다차원 인덱스를 위한 벡터형 태깅 연구

(A Vector Tagging Method for Representing Multi-dimensional Index)

정 재 윤 † 진 현 철 †† 김 종 근 †††
(Jae-Youn Jung) (Hyeon-Cheol Zin) (Chonggun Kim)

요 약 인터넷 사용이 대중화되면서 개인이 정보의 키워드 또는 검색할 주제에 따라 원하는 정보에 쉽게 접근할 수 있다. 이때 다양한 구조를 갖는 자료들의 속성을 잘 나타내는 메타데이터를 이용하면 검색의도에 보다 정확하게 부합하는 검색 결과를 얻을 수 있어 다양한 연구가 지속되고 있다. 본 연구는 소그룹의 사용자들이 공동으로 관심 있는 웹 콘텐츠의 즐겨찾기를 공동으로 유지 관리하는 용도로 다차원 벡터형 태그를 제안한다. 제안하는 벡터형 태그는 정보 유용성을 나타내는 색인을 벡터방식으로 기술하고 이것을 활용해 정보의 분류·관리·재활용의 효율을 높이는 표현법이다. 벡터방식 태깅은 대상 키워드에 사용자들이 두 개 이상의 요소에 대한 우선순위를 부여하고 벡터 방식으로 표현한다. 이 때 벡터의 기본이 되는 벡터공간은 정보생성시간, 선호순위 등으로 구성한다. 벡터성분으로 산출할 수 있는 벡터크기가 정보의 유용성을 나타내며 순위측정의 기준이 된다. 제안방식에 의한 순위측정은 단순한 링크구조에 의해 측정된 순위와 비교하였을 때, 사용자의 검색의도에 부합하는 순위 정보를 제공하고 있다.

키워드 : 인터넷검색, 벡터형 태그, 정보 유용성, 순위 측정

Abstract A Internet user can easily access to the target information by web searching using some key-words or categories in the present Internet environment. When some meta-data which represent attributes of several data structures well are used, then more accurate result which is matched with the intention of users can be provided. This study proposes a multiple dimensional vector tagging method for the small web user group who interest in maintaining and sharing the bookmark for common interesting topics. The proposed method uses vector tag method for increasing the effect of categorization, management, and retrieval of target information. The vector tag composes with two or more components of the user defined priority. The basic vector space is created time of information and reference value. The calculated vector value shows the usability of information and became the metric of ranking. The ranking accuracy of the proposed method compares with that of a simply link structure, The proposed method shows better results for corresponding the intention of users.

Key words : Internet Searching, Vector Tag, Information Value, Ranking

· 이 논문은 2009년 지역 인력 양성 사업의 연구비 지원으로 수행된 결과입니다.

† 학생회원 : 영남대학교 컴퓨터공학과
e-mail@ynu.ac.kr
†† 비 회 원 : 영남대학교 컴퓨터공학과
cyberzin@ynu.ac.kr
††† 종신회원 : 영남대학교 컴퓨터공학 교수
cgkim@yu.ac.kr
(Corresponding author임)

논문접수 : 2008년 1월 3일
심사완료 : 2009년 8월 3일

정보과학회논문지: 소프트웨어 및 응용 제36권 제9호(2009.9)

1. 서 론

웹의 정보가 급격하게 늘어나면서 인터넷 검색엔진이 처리할 수 있는 고유의 웹 처리 영역의 범위가 점차 확대되고 있다. 이러한 사실은 방대한 웹 환경에서 단순히 검색결과의 리스트만을 제공하는 한계를 극복해야 함을 보여준다. 현재 사용자의 요구도 불필요한 정보의 구분이나 검색의도에 부적합한 정보의 퇴출 등에 집중되어 있으므로 이러한 요구에 대응할 수 있는 기술이 필요하다.

정보를 이용하려는 사용자는 먼저 다양한 방법으로 정보를 획득하고 원하는 목적에 따라 분석 및 가공하여 정보를 통합한다. 사용자는 통합된 정보를 바탕으로 정

보의 중요성을 결정하고 활용하게 된다. 이 때 정보의 중요성을 결정하기 위해 사용자는 검색시스템이 제공하는 정보를 활용하게 된다. 그러나 검색시스템이 제공하는 정보는 단순링크기반을 이용한 색인어에 의한 검색이나 주제에 따라 분류하는 검색을 통해 이루어진 것이므로 사용자가 유용한 정보를 획득하는 것이 쉽지 않다.

일반적으로 정보의 중요성을 나타내는 지표로 검색엔진 사용자나 방문객의 페이지 뷰, 링크 수, 추천 수, 페이지 랭크 방식 등이 있다. 그러나 이것은 일반적인 지표일 뿐 사용자가 원하는 정보에 대한 가치 판단과 반드시 일치하지는 않는다. 또한, 기존의 검색 기술은 낱말의 형태만을 가지고 찾는 기계적인 것으로 낱말의 가치를 평가하지 못할 뿐만 아니라 검색어가 웹 컨텐츠 내부에 존재하지 않으면 검색에서 제외되는 경우 문제점도 있다.

이러한 기계적인 검색 방법을 보완하기 위해 문서나 낱말에 가치를 부여해 정보 검색을 돕는 방법이 웹 컨텐츠에 태그(tag)를 붙이는 것이다. 컨텐츠에 부가정보를 부여하기 위해 태그를 붙이는 태깅(tagging)은 웹의 등장 이전부터 사용된 기술로 예전보다 다양한 분야에 적용되어 사용되고 있으며 인터넷을 통해 공유가 쉽다는 점이 장점이다. 그러나 단순히 컨텐츠에 태그를 붙이는 것만으로는 특정 엘리먼트가 포함된 속성에 대한 검색을 효과적으로 수행하기 어렵다. 특히, 많은 수의 키워드가 정의되어 있는 경우 복잡한 연산에 따른 자료관리 및 검색 성능을 저하 시킬 수도 있다.

학교나 회사와 같은 단위 조직에서 유사한 업무를 수행하는 작업자의 정보 검색 범주는 매우 유사하다. 예를 들면 회사, 학교의 특정 부서에서 구성원들이 업무와 관련된 정보를 검색하면 그 검색결과를 보고 그 유용성에 가치를 두어 시간, 선호순위에 태깅하여 둔다면 동일한 부서의 다른 구성원이 검색을 통해 얻을 수 있는 검색결과는 처음 검색할 때 보다 더 가치 있는 정보가 될 수 있다. 즉 부여된 태그를 가진 정보를 검색할 경우 태그가 없는 검색 결과보다 정확히 사용자에게 제공될 수 있다. 이는 첫 검색 시 태깅이라는 작업에 따라 오버헤드가 생기더라도 그 이후에 얻는 검색결과의 유용성으로 인해 오버헤드를 상쇄할 수 있다.

본 연구에서는 웹 컨텐츠의 중요성에 대한 개인의 판단 및 사용자 정보 이력을 공동의 관심사를 가지는 사용자 그룹이 공유하기 위해 다차원 벡터를 이용하여 인덱싱하는 태깅방법을 제안한다. 본 논문의 구성은 제2장에서 연구배경을 언급하고, 제3장에서 제안 방법의 적용과 개념적 구조를 제시하고, 제4장에서 제안방식의 순위측정에 대한 적합성을 검증하며, 제5장에서는 성능평가를 하고, 제6장에서 결론을 맺는다.

2. 웹의 순위제공 연구

적절한 검색결과에 대한 사용자의 요구로 페이지랭크(PageRank) 알고리즘이 제안되어 적합도 연산을 수행하면서 좋은 검색엔진이 많은 문서를 검색해주는 것이라는 인식이 바뀌었다. 페이지랭크 알고리즘의 특징은 검색과는 독립적으로 순위측정 시스템을 사용한다는 점이다[1]. 웹에서 가져온 문서들에 대해 미리 문서의 적합도를 연산해 놓고 검색시 연산값을 참조하는 방식으로 문서정렬 결과를 제공하여 검색속도가 빠르고 방대한 문서량을 소화할 수 있다. 이 방법의 장점은 이전에 적합도 연산 알고리즘으로 발표된 HITS알고리즘이 적합도 연산에 필요한 연산량이 많아 제한적인 환경에서 소량의 문서들에만 적용되던 것과 비교된다[2]. 웹문서의 수가 기하급수적으로 증가하면서 검색결과 중요도의 신뢰성이 저하되고 의도적인 중요도 조작이 문제가 된다. 문서의 중요도 혼란을 초래하는 문제로 대표적인 예는 사이트간 배너교환과 스팸문서의 생성이 있다.

페이지랭크의 안정적인 문서중요도 측정을 위한 해결책으로 주제특성을 이용하는 방법이 있다. 이는 오픈디렉토리에 의해 문서를 분류하고 문서에 출현하는 단어들의 각 카테고리에 대한 출현 빈도수를 계산하는 방법이다[3]. 검색시에는 추가적인 프로세스를 거쳐 검색어로 사용된 단어의 주제특징에 따라 페이지랭크에 의도된 값(bias)을 준다. 이러한 시도는 문서의 인기도에만 의존하는 문서 중요도 혹은 적합도의 척도에서 비롯되는 문제의 해결책으로 제시되었다[4]. 하지만 자동적인 추론을 원활하게 수행할 수 없는 단점이 있다.

한편, 검색엔진의 역색인(inverse index) 데이터베이스를 활용하여 문서의 중요도 혼란에 관한 문제를 해결할 수 있다는 연구도 있다[5]. 하지만 대다수 웹 검색엔진에 있어 역색인 데이터베이스와 순위연산을 수행하는 랭크 데이터베이스가 분산되어 있어 근원적인 해결책으로 사용될 수는 없다. 또 다른 방편으로 중요단어 사전을 구성하여 사전에 명시된 검색어를 사용하는 제안도 있다. 그러나 수 만개의 검색어에 대한 문서 랭크를 따로 연산해야 하는 문제가 동반되어 처리시간이 길어지는 약점을 보인다[6].

최근에는 검색시 추가 연산이 필요하거나 데이터베이스와 같은 지나치게 많은 저장공간을 필요로 하는 등의 문제극복을 위해 알고리즘 개선뿐만 아니라 의미처리를 할 수 있는 정보접근에 관한 연구가 활발하다. 웹에서 문서의 중요도를 안정적으로 측정하는 의미기반의 정보접근(semantic-driven information access)은 온톨로지를 바탕으로 하는 연구의 주류를 이루어왔다[7,8]. 이러한 접근방식은 자동화된 정보처리(automated informa-

tion processing), 정보통합(information integration), 지식관리(knowledge management) 등으로 알려져 있다. 대량의 정보를 자동화된 에이전트를 통해 처리될 수 있는 방법을 연구하게 되고 그 대표적인 것이 자동으로 주석을 다는 태깅이다. 태그는 정보자원을 의미적으로 정확히 분석하고, 관리할 수 있도록 문서나 웹 정보들에 주석을 붙여 사용자에게 더욱 정확한 정보를 제공할 수 있게 한다[9]. 본 연구에서는 다차원의 인덱스를 제공하기 위해 벡터형 태그를 효율적으로 구성하는 방법을 제안함으로써 신속한 의미기반의 정보접근이 가능하도록 하여 시맨틱 웹의 기반이 될 수 있다. 또한, 현재 온톨로지 관리와 추론을 향상시키기 위한 상당한 연구진척에도 불구하고 데이터베이스만큼 활성화되지 못한 문제해결에도 기여할 수 있다.

3. 다차원 인덱싱 기능의 벡터방식 태그정의와 처리구조

3.1 다차원 인덱싱 기능의 벡터방식 태그의 개념

웹에서의 정보 교환은 수신자가 송신자의 의도와 동일하게 정보를 해석한다는 조건에서 이루어진다. 이때 정보의 상호 운용성을 높일 수 있도록 메타데이터를 잘 활용하면 정확한 정보해석이 가능하다. 정확한 정보해석은 메타데이터를 정확하게 해석하는 것으로 웹에서의 정확한 정보검색 수행을 의미한다. 그러나 기계나 프로그램이 자동적으로 메타데이터의 의미를 이해하기는 어렵다. 이런 단점을 극복하기 위해서 웹 콘텐츠에 주석을 달게 된다. 본 논문에서는 주석을 다는 태깅 작업을 할 때 벡터를 이용한 정보의 유용성 표현방안으로 인덱싱 기능의 벡터형 태그를 도입한 연구[10]를 확장하여 다차원으로 정의할 것을 제안한다. 제안하는 다차원 인덱스 표현용 벡터형 태그의 형식은 다음과 같다.

KEYWORD [X, Y, Z]

제안하는 태그 방식은 검색 의도를 표현한 KEYWORD 부분과 정보의 가치를 표현한 [X, Y, Z] 부분으로 구성된다. KEYWORD 부분은 사용자가 어떤 검색정보에 대해 자신의 정보활용 가치를 표현한 태그를 부여할 때 색인기능을 겸한 검색의도를 표현한다. [X, Y, Z] 부분은 정보의 가치를 표현하며, 각 X, Y, Z는 가상의 벡터공간(vector space)에서 검색의도인 KEYWORD를 시작점으로 하는 위치벡터의 성분이다. 이것을 그림 1로 설명하면 벡터공간은 위치벡터의 각 성분으로 구성되는 좌표계이며, 이에 따른 위치벡터는 검색된 문서들이 얼마만큼 검색의도에 부합하는지를 표현한다.

그림 1의 가상 X, Y, Z축에서 X축은 사용자가 해당 콘텐츠의 활용 가치를 나타내고자 부여한 정보순위(rank)이고, Y축은 사용자가 생각하는 해당 콘텐츠의 시간속성을 나타낸다. 시간속성은 어떤 콘텐츠를 활용하려는 사용자가 체감하는 시간이다. 사용자의 시간속성을 부여하는 이유는 분야에 따라 10년 전 정보가 최신일 경우도 있고 1일된 정보가 사장되어야할 경우가 있기 때문이다. Z축은 다른 사용자들이 부여한 태그의 수치를 참조하는 값이다. 이 값은 태그가 사용자의 주관적 평가에 의해 부여되므로 고의성을 배제하기 위해 사용한다. 이로 인해 태그를 부여하는 사용자가 많을수록 신뢰성을 증가시킬 수 있다.

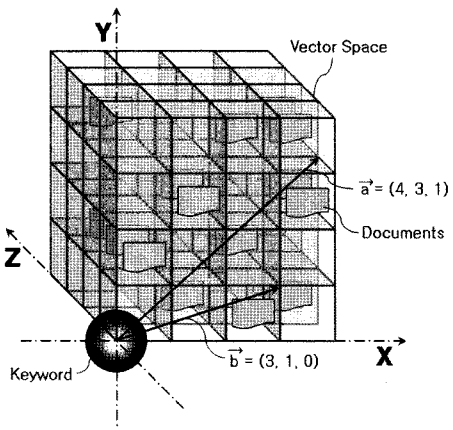


그림 1 태그정의에 다차원 벡터를 이용하는 인덱싱구현 개념도

태그의 위치벡터가 가지는 벡터량에서 방향성을 제외한 성분 [X, Y, Z]로 알 수 있는 스칼라에 의해 검색된 정보의 가치를 표현할 수 있다. 정보가치는 검색된 정보가 정보검색자의 검색의도에 얼마만큼 근접한지를 표현한다. 검색의도에 대한 근접 표현은 그림 1에서 볼 때 벡터공간이 가상 X, Y, Z축으로 설명되는 좌표계를 이용해 위치벡터를 방향이 있는 선으로 나타낼 수 있어 가능하다. 여기서 핵심어는 정보 검색의 용도 및 분야를 표현할 수 있는 단어를 사용자가 부여한 것으로 위치벡터의 시작점이며, 각 좌표는 사용자가 평가한 수치로 위치벡터의 끝점이 된다. 따라서 이를 통해 산출될 수 있는 벡터 크기가 검색의도와의 거리가 된다. 예를 들어 그림 1에서 웹페이지가 키워드가 "온톨로지"라는 태그를 가진다고 가정하여 두 가지 벡터 값을 구하면 $\vec{a}$ =(4, 3, 1), $\vec{b}$ = (3, 1, 0)가 된다. 여기서 벡터 $\vec{a}$, $\vec{b}$가 가진 값은 X성분은 해당 콘텐츠의 활용 순위, Y성분은 해당 콘텐츠의 시간 속성, Z 성분은 다른 사용자

들이 참조하는 태그의 수치를 참조하는 값으로부터 구할 수 있고 벡터 $\vec{a}$, $\vec{b}$ 중 어느 것이 검색의도에 더 근접한 가치 있는 정보인지를 알기위해서 3.3절에서 논의할 내용에 따라 수식을 통해 벡터 값을 구해보면 벡터 $\vec{b}$를 가지는 태그가 검색의도에 더 근접한 가치 있는 정보임을 알 수 있다. 즉 값이 작을수록 키워드에 가깝게 되고 이를 통해 원하는 정보에 근접해 있음을 직관적으로 알 수 있다.

3.2 제안방식을 이용한 검색결과의 순위측정

제안한 다차원 벡터형 태그를 바탕으로 하는 검색결과의 순위측정은 3가지 벡터성분, 즉 정보순위, 정보시간, 참조값에 따라 수행한다. 검색의도에 가깝고 가치가 있는 정보를 검색결과의 상위에 유지하기 위해서 각 벡터성분으로 산출되는 벡터의 크기를 통한 양적비교를 수행한다. 위치벡터는 가상의 X, Y, Z축이 직교하는 벡터공간에서 만들어진다(그림 1 참조). 그러므로 양적비교를 위한 위치벡터의 크기(길이) R은 식 (1)과 같이 정립된다.

$$R=\sqrt{\alpha x^2+\beta y^2+\gamma z^2}. \quad (1)$$

식 (1)에서 x, y, z는 각 벡터의 성분값이며, 표준바이어스로 α, β, γ값을 사용한다. R값이 작을수록 검색의도에 근접하는 정보이다. 동일한 웹 콘텐츠를 여러 사람들이 열람을 하고 태그를 부여하면 1개의 콘텐츠에 여러 개의 벡터형 태그가 선언된다. 이때 순위를 측정하기 위해서는 벡터량 R을 누적하여 사용해야 한다. 따라서, 식 (1)을 여러 개의 태그가 적용될 경우로 변형하여 개인별 [X, Y, Z]에 따른 각 벡터 성분을 누적하는 종합적 정보유용성(V)을 나타내면 식 (2)와 같다.

$$V=\frac{\sum_{i\in T_N}\sqrt{X_t^2+Y_t^2+Z_t^2}}{N_{total}}. \quad (2)$$

식 (2)에서 T_N은 동일한 카테고리로 분류될 수 있는 N개의 태그집합을 말하며, i는 N개의 태그들 중 하나이다. X_t는 태그를 정의한 어떤 시점(t)의 x성분값이고, Y_t는 태그를 정의한 어떤 시점(t)의 y성분값이다. 또한 Z_t는 태그를 정의한 어떤 시점(t)의 z성분값이다. 이들 값들이 누적되는 위치벡터의 크기가 정보유용성(V)을 나타내게 된다. 이때 누적된 모든 벡터량은 참조한 전체 태그개수(N_{total})로 나눈 값을 취한다.

검색엔진을 통해 전달받은 검색질의 결과는 링크기반 순위이다. 여기에 사용자가 부여한 벡터기반 순위연산을 적용하여 검색이력으로 유지하게 된다. 이것을 종합하여 정리하면 (3)과 같이 나타낼 수 있다.

$$PR(p)=\frac{(1-\delta)V}{N_{total}}+\delta\times\sum_{i=1}^{n}\frac{PR(W_i)}{C(W_i)}. \quad (3)$$

웹 환경에서는 한 사용자가 동일한 정보에 대해 다른 사용자들과 동시에 벡터형 태깅을 하게 된다. 이 때 부여된 다른 사용자들의 바이어스를 연산에 포함시킨다. 연산에 포함시킬 바이어스는 $(1-\delta)/N_{total}$ 값으로 사용하며, 이는 다른 검색자들의 비율($1-\delta$, $0\le\delta\le1$)을 전체 N_{total} 개의 태그에 대한 비로 나타낸 것이다. 이렇게 연산에 의해 벡터성분을 축적하고 바이어스를 포함하면 링크에만 의존하던 순위측정 오류를 최소화할 수 있다. 또한 바이어스는 카테고리에 따른 정보이용성을 추론할 수 있는 자료로 활용될 수 있어 정보의 유용성 표현에 좋은 지표가 된다. 한편 $PR(W_i)$는 웹페이지 W_i로부터의 하이퍼링크수를 나타내고, $C(W_i)$는 웹페이지 W_i가 가지는 순위 값이다.

3.3 제안방식의 운용시스템 구성

제안방식의 운용시스템은 인터페이스, 메타엔진, 주석기(annotator), 연산기(operator), 데이터베이스 등으로 구성되며 외부의 검색엔진을 연동한다. 시스템의 구조는 그림 2와 같다. 제안 방식은 링크기반의 순위 연산에 부가적인 연산을 요구하면서도 시스템 자원의 소요가 적고, 주제특성과 비슷한 효과를 낼 수 있는 효과적인 방식임을 알 수 있다. 특히 4000개 이내의 제한된 영역에서 정보활동에 높은 효율을 낼 수 있다는 것을 실험을 통해 알 수 있다.

사용자 인터페이스에서는 시스템과 사용자간 의사교환이 수행된다. 메타엔진은 검색질의를 수용하여 외부의 검색엔진에 검색을 의뢰한다. 주석기는 기본적인 검색결과에 사용자가 부여하는 태깅을 수행한다. 연산기는 사용자가 부여한 태그를 기반으로 순위를 측정하며, 순위측정 결과를 사용자가 전달받게 된다. 검색질의를 하고 최종 검색결과를 제공받는 과정은 7단계의 절차를 통해서 이루어지며, 그 과정을 그림 2에 따라 각 단계별로 자세히 설명하면 다음과 같다.

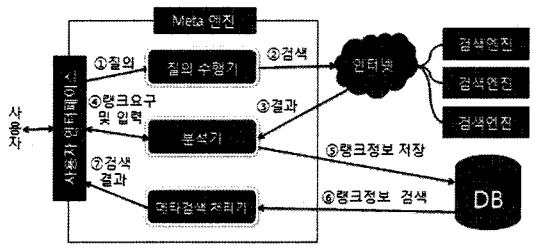


그림 2 제안방식의 운용시스템 구조

- 1단계: 사용자가 검색하고자 하는 요구질의를 인터페이스를 통해 질의수행기에 전달한다. 이 단계에서 검색이 시작된다.

- 2단계: 질의수행기는 검색자의 질의를 조합하여 검색엔진에 의뢰한다. 검색질의를 넘기는 것과 동시에 검색엔진이 웹 문서를 검색한다.
- 3단계: 웹에서 검색된 콘텐츠들의 목록을 분석기에 전달한다. 이때 전달되는 문서들은 링크구조에 의해 측정된 순위를 가지고 있다.
- 4단계: 전달받은 검색결과 목록을 사용자에게 전달하여 랭크를 요구하며 사용자는 이 요구에 대해 콘텐츠에 중요도 검토하여 랭크를 부여한다.
- 5단계: 분석기는 그림 3에 나타낸 의사코드와 같이 사용자가 부여한 태그를 기반으로 순위를 측정하여 다차원 태그기반 랭크정보 DB에 저장한다.
- 6단계: 메타검색 처리기는 DB로부터 랭크정보를 검색한다.
- 7단계: 메타검색 처리기는 검색한 랭크 정보를 이용하여 순위를 부여하여 검색결과를 최초 검색질의를 요구한 사용자에게 사용자 인터페이스를 통해 제공한다.

그림 3은 제안된 방식에 따른 순위측정을 위한 의사코드를 나타낸 것으로 벡터 값 X, Y, Z를 이용하여 먼저 해당 콘텐츠에 대한 R을 구하고, 해당 콘텐츠에 다수의 태그가 부여되어 있으면 부여된 태그 수로 R을 나누어 V를 구한다. 만약 부여된 태그가 1개라면 R을 V로 확정한다. V가 구해지면 정렬을 수행한다.

```
Rank_List( )
{
        input Vector value X, Y, Z;
        R = sqrt(X^2+Y^2+Z^2); //Calculate Vector value
        if contents have other tags
        {
                V = sum R / number of tags;
        }
        else
        {
                V = R;
        }
        compare V of contents;
        sorting;
        return;
}
```

그림 3 순위측정 의사코드

4. 벡터태그 방식에 의한 순위측정의 적합성 평가

본 절에서는 제안방식을 실재 웹문서에 적용하여 측정순위에 대한 적합성을 검증하고자 한다. 제안방식이 가지는 순위측정 효과는 앞서 언급한 것처럼 대표적인 링크기반의 순위측정 방법인 페이지랭크 알고리즘과 비교한다. 측정순위 적합성 검증을 위한 분야는 Plasmid DNA 관련분야로 한정한다. 실험용 문서는 해당분야의 mini-prep protocol에 관련된 웹문서 20여 종을 무작위로 선발하여 표 1에 나타내었다.

표 1 제안방식에 의한 측정순위의 적합성 검증에 사용한 웹문서

No.	URL
1	http://cat.inist.fr/?aModele=afficheN&cpsidt=4181477
2	http://openwetware.org/wiki/Miniprep/Kit-free_high-throughput_protocol
3	http://people.morehead-st.edu/fs/d.peyton/protocols.html
4	http://sosnick.uchicago.edu/DNA_miniprep.html
5	http://userwww.service.emory.edu/~kressle/protocols/BAC%20miniprep%20protocol.doc
6	http://wolverton.owu.edu/lab/2006/01/eppendorf-miniprep-protocol
7	http://www.bioinformatics.vg/Methods/miniprep.shtml
8	http://www.bio.brandeis.edu/haberlab/jehsite/pdfs/ZymoPrep.pdf
9	http://www.bio.indiana.edu/~chenlab/protocols/qiagenmini.pdf
10	http://www.bio.net/hypermail/chlamydomonas/1993-December/000121.html
11	http://www.genetics.ucla.edu/labs/fan/Protocols_Miniprep.htm
12	http://www.genome.arizona.edu/agi/seq/QIAprep%20Spin%20Miniprep%20Kit%20Protocol.doc
13	http://www.genomed-dna.com/pdf/Quick-PDFs/Protocol%20Quick-Plasmid%20(Vacuum).pdf
14	http://www.gerardbiotech.com/documents/Protocols/GerardBiotech_HurricaneMiniPrep250_gbt010705.pdf
15	http://www.mbi.ufl.edu/~rowland/protocols/minipreps.htm
16	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7946306&dopt=Abstract
17	http://www.protocol-online.org/prot/Molecular_Biology/Plasmid/Miniprep/more2.html
18	http://www.tracy.k12.ca.us/thscdna/miniprep.html
19	http://www.umich.edu/~wakil/protocols/miniprep.html
20	http://www1.qiagen.com/HB/QIAprepMiniprep

평가실험은 실험용 문서(표 1)에 페이지랭크 알고리즘을 적용한 링크기반 측정순위와 제안방식에 의한 측정순위의 비교로 수행한다. 실험 환경은 식 (3)의 정의에 있어 $\alpha=\beta=\gamma=1$, $\delta=0.5$로 한다. 또한, 제안방식에서 태그 정의를 위한 벡터공간(10×10×10)을 정보순위 10등위(X성분)와 정보시간 10등위(Y성분)로 구성하고 참조값(Z성분)을 취한다. 참조값이 없을 경우는 기본값으로 5를 설정한다. 측정된 문서순위의 적합성검증은 현재 DNA 관련 연구에 참여하고 있는 영남대학교, 생명공학연구원의 연구원 35인을 수행평가단으로 활용하였다.

검증은 평가대상 웹문서(표 1)에 대해 제안방식과 링크기반으로 측정한 각 순위를 수행평가단이 5점 등간척도(부적절→적절)에 따라 적절성을 평가하였다.

표 2에서 보면 페이지랭크에 의한 문서순위와 제안방식의 문서순위가 서로 현저한 차이를 보인다. 모집단의 평가조사에서 제안방식에 의한 웹문서의 측정순위가 적합한 것으로 나타났고, 각 문서의 순위에 대한 지지율(Accuracy)은 평균 79.8%에 달했다. 표 2에 따르면 문서 12는 양쪽의 측정순위가 동일함을 보였으나 그 외에는 문서 7, 17, 18 만이 다소 유사한 경향을 보였다. 한편, 불필요한 정보로 제외시킨 문서 1, 4, 9, 16, 19는 99.4%가 삭제에 동의하였다. 여기서 말하는 지지율(Accuracy)은 제안된 방식에 따른 웹 문서 순위의 적절성을 나타내는 비율로 실험에 참가한 연구원들이 제시된 순위의 적절성 여부에 대해 비율로 나타낸 것이다. 지지율이 높을수록 제안된 방식으로 제시된 웹 문서의 순위 적절하다는 것을 나타낸다.

그림 4는 표 1의 웹문서가 가지는 정보의 순위를 제안방식과 링크기반에 따른 각각의 순위를 비교 평가한 그래프이다. 수행평가단이 5점 등간척도에 따라 평가한 각 순위의 평가지수에 로그를 취한 추세선을 비교해 보면 거의 1점 척도만큼 제안방식이 우수하다.

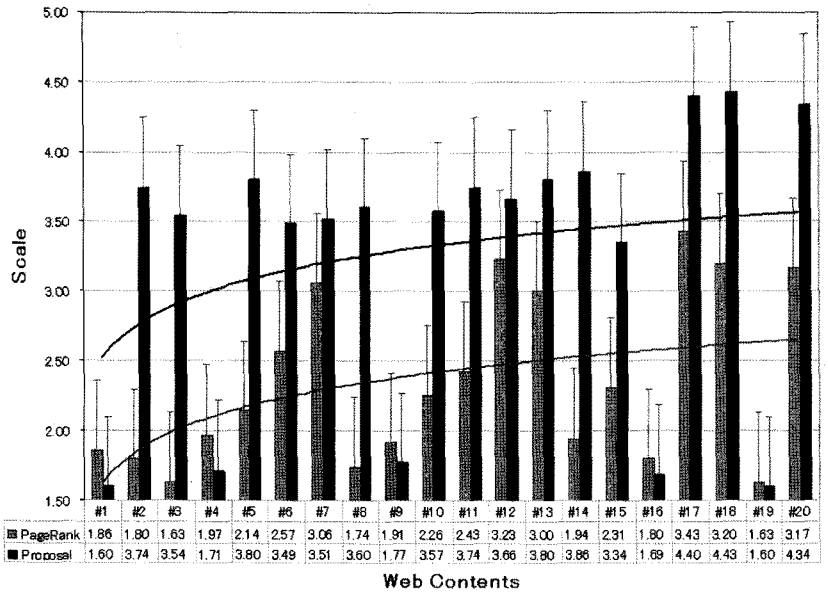
또한, 제안방식에서 사용자들이 불필요한 문서로 결정한 문서 1, 4, 9, 16, 19에서만 링크기반의 순위가 제안방식에 비해 근소하게 앞선다. 이는 링크기반에 의한 순위측정이 불필요한 정보에 대해 관대하다는 약점을 지적하는 증거이다. 경우에 따라 오차범위 내에서 평가가 근소하게 앞서는 문서(문서 7, 12)도 있으나 링크기반 순위에 비해 제안방식이 정보 중요도를 표현하는 순위측정에 적합함을 알 수 있다(그림 4).

제시한 표 1의 문서에 대한 문서순위척도의 통계량은 평균 136.75, 분산 223.566, 표준편차 14.952를 보였다. 각 문서 순위측정의 일관성이나 동질성 정도를 측정하는 척도화 분석에서 내적 일관성(internal consistency) 등간척도를 구성하고 있는 각 문서들의 신뢰성을 측정한 Chronbach's α계수는 0.83이었다. 신뢰도는 측정도

표 2 제안방식의 순위측정결과와 링크기반(페이지랭크)의 순위측정결과 비교

No.	URL	PageRank	Proposed Rank	Accruacy	Remark
1	http://cat.inist.fr/?aModele=afficheN&cpsidt=4181477	16	-	1.00	*
2	http://openwetware.org/wiki/Miniprep/Kit-free_high-throughput_protocol	17	2	0.83	
3	http://people.morehead-st.edu/fs/d.peyton/protocols.html	18	1	0.89	
4	http://tosnick.uchicago.edu/DNA_miniprep.html	10	-	1.00	*
5	http://userwww.service.emory.edu/~kressle/protocols/BAC%20miniprep%20protocol.doc	5	14	0.74	
6	http://wolverton.owu.edu/lab/2006/01/eppendorf-miniprep-protocol	11	9	0.77	
7	http://www.bioinformatics.vg/Methods/miniprep.shtml	4	3	0.80	
8	http://www.bio.brandeis.edu/haberlab/jehsite/pdfs/ZymoPrep.pdf	19	4	0.86	
9	http://www.bio.indiana.edu/~chenlab/protocols/qiagenmini.pdf	2	-	0.97	*
10	http://www.bio.net/hypermail/chlamydomonas/1993-December/000121.html	7	15	0.71	
11	http://www.genetics.ucla.edu/labs/lenu/Protocols_Miniprep.htm	14	7	0.77	
12	http://www.genome.arizona.edu/legi/seq/QIAprep%20Spin%20Miniprep%20Kit%20Protocol.doc	12	12	0.83	
13	http://www.genomed-dna.com/pdf/Quick-PDFs/Protocol%20Quick-Plasmid%20(Vacuum).pdf	15	13	0.83	
14	http://www.gerardbiotech.com/documents/Protocols/GerardBiotech_HurricaneMiniPrep250_gbt010705.pdf	20	10	0.77	
15	http://www.mbi.ufl.edu/~rowland/protocols/minipreps.htm	1	8	0.80	
16	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7946306&dopt=Abstract	13	-	1.00	*
17	http://www.protocol-online.org/prot/Molecular_Biology/Plasmid/Miniprep/more2.html	3	5	0.80	
18	http://www.tracy.k12.ca.us/thecdna/miniprep.html	9	6	0.71	
19	http://www.umich.edu/~wakil/protocols/miniprep.html	8	-	1.00	*
20	http://www1.qiagen.com/HB/QIAprepMiniprep	6	10	0.86	

*: 불필요한 문서로 분류되어 제안방식의 순위측정에서 제외된 웹문서



	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20
PageRank	1.86	1.80	1.63	1.97	2.14	2.57	3.06	1.74	1.91	2.26	2.43	3.23	3.00	1.94	2.31	1.80	3.43	3.20	1.63	3.17
Proposal	1.60	3.74	3.54	1.71	3.80	3.49	3.51	3.60	1.77	3.57	3.74	3.66	3.80	3.86	3.34	1.69	4.40	4.43	1.60	4.34

그림 4 제안방식과 링크기반 순위측정간 문서별 순위 평가 비교

구의 정확성이나 정밀성을 나타내는 것으로 신뢰도 분석은 동일한 개념을 독립된 측정 방법에 의해 측정한 경우 결과가 동일하게 나타나야 한다는 것을 전제로 한다. 따라서 Chronbach's α계수가 1에 가까운 0.83으로 신뢰도가 상당히 높아 순위측정에 대한 제안방식의 측정순위가 적합함을 뒷받침한다.

5. 순위연산에 대한 성능 평가

본 절에서는 제안한 방식을 실재 검색모델에 적용할 것을 고려하여 순위 연산에 대한 성능을 기존 방식과 비교하여 평가한다. 비교대상인 기존 방식은 페이지 랭크 알고리즘과 주제특성 페이지 랭크 알고리즘으로 한다. 제안하는 방식은 실제 웹 데이터에 벡터방식의 순위 측정을 할 수 있는 태그를 부여한 가상 웹 환경을 구축하고 이를 이용한다.

실험을 진행할 가상의 웹 환경은 스탠포드 매트릭스를 기반으로 8,000개의 사이트를 구축하고 성능평가로 도구로는 매트랩(MATLAB ver. 2007a)을 이용하였다. 제안방식, 페이지랭크, 주제특성 페이지 랭크의 세 가지 방식에 대한 성능평가를 위한 비교대상 파라미터는 가상 웹 환경에서 전체사이트에 대한 순위연산의 수행시간이다.

먼저 제안방식과 페이지 랭크 알고리즘을 비교한다. 페이지 랭크 알고리즘은 본 연구에서 제안한 방식의 기본적인 시스템 자원의 소요를 측정할 수 있다. 다음으로 제안방식과 주제특성 페이지 랭크 알고리즘을 비교한다.

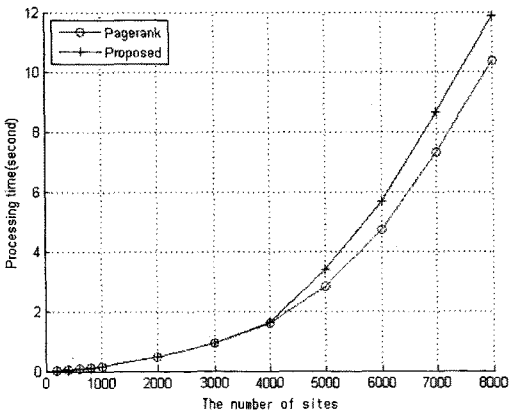


그림 5 제안방식과 페이지랭크 알고리즘의 처리 시간 비교

주제 특성 페이지 알고리즘은 제안 방식과 동일하게 링크구조의 순위 연산을 기본으로 채택하고 있어 유사한 연산구조를 갖는 시스템간 비교를 수행할 수 있다.

그림 5는 제안 방식과 페이지랭크 알고리즘간 성능평가를 수행한 결과를 나타낸다. 결과에서 알 수 있듯이 제안방식은 페이지 랭크 알고리즘에 비해 다소 높은 연산처리 시간을 소요하지만, 두 시스템의 전반적인 시스템 부하는 유사한 경향을 보인다. 이는 제안방식이 벡터 성분을 기반으로 단순 비교 연산을 수행하므로 전체적인 연산량에 큰 영향을 미치지 않는 다는 것을 입증한다. 특이한 점은 4,000개 미만 사이트에서는 거의 동일한 성능을 보이는데 이는 제안방식이 사용자의 참여가

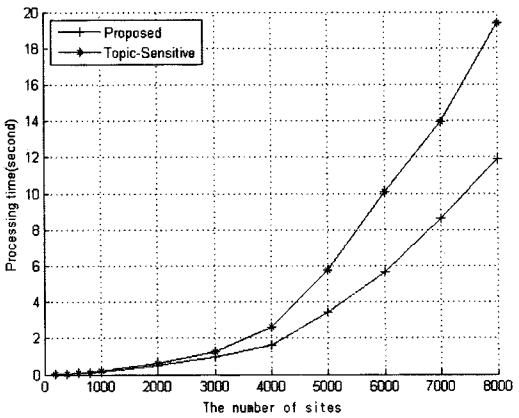


그림 6 제안방식과 주제특성 페이지랭크 알고리즘의 처리시간 비교

이루어지는 한정된 영역에서 정보활동이 수행되므로 유용하다.

한편, 그림 6은 제안 방식과 주제특성 페이지 랭크 알고리즘 간 성능 평가를 수행한 결과를 나타낸다. 결과에서 알 수 있는 것처럼 주제특성 페이지 랭크 알고리즘이 전반적으로 제안 방식에 비해 많은 시스템 부하를 유발한다. 주제 특성 페이지랭크 알고리즘과 유사한 연산구조를 가지면서 제안방식이 시스템 자원을 덜 소모한다는 것을 알 수 있다. 특히 사이트 수가 5,000개를 넘어가면서 점점 더 큰 차이를 보이고 있다. 그러나 3,000개의 사이트 이내에서는 그리 큰 차이를 보이지 않으므로 한정된 영역에서는 시스템 간 성능이 큰 문제가 되지 않는다.

제안 방식은 링크기반의 순위 연산에 부가적인 연산을 요구하면서도 시스템 자원의 소요가 적고, 주제특성과 비슷한 효과를 낼 수 있는 효과적인 방식임을 알 수 있다. 특히 4000개 이내의 제한된 영역에서 정보활동에 높은 효율을 낼 수 있다는 것을 실험을 통해 알 수 있다.

6. 결론 및 추후 연구

본 연구는 인터넷의 정보를 이용할 때 등록된 태그(tag)기반의 사용자 평가를 검색에 활용할 수 있도록 다차원 벡터를 이용한 정보의 유용성 표현을 제안하는 연구이다. 다양한 시스템으로 구성되는 웹에서 정보에 대한 즐겨찾기 등을 유지하는데 제안한 방식을 활용하면 정보의 분류·관리·재활용에 효율성을 기대할 수 있다. 그 이유는 색인으로 사용할 핵심어가 태그로 정의되어 있어 재검색을 위한 색인기능을 제공하고 정보검색 결과에 대해 벡터성분의 계산으로 순위를 정하기 때문이다. 이 기능은 특정 정보의 속성에 대한 직접적인 접근 없이 웹 검색 메타엔진들이 개인화된 평가정보를 검색할 수 있도록 도움을 준다. 또한, 유사한 콘텐츠에 대한 수요자들의 접근효율을 높인다.

본 연구의 결과는 소규모 동일 관심 그룹에서 특히 그 효과를 발휘할 수 있다. 일반적인 상황에서 효과를 보기 위해서는 다양한 분야에 대한 메타엔진 클러스터가 생성될 필요가 있다. 향후에는 다차원 벡터형 태그기반의 자동 주석 시스템을 위해 시멘틱 웹 기반 메타 데이터 레지스트리(Metadata Registry)[11] 설계를 고려한 데이터 요소의 개념과 표현에 따른 시맨틱 마크업 언어의 정의적인 부분을 연구할 것이다. 또한, 시맨틱 웹 환경에서 검색 정보에 대한 브라우징 시스템을 구현하는데 제안방식이 적용될 수 있도록 할 것이다.

참 고 문 헌

[1] Richardson, M. and P. Domingos, "The intelligent surfer: Probailistic combination of link and content information in pagerank," In *Advances in Neural Information Processing Systems*, volume 14, MIT Press, 2002.

[2] Haveliwala, T.H. "Topic-sensitive pagerank," In *Proceedings of the elveventh international conference on World Wide Web*, pp.517-526, ACM Press, 2002.

[3] "ODP-Open Directory Project," http://dmoz.org/

[4] Sullivan, D. "More Evil Than Dr. Evil?," 1999, http://searchenginewatch.com/sereport/article.php/2167621

[5] Brin, S. and L. Page, "The anatomy of a large-scale hypertextual web search engine," In *Proceedings of the 7PthPInternational World Wide Web Conference*, pp.107-117, Elsevier Science, 1998.

[6] "Theme-based PageRank," http://pr.efactory.de/e-pagerank-themes.shtml

[7] 이재호, "시멘틱 웹의 온톨로지 언어", *정보과학회지*, 21권, 3호, pp.18-27, 2003.

[8] 최호섭, 옥철영, "정보검색 시스템과 온톨로지," *정보과학회지*, 22권, 4호, pp.62-71, 2004.

[9] Ovsiannikov, I.A., "Annotation Technology," International Journal of Human-Computer Studies vol.50(4), 1999.

[10] Jung, J.Y., S.H. Kim, and C.G. Kim, "A Study of Personal Ranking Vector for Searched Information on Web," *The 4th Conference on New Exploratory Technologies*, pp.231-234, 2007.

[11] "ISO/IEC 11179 - Metadata Registry(MDR)," http://metadata-stds.org/11179/

정 재 윤

1998년 영남대학교 농학과(학사). 2000년 영남대학교 농학과(석사). 2005년 영남대학교 농학과(박사). 2008년 영남대학교 컴퓨터공학과(박사). 현재 영남대학교 컴퓨터공학전공 연구원. 관심분야는 BIT융합기술, 시뮬레이션, 웹기반 기술, 온톨로지, 환경생태 제어

진 현 철

1994년 영남대학교 전자공학과(학사). 2001년 경북대학교 교육대학원(석사). 2009년 영남대학교 대학원 컴퓨터공학과(박사과정). 관심분야는 모바일 네트워크, USN, 클라우드 컴퓨팅

김 종 근

1981년 영남대학교 전자공학과(학사). 1987년 영남대학교 전자공학과(석사). 1991년 (일본) 전기통신대학 박사. 1997년(미국) Virginaia Tech. 연구교수. 2003년(미국) UCSC 연구교수. 현재 영남대학교 컴퓨터공학전공 교수. 관심분야는 컴퓨터 네트워크, 무선 모바일 네트워크, 분산처리, 운영체제, 멀티미디어기반 가상강의 시스템, USN, 클라우드 컴퓨팅