퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 6, pp. 745-751

지역적 문맥 분석 피드백을 이용한 웹 정보검색에 관한 연구

A Study on Information Retrieval of Web Using Local Context Analysis Feedback

김영천*, 이성주**
Young-cheon kim, Sung-joo Lee

*서정대학 정보통신과
**조선대학교 전자계산학과

요 약

순수한 부울 검색 시스템은 문서와 질의 사이의 유사 도를 나타내는 문서 값을 계산할 수 없기 때문에 검색된 문서들을 질의를 만족하는 정보에 따라 정렬할 수 없다. 부울 검색 시스템의 이러한 단점을 보완하는 방법으로 MMM 모델, Paice 모델, P-norm 모델이 개발되었다. 이러한 방법들은 부울 연산자를 유연하게 연산하는 공통된 특성을 지니고 있다. 본 논문에서는 높은 검색 효과를 제공하는 지역적 문맥 분석 피드백(Local Context Analysis Feedback)을 이용한 웹 정보 검색 모델을 이용한다. 지역적 문맥 분석 피드백 모델의 연산 특성이 MMM(Max and Min Model), Paice, P-norm 모델보다 우수함을 설명하고, 또한 성능 비교를 통하여 이를 입증한다.

Abstract

In conventional boolean retrieval systems, document ranking is not supported and similarity coefficients cannot be computed between queries and documents. The MMM(Max and Min Model), Paice and P-norm models have been proposed in the past to support the ranking facility for boolean retrieval systems. They have common properties of interpreting boolean operators softly. In this paper we propose a new soft evaluation method for web Information retrieval using local context analysis feedback model. We also show through performance comparison that local context analysis feedback is more efficient and effective than MMM, Paice and P-norm.

Key word : 웹 정보검색, 지역적 문맥 분석 피드백, 질의 분해, 부울 검색, 유사도

Ⅰ. 서 론

순수한 부울 검색 모델은 문서와 질의 사이의 유사 도를 나타내는 문서 값을 계산할 수 없기 때문에 검색된 문서들을 질의를 만족하는 정도에 따라 정렬할 수 없는 단점을 지니고 있다[1].

순수한 부울 검색 시스템의 단점을 보완하기 위하여 퍼지 집합 모델(Fuzzy Set Model)이 개발되었다[2]. 퍼지 집합 모델은 색인어가 문서 내에서 갖는 중요성을 반영하는 색인어 가중치를 이용하여 문서 값을 계산함으로써 부울 검색 시스템의 문제점을 극복하였다. 그러나 퍼지 집합 모델은 많은 경우에 부정확한 문서 값을 생성하기 때문에 정보 검색 모델로서 부적합하다고 비판되어 왔다. 이것은 AND와 OR 연산을 위하여 사용하는 MIN과 MAX 연산자가 단일 피연산자 의존 문제(Single Operand Dependency Problem)를 발생시키기 때문이다.

퍼지 집합 모델의 단일 피연산자 의존 문제를 극복하기 위하여 MMM 모델, Paice 모델, P-norm 모델이 개발되었다. 이들 모델들은 AND와 OR 연산을 위하여 MIN과 MAX 대신에 부울 연산자를 유연하게 연산하는 새로운 연산자를 사용함으로써 단일 피연산자 의존 문제를 극복하였을 지라도 다음과 같은 단점을 지니고 있다.

첫째, MMM 모델은 빠른 검색 시간을 제공하더라도, 부정확한 문서 값을 생성할 수 있는 요인을 지니고 있다.

둘째, MMM 모델의 문제점이 Paice, P-norm 모델에서는 발생하지 않을 지라도, 이들 모델은 검색 시간이 느리다는 단점을 가지고 있다[2].

정보검색에서 가장 중요하면서도 어려운 문제 중의 하나는 사용자가 원하는 정보를 찾기 위한 효율적인 질의를 작성하는 일이다. 하지만 전체 문서집합의 구성에 대해 미리 알고 있지 않는 한 이상적인 최적의 질의를 작성할 수 없다. 대신 최초에는 시험적 질의(tentative query)로 검색을 수행한 후, 이전의 검색 결과에 대한 평가를 기반하여 다음 번 검색의 질의를 개선시키는 방법이 적합성 피드백(relevance feedback)이다[3,4].

적합성 피드백은 특정 질의와 적합한 문서들은 유사한 벡터로 표현된다고 가정한다. 따라서 어떤 문서가 주어진 질의에 적합하다고 판단되면 질의를 적합한 문서와의 유사도가 증가하도록 변환하여 질의를 개선시킨다. 이렇게 개선된 질의는 최초 적합하다고 판단된 문서와 유사한 문서들을 추가적으로 검색하여 더 많은 양의 적합 문서를 검색해 낼 수 있다[5].

실제 이 적합성 피드백을 이용할 때에는 전체 문서집합에 대해 적합문서와 부적합문서를 미리 알 수 없으므로, 이미 적합성이 알려져 있는 문서들의 정보에 기반하여 질의확장을

접수일자 : 2004년 4월 7일
완료일자 : 2004년 9월 17일

수행한다. 이때의 적합성을 사용자가 알려주는 방법인 사용자 적합성 피드백(user relevance feedback)과 사용자의 개입 없이 초기질의로 검색된 결과 문서 중 상위 문서를 적합한 문서로 간주하여 적합성 피드백을 적용하는 방법인 의사 적합성 피드백(pseudo relevance feedback)이 있다[2,7].

지역적 방법에는 질의어에 의해 검색된 문헌들을 이용하여 질의 시간에 질의 확장을 위한 용어들을 선택한다. 이 방법은 적합성 피드백 과정과 유사하지만 사용자의 도움이 필요 없다는 점에서 다르다.

지역적 문맥 분석 피드백 모델을 이용하는 본 논문의 구성은 다음과 같다. 2장에서는 부울 연산자를 유연하게 연산하는 기존의 방법들인 MMM, Paice, P-norm 모델에 대하여 기술한다. 3장에서는 높은 검색 효과를 제공하는 지역적 문맥 분석 모델을 이용한다. 4장에서는 지역적 문맥 분석 피드백 모델과 MMM, Paice, P-norm 모델의 성능을 비교한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

II. 관련연구

퍼지 집합 모델의 단일 피연산자 의존 문제를 극복하기 위해 MMM 모델, Paice 모델, P-norm 모델이 개발되었다. 이들 모델들은 AND와 OR 연산을 위하여 MIN과 MAX 대신에 부울 연산자를 유연하게 연산하는 새로운 연산자를 사용함으로써 퍼지 집합 모델, MMM 모델, Paice 모델, P-norm 모델을 기반으로 하는 정보 검색 시스템은 <T, Q, D, F>로 정의되는 확장된 부울 검색 체계(Extended Boolean Retrieval Framework) 내에서 설명될 수 있다[9,10].

① T는 질의와 문서를 표현하기 위해 사용되는 색인어들의 집합이다.
② Q는 시스템이 인식할 수 있는 질의들의 집합이다. Q에 속하는 각각의 질의 q는 색인어들과 부울 연산 자 AND, OR, NOT으로 구성된 부울 수식이다.
③ D는 문서들의 집합이다. D에 속하는 각각의 문서 d 는 w_i가 색인어 t_i의 가중치일 때, $\{(t_1, w_1), ..., (t_n, w_n)\}$와 같이 표현되며. 색인어 가중치 w1는 0부터 1 사이의 값을 갖는다.
④ F는 문서 값을 계산하는 순위 결정 함수(Ranking Function)로서 다음과 같이 정의된다.

$$F: D \times Q \rightarrow [0, 1]$$

검색함수 F는 각 쌍의 (d, q)에 0부터 1사이의 값을 지정한다. 이 값은 문서 d와 질의 q사이의 유사 도를 의미하며, 질의 q에 대한 문서 d의 문서 값이다.

검색 함수 F(d, q)는 다음과 같은 2단계 과정을 거쳐서 계산된다.

(i) 질의에 나타난 각각의 색인어 t_i에 대하여, $F(d, t_i)$는 문서 d에서 색인어 t_i의 가중치 wi로 정의 된다.
(ii) 부울 연산자 AND와 OR는 주어진 식들을 이용하여 계산되고, NOT은 $F(d,NOT\ t_i)=1-w_i$로 계산된다.

두 개 이상의 부울 연산자를 포함하는 부울 질의는 가장 안쪽에 위치하는 절부터 순환적으로 계산된다.

퍼지 집합 모델의 부울 연산자 계산식 (a)는 두 개의 피연산자를 갖는 이항연산이고, MMM, Paice, P-norm 모델의 연산자 계산식은 2개 이상의 피연산자를 갖는 다항연산이다. 이것은 퍼지 집합 모델의 MIN과 MAX 연산자가 결합법칙을 만족하는데 비하여 MMM, Paice, P-norm 모델의 연산자는 결합법칙을 만족하지 못하기 때문이다. 결합법칙을 만족하지 못할 경우, 임의의 문서에 대하여 두 개의 동일한 질의($(t_1$ AND $t_2)$AND t_3와 t_1 AND (t_2 AND t_3))의 문서 값이 서로 다르다. MMM, Paice, P-norm 모델은 이러한 문제점을 다항연산을 가능하게 함으로써 극복하였다.

(a) 퍼지 집합 모델

$$F(d,\ t_1\ AND\ t_2) = MIN(w_1, w_2) \tag{1}$$

$$F(d,\ t_1\ OR\ t_2) = MAX(w_1, w_2) \tag{2}$$

(b) MMM 모델

$$F(d,\ t_1\ AND\ \ldots\ AND\ t_n) = r.MAX(w_1 \ldots w_n) + (1-r).MIN(w_1 \ldots w_n) \quad (0 \le r \le 0.5) \tag{3}$$

$$F(d,\ t_1\ OR\ \ldots\ OR\ t_n) = r.MIN(w_1 \ldots w_n) + (1-r).MAX(w_1 \ldots w_n) \quad (0.5 \le r \le 1) \tag{4}$$

(c) Paice 모델

$$F(d, t_1\ AND \ldots AND\ t_n) = \frac{\sum_{i=1}^{n}(r^{i-1}.w_i)}{\sum_{i=1}^{n}(r^{i-1})} \tag{5}$$

($0 \le r \le 1$, w_i'는 오름차순정렬)

$$F(d, t_1\ OR \ldots OR\ t_n) = \frac{\sum_{i=1}^{n}(r^{i-1}.w_i)}{\sum_{i=1}^{n}(r^{i-1})} \tag{6}$$

($0 \le r \le 1$, w_i'는 내림차순정렬)

(d) P-norm 모델

$$F(d,\ t_1\ AND \ldots AND\ t_n) = 1 - \left[\frac{(1-w_i)^p + \ldots + (1-w_n)^p}{n}\right]^{\frac{1}{p}} \quad (1 \le p \le \infty) \tag{7}$$

$$F(d,\ t_1\ OR \ldots OR\ t_n) = \left[\frac{w_i^p + \ldots + w_n^p}{n}\right]^{\frac{1}{p}} \quad (1 \le p \le \infty) \tag{8}$$

III. 지역적 문맥 분석 피드백 모델

확장된 부울 검색 체계를 기반으로 하는 검색 모델은 문서 값을 계산하기 위하여 색인어 가중치를 사용한다. 색인어 가중치는 역 문헌빈도(Inverse Document Frequency)와 색인어 출현빈도(Term Frequency)로부터 유도될 수 있다. 확장된 부울 검색 체계에서 색인어 가중치는 0부터 1 사이의 값이어야 하기 때문에 Wij는 식(9)와 같이 정규화 된다.

문서의 제목과 내용 문장만을 시스템의 입력으로 한다. 입력된 문서에 대해의 한국어 품사 태거를 이용하여 명사만 추출한 후 벡터를 구성한다. 이 때 벡터의 가중치 부여 방법을 위한 가중치 사이의 성능 실험은 다음과 같다.

실험 대상이 되는 가중치 부여 방법은 이진벡터(TF_{bin}), 문자 내에서의 단어의 빈도로 벡터를 구성하는 방법(TF), 그리고 정규화(normalization) 시켜서 벡터를 구성하는 방법(TF_{norm}) 등이 있다.

문장벡터가 다음과 같이 구성된다고 가정하면,

$$S_j = (w_{1j}, w_{2j}, ..., w_{nj})$$

각 각의 가중치 부여 방법들은 다음과 같다.

(1) 이진벡터(TFbin)

$$w_{ij} = \begin{cases} 1 & (\text{단어 } i\text{가 문장 } S_j\text{에 나타나면}) \\ 0 & (otherwise) \end{cases} \quad (10)$$

(2) 단어 빈도로 벡터를 구성하는 방법(TF)

$$w_{ij} = freq_{ij} \quad (11)$$

(3) 정규화시 켜서 벡터를 구성하는 방법(TFnorm)

$$f_{ij} = \frac{freq_{ij}}{\max freq_j} \quad (12)$$

$freq_{ij}$는 단어 i의 문장 S_j에서의 출현 빈도를, $\max freq_j$는 문서 S_j에서 출현한 단어 중 최대 빈도를 의미한다.

(4) 역 문헌 빈도수의 수식은 식(13)과 같다.

$$idf_i = \log \frac{N}{n_i} \quad (13)$$

N : 시스템 내의 총 문헌 수
ni : 색인어 k가 출현한 문헌 수

(5) 용어 가중치 할당 수식은 식(13)과 같다.

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \quad (13)$$

용어 가중치 할당은 용어 빈도수와 역문헌 빈도수를 곱하여 계산한다.

(6) 질의 용어 가중치 할당 수식은 식(14)와 같다.

$$w_{i,q} = \left(0.5 + \frac{0.5\ freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i} \quad (14)$$

역 문헌 빈도수, 용어 빈도수, 용어 가중치 할당을 구하는 간단한 예는 다음과 같다.

Q : "gold silver truck"
D1 : Shipment of gold damaged in a fire
D2 : Delivery of silver arrived in a silver truck
D3 : Shipment of gold arrived in a truck

* 역 문헌 빈도수 계산

용어	a	arrived	damaged	delivery	fire	gold
idf	0	.176	.477	.477	.477	.176
용어	in	of	silver	shipment	truck	
idf	0	0	.477	.176	.176	

* 용어 출현 빈도수 계산

용어	arrived	damaged	delivery	fire
D1	0	1/1	0	1/1
D2	1/2	0	1/2	0
D3	1/1	0	0	0
용어	gold	silver	shipment	truck
D1	1/1	0	1/1	0
D2	0	2/2	0	1/2
D3	1/1	0	1/1	1/1

* 질의 용어 가중치 계산

용어	gold	silver	truck
Wiq	0.176	0.477	0.176

* 유사 도를 구하는 식은 식(15)와 같다.

$$SIM = \sum_{i=1}^{t} \overline{w}_{i,q}\ \overline{w}_{i,j} = \frac{\sum_{i=1}^{t} w_{i,q}\ w_{i,j}}{\sqrt{\sum_{i=1}^{t} w^2_{i,q}} \sqrt{\sum_{i=1}^{t} w^2_{i,j}}} \quad (15)$$

용어	a	arrived	damaged	delivery	fire	gold
D1	0	0	.477	0	.477	.176
D2	0	.088	0	.239	0	0
D3	0	.176	0	0	0	.176
Q	0	0	0	0	0	.176
용어	in	of	silver	shipment	truck	
D1	0	0	0	.176	0	
D2	0	0	.477	0	.088	
D3	0	0	0	.176	.176	
Q	0	0	.477	0	.176	

SIM(Q,D1) = 0.031
SIM(Q,D2) = 0.243
SIM(Q,D3) = 0.062

위의 예에서는 문헌 D2가 가장 질의어에 유사한 문헌으로 생각할 수 있다.

1. 적합성 피드백

정보검색에서 가장 중요하면서도 어려운 문제 중의 하나는 사용자가 원하는 정보를 찾기 위한 효율적인 질의를 작성하는 일이다. 하지만 전체 문서집합의 구성에 대해 미리 알고 있지 않는 한 이상적인 최적의 질의는 작성할 수 없다. 대신 최초에는 시험적 질의(tentative query)로 검색을 수행한 후,

이전의 검색 결과에 대한 평가에 기반 하여 다음 번 검색의 질의를 개선시키는 방법이 적합성 피드백(relevance feedback)이다. 일반적인 적합성 피드백은 그림 1과 같다.

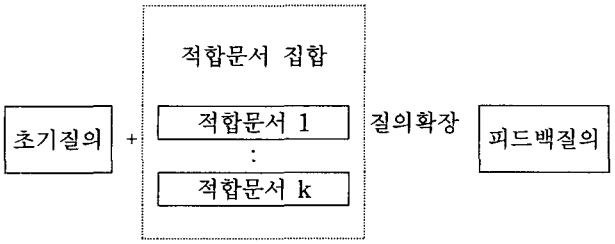

그림 1. 일반적인 적합성 피드백
Fig. 1. General Relevance Feedback

적합성 피드백은 특정 질의와 적합한 문서들은 유사한 벡터로 표현된다고 가정한다. 따라서 어떤 문서가 주어진 질의에 적합하다고 판단되면 질의를 적합한 문서와의 유사도가 증가하도록 변환하여 질의를 개선시킨다. 이렇게 개선된 질의는 최초 적합하다고 판단된 문서와 유사한 문서들을 추가적으로 검색하여 더 많은 양의 적합 문서를 검색해 낼 수 있다.

실제로 적합성 피드백을 이용할 때에는 전체 문서집합에 대해 적합문서와 부적합문서에 대한 사전 지식이 없으므로, 이미 적합성이 알려져 있는 문서들의 정보에 기반 하여 질의 확장을 수행한다.

적합성 피드백을 통해 질의를 확장해 가는 과정은 식(16)과 같이 표현할 수 있다[1].

$$Q^{new} = \alpha Q^{old} + \frac{\beta}{|R|} \sum_{D_i \in R} D_i - \frac{\gamma}{|N|} \sum_{D_i \in N} D_i \qquad (16)$$

여기서 Q^{new}는 새로 확장된 피드백 질의 벡터를 Q^{old}는 확장되기 전 단계의 질의 벡터를 의미한다. R과 N은 각각 초기 검색된 문서집합 중에서 적합하다고 판단된 문서집합과 부적합하다고 판단된 문서집합을 |R|과 |N|은 각각 해당 문서집합의 문서 개수를 뜻한다.

$\sum_{D_i \in R} D_i$ 은 검색된 문헌 중에서 적합한 문헌 용어 가중치의 합을 의미하며 $\sum_{D_i \in N} D_i$은 검색된 문헌 중에서 비 적합한 문헌 용어 가중치의 합을 의미한다. 질의어 확장에 관한 예를 다음과 같이 들어

질의어 Q={병렬(0.5) 프로그램(0.5)}로 검색한 결과가 표 1과 같다고 가정하자.

표 1. 검색 문서
Table 1. Retrieval Document

검색된 문서	문서벡터
	용어(가중치)
1	병렬(0.3) 프로그램(0.2) 시스템(0.5)....
2	병렬(0.7) 프로그램(0.0).... 처리(0.5).....
.....	.....
99	병렬(0.1) 프로그램(0.1) 시스템(0.9)...

검색된 문서 99번이 사용자가 진정으로 원하는 문서라면 질의 확장 수식을 통하여 수정된 질의어 Qm ={병렬(0.4) 프로그램(0.4) 시스템(0.2)}의 변경과 용어 가중치가 변경된다.

표 2. 질의 확장된 검색 문서
Table 2. Retrieval Document of Query Extension

검색된 문서	문서벡터
	용어(가중치)
1	병렬(0.3) 프로그램(0.2) 시스템(0.5)....
2	병렬(0.2) 프로그램(0.1) 시스템(0.9).....
.....	.....
99	.....

질의 확장을 통하여 검색하면 보다 정확한 문헌들을 검색할 수 있다.

2. 지역적 문맥 분석

적합성 피드백 과정에서 사용자는 검색된 상위 문헌을 검사하여 연관 문헌과 비 연관 문헌의 두 범주로 분류하는데 질의 확장을 위해 용어를 선택할 때 지역적 문맥 분석 정보가 사용된다.

질의 확장의 목적은 더 많은 연관 문헌을 검색하는 것이므로 피드백 전략은 클러스터링 개념을 기초로 하고 있다. 이 개념에 따르면 연관 문헌으로 판정된 문헌은 더 큰 연관 문헌 클러스터를 나타내는 용어를 포함하고 있으며 이 경우 더 큰 연관 문헌 클러스터를 나타내는 질의는 사용자의 도움을 받아서 점진적으로 작성된다.

전역적 방법에는 컬렉션 내 전체 문헌을 사용하여 용어 연관성을 나타내는 전역적 유사 시소러스(thesaurus) 구조를 작성하며 사용자는 자신에게 제시된 이 구조를 이용하여 질의 확장을 위한 용어를 선택한다.

지역적 방법에는 질의어에 의해 검색된 문헌들을 이용하여 질의 시간에 질의 확장을 위한 용어들을 선택한다. 이 방법은 적합성 피드백 과정과 유사하지만 사용자의 도움이 필요 없다는 점에서 다르다[11,12].

지역적 방법에서 자동 색인어 선정 방법에는 명사 집단 식별 방법이 있다. 자연 언어 텍스트 문장은 일반적으로 명사, 대명사, 관사, 동사, 형용사, 부사, 접속사 등으로 구성되어 있다. 각 문법 범주에 속한 단어들이 나름대로 특정 목적에 사용되는 반면, 대부분의 의미는 명사에 의해 전달된다고 볼 수 있다. 따라서 색인어 선정에서는 텍스트에 나타난 명사를 사용하는 것이 직관적으로 보아 타당한 전략이며 이 방법에서는 명사를 제외한 나머지 단어를 체계적으로 제거하면 된다.

두 세 개의 명사가 결합하여 단일 개념을 나타내는 경우가 흔하기 때문에 텍스트에서 가까이 있는 명사들을 모아 하나의 색인 개념으로 처리하는 것은 타당성이 있다. 그래서 색인어로 단순히 명사만 사용하는 대신 텍스트에서 구문적 거리 즉, 두 명사 사이의 단어 수로 측정하여 미리 정해놓은 임계 치를 초과하지 않는 명사들의 집합인 명사 집단(noun group)을 사용하는 방법을 취한다.

명사 집단을 색인어로 취할 경우, 단일 용어가 아니라 용어의 집단 집합으로 문헌의 개념적 논리상을 얻게 된다. 지역적 문맥 분석 피드백(Local Context Analysis Feedback: LCAF) 과정의 정보 검색은 그림 2와 같다.

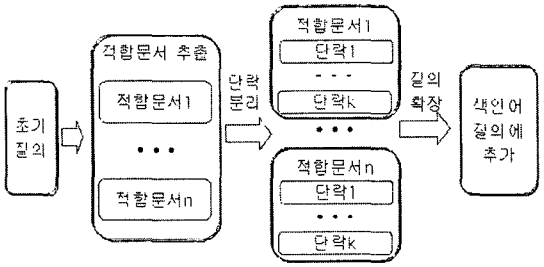

그림 2. LCAF의 정보 검색 모델
Fig. 2. Information Retrieval Model of LCAF

본 논문에서 적용한 지역적 문맥 분석 과정은 다음 3단계를 거쳐 수행된다.

첫째, 현재 질의를 사용하여 상위 n개의 단락을 검색한다. 이 과정은 현재 질의에 의해 초기 검색된 문헌을 300단어 길이의 단락으로 분할한 후 단락을 마치 문헌처럼 순위 화함으로써 수행된다.

둘째, 상위 순위 단락에 나타나는 각 개념 c에 대해 tf-idf 방법의 한 변형을 이용하여 해당 개념과 전체 질의와의 유사도 sim(q, c)를 계산한다.

셋째, m개의 상위 순위 개념이 원래 질의에 추가된다. 추가된 각 개념에 대해 1-0.9×i/m의 가중치가 부여되는데, 여기서 i는 최종 개념 순위에서 해당 개념의 위치이다. 원래 질의 q에 있던 용어들은 가중치를 2로 부여함으로써 강조될 수도 있다.

이 3단계 중 두 번째 단계는 각 연관 개념 c와 원래 질의 q 사이의 유사도 sim(q,c)는 식(17)에 의하여 산출된다.

$$sim(q,\ c) = \prod_{k_i \in q} \left(\delta + \frac{\log(f(c,k_i) \times idf_c}{\log n} \right)^{idf_i} \qquad (17)$$

식(14)에서 n은 상위 순위 단락의 수이다. 함수 $f(c,k_i)$는 개념 c와 질의 용어 k_i사이의 연관 도를 나타내며 식(18)을 이용하여 계산한다.

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j} \qquad (18)$$

식(18)에서 $pf_{i,j}$는 j번째 단락 내에서의 용어 ki의 출현 빈도를 나타내며 $pf_{c,j}$는 j번째 단락 내에서의 개념 c의 출현 빈도를 나타낸다. 식(18)은 연관 클러스터를 위해 정의된 표준 연관 척도 수식이지만 개념과 용어 출현 빈도를 측정하기 위해 단락에 적용하고 있다. 역 문헌 빈도인수는 식(19)에 의해 산출된다.

$$idf_i = \max\left(1, \frac{\log_{10} N/np_i}{5}\right) \qquad (19)$$

$$idf_c = \max\left(1, \frac{\log_{10} N/np_c}{5}\right) \qquad (20)$$

식(19)과 식(20)에서 N은 컬렉션 내의 단락 수를 나타내며, np_i는 용어 ki를 가진 단락의 수, nc_i는 개념 c를 포함하는 단락 수를 나타낸다. 인수 δ는 sim(q, c)가 0이 되는 것을 피하기 위한 상수인데 보통 0.1에 가까운 상수이다. 지수 부분의 idf_i인수는 저 빈도 질의 용어들을 강조하기 위함이다.

Ⅳ. 실험 및 결과

본 논문에서는 성능 평가 자료로서 CHODIC를 사용하였으며,. CHODIC은 500개의 문서와 21개의 질의로 구성되어 있다. 문서와 질의 사이의 연관성 평가는 문서 제목을 기준으로 설정하였다.

표 3. 문서 집합
Table 3. Document Collection

문서집합	문서내용
1.html	Comaromi,J.P. The present ...
...	...
500.html	Chernyi,A.I. At the present ...

표 4. 불 용어 집합
Table 4. Non-term Collection

불 용어
a about above accordingly across after ...

표 5. 질의어 집합
Table 5. Query Language Collection

질의번호	질의어
q1	Titles &(automatical | ...)
...	...
q21	Government&(information ...)

표 6. 적합문서 집합
Table 6. Collection of Relevance Document

질의번호	적합문서 집합
q1	28 35 42 52 65 76 86 ...
..	...
q35	12 18 54 75 92 119 126 ...

정보 검색 시스템의 검색 효과는 재현 율(Recall)과 정확율(Precision)을 이용하여 평가된다.

재현 율은 문서집합에서 사용자가 원하는 문서를 어느 정도 검색하였는가를 나타내고, 정확 율은 검색된 문서들 중에서 사용자가 원하는 문서가 얼마나 포함되어 있는가를 나타낸다.

A는 검색된 문서들, B는 검색된 문서 중 질의에 관련된 문서, C는 사용자가 원하는 문서들을 나타내고 있다. 정확율(P)과 재현 율(R)의 식(21), (22)과 같다.

$$P = \frac{B}{A+B} \qquad (21)$$

$$R = \frac{B}{B+C} \qquad (22)$$

예를 들어 질의 q에 대한 검색 문헌이 표 7과 같다고 가정하자.

표 7. 검색 문헌
Table 7. Retrieval Document

1	D123*	6	D9*	11	D36	* : 질의 q의 적합 문헌을 가정함
2	D84	7	D511	12	D48	
3	D56*	8	D129	13	D250	
4	D6	9	D187	14	D113	
5	D8	10	D25*	15	D3*	

질의 q에 대한 적합 문헌 집합 Rq ={D3, D56, D129}라면 재현율과 정확 율은 표 8과 같다.

표 8. 재현 율, 정확 율
Table 8. Recall and Precision

적합문헌	재현 율	정확 율
D56	33.3(1/3)	33.3(1/3)
D129	66.6(2/3)	25(2/8)
D3	100(3/3)	20(3/15)

문헌수가 많을수록 질의 q에 대한 적합문헌 집합이 검색될 비율이 높아지므로 재현 율은 증가하나 정확 율은 떨어진다. 표11, 12에서 보면 RF보다 LCAF에서 재현 율, 정확 율의 성능이 향상되었고, 정확 율 하강 현상이 완만하였다.

표 9는 MMM, Paice, P-norm, RF 모델의 검색효과를 보여준다. 본 논문에서는 검색 효과를 평가하기 위하여 질의들에 대한 평균 정확 율을 계산한다. 각각의 질의에 대한 정확율은 재현 율을 0.25, 0.5, 0.75에 고정시켜 계산된 정확 율들의 평균값이다. 또한 표에 나타난 검색 효과는 가장 높은 검색 효과를 나타내는 매개변수에 대한 것이다. RF(Relevance Feedback), P-norm 모델이 MMM, Paice 모델보다 높은 검색 효과를 제공한다.

MMM 모델은 Paice 모델보다 높은 검색 효과를 나타내고 있다. 이는 색인어 가중치가 역 문헌빈도와 출현 빈도로부터 유도하기 때문에 검색 효과를 저하시키는 요인이 발생하지 않았기 때문이다.

표 9. 검색 효과 비교(단위 : 정확 율)
Table 9. Retrieval Effect Comparison(Unit: Precision)

구분	평균
MMM	0.327
Paice	0.318
P-norm	0.362
RF	0.730

표 8에서는 지역적 문맥 분석 피드백(Local Context Analysis Feedback)의 검색결과가 적합성 피드백(Relevance Feedback) 결과에 비해 정확 율이 12.82% 향상을 보여주고 있다.

표 10. 질의 분해 피드백
Table 10. Local Context Analysis Feedback

구분	평균
Relevance Feedback	0.73
LCAF	0.78

표11과 표12에서는 검색문서수를 다르게 제한했을 때 검색효율이 어떻게 달라지는가를 분석한 것이다. 적합성 피드백(RF), 지역적 문맥 분석 피드백을 이용하여 검색문서수를 10건과 20건으로 제한한 경우 재현율과 정확 율을 보여 주고 있다.

검색문서수를 10건으로 제한하였을 때 지역적 문맥 분석 피드백 결과는 적합성 피드백 결과에 비해 재현율과 정확 율이 각각 8.33%, 2.67%가 향상되었다.

검색문서수를 20건으로 제한하였을 때 지역적 문맥 분석 피드백 결과는 적합성 피드백 결과에 비해 재현 율과 정확 율이 각각 1.3%, 4.23%가 향상되었다.

표 11. RF, LCAF의 재현율 비교
Table 11. Recall Comparison of RF, LCAF

척도 / 검색 문서 수	재현 율	
	RF	LCAF
문서 수 ≤ 10	0.48	0.52
문서 수 ≤ 20	0.78	0.79

표 12. RF, LCAF의 정확 율 비교
Table 12. Precision Comparison of RF, LCAF

척도 / 검색 문서 수	정확 율	
	RF	LCAF
문서 수 ≤ 10	0.75	0.78
문서 수 ≤ 20	0.71	0.74

그림 3은 검색 문서 수 20건으로 제한하였을 때 적합성 피드백, 지역적 문맥 분석 피드백 검색결과를 재현율과 정확 율로 표현한 성능곡선으로 비교한 것이다. 재현율의 증가에 따른 정확 율의 하강 현상이 두드러지지 않고 있음을 볼 수 있다.

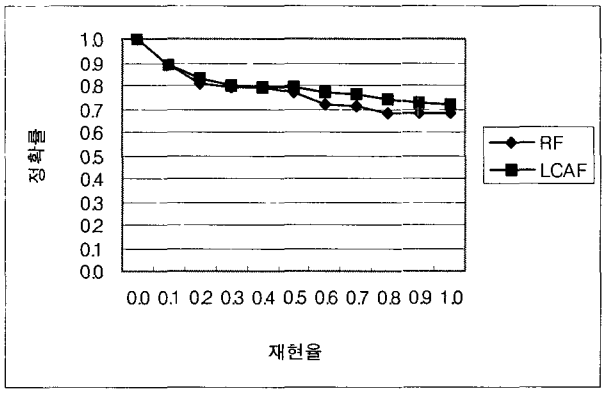

그림 3. RF, LCAF의 실험 결과
Fig. 3. Experimentation Result of RF and LCAF

V. 결 론

정보 검색 시스템의 중요한 목적중의 하나는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여함으로써 사용자들이 필요한 정보를 얻는데 소모되는 시간을 최소화시키는 것이다.

이러한 부울 검색 시스템의 단점을 보완하기 위해 퍼지 집합 모델이 제안되었으나 퍼지 집합 모델은 단일 피연산자 의존 문제로 인하여 많은 경우에 부정확한 문서 값을 생성하는 것으로 알려져 왔다.

퍼지 집합 모델의 문제점을 개선하는 방법으로서 MMM 모델, Paice 모델, P-norm 모델이 개발되었다.

본 논문에서는 웹 정보검색 분야에서 사용되는 적합성 피드백에 기초하여 높은 검색 효과를 제공하는 지역적 문맥 분석 피드백 모델을 사용하였다. 실험 결과 지역적 문맥 분석 피드백 방법으로 정보 검색 하는 경우 보다 더 좋은 정밀도를 보였고, 하강 현상이 완만하였다.

결과를 통하여 지역적 문맥 분석 피드백 모델은 기존의 방법들보다 높은 검색 효과를 제공됨을 알 수 있다. 본 연구에서 어려웠던 점은 질의 시간에 문헌을 분석함으로 인하여 검색 엔진에 부하가 걸리는 점이다. 따라서 검색 엔진의 질의 처리 속도를 향상시키는 연구가 필요하다고 생각된다.

참 고 문 헌

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley Publishing Company, 1999.

[2] Daniel Marcu, "Discourse trees are good indicators of importance in text", In Inderjeet Mani and Mark Maybury, eds, Advances in Automatic Text Summarization, pp.123-136, The MIT Press, 1999.

[3] Mark Sanderson, "Accurate User Directed Summarization from Existing Tools", In Proceedings of the 7th International Conference on Information and Knowledge Management, pp.45-51, 1998.

[4] Regina Barzilay and Michael Elhadad, "Using Lexical Chains for Text Summarization", In Inderjeet Mani and Mark Maybury, eds, Advances in Automatic Text Summarization, pp.111-121, The MIT Press, 1999.

[5] Anastasios Tombros and Mark Sanderson, "Advantages of Query Biased Summaries" in Information Retrieval, In Proceeding of ACM-SIGIR'98, pp.2-10, 1998.

[6] J.H. Lee, M.H. Kim and Y.J. Lee, "Information Retrieval Based on Conceptual Distance", in Is-a Hierarchies, Journal of Documentation, Vol. 49, No. 2, pp.188-207, 1993.

[7] M.H. Kim and J.H. Lee and Y.J Lee, "Analysis of Fuzzy Operators for High Quality Information Retrieval", Information Processing Letters, Vol. 46, No. 5, pp.251-256, 1993.

[8] G.Salton, Automatic Text Processing "The Transformation, Analysis, and Retrieval of Information by Computer", Addison Wesley, 1989.

[9] J.H.Lee, W.Y. Kim, M.H. Kim and Y.J. Lee, "Enhancing the Fuzzy Set Model with Positively Compensatory Operators", Proceedings of the 3rd International Symposium on Database Systems on Advanced Applications, Taejon, Korea, pp. 368-375, 1993.

[10] Inderjeet Mani, David House, Gary Kein, Lynette Hirschman, and Leo Obrst, "The TIPSTER SUMMAC Text Summarization", Evaluation Final Report, Technical Report MTR98 W0000138, MITRE, 1998.

[11] Gerard Salton, Automatic Text Processing, "The Transformation, Analysis, and Retrieval of Information by Computer", Addison-wesley Publishing Company, 1989.

[12] K.s, Han, "Automatic Text Summarization Based on Relevance Feedback with Query Splitting", 6, 2000.

[13] J.H, Lee, "An Efficient and Effective Evaluation Method for Boolean Operators", KISS, Vol. 21, No. 3, pp.440-445, 1994.

저 자 소 개

김영천(Young-Cheon Kim)

1992년 : 광주대 전자계산학과 졸업
1996년 : 조선대 컴퓨터공학과 졸업 (공학석사)
2002년 : 조선대 전자계산학과 졸업 (이학박사)
2003년~현재 : 서정대학 정보통신과 교수

관심분야 : 객체지향시스템, 소프트웨어 공학, 유전자 알고리즘, 정보검색

이성주(Sung-Joo Lee)

1970년 : 한남대학교 물리학과 (이학사)
1992년 : 광운대학교 전자계산학과 (이학석사)
1998년 : 대구가톨릭대학교 전자계산학과 (이학박사)
1988년~1990년 : 조선대학교 전자계산소 소장
1995년~1997년 : 조선대학교 정보과학대학장
1981년~현재 : 조선대학교 컴퓨터공학부 교수

관심분야 : 소프트웨어 공학, 프로그래밍 언어, 객체지향 시스템, 러프 집합