2002 년 한국정보처리학회 춘계학술발표논문집 제 9 권 제 1 호

스타일 기반 키워드 추출

이준휘*, 이원석*
*연세대학교 컴퓨터과학과
e-mail :{juni, leewo}@amadeus.yonsei.ac.kr

Keyword Extraction based on Style

Joon-Hwi Lee*, Won-Suk Lee*
*Computer Science Department, Yonsei University

요 약

기존의 키워드 추출 방법은 출현회수(frequency)에 기반한 가중치(weight) 부여 방식이 많이 쓰였다. 본 논문에서는 HTML 문서와 같이 스타일이 적용된 문서의 경우 출현회수와 함께 단어에 적용된 스타일을 고려하여 가중치를 부여해 키워드를 추출하는 방법을 제안한다. 가중치를 부여할 스타일 항목과 항목별 가중치 부여방법을 정의하고 이를 단어별로 합산하고 정규화(normalization)하는 방법을 정의하여 스타일에 기반해 키워드를 추출하였다. 내용이 특정된 도메인으로부터 순위(ranking)가 매겨진 도메인 키워드 리스트를 뽑아서 이를 기준으로 삼아 기존의 출현회수 기반의 키워드 추출 방식과 양적, 질적인 비교를 수행하여 우월함을 보였다.

1. 서론

World-Wide Web(WWW)이 널리 쓰이게 되면서 HTML 문서[1]가 급증하게 되었다. HTML 문서는 기존의 plain-text 와 달리 semi-structured 문서로서 다양한 스타일을 적용하여 문서를 꾸밀 수 있다. 이렇게 다양한 스타일이 적용된 HTML 문서의 키워드를 추출할 때 스타일에 기반한다면 더 나은 결과를 얻을 수 있을 것이다. 이에 기존의 단어 빈도수, 곧 tf factor(term frequency factor)기반의 가중치 부여 방식[2][3]에 스타일에 기반한 가중치 부여를 추가한 스타일 기반 키워드 추출 방식을 제안하고자 한다. 또한 문서 길이 정규화(Document Length Normalization)에는 최대 빈도수 정규화와 코사인 정규화의 두 가지 방법[4]이 많이 쓰이는데 정규분포에 기반한 새로운 정규화 방법을 제안하고자 한다.

2. 스타일 기반 키워드 추출

문서를 이루는 단어들 가운데 문서의 전반적인 스타일에서 벗어나는 스타일을 가진 단어들은 보다 중요한 의미를 가져서 강조하고자 하는 단어일 가능성이 높다. 이러한 스타일 차이에 따른 단어의 중요도를 파악하여 키워드를 추출하는 것을 스타일 기반 키워드 추출이라 한다.

2.1 스타일별 가중치(weight) 부여 방식

HTML 태그들 가운데 Text Formatting 태그들은 적용된 단어를 어떤 식으로 렌더링 해야 하는 지를 지시한다. 한 단어에 여러 태그가 적용될 수 있고 이들의 종합된 결과로 단어의 스타일이 결정된다. 따라서 각 Text Formatting 태그별로 가중치를 계산하지 않고 최종 결과물로서 브라우저를 통해 실제로 렌더링되는 스타일로부터 가중치(weight)를 계산한다. 스타일을 다음의 7 가지 항목[5]으로 나누어 각 항목별로 가중치를 부여한다.

- Font-Family : 글꼴 종류
- Font-Size : 글꼴 크기
- Font-Style : italic, normal, 또는 oblique
- Font-Weight : 글꼴 두께
- Text-Align : 텍스트 정렬방식 left-aligned, right-aligned, centered, 또는 justified
- Color : 색상
- Text-Decoration : blink, line-through, overline, 또는 underline decorations

2.2 정규화(Normalize)된 가중치 수식

단어가 문서에 나타나는 출현회수(frequency)에 근거해 문서의 키워드를 찾아내되 단순히 문서의 출현회

수를 모두 동일한 출현회수로 보는 것이 아니라 적용된 스타일에 따라서 가중치를 부여한다. 앞서 제시한 7 가지 스타일 항목별로 각각 정규화하여 계산된 가중치를 합하여 단어의 가중치가 부여된 출현회수(weighted frequency)를 구한다. 문서의 단어추출에는 한국어 분석모듈[6]을 이용하였다.

문서 내에서 실제로 적용된 스타일을 스타일 인스턴스라 한다. 예를 들어 어떤 문서에서 Font-Size 가 "12", "14", "24"의 세 종류가 사용되었다면 Font-Size 스타일 인스턴스는 "12", "14", "24" 세 가지가 된다. Font-Style 스타일과 같은 경우는 italic, noramal, 또는 oblique 이 세 가지가 스타일 인스턴스가 될 수 있다.

각 스타일 항목별 스타일 인스턴스 i 에 대한 가중치를 SW_i 라 하고 문서의 총 단어수는 TC, 스타일 인스턴스 i 가 적용된 단어의 수를 SC_i 라 하면, SW_i 는 다음에 제시될 식(1),(2)의 두 가지 방법에 따라 결정된다.

식(1)은 글꼴 크기(Font-Size), 두께(Font-Weight)와 같이 스타일 인스턴스의 값에 따라 중요도를 판단할 수 있는 경우에 적용한다. 스타일 인스턴스의 값을 대표값으로 보고 스타일 인스턴스가 적용된 단어수 SC_i 를 도수로 보아 평균과 표준편차를 구한다. 이 평균을 SC_{avg}, 표준편차를 SC_{sd} 라고 하고 스타일 인스턴스의 값을 SV 라 하면, SW_i 는 식(1)과 같다.

$$SWi = \frac{SV_i - SC_{avg}}{SC_{sd}} \qquad (1)$$

식(2)는 글꼴 종류(Font-Family), 글꼴 스타일(Font-Style), 색(Color), 정렬(Text-Align), 텍스트 데코레이션(Text-Decoration)과 같이 스타일 인스턴스의 값에 따른 중요도를 판단할 수 없는 경우, 스타일 인스턴스의 대표값을 문서의 총 단어수 대비 해당 스타일 인스턴스가 적용된 단어수의 비율로 삼고 스타일 인스턴스가 적용된 단어수를 도수로 잡아서 평균과 표준편차를 구한다. 이 평균을 SC_{avg}, 표준편차를 SC_{sd} 라고 하고 스타일 인스턴스의 대표값을 SV 라 하면, SW_i 는 식(2)와 같다.

$$SW_i = \frac{SV_i - SC_{avg}}{SC_{sd}}, SV_i = \frac{SC_i}{TC} \qquad (2)$$

단어 w_i 가 문서에 출현한 인스턴스 j 의 가중치가 부여된 출현회수 $fv_{i,j}$ 는 출현에 대한 출현회수 1 에 적용된 스타일 인스턴스의 합으로 구해지며, 이는 식(3)과 같다.

$$fv_{i,j} = 1 + \sum_{k=\text{적용된태그인스턴스}} SW_k \qquad (3)$$

식(3)으로 구해진 단어 w_i 의 인스턴스들의 fv 값의 총합이 단어 w_i 의 가중치가 부여된 출현회수 fv_i 가 되며 이는 식(4)와 같다.

$$fv_i = \sum_{j=w_i\text{의 인스턴스}} fv_{i,j} \qquad (4)$$

단어 w_i 의 정규화된 가중치가 부여된 회수(normalized weighted frequency) f_i 는 fv_i 값들이 표준 정규 분포를 이룬다고 보고 정규화를 수행하여 구한다. fvi 값들의 평균과 표준편차를 각각 Avg(fv)와 Sd(fv)라 하면, f_i 는 식(5)로 구하여진다. 이 f_i 가 단어 w_i 의 가중치이다.

$$fi = P(Z < a), a = \frac{fv_i - Avg(fv)}{Sd(fv)} \qquad (5)$$

3. 스타일 기반 키워드 추출 방식의 검증

스타일 기반의 키워드 추출 방식의 검증은 [그림 1]의 순서도에 따라서 행한다. 기준이 되는 키워드 리스트를 뽑아내기 위해 한 도메인에 선택하여 그에 속한 다수의 사이트를 선택하여 도메인 키워드 리스트를 추출한다. 이를 기준으로 삼아 스타일 기반 키워드 추출 방식과 출현회수 기반 키워드 추출 방식을 비교한다.

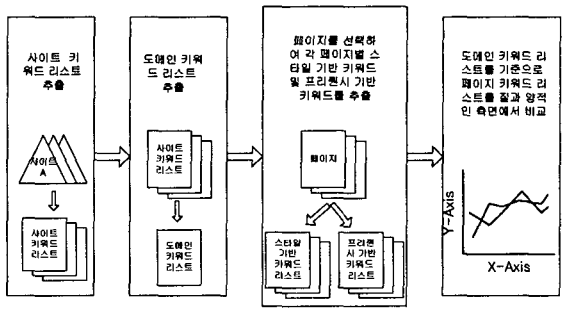


그림 1. 검증 순서도

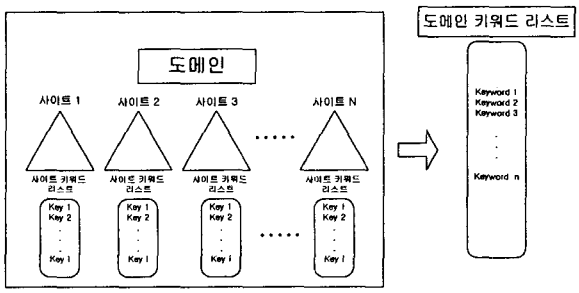


그림 2. 도메인 & 도메인 키워드 리스트

3.1 도메인에 속한 사이트들의 사이트 키워드 추출

도메인과 도메인 키워드 리스트는 [그림 2]와 같은 구조를 가진다. 먼저, 특정한 도메인을 선택하여 그에 해당되는 다수의 웹사이트(이하 사이트)들을 선택한다. 이 사이트내의 각 페이지의 모든 단어를 추출해 이를 페이지 키워드 리스트로 삼는다. 이 페이지별 키워드 리스트를 가지고, 사이트내의 한 페이지 이상에서 출현한 키워드를 사이트 키워드로 삼는다. 곧, 이는 하

나 이상의 페이지에 출현한 키워드를 모두 사이트 키워드 리스트로 삼게 된다. 이렇게 뽑아낸 사이트 키워드 리스트를 이용하여 도메인 키워드 리스트를 생성한다.

3.2 도메인 키워드 리스트 추출

[그림 2]는 도메인의 구성과 도메인 키워드 리스트를 나타내고 있다. 3.1 절의 기준에 따라서 선택한 도메인은 당연히 밀접한 내용적 유사성을 보이는 사이트 N 개로 구성되며, 3.1 절의 사이트 키워드 추출 단계로부터 도메인에 속하는 사이트들의 키워드 리스트를 가지고 있다. 이 사이트 키워드 리스트들을 이용하여 도메인 키워드 리스트를 생성한다. 사이트 키워드 리스트에 속하는 키워드 가운데 주어진 최소 서버 서포트(minimum server support s, $0 \le s \le 1$)를 만족하는 키워드만을 도메인 키워드로 삼는다. 키워드의 서버 서포트는 전체 사이트 수 대비 출현한 사이트의 비율로 구한다. 분명하게 한정된 도메인을 선택한다면 일정 수 이상의 사이트를 바탕으로 도메인 키워드 리스트를 뽑아냈을 때, 이 리스트가 도메인을 대표하는 키워드 리스트로 보아도 무방할 것이다.

3.3 도메인 키워드 리스트의 순위 부여

도메인 키워드 리스트의 순위를 부여하기 위해 먼저 도메인내의 사이트 별로 사이트 키워드의 순위를 부여한다. 그리고 사이트 키워드 순위를 바탕으로 도메인 키워드의 순위를 결정한다.

3.3.1 사이트 키워드 순위 부여

한 사이트내의 여러 페이지에 출현한 키워드들은 한 페이지를 대표하는 단어이기 보다는 일반적인 단어일 가능성이 높다. 따라서 사이트 내에서 여러 페이지에 나온 단어일수록 낮은 가중치를 부여하도록 한다. 사이트 키워드의 가중치는 식(6)에 따라서 부여한다.

$$\text{SiteKeywordWeight} = \frac{\text{TotalSitePageCount}}{KP} \quad (6)$$

TotalSitePageCount 는 사이트내의 전체 페이지의 수이고, KP 는 키워드가 출현한 페이지의 수이다. 위 식에 의해 적은 페이지에 출현한 키워드가 높은 가중치를 가지게 된다.

3.3.2 도메인 키워드 순위 부여

3.2 절에서 추출된 도메인 키워드에 한해서 다음과 같은 방법으로 순위를 부여한다. 키워드의 가중치는 출현한 사이트의 가중치의 합을 전체 사이트 수로 나눈 값으로 삼고 이 가중치의 내림차순으로 순위를 부여한다. 도메인 키워드의 가중치 공식은 식(7)과 같다.

$$DomainkeywordWeight = \frac{\sum_{\text{site}} SiteKeywordWeight}{\text{TotalSiteCount}} \quad (7)$$

TotalSiteCount 는 도메인의 사이트의 총 개수이다.

3.4 페이지별 키워드 추출의 비교 대상

앞에 제안한 스타일에 기반한 키워드 추출 방법의 적합성을 검증하기 위한 비교 대상으로써 단어의 출현회수만 고려한 추출 방법과 비교한다. 이는 단어의 각 출현 인스턴스 별로 적용된 스타일은 고려하지 않고 출현 빈도만으로 가중치를 구하는 방법이다. 이에 따라 출현회수 기반 키워드 추출 방식에 따른 키워드별 가중치는 다음과 같이 구한다.

문서상의 단어 w_i 의 한 인스턴스 j 의 출현회수 $fv_{i,j}$ 는 식(8)과 같다.

$$fv_{i,j} = 1 \quad (8)$$

단어 w_i 의 총 출현회수 fv_i 는 식(9)와 같다.

$$fv_i = \sum_{j = w_i\text{의 인스턴스}} fv_{i,j} \quad (9)$$

정규화는 스타일 기반의 방법과 동일한 방법인 식(5)를 이용하여 수행한다.

3.5 페이지별 키워드 추출 비교 방법

도메인내의 모든 페이지에 대하여 페이지별 키워드 추출을 비교한다. 이 비교에는 threshold 값이 상수로 주어지며 키워드의 정규화된 가중치 값이 threshold 를 넘어야만 페이지의 키워드로 삼는다. 먼저 출현회수 기반한 방식으로 threshold 이상인 키워드를 페이지 키워드로 추출한 후, 스타일 기반 키워드 추출 방식으로 가중치가 높은 순으로 상위에서부터 출현회수 기반의 방식으로 뽑은 것과 같은 수의 키워드를 추출하여 비교한다.

- 양적인 비교

뽑아낸 키워드들 가운데 도메인 키워드 리스트에 속한 키워드의 비율을 측정한다. 이 비율이 높을수록 도메인 키워드를 더 많이 뽑아내는 것이고 따라서 더 양적으로 더 우수하다고 볼 수 있다.

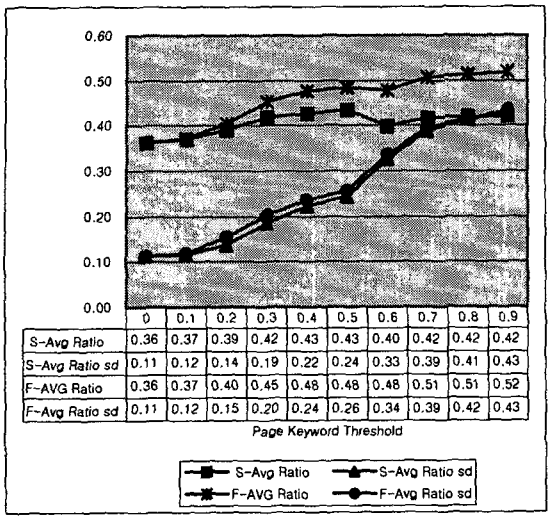
- 질적인 비교

키워드들 중에 도메인 키워드에 속한 키워드들의 도메인 순위(Ranking)의 평균을 구해 비교한다. 평균 순위가 작을수록 뽑아낸 키워드들의 도메인 순위가 높다는 것을 의미하고, 이는 곧 더 양질의 키워드를 뽑아낸 것이라 볼 수 있다.

이와 같은 방법으로 페이지별 비교를 수행하고 페이지별 비교 결과를 평균 내어 전체 페이지 비교 결과로 삼는다.

4. 실험 결과

실험 도메인은 구청의 민원관련 페이지로 잡았고, 총 16 개의 구청 사이트의 227 개 페이지를 대상으로 실험을 수행하였다. 서버 서포트는 0.6 이다.



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
S-Avg Ratio	0.36	0.37	0.39	0.42	0.43	0.43	0.40	0.42	0.42	0.42
S-Avg Ratio sd	0.11	0.12	0.14	0.19	0.22	0.24	0.33	0.39	0.41	0.43
F-AVG Ratio	0.36	0.37	0.40	0.45	0.48	0.48	0.48	0.51	0.51	0.52
F-Avg Ratio sd	0.11	0.12	0.15	0.20	0.24	0.26	0.34	0.39	0.42	0.43

그림 3. 양적인 비교 결과 그래프

4.1 양적인 비교

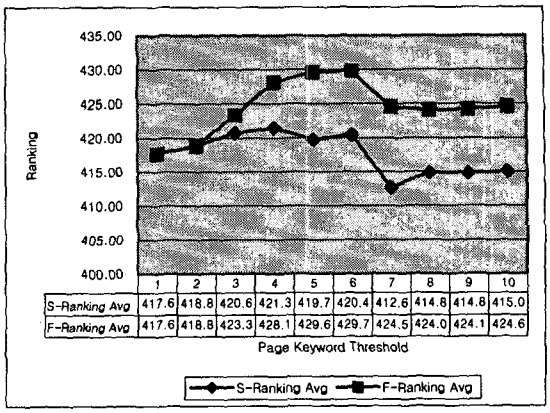
스타일 기반의 결과는 S-로 시작하고 출현회수 기반의 결과는 F-로 시작한다. Avg Ratio 가 페이지의 키워드 중 도메인 키워드 리스트에 속한 키워드의 비율의 페이지 평균을 나타내고 Avg Ratio sd 는 이 비율의 페이지당 표준편차를 나타낸다. [그림 3]의 결과에 보이듯이 양적인 비교인 페이지당 추출한 키워드 중 도메인 키워드의 비율은 threshold 가 올라갈수록 스타일 기반이 크게는 0.1 까지 떨어지게 나타났다. 그러나 페이지당 추출되는 키워드의 숫자가 많아야 10 개 내외란 점을 감안하면 이 같은 차이는 키워드 한 두개 정도 차이이다.

4.2 질적인 비교

[그림 4]의 그래프의 결과로 알 수 있듯이 질적인 측면에서는 스타일 기반의 키워드 추출이 평균 순위 10 정도의 차이를 보이면서 작게 나타났다. 이는 스타일에 기반한 추출 방법이 더 상위 순위의 키워드를 추출한다는 것으로 질적으로 더 우수하다는 것을 의미한다. [그림 5]의 순위 표준편차 결과를 보면 평균랭킹의 차이 만큼 스타일 기반 키워드 추출 방식의 랭킹 표준편차가 더 크게 나타난다. 이는 스타일 기반의 키워드 추출 방식의 순위가 더 넓게 분포하나 출현회수 기반의 방식보다 최소한 같거나 더 낮은 순위로 분포한다고 볼 수 있다.

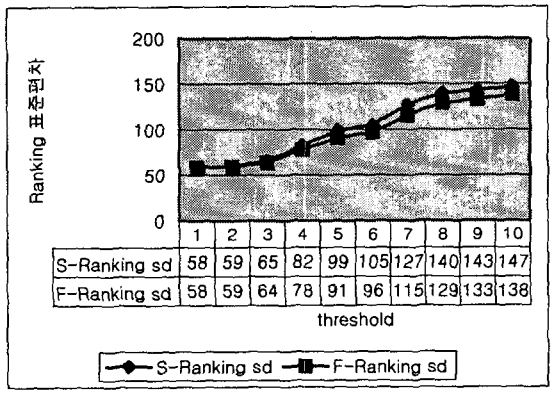
5. 결론 및 향후 연구방향

스타일 기반 키워드 추출이 출현회수에 기반한 키워드 추출 방식과 비교하여 양적인 부분에서는 비슷하나 질적인 면에서는 더 우수한 결과를 보임을 보였다. 향후 연구 방향으로는 각 스타일 항목별 가중치를 부여함으로써 더 나은 결과를 보일 수도 있을 것이다. 또한 HTML 문서뿐 만이 아니 HWP, 워드 문서 등 다양한 문서에도 확대 적용할 수 있을 것이다



	1	2	3	4	5	6	7	8	9	10
S-Ranking Avg	417.6	418.8	420.6	421.3	419.7	420.4	412.6	414.8	414.8	415.0
F-Ranking Avg	417.6	418.8	423.3	428.1	429.6	429.7	424.5	424.0	424.1	424.6

그림 4. 질적인 비교 결과 그래프 1



	1	2	3	4	5	6	7	8	9	10
S-Ranking sd	58	59	65	82	99	105	127	140	143	147
F-Ranking sd	58	59	64	78	91	96	115	129	133	138

그림 5. 질적인 비교 결과 그래프 2

참고문헌

[1] http://www.w3c.org/MarkUp/

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval", ADDISON WESLEY, pp. 29-30, 1999

[3] I. Aalbersberg,, "A Document Retrieval Model Based on Term Frequency Ranks", 17th international ACM SIGIR Conference on Research and Development in Information Retrieval, 163-172, 1994

[4] Amit Singhal, Chris Buckley, and Mandar Mitra, "Pivoted Document Length Normalization", Proceedings of 19th ACM International Conference on Research and Development in Information Retrieval, 1996

[5]http://msdn.microsoft.com/library/default.asp?url=/workshop/author/dhtml/reference/properties.asp

[6] 강승식, "한국어 분석 모듈 5.0.0a", http://nlp.kookmin.ac.kr/