마크업 패턴을 이용한 웹 검색

(Web Information Retrieval Exploiting Markup Pattern)

김 민 수 † 김 민 구 ††
(Minsoo Kim) (Minkoo Kim)

요 약 HTML은 웹 페이지의 시각적 표현을 목적으로 하고 있기 때문에, HTML로 작성된 웹 문서에 대한 색인과 질의는 쉬운 문제가 아니다. 그러나 웹 페이지를 표현하는 태그들이 가진 내재적 의미들은 검색 엔진의 성능을 향상시킬 수 있는 가능성을 가지고 있다. 본 논문은 이러한 HTML 태그의 내재적 의미를 이용하기 위해 마크업 패턴을 정의하고, 이를 웹 검색에 응용함으로서 검색 성능을 향상하고자 한다. 마크업 패턴은 웹 페이지 작성자의 표현 의도를 담고 있으며, 명시적으로 하나 이상의 HTML 태그의 연속으로 표현된다. 웹 페이지에서 마크업 패턴을 찾아내고, 이를 웹 검색에 응용하기 위해 본 논문에서는 웹 문서를 재색인하는 방법을 제안한다. 제안하는 방법을 적용한 웹 검색의 성능 향상을 증명하기 위해, BBC와 CNN 웹 사이트의 문서들을 대상으로 실험을 진행하였다. 대상 문서들은 제안한 방법을 통해 가중치를 갖게 되며, 특정 질의에 대한 정확도를 기존 검색 엔진과 비교하여, 본 논문에서 제안하는 마크업 패턴을 이용한 웹 검색의 성능 향상을 증명할 것이다.

키워드 : 마크업 패턴, 웹 검색, 정보 검색

Abstract Over the years, great attention has been paid to the question of exploiting inherent semantic of HTML in the area of web document retrieval. Although HTML is mainly presentation oriented, HTML tags implicitly contain useful semantics that can be catch meaning of text. Focusing on this idea, in this paper we define 'markup pattern' and try to improve performance of web document retrieval using markup patterns. Markup pattern is a mirror of intends of web document publisher and an internal semantic of text on web document. To discover the markup pattern and exploit it, we suggest a new scheme for extracting concepts and weighting documents. For evaluation task, we select two domains-BBC and CNN web sites, and use their search engines to gather domain documents. We re-weight and re-score documents using proposed scheme, and show the performance improvement in the two domains.

Key words : Markup Pattern, Web Document Retrieval, Information Retrieval

1. 서 론

오늘날 웹은 방대한 양의 정보를 가지고 있으며, 이러한 정보는 컴퓨터 과학 뿐 아니라 다양한 분야에서 널리 활용되고 있다. 웹이라는 분산 환경에 존재하는 이 정보들은 HTML(Hyper Text Markup Language)를 기반으로 표현되어 있다. HTML은 정보를 어떻게 표현하는가의 방법을 제공하는 것이기 때문에, HTML로 쓰여진 문서들 속에서 정보들을 찾고 걸러내는 일은 쉬운 일이 아니다. 이를 위해 웹 검색 분야에서 웹에 존재하는 유용한 정보를 찾기 위해 다양한 방법들이 제안되었고, 특히 HITS, Page-Rank 알고리즘 등과 같이 HTML의 특성을 파악하여 검색에 응용하는 지능적인 방법들은 긍정적 평가를 받고 있으며, Google 등의 웹 검색 엔진에 활용되고 있다.

그러나 긍정적 평가를 받고 있는 검색 엔진이라 하더라도, 일반적으로 하나의 질의에 대해 많은 결과 문서를 사용자에게 보여주며, 이들 중 많은 수는 사용자 질의와 관계없는 문서들이다. 많은 연구들은 이 문제에 대해 두 가지 큰 원인을 분석하고 있는 데[1-3], 첫 번째는 웹 문서에 대한 색인의 어려움이며, 두 번째는 사용자 질의의 모호함이다.

HTML을 기반으로 작성된 웹 문서는 사용자에게 보이는 내용 이외에도, 표현을 위한 태그와 메타 정보들을 포함하고 있기 때문에, 이 중에서 필요한 정보를 뽑아내기 쉽지 않다. 사용자가 웹 문서에서 보게 되는 하나의 문장은 실제 표현되는 단어와 많은 태그들이 혼재된 형태로 존재할 수 있으며, 따라서 이는 문장이 아니라 각각의 단어로 분석되고 색인될 수 있다. 이는 결국 검색 결과의 부정확성의 원인이 되며, 검색 성능의 저하를 가져온다.

· 본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅및네트워크원천기반기술개발사업의 지원에 의한 것임
· 이 논문은 2007 한국컴퓨터종합학술대회에서 '마크업 패턴을 이용한 웹 검색'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 아주대학교 정보통신학
visual@ajou.ac.kr
†† 종신회원 : 아주대학교 정보통신대학 교수
minkoo@ajou.ac.kr

논문접수 : 2007년 9월 28일
심사완료 : 2007년 11월 8일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제13권 제6호(2007.11)

또한 모호한 사용자 질의는 정확한 검색을 보장하는 엔진의 경우라 하더라도, 검색의 정확도를 낮추는 원인이다[4][5]. 이 문제의 해결을 위해 질의 확장 및 수정을 위한 다양한 방법이 제안되으며, 특히 Udo의 연구[6]는 웹 문서의 특성을 규정하고, 마크업 언어의 중요한 개념을 파악하여 문서로부터 용어들을 추출하고 이를 질의 확장에 응용하였다.

본 논문도 Udo의 연구와 같은 동기에서 출발한다. 즉 일반적 지식은 웹 검색을 위해 부적절하고, 특정 문서에 존재하는 지식은 그 문서를 구성하는 마크업 언어의 특성으로 묘사될 수 있다는 것이다. 즉 웹 문서의 내용은 특정 마크업 언어에 의해 강조되거나 특징지어지고, 이를 이용하여 웹 검색 성능을 향상시킬 수 있다는 것이다. 본 논문에서는 이러한 특징을 이용하여 질의를 확장/수정하는 것이 아니라, 웹 문서의 가중치 재설정을 통한 성능 향상을 꾀한다. 이를 위해 마크업 패턴을 정의하고, 검색 성능 향상을 위한 방법들을 제안한다.

2. 웹 문서의 마크업 컨텍스트

마크업 패턴에 대해 기술하기 전에, 이미 존재하는 웹 검색 및 관련 연구를 살펴보는 것이 연구의 이해에 도움이 될 것이다. 먼저 웹 문서에서 용어를 추출하기 위해 HTML 태그를 이용하는 다수의 연구가 존재한다. Jonathan은 그의 논문[1]에서 HTML 태그들이 가지고 있는 의미를 분석함으로서, 태그들을 이용한 의미 검색(Semantic Retrieval)의 가능성을 제기하였다. 그는 일반적 문서 검색과 웹 검색을 비교하면서, Link 태그, Heading 태그, comment 태그 등이 웹 검색의 성능을 향상시킬 수 있음을 증명하였다. 특히 'Anchor Text'는 웹 검색 분야의 많은 연구가들에 의해 주목받았다[4,7]. 전통적으로, 'Anchor'는 웹 문서의 중요한 특징으로 여겨졌는데, 웹 검색 성능의 뛰어난 향상을 가져온 HITS [7], Page-Rank[8] 알고리즘은 'Anchor'를 응용한 중요한 성과이다. 또한 'Anchor'는 사용자 질의와 유사하다는 분석도 제시되었고, 이를 이용해 질의 확장 및 수정 기법들이 제안되기도 하였다. 이러한 연구들은 웹 검색 성능의 향상을 가져왔고, 본 논문에서 정의하는 마크업 패턴 역시 'Anchor'의 특징을 내포하고 있다.

둘째로, 웹 문서로부터 용어를 추출하고, 용어 사이의 계층 구조를 발견하는 것은 일반 문서 검색 뿐 아니라 웹 검색 분야에서도 중요한 주제의 하나로 연구되어 왔다. 용어의 계층 구조는 문서에 존재하는 정보를 기계가 이해할 수 있는 형태로 제공하고 중요한 정보를 구조화하기 때문에 검색의 성능에 중요한 영향을 미친다. 그러나 HTML의 특성은 웹 문서로부터 용어들을 추출하고 계층 구조를 구축하는 것을 어렵게 하는 요인으로 작용하였다. 이를 극복하기 위해, Udo[6]는 웹 문서에 존재하는 용어의 계층 구조를 구축하기 위한 새로운 방법을 제안하였다. 그가 제안한 방법은 용어의 의미를 고려하는 것이 아니라, 웹 문서의 마크업 정보를 이용한 것이다. 즉 웹 문서에서 강조되고 있는 내용들을 위해 사용된 태그들을 '마크업 컨텍스트(Markup Context)'로 정의하고, 마크업 컨텍스트로 강조된 용어를 '개념(Concept)'으로 정의하였다. 한 문서에서 발견된 개념들은 서로 연결되어 계층 구조를 형성하고, 특정 사이트에 존재하는 문서들에서 발견된 모든 개념들이 다수의 계층구조를 형성한다.

Udo의 마크업 컨텍스트는 HTML의 특성을 고려할 때 매우 인상적인 연구이다. 이는 HTML 태그들의 의미 정보를 담고 있으며, 암묵적으로 웹 문서를 작성한 사람의 의도를 표현한 것이다. 이에 더하여, 웹 검색의 영역을 웹 전체를 대상으로 하는 것이 아니라, 특정 사이트로 한정한 것은 지능적인 웹 검색을 위한 단초를 보여주었다. 본 논문에서는 이러한 점에 착안하여, 마크업 정보를 이용한 웹 검색 성능 향상을 보이려 한다. 우리는 마크업 개념을 확장하여 마크업 패턴을 정의하고, 마크업 패턴을 이용하여 특정 사이트에서 웹 문서를 색인하고 가중치를 설정하는 웹 검색 성능 향상 방법을 제안한다.

3. 마크업 패턴(Markup Pattern)의 활용

3.1 마크업 패턴(Markup Pattern)

웹 문서에 존재하는 단어들은 굵음, 기울임, 큰 글자체 등 서식에 따라 분류될 수 있다. 일반적으로 웹 문서를 작성하는 사람은 자신이 사용자에게 보여주고 싶은 단어들에 대해 특정 서식을 적용한다. 즉 특정 서식이 적용된 단어들은 웹 문서 작성자의 의도를 담고 있다고 할 수 있다. 위장에서 설명한 Udo의 마크업 컨텍스트 역시 특정 서식을 적용하기 위한 태그들이다. 이러한 사고로부터, 마크업 패턴은 다음과 같이 정의될 수 있다.

정의 1. **Markup Pattern**

i) Implicitly, Markup Pattern is a mirror of intention of person who designs a web document.

ii) Explicitly, Markup Pattern is a style which decorates text in a web document

특정 태그들은 문서의 제목, 키워드 혹은 단락의 주제를 표현할 수 있고, 이 태그들에 의해 표현된 단어들은 문서에서 중요하게 여겨질 필요가 있다. BBC 뉴스 페이지의 기사 제목은 그림 1에 보이는 태그들을 이용해 표현될 수 있다. 혹은 몇몇 태그들은 그림의 제목을 위해 사용될 수 있다. 실제로, 특정 사이트에서 같은 종류의

표 1 BBC 사이트에서 사용된 CSS의 예

CSS id	Style	Description of text enclosed style
sh	FONT-WEIGHT: bold; FONT-SIZE: 18px; COLOR: #000000	Title of article (not page)
cap	FONT-WEIGHT: normal; FONT-SIZE: 10px; COLOR: #666666	Description text of image in article

내용을 표현할 때 같은 태그들을 사용하여 같은 서식을 적용한다. 즉 이 태그들을 마크업 패턴으로 고려한다면, 이 태그들에 의해 표현된 단어들은 웹 문서에서 중요한 용어로 작용하는 것이다. 최근에는 CSS(Cascading Style Sheet)를 이용하여 웹 문서의 서식을 표현하고 있다. CSS는 글자체, 색, 문서 여백 등을 지정하기 위한 간단한 방법을 제공한다. 표 1에서는 BBC 뉴스 사이트에서 실제로 사용되고 있는 CSS의 예를 보여주고 있다.

서식에 적용되는 태그들이나 CSS에 의해 표현되는 용어들은 웹 문서에 포함된 중요한 용어라 할 수 있다. 우리는 이 개념을 확장하여, 이 용어들을 개념으로 정의한다. 즉 마크업 패턴에 의해 표현되는 용어는 문서에 존재하는 개념인 것이다. 이러한 개념의 정의는 웹 문서로부터 중요한 용어들만을 추출할 수 있는 방법을 제공하며, 따라서 추출된 개념은 중요하지 않은 용어들을 포함하지 않음으로서, 결국 웹 검색 성능에 긍정적이 영향을 끼친다. 본 연구에서는 어떤 서식도 적용되지 않은 용어들을 중요하게 생각하지 않고, 마크업 패턴으로 표현된 용어들을 이용한 검색을 고려하는 것이다.

정의 2. **Concept**

Concept is a single-word or multi-words in markup pattern text. Concept is distinguished from other words which do not in markup pattern text. Concept has a weight

마크업 패턴을 구성하는 첫 번째 요소, 즉 서식에 적용되는 태그들을 고려할 때, <a>, <title>과 같은 태그들이 사용될 수 있다.

웹 문서로부터 유용한 용어들을 추출하기 위해 많은 연구가들은 몇몇 태그들의 의미 내용을 활용하였고, 위에서 설명한 태그들 역시 이와 같은 맥락이다. 일부 연구는 'meta-tag'의 활용에 주목하고 있는데, 본 연구에서는 'meta-tag'를 마크업 패턴을 구성하는 요소에서 제외하였다. 그 이유는 'meta-tag'의 종류가 고정되어 있지 않을 뿐 아니라, 'meta-tag'가 표현하는 의미를 활용하는 검색 방법은 지능적 검색 방법이라기보다, 시맨틱 검색의 한 종류라 판단되기 때문이다. 즉 본 논문은 웹 문서 스스로 가지는 잠재적 의미를 이용한 지능적 검색에 관한 연구이며, 의미 사전이나 온톨로지를 사용한 시맨틱 검색에 관한 것이 아니다.

서식에 적용되는 태그들과 CSS로 구성되는 마크업 패턴은 다음과 같이 구체적 형태로 재정의 된다.

정의 1. **(revisited) Markup Pattern**

Markup Pattern is a mirror of web designer's intention which shows his article efficiently. It is a non-ordered sequence of HTML tags and CSS elements. Next elements can make markup pattern.

*font-style tags : <font>, <b>, <i>, <u>, *
title tag : <title>
heading tags : <h1>…<h6>
CSS element which is related font style

3.2 마크업 패턴을 이용한 문서의 가중치 부여

특정 사이트에서 정보 검색을 빠른 시간에 할 수 있는 사람은 웹 페이지를 작성한 사람이거나 그 사이트를 자주 이용하는 사람일 것이다. 그들은 질의에 대한 답을 가진 페이지를 찾기 위한 지름길을 가지고 있다. 우리는 마크업 패턴이 이러한 지름길 역할을 할 수 있다고 예상한다. 즉 각 페이지의 특성을 마크업 패턴이 내재하고 있기 때문에 마크업 패턴을 이용해 가중치가 부여된 문서들은 질의에 대해 보다 정확히 답할 수 있다. Udo의 연구는 마크업 컨텍스트를 활용하여 후보 질의어들을 생성하였지만, 본 연구에서는 웹 문서의 가중치를 다시 부여한다. 웹 문서로부터 추출된 개념은 마크업 패턴의 중요도에 따라 가중치 값을 가지며, 문서의 가중치는 문서가 가진 개념의 가중치의 총합으로 표현된다. 마크업 패턴은 그 중요도에 따라 다른 중요도 값을 가질 수 있으며, 이는 시스템 사용자가 웹 작성자의 도움으로 설정할 수 있다.

문서의 초기 가중치 부여를 위해 잘 알려진 TF/IDF 방법이나 언어 모델(Language Model)에서 사용하는 가중치 부여 방법을 사용할 수 있다. 그러나 본 연구에서는 실험에 사용된 두 사이트, 즉 BBC와 CNN에서 사용되고 있는 검색 엔진의 가중치 부여 방법을 추가 작업 없이 사용하였다. 즉 질의에 대해, 각 검색 엔진이 돌려주는 결과 문서들을 대상으로 새로이 가중치 부여를 하여 질의에 대해 재평가를 하는 것이다. 우리는 특정 사이트에서 사용되고 있는 검색 엔진이 그 사이트에서 최적의 성능을 보이는 것이라 가정한다. 즉 본 연구의 목적에 비추어 볼 때, 기존 검색 엔진의 성능은 중요한 요소가 아니며, Google 혹은 Yahoo 등 어느 것이라도 될 수 있다. 검색 엔진이 돌려주는 결과 문서들은 이미 최적의 가중치가 부여된 문서들이며, 즉 본 연구에서는 추가 작업 혹은 비용 없이 초기 가중치를 부여할 수 있다.

4. 평가 방법 및 시스템

4.1 질의어 및 분석과정

마크업 패턴을 이용한 웹 검색의 성능을 평가하기 위해 본 연구에서는 BBC[1]와 CNN[2] 뉴스 사이트의 문서들을 실험 대상으로 하였다. 또한 2006 Google News Top 10 질의어[3]를 실험의 질의로 선택하였다. 평가는 다음과 같은 6단계를 거쳐 진행된다.

① 각 질의에 대해 두 사이트의 검색엔진이 돌려주는 상위 1000개 혹은 그 이하의 문서를 얻음(1000개 이하의 결과를 돌려주는 질의에 대해서는 1000개 이하).
② 각 사이트에서 얻어진 문서에서 마크업 패턴을 추출한다. 이를 위해 문서 분석, HTML 태그 분석, CSS 분석 등의 복잡한 작업이 요구됨.
③ 마크업 패턴에 의해 표현된 텍스트, 즉 개념을 추출.
④ 마크업 패턴의 중요도 값을 설정한다.
⑤ 마크업 패턴의 중요도 값에 따른 개념의 가중치를 부여. 즉 문서의 가중치를 재부여.
⑥ 질의를 가중치가 재 부여된 문서에 적용하고, 얻어진 결과를 처음 검색엔진이 돌려준 결과와 비교한다. 이 때 좀 더 좋은 결과를 얻기 위해 ③단계로 돌아가 마크업 패턴의 중요도 값을 재설정 할 수 있음.

4.2 문서의 순위부여 기법

실험 대상 사이트에서 사용되는 검색 엔진의 결과와 본 연구에서 제안한 방법으로 얻어진 결과를 비교하고 문서의 순위를 부여하기 위해 그림 2와 같은 간단한 점수 부여 식을 사용한다. 두 결과를 비교는 상위 10 문서에 대해서만 진행된다.

$$scoreofresult = \sum_{i=1}^{10} scoreof i^{th} rankeddocument \times (10-i)$$

그림 2 문서 순위부여 기법

각 문서는 0부터 10까지의 점수를 가질 수 있으며, 0은 질의와 관계없음을, 10은 질의와 강한 관계가 있음을 나타낸다. 두 결과를 비교하여, 상위 10개의 문서가 가지는 점수의 합이 높은 시스템의 성능이 높다고 판단한다.

4.3 검색 시스템

그림 3은 본 연구에서 개발한 검색 시스템 및 실험 단계를 보여주고 있다. HTML Parser와 CSS Parser를 이용하여 Markup Pattern Extractor는 문서로부터 마크업 패턴을 추출한다. 추출된 패턴은 유일한 식별자를 가진다. Indexer는 마크업 패턴을 이용하여 문서로부터 개념을 추출하게 되고, 패턴의 중요도 설정 및 개념의 가중치 부여로 문서들의 가중치가 재부여 된다. 이 문서들은 개념과 가중치의 쌍으로 간단하게 구조화된다. 이상의

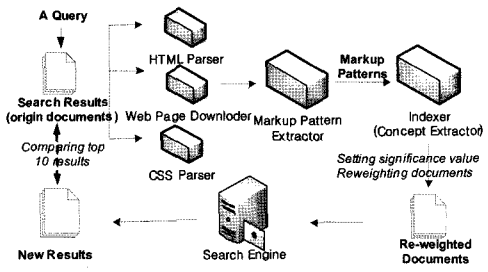

그림 3 검색 시스템 및 평가 방법

표 2 BBC와 CNN 사이트에서 수집된 문서 통계

Query	BBC		CNN	
	# of retrieved docs	# of concepts	# of retrieved docs	# of concepts
Paris Hilton	574	887	97	287
Orlando bloom	664	1012	35	-
Cancer	1000	1402	1000	2650
podcasting	994	1424	15	-
hurricane Katrina	998	1855	1000	1912
bankruptcy	995	1382	1000	2033
Martina hingis	998	1881	118	424
Autism	987	1412	50	203
2006nfl draft	253	803	1	-
celebrity big brother 2006	994	2319	22	-

과정은 모두 off-line으로 진행되고, 따라서 실제 검색에 소요되는 시간에는 영향을 미치지 않는다. Search Engine은 Re-weighted Documents를 토대로 질의에 대해 관련 있는 문서를 검색하여 사용자에게 결과를 돌려준다.

5. 실험 및 결과

5.1 실험 초기 설정

위장에서 설명한 검색 시스템은 J2SE 1.5 플랫폼에서 개발되었고, HTML 분석과 CSS 분석을 위해 Open Source 라이브러리를 사용하였다. 표 3은 각 질의에 대해 BBC와 CNN으로부터 얻어진 문서의 수와 개념의 수를 보여주고 있다.

BBC 검색 엔진은 10개의 질의에 대해 고른 수의 결과 문서를 돌려주었지만, CNN 검색 엔진은 4개의 질의에 대해 소수의 문서만을 결과로 보여주었다. 그러나 이는 실험에 문제가 되지 않는다. 즉 위에서 설명했듯이, 본 연구는 존재하는 검색 엔지의 우열을 가리는 것이 아니며, 질의에 대한 결과 문서의 수로 검색 엔진의 성능을 평가할 어떠한 근거도 없다. 또한 실험에 필요한 것은 두 사이트에서 질의에 관련 있는 문서들이며, 문서의 수와는 관계없다. 그러나 결과 문서가 적은 CNN의 4개의 질의(Orlando Bloom, podcasting, 2006 nfl draft, celebrity big brother 2006)에 대해서는 실험을 진행하지 않았는데, 이는 적은 수의 문서들을 비교하고 평가하는 것이 의미 없는 일이기 때문이다.

1) http://www.bbc.co.uk/
2) http://www.cnn.com/
3) http://www.google.com/intl/en/press/zeitgeist2006.html

표 3 BBC사이트에서의 성능 평가

Weighted markup pattern	9 patterns with weight value 3 15 patterns with weight value 2 22 patterns with weight value 1		
Query	Average Score		Improvement ratio (%)
	BBC search engine	Proposed technique	
Paris Hilton	132.8	149.1	12.3
Orlando bloom	408.1	396.3	-2.9
Cancer	99.4	155.2	56.1
Podcasting	183.0	190.7	4.2
hurricane Katrina	396.5	401.3	1.2
Bankruptcy	120.9	130.8	8.2
Martina hingis	222.0	296.8	33.7
Autism	140.2	150.6	7.4
2006 nfl draft	84.6	146.8	73.5
celebrity big brother 2006	486.2	473.4	-2.6

5.2 분석 및 토의

10개의 질의가 BBC 사이트에서 평가되었다. <title> 태그 등 9개의 마크업 패턴이 3의 중요도를 가졌으며, 15개의 패턴이 중요도 2, 22개의 패턴이 중요도 1을 가졌다. 전체 83개의 패턴 중 37개의 패턴은 무시되었다. 표4에서 보듯이, 제안한 방법의 검색은 10개의 질의 중 8개의 질의에 대해 성능 향상을 가져왔다. 특히 'cancer'와 '2006 nfl draft'에 대해 주목할 만한 성능 향상을 보였다. 반면 2개의 질의에 대해 약간의 성능 저하를 나타내었다.

CNN 사이트에서 10개의 질의 중 대상 문서의 수가 적절한 6개의 질의에 대해 평가되었다. 77개의 마크업 패턴 중 34개의 패턴이 중요도를 가졌고, 나머지 43개의 패턴은 무시되었다. 6개의 질의에 대해 제안한 검색 방법은 성능 향상을 보였으며, 특히 'Paris Hilton'에 대해 주목할 만한 성능 향상을 보였다. 표 5는 CNN 사이트에서의 실험 결과이다.

표 4 CNN사이트에서의 성능 평가

Weighted markup pattern	5 patterns with weight value 3 10 patterns with weight value 2 19 patterns with weight value 1		
Query	Average Score		Improvement ratio(%)
	CNN search engine	Proposed technique	
Paris Hilton	112.6	169.7	50.7
Cancer	138.3	144.2	4.2
hurricane Katrina	264.7	346.9	31.1
bankruptcy	120.9	150.0	24.1
Martina hingis	233.6	272.1	16.5
Autism	98.2	135.2	37.7

6. 결 론

본 연구에서는 지능적인 웹 검색을 위해 마크업 패턴을 활용한 검색 기법을 제안하였다. 마크업 패턴은 암묵적으로 웹 문서를 작성한 사람의 의도이며, 이는 명시적으로 HTML 태그 혹은 CSS을 통해 표현된다. 마크업 패턴을 활용하기 위해, 본 연구에서는 마크업 패턴으로 표현된 단어들을 개념으로 정의하였다. 개념은 웹 문서에 포함된 중요한 용어를 의미하며, 이 개념에 가중치를 부여함으로서 웹 문서의 가중치를 부여하였다. 제안한 방법의 평가를 위해서 웹 문서를 분석하고 마크업 패턴을 추출할 수 있는 검색 시스템을 개발하였다. 제안하는 방법의 우수성을 입하기 위해, 본 연구에서는 BBC와 CNN 두 사이트의 웹 문서들을 대상으로 실험하여 제안하는 시스템이 좋은 결과를 돌려줌을 보였다. 이러한 결과를 볼 때 마크업 패턴의 활용은 지능적인 웹 검색에 도움을 줄 것으로 여겨진다.

그러나 이 연구는 마크업 패턴을 활용하기 위한 초기 연구로서, 개선되어야 할 문제가 존재한다. 첫째, 웹 문서의 서식을 정확히 분석하여 다양한 마크업 패턴을 추출하는 문제, 둘째, Multi-word 개념 추출 방법, 셋째, 다양한 질의를 통한 실험 등이 향후 연구 과제로 요구된다.

참 고 문 헌

[1] Hodgson, J. 2001. Do HTML Tags Semantic Content? IEEE Internet Computing, 5(1):20-25.

[2] Sanderson, M. and Croft, W. B. 1999. Deriving Concept Hierarchies from text. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 206-213, Berkeley, CA.

[3] Lawrie, D. J. and Croft, W. B. 2003. Generating Hierarchical Summaries for Web Searches. In Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pages 457-458, Toronto, Canada.

[4] Reiner, K. and Jason, Z. 2004. Mining Anchor Text for Query Refinement. In Proceedings of WWW2004, New York, USA.

[5] Silverstein, C., Marais, H., Henzinger, M., Morics, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum, 33(1):6-12.

[6] Udo, K. 2005. Intelligent Document Retrieval Exploiting Markup Structure. : Springer, Berlin Heidelberg New York.

[7] Ruth, Y. Z., Laks, V. S. L., Ruben, H. Z. 2004. Extracting Relational Data from HTML Repositories. ACM SIGKDD Explorations Newsletter, 6(2): 5-12.

[8] Kleinberg, J. M. 1998. Authoritative Sources in Hyperlinked Environment. In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, pages 668-677, ACM.

[9] Brin, S. and Page, L. 1998.The anatomy of a large-scale hypertextual web search engine. In Proceedings of the seventh international conference on World Wide Web 7 (WWW7), Brisbane, Australia.