 하위 사이트의 영향력을 조절하기 위한 가중치 W_E와 W_S를 사용한다.

$$Rel(S_k)=\alpha\times Rel(P_E)+(1-\alpha)\left(\frac{\sum_{p_i\in S_k} W_E\times Rel(P_i)+\sum_{S_{kj}\in S_k} W_S\times Rel(S_{kj})}{n(S_k)-1}\right)$$

위 식에서 α는 상수로, 사이트 적합성 점수 $Rel(S_k)$안에서 엔트리 페이지의 영향력을 조절한다. $n(S_k)$는 사이트 S_k의 원소의 수를 말한다. S_k는 엔트리 페이지와 하위 페이지 또는 하위 사이트로 구성되었기 때문에 $n(S_k)-1$은 엔트리 페이지를 제외한 하위 구성 요소의 수를 말한다. 즉 $n(S_k)-1$은 하위 페이지의 수 i와 하위 사이트의 수 j의 합을 의미한다.

W_E와 W_S로는 P_E와 P_i, P_E와 S_{kj}의 유사도 값이나 1 이상의 상수 값을 이용하였다. 하위 페이지의 가중치 W_E를 기준으로 $W_E=SIM(P_E, P_i)$로 사용하는 유사도 가중치 모형(SW)과

〈표 3〉 하위 페이지 가중치(W_E) 및 하위 사이트 가중치(W_S) 값

모형	W_E	W_S	약자
유사도 가중치 모형	$SIM(P_E, P_i)$	1	SW_1
		1.5	SW_1.5
		$SIM(P_E, S_{kj})$	SW_SW
고정 가중치 모형	1	1	FW_1
		1.5	FW_1.5

W_E=1로 고정하는 고정 가중치 모형(FW)의 두 가지 모형을 사용하였다. 각각의 모형 안에서 W_S의 값을 변화시키면서 성능을 평가하였다. 〈표 3〉에 각 모형에서 사용한 가중치 값을 제시하였다.

유사도 가중치 모형에서 P_E와 P_i, P_E와 S_{kj}의 유사도를 이용한 것은, 주어진 사이트가 주제 이외의 다른 내용도 함께 담고 있을 때 이 사이트의 주제에 대한 적합성을 낮추기 위함이다. 즉, 엔트리 페이지와의 주제적 유사성을 계산하여 하위 페이지나 하위 사이트의 내용이 엔트리 페이지 안에서 차지하는 비율을 추정하고자 하였다.

엔트리 페이지 P_E와 하위 페이지 P_i 및 하위 사이트 S_{kj}간의 유사도는 자카드 계수로 산출하였다. 이를 위해 각 페이지/사이트를 대표하는 고빈도어를 추출하였는데 하위 페이지의 경우에는 해당 페이지 안에서, 하위 사이트의 경우에는 하위 사이트 내 검색된 웹 페이지 전체에서 고빈도어를 추출하였다. 실험에서는 각 페이지/사이트에서 10개의 고빈도어를 추출하였고, 유사도 값이 0이 되지 않도록 하기 위하여 최하의 유사도 값으로 0.05를 부여하였다.

P_E와 P_i 또는 S_{kj}에서 동시에 출현한 색인어의 수를 a, P_E에만 출현한 색인어 수를 , b또는 P_i에만 출현한 색인어를 c라고 할 때 P_E와 P_i, P_E와 S_{kj}의 유사도는 다음과 같은 자카드 계수 공식에 의해 산출하였다.

$$SIM(P_E, P_i) = \frac{a}{a+b+c}$$

$$SIM(P_E, S_{kj}) = \frac{a}{a+b+c}$$

적합성 점수 $Rel(S_k)$는 S_k의 중요성 정도를 측정하는 인링크의 수($Num-Inlink$)와 결합되어 최종 토픽 점수 $TScore(S_k)$를 만든다. 웹 페이지가 받는 링크는 동일 사이트 내에서의 링크와 사이트 외부로부터의 링크로 나눌 수 있다. 이 중 동일 사이트 내에서의 링크는 제외하고, 외부로부터의 링크만을 포함하였다. 링크를 계수할 때는 링크를 단위로 사용하지 않고, 웹 페이지를 단위로 하였다. 즉, 사이트 S_k에 포함된 모든 웹 페이지로 링크하는 다른 웹 페이지의 수를 $Num-Inlink$ 로 하였다.

적합성 점수 $Rel(S_k)$와 링크 점수를 결합할 때 두 점수의 범위를 동일하게 맞추기 위하여 각 점수 값을 정규화하였다. 두 최대값의 차이가 크면 한 점수가 전체 점수를 지배하게 될 수 있기 때문이다. 이 연구에서 토픽 점수 $TScore(S_k)$는 두 값의 최소값은 0, 최대값은 $\max(Rel(S))-\min(Rel(S))$이 되도록 다음과 같은 공식을 사용하여 산출하였다.

$$TScore(S_k) = \beta \times (Rel(S_k) - \min(Rel(S))) + (1-\beta) \times \frac{Num-Inlink}{\max(Num-Inlink)} \times (\max(Rel(S)) - \min(Rel(S)))$$

4. 실험 결과 분석 및 평가

4.1 실험 결과 분석

실험 문헌 집단에서 30개의 질의에 대하여 웹 페이지 검색을 수행하고, 이 결과를 대상으로 토픽 검색을 수행하였다. 질의별 검색 결과 상위 1,000개 페이지의 URL에서 도메인과 디렉토리 경로를 분석하여 사이트 선정 규칙에 따라 사이트/하위 사이트를 선정하였다. 30개

질의에 대하여 총 8,359개, 질의별 평균 279개의 사이트/하위 사이트가 선정되었으며, 사이트의 형태로 나누어 보았을 때 사이트가 약 17%, 하위 사이트가 약 83%를 차지하였다.

토픽 검색 결과 성능은 정확률 척도인 MAP와 n-순위 정확률을 이용하여 평가하였다. MAP는 전체 검색 결과에서의 적합문헌의 출현과 순위를 모두 고려한 성능 평가 방법으로, 적합문헌의 출현 시마다 산출한 정확률의 평균값을 말한다. n-순위 정확률은 검색 결과 상위에 적합문헌이 포함되었는가를 측정하여 성능을 평가하는 것으로, 일정한 지점에서 적합문헌이 포함된 비율로 측정된다. 토픽 검색의 평가에서는 10위에서의 정확률을 산출하여 성능을 평가하며 이를 P(10)으로 표현한다.

30개 질의에 대하여 선정된 총 8,359개의 사이트/하위 사이트에서 엔트리 페이지를 선정한 결과는 〈표 4〉와 같다. 파일명이 없는 URL이 엔트리 페이지로 선정된 경우가 가장 많았고, 질의어를 이용하거나 마지막 디렉토리명이 파일명에 포함되는 URL이 index 등의 6개의 파일명이 포함되는 경우보다 많았다.

〈표 5〉는 가중치 값에 따른 다섯 가지 모형을 통한 토픽 검색 결과의 성능을 MAP와 P(10)으로 평가한 결과이다. $Rel(S_k)$공식과 $TScore(S_k)$공식에서 α 와 β 를 각각 0.3, 0.5, 0.7로 변화시켜 토픽 점수를 산출하였다.

P(10)으로 성능을 평가하였을 때 유사도 가중치 모형에서는 하위 사이트 가중치(W_S)로 유사도를 이용한 경우(SW_SW)와 1을 부여한 경우(SW_1) 모두 α=0.7, β=0.5에서 각각 0.177과 0.183의 P(10) 값을 가지며 가장 좋은 성능을 보였다. W_S에 1.5의 값을 부여한 경우(SW_1.5)에는 α=0.5, β=0.7에서 P(10)이 0.193로 가장 좋은 결과를 얻을 수 있었다. 고정 가중치 모형의 실험 결과 하위 사이트 가중치 W_S=1에서 가장 높은 P(10)은 0.210으로 α=0.5, β=0.5일 때와 α=0.7, β=0.7일 때 이런 결과를 얻을 수 있었다. W_S=1.5에서 가장 높은 P(10)은 0.217로 고정 가중치 모형 전체에서 가장 좋은 성능을 보인 결과이며 이 때 α와 β는 모두 0.5의 값을 가지고 있다.

〈표 4〉 각 엔트리 페이지 선정 규칙이 적용된 사이트/하위 사이트의 수

엔트리 페이지 선정 규칙		해당 사이트/하위 사이트의 수
파일명 없음		6,185
파일명 정보	index	131
	main	54
	default	29
	home	113
	welcome	64
	homepage	22
마지막 디렉토리명		870
질의어		891

〈표 5〉 토픽 검색 결과

모형	α	β	MAP	P(10)
SW_SW	0.3	0.3	0.056	0.150
		0.5	0.057	0.143
		0.7	0.053	0.140
	0.5	0.3	0.059	0.167
		0.5	0.063	0.163
		0.7	0.060	0.150
	0.7	0.3	0.061	0.170
		0.5	0.067	0.177
		0.7	0.066	0.157
SW_1	0.3	0.3	0.058	0.153
		0.5	0.060	0.150
		0.7	0.059	0.150
	0.5	0.3	0.061	0.170
		0.5	0.065	0.167
		0.7	0.063	0.163
	0.7	0.3	0.062	0.167
		0.5	0.065	0.183
		0.7	0.065	0.160
SW_1.5	0.3	0.3	0.058	0.170
		0.5	0.061	0.153
		0.7	0.059	0.150
	0.5	0.3	0.061	0.163
		0.5	0.063	0.193
		0.7	0.064	0.163
	0.7	0.3	0.063	0.173
		0.5	0.065	0.190
		0.7	0.066	0.180
FW_1	0.3	0.3	0.053	0.160
		0.5	0.056	0.173
		0.7	0.056	0.167
	0.5	0.3	0.059	0.193
		0.5	0.063	0.210
		0.7	0.065	0.190
	0.7	0.3	0.063	0.200
		0.5	0.067	0.207
		0.7	0.071	0.210
FW_1.5	0.3	0.3	0.055	0.160
		0.5	0.058	0.197
		0.7	0.059	0.190
	0.5	0.3	0.061	0.200
		0.5	0.065	0.217
		0.7	0.064	0.203
	0.7	0.3	0.064	0.200
		0.5	0.068	0.207
		0.7	0.069	0.187

α와 β 값의 변화에 관계없이 하위 사이트 가중치 값이 높아질수록 성능이 향상되는 경향이 있는 것으로 나타났다. 즉 유사도 가중치 모형의 실험에서 하위 사이트 가중치로 유사도를 부여한 경우(SW_SW)의 P(10)이 가장 낮으며, 그 다음이 가중치로 1을 준 경우, 전체적으로 가장 좋은 성능을 보이는 것이 1.5를 가중치 값으로 이용하는 SW_1.5인 것으로 나타났다. 이 결과는 고정 가중치 모형에서도 비슷하게 나타난다. 하위 사이트 가중치로 1을 부여한 FW_1보다 1.5를 부여한 FW_1.5에서 더 좋은 성능을 보여주었다. 따라서 하위 사이트에 높은 가중치를 주어, 하위 사이트 적합성 값을 많이 반영하는 것이 토픽 검색에서 유리하다는 것을 확인하였다. 이 결과는 토픽 검색을 위한 전략에서 중요한 의미를 갖는다. 하위 사이트 가중치가 높다는 것은 하위 사이트의 적합성 값의 반영 비율이 높아진다는 것을 의미하기 때문이다. 즉, 사이트의 계층적 구조를 이용하여 질의에 대한 사이트/하위 사이트의 적합성 값을 계산하고자 하였던 것이 토픽 검색에서 유효하게 작용하였음을 보여준다.

동일한 하위 사이트 가중치 값을 부여한 실험에서, α 값의 변화에 따라 성능의 차이가 나타난다. α=0.3일 때의 성능이 0.5나 0.7의 값을 가진 경우보다 대체로 더 낮은 것을 볼 수 있다. 모든 경우에 가장 낮은 P(10)은 α=0.3인 경우 나왔다는 점을 통해서도 이를 확인할 수 있다.

β 값의 특성은, β=0.5 일 때 가장 좋은 결과를 가져올 수 있다는 것이다. β 는 토픽 점수 $TScore(S_k)$를 계산하는 공식에서 나온 것으로, β 값은 링크 점수의 반영 비율을 조절한다.

각 모형에서의 결과를 보면, 대부분의 경우 β =0.5일 때 토픽 검색의 성능이 가장 좋았다. 따라서 링크 점수가 전체 토픽 점수에서 0.5의 비율을 차지하는 것이 가장 유리하였음을 알 수 있다.

유사도 가중치 모형과 고정 가중치 모형을 비교하여보면, 고정 가중치 모형에서 더 나은 토픽 검색 결과를 얻을 수 있음을 확인할 수 있다. 유사도 가중치 모형에서 최고의 P(10)은 0.193이나, 고정 가중치 모형에서는 0.217의 값을 가져, 고정 가중치 모형이 유사도 가중치 모형보다 12.4% 향상된 성능을 보이고 있다.

유사도 가중치 모형과 고정 가중치 모형은 각 사이트/하위 사이트의 하위 페이지에 부여하는 가중치 값을 결정하는 것으로, 유사도 가중치 모형에서는 엔트리 페이지와 하위 페이지와의 유사도를 이용하여 두 페이지 간의 주제적 유사성을 평가하였다. 그런데 유사도 가중치 모형보다 고정 가중치 모형이 더 좋은 성능을 보인다는 것은 사이트의 적합성 계산에서 엔트리 페이지와 하위 페이지 간의 주제적 유사성을 반영하는 것이 토픽 검색에서 유효하지 않았음을 의미한다. 유사도에 대한 결과는 하위 사이트의 가중치의 변화에 따른 성능의 차이에서도 드러난다. 〈표 5〉에서 하위 사이트 가중치로 엔트리 페이지와의 유사도를 이용한 경우의 토픽 검색의 성능이 고정된 가중치를 이용한 경우보다 떨어진다는 것을 확인하였다. 오히려 하위 사이트의 적합성 값에 높은 가중치를 주는 것이 토픽 검색에서 유리하였다. 따라서 엔트리 페이지와 하위 페이지, 엔트리 페이지와 하위 사이트의 유사도를 통하여 사이트의 질의에 대한 적합성 정도를 평가할 수 있을 것으로 보았던 유사도 가중치 모형의 가정이 옳지 않았음을 알 수 있다.

4.2 토픽 검색 성능 비교

토픽 검색 알고리즘의 성능을 평가하기 위하여 TREC-2004의 토픽 검색 연구 결과와 본 연구에서의 토픽 검색 실험 결과를 비교하였

〈표 6〉 TREC-2004 토픽 검색 결과와의 비교 평가

기관		MAP	P(10)
기관명	약자		
Microsoft Research Asia	MSRA	0.178	0.251
University of Glasgow	UOG	0.179	0.249
Microsoft Research Cambridge	MSRC	0.165	0.231
Hummingbird	HUM	0.163	0.231
University of Amsterdam	UAMS	0.146	0.209
Chinese Academy of SciencesInstitute of Computing Technology	ICT	0.141	0.208
Tsinghua University	THU	0.147	0.205
본 연구의 실험 결과		0.065	0.217

다. TREC-2004에 참여한 18개 기관의 공식적인 토픽 검색 성능 평가 결과 MAP는 최소 0.003에서 최대 0.179, P(10)은 최소 0.011에서 최대 0.251 사이의 값을 보였다(Craswell and Hawking 2004). P(10) 값이 상위에 오는 7개 기관의 성능 평가값을 본 연구에서의 결과와 비교하여 〈표 6〉에 제시하였다. 결과적으로 본 연구에서 구현한 토픽 검색 알고리즘은 TREC-2004의 연구 결과들과 비교하였을 때 상위 5위에 해당되는 우수한 성능을 보였다.

TREC-2004와의 성능 비교에서 MAP를 이용한 평가는 P(10)에서처럼 높은 순위를 갖지 못한다. 이는 TREC과 본 연구에서의 토픽 검색의 단위가 서로 다르기 때문이다. 단위가 서로 다르다는 것은 두 가지 문제를 야기한다. 하나는 검색 결과 리스트의 크기가 서로 다르다는 것이며, 다른 하나는 평가를 위한 적합문헌의 단위가 본 연구 결과와 서로 달라 토픽 검색의 성능을 제대로 평가할 수 없다는 것이다.

TREC-2004에서는 각 참여 기관의 검색 결과를 평가할 때, 상위 1,000개의 웹 페이지를 대상으로 하였다(Craswell and Hawking 2004). 반면 본 연구에서는 1,000개의 웹 페이지 검색 결과에서 사이트를 선정하도록 하였기 때문에 검색 결과 웹 페이지는 1,000개가 될 수 없었다. 결과적으로 토픽 검색 결과 상위 10개의 웹 페이지를 대상으로 평가한 P(10)에서는 이전의 연구 결과와 본 연구 결과를 비교하는 것이 타당하지만, 1,000개의 결과를 대상으로 하는 TREC-2004의 MAP와 본 연구의 MAP를 비교하는 것은 의미가 없다고 볼 수 있다.

또 다른 이유는 TREC의 적합문헌의 리스트가 역시 사이트를 단위로 하지 않는다는 것이다. 즉 이 연구에서는 사이트/하위 사이트를 선정하고, 각각의 엔트리 페이지를 선정하는 방법으로 토픽 검색을 수행하였으나 TREC의 적합문헌 리스트는 사이트를 단위로 하지 않고 하위의 웹 페이지를 함께 포함하고 있어, 토픽 검색의 대상이 되는 적합 사이트의 리스트라고 보기 어렵다.

본 연구에서 사이트를 단위로 한 검색 결과를 TREC에서 제공하는 적합문헌 리스트와 비교하였기 때문에 토픽 검색의 성능이 낮게 평가되었을 가능성이 있다. 따라서 성능 평가 기준이 되는 TREC의 적합문헌 리스트를 분석한 결과를 반영하여 본 연구의 토픽 검색 알고리즘의 성능을 재평가하였다.

4.3 TREC 적합 페이지/사이트 URL과의 비교 분석

TREC-2004의 토픽 검색 질의에 대한 적합문헌 리스트에서 각 문헌, 즉 웹 페이지의 URL 리스트를 만들어 이를 분석하였다. 토픽 검색의 정의에 초점을 맞추어, 이들이 사이트의 엔트리 페이지로 구성되어 있는가를 판단하였고, 엔트리 페이지가 아닌 하위 페이지가 포함된 경우, 이들의 웹 페이지를 확인하였다. 실험에서 이용하였던 질의 전체를 분석하기에는 이용한 질의의 수가 많아, 이들 30개의 질의 중에서 Q5와 Q136, Q153, Q157의 4개 질의를 임의로 선택하였다. 이들 질의의 적합문헌 리스트에서 동일한 사이트/하위 사이트로부터의 다수의 웹 페이지가 적합문헌으로 선정

되어 있는 경우를 찾아, 각각의 웹 페이지를 확인하였다. 동일한 사이트/하위 사이트에서 다수의 웹 페이지가 적합문헌으로 평가되어 있는 경우를 23개 사이트/하위 사이트에서 찾을 수 있었으며, 이들이 적합문헌으로 평가된 이유를 분석하였다.

23개의 사이트/하위 사이트에 포함된 37개의 하위 페이지를 분석한 결과 28개의 웹 페이지가 엔트리 페이지가 아닌, 질의에 관련된 내용을 가진 페이지였다. 일부 사이트의 경우에는 질의와 관련된 내용을 가진 두 개의 웹 페이지가 적합 페이지의 리스트에 포함되어 있었다. 이는 전체 사이트/하위 사이트가 질의를 포함하지만 더 넓은 주제에 대해서 다루고 있기 때문이었다. 그러나 토픽 검색의 목적을 생각할 때, 이 사이트 전체가 질의에 대하여 다루고 있지는 않다 하더라도 나름의 관점에서 질의 및 관련 내용을 구성하였을 것이기 때문에 이들의 엔트리 페이지가 포함되는 것이 옳다. 엔트리 페이지를 제공해야 이 사이트의 관점에서 질의에 관련된 내용을 전체적으로 볼 수 있으리라는 것이다.

그 외 2개 페이지는 URL로 확인하였을 때 엔트리 페이지가 아니었으나, 해당 하위 사이트의 링크 구조 등을 통하여 보았을 때 이들이 실질적으로 엔트리 페이지로 기능하는 것으로 나타났다.

따라서 토픽 검색을 위한 평가의 단위가 페이지가 아닌 사이트가 되어야 한다는 관점에서, 제공된 적합문헌 리스트에서 평가의 단위를 사이트로 바꾸어 토픽 검색의 성능을 재평가하였다. 원래의 적합문헌이 사이트/하위 사이트의 엔트리 페이지인 경우에는 수정이 필요하지 않았지만 www.nea.gov/artforms/Music/01music.htm과 같은 하위 페이지의 경우에는 이를 하위 사이트 즉 www.nea.gov/artforms/Music/로 수정하였다. 이 때 동일한 하위 사이트의 둘 이상의 하위 페이지가 적합문헌으로 포함된 경우, 이를 사이트 단위로 수정하면 이들의 사이트 URL 하나만 적합 사이트의 리스트에 포함된다.

적합문헌 리스트를 수정한 적합 사이트 리스트와 토픽 검색 결과를 비교하고 토픽 검색 결과 성능을 다시 평가하였다. 비교를 위해 전체 실험 결과 중에서 가장 성능이 좋았던 실험 환경을 선택하였다. 〈표 7〉은 이를 이용한 토픽 검색 성능 평가 결과를 정리한 것이다.

새로운 성능 평가 결과, 분석 대상 4개의 질

〈표 7〉 적합문헌 리스트 수정 전후의 성능 평가 결과

질의번호	적합문헌 리스트		MAP		P(10)	
	수정 전	수정 후	수정 전	수정 후	수정 전	수정 후
Q 5	27	23	0.031	0.017	0.2	0.1
Q136	114	98	0.031	0.118	0.1	0.4
Q153	75	71	0.173	0.286	0.7	0.8
Q157	147	139	0.011	0.058	0.0	0.2
평균	90.750	82.750	0.061	0.120	0.250	0.375

의에 대한 평균 P(10)과 MAP가 향상하였음을 확인할 수 있다. MAP는 0.061에서 0.120으로 97%, P(10)은 0.25에서 0.375로 50% 향상하였다. 이는 적합문헌에는 하위 페이지가 포함되었으나 실험에서는 엔트리 페이지를 검색한 등의 경우를 올바르게 평가할 수 있었기 때문이다. 또한 적합문헌 리스트에 동일한 하위 사이트에서 엔트리 페이지와 하위 페이지가 동시에 포함된 경우, 토픽 검색 결과에 포함될 수 없었던 하위 페이지를 성능 평가 과정에서 제외시킬 수 있었던 것 때문이기도 하다.

질의별 결과를 보면 Q136과 Q153, Q157에서는 MAP와 P(10)이 향상하였으나 Q5에서는 오히려 하락하였다. Q5의 성능이 오히려 하락한 것은 엔트리 페이지가 없는 사이트가 적합 사이트로 판정되었기 때문이다.

TREC의 적합문헌 리스트를 기준으로 한 평가에서는 토픽 검색 결과의 적합성을 적절하게 평가할 수 없었지만, 이를 다시 사이트 수준으로 수정하여 결과를 평가함으로써 본 연구에서의 토픽 검색에 대한 접근 방법이 질의에 적합한 사이트 및 하위 사이트를 검색하는 데에 효과적이었음을 확인할 수 있다. 특히 재평가 결과 상위 10위 안에서 평균 3~4개의 적합 사이트를 찾을 수 있었던 것은, 적합한 정보가 상위에 오도록 하는 것이 중요한 웹 환경에서 매우 의미 있는 결과라고 볼 수 있다.

5. 결 론

이 연구에서는 질의에 대한 사이트를 검색하는 토픽 검색을 위하여 웹 사이트의 구조를 이용, 사이트 및 하위 사이트의 질의에 대한 적합성을 평가하는 알고리즘을 제안하고 성능을 평가하였다.

토픽 검색 실험을 통하여 발견한 사실은 다음과 같다.

첫째, 하위 페이지 적합성 값의 영향력을 반영하는 두 가지 가중치 모형 중에서 고정 가중치 모형이 유사도 가중치 모형보다 12.4% 향상된 성능을 보였다. 즉, 하위 페이지의 적합성 값에 보다 높은 가중치를 줌으로써 사이트의 적합성 점수 산출시 하위 페이지 적합성 값의 반영 비율을 높이는 것이 토픽 검색의 성능 향상에 더 유리한 것으로 나타났다. 그리고 엔트리 페이지와 하위 페이지 간의 유사도를 통하여 두 웹 페이지의 주제적 유사성을 측정, 전체 사이트의 주제 구조를 추정하고자 하였던 유사도 가중치 모형은 비효과적인 것으로 나타났다.

둘째, 하위 사이트의 가중치를 높일수록 토픽 검색 결과 성능이 향상되었다. 따라서 하위 페이지의 경우와 마찬가지로 사이트의 적합성 점수에서 하위 사이트 적합성 값이 중요하게 작용함을 알 수 있었다.

셋째, 하위 페이지와 하위 사이트 적합성 값의 반영 비율을 높일 경우 더 좋은 성능을 보인 실험 결과를 통하여 사이트의 계층 구조를 반영하여 질의에 대한 사이트의 적합성 정도를 평가하는 것이 토픽 검색에 적절한 전략임을 입증하였다.

넷째, 링크 점수의 반영 비율을 조절하는 β 값이 0.5일 때 가장 좋은 성능을 나타냈다. 즉, 사이트의 토픽 점수에서 링크 점수가 적합성 점수와 동일한 비율로 반영될 때 가장 좋은 성

능을 보이며, 이를 통하여 인링크의 수를 토픽 검색에서 이용하는 것이 적절하다는 것 또한 알 수 있었다.

다섯째, 토픽 점수를 구성하는 세 가지 값, 즉 엔트리 페이지의 적합성 값과 하위 페이지/하위 사이트의 적합성 값, 링크 점수 모두가 검색 성능에 영향을 주었다. 이들 세 가지 값이 적절히 반영되어 토픽 점수를 구성하여야 토픽 검색의 성능이 향상될 수 있었다.

이 연구에서 제안한 알고리즘의 성능을 TREC-2004의 토픽 검색 실험 결과와 비교하였다. TREC-2004에 참여한 18개 기관들 중 P(10) 값이 가장 높은 상위 7개 기관과의 성능 비교에서 이 연구의 토픽 검색 알고리즘은 P(10)이 0.217로서 상위 5위 수준의 우수한 성능을 보였다.

토픽 검색 성능 평가를 위한 TREC의 적합 문헌 리스트를 토픽 검색의 정의에 따라 수정하여 실험 결과를 다시 평가한 결과 P(10)이 0.250에서 0.375로 50%, MAP가 0.061에서 0.120으로 97% 향상되었다. 결론적으로 이 연구에서 제안한 토픽 검색 알고리즘의 성능은 TREC에 참여한 연구기관들의 실험 결과와 비교할 때 상위 수준에 속하는 것으로 나타났으며, 토픽 검색의 대상이 되는 사이트의 정의에 충실한 적합문헌 리스트가 제공된다면 그 성능이 더욱 높아질 것으로 기대된다.

참 고 문 헌

박기림, 장유진, 김민구, 박승규. 2003. "문서 내의 주제정보를 이용한 개선된 링크 분석 알고리즘." 『한국정보과학회 학술발표논문집』 30(2): 7-9.

Bahrat, K., and Henzinger, M. R. 1998. "Improved Algorithms for Topic Distillation in a Hyperlinked Environment." In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, 104-111.

Bahrat, K., and Mihaila, G. A. 2002. "When experts agree: Using non-affiliated Experts to rank popular topics." *ACM Transactions on Information Systems*, 20(1): 46-58.

Chakrabarti, S., Berg, M., and Dom, B. 1999. "Focused Crawling: A new approach to topic-specific web resource discovery." *Proceedings of Eighth International World Wide Web Conference*. 〈http://www8.org/w8-papers /5a-search-query/crawling/index.html〉

Craswell, N., and Hawking, D. 2003. "Task Descriptions: Web Track 2003." In *Proceedings of the*

Twelfth Text Retrieval Conference (TREC-12). 〈http:trec.nist.gov/pubs/trec12/papers/web03.guidelines.pdf〉.

Craswell, N., and Hawking, D. 2004. "Overview of the TREC-2004 Web Track." In : *Proceedings of the Thirteenth Text Retrieval Conference (TREC-13)*. 〈http :trec. nist.gov/pubs/trec13/ papers/WEB.OVERVIEW.pdf〉.

Kamps, J., Monz, C., Rijke, M., and Sigurbjörnsson, B. 2003. "Approaches to Robust and Web Retrieval." In *Proceedings of the Twelfth Text Retrieval Conference(TREC-12)*. 〈http:trec.nist.gov/pubs/trec12/papers/uamsterdam.web.robust.pdf〉.

Kleinberg, J. M. 1999. "Authoritative sources in a hyperlinked environment." *Journal of ACM* 46(5): 604-632.

Lim, C. S., Lee, K. J., and Kim, G. C. 2005. "Multiple sets of features for automatic genre classification of web documents." *Information Processing and Management,* 41(5): 1263-1276.

MacFarlane, A. 2002. "Pliers at TREC 2002." In *Proceedings of the Eleventh Text Retrieval Conference (TREC-11)*. 〈http:trec.nist.gov/pubs/trec11/papers/cityu.pliers.pdf〉.

Plachouras, V., Cacheda, F., Ounis, I., and Rijsbergen, C. J. 2003. "University of Glasgow at the Web Track: Dynamic Application of Hyperlink Analysis using Query Scope." In *Proceedings of the Twelfth Text Retrieval Conference (TREC-12)*. 〈http:trec.nist.gov/pubs/trec12/papers/uglosgow.web.pdf〉.

Qin, T., Liu, T., Zhang, X., Feng, G., Wang, D., and Ma, W. 2007. "Topic distillation via sub-site retrieval." *Information Processing & Management* 43(2): 445-460.

Robertson, S.E. and Sparck Jones, K. 1976. "Relevance weighting of search terms." *Journal of the American Society and Information Science*, 27(3): 129-146.

Robertson, S.E., Walker, S. Beaulieu, M. 2000. "Experimentation as a way of life: Okapi at TREC." *Information Processing & Management* 36(1): 95-108.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., and Gatford, M. 1994. "Okapi at TREC-3." In *Proceedings of*

the Third Text Retrieval Conference (TREC-3). 〈http:trec.nist.gov/pubs/trec3/papers/city.ps.gz〉.

Song, R., Wen, J., Shi, S., Xin, G., Liu, T., Qin, T., Zheng, X., Zhang, J., Xue, G., and Ma, W. 2004. "Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004." In *Proceedings of the Thirteenth Text Retrieval Conference (TREC-13).* 〈http:trec.nist.gov/pubs/trec13/papers/microsoft-asia.web.tera.pdf〉.

Sun, A. and Lim, E. 2003. "Web Unit Mining - Finding and Classifying Subgraphs of Web Pages." In *Proceedings of the twelfth ACM CIKM* : 108-115.

Tomlinson, S. 2002. "Experiments in Named Page Finding and Arabic Retrieval with Hummingbird SearchServer™ at TREC 2002." In *Proceedings of the Eleventh Text Retrieval Conference(TREC-11).* 〈http:trec.nist.gov/pubs/trec11/papers/hummingbird.tomlinson.pdf〉.

Tomlinson, S. 2003. "Robust, Web and Genomic Retrieval with Hummingbird SearchServer™ at TREC 2003." In *Proceedings of the Twelfth Text Retrieval Conference(TREC-12).* 〈http:trec.nist.gov/pubs/trec12/papers/hummingbird.robust.web.genomics.pdf〉.

Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. 2004. "Microsoft Cambridge at TREC-13: Web and Hard Tracks." In *Proceedings of the Thirteenth Text Retrieval Conference(TREC-13).* 〈http:trec.nist.gov/pubs/trec13/papers/microsoft-cambridge.web.hard.pdf〉.

Zhang, M., Lin, C., Liu, Y., Zhao, L., and Ma, S. 2003. "THUIR at TREC 2003: Novelty, Robust and Web." In *Proceedings of the Twelfth Text Retrieval Conference(TREC-12).* 〈http:trec.nist.gov/pubs/trec12/papers/tsinghuau.novelty.robust.web.pdf〉.

Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y., and Zhao, L. 2002. "THU TREC-2002 Web Track Experiments." In *Proceedings of the Eleventh Text Retrieval Conference (TREC-11).* 〈http:trec.nist.gov/pubs/trec11/papers/tsinghuau.web2.pdf〉.