엘리먼트 기반 XML 문서검색의 성능에 관한 실험적 연구*

An Experimental Study on the Performance of Element-based XML Document Retrieval

윤 소 영 (SoYoung Yoon)**
문 성 빈 (Sung-Been Moon)***

초 록

이 연구에서는 가장 적합한 엘리먼트 기반 XML 문서검색 기법을 제시하기 위해 언어모델 검색 접근법으로 다이버전스 기법, 보정 기법 그리고 계층적 언어모델의 검색성능을 평가하는 실험을 수행하였다. 실험 결과, 가장 효율적인 검색 접근법으로 문서의 구조정보를 적용한 계층적 언어모델 검색을 제안하였다. 특히, 계층적 언어모델은 실제 검색에서 중요성을 가지는 검색순위 상위에서 뛰어난 성능을 보였다.

ABSTRACT

This experimental study suggests an element-based XML document retrieval method that reveals highly relevant elements. The models investigated here for comparison are divergence and smoothing method, and hierarchical language model. In conclusion, the hierarchical language model proved to be most effective in element-based XML document retrieval with regard to the improved exhaustivity and harmed specificity.

키워드 : XML, 엘리먼트 검색, 내용기반 검색, 계층구조, 언어모델, 다이버전스, 보정, 계층적 언어모델

---

* 본 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.
** 국사편찬위원회 자료정보실(syoon@history.go.kr)
*** 연세대학교 문헌정보학과 교수(sbmoon@yonsei.ac.kr)

1. 서 론

XML 문서는 문서의 논리적 구조에 기반하여 계층적인 구조를 갖도록 구성되어 있으며, 각 구성요소인 XML 엘리먼트가 실제 정보검색의 검색단위가 될 수 있다. 따라서 원칙적으로 XML 문서를 구성하는 텍스트 엘리먼트 각각이 질의에 대한 검색 결과가 될 수 있다.

정보검색 측면에서는 XML 문서의 구조정보를 이용한 정보검색 성능의 향상을 기대할 수 있다. 전통적인 정보검색의 방식은 정보 요구가 발생했을 때 전체 문서를 검색하여 결과를 보여주는 반면에, 최근 정보검색 시스템은 문서의 구조와 검색결과 측면에 제한을 두는 더 세분화된 검색 패러다임을 구현하고 있다. 즉 전체 문서를 검색하는 것이 아니라 이용자의 정보 요구에 부응하면서도 문서보다 작은 크기의 특정 부분을 검색하려는 목표를 가지고 있다. 이러한 방법으로 정보검색이 수행되면 이용자는 더 작은 단위, 즉 문서의 일부를 검색결과로 얻을 수 있게 되어 검색 결과를 살펴봐야 하는 시간과 수고를 줄일 수 있게 된다.

엘리먼트 기반으로 한 XML 문서검색 실험 이전에도, 문서의 구조를 이용하여 문서의 일부인 섹션이나 문단 등의 단락(passage)을 검색단위로 하여 정보검색의 성능을 향상시키고자 한 연구들이 있었다(Salton, Allan, and Buckley 1993; Wilkinson 1994; Moffat et al. 1994). 연구 결과, 문서의 물리적 구조에 기반하여 선형적으로 문서를 나누어 검색을 수행하였을 경우에는 문서검색보다 오히려 성능이 좋지 않았으나(Wilkinson 1994), 구조정보를 추가정보로 사용하였을 경우에는 검색성능이 향상되는 결과를 보여주었다(Salton, Allan, and Buckley 1993).

이 연구에서는 전통적인 정보검색 질의인 내용질의(content-only)를 가지고 XML 엘리먼트 검색에 초점을 맞추어 XML 문서검색을 위한 효율적인 정보검색방법을 제시하고자 한다. 이와 같은 목적으로 벡터공간모델 검색과 확률모델검색을 기초로 한 언어모델 검색, 그리고 이 두 가지 검색 모델 각각에 XML 문서의 구조 정보를 적용한 계층적 검색기법의 성능을 평가하는 실험을 수행하여 가장 효과적인 엘리먼트 기반 XML 문서검색 알고리즘을 제시하고자 하였다.

2. 이론적 배경

2.1 XML 문서검색

전통적인 정보검색과 XML 문서검색의 기본적인 차이는 바로 검색단위라 할 수 있다. 전통적인 정보검색에서는 검색단위가 대부분의 문서 전문

(full-text)으로 고정되어 있는 반면, XML 검색에서는 모든 XML 엘리먼트가 검색 가능한 단위가 된다. 이러한 차이점이 XML 검색을 더 복잡하고 어렵게 만드는데 그 이유는 검색결과가 적합해야 할뿐만 아니라 검색된 결과로 제시되는 엘리먼트의 크기에 있어서도 그 크기가 너무 크면 문서검색과 거의 동일하고, 너무 작으면 의미를 충분히 전달할 수 없다는 문제가 발생하기 때문이다.

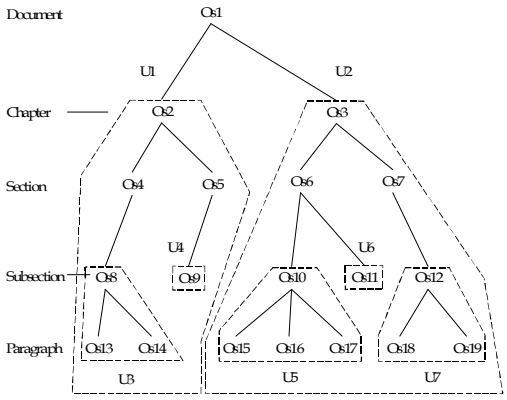

그림 1. FERMI 모델에서 규정한 문서의 구조 (Chiaramella, Mulhem, and Fourel 1996, 33)

정보검색은 계층적 내부 구조를 가지는 문서의 검색이라는 관점에서 볼 수 있는데, 예를 들어 문서는 섹션(section), 문단(paragraph) 그리고 문장(sentence)을 가지며 각 문서 구조의 요소들은 검색에 이용될 수 있는 검색 단위가 될 수 있다(Callan 1994). FERMI(Formalisation and Experimenta-tion on the Retrieval of Multimedia Information) 모델에서는 그림 1과 같이 문서를 구조적인 측면에서 분석하여 트리구조로 표현하고 트리의 수준(level)을 구분하여 검색단위가 될 수 있는 요소를 점선으로 그룹화하여 표현하고 있다(Chiaramella, Mulhem, and Fourel 1996).

문서보다 작은 단위의 하위 구성요소를 검색하려는 시도는 기존 정보검색의 단락검색(passage retrieval)에서 이미 찾아볼 수 있다. XML 엘리먼트 검색도 문서의 일부를 검색한다는 의미로 보면 단락검색에 포함되는 것으로 볼 수 있다. 단지 차이라고 한다면 XML 문서검색에서는 단락을 구분하는 방법으로 XML 문서의 엘리먼트를 그대로 이용한다는 것이다. 따라서 이 연구에서는 단락과 엘리먼트를 동일한 의미로 사용하였다.

2.2 언어모델 기반 검색

정보검색의 언어모델링 접근법은 웹검색과 교차언어 검색과 같은 다양한 검색과정을 기술하기 위한 공식적인 프레임워크를 제공하며, 구조 문서를 모델링하기 위한 자연스러운 환경을 제공한다. 기본적인 개념은 각 문서의 언어모델을 추정하여 추정된 모형에 따른

가능성 있는 질의로 문서의 순위를 매기는 것이다.

1998년 이후로 통계적 언어모델이 정보검색에 적용되었는데, Ponte와 Croft(1998)가 처음 정보검색에 언어모델의 이용을 제안하였으며 작은 문서 크기 문제를 해결하기 위해 위험함수(risk function)를 이용하였다.

2.2.1 다이버전스(divergence) 기법

Amati와 Rijsbergen(2002)은 정보검색의 확률 모델을 유도하기 위한 프레임워크를 소개하였는데 이 모델은 언어모델 접근법이므로 비-매개변수 정보검색 모델이라고 할 수 있다. 용어가중치 모델은 무작위로 얻어낸 실제 용어 분포의 다이버전스를 측정하여 유도하였으며 이를 위해 이항분산과 Bose-Einstein 통계를 적용하였다. 문서에 질의를 매칭하는 과정에서 용어 가중치를 조정하기 위해서 두 가지의 단어빈도 정규화를 이용하였다. 첫 번째 정규화는 문서가 동일한 길이를 가진다고 가정하고 적합문서에 출현하는 용어의 확률(information gain)을 측정하는 것이다. 두 번째 정규화는 문서 길이 통계값을 적용하는 방법이다.

2.2.2 보정(smoothing) 기법

보정 기법은 값들이 어떤 부드러운 곡선을 기준으로 랜덤(random)하게 이탈하여 얻어진 것이라는 전제하에 그 부드러운 곡선의 패턴을 찾아내자는 방법이다. 이 기법은 모델의 정확성을 향상시키기 위해 언어모델의 확률을 재추정하는 것으로 일반적으로 추정하는 많은 언어모델이 확률 분포의 작은 샘플에 기반하고 있다는 사실을 바탕으로 생겨났다. 정보검색에서 보정 기법은 매우 작은 양의 텍스트에서 언어모델을 구축할 수 있도록 하기 때문에 매우 유용하다.

관찰된 텍스트가 있을 때 언어모델을 추정하는 가장 직접적인 방법은 기본적으로 다형모형(multinomial model)을 가정했을 때 최우추정(maximum-likelihood estimate)을 이용하는 것이다. 그러나 최우추정은 단문 텍스트의 저빈도어를 추정할 때는 0확률을 부여하는 문제가 발생된다. 0의 로그값은 음의 무한대가 되므로 문서 언어모델을 추정하는데 있어 문제가 발생하게 된다. 이러한 문제를 해결하기 위한 방법이 보정 기법으로 역문헌빈도와 유사한 영향력을 가진다.

문서나 엘리먼트 점수는 일반적으로 문서나 엘리먼트 모형이 주어지면 단어 확률 로그값의 합이 되므로, 검색성능은 일반적으로 보정 기법(smoothing)의 매개변수에 영향을 받게 된다. 검색에서 가장 효과적인 보정 기법 프로시저(smoothing procedure)는 선형보간법으로 알려져 있다(Zhai and Lafferty

2001; Miller, Leek, and Schwartz 1999).

2.2.3 계층적 언어모델

구조문서 검색을 위한 다양한 모델들이 소개되었으나, 계층적 언어모델은 트리의 노드로부터의 정보를 결합하고 용어가중치를 추정하는 방법 등을 제공한다는 점에서 차이가 있다(Ogilvie and Callan 2003).

계층적 구조를 가진 문서는 트리로 표현될 수 있는데 그 트리의 노드가 문서의 요소에 대응한다. 트리의 각 노드를 위해 문서 내용에서 생성분포를 추정할 수 있으며 노드의 분포는 노드의 텍스트, 자식노드의 분포, 또는 부모노드의 분포 등을 이용하여 추정될 수 있다. 이러한 방식을 이용하면 계층적 구조를 가진 문서를 간단하고 유연하게 표현할 수 있다.

문서의 구문분석 트리는 구조를 따라 만들어지는데, 예를 들어 XML 스키마에서 문서를 제목, 초록, 그리고 본문 텍스트라고 정의하고 문법에 대응시키면 다음과 같다.

문서 → 제목 초록 본문

여기에서 제목이나 초록과 같은 노드는 말단 노드가 되는데, 전통적인 문맥자유문법에서 말단 노드는 그 안에 부가적인 구조를 가지지 않는 텍스트 단위가 되며 이 말단 노드를 위한 언어모델은 그 노드의 텍스트로부터 추정된다.

앞의 예와 같이 문서 언어모델이 제목, 초록, 그리고 본문 언어모델로 구성된다고 명시했으면 그 다음에는 그 자식 언어모델을 어떻게 결합할지 명시해야 한다. 이 모델에서는 선형보간법을 이용하여 요소, 자식 요소, 그리고 부모 요소 등에서 추출된 정보를 결합하여 언어모델을 추정한다.

3. 엘리먼트 기반 XML 문서검색 실험

3.1 실험 개요

실험과정은 그림 2에서 보는 바와 같이 XML 문서를 만들고 기초 실험 결과로 이용할 문서검색을 위한 문서색인과 XML 문서검색을 위한 엘리먼트 색인을 구축하였다. 엘리먼트 기반 XML 문서검색을 위한 정보검색 결과의 성능을 비교, 평가하기 위하여 언어모델 검색 실험을 수행하였다.

언어모델 검색은 다이버전스 기법, 보정 기법, 그리고 XML 문서의 트리구조에 기반하여 구조정보를 추가한 계층적 언어모델 검색 실험으로 구성된다. 다이버전스 기법 실험에서 중복색인을 허용하는 서브트리 색인과 언어모델을

결합하여 내용만을 대상을 하는 XML 문서검색을 수행하였다. 다이버전스 기법을 적용한 언어모델 검색은 정보검색 측면에서 보면 비-매개변수 모델이라 할 수 있다. 용어가중치는 무작위 과정에서 얻어진 값에서 실제 용어 분포의 다이버전스를 계산하여 구한다. 다이버전스 기법을 적용한 검색 실험을 위하여 이항모형(binomial model)과 Bose-Einstein 모델을 선택하였다.

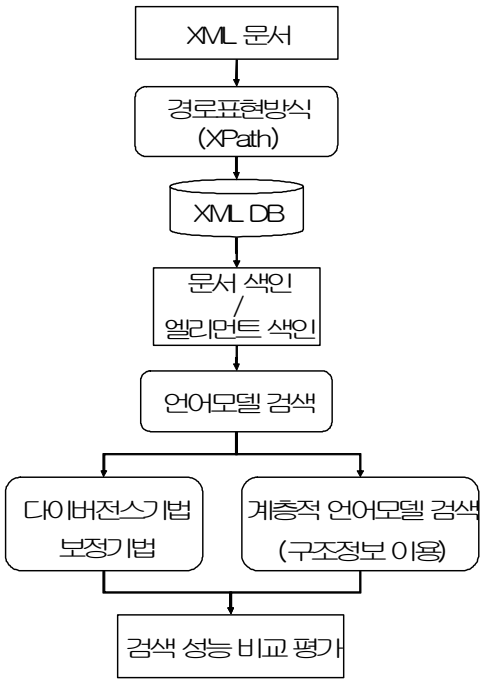

그림 2. XML 문서검색의 실험과정

보정 기법을 적용한 검색 실험에서는 다항언어모델(multinomial language model)인 Jelinek-Mercer 보정을 적용하였다. 실제 검색은 먼저 엘리먼트 각각을 위한 언어모델을 추정한 후, 컬렉션 언어모델을 이용하여 재추정한 언어모델이 주어지면 문서에 질의가 나타날 가능성에 따라 엘리먼트의 순위를 매기는 방법으로 진행된다.

마지막으로, 계층적 언어모델을 이용한 XML 엘리먼트 검색 실험을 수행하였다. 이 언어모델은 언어모델의 확률을 추정할 때 트리의 노드로부터의 정보를 결합하여 재추정한다는 점에서 앞서의 언어모델 검색과 차이가 있다.

엘리먼트 기반 XML 문서검색 실험을 위한 실험문서집단은 학위논문이나 학술지와 같이 논리적 계층 구조를 가진 문서의 전문(full-text)이며 XML 구조로 표현되어 있어야 한다. 그러나 일반적인 정보검색에서 사용하는 실험문서는 XML 구조를 지원하지 않을 뿐 아니라 실험에서 필요한 명확한 논리적 계층 구조를 가진 경우가 거의 없다.

이 실험에서는 실험환경에 맞는 실험집단을 구축하기 위해 ACM이 출판하는 학술회의 논문집 중에서 1998년에서 2004년 사이에 출판된 정보검색 및 정보기술 관련 주제를 다루고 있는 SIGIR 281건, SIGMOD 396건, CIKM 228건, 그리고 ACM DL 125건 등 총 1030건을 수집하였다. 실험문서집단의 통계적 특성은 표 1과 같다.

표 1. 실험문서집단에 대한 통계치

문서 수	1030
평균 XML 노드 수	810
평균 XML 트리 깊이	6.4
섹션(section) 수	10,356
서브섹션 수	7,383
질의문 수	30
질의문 당 평균 적합 문서 수	10.5

3.2 다이버전스 기법 실험

이 실험에서는 문서의 단어가중치를 확률로 구하는 정보검색의 언어모델 접근법과 같이 이질적인 확률분포를 조합하여 비-매개변수 모델을 유도하는 실험을 수행하였다.

$$w = (1 - \mathrm{P}rob_2) \cdot (-\log_2 \mathrm{P}rob_1) = Inf_2 \cdot Inf_1 \quad (1)$$

비-매개변수 접근법이 갖는 이점은 학습집단으로부터 매개변수를 학습해야 하는 데이터 유도방식이 필요 없다는 것이다. 다이버전스 기법의 실험을 위하여 이항모델(binomial model)과 Bose-Einstein 모델을 적용하였다.

첫째, 다이버전스를 가진 이항모델의 추정으로 F용어의 발생은 N문서에서 독립적으로 분포되어 있다고 가정한다.

둘째, Bose-Einstein 모델은 N문서 내 F단어빈도의 모든 가능한 분포와 현재 문서가 tf을 가지는 모든 사건을 고려하게 된다.

3.2.1 1차 정규화

공식 (1)에서 $Inf_2 = (1 - Prob_2)$은 첫 번째 정규화로 단어빈도 정규화라고 할 수 있는데, $Prob_2$는 tf가 있을 때, 엘리먼트의 또 다른 단어의 발생을 관찰할 수 있는 확률로 정의된다. 이를 위해 Lapalce의 계승 규칙(Feller 1968)과 Bernoulli 프로세스 비율을 고려하였다.

정규화 L : Laplace의 계승 규칙에 근거하여 다음과 같이 변형된다. 정규화 L에서는 단어의 빈도만을 가지고 정규화를 수행하게 된다.

$$Prob_2(tf) = \frac{tf}{tf+1}$$

공식(1)에 정규화 L을 적용하면 다음과 같은 첫 번째 정규화 L을 얻을 수 있다.

$$w(t,E) = \frac{1}{tf+1} \cdot Inf_1(tf)$$

정규화 B : 공식(1)과 같은 Bernoulli 프로세스 비율을 고려하면 다음과 같은 공식이 된다. 정규화 B에서는 단어빈

도 외에 엘리트 집합의 크기까지 고려하여 단어빈도의 정규화를 수행하게 된다.

$$Prob_2(tf) = 1 - \frac{B(n, F+1, tf+1)}{B(n, F, tf)} = 1 - \frac{F+1}{n \cdot (tf+1)}$$

n : 단어의 엘리트 집합의 크기.
단어가 출현한 문서의 수.

공식(1)에 정규화 B을 적용하면 다음과 같은 첫 번째 정규화 B를 얻을 수 있다.

$$w(t,E) = \frac{B(n, F+1,\ tf+1)}{B(n,\ F,\ tf)} \cdot Inf_1(tf) = \frac{F+1}{n \cdot (tf+1)} \cdot Inf_1(tf)$$

3.2.2 2차 정규화

위의 매개변수들은 엘리먼트의 길이를 고려하지 않았으므로 2차 정규화를 통해 엘리먼트 길이와 단어빈도간의 관계를 고려하여 검색 실험을 수행하였다. 엘리먼트 길이에 따라 단어빈도를 재산출하기 위해 다음과 같은 두 가지 가정 하에서 실험을 수행하였다.

$H1$: 단어의 분포는 엘리먼트 내에서 단일하다. 이 가정은 Robertson과 Walker(1994)의 원칙을 변형하여 적용하였다. 이를 위한 단어빈도 밀도함수 $\rho(l)$은 상수 ρ가 된다.

$$\rho(l) = c \cdot \frac{tf}{l} = \rho$$

c : 상수
l : 문서길이

엘리먼트 길이의 정규화를 위하여 tf 빈도를 정규화 단어빈도 tfn으로 대체한 공식은 다음과 같다.

$$tfn = \int_{l(E)}^{l(E)+avg\_l} \rho(l) El = \rho \cdot avg\_l = tf \cdot \frac{avg\_l}{l(E)}$$

$El, l(E)$: 엘리먼트 E의 길이
avg_l : 엘리먼트 길이의 평균

$H2$: 단어빈도 밀도함수 $\rho(l)$은 길의 l의 감소함수인 단어빈도함수는 길이와 역의 관계에 있다고 가정하였다. $H2$의 정규화 단어빈도는 다음과 같다.

$$tfn = \int_{l(E)}^{l(E)+avg\_l} \rho(l) El = c \cdot tf \cdot \int_{l(E)}^{l(E)+avg\_l} \frac{El}{l}$$

$$= tf \cdot \log_2(1 + \frac{avg\_l}{l(E)})$$

$$c \quad : 1/log_e 2 = log_2 e$$

$H2$의 정규화 단어빈도는 로그값으로 엘리먼트 길이의 영향력을 낮추는 효과가 있어 $H1$의 정규화 단어빈도보다 엘리먼트 길이에 따른 값의 편차가 작다. 예를 들어 단어빈도가 5이고 엘리먼트의 평균길이가 1,000단어이며 대상 엘리먼트의 길이를 각각 500과 1,500이라고 했을 때, $H1$의 정규화 단어빈도는 각각 10과 3.33이 되고 $H2$는 각각 7.92와 3.68이 된다.

3.3 보정 기법 실험

XML 엘리먼트는 상대적으로 작은 크기로 최우추정을 통해 언어모델을 추정하는데 있어 0 확률값이나 0에 가까운 확률값이 부여되는 문제가 발생한다.

이 실험에서는 이러한 문제를 해결하기 위해 Jelinek-Mercer(1980)의 보정 기법과 조정이 가능한 길이 사전확률을 적용하는 다항 언어모델(multinomial language model)에 기초하여 검색 실험을 수행하였다.

Jelinek-Mercer(1980)의 보정 기법은 다항언어모델로 먼저 컬렉션의 XML 엘리먼트 각각을 위한 언어모델을 추정한 후 엘리먼트 언어모델이 질의를 생성할 가능성, 즉 엘리먼트에 질의가 나타날 확률에 따라 엘리먼트의 순위를 매기는 방법이다. 다시 말해 다음과 같은 확률을 추정하는 것이다(Sigurbjörnsson, Kamps, and Rijke, 2003).

$$P(E,Q) = P(E) \cdot P(Q|E)$$

E : 엘리먼트
Q : 질의

여기에서 두 가지 확률을 추정해야 하는데 엘리먼트의 생성확률인 $P(E)$와 질의를 생성하는 확률인 $P(Q|E)$이다.

엘리먼트는 상대적으로 작은 양의 텍스트를 포함하므로 언어모델을 추정하기 위한 유일한 기본값이 되기에는 너무 작다. 이러한 데이터의 희소성을 해결하기 위해서 두 언어모델, 즉 엘리먼트 데이터를 기본으로 하는 언어모델과 컬렉션 데이터를 기본으로 하는 또 하나의 언어모델을 선형보간법에 의해 결합하여 엘리먼트 언어모델로 추정하게 된다. 여기에서 질의는 독립적이라고 가정하여 엘리먼트 언어모델이 주어지면 질의의 확률을 다음과 같은 공식을 이용하여 추정하였다.

$$P(Q|E) = \prod_{i=1}^{n} (\lambda \cdot P_{mle}(t_i|E) + (1-\lambda) \cdot P_{mle}(t_i|C))$$

Q : 단어 $t_1,...,t_k$ 로 구성되는 질의

E : 엘리먼트

C : 컬렉션

λ : 보간법 삽입인자 (interpolation factor)로 보정 매개변수.

$P_{mle}(\bullet|\bullet)$: 최우추정을 이용하여 추정되는 언어모델.

이 공식에서 $P_{mle}(t_i|E)$는 엘리먼트내 단어의 확률이고 $P_{mle}(t_i|C)$는 컬렉션내 단어의 확률로 모두 최우추정을 이용하여 추정된다.

보정 매개변수 λ는 컬렉션 모델을 이용하여 보정할 때 엘리먼트 언어모델에 주어지는 가중치가 된다. λ의 역할이 중요한데 Zhai와 Lafferty(2001)는 큰 문서보다 작은 문서에서 보정이 더 요구될 것이라고 하였다. Kamps 등(2003)은 XML 검색시 검색 단위에 대한 보정의 효과에 대해 보정 매개변수의 값과 검색된 엘리먼트 크기 사이에는 상관관계가 있다고 하였다. 보정을 조금 적용시켜 더 높은 보정 매개변수 λ을 사용하면 검색된 엘리먼트의 평균크기는 아주 커지게 된다. 다시 말해 언어모델에서 λ값을 증가시키면 용어가 급격하게 늘어나게 되고 그 결과 많은 용어를 가지는 엘리먼트가 적은 용어를 가지는 엘리먼트보다 검색기회가 높아진다. 색인 서브트리 방식에 의한 중복 엘리먼트 색인의 경우에 높은 λ값은 article과 같은 큰 단위의 엘리먼트를 검색하는 편향성을 보여주고, 낮은 λ 값은 절이나 문단과 같은 더 작은 단위의 엘리먼트를 검색하는 편향성을 보여주었다.

길이에 따른 검색 요소의 사전 확률(prior probability)의 유용성에 대한 연구가 있었는데(Hiemstra and Kraaij 1999; Singhul, Buckley, and Mitra 1996), XML 검색에서는 특히 길이와 관련된 사전 확률이 유용하여 길이에 비례하는 요소의 사전 확률을 가지는 것이 가장 공통적이다. 다음과 같은 공식을 이용하여 길이와 관련된 사전 확률을 계산할 수 있다.

$$lp(E) = \beta \cdot log(\sum_t tf(t, E))$$

β : 길이를 위한 매개변수

3.4 계층적 언어모델 검색 실험

이 실험에서는 XML 문서의 트리구조에서 색인 노드의 전후문맥 정보를 이용하여 언어모델을 생성하였다. 즉 엘리먼트 요소의 언어모델은 그 엘리먼트 요소와 자식 엘리먼트, 그리고 부모 엘리먼트의 언어모델을 보정 기법을 이용하여 추정하였다.

요소의 언어모델은 선형보간법을 이용하여 요소, 자식의 모델, 그리고 그 부모 모델로부터 직접적으로 추정하여

생성된다. 문서의 요소를 위한 생성 모델의 추정은 다음과 같이 세 단계로 나뉘어 수행된다.

첫 번째 단계에서 언어모델 θ_{v_i}는 그 해당 색인노드 v_i의 요소만을 관찰하여 추정한다.

이 모델은 문서 요소내의 관찰된 텍스트로부터의 분포를 추정하는 것으로 $\theta_{collection}$모델은 이 추정을 보정하는 컬렉션 수준의 배경 모델 혹은 보편적 모델이다. Ogilvie와 Callan(2004)은 컬렉션 언어모델 $\theta_{collection}$을 $\theta_{type(v_i)}$로 사용하는 데 이때 type함수는 이 실험에서는 고려하지 않았으나 문서 요소의 특정 모델을 명시하는데 이용될 수 있다. 예를 들어 제목 노드의 언어모델은 다른 노드의 텍스트와 다르게 설정하기 위해 type 함수를 이용하거나 말뭉치 모델을 추정하기 위해 대량의 텍스트를 제공하는 모든 요소를 위한 하나의 언어모델을 추정하는데 이용할 수 있다.

두 번째 단계는 트리의 말단에서 상위로 가는 중간 θ'_{v_i}모델을 추정하는 것이다.

$|v_i|$는 노드 v_i의 텍스트 토큰의 길이로 그 자식의 토큰은 포함하지 않으며 θ'_{v_i}는 컬렉션 모델에 삽입된 노드 v_i의 텍스트로부터 추정된 텍스트 모델과 동일하다. 예를 들어 제목 노드의 텍스트는 문서의 본문보다 질의로 표현될 확률이 더 높으므로 제목 모델에 더 높은 가중치를 주는 것이 검색 성능을 향상시킬 수 있다.

마지막 단계는 θ'_{v_i}를 추정한 후에 요소의 순위를 매기기 위해서 이용되는 θ''_{v_i} 모델을 트리의 루트에서부터 시작하여 추정하는 것이다.

순위를 매기는데 이용되는 θ''_{v_i}모델은 θ'_{v_i}모델과 $\theta'_{parent(v_i)}$모델을 보간법을 이용하여 추정한다. 예를 들어 θ''_{v_1}은 부모가 없기 때문에 간단하게 θ'_{v_1}로 추정되지만 부모를 가진 노드는 두 번째 단계에서 추정된 언어모델과 θ''_{v_i}모델을 결합하여 추정한다.

엘리먼트의 순서는 질의 언어모델의 확률에 기초하여 매겨지는데 다음과 같이 $P(Q|\theta_{v_i})$를 이용하여 간단하게 정렬을 할 수 있다(Ponte and Croft 1998; Hiemstra 2001; Zhai and Lafferty 2004).

색인 노드의 텍스트의 길이를 고려하는 사전확률은 다음과 같이 적용하였는데 이 사전확률은 질의가 주어졌을 때 문서 요소 노드의 확률에 의해 순위를 매기는 결과로 나타나게 된다. 길이에 관련된 사전확률은 다음과 같은 세 가지 선택조건을 가질 수 있다.

$$P(v_i) = P(v_i|\ lenght(v_i))$$

·선형 - $P(v_i|\ lenght(v_i)) \propto x$

스퀘어 - $P(v_i|\ lenght(v_i)) \propto x^2$

큐빅- $P(v_i|\ lenght(v_i)) \propto x^3$

사전확률을 결합한 문서의 순위화를 위한 공식은 다음과 같다(Ogilvie and Callan 2003).

$$s(Q,\ v_i) = \log(P(Q|\theta_{v_i})P(v_i)$$

이러한 방법으로 부모 언어모델을 결합하면 언어모델에 영향을 주는 문서 요소간의 계층정보를 이용할 수 있게 된다. 이것을 축소(shrinkage)라 부르기도 하는데, McCallum과 Nigam(1999)이 정보검색에 이 개념을 소개하였다. 클래스는 계층적으로 모델화되고 클래스 언어모델은 클래스에 할당된 문서의 텍스트로부터 그리고 모든 조상의 언어모델은 클래스 계층의 텍스트로부터 추정된다. 이 실험에서는 클래스를 위한 계층적 모델의 개념을 문서 수준으로 적용시켜 모델을 추정하였다.

4. 실험 결과 분석 및 평가

4.1 다이버전스 기법 실험 결과

Ponte와 Croft(1998)가 정보검색에 언어모델을 처음 적용한 이래로 많은 연구들이 있었다. 그 중 Amati와 Rijsbergen(2002)은 용어가중치 모델을 무작위로 얻어낸 실제 용어 분포의 다이버전스를 측정하여 유도하였다.

이 실험에서는 문서에 질의를 매칭하는 과정에서 용어 가중치를 조정하기 위해서 두 가지의 단어빈도 정규화를 이용하였다. 첫 번째 정규화는 문서가 동일한 길이를 가진다고 가정하고 적합 문서에 출현하는 용어의 확률(information gain)을 측정하는 것이다. 두 번째 정규화는 문서 길이의 통계값을 적용한 방법이다.

이 실험에서는 Amati와 Rijsbergen (2002)의 언어모델 기법에서 제안한 가중치 함수 서브셋인 Inf_1와 Inf_2에 대해 1, 2차 정규화를 적용하여 네 가지의 기본적인 가중치 공식을 적용하여 수행한 실험 결과를 비교, 평가하였다. Amati와 Rijsbergen의 언어모델은 문서를 검색단위로 하는 반면, 이 실험에서는 검색결과로 제시되는 의미 있는 단위가 XML 엘리먼트이므로 내용질의 검색은 색인 노드를 탐색하게 된다.

다이버전스 기법의 실험 결과를 나타내기 위해 세 자리 기호를 적용하였는데 첫 번째 자리는 적용한 기법을 나타내고 두 번째 자리는 1차 정규화를 그리고 마지막 숫자는 2차 정규화를 나타냈다. 예를 들어 BB1은 binomial 기법, 1차 정규화 B(Bernoulli 프로세스), 그리고 2차 정규화 H1이며, EL2는 Bose-Einstein 기법, 1차 정규화

L(Laplace의 계승규칙), 그리고 2차 정규화 H2를 나타낸다.

1차 정규화, 즉 단어빈도만을 정규화한 검색 실험 결과는 모두 문서검색보다 좋지 않은 성능을 보여주었다. 그 이유는 중복색인 방식에 따른 엘리먼트 길이의 편차에 있는 것으로 분석되었다. 예를 들어 섹션이 여러 개의 서브섹션으로 구성되어 있다고 하면 상위 엘리먼트 섹션과 하위 엘리먼트 서브섹션의 길이는 상당히 차이가 나게 된다. 이러한 1차 정규화 실험의 분석결과에 따라 엘리먼트의 길이를 고려하여 단어빈도를 재계산하는 실험을 2차 정규화 실험에서 수행하였다.

표 2. 다이버전스 기법의 성능 평가

	평균정확률	향상률(%)	p(10)	p(20)
문서검색	0.4072	-	0.4710	0.3805
BB1	0.4029	-1.05	0.5400	0.4667
BL1	0.4055	-0.41	0.5800	0.4873
BB2	0.4755	16.77	0.6660	0.5273
BL2	**0.5001**	22.81	**0.6780**	**0.5573**
EB1	0.3971	-2.49	0.6580	0.5200
EL1	0.4464	9.63	0.6680	0.5393
EB2	0.4375	7.44	0.6520	0.5260
EL2	0.4890	20.08	**0.6780**	0.5447

엘리먼트 길이와 단어빈도간의 관계를 고려하는 2차 정규화까지 적용한 실험 결과는 표 2에 나타난 것처럼 이항 모델에 1차 정규화 L(Laplace의 계승규칙)을 적용한 후, 단어빈도 밀도함수가 길이와 역의 관계있다고 가정하는 2차 정규화 H2를 적용했을 때 0.5001로 가장 좋은 성능을 보여 문서검색 대비 22.81%의 성능 향상률을 나타냈다.

2차 정규화 실험 결과는 H1보다 로그값을 적용하여 엘리먼트 길이에 따른 영향력을 더 감소시킨 H2를 적용한 실험 결과가 전반적으로 더 좋은 것으로 나타났다. 이 결과로 보아 엘리먼트 길이의 편차가 클 경우 그 길이에 따른 단어빈도의 영향력을 최소화하는 것이 검색성능을 향상시키는 것으로 분석된다. 이 실험 결과는 Amati와 Rijsbergen(2002)의 실험 결과와 Abolhassani 등(2004)의 실험 결과와도 일치한다.

4.2 보정 기법 실험 결과

정보검색에서 보정 기법은 매우 작은 양의 텍스트에서 언어모델을 구축할 수 있게 한다는 장점이 있으므로 문서보다 작은 엘리먼트를 검색단위로 하는 XML 문서검색에 적합하다고 할 수 있다.

보정 기법은 먼저 컬렉션의 XML 엘리먼트 각각을 위한 언어모델을 추정한 후 엘리먼트 언어모델이 질의를 생성할 가능성인 엘리먼트에 질의가 나타날 확

률에 따라 엘리먼트의 순위를 매긴다. 보정 기법 실험에 앞서 두 가지의 사전 실험을 수행하였는데 먼저 보정의 영향력을 알아보기 위해 보정 매개변수 λ를 [0.1, 0.9] 범위의 값으로 변화시켜 가며 적절한 크기의 λ을 찾는 실험을 수행하였다. 그 다음으로 길이 사전확률의 중요성을 결정하기 위해 사전확률 매개변수 β를 0.0에서 5.0 사이의 값을 가지고 검색 성능의 변화를 관찰하였다. 보정 기법 검색 실험에서는 사전 실험을 통해 알아낸 보정 매개변수와 길이 사전확률 매개변수를 교차시켜 λ와 β에 의한 이차원 탐색공간에서 실험을 수행하였다.

사전실험을 통해 길이 사전확률 매개변수 1일 때 가장 좋은 결과를 보인 보정 매개변수 0.2를 가지고 길이 사전확률 매개변수 β의 값을 조정하여 검색 실험을 수행하였다. 실험 결과는 표 3에서 보이듯이 β = 3.0 일 때 0.5420으로 가장 좋은 검색 결과를 나타내어 문서검색 대비 33.10%의 검색 성능 향상률을 보였다. 여기에서 β = 1.0은 정규 길이 사전확률인 선형 사전확률을 의미하며 β = 2.0은 길이의 제곱인 스퀘어를 적용한 사전확률을 의미하고 β = 3.0은 길이의 세제곱인 큐빅을 적용한 사전확률을 의미한다. β = 3.0 일 때 평균정확률이 0.5420으로 가장 좋은 검색 성능을 보였으며, 검색순위 상위에서도 0.7002(p(10)), 0.6246(p(20))으로 가장 높은 정확률을 나타냈다.

표 3. 사전확률변화에 따른 검색 성능 평가

보정 매개변수 (λ)	길이 매개변수 (β)	평균정확률	향상률 (%)	p(10)	p(20)
문서검색		0.4072	-	0.4710	0.3805
0.2	1.0	0.5210	27.95	0.6485	0.5673
0.2	2.0	0.5280	29.67	0.6817	0.6177
0.2	3.0	**0.5420**	33.10	**0.7002**	**0.6246**

그러나 이때 검색 결과로 제시되는 엘리먼트는 대체로 큰 단위의 상위 엘리먼트가 되어 색인 노드간의 중복도는 작아지지만 XML 검색에서 기대하는 작은 단위의 엘리먼트가 아니므로 검색효율성 측면에서는 이 점을 고려해야 한다. 실험 결과에서 보면 선형길이 사전확률을 적용하였을 때 0.5210이고 스퀘어길이 사전확률을 적용하였을 때 0.5280으로, 큐빅길이 사전확률을 적용한 실험 결과인 0.5420과 검색성능에서 크게 차이가 없으므로 망라성과 특정성을 고려하여 선형길이나 스퀘어길이 사전확률을 채택하는 것이 바람직하다.

4.3 계층적 언어모델 검색 실험 결과

계층적 언어모델 검색 실험에서는

XML 구조의 트리기반 계층정보를 이용하여 엘리먼트 언어모델을 재추정하여 검색을 수행하였다. 이 모델은 엘리먼트의 언어모델을 추정함에 있어 자식과 부모의 언어모델의 확률을 보정 기법을 이용하여 선형보간법으로 결합하였다.

부모 언어모델을 결합을 위한 보정 매개변수인 축소(shrinkage) 매개변수의 변화에 따른 검색결과도 비교 평가하였다. 사전실험을 통해 너무 큰 축소 매개변수는 재현율이 낮은 구간에서 정확률이 좋지 않은 결과를 보여 계층적 언어모델 실험에서 작은 축소 매개변수 값을 적용하였으나 결과에 큰 차이가 없었다. 이러한 이유로 축소 매개변수를 적용하지 않은 경우와 0.1을 적용한 경우에 대해서만 검색결과를 비교하였다. 표 10에서 알 수 있듯이 축소 매개변수 λ^p = 0.1을 적용한 경우가 적용하지 않은 경우보다 전반적으로 약간 좋은 결과를 보였다.

사전확률 측면에서 보면 사전확률을 적용한 결과가 적용하지 않은 경우보다 좋게 나타나며, 엘리먼트 크기의 제곱인 스퀘어 길이를 적용한 경우가 세 가지 사전확률 중 가장 좋은 성능을 보였다. 축소 매개변수 0.1과 스퀘어 길이 사전확률을 적용한 결과가 0.5311로 문서검색 대비 30.43%의 성능 향상률을 보여 가장 좋은 결과를 나타냈다. 그 다음으로 축소 매개변수 0.0과 스퀘어 길이 사전확률을 적용한 결과가 0.5254로 좋은 성능을 보였다.

표 4. 계층적 언어모델 검색 결과

λ^p	길이 사전확률	평균정확률	향상률 (%)	p(10)	p(20)
문서검색		0.4072	-	0.4710	0.3805
0.0	-	0.4154	2.01	0.6401	0.5761
0.0	선형	0.4851	19.13	0.6873	0.6186
0.0	스퀘어	0.5254	29.03	0.7012	0.6311
0.0	큐빅	0.5066	24.41	0.6911	0.6220
1.0	-	0.4652	14.24	0.6458	0.5812
1.0	선형	0.4975	22.18	0.6891	0.6202
1.0	스퀘어	**0.5311**	30.43	**0.7133**	**0.6420**
1.0	큐빅	0.5163	26.79	0.6971	0.6274

앞서의 언어모델 검색 결과와 성능을 비교해 보면 표 4에서 알 수 있듯이 계층적 언어모델의 축소 매개변수 1.0과 스퀘어 길이 사전확률을 적용한 결과가 0.5311로 문서검색 대비 30.43%의 성능 향상률을 보여 보정 기법의 사전확률 큐빅을 적용한 결과 0.5420에 이어 두 번째로 좋은 검색결과를 보였다.

4.4 XML 문서검색 실험 결과의 비교 분석

XML 검색을 위해 각 단계별로 수행된 실험의 결과는 다음과 같다.

문서를 검색단위로 하는 문서검색 실험 결과, 로그단어빈도*역문헌빈도 가중치를 적용한 검색결과가 가장 높은 검색성능을 보였다. 이 실험 결과는 엘리먼트 기반 XML 문서검색 실험의 검색성능 비교, 분석을 위한 기초 실험 결과로 이용되었다.

다이버전스 기법 실험에서는 문서에 질의를 매칭하는 과정에서 용어 가중치를 조정하기 위해서 두 가지의 단어빈도 정규화를 이용하였다. 첫 번째 정규화는 문서가 동일한 길이를 가진다고 가정하고 적합문서에 출현하는 용어의 확률(information gain)을 측정하는 것이다. 두 번째 정규화는 문서 길이의 통계값을 적용한 방법이다. 적용한 기법으로는 이항모델(binomial)과 Bose-Einstein 기법이며, 1차 정규화를 위해서는 B(Bernoulli 프로세스)와 L(Laplace의 계승규칙)을 적용하였다. 2차 정규화는 엘리먼트의 길이를 고려하기 위한 방법으로 단어분도에 관한 H1(단어의 분포는 문서내에서 단일)과 H2(단어빈도 밀도는 길이와 역의 관계) 가정 하에 검색 실험을 수행하였다. 가중치 함수 서브셋 Inf_1와 Inf_2에 대해 각각 두 가지를 계산하여 네 가지의 기본적인 가중치 공식을 적용하였다. 1차 정규화 결과는 이항모델 기법에 1차 정규화 L(Laplace의 계승규칙)을 적용한 경우가 가장 좋은 성능을 보였으며, 검색순위 상위에서는 B(Bernoulli 프로세스)를 적용한 결과가 가장 좋았다. 2차 정규화까지 적용한 결과는 이항모델 기법에 1차 정규화 L(Laplace의 계승규칙)과 단어빈도 밀도함수가 길이와 역의 관계있다고 가정하는 2차 정규화 H2를 적용했을 때 가장 좋은 결과를 보였다.

보정 기법 실험에서는 XML 엘리먼트는 상대적으로 작은 크기로 최우추정을 통해 언어모델을 추정하는데 있어 0 확률값 이나 0에 가까운 확률값이 부여되는 문제를 해결하기 위해 Jelinek-Mercer(1980)의 보정 기법과 길이 사전확률을 적용하는 다항 언어모델(multinomial language model)에 기초하여 검색 실험을 수행하였다. 보정 매개변수 λ값이 작으면 검색되는 엘리먼트의 크기가 상대적으로 작으며 반대로 큰 값의 λ을 적용하면 엘리먼트의 크기는 커지게 된다. 그러나 실험 결과 λ값을 조정하여 엘리먼트 크기를 크게 하는 것이 더 나은 검색결과를 보이지는 않는 것으로 나타났다. λ = 0.2 일 때 길이 사전확률 매개변수 β의 값을 조정하여 실험한 결과는 β = 3.0(큐빅 길이) 일 때 0.5420으로 가장 좋은 성능을 나타냈다.

계층적 언어모델 기법 실험에서는 XML 구조의 계층정보를 이용하여 언어모델을 재추정하는 기법을 적용하였다. 이 기법은 엘리먼트의 언어모델을 추정함에 있어 자식과 부모의 언어모델을

보정 기법을 이용하여 선형보간법으로 결합하였다. 자식 엘리먼트를 결합하기 위해서는 엘리먼트 길이로 계산하는 보정 매개변수를 적용하고 부모 엘리먼트를 결합하기 위해서는 축소 매개변수를 적용하였다. 사전실험을 통해 너무 큰 축소 매개변수는 재현율이 낮은 구간에서 정확률이 좋지 않은 결과를 보여 계층적 언어모델 실험에서 작은 축소 매개변수값을 적용하였다. 그러나 축소 매개변수를 적용한 경우가 적용하지 않은 경우보다는 전반적으로 약간 좋은 결과를 보였다. 사전확률 측면에서 보면 사전확률을 적용한 결과가 적용하지 않은 경우보다 좋게 나타나며 엘리먼트 크기의 제곱인 스퀘어를 적용한 경우가 세 가지 사전확률 중 가장 좋은 성능을 보였다. 축소 매개변수 0.1과 스퀘어 사전확률을 적용한 결과가 0.5311로 가장 좋은 결과를 보였다.

실제 정보검색에서 유용성을 가지는 검색순위 상위에서의 검색 성능은 표 5에서 알 수 있듯이 검색결과 순위 10위 내에서는 계층적 언어모델의 검색 결과가 0.7133으로 문서검색 결과 0.4710보다 51.44%의 성능 향상률을 나타냈으며, 그 다음으로 보정 기법을 적용한 실험 결과가 0.7002로 48.67%의 성능 향상률을 보여주었다. 검색순위 20위내에서도 계층적 언어모델이 가장 좋은 결과를 나타내어 0.6420으로 문서검색 결과 0.3805 보다 무려 68.72%의 성능 향상률을 나타냈다.

표 5. 언어모델 검색 실험 성능 비교

	평균정확률	향상률(%)	p(10)	p(20)
문서검색	0.4072	-	0.4710	0.3805
다이버전스 기법(BL2)	0.5001	22.81	0.6780	0.5573
보정 기법 (λ=0.2, β=3.0)	0.5420	33.10	0.7002	0.6246
계층적 언어모델 (λ^p=1.0, 스퀘어)	0.5311	30.43	**0.7133**	**0.6420**

5. 결 론

이 연구에서는 정보검색모델을 적용하여 정보검색에서 문서단위가 아닌 더 세분화된 단위를 검색하기 위한 엘리먼트 기반 XML 문서검색의 성능에 대해 연구하였다.

단계별로 수행된 실험과 평가의 결과는 다음과 같다.

첫째, 다이버전스 기법을 적용하였을 때 보다 보정 기법을 적용하였을 때 전반적으로 검색성능이 좋은 것으로 나타났다. 다이버전스 기법은 이질적인 확률분포를 조합하여 비-매개변수 모델을 유도하여 수행되므로 학습 집단으로부

터 매개변수를 학습해야하는 데이터 유도 방식이 필요 없다는 장점을 가지고 있다.

둘째, 보정 기법은 컬렉션 언어모델 등과 같은 다른 언어모델을 결합하여 엘리먼트 언어모델을 추정하도록 하여 검색을 수행한다. 보정 매개변수와 엘리먼트의 길이를 위한 사전확률이 중요한 변수로 작용한다. 큰 보정 매개변수를 이용하면 검색결과로 제시되는 엘리먼트의 크기가 커지게 되는 결과를 보이며 사전확률의 경우에도 적용되는 사전확률값이 클수록 엘리먼트의 크기가 커지는 결과를 보였다. 실험 결과 대체로 큐빅 길이나 스퀘어 길이 사전확률이 약간 좋은 결과를 보여 망라성은 좋아졌으나 특정성이 떨어지는 결과를 보였다.

셋째, 계층적 언어모델 검색 실험에서는 XML 구조의 계층정보를 이용하여 언어모델을 재추정하는 기법을 적용하였다. 엘리먼트의 언어모델을 추정함에 있어 자식과 부모의 언어모델을 보정 기법을 이용하여 선형보간법으로 결합하였다. 문서 트리의 언어모델을 결합하기 위해 매개변수를 이용하였는데 자식 엘리먼트를 결합하기 위해서는 자식 텍스트의 축적된 길이를 노드의 축적된 길이로 나눠서 자식노드에 가중치를 주는 보정 매개변수를 적용하였다. 부모 엘리먼트를 결합하기 위해서는 축소 매개변수를 적용하였다. 길이에 따른 사전확률 실험 결과, 사전확률을 적용한 실험 결과가 적용하지 않은 경우보다 좋게 나타났으며 엘리먼트 크기의 제곱인 스퀘어 길이를 적용한 경우가 세 가지 사전확률 중 가장 좋은 성능을 보였다. 축소 매개변수 0.1과 스퀘어 길이 사전확률을 적용한 결과가 0.5311로 가장 좋은 결과를 보였는데 축소와 사전확률을 모두 적용하지 않은 결과 0.4154보다는 27.85%, 문서검색 결과 0.4072 보다는 30.43%의 성능 향상률을 나타냈다.

이상의 실험 결과를 통해서 이 연구에서는 엘리먼트 기반 XML 문서검색을 위한 가장 효율적인 검색 접근법으로 문서의 구조정보를 적용한 계층적 언어모델 검색을 제안한다. 계층적 언어모델은 문서의 트리구조에 기반하여 XML 문서의 계층구조를 효과적으로 적용하여 엘리먼트 기반 XML 문서검색의 성능이 가장 뛰어난 것으로 나타났다. 특히 검색순위 상위에서 뛰어난 성능을 보였으며, 언어모델의 길이 사전확률 적용에 있어서도 엘리먼트의 망라성과 특정성의 수준을 적절히 유지하는 것으로 나타났다.

이 연구의 제한점 및 향후 과제는 다음과 같다.

첫째, 이 연구에서 제안한 검색 실험 기법들은 문서의 논리적 구조가 비교적 분명한 논문기사를 대상으로 하였으므로 대규모의 일반적인 실험집단에 적용하여 검증할 필요가 있다. 실제 XML 문서가 사용되고 있는 웹 환경에서 이 실

험에서 제안한 기법들이 실용성이 있는지 평가하는 연구도 필요하다.

둘째, XML 색인 방식의 문제로, 이 연구에서는 색인 서브트리 방식과 독립색인 방식을 기법마다 각각 적용하여 검색 실험을 수행하였다. 독립색인 방식으로 색인을 작성하여 검색을 수행한 계층적 검색 실험의 성능이 상대적으로 좋은 결과를 보여주었다. 이 결과를 바탕으로 하여 색인 서브트리 방식의 엘리먼트 중복색인으로 실험을 진행한 언어모델에 대해 독립색인 기반으로 검색 실험을 수행하여 색인 방식간의 검색 성능의 변화를 관찰할 필요가 있다.

마지막으로, 실질적으로 XML 문서검색 실험에서 검색결과에 대한 적합성 판정은 문서 수준이 아닌 엘리먼트 수준에서 이루어져야 하므로 XML 엘리먼트에 대한 적합성 판정을 바탕으로 검색성능을 평가할 필요가 있다.

참고문헌

Abolhassani, M, and N, Fuhr. 2004. “Applying the Divergence From Randomness Approach for Content-Only Search in XML Documents". *26th European Conference on Information Retrieval Research (ECIR 2004)*. Springer.

Amati, G, and C. J. Rijsbergen. 2002. "Probabilistic models of information retrieval based on measuring the divergence from randomness". *ACM Transactions on Information Systems(TOIS)*, 20(4).

Chiaramella, Y., P. Mulhem, and F.Fourel. 1996. "A Model for multimedia information retrieval". *Technical report, FERMI SPRIT BRA E8314*, University of Glasgow.

Hiemstra, D. and Q. Kraaij. 1999. Twenty-One at TREC-7: Ad-hoc and cross-language track. In *The Seventh Text REtrieval Conference (TREC-7)*, 227-238.

Jelinek, F. and R. Mercer. 1980. "Interpolated estimation of Markov source parameters from sparse data". In *Proceedings of the Workshop on Pattern Recognition in Practice*.

McCallum, A. and K. Nigam. 1999. Text classification by bootstrapping with keywords, em and shrinkage. In *Proceedings of the ACL 99 Workshop for Unsupervised Learning in Natural Language Processing*, 52-58.

Miller, D. R. H., T. Leek, and R. M.

Schwartz. 1999. "A hidden Markov model information retrieval system". In *Proceedings of the 22nd ACM SIGIR Conference*, 214-221.

Moffat, A., R. Sacks-Davis, R. Wilkinson, and J. Zobel. 1994. "Retrieval of partial documents". In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*.

Ogilvie, P. and J. Callan. 2003. Using Language Models for Flat Text Queries in XML Retrieval. In *Proceedings of the Second Annual Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*.

Ogilvie, P. and J. Callan. 2004. Hierarchical Language Models for XML Component Retrieval. In *Proceedings of the Third Annual Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*.

Ponte, J. M. and W. B. Croft. 1998. "A language modeling approach to information retrieval", In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR '98)*.

Salton, G, J Allan, and C. Buckley. 1993. "Approach to passage retrieval in full text information systems". *Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval*.

Sigurbjörnsson, B., J. Kamps and R. Maarten. 2003. "An Element-based Approach to XML Retrieval". *Proceedings of the third Workshop of the INitiative for the Evaluation of XML Retrieval*.

Singhul, A., C. Buckley, and M. Mitra. 1996. "Pivoted document length normalization". In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development Information Retrieval*, 21-29.

Wilkinson, R. 1994. “Effective retrieval of structured documents”. *Proceedings of SIGIR Conference*, 311-317.

Zhai, C. and J. Lafferty. 2001. "A study of smoothing methods for language models applied to ad hoc information retrieval". In *Proceedings of the 24th ACM SIGIR Conference*, 334-342.