검색 질의 확장을 위한 인기도 기반 단어 가중치 측정

(A Term Weight Mensuration based on Popularity for Search Query Expansion)

이 정 훈 † 전 서 현 ††
(Jung-Hun Lee) (Suh-Hyun Cheon)

요 약 인터넷의 활용이 보편화 됨에 따라 사람들이 많은 정보를 웹을 통해 접할 수 있게 되었다. 정보의 양이 급격히 늘어나면서 검색 엔진은 사용자가 필요로 하지 않는 정보까지 보여주는 검색 성능의 한계를 가져왔다. 따라서 사용자는 원하는 정보를 검색하기 위해 과거보다 더 많은 시간과 노력이 필요하게 되었다. 이 연구에서는 질의 확장을 이용하여 사용자가 필요로 하는 정확한 정보를 신속하게 찾아서 제공할 수 있는 방법을 제안한다. 제안된 단어 가중치 평가방법은 검색 주제의 변동 없이 하나의 검색 주제를 검색할 경우 TF-IDF 또는 단순 인기도 측정법 보다 우수한 성능을 보인다. 또한 검색 중 주제를 변경하였을 때에도 검색 주제 변경 전과 유사한 성능으로 기존의 측정법 보다 빠르게 새로운 주제와 관련된 단어를 추출하고 정확한 가중치를 측정한다.

키워드 : 개인화 검색, 질의 확장, 인기도 기반 키워드 랭킹, 클러스터링, TF-IDF 검색 질의 확장을 위한 단어 추출 및 가중치 측정법

Abstract With the use of the Internet pervasive in everyday life, people are now able to retrieve a lot of information through the web. However, exponential growth in the quantity of information on the web has brought limits to online search engines in their search performance by showing piles and piles of unwanted information. With so much unwanted information, web users nowadays need more time and efforts than in the past to search for needed information. This paper suggests a method of using query expansion in order to quickly bring wanted information to web users. Popularity based Term Weight Mensuration better performance than the TF-IDF and Simple Popularity Term Weight Mensuration to experiments without changes of search subject. When a subject changed during search, Popularity based Term Weight Mensuration's preformance change is smaller than others.

Key words : personalized search, query extraction, clustering, TF-IDF

1. 서 론

정보의 양이 늘어 날수록 검색 엔진은 사용자가 필요로 하지 않는 정보까지 검색하여 사용자가 원하는 정보를 찾는데 더 많은 시간과 노력이 필요하다. 따라서 인터넷을 통하여 정확한 정보를 신속하게 찾아서 제공하는 방법에 관심이 점점 증대되고 있다. Sergey Brin[1]의 논문에서 이러한 기술의 필요성을 예견하였다. 현재의 일반 적인 정보 검색 시스템은 수억 건 이상의 웹 문서를 데이터베이스화하여 사용자가 입력한 질의(query)에 대해 유사도가 높은 문서를 보여주는 방식이다. 이때 검색 속도를 향상하기 위해 검색어들을 추출하여 사용하기 때문에 하나의 질의에 의해 다양한 주제의 문서가 검색될 수 있다. 이것은 용어가 표현하는 의미의 모호성 등에 의해 발생하는 것으로 문서를 대표하는 용어를 찾는 것은 용이한 일이 아니다. 이에 따라, 사용자가 입력한 질의에 부가적으로 질의를 추가 하거나 질의

† 학생회원 : 동국대학교 컴퓨터공학과
leeye123@naver.com
†† 종신회원 : 동국대학교 컴퓨터공학과 교수
shcheon@dgu.edu
논문접수 : 2010년 2월 17일
심사완료 : 2010년 6월 5일

정보과학회논문지: 소프트웨어 및 응용 제37권 제8호(2010.8)

를 확장하는 방식을 통해 질의를 구체화 하여 검색의 효용성을 증가시킬 필요가 있다.

질의 확장(query expansion)을 이용한 검색의 효용성 증가 방식은 개인화 검색(Personalized search)[2]을 구현하기 위한 한 가지 방식으로 볼 수 있다. 개인화 검색은 사용자에게 정보를 보다 빠르고 손쉽게 보여주기 위한 방식이다.

이 연구에서는 사용자가 방문한 웹 문서를 이용하여 사용자가 원하는 주제의 문서를 자동으로 분류하고 문서에서 질의와 유사성이 높은 용어를 추출하여 사용자의 질의를 구체화 할 수 있는 방법을 제시하고 질의 확장을 통하여 개인화 검색 시스템을 구축할 수 있음을 보인다.

2. 관련 연구

2.1 질의 확장

Anick and Vaithyanthan[3]은 사용자들의 정보 검색에서 질의어는 다양한 요소에 영향을 받으며, 질의 용어 자체가 모호할 경우에는 그것을 질의어로 표현하는 데 어려움을 겪는다고 주장하였다. 질의 확장은 효과적인 정보검색이 가능하게 하며, 적절한 질의를 만들어야 하는 사용자들의 어려움을 덜어줄 수 있다.

질의 확장을 위해서는 질의어로 사용가능한 용어를 문서에서 추출할 수 있어야하며, 질의어와의 유사도를 측정할 수 있어야한다. 텍스트 마이닝(text mining)은 비 구조화된 텍스트에서 데이터 마이닝(data mining) 기법인 연관규칙 마이닝을 텍스트 처리에 응용하여, 텍스트로부터 연관용어 집합을 생성하고, 이를 통해 유용한 지식 정보를 효과적으로 얻으려는 시도를 한 것이다[4]. 이러한 방식들을 이용하여 질의어와 유사성이 높은 용어를 추출하여 질의를 확장하게 된다.

질의를 확장하는 방법으로는 시소러스(thesaurus)를 이용하는 방법과 문서들에서 고르게 추출되는 용어들을 클러스터링(clustering)하여 사용하는 방법과 질의어로 검색된 문서에서 질의어와 관련된 용어를 추출하여 추가하는 방법 등이 있다.

시소러스를 이용하는 방법은 용어간의 관계에 따라 어휘사전을 구축한 다음 동의어나 관련어를 질의어에 추가한다. 이 방식은 시소러스를 먼저 구축하여야 하는데 이는 컴퓨터에 의해 자동으로 구현되기가 어렵다는 문제가 있으며, 모든 웹 문서들의 용어들에 대한 시소러스를 구축하는 것은 불가능하다. 특정영역의 시소러스를 수작업으로 구축하여 활용하는 방법이 연구된 적이 있다[5].

클러스터링을 이용하는 방법은 문서들에서 발생되는 용어의 동시 출현빈도(co-occurrence frequency)를 측정하여 색인어간 유사도(analogous map)를 측정하고 질의어와 유사한 용어를 추출하는 것이다.

적합도 피드백(relevance feedback) 방법은 검색 결과에 대한 사용자의 피드백을 이용하여 문서로부터 새로운 질의를 추출하여 질의에 추가한다[2]. 이 방식을 사용할 경우 정보검색의 성능을 향상 시키는 것으로 나타났다[6,7]. 하지만 질의에 적합한 문서가 존재하지 않거나 적합한 문서를 사용하지 못할 경우 질의 확장의 효과가 현저히 떨어진다.

2.2 시멘틱 웹(Semantic Web)

일반적으로 웹 문서는 모두 다른 이형적(allogeneric)인 구조를 가지는데, 시멘틱 웹은 이런 구조의 웹 문서들을 메타데이터(metadata)를 이용하여 검색이 가능하게 한다[8].

메타데이터는 구조화된 데이터로, 웹 문서를 설명해주는 데이터를 의미한다. 시멘틱웹에서 중추적인 역할을 하는 온톨로지(Ontology)는 각 웹 문서의 메타데이터를 구축하기 위한 구조등을 제공한다. 사용자가 사용한 메타데이터를 이용하여, 사용자의 행위 패턴을 분석하며, 이 정보로 개인화 검색 시스템을 구축할 수 있다. 하지만 메타데이터는 웹 페이지의 개발자에 의해서 또는 자동으로 생성되는 것으로써 상업적으로 악용될 수도 있다[1]. 또한 모든 웹 문서의 메타데이터를 구축하기 어려우며, 기존의 HTML 문서 기반의 웹에서는 사용하기가 쉽지 않다. 이 논문에서는 시멘틱 웹이나 HTML 문서기반을 웹 등과 같은 웹 환경에 영향을 받지 않고 웹 문서의 출력 결과에서 텍스트를 추출하여 사용자의 질의를 확장할 수 있는 방법을 제시한다.

2.3 사용자 정보를 이용한 개인화 검색

개인화 검색을 위해 사용자의 프로파일 데이터(Profile Data)를 이용하여 사용자의 패턴 또는 관심 주제를 판단하여 검색 효율을 증가 시키는 방법이 있다. 프로파일 데이터는 사용자가 웹 사이트에 가입하면서 작성하는 개인정보로부터 사용자가 방문한 로그정보를 저장하는 것까지 다양한 종류가 있다.

2.3.1 프로파일 데이터를 이용한 개인화 검색

프로파일 데이터를 효율적으로 이용하여 사용자에게 보다 유용한 검색 결과를 제공하기 위해 많은 연구가 진행되고 있다. 프로파일 데이터의 사용범위는 광범위하다. 구글(google)의 경우 사용자가 과거 사용하였던 질의나 로그정보 등의 과거 정보(history data)를 이용하여 '사용자들이 이전에 질의했던 것 중 계속 관심 있는 질의에 새로운 검색결과가 추가되었을 때 적절하게 추천하는 방법은 무엇인가?'를 고려하는 경우도 있으며[9], 사용자의 클릭정보(Click History data)를 이용하여 사용자의 패턴 및 주제를 파악하는 방법[10,11]이 제안

되기도 하였다. 이러한 방식들은 사용자의 패턴을 미리 파악하여 사용함으로 실행 속도가 빠르며 특정 분야에서 정확도가 높을 수 있다. 하지만 모든 사용자의 데이터를 저장하는 것이 불가능하며, 정확한 정보를 이용하였는지 파악하기 힘들다. 또한 프로파일 데이터로 만들어진 주제 이외의 주제를 검색할 경우 프로파일 데이터를 기반으로 결과를 보여 주게 되므로 잘못된 결과가 나올 수 있으며, 프로파일 데이터를 이용하여 사용한다 하여도 개발자의 의도에 따라 변질될 가능성도 있다.

2.3.2 웹 페이지를 이용한 개인화 검색

야후(Yahoo)의 Y!Q!는 사용자가 현재 방문한 웹 페이지에서 텍스트를 추출하여 사용자의 질의를 확장한다[12]. 이러한 방식들은 사용자가 보고 있는 문서에서 단편정보를 추출하여 이용함으로써 특정 주제에 영향을 받지 않기 때문에, 사용자의 프로파일 데이터를 이용할 때 발생되는 문제점들을 일부 해결할 수 있다. 하지만 사용자가 주제와 관련되지 않은 문서를 방문하여 시스템이 용어를 추출하면 잘못된 결과를 보인다. 또한 항상 같은 문서에서는 같은 결과만을 보여준다. 즉, 사용자의 데이터가 쌓여서 점점 좋은 성능을 내는 프로파일 데이터를 사용하는 방식에 비해 항상 같은 성능만을 보인다는 것이다. 또한 사용자의 패턴을 인식하지 않음으로 사용자가 검색하는 주제를 파악하기 힘들다는 문제가 있다.

3. 검색어 질의 확장을 이용한 검색 시스템

이 연구에서는 방문한 문서를 이용하여 단어를 추출하고 가중치를 측정한다. TF-IDF는 핵심어를 추출하는 측면에서는 학습을 이용한 단어 추출방식과 유사한 성능을 보이지만 문서의 수가 많거나 분류할 문서가 뚜렷한 특징을 나타내지 못할 경우 문서 분류는 가능하지만 의미를 가지지 못하는 단어를 핵심어로 제공하게 된다. 인기도 측정법은 문서를 분류할 수 있는 핵심어 이면서 주제와 관련된 의미를 가진 단어를 추출하는 방법이다. 이 연구에서는 인기도 측정을 위한 수식 및 방법을 제시한다.

연구에서 구축한 시스템은 텍스트 문서에서 명사 등을 추출하는 자연어 처리 부분(part)과 특징(feature)이 되는 용어를 추출하는 부분 그리고 추출된 용어를 이용하여 문서를 분류하는 부분, 마지막으로 추출한 용어의 랭킹을 정하는 부분으로 되어 있다. 명사를 추출하는 부분은 KLT(구 HAM: Hangul Analysis Module)[13]을 이용하여 처리하였다. 용어 분류에는 Aporiori 알고리즘을 이용하였으며, 관심 주제별 문서 분류는 ANN(Approximate Nearest Neighbor)[14]를 이용하였다.

3.1 연관 규칙 알고리즘

연관 규칙 알고리즘은 Aporiori 알고리즘을 적용해서 질의에 추가할 용어를 선정하였으며, 용어의 공간 빈도를 기반으로한 클러스터링 기법을 적용하여 문서를 분류하였다. 연관규칙 마이닝에서 용어간의 연관관계(association relationship)는 항목들 사이에 존재하는 유사성 또는 패턴을 의미하는 것으로 한 단락이나 한 문장을 하나의 트랜잭션으로 하여 용어간의 연관성을 측정하는 것이다.

Aporiori 알고리즘은 Agrawal and Srikant[15]이 제안한 것으로 후보항목 집합을 생성하고, 발생 빈도를 계산한 후 사용자가 정의한 최소지지도를 가진 빈발 항목집합을 결정하는 것이다. Aporiori 알고리즘은 각 패스에서 빈발 항목집합들의 후보 항목집합을 구성한 후에 각후보 항목집합의 발생 빈도를 계산하고, 사용자가 정의한 최소지지도를 기준으로 하여 빈발 항목집합들을 결정한다. 다음 단계에서는 이들 빈발 항목집합들로부터 최소신뢰도 임계치를 만족하는 연관규칙을 모두 찾는다. 최소신뢰도 임계치는 발생된 항목집합에서 사용자가 지정한 최소한의 빈도수를 뜻한다. 최소한의 빈도수를 넘지 못하는 경우 항목집단에서 제외된다.

3.2 키워드 추출 시스템구조

현재 개인화 검색엔진은 개인의 프로파일 데이터를 이용하여 개인의 패턴을 인식하고 사용자가 원하는 문서를 먼저 추천하는 방식을 취한다. 또는 사용자의 패턴을 학습하여 유사한 사용자가 찾았던 결과를 상위에 보여줌으로써 개인의 관심사를 만족시켰다. 하지만 이 방식은 프로파일 데이터를 누적시켜 사용하는 방식으로 특정 분야에 대한 정확도 및 추천 시간에 대한 성능은 좋아지지만 2.3.1에서 제시된 문제를 여전히 가지고 있다. 이러한 이유로 개인의 프로파일 데이터를 사용하지 않고 사용자의 현재 패턴을 분석하여 질의 확장을 통해 사용자의 검색 질의어를 조금 더 명확하게 해 줌으로써 사용자가 원하는 정보를 빠르게 찾도록 도와주는 것이다. 개인이 원하는 정보를 빠르게 검색하여 검색의 효용성을 높인다는 점에서 개인화 검색에 가깝다.

사용자들은 웹 문서를 검색할 경우 검색엔진에 질의를 던지고 결과를 받게 된다. 받은 결과는 아래 그림 1과 같이 헤드라인(headline)과 그 요약된 정보 문서로 나뉘어서 보이게 된다.

사용자는 헤드라인과 요약된 정보를 이용하여 자신이

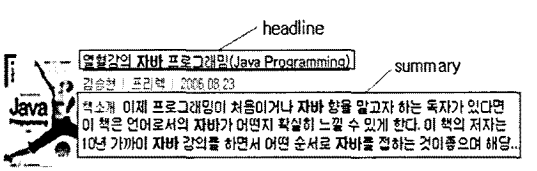


그림 1 검색된 문서의 표현 형식

찾고자 하는 문서와 유사한 문서를 클릭하여 읽게 된다. 이렇게 사용자가 자신이 원하는 주제만을 읽게 되므로 시스템 자체에서 도메인(domain)을 제한할 필요가 없게 된다. 사용자가 자신이 원하는 문서를 읽으면 문서에서 특징이 되는 용어들을 추출한다. 시스템이 수집한 모든 정보는 모든 검색이 끝날 때까지만 유지시킨다. 이 정보를 이용하여 사용자의 패턴 또는 관심 주제를 추적하게 되는데, 이 방식은 프로파일 데이터를 이용하는 것과 유사하지만 검색되는 동안만 잠시 유지되며, 검색이 종료되면 자동 삭제된다는 점에서 차이가 있다.

데이터베이스에 저장된 정보를 이용하여 시스템은 각 문서를 분류하게 된다. 분류의 기준은 사용자가 입력한 검색어이며, 각 검색어를 축으로 문서를 분류한다. 그림 2는 질의로 사용된 용어가 2개일 경우 문서가 분류되는 예이다. 데이터베이스의 저장된 각 문서에서 질의로 사용되는 용어가 출현하는지를 검사하고 출현할 경우 문서에서 차지하는 비중이 얼마인지를 이용하여 그림 2와 같이 좌표로 웹 문서를 표시한다. 문서에서 질의어가 출현하지 않을 경우 좌표 값을 0으로 한다. 용어가 문서에서 차지하는 비중은 문서에서 특징을 추출하여 데이터베이스에 저장할 때 사용되는 Apriori 알고리즘의 결과값을 이용한다.

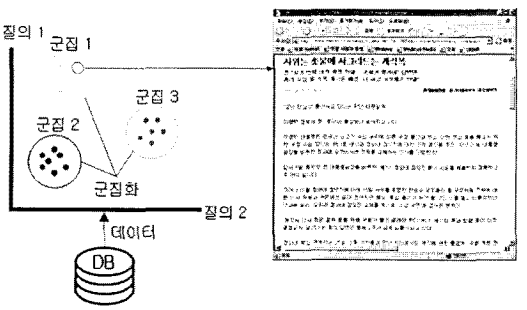


그림 2 질의를 이용한 문서 분류 방식

식 (1)을 이용하여 문서에서의 비중을 측정하게 된다. 측정된 값은 데이터베이스에 저장되며, 질의어가 출현하는지 검사한 후 좌표에 표시할 때의 수치로 사용되는 것이다.

$$\frac{\text{해당용어가출현하는문장의수}}{\text{전체문장의수}} * 100\% \qquad (1)$$

사용되는 좌표의 차원(Dimension)은 검색어로 사용된 질의어의 수와 동일하다. 그림 2와 같이 질의어가 2개일 경우 2차원의 좌표에 각 문서를 표기하는 것이다. 좌표에 표시된 웹 문서는 최근접점 탐색 알고리즘으로 분류한다.

최근접점 탐색 알고리즘(Nearest Neighbor Search)으로는 ANN(Approximate Nearest Neighbor)[14]을 이용하였다. 이렇게 나누어진 문서들은 군집(class)을 이루게 되며, 각 군집은 하나의 주제를 나타내게 된다. 하지만 질의의 수가 2개 이하일 경우, 문서가 세밀한 주제로 분류되지 않고, 큰 주제 별로 군집을 이루게 된다. 이러한 문제는 군집에서 용어를 추출하여 질의를 확장할 때 검색 주제와 밀접한 용어를 추출하여 사용하기 힘들게 한다.

논문에서 군집의 주제를 대표할 수 있는 용어는 각 군집의 문서들에서 추출하여 사용하며, 군집의 모든 문서의 텍스트를 사용하지 않고 이전에 추출하였던 각 문서의 용어들을 이용한다. 그리고 특정 문서에서 하나의 단어가 많이 출현하는 경우를 고려하여 하나의 문서보다 여러 문서에서 고르게 출현하는 용어를 추출하도록 한다. 각 군집은 하나의 주제를 뜻하게 되는 것이며 가장 군집이 큰 것은 사용자가 관심을 가지는 주제와 높은 연관성을 가진다. 이것은 앞에서 말한 것과 같이 사용자가 검색 결과에서 관심을 가지고 있는 문서만을 클릭하여 읽기 때문이다. 이 방식은 프로파일 데이터를 사용할 때 발생하는 문제 중 질의의 모호성에 의해 발생되는 문제와 프로파일 데이터로 만들어진 주제에 편중되어 결과를 생성하는 문제를 완화시킬 수 있다. 2.3.2절에서 설명한 방식과 유사하지만 군집을 이용하여 사용자의 패턴을 분석한다는 점에서 사용자의 의도를 파악하기 쉬우며, 사용자가 일부 잘못된 문서를 읽는다 하더라도 군집에 의해 분류되므로 시스템 성능에 영향이 적다.

3.3 키워드 랭킹(Ranking)

지금까지 사용자가 방문한 문서를 통하여 사용자의 검색에 도움이 되는 질의어를 추출하였다. 하지만 방문한 문서에서 자주 출현되고, 많은 문서가 포함된 군집에서 출현하는 질의어라 하더라도 유용성(quality)이 떨어지는 경우가 허다하다. 그렇기에 방문한 페이지를 통해 수집된 단어의 유용성을 측정하여 질의어들의 랭킹을 정하는 것이 필요하다. 또한 새로운 주제 관련 단어를 빠르게 상위에 랭크 시킴으로써 검색 주제의 변화 또는 새로운 주제어를 상위랭킹에 빠르게 반영한다.

3.3.1 인기도를 이용한 유용성 측정

단어의 유용성은 문서에서 자주 출현하는 문서일수록 유용성이 높을 수도 있으나 자주 출현하지 않지만 사용자가 찾는 웹 문서들을 상위의 검색 결과로 이끌어낼 수 있는 단어들 또한 존재한다. 이러한 단어들은 단순하게 출현빈도로 측정할 경우 낮은 랭크를 가지게 되므로 빈도수가 아닌 사용자가 선호하는 인기도(popularity)를 이용하여 측정하여야한다. Junghoo Cho[16]은 웹 문서의 인기도와 유용성의 상관관계를 증명하고, 사용자가 인식한 시점으로부터의 인기도와 시간당 인기도 변화량을 이용하여 유용성을 측정하였다. 이러한 인기도 또는

가중치 측정에서 가장 우려되는 문제는 빈도수만으로 측정할 경우 광고 문자 또는 대명사 등의 랭킹이 높아진다. 새롭게 등장한 단어 또는 자주 발생되지 않는 단어를 높은 가중치로 상위권에 넣는 방식은 기존에 발생된 단어들이 모두 하위 랭크로 밀려나게 된다. 그러므로 이 논문에서는 식 (2)를 응용하여 단어 간의 인기도를 측정할 수 있도록 한다.

$$\frac{dP(w_p,t_i)/\triangle t_i}{P(w_p,t_i)}+P(w_p,t_i) \quad (2)$$

t_i : 현재 시간

$P(w_p, t_i)$: 시간 t_i 일 때 단어 w의 인기도 값

3.3.2 단어 유용성 측정

분류된 문서 집합에서 식 (2)를 유용성 값으로 이용하여, 질의로 사용 가능한 단어를 추출한다. 단어의 인기도는 그 단어가 얼마나 검색에 유용한지를 뜻한다.

$$n=\frac{TF-IDF(d,w)}{y} \quad (3)$$

n : 문서의 유용성

y : 0 <= y <= 1 의 값. 문서와 군집간의 관계 수치 값

d : 새로 방문한 문서

w : 문서에서 추출된 단어

식 (3)은 단어와 문서간의 연관 관계를 고려한 것이다. y는 문서와 군집 간의 관계 값으로써, 단어 w가 들어있는 문서가 모든 군집에 분포할 경우 중요도를 떨어트린다. 또한 TF-IDF값을 이용하여 단어가 속한 문서들에서 얼마만큼의 중요도를 차지하는지를 측정하고, 단어 w가 들어있는 문서가 여러 군집에 분포할 경우 중요도가 낮으므로 측정된 TF-IDF값을 조정할 수 있도록 한다.

$$y=\frac{x-1}{c-1}+1 \quad (4)$$

c : 총 군집 개수.

x : 단어 w가 포함된 문서가 출현한 군집 수. 1<=x<= c

식 (4)는 문서의 유용성을 판단하는 수식으로 전체 군집에 다수 분포된 단어일 경우 개인화를 위한 문서 분류를 할 수 없다. 이러한 단어들의 n값을 조절하기위해 y를 사용한다.

$$\begin{aligned} TF-IDF(d,w) &= r(d,w)*weight(w) \\ weight(w) &= \log(1+N/f(w)) \\ r(d,w) &= 1+\log(f(d,w)) \end{aligned} \quad (5)$$

N : 총 방문한 문서 수

w : 단어

d : 새로 방문한 문서

f(d, w) : 문서 d에 키워드 w가 몇 번 나오는지를 나타낸다.

r(d, w) : 키워드 빈도를 바탕으로 문서 d와 키워드 w 사이의 연관성을 수치화

f(w) : 단어 w가 등장하는 문서의 개수

weight(w) : 단어 w의 가중치

식 (5)는 TF-IDF를 측정하기 위한 식이다. 기존의 TF-IDF의 경우 f(d, w), N 또는 f(w)값이 0일 경우 전체 값이 0이 된다. 이것을 피하기 위해 수식을 0이라는 값이 발생할 경우 최종값이 1이 될수 있도록 수정된 수식을 사용하였다. 또한 weight(t)에서 f(t1)은 1이고, f(t2)는 2라고 할 때, t1과 t2 사이에는 차이가 작지만, 단순히 1 / f(t)로 계산하면 w(t1)은 w(t2)의 두 배가 된다. 이런 문제를 해결하기 위해서 로그 함수를 사용한다. 이것은 r(d, w)와 식 (6)에서 단어의 인기도 값에서도 같은 이유로 적용되었다.

$$Q(w_p,t_i)=\frac{n}{r}\left[\frac{\triangle WQ(w_p,t_i)/\triangle t_i}{WQ(w_p,t_i)}\right]+\log(WQ(w_p,t_i)) \quad (6)$$

$WQ(w_p,t_i)$: 시간 t_i 일 때 단어 w의 유용성

n : 군집들에서 단어 w가 들어있는 문서의 중요도

r : Vp(w,v_c)

Vp(w,v_c) : 방문한 문서수가 v_c가 될 때마다, v_{c-1}에서부터 v_c 사이에 방문한 문서에서 단어 w가 출현한 문서의 비율

v_c : 일정한 방문 수마다 Vp(w,v_c)을 연산하도록 하는 기준 값

식 (6)은 식 (2)를 단어의 유용성을 측정할 수 있도록 변형한 수식으로 n값은 식 (3)을 이용한다. 하지만 r의 경우 단순히 고정된 시간동안 w라는 단어가 출현하는 문서의 방문한 횟수로 사용할 경우 사용자가 하나의 페이지를 읽는 시간이 일정하지 않으므로 고정 값 기준시간 t를 결정하기 힘들다. t값이 정확하지 않다면 사용자가 문서를 읽는 시간에 따라 기준시간 t동안 하나의 문서도 방문하지 않는 경우가 발생한다. 또한 사용자가 검색을 일시 중단할 경우 기준시간 t로 인하여 유용성을 정확히 측정할 수 없다. 그러므로 일정한 문서수를 방문하는 동안 단어 w가 출현하는 문서의 방문 비율을 측정하여 사용하도록 한다. 식 (6)의 방식을 이용한 단어 랭킹 알고리즘은 인기도 기반 키워드 랭킹 알고리즘(popularity-based keyword ranking)이라고 한다.

4. 실험 및 결과

이 장에서는 질의 확장 시스템의 성능을 분석한다. 질의 확장시스템의 성능은 검색어와 연관성 있는 단어가 상위에 랭크 될수록 더 높은 성능을 나타내게 된다. 인기도 기반 키워드 랭킹 알고리즘의 성능을 평가하기 위해 HL(Half-Life)[17]을 이용하며, TF-IDF을 이용한 단어 가중치 추출 방식과 비교한다. 또한 NDCG(Normalizing Discounted cumulative gain)[18] 성능 평가 방법을 이용하여 추천단어를 이용하였을 때 검색 시스템의 성능을 평가한다.

$$\sum_{j=1}^{R} \frac{1}{2^{(j-1)/(\alpha-1)}} \quad (7)$$

식 (7)은 HL을 나타낸 공식으로 R은 주제와 관련된 단어들의 랭킹을 뜻하며, α 는 half-line을 뜻하는 고정값이다. j는 주제와 관련된 단어의 랭킹이다. 논문에서는 HL[17]를 참고하여 half-line을 5로 고정한다. α 를 5로 하는 half-Life를 HL@5로 표기한다. 식 (7)의 결과 값이 높기 위해서는 정확한 키워드 추출이 밑바탕이 되어야 한다. 또한 키워드들이 상위에 랭크될수록 높은 결과값을 가진다. 단어 추출에서 기본적으로 많은 단어를 추출할수록 주제와 관련된 단어를 많이 가질 수 있다. 하지만 HL에서는 기준 α 보다 높은 랭킹을 가지지 못할 경우 전체적인 합산 값이 높을 수 없다. 그러므로 HL을 이용한 측정방식이 논문에서 제시하는 키워드 추출방식과 랭킹방식을 종합적으로 측정할 수 있는 실험 방식이다.

$$DCG_p = rel_1 + \sum_{i=1}^{p} \frac{rel_i}{\log_2 i}$$

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

p : 총 검색 결과 문서들 수.

DCGp : p개의 실제 검색 결과 순서에 따른 가중치를 적용한 점수 값

reli : 랭크 i번째 문서의 중요도 값

IDCGp : 가장 이상적인 검색 결과 순위일 때의 DCG값

NDCG는 식 (9)를 이용하여 측정한다. NDCG를 측정하기 위해 검색된 결과의 모든 문서를 사용하지 않고 P개의 문서를 사용할 때 NDCG@P라고 표현한다. 실험에서는 랭킹 30위까지의 문서를 이용하여 측정하므로 NDCG@30으로 표현한다. 이 성능 측정법은 야후 및 구글 등의 검색 엔진 성능 평가에도 사용되며, 결과값이 1에 가까울수록 좋은 성능을 나타낸다.

실험은 자바관련 책을 검색하는 것으로 한다. 검색결과 100개중 자바 네트워크 프로그래밍 관련 책에 관한 사이트를 방문하고 각각의 단어 추출 방식의 성능을 측정한다. 주제와 관련된 평가단어(item set)는 방문한 웹 문서의 모든 단어 중에서 주제와 관련된 9개의 단어를 추려서 사용하였다.

같은 주제를 검색하더라도 사용자는 검색 할 때마다 이전 검색에서 방문한 문서와는 조금씩 다른 문서를 방문하게 된다. 그러므로 실험에서는 일정하게 고정된 평가단어를 사용하지 않고 실험이 종료된 이후에 TF-IDF로 추출된 모든 단어 중 자바 네트워크와 관련된 단어 9가지를 선정하여 사용한다. TF-IDF는 핵심단어의 가중치를 측정하기 위해 문서에서 발생되는 대부분의 명사 단어를 모두 사용한다. 또한 인기도 기반 알고리즘에서 추출된 단어에서 평가단어를 구성할 경우 인기도 기반 알고리즘에 치우친 단어가 평가단어가 될 수 있으므로 TF-IDF에서 추출된 단어 중 일부를 사용한다.

그림 3은 검색하는 주제의 변경 없이 "자바 네트워크 프로그래밍 책" 관련 웹 문서를 방문하였을 때 HL@5를 측정한 것이다. 세로축은 HL@5의 합산 값이며, 가로축은 방문한 문서의 개수를 뜻한다. '4'번째 웹 문서를 방문한 이후부터 인기도 기반 키워드 랭킹 알고리즘의 성능이 상승한다. TF-IDF의 경우 '3'번째 웹 문서 이후에 중요단어가 자주 발생되고 추출한 단어의 수가 기하급수 적으로 늘어남에 따라 오히려 주제와 관련 없는 단어들이 상위 랭킹에 나타난다. '1'번째 웹 문서에서 인기도 기반 키워드 랭킹 알고리즘이 0인 이유는 알고리즘 특성상 단어에 의한 연관 관계로 누적된 값을 사용하는 것이므로 처음 방문한 문서로는 랭킹을 정할 수 없기 때문이다. 그러므로 HL@5의 값이 0인 것이다.

그림 4는 자바 네트워크 외에 자바와 관련된 웹 문서를 검색한 이후 자바 네트워크와 관련된 웹 문서를 방문하였을 때 HL@5의 결과 값이다. 가로축의 '1'은 검색 주제가 변경된 이후 변경된 주제와 관련된 첫 번째 문서의 방문을 뜻한다. 그림 4와 같이 TF-IDF의 경우 검색 주제의 변경시 빠르게 주제와 관련된 단어를 추출하지 못한다. 하지만 인기도 기반 키워드 랭킹 알고리즘

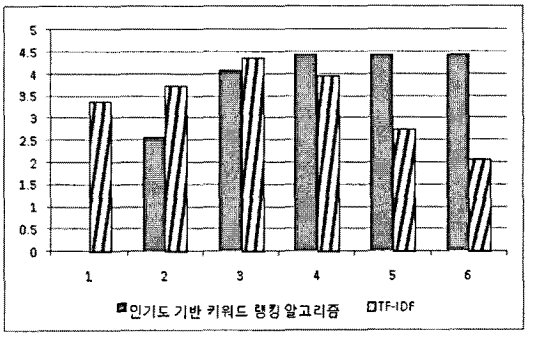


그림 3 검색 주제 변경 없이 실험

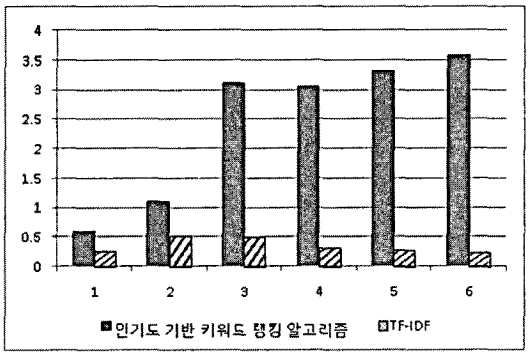


그림 4 검색 주제 변경 후 실험

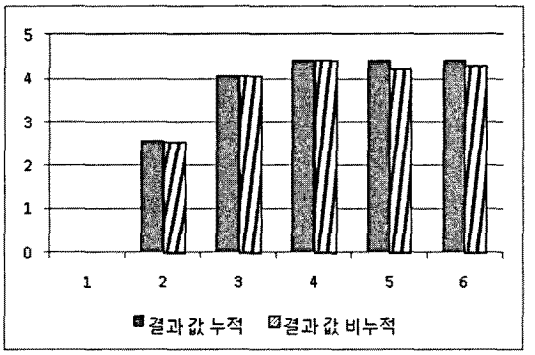


그림 6 검색 주제 변경 없이 실험

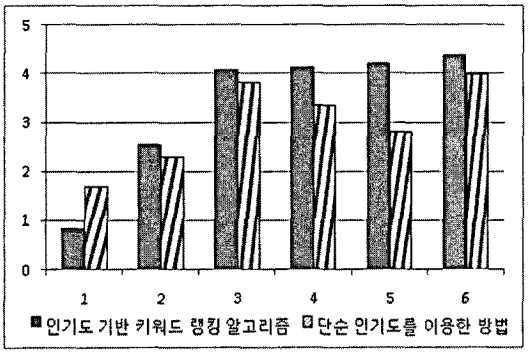


그림 5 단순 인기도와의 성능 비교

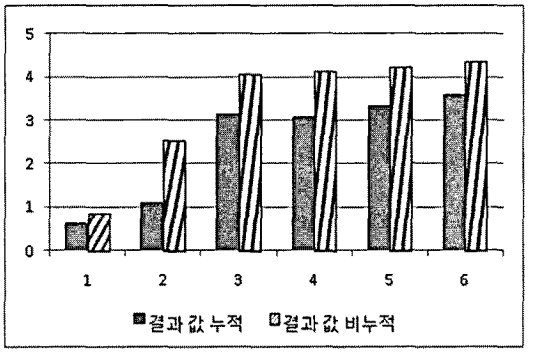


그림 7 검색 주제 변경 후 실험

의 경우 3번째 웹 문서 방문한 이후부터 정상적으로 주제와 관련된 단어를 빠르게 상위 랭킹에 랭크시킨다. 그림 3의 경우에도 TF-IDF가 초기 높은 성능을 보이는 것으로 보이지만 단순히 단어를 많이 추출하였기 때문이다.

그림 5는 그림 4와 같은 실험 방법을 이용하여 단순 인기도를 이용하는 식 (2)의 방법과 논문에서 제시된 알고리즘 식 (6)을 비교한 것이다. 논문에서 제시된 단어 가중치 측정 방법인 식 (6)은 기존의 수식이나 방법에 비해 빠르게 주제와 관련된 단어만을 추출하여 상위의 랭킹에 오를 수 있도록 단어의 가중치를 조절하였다. 또한 주제가 변경될 경우에도 빠르게 변경된 주제에 대한 단어를 추출하고 단어에 적합한 가중치를 부여한다.

그림 6은 검색 주제를 변경 없이 하나의 주제에 대하여 검색하였을 경우 식 (6)의 값을 누적 하였을 경우와 누적하지 않았을 경우에 대한 실험 결과이다. 유사한 성능을 보이지만 4번 째 웹 문서 방문 이후 누적 값을 사용하는 것이 성능이 높아진다. 이것은 주제와 관련된 단어일수록 출현 빈도가 높아지므로 알고리즘 특성상 모든 문서에서 출현빈도가 높을 경우 식 (3)에 의해 가중치 값을 낮추게 된다. 이러한 현상은 주제와 관련 되었지만 모든 문서에서 자주 발생하는 단어일 경우 사용자가 이미 인식하고 있는 단어이므로 이러한 단어보다 사용자가 인식하지 못하는 단어를 높은 랭크에 올리기 위한 방법이다. 하지만 누적 값을 사용할 경우 식 (3)의 영향을 적게 받게 된다. 이것이 성능상 좋게 보일 수 있지만 그림 7과 같이 주제가 변경될 경우 새롭게 변경된 주제의 관련단어들이 빠르게 상위 랭크에 오르지 못하는 치명적인 문제를 발생시킨다.

그림 7은 검색 주제가 변경되었을 경우 누적 값을 사용하는 것에 비해 사용하지 않는 것이 더 높은 성능을 보인다. 누적된 값에 의해 새롭게 등장한 주제와 관련된 단어들이 빠르게 상위 랭크에 오르지 못하는 것이다. 그림 7의 실험을 통해 단순히 하나의 주제를 검색 할 때 높은 성능을 보이는 것 외에도 변경된 주제에서 높은 성능을 고려할 필요가 있음을 보인다.

표 1은 사용자가 검색어를 입력하여 얻은 검색결과에서 검색 결과의 문서를 방문 할 때 마다 추천되는 단어들을 이용하여 검색 결과를 재 랭킹 하였을 때 각 알고리즘들의 성능을 나타낸 것이다. Avg_NDCG@30은 각 문서 방문 횟수에 맞게 문서들을 방문하고 추천되는 단어들로 재 랭킹하였을 때 결과를 평균한 것이다. 검색

표 1 하나의 검색 주제에서 알고리즘별 성능

문서 방문 횟수	단어 추천 알고리즘	Avg_NDCG@30
순수 검색엔진의 검색 결과 성능		0.222746
1	popularity-based keyword ranking	0.287094
	Aporiori	0.110498
	tf-idf	0.281512
2	popularity-based keyword ranking	0.406245
	Aporiori	0.110498
	tf-idf	0.281512
3	popularity-based keyword ranking	0.768737
	Aporiori	0.50936
	tf-idf	0.220996
4	popularity-based keyword ranking	0.768737
	Aporiori	0.110498
	tf-idf	0.110498
5	popularity-based keyword ranking	0.768737
	Aporiori	0.110498
	tf-idf	0

문서 방문 횟수가 3이면, 검색결과들에서 3개의 문서를 방문하게 된다. 이때 검색결과 상위 30개 중에서 주제와 관련된 문서들을 랜덤하게 3개를 방문하게 된다. 이러한 과정을 5회 실행하여 얻은 결과를 평균한 것이다. 5회를 기준으로 평균한 것은, 5회 이상일 때에는 추천단어들의 조합이 서로 유사하게 발생되기 때문이다. Avg_NDCG @30의 값은 0과 1사이의 값을 가지며, 1에 가까운 값을 가질수록 높은 성능을 나타낸다.

하나의 검색 주제에 대해서 인기도 기반 키워드 랭킹 알고리즘(popularity-based keyword ranking)이 높은 성능을 나타낸다. 표 1과 같이 3개 이상의 문서를 방문하였을 경우 인기도 기반 키워드 랭킹 알고리즘은 0.7이상의 성능을 보인다.

5. 결 론

논문에서 제시한 인기도 기반 키워드 랭킹 알고리즘은 TF-IDF처럼 문서의 그룹을 구분지을 수 있는 단어일 경우 빠르게 가중치를 높이며, 단순히 자주 발생되는 단어라도 주제와 밀접한 단어일 경우 일정한 인기도 만큼의 가중치를 유지할 수 있는 방법을 제시하고 실험하였다. 실험과 같이 인기도 기반 키워드 랭킹 알고리즘은 주제와 관련된 새로운 단어를 빠르게 상위에 랭크 시키며, 주제와 관련 되었지만 모든 문서에서 발생되는 단어에 대해서도 가중치 조절이 가능함을 보였다. 또한 검색 주제의 변화를 인식할 수 있으므로 검색 주제의 변경에 대처할 수 있으며, 질의 확장을 통해 검색 성능의 향상을 가져왔다. 이것은 이전의 일반적인 개인화 검색의 단점을 보완한 것이다.

본 논문에서 인기도 기반 키워드 랭킹 알고리즘은 주제의 변화에 빠르게 적응하면서, 주제의 변화를 시스템이 인지하여, 질의 확장 외에도 프로파일 데이터를 이용하는 방법에서도 정확한 단어 정보를 이용하여 사용자가 원하는 주제와 관련된 프로파일 데이터를 정확히 검색하고 세밀한 주제 분류가 가능하도록 도와줄 것이다.

6. 향후 과제

질의 확장의 경우 자동으로 확장되는 질의어의 수가 늘어날수록 사용자가 검색에 사용하는 사용자 질의에 비해 확장 질의가 부자연스러워 진다. 그러므로 무작정 질의어를 늘이는 방법보다는, 사용자의 검색 초기에는 질의어 확장을 통해 사용자의 검색을 도우며, 사용자의 검색 주제가 파악될 경우, 검색 주제에 해당하는 프로파일 데이터를 통하여 자연어 형태의 검색어를 사용자에게 제시하는 방식이 필요할 것이다.

참 고 문 헌

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Technical report*, Stanford University, 1998.

[2] Buckley C., Salton G., and Allan J., "The Effect of Adding Relevance Information in a Relevance Feedback Environment," *Proceedings of 17th annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, pp.292-300, 1994.

[3] Anick, P. G. and Vaithyanathan, S., "Exploiting Clustering and Phrases for Context-Based Information Retrieval," *Proceeding of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.314-323, 1997.

[4] Tribula, W. J., "Text Mining," *Annual Review of Information Science and Technology*, pp.385-419, 1999.

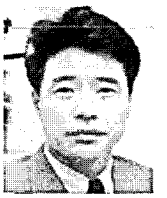
[5] Kristensen, J., "Expanding End-Users," Query Statements for Free-text Searching with a Search-aid Thesaurus," *Information Processing and Management*, vol.11, pp.22-33, 1968.

[6] Salton, G., and Buckley, C., "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science*, vol.41, pp.288-297, 1990.

[7] Harman, D., "Relevance Feedback Revisited," *Proceedings of 15th annual International ACM-SIGIR Conference on Research and Development in*

Information Retrieval, Copenhagen, pp.1-10, 1992.
[8] Li Ding, Tim Finin, and Anupam Joshi, "Swoogle: A search and metadata engine for the semantic web.," In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, pp.58-61, 2004.
[9] B. Yang and G. Jeh, "Retroactive answering of search queries," *Proceedings of the 15th international conference on World Wide Web*, pp.457-466, 2006.
[10] Qiu, F., and Cho, J., "Automatic identification of user interest for personalized search," In *Proceedings of the 15th International Conference on World Wide Web.*, pp.727-736, 2006.
[11] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," In *Proceedings of the 15th International Conference on World Wide Web.*, pp.675-684, 2004.
[12] Reiner Kraft, Chi Chao Chang, Farzin Maghoul, and Ravi Kumar, "Searching with context," In *15th International CIKM Conference Proceedings*, pp.477-486, 2006.
[13] S. S. Kang, "A Rule-Based Method for Morphological Disambiguation," *Proceedings of the NLPRS (Natural Language Processing Pacific Rim Symposium)*, pp.67-72, 1999.
[14] D. Mount, "ANN: Library for Approximate Nearest Neighbor Searching," http://www.cs.umd.edu/~mount/ANN/, 2006.
[15] Agrawal, R., and Srikant, R., "Fast Algorithms for Mining Association Rules," *Proceeding of the 20th International Conference on Very Large Databases*, pp.487-499, 1994.
[16] J. Cho, S. Roy, and R. Adams, Page quality: In search of an unbiased web ranking. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data 2005*, Baltimore, Maryland, June, pp.14-16, 2005.
[17] Zhicheng Dou, Ruihua Song, Ji-Rong Wen, "A Largescale Evaluation and Analysis of Personalized Search Strategies," *Proceedings of the 16th international conference on World Wide WebNew York*, NY, USA: ACM, pp.581-590. 2007.
[18] Kalervo Jarvelin, Jaana Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, 20(4), pp.422-446 (2002).

이 정 훈
2005년 동국대학교 컴퓨터공학과 학사
2007년 동국대학교 컴퓨터공학과 석사
2007년~현재 동국대학교 컴퓨터공학 박사과정. 관심분야는 개인화검색, 검색엔진, Web crawler



전 서 현
1978년 경북대학교 전자계산공학과 학사
1980년 한국과학기술원 전산학과 석사
1991년 한국과학기술원 전산학과 박사
1980년~1983년 계명대학교 전임강사. 1983년~현재 동국대학교 컴퓨터공학과 교수
관심분야는 함수언어, 심볼릭 컴퓨팅, 분산컴퓨팅, 개인화 검색