단락을 분류에 따른 XML 키워드 가중치 결정 기법

정혜진*, 김형진*
*전북대학교
e-mail:hi-jin@hanmail.net, kim@chonbuk.ac.kr

An XML Keyword Indexing Method Using on Lexical Similarity

Hye-Jin Jeong*, Hyoung-Jin Kim*
*Chonbuk National University

요 약

보다 효과적인 키워드 추출 및 키워드 가중치 결정을 위하여 문서의 내용뿐 아니라 구조를 이용하여 색인을 추출하는 연구가 이루어지고 있는데, 대부분의 연구들이 XML 단락별 중요도가 아닌, 문맥상의 단락에 대한 중요도를 계산하는게 일반적이다. 이러한 기존 연구들은 대부분이 객관적인 실험을 통해서 중요도를 입증하기보다는 일반적인 관점에서 단순한 수치로 중요도를 결정하고 있다. 본 논문에서는 웹 문서 관리를 위한 표준으로 자리잡아가고 있는 XML 문서의 자동색인을 위하여, 논문을 구성하는 주요 단락을 세분하고, 단락에서 추출된 용어의 가중치를 갱신해 가면서 최종 색인어 가중치를 계산하는 방법을 제안한다.

1. 서론

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, XML의 엘리먼트는 데이터를 해석하는 데에 사용할 수 있기 때문에 XML의 역할과 중요성이 인식되고 있다[1]. 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[2]. 또한 문서의 제목이나 요약부분에서는 찾을 수 없지만 차선의 핵심 키워드인 경우 그 키워드를 통해 문서가 검색될 수 있는 효율성도 필요하다.

따라서 본 논문에서는 XML의 구조 정보를 바탕으로 하여 단락(태그)을 세분화하고, 단락별로 용어 가중치를 측정하여 제목에 큰 비중을 두고 있는 문서가 아닌 문서라 할지라도 효율적인 검색이 이루어 질 수 있는 방안을 제시하고자 한다.

먼저 2장에서 [3]과 [4]가 제안한 키워드 가중치 계산법에 대해 간단히 살펴보고 3장에서는 본 논문에서 제안하는 XML 문서의 가중치 계산 방법을 설명한다. 4장에서는 본 논문에서 제안하는 방법으로 계산된 가중치가 검색성능에 미치는 영향을 알아보기 위해 실험하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

XML 문서에서 각 색인어에 부여되는 가중치는 XML 문서의 기본 단위인 엘리먼트별로 계산되는 것이 바람직하다[4]. 단어에 가중치를 부여하는 목적은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라 키워드로서 상대적 가치를 표현하기 위함이다. 자동색인 기법에서 키워드 가중치 결정은 주로 통계적 기법을 이용하는데, 통계적 기법의 통계적 기준은 모두 단어의 출현빈도에 근거하고 있다. 단어빈도를 문헌빈도로 나누어주는 역문헌빈도($tf \bullet idf$)에 의한 키워드 후보의 가중치 기법[4]과 표제 키워드 후보의

가중치 기법 등이 많이 사용되고 있다. 이러한 방법과 XML 단락의 중요도를 이용한 방법을 설명하고자 한다.

2.1 [3]이 제안한 키워드 가중치 계산

XML 문서의 경우 텍스트 엘리먼트에서 키워드의 가중치($eW_{i,j,k}$)는 기존의 $tf \bullet idf$ 공식[5]를 변형한 식 1을 적용하여 계산한다.

$$eW_{i,j,k} = etf_{i,j,k} \bullet ief_i \bullet es_l \quad (1)$$

- $eW_{i,j,k}$: 문서 k의 엘리먼트 j에서 키워드 i의 가중치
- $etf_{i,j,k}$: 문서 k의 엘리먼트 j에서 키워드 i의 빈도수
- ief_i : 키워드 i가 나타나는 엘리먼트의 수에 대한 역 엘리먼트 출현 빈도(inverse element frequency)

$$ief_i = log_e \frac{(eN+1)}{ef_i}$$

- ef_i : 키워드 i가 나타나는 엘리먼트의 수
- eN : 전체 엘리먼트의 수
- es_l : 엘리먼트 타입 l에 주어지는 중요도

이 경우 키워드 가중치가 색인처가 출현된 엘리먼트 수에 영향을 받기 때문에 본문과 같은 경우 출현빈도(ief_i)가 높아지게 되면, 그 엘리먼트 타입에 낮은 중요도(es_l)를 주게 된다 하더라도 자칫 문서의 제목이나 키워드처럼 발생 빈도는 낮아도 해당 엘리먼트에서 타입에 높은 중요도(es_l)를 준 경우와 별 차이가 없는 상황이 발생할 수 있다.

2.2 [4]가 제안한 키워드 가중치

단어의 빈도수(Term Frequency)와 각각의 단락에 부여된 가중치(Weight)로 문서에 대한 단어의 지지도(Term Support)를 측정하여 문서에 대한 단어들의 중요도가 높은 단어들로 연관 규칙을 적용한다.

$$Sup_{t_i},p = \frac{sup'_{t_i,p}}{MAX\{sup'_{t_i,p}\}} \quad (2)$$

- $Sup'_{t_i},p = \sum_{s_j} tf_{ij} \bullet ws_j$
- tf_{t_i} : 단락(엘리먼트)에 있는 t_i의 단어 빈도수
- ws_j : 단락(엘리먼트) s_j의 가중치 값

이 경우 단락(태그)의 가중치에 많은 영향을 받고 있기 때문에 단락에 좀 더 정확한 가중치 값이 주어지지 않는 경우 적용률이나 재현율에 치명적인 악영향을 미칠 수 있다.

따라서 본 논문에서는 XML 문서의 구조정보를 이용하여 단락(태그)을 나누고 태그별로 가중치를 정해진 태그별 가중치에 의존하지 않고 가중치를 계산할 수 있는 방법을 제안하고 이를 적용하여 좀 더 신뢰할 수 있는 키워드 가중치 계산이 가능하다.

3. XML 태그 가중치를 이용한 키워드 가중치 계산

본 장에서는 학위 논문을 대상으로 XML 문서의 태그 가중치를 계산하고, 이 태그의 중요도를 이용하여 키워드 가중치를 계산하는 방법을 기술한다.

학위 논문의 구성도를 보면 그림 1과 같다고 할 수 있다.

본 논문에서는 문서를 그림 1과 같이 단락을 세분화한다. 그리고 각 단락별로 키워드를 추출하고 가중치를 계산한다. 측정된 가중치를 이용하여 해당 단락에서의 지지도를 계산한다. 문서의 크기가 크면 클수록 단어 빈도수가 높게 나타나게 되기[6] 때문에 연구내용이나 실험 및 평가부분에서 용어의 빈도수가 높게 나타난다고 해서 문서 전체의 지지율에 영향을 줘서는 안 되기 때문이다.

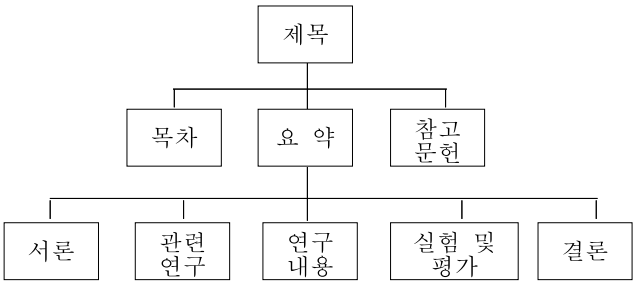

그림 1. 논문의 구성도

따라서 본 논문에서 제안하고자 하는 가중치 결정 기법의 흐름도는 그림 2와 같다.

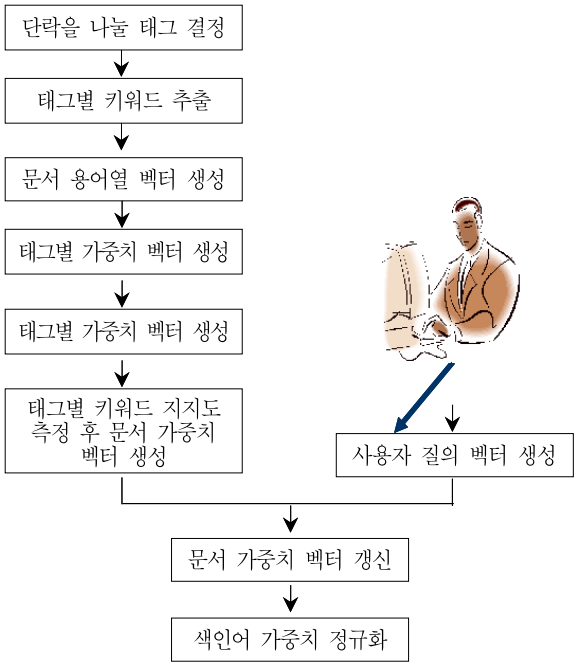

그림 2. 가중치 결정 기법의 흐름도

3.1 문서 용어열 벡터 생성

색인어 선정에 사용하기 위하여 문서집합에서 추출한 용어로 문서 용어열 $T_doc = (dt_1, dt_2, ..., dt_n)$을 생성한다. 이때 n은 XML 문서 집합을 구성하는 문서에서 추출한 키워드의 수이다.

3.2 태그별 가중치 벡터 생성

태그별 가중치 벡터는 문서 용어열 $T_tag_i = (tag_t_{i1}, tag_t_{i2}, ..., tag_t_{in})$과 쌍을 이루는 가중치 벡터 $W_tag_i = (tw_{i1}, tw_{i2}, ..., tw_{in})$을 생성한다. 이때 벡터의 크기는 문서 용어열 T_doc와 같다. 그리고 각 태그별 용어열에 대응되는 가중치를 나타내는 태그별 가중치 벡터 W_tag_i는 T_tag_i에 대응되는 가중치 벡터이다.

태그별 가중치 벡터는 문서 용어열의 용어를 포함하지 않는 부분은 0값을 가지고, 문서 용어열의 용어를 포함하는 부분은 역문헌빈도($TF \cdot IDF$) 방법을 이용하여 태그별 용어 가중치 벡터 $W_tag'_{ij}$을 생성한다. $W_tag'_{ij}$는 i번째 태그(tag)의 j번째 용어 가중치를 의미하는 것이다.

3.3 태그별 키워드 지지도 측정 후 문서 가중치 벡터 생성

측정된 $W_tag'_{ij}$는 해당 단락에 지지도 값 W_tag_{ij}을 식 3로 계산할 수 있다. 이때 계산된 태그별 키워드 지지도는 키워드별로 더해서 문서 가중치 벡터 $V_doc'_{kj}$을 식 4에 따라 생성한다.

$$W\_tag_{ij} = \frac{W\_tag'_{ij}}{MAX\{W\_tag'_{ij}\}} \quad (3)$$

$$V\_doc'_{kj} = \sum W\_tag_{ij} \quad (4)$$

- W_tag_{ij} : i번째 태그의 j번째 용어에 대한 지지도
- $V\_doc'_{kj}$: k번째 문서의 j번째 용어에 대한 지지도

3.4 질의 벡터 생성

사용자가 입력한 질의로 질의열 $T_query_m = (q_{m1}, q_{m2}, ..., q_{mn})$와 질의 가중치 벡터 $W_query_m = (qw_{m1}, qw_{m2}, ..., qw_{mn})$을 생성한다. 이때 m은 사용자가 입력한 질의의 수이다.

질의열과 질의 가중치 벡터는 문서 용어열의 크기와 같다. 질의 가중치 벡터의 값은 0 또는 1을 가진다. 질의열을 구성하는 용어가 문서 용어열을 구성하는 용어와 일치하면 가중치를 1로 주고, 그렇지 않은 용어의 가중치는 0을 준다.

3.5 문서 가중치 벡터 생성

문서의 가중치 벡터와 질의 벡터가 생성되면, 두 벡터를 비교하여 문서 가중치 벡터의 값을 갱신한다. 질의를 구성하는 용어와 일치하는 용어에 해당하는 키워드에 대해 가중치 벡터의 가중치를 갱신되고 질의와 일치하지 않는 가중치는 갱신되지 않는다. 가중치 갱신 방법은 식 5와 같다.

$$W\_doc'_{kj} = V\_doc'_{kj} \bullet W\_query_m \quad (5)$$

3.6 문서 가중치 벡터 생성

용어가 여러 태그에 중복 위치한 경우 각 태그별 지지도 값인 W_tag_{ij}는 키워드별로 모두 더해지므로 키워드 가중치가 상대적으로 높아진다.

키워드 가중치는 [0, 1] 범위의 값이 나올 수도 있지만, 최종 색인어의 가중치는 [0, 1] 범위를 넘을 수도 있다. 식 6에 의해 용어의 가중치를 정규화 시킴으로써 [0, 1] 범위의 값을 갖는 각 문서에 대한 최종 키워드 가중치 W_doc_{kj}을 계산한다.

$$W\_doc_{kj} = \frac{W\_doc'_{kj} - W_{\min}}{W_{\max} - W_{\min}} \quad (6)$$

- $W_{\min}$: $W_doc'_{kj}$중에서 가장 작은 값
- $W_{\max}$: W_tag_k중에서 가장 큰 값

4. 실험 및 평가

본 장에서는 실험평가를 실시한다. 용어의 태그별 지지도를 반영하여 키워드 가중치를 결정한 후, 일반적인 검색 성능 평가 척도를 이용하여 성능을 평가한다. 또한 본 논문에서 제안하는 용어의 태그별 지지도를 반영하여 계산된 키워드 가중치의 성능을 확신하기 위하여 문서 순위 결정을 수행한 후, 상위 문서의 적합성 정도를 평가한다. 본 논문에서는 실험 평가를 위하여 문서 제목이나 요약에 비중을 크게 두고 있는 문서와 적게 두고 있는 문서의 정확률(precision ratio)과 재현율(recall ratio), 그리고 적합률(relevance ratio)[7]을 평가한다.

XML 태그별 키워드 지지도를 이용한 색인어 가중치 계산에 대한 검색 성능을 실험 평가하기 위한 실험 환경은 다음과 같다.

① 실험 데이터 : 웹에서 검색한 컴퓨터과학 및 정보통신 분야의 논문과 기사를 XML로 재구성한 데이터 약 300개

② 실험 참가자 : 컴퓨터과학 및 정보통신 분야의 석·박사학위 과정 이상인 전공자 25명

③ 키워드 추출범위와 방법 : 키워드 추출 범위는 문서의 전체이고 키워드 추출 방법은 자동 색인을 이용한다.

본 논문에서는 실험 평가를 위해 KT-Set의 문서를 XML로 변환하여 사용한다. KT-Set은 컴퓨터과학

및 정보통신 분야의 논문과 기사를 합하여 총 4,414개의 요약으로 구성되어 있다.

4.1 실험 평가

실험은 태그별 키워드 가중치를 반영한 검색 성능 평가를 위한 실험이다.

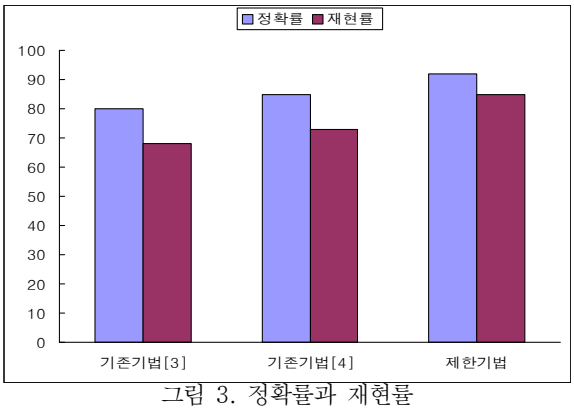

그림 3. 정확률과 재현률

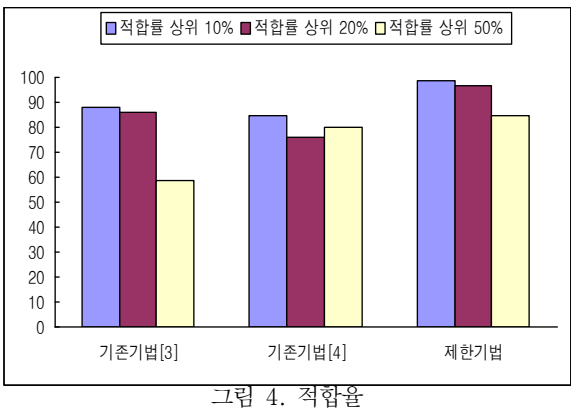

그림 4. 적합율

일반 정보검색 평가 척도인 정확률(precision ratio)과 재현율(recall ratio)을 평가한 결과는 그림 3과 같고, 문서순위결정 평가에 맞는 적합률(relevance ratio)은 그림 4와 같다.

본 논문은 [3]과 [4]의 기존 기법과 본 논문에서 제한한 기법을 비교하여 정확률과 재현률 그리고 적합율의 우수함을 보였다.

5. 결론

본 논문에서는 XML을 구성하고 있는 단락을 분석하고 단락별로 추출된 키워드의 지지도를 측정하였다. 이에 단락별 키워드의 지지도를 반영하여 키워드 가중치를 결정한 후 검색 성능을 평가해 본 결과, 출현빈도에 영향을 크게 받지 않아도 단락별 키워드의 지지도를 적용하여 키워드 가중치를 결정하면, 사용에게 보다 적합한 검색 결과를 제공할 수 있다. 또한 문서순위결정 방법과 같이 사용되어 사용자에게 검색 편의성을 제공할 수 있을 것이다.

앞으로 복잡한 가중치 계산으로 인한 연산 시간에 미치는 영향과 사용자의 선호도를 반영할 수 있는 클러스터링 기법을 적용하여 가중치로 표현하기 위한 방법에 관하여도 연구를 계속할 것이다.

참고문헌

[1] 김영란, “XML DTD의 효율적인 검색을 위한 구조 정보 및 인덱스 메카니즘”, 컴퓨터정보학회 논문지 제 8권 제 2호, 2003.

[2] 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.

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

[4] 한예지, 한창우, 서동혁, 김수희 “XML 문서의 내용 기반 검색을 위한 인덱싱 모델 및 색인어의 가중치 부여”, 한국정보과학회 발표논문집 제31권 제1호 (B), pp. 103~105, 2004. 4.

[5] Salton G., “Automatic Text Processing : The Transformation , Analysis, and Retrieval of Information by Computer”, Addison- Wesley Publishing Company, 1989.

[6] 박기선, 정혜경, 이근용, 이용석 “자연어 질의 문맥 구조를 이용한 효과적인 정보검색”, 한국인터넷정보학회 논문집 제6권 제2호, pp. 427 ~ 431, 2005. 11.

[7] 우선미, 유춘식, 김용성, “용어 연관성 분석을 이용한 사용자 위주의 문서순위결정 기법”, 한국정보과학회 논문지, 제28권, 제2호, pp. 149-156, 2001.