태그결합을 이용한 불리언 검색에서 순위화된 검색결과를 제공하기 위한 시스템 설계 및 구현*

Design and Implementation of Tag Coupling-based Boolean Query Matching System for Ranked Search Result

김 용 (Yong Kim)**
주원균 (Won-Kyun Joo)***

초 록

불리언 검색만을 제공하는 정보시스템들은 순위화된 검색 결과를 제공하지 않아 이용자들이 많은 시간을 들여 수많은 결과를 일일이 확인해야하는 단점이 있다. 따라서 본 연구에서는 불리언 검색 모델의 단점을 극복하기 위한 방법으로써 불리언 검색에서 적용되고 있는 색인 가중치 정보 대신에 태그 간의 결합 관계 정보를 이용하여 순위화된 검색 결과를 제공하기 위한 시스템을 제안한다. 본 연구에서 제안하고 있는 방법은 일반적인 키워드 질의 대신에 문서를 질의로 사용하기 때문에 해당 문서에서 질의로 사용하는 핵심태그를 추출한다. 질의 생성 과정에서는 태그결합도에 따라 다양한 그룹의 불리언 질의를 생성하고, 매칭 과정에서는 해당 질의어 그룹 간에 차별성 정보와 태그 중요도 정보를 이용하여 순위화를 처리한다. 본 연구에서 제안하고 있는 방법의 유용성을 평가하기 위하여 선정된 연구정보와 관련된 동향분석정보를 추출하는 과정에 적용하여 실험을 수행하였다. 또한 제안된 방법에 대한 이용자 평가를 위하여 다수의 이용자들을 대상으로 약 1년간 서비스를 제공하였으며 그 결과 높은 이용자 만족도를 확보할 수 있다고 조사되었다.

ABSTRACT

Since IR systems which adopt only Boolean IR model can not provide ranked search result, users have to conduct time-consuming checking process for huge result sets one by one. This study proposes a method to provide search results ranked by using coupling information between tags instead of index weight information in Boolean IR model. Because document queries are used instead of general user queries in the proposed method, key tags used as queries in a relevant document are extracted. A variety of groups of Boolean queries based on tag couplings are created in the process of extracting queries. Ranked search result can be extracted through the process of matching conducted with differential information among the query groups and tag significance information. To prove the usability of the proposed method, the experiment was conducted to find research trend analysis information on selected research information. Aslo, the service based on the proposed methods was provided to get user feedback for a year. The result showed high user satisfaction.

키워드: 불리언 검색, 태그 기반 매칭, 질의어 분해 및 확장, 태그 결합도, 태그 기반 검색
boolean model, tag-based matching, query decomposition and extension, tag coupling, tag-based IR

---

* 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2011-327-H00017).
** 전북대학교 문헌정보학과 부교수, 비판적사고와 논술연구소 연구원(yk9118@jbnu.ac.kr) (제1저자)
*** 한국과학기술정보연구원(KISTI) NTIS센터 선임연구원(joo@kisti.re.kr) (교신저자)
■ 논문접수일자 : 2012년 11월 17일 ■ 최초심사일자 : 2012년 11월 23일 ■ 게재확정일자 : 2012년 12월 19일
■ 정보관리학회지, 29(4), 101-121, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.4.101〕

1. 서 론

정보기술과 인터넷의 발전은 폭발적인 정보 생산에 따른 정보유통 환경의 변화를 초래하였다. 특히, 정보검색에 있어서 폭발적으로 증가하는 대용량의 정보처리를 위한 다양한 검색기술이 발전되어져 왔다. 전통적인 정보검색방법에 있어서 문서들은 키워드 또는 색인어의 집합으로 식별되어 저장된다. 정보의 대량 생산과 유통 및 이용자의 다양한 검색요구의 변화에 따라 다양한 검색기술들이 등장하였다. 이러한 검색기술의 발전은 전형적인 불리언 검색(Boolean Search)과 벡터공간모델(Vector Space Model) 등의 대상문서의 색인어와 질의어와 같은 단어들 간의 유사도 즉, 단어매칭에 대한 평가를 통한 검색기법 중심으로 발전하였다(Salton & McGill, 1983). 이러한 검색기술의 발전과 함께, 2000년대 웹의 활성화에 따른 구글의 PageRank (Page et al., 1998)와 같은 가중치 정보를 기반으로 하는 유사검색, 링크관계 등의 각종 부가정보를 사용하는 검색방법이 보편화되고 있다(Baeza-Yates & Riberiro, 1999). 그러나 여전히 대부분의 정보검색 서비스를 제공하고 있는 정보검색 시스템들 특히, 대용량의 정보처리를 통한 검색서비스를 제공하기 위한 검색시스템들은 전통적인 문헌내의 단어와 질의어에 대한 매칭을 통한 검색에 중점을 두고 있으며 색인 가중치 정보를 사용하지 않는 순수 불리언 검색만을 제공한다. 이러한 환경적 요인에 따라 학술정보를 필요로 하는 연구자를 포함하여 도서관 분야의 검색 전문가들은 불리언 검색모델에 적용가능한 불리언 연산자를 사용한 매우 정교한 불리언 질의문을 효과적으로 생성하면서 정확률에 기반한 검색을 수행하여 왔다. 특히 불리언 모델은 가장 대중적인 색인어와 질의어에 대한 매칭검색 모델로써 대용량의 상업적 검색시스템에서 가장 많이 적용되었다(이수상, 이순영, 2009). 실질적으로 많은 분야에서 색인어와 질의어 처리의 상대적인 단순성 및 효과성으로 인하여 불리언 검색이 여전히 유용하게 사용되고 있으나, 전통적인 불리언 검색은 몇 가지의 단점을 지니고 있다(Salton et al., 1983). 첫째, 단순 단어기반의 검색을 통하여 색인어와 질의어의 매칭을 통하여 검색이 이루어지기 때문에 검색결과집합의 크기를 제어하기가 힘들다. 즉, 경우에 따라서 너무 많은 결과를 얻거나 또는 아무 결과도 얻을 수 없다. 둘째, 검색 결과가 순위화되지 않기 때문에 모든 검색 문서들은 동일한 우선순위를 갖는다. 셋째, 문서 또는 질의와 연결된 단어에 가중치를 부여할 수 없다. 문서 또는 질의에 포함된 단어는 모두 동일한 우선순위를 갖는다. 넷째, 직관적이지 않은 검색 결과를 보일 수 있다. 예를 들어, OR 질의문에 있어서 최소한 한 개의 단어만을 포함하면 되기 때문에 한 개의 단어를 포함한 것과 전체 단어를 포함한 결과가 동일한 중요도를 갖는다. 유사하게 AND 질의문에서는 전체 질의어를 구성하는 단어 중 단 한 개를 포함하지 못해 검색결과에서 누락된 문서와 질의 단어 중 아무것도 포함하지 않는 문서가 동일하게 평가됨으로써 검색결과의 재현율을 떨어뜨리는 결과를 가져올 수 있다. 불리언 검색기법과는 달리 유사도 기반의 검색기법은 불리언 검색기법에 비하여 상대적으로 다양한 측면에서 장점을 제공한다. 그러나 단어들 간의 유사도를 기반으로 검색을 수행하는 검색 기법들

은 불리언 검색에서 사용하는 구조적인 질의문을 처리할 수 없으며 검색모델에 따라서 매우 높은 연산량을 요구한다. 따라서 대용량의 정보를 색인하고 탐색하는데 있어서 높은 계산량으로 인하여 검색의 효율성 측면에서 문제점이 되고 있다. 물론 컴퓨팅 기술의 비약적인 발전에 따라 이러한 문제점은 부분적으로 해결되고 있으나 여전히 웹과 같은 대규모의 정보를 하는데 있어서는 문제점으로 지적되고 있다. 또한 질의어나 문서를 표현할 때 단순 불리언 모델에서는 AND와 OR 연산자와 가중치가 없는 용어를 사용하여 표시한다. 그러나 이러한 질의어는 가중치를 사용하지 못하므로 그 표현력이 약해 정확한 검색을 꾀할 수 없다. 이를 문제를 해결하기 위하여 다양한 연구들이 수행되었으며 특히 불리언 검색의 장점을 효과적으로 살리면서 단점으로 지적되고 있는 검색결과를 순위화 하기 위한 방법으로써 확장 불리언 모델이 제안되었다(Bookstein, 1980; Salton et al., 1983; Waller & Kraft, 1979). 확장 불리언 모델은 가중치 정보를 이용하면서 특정 요소에 대한 통제를 통하여 순수 불리언 모델과 유사도 검색 모델을 모두 포괄할 수 있다. 그러나 유사도 기반의 검색과 확장 불리언 모델은 모두 가중치를 기반으로 하고 있기 때문에 사전에 가중치 정보가 구성되어 있어야 한다. 단어에 대한 가중치 정보를 가지고 있지 않은 경우에 있어서 가중치 정보를 대체할만한 정보가 필수적으로 제공되어야 한다. 또 다른 방법으로 시소러스를 기반으로 하여 질의를 확장하는 방법에 관한 연구(Jing & Croft, 1994; Lee, 1999)가 있으나, 시소러스라는 방대한 외부 자원을 필요로 한다는 제약사항이 존재한다. 또한 시소러스를 사용하는 일은 많은 검색시간을 필요로 하며, 심지어는 검색 과정에서 원하는 정보를 찾지 못하는 경우도 발생한다. 또한 질의와 시소러스 용어 사이의 불일치 문제가 발생하여 검색의 재현율을 감소시키는 주된 원인으로 작용하기 때문에 이용자의 효과적인 검색을 저해하게 된다(Chen et al., 1995).

최근에는 웹의 폭발적인 성장에 따라 웹 검색에 대한 다양한 연구가 수행되고 있다. 일반적으로 웹 검색 결과의 품질을 높이기 위한 방법으로써 이용자가 입력한 질의 자체에 대한 분석으로써 질의 확장(query expansion) 기술이 대표적이라고 할 수 있으며 두 번째는 검색 알고리즘에 사용되는 문서 속성에 대한 접근 방법으로써 웹 문서는 작성자가 입력한 단어나 어구뿐 아니라 타 문서로의 링크관계, 최신성, 이용자 클릭 수 등 다양한 속성을 가지고 있는데 이에 대한 분석을 통해 검색 결과를 향상시키고자 하는 접근이다. 마지막으로 실제 수식화를 위한 방법에 대한 접근으로써 전통적인 TF-IDF와 같은 문서와 단어의 가중치(weight)를 이용한 방법, 질의와 문서의 벡터 비교 방법 등이 있으며 최근에는 각 속성의 가중치를 학습을 통해 결정하는 기계학습(machine learning)방법이 널리 사용되고 있다(임영석 등, 2011). 이러한 접근방법들 중에서 문서속성에 대한 접근으로써 태그를 활용하는 방법이 각광을 받고 있다.

이러한 웹 정보검색에 있어서 이용자 또는 특정목적에 의하여 기술되는 태그에 대한 중요성을 고려하여 본 연구의 목적은 태그정보를 활용하여 검색결과를 순위화할 수 있는 방법을 제안하였다. 특히, 전통적인 불리언 검색모델

의 최대 약점으로 지적되고 있는 검색결과의 순위화를 제공하지 못하는 단점을 극복하면서 검색의 단순성 및 효율성의 효과를 확보할 수 있는 방법으로써 웹 상에서 제공되고 있는 태그정보를 이용하여 검색결과를 순위화하여 제공하는 방법을 제안하고자 한다. 이를 위하여 태그의 유형으로써 하나의 태그로 구성된 단순태그와 여러 개의 태그로 구성된 복합태그에 대한 가중치에 대한 차별화를 통하여 검색결과에 대한 순위화를 수행하였다. 제안된 방법을 적용한 시스템이 설계 및 구현을 수행하였으며 제안된 시스템이 정보검색 과정에서 검색 효율성에 대한 개선 및 유용성에 대한 평가를 수행하였다.

2. 선행연구

웹 정보의 폭발적인 성장은 전통적인 정보검색방법 및 정보유통에 있어서 변화를 요구하고 있다. 정보검색에 있어서 중요한 기준이 되고 있는 효과성(effectiveness)과 효율성(efficiency)에 대한 중요성에 있어서 전통적으로 효과성에 대한 중요도가 상대적으로 높이 고려되어져 왔다. 그러나 웹 환경에서 대용량의 정보환경에서 정보처리를 위한 효율성에 대한 중요도가 상대적으로 높아져 가고 있는 상황이다. 또한 웹의 특성에 따라 전통적인 학술정보에서는 나타나고 있지 않은 링크에 대한 중요성이 매우 높아져 가고 있다. 이와 같은 링크정보를 활용한 대표적인 기법으로써 90년대 후반부터 Google의 검색엔진에 사용되는 알고리즘으로 널리 알려진 PageRank와 같은 문서의 전역 순위를 계산하는 방법들이 연구되고 있다. PageRank는 웹 문서들의 링크관계를 사용해서 문서들의 순서를 정렬하게 된다. 문서에 대한 평판에 따른 전역 순서를 웹 문서 작성자들의 평판에 의하여 작성된다. 또 다른 웹 문서의 중요한 특징으로써 이용자에 의하여 추가되는 태그를 검색에 활용하는 연구들이 시도되고 있다. 최근에는 웹 2.0과 함께 등장한 태그를 많은 분야에서 활용하고 있다(Carmagnola et al., 2007). 웹 2.0의 대표적 기술중의 하나인 태깅은 다수의 일반 이용자에 의해 만들어지는 매우 유연하고 역동적인 분류체계를 제공하지만 유연성과 역동성의 확보로 인하여 발생하는 근본적인 한계 또한 안고 있는 것이 사실이다(이정미, 2007). 예를 들어 검색분야에 있어서 단순한 태그기반의 검색은 동일한 자원에 기술되어져 있다 하더라도 그들의 연관성을 찾기가 어렵기 때문에 많은 한계점을 드러낸다(이성재, 조수선, 2011). 따라서 태그들 간의 연관성을 분석하기 위한 방법으로써 연관된 태그를 찾아서 그에 따른 분류를 클러스터로 표현하고 하나의 자원에 동시출현(co-occurrence)하는 태그들끼리의 포함관계를 이용하여 클러스터링하거나(Schmitz, 2006) 태그들을 동시출현하는 횟수를 가중치로 가지는 에지(edge)로 연결함으로써 태그공간을 그래프로 구성한 후에 결과를 클러스터링하는 방법 등이 있다(Begelman et al., 2006). 태그를 기반으로 하는 검색은 일반적인 웹 검색의 특별한 형태로써 웹 검색에서 질의와 문서의 관련성을 판단하기 위해 기본적으로 사용하는 문서의 단어나 어구 대신 소셜 태깅 시스템에서의 정보자원에 부여된 태그를 검색에 활용한다. 특히, 태그는 그 자체로도 문서를 잘 추상화하면서도

간결하기 때문에 이용자가 이해하기에 더욱 쉽다. 이러한 태그가 가장 많이 사용되는 곳이 바로 소셜 태깅 서비스이다. 소셜 태깅 서비스에서는 지금도 믿을 수 없는 속도로 많은 양의 태그가 생산되고 있다. Yanbe 등(2007)은 소셜 태깅 시스템에서의 태그 검색의 가능성을 제시하고 PageRank와의 결합 가능성에 대한 결과를 보여주었다. 소셜 태깅 시스템에서의 한 문서에 대한 이용자들의 총 태깅 횟수를 SBRank로 정의하고 이를 PageRank와 선형 결합하는 방법을 제안하였다. Bao 등(2007)은 웹 검색 결과의 향상을 위해 태그를 이용하기 위하여 기존의 SimRank(Jeh & Widom, 2002)를 활용하여 질의와 유사 태그간의 매칭을 시도하는 SocialSimRank와 소셜 태깅 시스템에서의 문서-이용자-태그 삼각관계를 상호강화관계로 보고문서의 전역 순위를 구하는 SocialPageRank를 제안하였다. 이용자에 의해서 추가되는 소셜 태그의 검색에 대한 활용 연구도 매우 활발하게 진행되고 있다. Heymann 등(2008)은 소셜태그를 기반의 웹 검색 질의어의 확장에 적용하여 유용성을 확인하였으며 Yi와 Chan(2009)은 잠재의미색인방법(Latent semantic indexing method)을 사용하여 소셜태그의 색인어로써의 가치를 평가하였다.

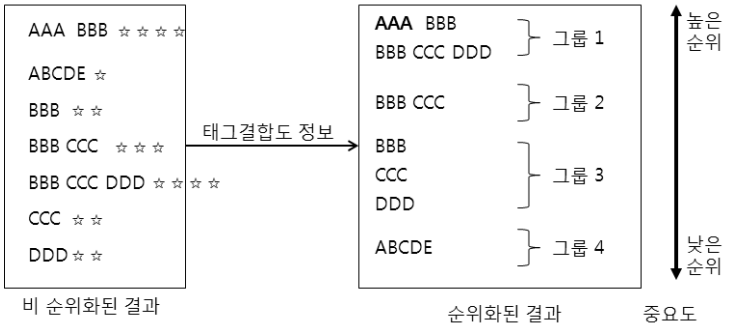
이와 같은 태그정보를 활용하여 검색을 수행하고자 하는 연구들과 함께, 태그정보를 활용하여 검색된 결과에 대한 랭킹의 효율을 높이는 방법들이 제안되고 있으나 상대적으로 태그정보를 활용한 실질적인 검색에 관련된 연구에 비하여 매우 적다. 국내에서 엄태영 등(2010)은 북마크 검색에 대해 개인화된 검색결과를 추천하기 위하여 사용자 태그를 기반으로 하여 딜리셔스가 제공하는 북마크들의 순위를 재순위화 하는 방법론을 제안하였다. 또한 태그유사도를 기반으로 한 태그 네트워크를 이용하여 사용자의 검색어에 의미적으로 유사한 다른 태그들도 순위에 반영될 수 있도록 하였다. 새로운 정보원으로써 고려되고 있는 블로그에 대한 검색 성능향상을 위하여 김은희와 정영미(2011)는 4,908개의 블로그페이지와 페이지에 트랙백으로 연결된 다른 블로그페이지의 URL을 수집하여 본문 용어와 이용자 태그를 검색자질로써 활용하면서 네트워크의 중심값을 반영함으로써 검색성능이 향상됨을 확인하였다. Choi(2010)는 웹 환경에서 이용자 협업에 의하여 생성된 소셜태깅이 웹 자원을 위한 디지털지식생성에 대한 활용성을 증명하기 위하여 이용자에 의한 소셜태깅의 색인 일관성과 전문가들에 의한 통제어 기반의 색인 일관성과의 비교를 위하여 VSM에 기반한 평가를 수행하였으며 이를 통하여 이용자에 의하여 생성된 색인어가 전문가에 의한 통제 색인어에 비하여 결코 뒤지지 않는 높은 품질임을 입증함으로써 검색에 도움이 될 수 있음을 확인하였다. Nakamoto 등(2008)은 유사도 검색의 개선 측면에서 태그정보를 활용하는 방법을 제안하고 있으나 태그정보와 함께 가중치 정보를 필수적으로 사용해야 한다는 제한점이 있다. 따라서 본 연구에서는 전통적인 불리언 검색방법의 빠른 처리 및 구조화된 검색식 등의 장점을 수용하면서 태그정보의 활용을 통하여 웹 상에서의 검색결과의 순위화를 위한 태그결합도에 기반한 검색결과 순위화 방법을 제안하였다.

3. 태그특성을 이용한 검색결과 순위화

3.1 태그특성을 이용한 불리언 검색결과 순위화 개요

전통적인 불리언 검색 기법은 검색결과에 대한 순위화를 제공하지 못한다. 따라서 본 연구에서는 불리언 검색모델을 기반으로 하면서 검색결과에 대한 순위화를 위하여 이용자 또는 특정목적을 위하여 기술되어진 태그정보를 활용하여 검색결과에 대한 순위화를 제공하는 방법을 제안하였다. 제안된 방법에 따라 검색결과의 순위화 과정에 있어서 해당 문서에 부여한 태그정보는 검색결과의 순위화를 위한 주요한 요소가 된다. 일반적으로 이용자에 의하여 기술된 태그를 살펴보면 한 문서에 부여되는 태그의 양이 다르다는 것을 알 수 있다. 한 개 또는 두 개의 태그가 부여된 문서가 있는 반면에 다섯 개 이상 혹은 열개가 넘는 태그가 부여된 문서가 있음을 쉽게 볼 수 있다. 한편 추가된 태그의 양적 측면과 함께, 질적 측면에 있어서 문서의 내용을 고려하지 않는 태그 검색의 특성으로 인하여 태그가 문서의 내용을 잘 추상화할수록 검색의 품질이 향상된다. 따라서 문서의 내용을 잘 추상화하는 태그들로 구성된 태그 집합이 필요하다는 점에서 볼 때 여러 개의 태그를 통해 문서를 표현하는 이용자들의 태깅 활동이 검색의 품질을 높인다고 볼 수 있다. 그러나 문서에 기술된 태그의 질을 평가하는데 있어서 다양한 기준이 적용될 수 있으며 판단하는 전문가에 따라 주관적 기준이 적용될 수 있기 때문에 태그의 질적 측면을 검색의 품질 및 결과의 순위화에 직접적으로 적용하는데 있어서 부분적으로 한계점이 존재한다. 따라서 보다 객관적이고 신뢰성있는 문서의 검색 또는 결과를 순위화하는 경우에 있어서 많은 양의 태그가 부여된 문서와 적은 양의 태그가 부여된 문서에 대한 태그에 대한 가중치를 통하여 차별화해야 한다. 본 연구에서는 이와 같은 태그의 특성과 함께, 부여된 태그에 대한 결합도를 기반으로 하는 태그정보를 이용하여 불리언 검색결과에 대해 순위화 방법을 제안한다. 태그정보 기반의 검색결과의 순위화 방법에 대한 기본 개요를 〈그림 1〉에서 도식으로 표현하고 있다. 〈그림 1〉에서 볼 수 있는 별표(☆)는 검색결과에 대한 우선순위를 표현하고 있으며 별표가 많을수록 순위가 높은 결과임을 의미한다. 일반적인 불리언 검색에서는 〈그림 1〉 왼쪽의 "비 순위화된 결과"에서 볼 수 있듯이 검색결과에 대한 우선순위를 식별할 수 없으나, 제안한 방법에서는 태그결합도를 이용하여 오른편과 같이 검색결과를 그룹화 함으로써 중요 문서에 대한 순위화를 확보할 수 있다. 즉, 일반적인 불리언 모델을 기반으로 추출된 검색결과에서는 문서의 우선순위를 구분하지 못하고 검색된 모든 문서는 동일한 우선순위를 가지고 있으나 본 연구에서 제안하고 있는 방법은 검색을 수행하는 과정에서 태그정보를 기반으로 하는 태그결합도를 기반으로 검색결과에 대한 순위화를 수행할 수 있다.

한편, 이용자 또는 특정 목적을 위하여 추가된 태그는 다양한 특성을 가지고 있으며 본 연구에서는 태그 결합도라고 정의한 특성을 이용하여 불리언 검색에 대한 순위화를 제공한다. 태그 결합도는 태그가 결합한 정도를 의미하는 것으로 몇 가지로 세부 속성으로 구분할 수 있



〈그림 1〉 태그정보를 이용한 순위화 방법

다. 본 연구에서는 기본적으로 속성 1과 2를 사용하여 불리언 검색결과에 대한 순위화를 수행하였다.

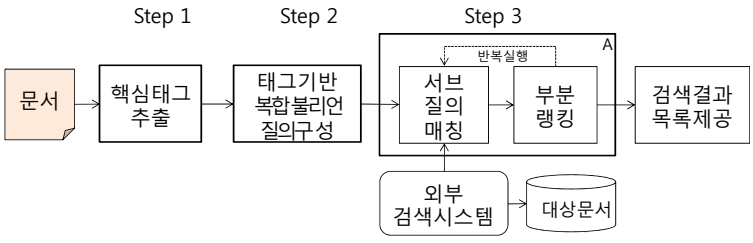
[속성 1] 단일 태그의 중요도: 일반적으로 하나의 태그에 대한 가중치정보를 나타내는 것으로 질의 태그 내에서 중요한 의미를 가진다.

[속성 2] 태그가 결합한 개수: 몇 개의 태그가 결합했는지를 나타내는 수치로서 일반적으로 많은 태그가 결합할수록 중요한 결과를 의미한다.

[속성 3] 두 태그 간의 강도: 두 태그가 동시에 출현하는 빈도를 의미한다.

이용자가 요청한 질의가 순위화된 검색결과로 제공되기까지의 전체 과정은 〈그림 2〉에 도식화하여 표현하였다. 이용자가 문서 형태의 질의를 제출하면 최종적으로 순위화된 검색결과 목록을 제공받는다. 이용자의 요청에 의하여 검색시스템이 최종적으로 검색결과를 순위화하여 보여주는 과정은 세 단계로 구성된다.

첫 번째 단계는 이용자가 제공한 질의문서에서 핵심태그를 추출하는 단계이다. 즉, 일반적인 검색에서의 색인어 추출과정이라고 할 수 있다. 이용자가 제공한 질의문서는 검색과정에 직접적으로 적용하기 어려운 형식이기 때문에 불리언 검색과정에 사용할 수 있는 단어를 추출하는 과정을 거쳐야 한다. 한편, 검색과정에서 주요한 요소로써 사용되는 단어는 키워드 또는



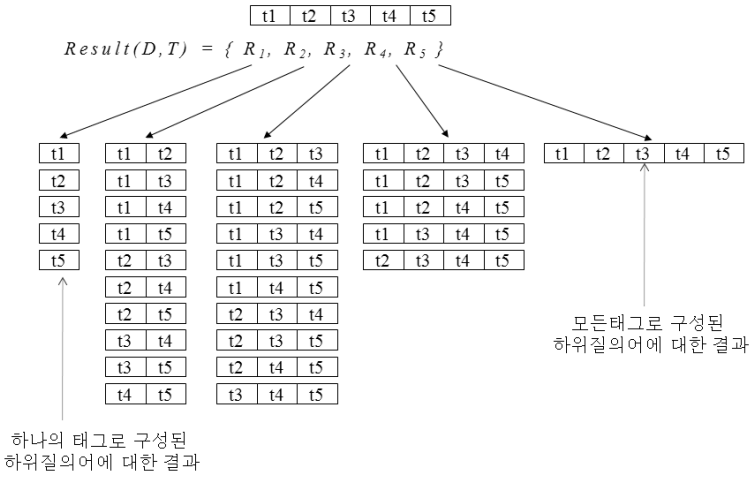
〈그림 2〉 태그특성을 이용한 검색 순위화

태그 형태로 표현될 수 있다. 그러나 키워드와 태그는 매우 차별화된 특성을 보여주고 있다. 즉, 키워드는 검색엔진이나 프로그램이 자동으로 추출하여 파악한 정보인 반면에 태그는 정보 작성자가 수동으로 직접 부여한 정보란 점에 차이가 있다. 일반적으로 검색단어 식별은 프로그램을 이용하여 자동으로 키워드를 추출하는 방법(Xu et al., 2006)을 사용할 수도 있으나, 관련 분야의 콘텐츠 전문가가 해당 문서를 가장 잘 표현할 수 있는 단어를 부여함으로써 문서와의 연관성을 높일 수 있다(Cattuto et al., 2008). 따라서 프로그램에 의하여 자동적으로 추출되는 키워드와는 다른 특성을 보여주는 단어 즉, 태그는 일반적으로 문서의 주제분야에 대한 분류뿐만 아니라 검색에도 효과적으로 사용할 수 있다.

본 연구에서는 두 가지 유형의 태그 추출 방법을 모두 사용할 수 있다. 다만, 특정 단어가 태그로 추출되었는지의 여부에 따라 시스템 성능에 일정 부분 영향을 미칠 수 있다. 일반적으로 프로그램에 의하여 자동으로 추출되는 방식에 비하여 내용 전문가에 의한 추출방법이 문서의 내용에 적합한 단어를 추출할 수 있다. 그러나 데이터의 양이 많은 경우에는 비용 및 시간적인 측면에 있어서 많은 시간과 비용이 요구된다. 따라서 데이터 규모가 커서 수동으로 태그를 추출하기 어려운 경우가 아니라면 전문가에 의한 태그 추출 방식을 사용하는 것이 검색의 효과성에 보다 적합하다고 할 수 있다. 본 연구에서 대상으로 하고 있는 실험집합은 상대적으로 규모가 작기 때문에 전문가에 의한 태그 추출 방식을 사용하고 있다. 실험에서는 기준이 되는 방법과 제안한 방법이 모두 동일한 태그를 이용한다. 따라서 태그 추출 시 자동 또는 전문가에 의한 수동으로 추출하는 방법의 차이점에도 불구하고 본 연구에서 제안하고 있는 방법에 대한 평가에 있어서 유의미한 영향을 미치지 않는다.

두 번째 단계는 태그기반의 차별화된 불리언 검색을 위한 질의어를 구성하는 단계이다. 따라서 본 단계에서는 〈그림 3〉에서와 같이 태그결합 관계를 이용하여 생성할 수 있는 모든 하위질의어를 생성하며 생성된 하위질의어를 실제 검색과정에서 활용함으로써 검색결과를 추출하게 된다. 특히, 생성된 하위질의어를 기반으로 연산을 수행하는데 있어서 본 연구에서는 질의문의 복잡성을 줄이고 검색결과의 정확도를 보장하기 위하여 질의어의 구성에 있어서는 AND 연산자만을 이용한다. 만일 일반적인 불리언 기법에서 사용되는 OR 연산자 등을 허용한다면 파생되는 하위질의어의 수와 연산처리에 따른 요구시간이 AND 연산자만을 사용하는 것에 비하여 몇 배로 증가할 수 있다. 〈그림 3〉에서 볼 수 있듯이 하나의 태그로 구성된 질의어도 검색결과에 포함되기 때문에 적합한 자료임에도 불구하고 검색되지 않는 상황은 발생하지 않는다.

세 번째 단계는 태그결합도에 따라 해당 부분질의어에 대한 순위화를 수행하는 단계이다. 특히, 부분질의어를 매칭하고 부분적인 순위결과를 통합하는 과정으로서 부분질의어의 수에 비례하여 반복적인 과정이 이루어진다. 한편, 부분질의어를 검색하는데 있어서 외부의 검색시스템을 이용하였다. 이러한 이유는 추출된 태그의 조합만큼의 부분질의어가 생성되기 때문에 질의어 처리에 많은 연산이 요구됨에 따라 검색



〈그림 3〉 태그 관계를 이용하여 하위질의어 생성방법

결과 추출에 따른 많은 시간이 요구되기 때문이다. 또한 검색에 따른 효율성(efficiency)의 확보를 통하여 이용자에게 빠른 검색 결과를 제공하기 위해서 부분질의어를 사전에 계산하여 저장한다. 해당 과정에서 부분질의어에 대한 이용통계정보를 기반으로 빈도가 높은 질의어를 중심으로 처리한다.

궁극적으로 두 번째 단계와 세 번째 단계에서 수행되는 과정이 본 연구에서 제안하고 있는 불리언 검색에 있어서 태그결합도를 기반으로 하는 '검색결과에 대한 순위화 알고리즘'으로써 보다 세부적으로는 '복합질의어의 집합 구성 및 순위화 방법'이라고 할 수 있다.

3.2 복합질의어 집합 구성 및 순위화

질의문서에서 추출한 태그정보를 이용하여 검색결과에 대한 순위화를 수행하기 위하여 먼저 복합 불리언 질의어집합을 구성하고 이를 기반으로 순위화를 실시하는 방법은 본 연구에서 제안하고 있는 주요 방법이라고 할 수 있다.

복합 불리언 질의어집합을 구성하는 방법에 대해서 알아보면 다음과 같다. 전체 검색 대상 문서(D)에 대한 전체 태그(T)에 대한 검색결과 집합 $Result\ (D,\ T)$는 수식 (1)과 같이 정의할 수 있다. $Result(D,\ T)$는 부분 검색결과 R_i의 집합으로 구성되는데, R_i는 i개의 태그로 구성된 부분 검색결과 집합을 의미한다. $Result\ (D,\ T)$는 총 $\sum_{i=1}^{n} C(N, i)$개의 부분집합으로 구성된다. 예를 들면, N이 5일 때 R_3은 이 중 3개의 태그만을 이용하였을 때의 부분 검색결과 집합을 의미하며, R_3는 조합 $C(5,\ 3)$에 따라 10개의 부분집합을 갖는다.

수식 (1)에서 T는 최대 N개의 태그로 구성

$$Result(D, T) = R_1(D, T), R_2(D, T), ..., R_i(D, T), ..., R_n(D, T) \qquad \text{수식 (1)}$$

$$sim(R_i)^p = 1 - \left[\frac{(1-wt_1)^p + (1-wt_2)^p + \cdots + (1-wt_i)^p + \cdots + (1-wt_n)^p}{(wt_1^p + wt_2^p + \cdots wt_i^p + \cdots + wt_n^p) + N} \right]^{1/p} \quad \text{수식 (2)}$$

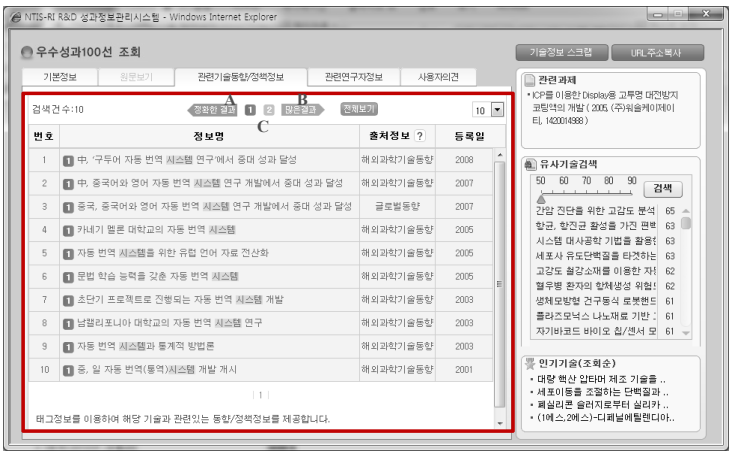
되고, t_1, t_2, … , $t_5 \in T$로 정의할 수 있다.

다음으로 복합 질의어 집합 내의 개별 질의어를 대상으로 태그매칭기법에 따라 부분유사도 값을 계산한다. 본 연구에서 제안하고 있는 태그매칭 알고리즘에는 위에서 정의한 태그의 두 가지 속성을 반영한다. 태그매칭 알고리즘은 확장 불리언모델에서 사용한 수식을 근간으로 하며, R_i 그룹에 대한 유사도 $sim(R_i)^p$는 수식 2와 같이 계산된다.

수식 (2)에서 wt_i는 t_i에 대한 단어 가중치를 의미하고, i와 wt_i는 각각 $1 \le i \le N, 0 \le wt_i \le 1$의 범위를 갖는다. 태그 단어 1개로 구성된 하위 그룹의 경우에 있어서 N값을 적용함으로써 해당 그룹에 대한 검색 유사도가 0부터 1사이의 정규화된 가중치를 유지할 수 있게 된다. 만약 수식에서 N을 생략한다면 정규화된 가중치 범위를 보장할 수 없게 된다. p값은 확장 불리언 모델에서 전형적인 불리언 검색과 유사도 검색 사이의 분별력을 제공하기 위해서 사용한다. 태그 단어 t_i의 가중치에 따른 검색결과 편차를 크게 할 목적이라면, p값을 적절히 조정하여 검색그룹 간의 변별력을 유지할 수 있도록 하여야 한다. 이러한 경우에 있어서 p값의 크기가 커질수록 검색 그룹간의 가중치 편차는 점차 커진다. 그러나 태그 가중치를 *1*로 고정한다면, 부분 검색결과는 p값에 의해 크게 좌우되지 않는다.

4. 태그 네비게이션을 이용한 검색 인터페이스 구현

본 연구에서 제안하는 태그정보를 이용한 검색 순위화 방법을 기반으로 구현된 서비스 시스템은 이용자의 편의성 및 이용자 만족도를 높이기 위하여 특별한 이용자 인터페이스를 제공하고 있다. 〈그림 4〉에서는 구현된 이용자 인터페이스를 보여주고 있다. 〈그림 4〉의 사각형 테두리 안의 인터페이스는 기본 기능으로써 태그결합도에 따라 검색결과를 순위화하여 제공하고 있으며 이를 통하여 이용자는 검색결과에 대한 조회를 할 수 있다. 전통적인 불리언 모델 방식에서는 이용자가 관심 대상으로 삼는 키워드에 대한 검색결과를 통하여 실질적으로 적합한 문서를 찾기 위해서는 검색결과 전체를 확인해야 하는 번거로움이 있었으나 본 연구에서 제공하고 있는 이용자 인터페이스를 통하여 이와 같은 번거로움을 제거할 수 있다. 즉, 제안한 방식에서는 이용자가 태그 결합도가 높은 순서에 따라 중요 문서를 먼저 확인하고, 태그결합도를 낮춰가면서 더 많은 문서를 열람할 수 있다. 〈그림 4〉에서 사각형 테두리 안의 상단은 태그결합도에 대한 네비게이션을 제공하고, 하단은 검색결과를 제공한다. 이용자는 네비게이션을 이용하여 가장 정확한 결과를 먼저 조회하고, 정확한 결과와 많은 결과를 적절히 조절하여 열람할 수 있다. 이용자 인터페이스를 통하여 제공되는 주요한 기능은 크게 세 가지로 구분할 수 있다.



〈그림 4〉 태그정보를 이용한 검색 순위화 인터페이스

먼저 상단에 있는 네비게이션에서 'A'의 기능은 검색결과에 있어서 태그결합도에 따라 높은 정확률을 보여주는 검색결과를 보여주는 기능을 수행하고 있다. 또 다른 네비게이션 기능을 제공하고 있는 'B'는 'A'에 비하여 재현율을 높인 결과를 확보하기 위한 기능으로써 좀 더 많은 검색결과로 이동할 수 있는 기능을 제공한다. 'C'는 태그결합도에 따라 검색결과에 대하여 그룹화된 결과를 표시하는 것으로써 그룹별 브라우징 기능을 제공한다. 예를 들어 〈그림 4〉에서 보여주는 것처럼 전체 검색결과는 2개의 그룹으로 구분해서 조회할 수 있으며 1이 보다 중요한 그룹을 나타낸다.

5. 실 험

5.1 실험 개요 및 환경

본 연구에서는 웹 상에 존재하는 정보에 대한 검색성능에 있어서 효율성 향상을 위하여 불리언 검색 모델을 기반으로 색인어와 질의어에 대한 가중치를 기반으로 검색결과에 대한 순위화를 제공하는 전통적인 확장불리언 검색방법이 아닌 이용자 또는 특정 목적을 위하여 추가된 태그정보를 활용하여 검색결과에 대한 순위화를 제공하는 방법을 제안하고 있다. 따라서 웹 상의 태그정보는 검색결과의 순위화를 위한 주요한 요소가 되며 이러한 이용자 태그 또는 특정 목적을 위하여 추가된 태그는 단일 태그 와 복합태그로 구분할 수 있다. 단일 태그는 하나의 태그에 대한 가중치정보를 나타내는 것으로 질의 태그 내에서 중요한 의미를 가지는 특성을 내포하고 있으며 여러 개의 태그가 결합된 복합태그는 몇 개의 태그가 결합된 것으로써 일반적으로 많은 태그가 결합할수록 중요한 결과를 의미한다.

이와 같이 질의문서에서 추출한 태그정보를 이용하여 검색결과에 대한 순위화를 수행하기 위하여 먼저 복합 불리언 질의어집합을 구성하고 이를 기반으로 순위화를 실시하였다. 본 연구에서 제안하고 있는 방법에 대한 유용성 평가를

위한 실험을 수행하기 위하여 사용된 데이터 컬렉션은 이용자가 질의어로써 제공하는 검색 질의용 컬렉션과 검색 대상이 되는 불리언 검색 시스템에서 사용하는 검색 대상 컬렉션의 두 가지로 구분한다. 검색 질의 컬렉션은 NTIS[1]에서 자체적으로 매년 구축하고 있는 우수성과 100선 정보 중에서 2011년도에 선정된 100건의 문서를 대상으로 하였다. 검색 대상 컬렉션은 NDSL을 통하여 현재 서비스되고 있는 동향분석정보를 대상으로 하였다.

세부적인 실험방법으로써 일반적인 정보검색 시스템의 성능평가를 수행하는 방식과 같이 실험데이터와 평가데이터를 구분하여 검색 질의 컬렉션을 대상으로 검색 대상 컬렉션을 매칭하는 방식을 사용하였다. 본 연구에서는 우수성과 100선 내의 각각의 기술정보와 관련된 NDSL 동향분석정보를 검색하는 방식으로 진행하였다. 실험을 위하여 먼저 검색 질의용 컬렉션인 우수성과 100선 데이터를 대상으로 태그를 추출하였다. 추출된 검색 질의용 컬렉션 내의 태그에 대한 특성은 〈표 1〉과 〈표 2〉에서 보여주고 있다. 과학기술 분야 콘텐츠 전문가는 검색 질의 컬렉션을 대상으로 태깅 작업을 거쳐 태그 정보를 구축하였다. 컬렉션 내의 문서는 각각 2~6개의 태그로 구성되고 이 중 대다수의 문서는 3~5개의 태그로 구성된다. 전체적으로 유일한 태그 개수는 333개이고, 280개의 태그는 단 1개의 문서에 출현한다. 이러한 태그 구성은 문서간의 변별력이 매우 높다는 것을 의미한다. 특히, '친환경'이라는 태그는 6개의 기술문서에 출현하여 가장 높은 출현빈도를 보이고 있으며 이 밖에도 2개 이상의 문서에서 출현하여 상대적으로 고빈도를 보이는 태그 단어는 약 53개로 확인된다.

다음으로 검색 대상 컬렉션에 포함된 데이터 특성에 대해 알아보면 다음과 같다. 검색 대상 컬렉션인 NDSL 동향분석정보[2]는 세계 주요국의 과학기술 및 과학기술정책분야의 최신 동향분석정보, 산학연 과학기술분야의 리더들이 자신들의 관점에서 기술한 전문적이고 차별화된 동향정보를 제공하고 있으며 구체적인 구성 및 특성을 〈표 3〉에서 보여주고 있다.

실질적인 실험은 순위화 수식에서 적용된 파라미터에 대한 결정과 제안한 알고리즘에 대한 성능평가로 구분하여 수행하였다.

〈표 1〉 개별 문서 내의 태그정보 수

태그 포함 개수	1	2	3	4	5	6	7	8	소계
해당 문서 수	-	5	26	32	26	11	-	-	100

〈표 2〉 태그 출현 빈도 수

태그 출현 빈도	1	2	3	4	5	6	7	8	소계
관련 태그 수	280	36	10	6	-	1	-	-	333

1) National Science & Technology Information Service(NTIS), http://www.ntis.go.kr
2) NDSL Trend Service, http://radar.ndsl.kr

〈표 3〉 검색 대상 컬렉션 개요

	해외과학기술동향	과학기술정책동향	정보서비스 글로벌 동향
수록건수	147,325	17,840	1,306
수록기간	1999~	2005~	2009~
갱신주기	매일	매일	매일
주제분야	과학기술전분야	과학기술정책	정보서비스

5.2 순위화 파라미터에 대한 결정

본 과정에서는 수식(2)에서 사용한 *p*값과 태그 가중치 *wt*의 임계치를 결정하는 방법으로써 순위화에 직접적인 영향을 주고 있는 파라미터에 대한 결정에 대한 방법을 설명하고 있다. 먼저 *p*값의 결정 과정을 시뮬레이션을 통해 도출한다. 예를 들어, *N=4*일 때, *p*값에 변화를 주고 그 때의 문서 유사도 값의 변화를 〈표 4〉에서 설명하고 있다. 이 때 각 태그에 대한 가중치는 각각 0.5, 09, 0.3, 0.8로 부여한다. 〈표 4〉에서 볼 수 있듯이 *p*값의 변화에 따른 부분 유사도 값의 변화를 확인한 결과, *p*값이 커질수록 그룹간의 변별력이 확연히 커지는 것을 알 수 있다. 따라서 결합한 태그의 개수가 많을수록 좋은 결과라는 접근 방법을 고려한다면 *p*값을 높게 가져가야 한다고 가정할 수 있다.

한편, 태그 가중치 *wt*는 그룹 간의 변별력보다는 태그 단어에 우선순위를 두기 위한 목적으로 사용된다. 〈표 4〉에서 밑줄로 표시한 부분을 보면, *p=1*일 때 태그를 3개 사용한 결과(0.571)가 태그를 2개 사용한 결과(0.596)보다 낮은 유사도를 보인다. 이것은 *t2*가 *t1*이나 *t3*보다 우선하는 단어임을 직접적으로 보여준다. 사용할 수 있는 다양한 태그 가중치 조합을 이용하여 실험한 결과, *p*값이 3이상일 경우에 태그 가중치 *wt*의 사용이 유의하지 않다는 것으로 확인되었다.

태그 가중치(*wt*)는 1로 고정하고, *p*값을 변경하였을 때의 부분 유사도의 변화에 대한 추이를 〈표 5〉에 보여주고 있다. 이와 같은 경우에는 *p*값에 무관하게 그룹간의 변별력이 유지되면서 *p*값의 변화는 단지 시각적으로 인지할 수 있는 유사도 수치의 범위에만 영향을 미친다.

결론적으로 태그 가중치(*wt*)는 콘텐츠 전문가가 태그를 선정하는데 있어서 태그의 중요도를 부여하는 용도로 사용할 수 있다. 따라서 중요한 태그를 식별하여 해당 검색결과를 상위에 제시하는 효과가 있다. 실험 결과, *p*값은 1 또는 2로 설정하여 사용하는 것이 적절하다는 결론을 도출하였다. 일예로써, "극초단 레이저 기반의 초정밀 절대거리측정기술"이라는 문서에서는 "극초단, 레이저, 거리, 측정"이라는 태그를 추출할 수 있으며 해당 네 개의 태그 중에서 측정이라는 태그에 높은 가중치를 부여하고, 레이저에 낮은 가중치를 부여할 경우에 좋은 성능을 보인다. 본 실험에서 적용된 테스트 컬렉션의 경우에 있어서 레이저가 포함된 문서들은 다른 주제와 관련된 경향이 높기 때문에 이러한 현상이 발생함을 알 수 있다. 따라서 태그 가중치(*wt*)는 대상 컬렉션의 특성을 고려하여 휴리스틱한 방법으로 사용하는 것이 적합할 것으로 판단된다.

〈표 4〉 부분 유사도를 계산하는 예시 1 (N=4, wt=0~1, p=1~100)

Tag Depth	Tag Weight				Sub-Similarity				
	t1	t2	t3	t4	p=1	p=2	p=3	p=10	p=100
4	0.5	0.9	0.3	0.8	0.769	0.631	0.554	0.395	0.310
3	0.5	0.9	0.3		0.596	0.417	0.330	0.134	0.014
	0.5	0.9		0.8	0.710	0.522	0.404	0.139	0.014
		0.9	0.3	0.8	0.667	0.473	0.365	0.136	0.014
	0.5		0.3	0.8	**0.571**	0.402	0.319	0.129	0.014
2	0.5	0.9			0.519	0.332	0.241	0.075	0.007
	0.5		0.3		0.333	0.205	0.159	0.066	0.007
	0.5			0.8	0.491	0.316	0.228	0.069	0.007
		0.9	0.3		0.462	0.286	0.210	0.073	0.007
		0.9		0.8	**0.596**	0.387	0.274	0.077	0.007
			0.3	0.8	0.431	0.269	0.197	0.068	0.007
1	0.5				0.222	0.126	0.088	0.028	0.003
		0.9			**0.367**	**0.209**	0.141	0.036	0.003
			0.3		0.140	0.076	0.060	0.027	0.003
				0.8	**0.333**	0.191	0.126	0.031	0.003

〈표 5〉 부분 유사도를 계산하는 방법 2 (N=4, wt=1, p=1~100)

Tag Depth	Tag Weight				Sub-Similarity				
					p=1	p=2	p=3	p=10	p=100
4	1	1	1	1	1	1	1	1	1
3	1	1	1		0.857	0.622	0.477	0.177	0.019
2	1	1			0.667	0.423	0.307	0.104	0.011
1	1				0.400	0.225	0.157	0.050	0.005

5.3 제안 방법에 대한 평가

일반적으로 순위화하지 않은 일반적인 검색시스템에 대한 평가에 있어서 사용되는 대표적인 기준으로써 재현율(recall)과 정확도(precision)라는 측정방법을 이용한다(Kent, Berry, Luehrs Jr., & Perry, 1955). 본 연구에서는 검색 대상을 웹 상에 존재하는 정보를 대상으로 하고 있으며 순위화된 검색 결과를 제공하는 시스템에 대한 평가를 실시하고자 하였다. 웹 정보의 대표적인 특성으로써 대용량의 정보환경에서 유사한 정보의 발견과 관련된 중요한 고려사항으로써 정보이용자의 검색시간을 줄여주는 데에 주안점을 두고 있다. 실험에서는 순위화된 검색 결과의 평가에 사용할 수 있으며, TREC(Text Retrieval Conference)[3]에서 웹 정보검색시스

3) TREC(Text Retrieval Conference), http://trec.nist.gov/

템의 평가를 위해서 널리 사용하는 k-문서 정확도(precision at k)라는 평가방법을 이용한다(Moffat & Zobel, 2008). k-문서 정확도는 적합 문헌의 총 개수가 정확도에 강한 영향을 미친다는 단점을 지니고 있지만, 전체 데이터에 대한 적합성 정보를 모두 구축할 필요가 없다는 장점을 지니고 있어 본 실험에 유용하게 사용할 수 있다. 제안한 방법의 평가를 위한 도구로는 TREC에서 사용하는 'trec_eval'이라는 프로그램을 이용하였다.

세부평가결과는 기존의 일반적인 불리언 모델과 제안한 방법의 비교를 통해 제시하였다. 비교 대상으로써 확장 불리언 모델과 같은 부분 랭킹 기반의 알고리즘이 아닌 전통적인 불리언 모델을 고려대상으로 하고 있는 이유는 다음과 같다. 본 연구에서 해결하려고 하는 대상이 일반적인 불리언 검색만을 제공하는 외부시스템을 이용하여 순위화된 검색결과를 제공하자는 의도에서 출발했다. 즉, 불리언 검색으로 고정된 외부의 검색엔진 모델을 직접 통제할 수 없는 상황을 고려하였다. 색인 정보를 통제할 수 있는 상황이라면 확장 불리언 모델보다 통상적으로 높은 효율을 보이는 유사도 검색 방법을 이용하는 것이 오히려 적합할 수 있다. 따라서 본 연구의 후속연구로써 색인정보를 통제하여 확장불리언 모델과 제안한 방법 간의 비교를 수행할 예정이다.

성능평가를 위해 사용한 세부 질의 규칙은 다음과 같다. 예를 들어 '나노', '기술', '연구'라는 3개의 태그가 주어졌을 때, 일반적인 불리언 검색에서는 "나노 or 기술 or 연구"라는 질의로 검색을 수행한다. 반면 제안한 방법에서는 앞서 설명한 태그결합도를 이용하여 3개의 태그를 조합하여 검색을 수행한다. "많은 정보를 사용하기 때문에 당연히 성능이 좋아질 것이다"라고 가정할 수 있으나 색인정보를 통제할 수 없는 확장 불리언모델이나 유사도 검색 모델을 사용할 수 없는 상황에서도 그와 비슷한 성능을 보일 수 있는 방법을 제안했다는 데어서 제안하고 있는 방법의 의미를 찾을 수 있다.

〈표 6〉에서는 제안한 방법에 대한 실험평가결과를 보여주고 있다. 이 때, 실험의 용이성을 고려하여 태그 가중치는 wt=1, p=2로 고정하였다.

평가결과, 제안한 방법의 검색 성능이 전 구간에서 우수한 것으로 측정되었으며 P@10에서는 약 5배 가량 우수한 성능을 보인다. 일반적인 불리언 검색에서 정확도가 낮은 이유는 불리언 자체가 순위화된 방식이 아니라 적합한 문서가 검색되었더라도 전체 검색결과에 포함되는 문서에 대한 순위화가 이루어지지 않고 있기 때문에 적합문서가 뒤쪽에 위치했기 때문인 것으로 고려된다.

다음으로, 질의문서를 중심으로 검색된 결과집합을 분석하였다. 태그를 이용하여 검색할 경우, 하나의 질의문서는 평균적으로 920여 개의

〈표 6〉 정확도를 이용한 검색 성능 평가 결과표

구분	P@5	P@10	P@15	P@20	P@30	P@100
일반적인 불리언 검색	0.1	0.05	0.035	0.025	0.017	0.005
제안한 방법	0.3	0.25	0.2	0.2	0.167	0.08

〈표 7〉 부분 질의에서 이용하는 검색 결과 개수에 따른 성능 평가 결과표

검색결과 개수	P@5	P@10	P@15	P@20	P@30	P@100
10	0.3	0.25	0.167	0.175	0.117	0.035
20	0.3	0.25	0.2	0.2	0.167	0.05
30	0.3	0.25	0.2	0.2	0.167	0.065
40	0.3	0.25	0.2	0.2	0.167	0.07
50	0.3	0.25	0.2	0.2	0.167	0.075
60	0.3	0.25	0.2	0.2	0.167	0.08
100	0.3	0.25	0.2	0.2	0.167	0.08

검색결과를 제공한다. 이용자가 검색 결과를 조회하는데 있어서, 태그 결합도가 높은 것(4~6)을 중심으로 조회할 때에는 문제가 없겠지만, 태그 결합도가 낮은(2~3) 결과를 조회할 경우에는 상대적으로 많은 열람 건수로 인해 불편을 초래할 수 있다. 이처럼 태그결합도가 낮을 경우에는 각 그룹 내의 부분질의에서 사용할 검색결과의 개수에 제한을 둘 필요가 있다. 부분질의에서 이용할 검색결과의 개수를 다양하게 할 경우의 검색 성능은 〈표 7〉에서 보여주고 있다.

결과적으로 부분 검색결과 개수를 너무 적게 잡으면 후반부의 검색 성능이 좋지 않은 것으로 확인되며 사용할 부분 검색결과의 개수를 30개 수준으로 제한할 경우 좋은 성능을 보이면서 조회해야 할 문서의 개수는 줄어드는 것으로 확인되었다.

본 연구에서 제안하고 있는 방법을 적용한 검색시스템을 구현하여 서비스를 제공하였으며 이를 통하여 이용자 만족도를 간접적으로 조사하였다. 실질적으로 NTIS 내의 R&D성과정보서비스[4]에서 해당 기법을 구현하여 특정 우수 유망기술정보와 관련 있는 동향분석정보의 제공을 위한 서비스에 적용하였으며 약 1년간 해당 서비스를 이용한 경험이 있는 이용자를 대상으로 이용자 만족도에 대한 조사를 수행하였다. NTIS 내에서는 다양한 정보서비스를 제공하고 있으며 본 연구에서 제안하고 있는 방법을 적용하고 있는 R&D성과정보서비스에 대한 이용자 만족도가 다른 서비스들에 비하여 상대적으로 높은 서비스 이용률을 보였다.

6. 결 론

본 연구에서는 색인어에 대한 가중치정보를 제공하지 않는 불리언 검색에서 태그결합도라는 정보를 이용하여 순위화된 검색결과를 제공할 수 있는 방법을 제안하였다. 웹과 같은 대용량의 정보환경에서 검색의 효율성 확보는 매우 중요한 고려사항이 되고 있다. 웹 상의 정보를 검색하는데 있어서 일반적인 문헌검색과는 다른 특징을 포함하고 있다. 웹을 통하여 제공되는 정보자원의 대표적인 특징으로써 링크정보와 함께, 이용자의 직접적인 참여에 의한 태그

4) National R&D Outcome Service, http://roots.ntis.go.kr

정보의 제공이라고 할 수 있다. 따라서 태그정보는 웹 정보에 대한 검색에 있어서 중요한 요소로써 작용하고 있으며 본 연구에서는 이와 같은 태그정보를 활용하여 검색의 효율성을 확보할 수 있는 방법을 제안하였다. 특히, 불리언 모델을 기반으로 하는 검색시스템에서 검색결과의 순위화를 제공하지 못하는 문제점을 해결하면서 불리언 모델에서의 연산의 효율성을 확보할 수 있는 방법으로써 태그정보를 활용하여 검색의 효율성을 확보하면서 검색결과에 대한 순위화를 제공함으로써 이용자의 검색결과에 대한 브라우징의 복잡성을 해결하였다. 본 연구에서 중요한 요소로써 고려할 수 있는 이용자 태그 또는 특정 목적을 위하여 추가된 태그는 단일 태그와 복합태그로 구분할 수 있다. 단일 태그는 하나의 태그에 대한 가중치정보를 나타내는 것으로 질의 태그 내에서 중요한 의미를 가지는 특성을 내포하고 있으며 여러 개의 태그가 결합된 복합태그는 몇 개의 태그가 결합된 것으로써 일반적으로 많은 태그가 결합할수록 중요한 결과를 의미한다.

이와 같이 질의문서에서 추출한 태그정보를 이용하여 검색결과에 대한 순위화를 수행하기 위하여 먼저 복합 불리언 질의어집합을 구성하고 이를 기반으로 검색결과에 대한 순위화를 수행하는 방법을 설계하고 이를 구현하였다. 제안된 방법에 대한 유용성평가와 함께, 전통적인 불리언 검색과의 비교를 수행하였다. 이를 위하여 먼저 이용자가 제시한 질의문서를 전문가가 부여한 태그로 표현하고, 해당 태그정보를 이용하여 차별화된 복합 불리언 질의를 생성한다. 태그의 특성 중 태그결합도에 따라 복합 불리언 질의를 매칭하는 알고리즘을 이용하여 검색결과에 대한 그룹순위를 부여한다. 제안된 방법을 통하여 최적의 유사한 문서에 대한 우선 열람과 최소한의 결과 열람을 보장함으로써 이용자의 검색 만족도를 향상시킬 수 있었다.

본 연구는 불리언 검색에서 태그정보만을 이용하여 순위화된 검색결과를 제공하였다는 점에 의미를 부여할 수 있다. 또한, 확장 불리언 검색, 유사도 검색 또는 시소러스를 확장한 불리언 검색과 달리 통제된 색인정보를 이용하지 않고 보다 적은 정보로 간편하게 순위정보를 제공할 수 있다는 점에서 효과적인 방법으로 고려될 수 있다. 실험을 통하여 검색성능을 입증하였으며 검색결과를 효과적으로 제공하기 위한 특별한 이용자 인터페이스를 구현하였다. 실제로 NTIS 내의 R&D성과정보서비스에서 본 연구에서 제안하고 있는 방법을 적용하여 시스템을 구현하여 특정 우수유망기술정보와 관련 있는 동향분석정보의 제공을 위한 서비스에 적용하였다.

향후 연구로써 두 태그 간의 강도를 이용하는 방법에 대해 연구를 수행할 예정이며 태그의 수가 많아질 경우에 있어서 대량의 부분질의가 생성되는데 이를 효과적으로 축소하는 방법에 대한 연구도 필요하다고 하겠다. 또한, 색인정보를 통제할 수 있는 환경을 구성하여 확장 불리언 모델 또는 기타 유사도 검색과의 성능도 비교할 예정이다.

참 고 문 헌

김은희, 정영미 (2010). 사용자 태그와 중심성 지수를 이용한 블로그 검색 성능 향상에 관한 연구. 정보관리학회지, 27(1), 61-77. http://dx.doi.org/10.3743/KOSIM.2010.27.1.061

이성재, 조수선 (2011). 위키피디아 기반의 의미 연관성을 이용한 태깅된 웹 이미지의 검색순위 조정. 멀티미디어학회논문지, 14(11), 1491-1499.

이수상, 이순영 (2009). 차세대 검색서비스의 속성에 관한 연구. 정보관리학회지, 26(4), 93-112. http://dx.doi.org.10.3743/KOSIM.2009.26.4.093

이정미 (2007). 폭소노미의 개념적 접근과 웹 정보서비스에의 적용. 한국비블리아학회지, 8(2), 141-159.

엄태영, 김우주, 박상언 (2010). 태그 네트워크를 이용한 개인화 북마크 추천시스템. 한국전자거래학회지, 15(4), 181-195.

임영석, 이강표, 김현우, 안재민, 김형주 (2011). 사용자 활동 점수에 기반한 태그 검색 개선. 정보과학회논문지: 컴퓨팅의 실제 및 레터, 17(3), 150-158.

Baeza-Yates, R., & Riberiro-Neto, B. (1999). Modern information retrieval. New York: Addison-Wesley Longman.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing web search using social annotations. Proceedings of the 16th international conference on World Wide Web (WWW '07), 501-510. http://dx.doi.org/10.1145/1242572.1242640

Begelman, G, Keller, P., & Smadja, F. (2005.2.22). Automated tag clustering: Improving search and exploration in the tag space. Paper presented at the Collaborative Web Tagging Workshop at WWW 06, Edinburgh, UK. Retrieved from http://www.semanticmetadata.net/hosted/taggingws-www2006-files/20.pdf

Bookstein, A. (1980). Fussy requests: An approach to weighted boolean searches. Journal of the American Society for Information Science, 31(4), 275-279.

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., & Torre, I. (2007). Towards a tag-based user model: How can user model benefit from tags? Lecture Notes in Computer Science, 4511, 445-449. http://dx.doi.org/10.1007/978-3-540-73078-1_62

Cattuto, C., Benz, D., Hotho, A., & Stummen, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. Lecture Notes in Computer Science, 5318, 615-631. http://dx.doi.org/10.1007/978-3-540-88564-1_39

Chen, H., Tim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. Journal of the American Society for Information Science, 46(3), 175-193.

Choi, Yun-Seon (2010). Implications of social tagging for digital libraries. Journal of the Korean Society for Information Management, 27(2), 225-239. http://dx.doi.org/10.3743/KOSIM.2010.27.2.225

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search? Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08), 195-206. http://dx.doi.org/10.1145/1341531.1341558

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), 538-543. http://dx.doi.org/10.1145/775047.775126

Jing, Y., & Croft, W. B. (1994). An association thesaurus for information retrieval. Proceedings of the RIAO '94 Conference, 146-160.

Kent, A., Berry, M. M., Luehrs Jr., F. U., & Perry, J. W. (1955). Machine literature searching Ⅷ: Operational criteria for designing information retrieval systems. American Documentation, 6(2), 93-101. http://dx.doi.org/10.1002/asi.5090060209

Lee, Bog-Gi (1999). A new document ranking algorithm in boolean retrieval system. Journal of Kyungwon College, 21, 159-165.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems, 27(1), 1-26. http://dx.doi.org/10.1145/1416950.1416952

Nakamoto, R. Y., Nakajima, S., Miyazaki, J., Uemura, S., Kato, H., & Inagaki, Y. (2008). Reasonable tag-based collaborative filtering for social tagging systems. Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW '08), 11-18. http://dx.doi.org/10.1145/1458527.1458533

National Discovery for Science Leaders (NDSL). Retrieved from http://www.ndsl.kr

National R&D Outcome Service. Retrieved from http://roots.ntis.go.kr

National Science and Technology Information Service (NTIS). Retrieved from http://www.ntis.go.kr

NDSL Trend Service. Retrieved from http://radar.ndsl.kr

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Proceedings of the 7th International World Wide Web Conference Brisbane, 161-172.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw Hill.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended boolean information retrieval. Communications of the ACM, 36(11), 1022-1036.

Schmitz, P. (2006.5.22). Inducing ontology from flickr tags. Paper presented at the Collaborative Web Tagging Workshop at WWW 06, Edinburgh, UK. Retrieved from http://www.semanticmetadata.net/hosted/taggingws-www2006-files/22.pdf

TREC (Text Retrieval Conference). Retrieved from http://trec.nist.gov/

Waller, W. G., & Kraft, D. H. (1979). A mathematical model for a weighted boolean retrieval system. Information Processing and Management, 15(5), 235-245.

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006.5.22). Towards the semantic web: collaborative tag suggestions. Paper presented at the Collaborative Web Tagging Workshop at WWW 06, Edinburgh, UK. Retrieved from http://www.semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2007). Can social bookmarking enhance search in the web? Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07), 107-116. http://dx.doi.org/10.1145/1255175.1255198

Yi, K., & Chan, L. M. (2009). Linking folksonomy to library of congress subject headings: An exploratory study. Journal of Documentation, 65(6), 872-900. http://dx.doi.org/10.1108/00220410910998906

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Eom, Tae Young, Kim, Wooju, & Park, Sangun (2010). Personalized bookmark recommendation system using tag network. Journal of Society for e-Business Studies, 15(4), 181-195.

Kim, Eun-Hee, & Chung, Young-Mee (2010). Enhancing the performance of blog retrieval by user tagging and social network analysis. Journal of the Korean Society for Information Management, 27(1), 61-77. http://dx.doi.org/10.3743/KOSIM.2010.27.1.061.

Lee, Jeong-Mee (2007). A conceptual access to the folksonomy and its application on the web information services. Journal of Korean Biblia Society for Library and Information Science, 18(2), 141-159.

Lee, Seongjae, & Cho, Soosun (2011). Tagged web image retrieval re-ranking with Wikipedia-based semantic relatedness. Journal of Korea Multimedia Society, 14(11), 1491-1499.

Lee, Soo-Sang, & Lee, Soon-Young (2009). A study on the features of the next generation

search services. Journal of the Korean Society for Information Management, 26(4), 93-112. http://dx.doi.org.10.3743/KOSIM.2009.26.4.093.

Lim, Young-Seok, Lee, Kang-Pyo, Kim, Hyun-Woo, Ahn, Jae-Min, & Kim, Hyoung-Joo (2011). Improving tag search based on user activeness scores. Journal of KIISE: Computing Practices and Letters, 17(3), 150-158.