질의분해 적합성 피드백을 이용한 검색시스템의 성능 증진에 관한 연구

A Study on Improving the Effectiveness of Retrieval System Using Query Splitting Relevance Feedback

김 영천, 박 병권*, 이 성주**
Young-cheon kim, Byung Gweun Park, Sung-joo Lee

*서강정보대학 정보통신과
**조선대학교 전자계산학과
E-mail : yckim@stmail.chosun.ac.kr

요약

순수한 부울 검색 시스템은 문서와 질의 사이의 유사도를 나타내는 문서값을 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. 본 논문에서는 높은 검색 효과를 제공하는 질의분해 적합성 피드백(QSRF) 모델을 제안한다. 질의 분해 적합성 피드백 모델의 연산 특성이 MMM, Paice, P-norm 모델보다 우수함을 설명하고, 또한 성능 비교를 통하여 이를 입증한다.

Keyword : 정보검색, 적합성 피드백, 질의 분해, 부울 검색, 유사도

Ⅰ. 서론

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 기반하여 다음 번 검색의 질의를 개선시키는 방법이 적합성 피드백(relevance feedback)이다[1].

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.

실제 이 적합성 피드백을 이용할 때에는 전체 문서집합에 대해 적합문서와 부적합문서를 미리 알 수 없으므로, 이미 적합성이 알려져 있는 문서들의 정보에 기반 하여 질의확장을 수행한다[2].

본 논문에서는 질의 분해 적합성 피드백 모델을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 부울 연산자를 유연하게 연산하는 기존의 방법들 MMM, Paice, P-norm 모델에

대하여 기술한다. 3장에서는 높은 검색 효과를 제공하는 질의 분해 적합성 피드백 모델을 제안한다. 4장에서는 질의 분해 적합성 피드백 모델과 MMM, Paice, P-norm 모델의 성능을 비교한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

II. 부울 연산자를 유연하게 연산하는 기존의 방법들

퍼지 집합 모델의 단일 피연산자 의존 문제를 극복하기 위해 MMM 모델, Paice 모델, P-norm 모델이 개발되었다.

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 문서값이 서로 다르다. MMM, Paice, P-norm 모델은 이러한 문제점을 다항연산을 가능하게 함으로써 극복하였다.

$$F(d,\ t_1\ AND\ t_2) = MIN(w_1, w_2) \quad \ldots \quad (1)$$

$$F(d,\ t_1\ OR\ t_2) = MAX(w_1, w_2) \quad \ldots \quad (2)$$

(a) 퍼지 집합 모델

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

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

(b) MMM 모델

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

($0 \le r \le 1$, w_i'는 오름차순정렬)

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

($0 \le r \le 1$, w_i'는 내림차순정렬)

(c) Paice 모델

$$F(d,\ t_1\ AND \ldots AND\ t_n) = 1 - \left[\frac{(1-w_1)^p + \ldots + (1-w_n)^p}{n}\right]^{\frac{1}{p}} \quad \ldots (7)$$

($1 \le p \le \infty$)

$$F(d,\ t_1\ OR \ldots OR\ t_n) = \left[\frac{w_1^{\,p} + \ldots + w_n^{\,p}}{n}\right]^{\frac{1}{p}} \quad \ldots (8)$$

($1 \le p \le \infty$)

(d) P-norm 모델

2.1 정규화

확장된 부울 검색 체계를 기반으로 하는 검색 모델은 문서값을 계산하기 위하여 색인어 가중치를 사용한다. 색인어 가중치는 역문헌빈도(Inverse Document Frequency)와 색인어출현빈도(Term Frequency)로부터 유도될 수 있다. N이 문서 집합을 구성하는 문서들의 수이고, n_k가 색인어 k가 출현하는 문서들의 수일 때, 색인어 k의 역문헌빈도 IDF_k는 $\log(N/n_k)$로 정의된다. 색인어 추현빈도 TF_{ik}는 문서 i에서 색인어 k의 출현빈도를 의미한다. 문서 i에서 색인어 k의 색인어 가중치 W_{ik}는 $IDF_k \cdot F_{ik}$로 정의될 수 있다. 확장된 부울 검색 체계에서 색인어 가중치는 0부터 1 사이의 값이어야 하기 때문에 W_{ik}는 식(9)와 같이 정규화된다.

$$W_{ik} = \frac{TF_{ik}}{\max TF\ in\ document\ i} \cdot \frac{IDF_k}{\max IDF\ in\ document\ i} \quad \ldots \quad (9)$$

III. 질의 분해 적합성 피드백 모델

문서의 제목과 내용 문장만을 시스템의 입력으로 준다. 입력된 문서에 대해의 한국어 품사 태거를 이용하여 명사만 추출한 후 벡터를 구성한다. 이 때 벡터의 가중치를 어떻게 부여할 것인지를 결정하기 위한 방법은 이진벡터(TF_{bin}), 문자내에서의 단어의 빈도로 벡터를 구성하는 방법(TF), 그리고 정규화(normalization) 시 켜서 벡터를 구성하는 방법(TF_{norm}) 등이 있다.

3.1 적합성 피드백

처음에는 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 [11]. 일반적인 적합성 피드백은 그림 1과 같다.

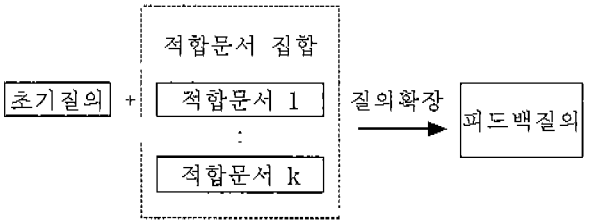

그림 1. 일반적인 적합성 피드백

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.

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

3.2 질의분해 적합성 피드백

적합 문서를 이용하여 양성 피드백(Positive Feedback)으로 초기질의를 확장해 나간다. 이 때 질의 확장을 이용한 기존의 문서 검색에서는 그림 1에서처럼 k개의 문서를 통합하여 초기질의를 확장하는데 사용한다. 따라서 질의확장이 적용된 후의 질의는 여전히 하나만 남게 된다. 이 경우 적합문서에 노이즈가 섞일 가능성이 크며, 이 적합문서를 이용해 질의확장을 적용할 경우 너무 포괄적인 질의로 확장이 된다. 이 문제를 완화하기 위해서 본 시스템에서는 통합하여 하나의 질의로 확장하지 않고 그림 2와 같이 k개의 적합문서들에 대해 각각 개별적으로 질의확장을 수행하여 k개의 피드백 질의를 생성한다. 이렇게 질의를 분해함으로써 노이즈 문서에 대해 좀더 배타적인 정보검색을 할 수 있다.

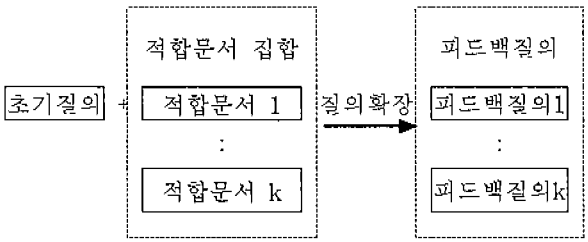

그림 2. 질의분해 피드백 질의

적합성 피드백을 통해 질의를 확장해 가는 과정은 다음과 같이 식으로 표현할 수 있다[1].

Standard Rocchio 는 식(10)과 같다.

$$Q^{new} = \alpha Q^{old} + \frac{\beta}{|R|} \sum_{D_i \in R} D_i - \frac{\gamma}{|N|} \sum_{D_i \in N} D_i \quad .. (10)$$

Ide Regular는 식(11)와 같다.

$$Q^{new} = \alpha Q^{old} + \beta \sum_{D_i \in R} D_i - \gamma \sum_{D_i \in N} D_i \quad ... (11)$$

Ide Dec Hi는 식(12)와 같다.

$$Q^{new} = \alpha Q^{old} + \beta \sum_{D_i \in R} D_i - \gamma \max_{n-r}(D_i) \quad .. (12)$$

여기서 $Q^{new는}$ 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 해

당 문서집합의 문서 개수를 뜻한다. $max_{n-r}(D_i)$는 부적합문서 중 최상위 문서를 나타낸다. α, β, γ는 이전 단계의 질의, 적합문서집합, 부적합문서집합 간의 중요도를 조율하는 상수이다.

3.3 적합 문서 추출

초기질의와 각 문서 사이의 유사도 계산은 정보검색에서 많이 사용하는 코사인 유사도 (cosine similarity) 식 (13)을 이용한다.

$$sim(S_j, Q^0) = \frac{S_j \cdot Q^0}{|S_j| \times |Q^0|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad \cdots (13)$$

여기서, Sj는 각 문서 벡터, Q^0는 초기질의 벡터를 의미하고, w_{ij}와 w_{iq}는 단어 i가 각각 문서와 초기질의에서 갖는 가중치이다. t는 각 문서 벡터와 초기질의 벡터를 생성하는데 사용된 단어의 총 개수이다. 식 (13)에 의한 유사도 값에 따라 문서를 내림차순 정렬한 후 유사도 값이 큰 상위 k개의 문서를 적합 문서로 간주한다.

Ⅳ. 실험 및 결과

4.1 성능 평가 결과

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

정보 검색 시스템의 검색 효과는 재현능력도와 검색정밀도를 이용하여 평가한다. 검색정밀도(P)와 재현능력도(R)의 식(14), (15)과 같다.

$$P = \frac{B}{A+B} \cdots (14) \qquad R = \frac{B}{B+C} \cdots (15)$$

표 1은 MMM, Paice, P-norm, RF 모델의 검색효과를 보여준다. 본 논문에서는 검색 효과를 평가하기 위하여 질의들에 대한 평균 검색정밀도를 계산한다. 각각의 질의에 대한 검색정밀도는 재현능력도를 0.25, 0.5, 0.75에 고정시켜 계산된 검색정밀도들의 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.

표 1. 검색 효과 비교(단위 : 검색정밀도)

	Average
MMM	0.327
Paice	0.318
P-norm	0.461
RF	0.602

표 2에서는 질의분해 적합성 피드백(Query Splitting Relevance Feedback)과 적합성 피드백(Relevance Feedback)을 이용하여 실험한 검색정밀도를 보여주고 있다.

표 2. 질의 분해 피드백(단위 : 검색정밀도)

	Average
Relevance Feedback	0.602
QSRF	0.623

표3과 표4에서는 검색문서수를 다르게 제한했을 때 검색효율이 어떻게 달라지는가를 분석한 것이다. P-norm, 적합성 피드백(RF), 질의분해 적합성 피드백(QSRF)를 이용하여 검색문서수를 10건과 20건으로 제한한 경우 재현능력도와 검색정밀도를 보여 주고 있다.

검색문서수를 10건으로 제한하였을 때 적합성 피드백 검색 결과는 P-norm 결과에 비해 재현능력도와 검색정밀도가 각각 56%, 50% 향상되었고, 질의분해 적합성 피드백 결과는 적합성 피드백 결과에 비해 재현능력도와 검색정밀도가 각각 2.6%, 3.2%가 향상되었다.

검색문서수를 20건으로 제한하였을 때 적합성 피드백 검색 결과는 P-norm 결과에 비해 재현능력도와 검색정밀도가 각각 46.81%, 48.72% 향상되었고, 질의분해 적합성 피드백 결

과는 적합성 피드백 결과에 비해 재현능력도와 검색정밀도가 각각 1.45%, 5.17%가 향상되었다.

표 3. P-norm, RF, QSRF의 재현능력도 비교

척도 / 검색문서수	재현능력도		
	P-norm	RF	QSRF
문서수≤10	0.25	0.39	0.40
문서수≤20	0.47	0.69	0.70

표 4. P-norm, RF, QSRF의 검색정밀도 비교

척도 / 검색문서수	검색정밀도		
	P-norm	RF	QSRF
문서수≤10	0.42	0.63	0.65
문서수≤20	0.39	0.58	0.61

그림 3은 검색 문서수 20건으로 제한하였을 때 P-norm 초기검색과 적합성 피드백, 질의분해 적합성 피드백 검색결과를 재현능력도와 검색정밀도로 표현한 성능곡선으로 비교한 것이다. 재현능력도의 증가에 따른 검색정밀도의 하강 현상이 두드러지지 않고 있음을 볼 수 있다.

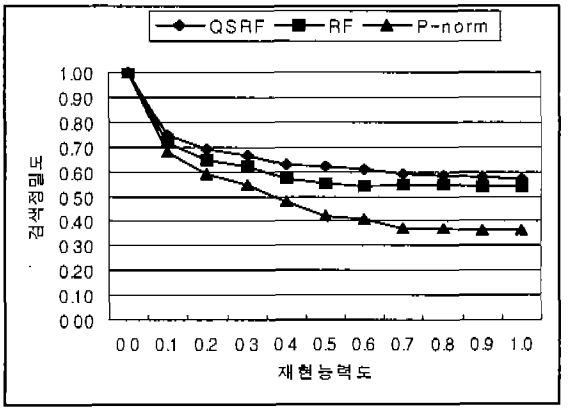

그림 3. P-norm 초기와 RF,QSRF의 실험 결과

V. 결론

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.

본 논문에서는 정보 검색 분야에서 사용되는 적합성 피드백에 기초하여 높은 검색 효과를 제공하는 질의 분해 적합성 피드백(QSRF) 모델을 제안하였다. 실험 결과 제안하는 질의 분해 적합성 피드백 방법으로 정보검색하는 경우가 질의분해 적합성 피드백을 이용하지 않는 경우 보다 더 좋은 정밀도를 보였다.

QSRF 모델은 기존의 방법들보다 높은 검색 효과를 제공함을 성능 비교를 통하여 입증되었다.

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] K.S. Lan, D.H. Baek and H.C. Rim, Automatic Text Summarization Using Query Expansion, KISS, v. 27, pp. 339-341, 2000.