마크업 패턴을 이용한 웹 검색*

김민수○ 김민구
아주대학교 인공지능연구실
visual○@ajou.ac.kr, minkoo@ajou.ac.kr

Web Information Retrieval Exploiting Markup Pattern

Minsoo Kim○ Minkoo Kim
Artificial Intelligence Lab. Ajou University

1. 서 론

오늘날 웹은 방대한 양의 정보를 가지고 있으며, 이러한 정보는 컴퓨터 과학 뿐 아니라 다양한 분야에서 널리 활용되고 있다. 웹 정보들은 HTML(Hyper Text Markup Language)를 기반으로 표현되어 있고 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, Google등의 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, 첫 번째는 웹 문서에 대한 색인의 어려움이며, 두 번째는 사용자 질의의 모호함이다. 이 문제를 해결하기 위해 다양한 색인 기법, 질의 확장 및 수정 등의 연구가 발표되었으며[1][2][3], 특히 Udo의 연구[4]는 웹 문서의 특성을 규정하고, 마크업 언어의 중요한 개념을 파악하여 문서로부터 용어들을 추출하고 이를 질의 확장에 응용하여 성능 향상에 큰 성과를 보였다. 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. 웹 문서의 내용은 특정 마크업 언어에 의해 강조되거나 특징지어지고, 이를 이용하여 웹 검색 성능을 향상시킬 수 있다는 것이다. 본 논문에서는 이러한 특징을 이용하여 웹 문서의 가중치 재설정을 통한 성능 향상을 꾀한다. 이를 위해 마크업 패턴을 정의하고, 검색 성능 향상을 위한 방법들을 제안한다.

2. 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. Udo의 연구[4]에서 정의한 마크업 컨텍스트 역시 특정 서식을 적용하기 위한 태그들이다. 특정 태그들은 문서의 제목, 키워드 혹은 단락의 주제를 표현할 수 있고, 이 태그들에 의해 표현된 단어들은 문서에서 중요하게 여겨질 필요가 있다. 예를 들어 BBC 뉴스 페이지의 기사 제목들은 <h1> 태그를 사용하여 표현될 수 있고, 그림의 제목들은 <b>태그로 표현될 수 있다. 최근에는 글자체, 색, 여백 등을 간단히 지정할 수 있는 CSS(Cascading Style Sheet)를 이용하여 웹 문서의 서식을 표현하고 있으며, 이 또한 사용자의 의도를 표현하는 것이다. 이러한 사고로부터, 마크업 패턴은 다음과 같이 정의될 수 있다.

***정의 1*. 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, title tag, heading tags. CSS element which is related font style>

또한 마크업 패턴에 의해 꾸며지는 단어들은 문서에서 중요한 의미를 가진다고 볼 수 있으며, 본 논문에서 이 단어들을 정의2와 같이 Concept으로 정의한다

***정의 2*. Concept**

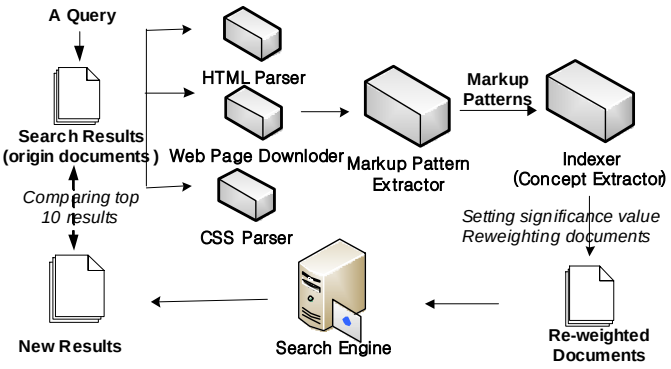
Concept is a single-word or multi-words in markup pattern text. Concept is distinguished from other words which do not in markup pattern text. Concept has a weight

---

본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅및네트워크원천기반기술개발사업의 지원에 의한 것임

3. 마크업 패턴을 활용한 웹 검색 시스템

그림 1은 웹 검색을 위한 과정과 함께 검색 시스템의 컴포넌트들을 보여준다. 다음 장에서 설명한 BBC, CNN 두 도메인에 대해서, 각 질의에 대한 결과문서들을 모아 저장한 후, HTML parser, CSS parser 등을 이용해 마크업 패턴을 추출한다. 추출된 마크업 패턴을 이용해 문서에서 Concept들을 추출하게 되고, 각 Concept은 이후 사용자 혹은 시스템 기본값으로 가중치가 부여된다. 즉 문서의 가중치가 재설정된 것이다. 가중치가 재설정된 문서는 이제 질의에 대한 새로운 대상 문서들이 되며, 질의에 대해 관계 높은 순서로 순위가 부여된다. 순위 부여 기법은 수식 1과 같다. 각 질의에 대해 도메인에 존재하는 검색 엔진과 제안하는 검색엔진의 상위 10개의 결과에 대해 점수가 부여된다. 각문서는 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.



[그림 1] 검색 시스템 구조도

$$scoreofresult = \sum_{i=1}^{10} score\,of\,i^{th}\,ranked\,doc \times (10-i)$$

[수식 1] 순위 부여 기법

4. 결과 및 결론

제안하는 방법을 평가하기 본 논문에서는 BBC*와 CNN** 뉴스 사이트의 문서들을 대상으로 각 사이트에서 사용되고 있는 검색엔진과 위장에서 기술한 검색 시스템을 대상으로 동일한 질의에 대한 정확도를 비교 평가하였다. 사용한 질의는 2006년 Google 뉴스 사이트에 입력된 상위 10개의 질의를 사용하였다. 평가는 결과의 출처를 공개하지 않은 상태에서 11명의 컴퓨터분야 대학생 및 대학원생들에 의해 진행되었다. 아래 표는 10개의 질의에 대해 각 도메인에 대해, 검색된 문서 수, 추출된 마크업 패턴 등과 함께 정확도를 비교하고 있고, 많은 질의에 대해 제안한 방법의 우수성을 보여주고 있다.

[표1] BBC,CNN 도메인에서 실험 결과

Domain	BBC			CNN			
Weighted markup pattern	9 patterns with weight value 3 15 patterns with weight value 2 22 patterns with weight value 1			5 patterns with weight value 3, 10 patterns with weight value 2 19 patterns with weight value 1			
Query	#doc	Average Score		Improvement ratio (%)	#doc	Average Score		Improvement ratio (%)
		BBC search engine	Proposed technique			CNN search engine	Proposed technique	
Paris Hilton	574	132.8	149.1	12.3	97	112.6	169.7	50.7
Orlando bloom	664	408.1	396.3	-2.9	35	-	-	-
Cancer	1000	99.4	155.2	56.1	1000	138.3	144.2	4.2
Podcasting	994	183.0	190.7	4.2	15	-	-	-
hurricane Katrina	998	396.5	401.3	1.2	1000	264.7	346.9	31.1
Bankruptcy	995	120.9	130.8	8.2	1000	264.7	346.9	31.1
Martina hingis	998	222.0	296.8	33.7	118	233.6	272.1	16.5
Autism	987	140.2	150.6	7.4	50	98.2	135.2	37.7
2006 nfl draft	253	84.6	146.8	73.5	1		-	-
celebrity big brother 2006	994	486.2	473.4	-2.6	22		-	-

5. 참고문헌

[1] Hodgson, J. 2001. Do HTML Tags Semantic Content? IEEE Internet Computing, 5(1):20-25,
[2] 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
[3] Reiner, K. and Jason, Z. 2004. Mining Anchor Text for Query Refinement. In Proceedings of WWW2004, New York, USA
[4] Udo, K. 2005. Intelligent Document Retrieval Exploiting Markup Structure. : Springer, Berlin Heidelberg New York

* http://news.bbc.co.uk
** http://www..cnn.com