검색 문서의 분류 정보에 기반한 용어 클러스터 질의 확장 모델

*강현수, *강현규, *박세영, **이용석

*한국전자 통신연구원 지식정보연구부 문서정보연구팀
**전북대학교 컴퓨터과학과

A Term Cluster Query Expansion Model Based on Classification Information of Retrieval Documents

*Hyun-Su Kang, *Hyun-Kyu Kang, *Se-Young Park, **Yong-Seok Lee

*Document Information Research Team, Dept. of Knowledge Information, ETRI
**Dept. of Computer Science, Chonbuk National University

요 약

정보 검색 시스템은 사용자 질의의 키워드들과 문서들의 유사성(similarity)을 기준으로 관련 문서들을 순서화하여 사용자에게 제공한다. 그렇지만 인터넷 검색에 사용되는 질의는 일반적으로 짧기 때문에 보다 유용한 질의를 만들고자 하는 노력이 지금까지 계속되고 있다. 그러나 키워드에 포함된 정보가 제한적이기 때문에 이에 대한 보완책으로 사용자의 적합성 피드백을 이용하는 방법을 널리 사용하고 있다. 본 논문에서는 일반적인 적합성 피드백의 가장 큰 단점인 빈번한 사용자 참여는 지양하고, 시스템에 기반한 적합성 피드백에서 배제한 사용자 참여를 유도하는 *검색 문서의 분류 정보에 기반한 용어 클러스터 질의 확장 모델(Term Cluster Query Expansion Model)*을 제안한다. 이 방법은 검색 시스템에 의해 검색된 상위 n개의 문서에 대하여 분류기를 이용하여 각각의 문서에 분류 정보를 부여하고, 문서에 부여된 분류 정보를 이용하여 분류 정보의 수(m) 만큼으로 문서들을 그룹을 짓는다. 적합성 피드백 알고리즘을 이용하여 m개의 그룹으로부터 각각의 용어 클러스터(Term Cluster)를 생성한다. 이 클러스터가 사용자에게 문서 대신에 피드백의 자료로 제공된다. 실험 결과, 적합성 알고리즘 중 Rocchio방법을 이용할 때 초기 질의보다 나은 성능을 보였지만, 다른 연구에서 보여준 성능 향상은 나타내지 못했다. 그 이유는 분류기의 오류와 문서의 특성상 한 영역으로 규정짓기 어려운 문서가 존재하기 때문이다. 그러나 검색하고자 하는 사용자의 관심 분야나 찾고자 하는 성향이 다르더라도 시스템에 종속되지 않고 유연하게 대처하며 검색 성능(retrieval effectiveness)을 향상시킬 수 있다.

1. 서 론

정보 검색 시스템은 사용자 질의의 키워드들과 문서들의 유사성(similarity)[3,7,13]을 기준으로 관련 문서들을 순서화하고 검색 결과를 사용자에게 제공한다. 그러나 사용자가 인터넷 정보 검색에서 이용하는 질의는 일반적으로 짧은 경향이 있다. 이는 검색 성능(retrieval effectiveness)을 저하시키는 주요 원인 중의 하나이다. 그렇지만 질의에 포함된 키워드의 중요성 때문에 보다 유용한 질의를 만들고자 하는 노력이 지금까지 계속 진행되고 있다.

질의 확장(Query Expansion)[8,12]은 사용자가 제시한 키워드들과 관련된 단어들을 추가하여 보다 많은 관련 문서들을 검색하고자 하는 방법이다. 이를 위하여 시소러스나 워드넷, 코퍼스에 나타난 단어의 관계(예를 들면 상호 정보)를 이용하여 질의를 확장하였다. 그러나 키워드에 포함된 정보가 제한적이기 때문에 이에 대한 보완책으로 사용자의 적합성 피드백을 이용하는 방법을 널리 사용하고 있다.

적합성 피드백(Relevance Feedback)[1,2,10]은, 질의 확장에서 이용한 키워드들 대상으로 하는 것이 아니라 검색된 문서를 대상으로 사용자가 판단한 적합성 정보에 근거하여 질의를 확장하여 검색하는 방법이다. 검색 시스템은 사용자가 적합하다고 판단한 적합 문서를 이용하여 이 문서와 유사한 새로운 문서를 검색할 수 있으며, 또한 사용자가 판단한 부적합 문서를 이용하여 이 문서와 이 문서와 유사한 문서의 순위는 낮춤으로 전체적인 검색 성능을 향상시킨다.

그러나 적합성 피드백의 가장 큰 단점[1,5]은 검색된 문서마다 사용자가 적합/부적합 정보를 판단해야 한다는 것이다. 이는 신속한 정보를 원하는 사용자에게 현실적으로 수용하기 힘든

요구 사항이다. 이러한 불편 때문에 시스템에 기반한 적합성 피드백[5]에 의한 방법도 연구되었다. 시스템에 기반한 적합성 피드백은 상위 n개의 문서가 모두 적합한 문서라고 가정하고, 상위 n개의 문서를 모두 적합성 피드백에 참여시키는 방법이다. 그러나 이 방법은 초기 검색 결과의 성능에 밀접하게 의존적이다. 상위 n개의 문서 중에서 대부분의 검색 문서가 사용자의 요구를 만족시키지 못했을 때에는 성능은 오히려 크게 감소한다.

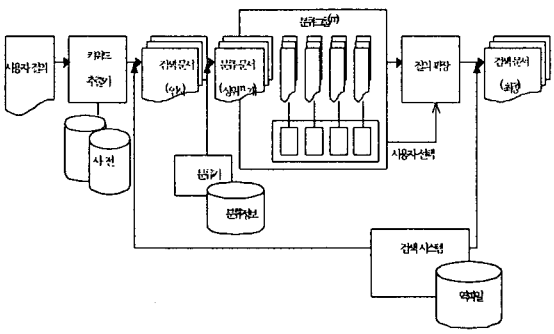
따라서, 본 논문에서는 적합성 피드백의 빈번한 사용자 참여는 지양하고, 시스템에 기반한 적합성 피드백에서 배제한 사용자 참여를 유도하는 검색 문서의 분류 정보에 기반한 용어 클러스터 질의 확장 모델(Term Cluster Query Expansion Model)을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 용어 클러스터 질의 확장 모델의 시스템 구성 및 각 모듈에 대하여 설명한다. 3장에서는 용어 클러스터 질의 확장 모델의 처리를 실제 예를 들어서 설명하고 4장에서는 실험 및 평가를, 그리고 마지막으로 5장에서는 결론을 맺는다.

2. 용어 클러스터 질의 확장 모델

본 논문에서 제안한 용어 클러스터 질의 확장 모델(Term Cluster Query Expansion Model, TCQEM)은 사용자의 적합성 피드백 정보를 기반으로 하여 초기 질의를 확장하여 검색 성능의 향상을 꾀한다. 그리고 일반적으로 적합성 피드백에서 가장 큰 단점인 사용자의 빈번한 참여를 최소화면서 사용자의 적극적인 참여를 유도한다.

2.1시스템 구성



[그림 1] 용어 클러스터 질의 확장 모델

본 논문에서 제안한 용어 클러스터 질의 확장 모델(Term Cluster Query Expansion Model, TCQEM)의 시스템 구성도는 [그림 1]과 같다.

용어 클러스터 질의 확장 모델에서는 먼저 사용자 질의로부터 키워드를 추출하여 1차 검색을 실행한다. 1차 검색된 상위 n개 문서에 대하여 분류기를 이용하여 각각의 분류 정보를 추출한다. 추출된 분류 정보에 따라 그룹들을 설정하고 각각의 그룹으로부터 그룹을 대표하는 용어 클러스터를 생성하여 사용자에게 적합한 후보 클러스터를 선택하도록 한다. 이 선택된 후보 클러스터를 이용하여 검색 시스템은 재검색을 하여 사용자에게 최종 검색 결과를 제공한다.

2.2 카테고리 설정

사용자의 초기 질의로부터 검색 시스템에 의하여 검색된 문서의 카테고리를 설정하기 위하여 문서 분류기(TAXON)[9]을 이용하였다. TAXON은 문서를 축소된 벡터로 표현하고, 문서와 주제에 대한 벡터의 관련도를 검사하여 문서분류를 한다. 또한 시소러스 도구를 사용하여 문서의 주제와 밀접하게 관련된 의미를 획득하여 개념 기반의 문서 분류가 가능하다. [표 1]은 실험에 사용된 TAXON의 성능이다.

[표 1] 문서 분류기(TAXON)의 성능

	학습 문서		비학습 문서	
	Best-1	Best-2	Best-1	Best-2
76주제 성공율	0.566	0.707	0.543	0.694
12주제 성공률	0.642	0.836	0.654	0.840

본 논문에서는 실험의 편이상 비학습 문서의 12주제의 Best-1을 이용하여 검색된 문서들을 분류하여 분류 그룹을 설정하였고, 각각의 분류 그룹에서 용어 클러스터를 생성하였다.

2.3 분류 그룹 설정 및 용어 클러스터 생성

분류 그룹을 설정하기 위하여 본 연구에서는 상위 n(30)개의 문서를 이용하였다. 상위 n개의 문서의 대하여 분류 그룹을 설정하고 용어 클러스터를 생성하는 과정은 다음과 같다. 검색된 상위 n개에 대하여 문서 분류기를 통하여 각각의 분류 정보를 획득한다. 획득된 분류 정보를 이용하여 같은 분류 정보를 가지는 문서들끼리 그룹핑을 하고, 분류된 각각의 그룹에서 그룹을 대표하는 용어를 추출하여 용어 클러스터를 생성하여 사용자에게 적합한 한 클러스터를 선택하도록 한다.

각각의 그룹에서 초기 질의를 확장하기 위한 용어들을 선택하기 위하여 적합성 피드백(Relevance Feedback)[7]을 이용하였다. 적합성 피드백은 크게 Rocchio방법, Ide Regular 방법, Ide dec-hi방법이 있다. 이 세 가지 방법의 기본 연산은 검색된 문헌 벡터와 초기 질의 벡터를 병합하는 것이다. Ide dec-hi 방법은 유일하게 사용자에게 보여진 첫번째 집합 내에서 검색되어진 비관련 문서 대신에 피드백에 대한 최상위 비관련 문서를 이용한다. Rocchio 방법은 관련 문서와 비관련 문서의 계수를 조정을 허락하여, 실제 문서 자체의 가중치보다는 문서 가중치의 정규화된 버전을 기본으로 하여 이용된다. [그림 2]는 각각의 방법에 대한 공식들이다.

이들 적합성 피드백 알고리즘을 적용하기 위하여 위해서는 분류된 그룹의 적합 정보 뿐만 아니라 부적합 정보도 알아야만 한다. 그러나 본 연구에서는 사용자의 빈번한 참여를 제한하기 위하여 시스템이 제공한 용어 클러스터 중에서 관련되는 하나의 클러스터만 선택하도록 하였기 때문에 부적합 정보를 얻을 수 없다. 따라서 본 연구에서는 분류된 각각의 그룹이 상호 배타적(mutual exclusive)이라고 가정하였다. 즉, 사용자가 자신의 검색 의도에 '적합하다'고 생각되는 한 그룹을 선택하면,

선택되지 않은 나머지 그룹들은 모두 사용자의 검색 의도에 '부적합하다'고 가정하였다.

Rocchio : $Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$

Ide Regular : $Q_1 = Q_0 + \sum_{i=1}^{n_1} R_i - \sum_{i=1}^{n_2} S_i$

Ide dec - hi : $Q_1 = Q_0 + \sum_{i=1}^{n_1} R_i - S_i$

Q_0 : 초기 질의에 대한 벡터
R_i : 관련 문서 i에 대한 벡터
S_i : 비관련 문서 i에 대한 벡터
n_1 : 관련 문서들의 수
n_2 : 비관련 문서들의 수

[그림 2] 적합성 피드백 방법

2.4 용어 클러스터 선택 및 검색

사용자는 검색 시스템이 제공한 용어 클러스터들로부터 하나의 용어 클러스터를 선택한다. 선택된 용어 클러스터에는 적합성 피드백 알고리즘을 이용하여 산출한 용어 가중치가 포함되어 있다. 이 용어 가중치는 초기 질의에서 제공하지 못했던 것이다. 또한 검색된 문서 전체에 대하여 적합/부적합 정보를 요구하는 것이 아니라 용어 클러스터들 안에서 하나의 클러스터를 선택하는 것이기 때문에 사용자의 적극적인 참여를 유도할 수 있다. 이는 시스템에 기반한 적합성 피드백의 단점인 초기 검색 결과의 의존성을 감소시킬 수 있다.

3. 용어 클러스터 질의 확장 모델의 처리 절차

본 장에서는 용어 클러스터 질의 확장 모델의 처리 절차를 ETRI-KEMONG SET의 46개 질의 중 40번째 질의를 이용하여 구체적으로 설명한다.

현재 용어 클러스터 질의 확장 모델은 stand-alone버전으로 구현되어 실험 중이기 때문에 인터페이스 상에게 개선되어져야 할 점들이 있다.

3.1 키워드 추출

ETRI-KEMONG SET의 40번째 질의는 "콜럼버스가 최초로 착륙한 곳은"이다. 이 질의에서 키워드를 추출하면 [콜럼버스] [최초] [착륙]의 3개 키워드가 추출된다.

3.2 1차 검색 결과

[표 2]에서는 1차 검색 결과의 상위 10개 문서의 타이틀과 분류 정보만을 보여주고 있지만, 실제 실험에서는 상위 30개의 문서를 이용하였다. (타이틀에 있는 *표는 실험 집합에서 제시한 관련 문서를 나타낸다.)

[표2] 1차 검색 결과의 상위 10개 문서

순위	타이틀	분류 정보
1	비행기	산업
2	인공위성	과학
3	우주여행	과학
4	베스푸치	지리
5	지리상의 발견*	지리
6	우주개발	과학
7	김포국제공항	산업
8	관제탑	산업
9	케네디우주센터	과학
10	콜럼버스*	지리

3.3 분류 그룹 설정 및 용어 클러스터 생성

1차 검색된 상위 30개 문서를 이용하여 각각의 분류 정보에 따라 분류 그룹을 나누고 각각의 분류 그룹을 대표하는 용어 클러스터를 생성한다. 용어 클러스터를 생성하기 위하여 적합성 피드백 알고리즘을 사용하였다. 그리고 아래 [표 3]는 적합성 피드백 알고리즘 중 Rocchio방법을 이용하여 각각의 분류 그룹에서 용어 클러스터를 생성한 결과이다. Rocchio방법에 사용된 β와 γ값은 각각 사용한 0.75와 0.25이다. 실제 사용자는 [표 3]의 분류 정보, 문서 비율, 용어 리스트 외에 검색 문서의 타이틀 정보를 참고하여 자신의 검색 요구를 만족하는 용어 클러스터를 선택하게 된다.

[표 3] 분류 그룹 설정 및 용어 클러스터 생성

순서	분류 정보	문서 비율	용어 클러스터
1	철학	3.33%	착륙/4.51876콜럼버스/4.15974방송/0라디오/0교육/0 KBS/0 간기/0
2	사회	6.67%	방송/5.06318착륙/4.51876콜럼버스/4.15974라디오/0.72603교육/0.658155KBS/0.590194간기/0.552569
3	과학	23.33%	착륙/4.35441콜럼버스/4.12529가스/0.444723대륙/0.425074개발/0.197292대기/0.184003계획/0.172801
4	생물	3.33%	콜럼버스/4.5519착륙/4.51876품종/1.83881만들기/1.70205줄기/1.65853먹기/1.61877옥수수/1.45038
5	산업	20%	착륙/4.42064콜럼버스/4.15974가스/0.922544로켓/0.506262가입/0.39892공기/0.298201고체/0.247425
6	지리	40%	착륙/4.51876콜럼버스/3.8253공업/0.12765개척/0.116071대륙/0.09839기후/0.0949593America/0.0821972
7	역사	3.33%	콜럼버스/4.84143착륙/4.51876여왕/2.50964아라곤/1.83277에스파냐/1.79755Isabel [illegible]/1.4063그라나다/1.33361

3.4 용어 클러스터 선택

사용자는 [표 3]에 제시된 다양한 정보를 보고 자신의 검색 의도에 일치하는 용어 클러스터를 선택할 수 있다. 그런데, 만약 사용자가 자신의 검색 의도를 모른다고 할 경우를 가정해 보자. 이 경우에도 사용자는 용어 클러스터 정보를 참고로 하여 자신의 요구를 구체화시킬 수 있다.

[표 3] 용어 클러스터 선택

Which category?(type id) : 6 You choose 6(지리)

3.5 검색어 조정

사용자는 제시된 용어 클러스터에서 자신의 검색 의도와 일치하는 하나의 클러스터를 선택한다. 선택된 용어 클러스터와 초기 질의를 비교해 보면 다음 [표 4]와 같다.

[표 4] 초기 질의와 최종 질의 비교

Original query : vsm=30:3:콜럼버스:최초:착륙: New query : Vsm=30:7:착륙/4.518758:콜럼버스/3.825295:공업/0.127650:개척/0.116071:대륙/0.098390:기후/0.094959:America/0.082197:

[표 4]에서 보는 것처럼 초기 질의에 비하여 사용자의 요구가 보다 구체화되어 있음을 알 수 있다. 또한 초기 질의에서 제공하지 못했던 용어 가중치가 부여되어 있음을 볼 수 있다.

현재 stand-alone버전에서는 검색어 조정 기능이 빠져 있다. 최종 질의를 구성하는 용어의 추가 또는 삭제, 용어 가중치 수정이 불가능하다. 그러나 인터페이스를 개선하여 사용자가 직접 용어를 추가하거나 삭제가 가능하도록 변경하고, 용어에 대한 가중치도 수정할 수 있도록 하여 검색어 조정 기능을 강화할 예정이다.

사용자의 검색어 조정이 끝난 후 시스템은 재검색을 실행한다.

3.6 재검색

사용자가 선택한 용어 클러스터를 이용하여 재검색한 결과는 [표 5]와 같다. (타이틀에 있는 *표는 실험 집합에서 제시한 관련 문서를 나타낸다.)

[표 5] 재검색 결과

순위	타이틀	분류 정보
1	북아메리카주*	지리
2	푸에르토리코	지리
3	유럽주	지리
4	도미니카연방	지리
5	도미니카공화국	지리
6	지리상의발견*	지리
7	옥수수	생물
8	우주개발	과학
9	비행기	산업
10	김포국제공항	산업

[표2]와 [표 5]를 비교하면 상위 10개 문서 안에서 관련 문서 수는 2개로 동일하지만 분류 정보에서 보여주는 것처럼 검색 된 문서들이 서로 구별됨을 알 수 있다. 그리고 검색된 관련 문서의 수에 따라 초기 질의와 확장된 질의의 검색 결과를 비교해 보면 [표 6]과 같다. 이 표에서는 총 관련문서 3개가 모두 상위 30개 문서 안에서 검색되고 있지만 그 시점이 서로 다름을 보여준다. (질의 40번 "콜럼버스가 최초로 착륙한 곳은"의 적합 문서의 타이틀은 '북아메리카주', '지리상의 발견', '콜럼버스'이다.)

[표 6] 질의 40번의 관련 문서 수 비교

관련문서수	1개	2개	3개
초기 질의	순위 5	순위 10	순위 17
질의 확장	순위 1	순위 6	순위 15

그리고 사용자가 용어 클러스터 선택 과정에서 '지리'를 선택하지 않고 '역사'를 선택했을 경우를 살펴보자. 초기 검색 결과에서 '역사'라는 그룹에 속한 문서는 1문서로 문서의 타이틀은 '이자벨1세'이고, 검색된 순위는 15이다. 실제 '역사' 분류를 선택하여 검색 결과를 살펴보면 최상위 문서로 '이자벨1세'가 검색됨을 볼 수 있다. 즉 사용자의 관심 분야나 찾고자 하는 성향이 다르더라도 시스템에 종속되지 않고 유연하게 대처하며 검색 성능(retrieval effectiveness)을 향상시킬 수 있다.

4. 실험 및 평가

4.1 평가 문서 집합

평가 집합은 ETRI-KEMONG SET[14]을 이용하였다. 이 평가 집합은 ㈜계몽사에서 편집 출판한 학생 백과 사전으로, 텍스트로 약 10M의 분량이며, 23,113개의 표제어와 이를 설명하는 내용으로 구성되어 잇다. 전체 문서의 평균 단어 수는 56개이다. 그리고 자연 언어 질의는 총 46개이고, 질의의 평균 길이는 3개의 단어이다. 아래 [표 7]은 평가 집합의 통계 정보이다.

[표 7] ETRI-KEMONG SET의 통계 정보

문서 수	23,113
질의 수	46
평균 문서 길이	56단어
평균 질의 길이	3 단어
평균 적합 문서 수	9

4.2 성능 평가

사용자의 초기 질의와 용어 클러스터 질의 확장 모델의 검색 결과는 [표 8]과 같다.

[표 8] 검색 성능 평가표

재현률	정확률				
	초기 질의	Tf*Idf	Rocchio	Ide	Ide Dec-hi
0.0	1.00	1.00	1.00	1.00	1.00
0.1	0.65	0.42	0.71	0.55	0.47
0.2	0.51	0.31	0.56	0.42	0.38
0.3	0.41	0.25	0.45	0.34	0.31
0.4	0.32	0.20	0.36	0.28	0.25
0.5	0.27	0.15	0.28	0.22	0.19
0.6	0.21	0.11	0.21	0.17	0.13
0.7	0.10	0.07	0.12	0.08	0.08
0.8	0.10	0.07	0.12	0.08	0.08
0.9	0.10	0.07	0.12	0.08	0.08
1.0	0.00	0.00	0.00	0.00	0.00
성능	*	-30.3%	+9.5%	-14.6%	-20.7%

[표 8]을 살펴볼 때 초기 질의에 비하여 Rocchio방법을 제외한 Tf*Idf, Ide-Regular, Ide Dec-hi 방법은 크게 성능이 하락됨을 볼 수 있다. 그 이유를 분석해 보면 다음과 같다.

평가 집합의 질의 5번(아침에 풀잎에 맺힌 물방울은 이슬인가 아니면 서리인가)을 예로 들어 설명한다. 질의 5번의 관련 문서는 총 3개로 각 문서의 타이틀은 "이슬", "서리", "이슬점"이다. 질의 5번에 대한 각 방법의 최종 질의와 관련 문서의 수는 [표 9]와 [표 10]과 같다.

[표 9] 질의 5번의 최종 질의 비교

방법	질 의
초기 질의	아침:풀잎:물방울:이슬:서리
Tf*Idf	기간/3.037520:공기/2.642276:액체/1.681248:기온/1.507776:방식/1.353879:서리/1.295035:진통/1.190115:
Rocchio	풀잎/4.886735:이슬/4.817478:서리/4.717742:물방울/4.179665:아침/3.886735:기간/1.058482:공기/0.975772:
Ide	풀잎/4.886735:아침/1.295578:기간/1.154902:이슬/1.096427:공기/1.060834:물방울/1.054840:서리/0.962049:
Ide Dec-hi	풀잎/4.886735:공기/3.494625:구름/3.287859:물방울/3.249951:수증기/2.373587:결정/1.796502:공기중/1.471193:

[표 10] 질의 5번의 관련 문서 수 비교

관련문서 수	1개	2개	3개
초기질의	순위1	순위2	순위9
Tf*Idf	X	X	X
Rocchio	순위1	순위2	순위13
Ide	순위1	순위12	X
Ide Dec-hi	X	X	X

[표 9]의 최종 질의들을 살펴 보았을 때, Tf*idf방법에 의해 만들어진 질의에서 '진통'이라는 용어를 제외하고는 직관상 거부감이 드는 용어는 없다(실제 Tf*idf에 의해 만들어진 질의를 살펴보면 직관상으로 거부감이 드는 용어가 포함되는 경우가 많다.). 그러나 이 질의들에 의하여 검색된 관련 문서 수[표 10]를 비교해 보면 상당한 차이가 있음을 발견할 수 있다. 즉, 적절하지 못한 용어 가중치는 오히려 성능을 저하시킴을 알 수 있다.

또한 각 방법에서 구해진 검색 결과 중 관련 문서 수가 0인 질의의 개수를 비교해 보면 [표 11]과 같다.

[표 11] 관련 문서 수가 0인 질의 개수 비교

	관련 문서 수가 0인 질의 개수
초기 질의	4개
Tf*Idf	15개
Rocchio	4개
Ide	6개
Ide Dec-hi	9개

[표 8]에 의하면 Rocchio방법에 의한 질의 확장 결과가 초기 질의에 비하여 9.5%의 검색 성능의 향상을 보여준다. 그러나 이러한 향상율은 다른 실험 결과와 비교하여 보았을 때 높은 향상 결과는 아니다[10]. 그 이유를 분석해 보면 실험에 사용한 분류기의 자체 오류 뿐만 아니라 문서 특성상 한 영역으로 규정짓기에는 애매한 문서가 존재하기 때문이다.

5. 결론

본 논문에서는 보다 유용한 질의를 만들기 위하여 사용자의 적합성 피드백을 이용하였다. 그러나 일반적인 적합성 피드백의 가장 큰 단점은 빈번한 사용자 참여이다. 이는 신속한 정보를 원하는 사용자에게 선택하기 힘든 요구 사항이다. 한편 시스템에 기반한 적합성 피드백 방법의 경우 성능은 초기 검색 결과에 밀접하게 의존적이다. 따라서 본 논문에서는 적합성 피드백의 빈번한 사용자 참여는 지양하고, 시스템에 기반한 적합성 피드백에서 적합성 피드백에서 배제한 사용자 참여를 유도하는 분류 정보에 기반한 용어 클러스터 질의 확장 모델(Term Cluster Query Expansion Model)을 제안하였다. 평가 집합을 가지고 실험한 결과, Rocchio, Ide-Regular, Ide Dec-hi의 3가지 방법 중에서 Rocchio방법을 이용하였을 때 초기 질의에 비하여 9.5%의 검색 성능(retrieval effectiveness)이 향상됨을 알 수 있었다. 이 결과는 다른 적합성 피드백 실험과 비교하여 보았을 때 높은 향상 결과는 아니다. 그 이유를 분석해 보면 실험에 사용한 분류기의 자체 오류 뿐만 아니라 문서 특성상 한 영역으로 규정짓기에는 애매한 문서가 존재하기 때문이다. 그러나 검색하고자 하는 사용자의 관심 분야나 찾고자 하는 성향이 다르더라도 시스템에 종속되지 않고 유연하게 대처하며 검색 성능을 향상시킬 수 있다. 차후 연구 과제는, 보다 정확한 용어 클러스터를 생성하기 위하여 분류기의 성능 향상을 고려하고 있다. 또한 클러스터링 알고리즘[1,6]에 의하여 생성된 용어 클러스터와의 비교 또한 연구 과제이다.

참고 문헌

[1] Chia-Hui Chang, Ching-Chi Hsu, "Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW," IEEE Transactions on Knoledge and Data Engineering, Vol. 11, No.4, pp.595-609, 1999.
[2] D. Harman, "Relevance Feedback Revisited," Proc. ACM SIGIR Int'l Conf. Researh and Devleopment in Information Retrieval, pp.1-10, 1992.
[3] G. Salton and M.J. McGill, "Introduction to Modern Information Retrieval," McGrawHill, 1983.
[4] Hyun-Kyu Kang, Key-Sun Choi, "Two-level Document Ranking Using Mutual Information in Natural Language Information Retrieval," Information Processing & Management, Vol. 33, No.3, pp.289-306, 1997.
[5] Mandar Mitra, Amit Singhal, Chris Buckley, "Improving Automatic Query Expansion," Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval, 1998.
[6] P.G. Anick and S. Vaithyanathan, "Exploiting Clustering and Phrases for Context-Based Information Retrieval," Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval, pp.314-323, 1997.
[7] W.B. Frakes and R. Baeza-Yates, "Information Retrieval : Data Structures and Algorithms," Prentice Hall, New Jersey, 1992.
[8] J. Xu and W.B. Croft, "Query Expansion Using Local and Global Document Analysis," Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval, 1996.
[9] 강원석, 강현규, 김영섬, "개념 기반 문서분류기 TAXON의 설계 및 구현", 1997년도 한국 정보 과학회 추계 학술발표 논문집(B), pp.197-200, 1997
[10] 박수현, 권혁철, "한국어 정보검색 시스템을 위한 다양한 적합성 피드백 방법의 실험", 정보과학회논문지(B) 제 26권 제5호, pp.682-691, 1999.
[11] 박수현, 박세진, 군혁철, "미리내 정보검색 시스템에서 Relevance Feedback 구현", 제 9회 한글 및 한국어 정보처리학술발표 논문집, pp.65-71, 1997.
[12] 윤보현, 백대호, 김상범, 한경수, 임해창, "대화적 질의 확장을 통해 의미적 용어불일치를 완화하는 정보검색 방안", 1999년도 한국 정보 과학회 봄 학술 발표 논문집(B), pp.345-347, 1999.
[13] 정영미, 정보검색론, 구미무역㈜출판부, 1993.
[14] 한국전자통신연구원, "ETRIKEMONG SET", 자연어처리 연구실, 한국전자통신연구원, 1997.