관계성 확률을 이용한 XML 태그의 가중치 결정

정혜진*
전북대학교 공과대학 전자정보공학부
e-mail : arayesO@chonbuk.ac.kr

Weight decision of the XML Tag using Relationship Probability

Hye-Jin Jeong*
*Division of electronics and information engineering, Chonbuk National University

요 약

보다 효과적인 색인어 추출 및 색인어 가중치 결정을 위하여 문서의 내용뿐 아니라 구조를 이용하여 색인을 추출하는 연구가 이루어지고 있는데, 대부분의 연구들이 XML 태그의 중요도가 아닌, 문맥상의 단락에 대한 중요도를 계산하거나 HTML 문서 태그의 중요도 결정에 관한 연구들이다. 이러한 기존 연구들은 대부분이 객관적인 실험을 통해서 중요도를 입증하기보다는 상식적인 관점에서 단순한 수치로 중요도를 결정하고 있다. 본 논문에서는 웹 문서 관리를 위한 표준으로 자리잡아가고 있는 XML 문서의 태그 정보를 이용한 자동색인을 위하여, 논문을 구성하는 주요 태그의 가중치를 계산하는 방법을 제안한다. 보다 객관적인 가중치 결정을 위하여 인용된 문서간의 관계를 알아보고 서로 연관이 있을 확률을 계산하여 그 기대치만큼 색인어에 대한 가중치에 반영한다. 그리고 기존 태그 중요도 결정 방법을 적용하여 계산된 색인어 가중치를 이용한 검색성능과 비교함으로써 본 논문에서 제안한 방법을 적용하여 계산된 색인어 가중치의 효과를 검증한다.

1. 서 론

웹의 발달과 인터넷의 보편화로 인하여 자신이 원하는 정보를 얻기가 점점 어려워지고 복잡해지므로 웹에서 보다 효과적으로 색인을 추출하고 검색 편의성을 제공하는 연구가 필요하다. 이러한 문제점을 해결하기 위한 대안 중의 하나가 정보를 XML(Extended Markup Language) 형태로 관리하는 것이다. XML은 문서의 구조정보를 제공할 뿐만 아니라, XML 태그(tag)는 데이터를 해석하는 데에 사용할 수 있기 때문에 XML의 역할과 중요성이 인식되고 있다. HTML이 하나의 고정된 DTD(Document Type Definition)를 사용하는 것과는 달리 XML은 논리적 구조를 나타내는 여러 DTD를 사용할 수 있다. XML 문서는 하나의 문서에 내용 정보와 구조정보를 가지고 있기 때문에 기존의 내용 정보에 대한 검색뿐만 아니라 논리적인 구조 정보를 검색할 수 있는 기능도 필요하다[1]. 또한 문서의 내용 정보뿐만 아니라 문서의 구조 정보를 이용하여 색인을 추출하고 색인어 가중치를 계산할 수 있다면 검색 효율성을 높일 수 있을 것이다. 일반적으로 중요 문서들은 많은 문서들로부터 참조된다. 즉 참조되는 문서가 많을수록 중요 문서일 가능성이 높다[2]. 따라서 보다 효율적인 XML 문서의 자동색인을 하여 본 논문에서는 각 문서들 간의 연결 구조에 기반으로 인용된 문서간의 관계성을 [3]가 제안한 확률로 계산하고, 확률에 따른 기대치를 태그별 가중치에 업그레이함으로써 색인어 가중치를 결정하는 방법을 제안한다.

2. 관련연구

본 장에서는 본 논문의 테스트 대상인 논문의 XML 구조와 태그별 가중치 기법과 확률에 대해 알아본다. Bob;opgraphic Reference Link, PageRank 기법[2]에서 알 수 있듯 한 문서가 다른 문서로 인용을 포함하고 있거나 포함되고 있을 때 그 문서의 가중치

를 달라진다. 따라서 본 논문에서는 태그별 가중치로 문서에서 발생되는 용어의 가중치를 계산하고, 그 용어의 가중치를 통해서 인용된 문서 간에 관련 있을 확률을 계산하게 된다.

2.1 태그별 색인어 가중치 결정 기법

XML 문서의 자동색인 및 색인어 가중치 결정을 위한 태그의 가중치를 계산하기 위하여, 일정한 XML 테스트 문서 집단(논문이나 연구 보고서)을 만들어 사용자의 검색 행위를 알아본다[4][5][6]. XML 문서에서 색인어를 추출하여 주요 태그마다 태그 용어 벡터와 태그 가중치 벡터를 생성한다. 이때 사용하는 태그는 논문을 작성할 때 자주 사용하는 제목, 목차, 저자, 출처, 키워드, 요약, 서론, 본론, 결론, 참고문헌이다.

2.2 확률 계산

[3]이 제안한 확률(probability)은 어떤 특정한 사건이 일어나는 경향을 말한다. 즉, 확률을 객관적으로 정의한다면, 어떤 결과가 나타날 빈도(frequency)를 확률이라 한다.

어떤 시행에서 일어날 수 있는 모든 경우의 수가 n 가지이고, 각각의 경우가 일어날 가능성이 같다고 할 때 사건 A가 일어날 경우의 수가 a가지이면 A가 일어날 확률 P는 식 2-1과 같다.

$$P(a \parallel n) = \sum_{x \in X} a(x) \cdot \log\left(\frac{a(x)}{n(x)}\right) \quad \cdots\cdots \text{(식 2-1)}$$

어떤 문서가 다른 문서에 인용되었다면 그 인용된 문서나 용어의 가중치는 업그레이드되어야 한다. 따라서 본 논문은 태그별 가중치를 이용해 용어별 가중치를 구한 후 인용된 문서와 인용을 한 문서간의 용어에 대한 관련성을 확률적으로 계산하고 얻어진 확률에 따라 기대치를 유출하여 그 기대치만큼 가중치를 업그레이드하여 좀 더 신뢰할 수 있는 색인어 가중치 계산이 가능하다.

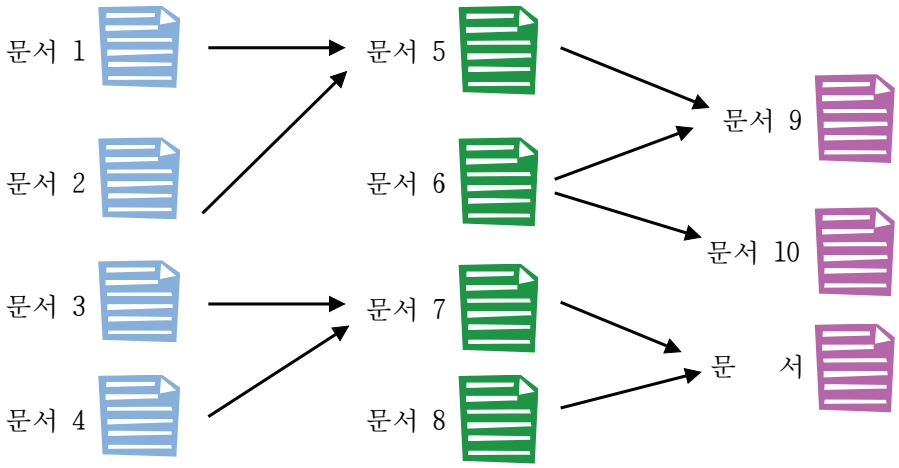
3. XML 태그 가중치를 이용한 색인어 가중치 계산

태그 가중치를 계산하기 위하여 테스트 집합을 준비하여 사용자들의 검색 행위를 관찰하였다. 사용자가 입력한 질의를 구성하는 용어와 일치하는 색인어가 위치하는 태그를 조사한 결과, 요약, 서론, 본론, 결론 태그에 질의가 가장 많이 포함되었다. 그러나 이러한 정보만으로 태그의 가중치를 결정하기엔 미흡한 점이 많다. 즉, 본문의 경우엔 추출된 용어 자체가 많아서 당연히 질의와 일치되는 경우가 많다. 본 논문에서는 [1], [2]에서 제안한 태그별 가중치를 계산하는 기법을 이용하여 문서의 용어에 대한 색인어의 가중치를 나타내는 색인어 가중치 벡터 $W_i = (w_{i1}, w_{i2}, ..., w_{io})$을 생성한다.

3.1 인용문서간의 확률 계산을 적용한 가중치에 대한 기대치 생성

문서 D_i가 문서 D_j에 인용되었을 때 문서 D_j에 인용된 D_i의 특정 키워드는 D_j 특정 키워드 가중치만큼 D_i 특정 키워드 가중치는 기대치를 가지고 있다 할 수 있다. 이 기대치를 계산하기 위해 서지적 참조(Bibliographic Reference)를 통해 서지 정보를 참고하여 유사성을 계산한다.

예를 들어 문서들이 (그림 1)과 같이 연결되어 있다고 가정하자. 또한 논문을 대상하고 하고 있기 때문에 인용관계가 링(Ring)형태를 이루지 않는 조건을 두고 있다. 문서 5가 문서 1에서 인용된 특정 키워드는 문서 1의 해당 키워드가 중요하다는 뜻을 내포하고 있게 된다. 그렇게 된다면 문서 1에서는 해당 키워드의 가중치를 갱신해야 할 필요성이 있게 된다.



(그림 1) 문서간의 인용 관계

3.2 인용된 문서 사이의 확률 계산

문서 간의 확률을 계산하기 위한 초기 값은 3.1절에서 생성한 가중치 벡터를 이용한다.

문서 D_i의 키워드 t이 문서 D_j에 인용됐을 확률을 계산한다.

문서 D_j와 D_i에서 같거나 유사한 키워드만 추출하고 추출된 키워드로부터 서로간의 유사성을 (식 3-1)에 의해 계산한다.

$$cos(D_{i\_t}, D_{j\_t}) = \frac{\sum_{i\_t=1, j\_t=1}^{t} D_{i\_t} \cdot D_{j\_t}}{\sqrt{\sum_{i\_t=1}^{t} D_{i\_t}^2 \cdot \sum_{j\_t=1}^{t} D_{j\_t}^2}} \quad \text{(식 3-1)}$$

여기서 D_{i_t}는 D_i문서의 키워드

D_{j_t}는 D_j문서의 키워드

계산 된 키워드를 대상으로 D_i의 키워드 t이 문서 D_j에 인용됐을 확률 (식 3-2)에 의해 계산한다.

$$P(D_j \parallel D_{i\_t_m}) = \sum_{x \in X} D_j(x) \cdot \log\left(\frac{D_j(x)}{D_{i\_t_m}(x)}\right) \quad \text{(식 3-2)}$$

여기서 D_{i_t}는 D_i문서의 키워드 t

$D_{j_t_i}$는 D_j문서의 m개의 키워드

3.2 확률 벡터 생성

(식 3-2)에 의해 계산된 키워드별 확률을 이용하여 확률이 계산된 키워드 가중치 벡터 $W_Pk = (P_{k1}, P_{k2}, ..., P_{kn})$을 생성한다. 이 때 k는 확률 계산에 사용된 인용 키워드 수이다.

확률 계산된 가중치 벡터는 문서 용어열의 크기와 같다. 확률 계산된 가중치 벡터의 값은 (식 3-2)에 의한다. 확률 계산되지 않은 키워드의 가중치 벡터의 값은 0으로 한다.

3.3 태그 가중치 벡터의 갱신

태그마다 태그 가중치 벡터를 생성하고, 확률벡터가 생성되면, 태그 가중치 벡터와 확률 벡터를 비교하여 태그 가중치 벡터의 값을 갱신한다. 가중치 갱신은 (식 3-3)에 의해 수행된다.

$$W\_doc_k = W\_doc_k + W\_Pk \quad \cdots\cdots\cdots\cdots \text{(식 3-3)}$$

단, W_tag_{jk} : 문서의 k번째 용어의 가중치

W_Pk : k번째 용어의 확률 계산된 가중치

확률 계산된 가중치 벡터의 값인 W_Pk는 k번째 용어의 확률 계산된 가중치 벡터의 값인 W_Pk 값만큼 더해줌으로써 가중치는 갱신된다.

이때, 확률 계산된 가중치는 모두 더해지므로 키워드 가중치가 상대적으로 높아진다.

$$N\_W_{ik} = \frac{W_{ik} - W_{\min}}{W_{\max} - W_{\min}} \quad \cdots\cdots\cdots\cdots \text{(식 3-4)}$$

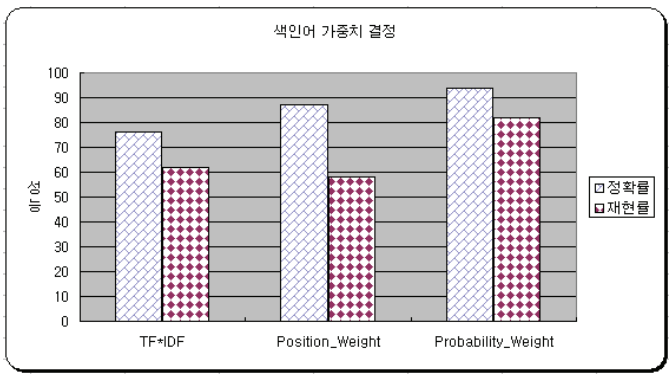
단, $W_{\min}$: W_{ik}중에서 가장 작은 값,

$W_{\max}$: W_{ik}중에서 가장 큰 값

$TF \cdot IDF$값으로 정해진 초기 색인어 가중치는 [0, 1] 범위의 값이 나올 수도 있지만, 문서에서 추출된 어떤 용어가 요약, 서론, 본론에도 포함되어 있을 수 있으므로 문서의 용어가 포함된 태그가중치를 반영하여 계산하면 최종 용어의 가중치는 [0, 1] 범위를 넘을 수도 있다. (식 3-4)에 의해 용어의 가중치를 정규화 시킴으로써 [0, 1] 범위의 값을 갖는 각 문서에 대한 최종 색인어 가중치 N_W_{ik}를 계산한다.

4. 실험 및 평가

용어 추출 후, 위 세 가지 방법에 의해 용어의 가중치를 계산한 후, 가중치를 [0,1] 범위로 정규화한 후, 0.5 이상의 가중치를 가진 용어들을 색인어로 선정하였다. 본 논문에서는 기본적으로 $TF \cdot IDF$를 이용하고 다른 방법들을 적용하고 있으므로, 색인어 선정 기준에 큰 비중을 두지 않았다. 평가자 집단은 전자정보통신 분야의 전공자 33명이 약 3개의 관심분야를 갖고 각 15회에 걸쳐서 실험을 실시하였다. 사용자가 입력한 질의와 일치하는 색인어의 가중치가 0.5이상인 문서만을 검색 결과로 했을 경우, 각 방법들을 이용한 검색결과 평균 정확률과 평균 재현률은 (그림 3)과 같다.



(그림 3) 성능 평가 - 정확률과 재현률

태그가중치의 성능 평가를 위한 두 번째 실험으로서 검색결과로 제시된 문서들을 대상으로 순위결정을 수행한다. 질의와 일치되는 색인어의 가중치가 0.5이상인 문서를 검색 결과로 제시되는 실험만으로는 [0.5, 1]인 값을 갖는 색인어 가중치들의 성능을 제대로 평가할 수 없으므로 문서순위결정을 수행한다. 문서순위 결정 방법은 본 논문의 연구 범위가 아니고, 색인어 가중치의 성능을 평가하기 위한 실험이므로 가장 기본적인 벡터기반 방법을 이용하였다. 질의와 일치하는 색인어의 가중치의 합을 구하여 각 검색된 문서들의 적합성 정도를 계산한다.

"Probability_Weight"는 본 논문에서 제안하는 방법으로 색인어 가중치를 계산한 것을 뜻하고[5], "TF*IDF"는 $TF \cdot IDF$ 방법으로 색인어 가중치를 계산한 것을 뜻한다[6]. "Position_Weight"는 [4]의 방법으로 색인어의 가중치를 계산한 것이다. "Probability_Weight"의 자체 성능 평가는 모든 텍스트로부터 중요한 색인어를 추출하는데 복잡하고 시간 비용이 다소 높았다. 본 논문에서 제안하는 태그가중치 결정 방법을 이용하면 자동색인 뿐만 아니라 검색·문서 순위 결정 기법의 성능 향상에 큰 도움이 될 것이다.

5. 결론 및 향후 연구 과제

웹(World Wide Web; WWW)이 1990년대 중반부터 폭발적으로 성장해 오면서 인터넷을 이용하는 인구의 수도 급격하게 증가하게 되었고, 웹에서 보다 효과적으로 색인을 추출하고 검색 편의성을 제공하기 위한 대안으로서 XML(Extended Markup Language)이 등장하였다.

XML의 태그 정보를 이용하여 검색성능을 향상시키고자, 본 논문에서는 사용자의 검색 행위를 반영하여 XML 태그가중치를 계산하는 방법을 제안하였다. 그리고 결정된 태그의 중요도를 성능 평가하기 위해 두 가지 실험을 실시하였다. 첫 번째 실험은 본 논문에서 제안한 방법으로 특정 분야의 테스트 문서 집합에서 태그가중치를 계산하여 태그의 중요 순위를 결정하는 실험이다. 태그의 중요도를 구하는 실험 결과, 태그의 중요 순위는 키워드, 제목, 초록, 참고문헌, 본론, 서론, 결론, 목차 순이었다. 또한 키워드와 제목 태그의 가중치 차이가 매우 적었다. 이 순서는 실험 대상 문서의 종류에 따라 태그가 다르기 때문에 다른 부류의 문서의 경우엔 본 실험과 다른 결과가 나올 수 있다. 두 번째 실험으로, 태그 가중치를 반영하여 색인어 가중치를 결정한 후 검색 성능을 평가해 본 결과, 본 논문에서 제안하는 방법의 정확률과 재현율, 그리고 적합률이 모두 좋은 성능을 나타냄을 알 수 있었다. 본 논문에서 제안하는 태그의 중요도 결정 방법을 이용하여 색인어 가중치를 결정하면, 사용에게 보다 적합한 검색 결과를 제공할 수 있을 것이고, 문서순위결정 방법과 같이 사용되어 사용자에게 검색 편의성을 제공할 수 있을 것이다.

본 논문에서 제안하는 방법이 기술 분야 논문이 아닌 일반적인 XML 문서의 태그들에도 적용될 수 있는가에 대한 검증이 필요하다. 그리고 태그가중치 결정 과정 중에서 태그 가중치 벡터의 생성과 갱신 부분에 문서빈도(DF)와 용어빈도(TF)가 미치는 영향에 관하여 연구를 계속할 계획이다. 또한 질의 가중치 벡터의 가중치를 0과 1이 아닌 사용자의 선호도를 반영할 수 있는 가중치로 표현했을 경우의 효과에 대해서도 연구를 계속할 계획이다.

참고문헌

[1] Brian Lowe, Justin Zobel, Ron Sacks-Davis "A Formal Model for Databases of Structured Text", Proceedings of the Fouth International Conference on Database Systems for Advanced Applications(Dasfaa '95), pp. 449-456, 1995.

[2]Taher H.Haveliwala, " Topic-Sensitive PageRank:A Context-Sensitive Ranking Algorithm for Web Search", IEEE transactions on knowledge and data engineering, VOL 15, NO 4, pp.784-796 2003

[3] A. Maedche "Ontology Learning for the Semantic Web", Kluwer Academic Publishers, pp. 251-263, 2002a

[4] 우선미, "사용자 질의를 이용한 XML 태그의 가중치 결정." 정보처리논문지 D(정보처리 응용), 2005

[5] 김홍남, 이기성, 조근식 "가중치가 부여된 규칙을 이용한 문서 분류}, 한국 정보 과학회지, 제 30권, 제 2-1호 pp. 0154~ 0156 2003

[6] 김종영, 김철수 "가중치를 가지는 웹문서 색인기법에 관한 연구", 한국정보처리학회, 제 09권, 제 02호 pp. 0000~0000, 2002