문서필터링을 위한 질의어 확장과 가중치 부여 기법

신 승 은[†]·강 유 환[†]·오 효 정[††]·장 명 길[††]
박 상 규[††]·이 재 성[†††]·서 영 훈[††††]

요 약

본 논문에서는 문서 필터링을 위한 질의어 확장과 가중치 부여 기법을 제안한다. 문서 필터링은 웹 검색 엔진들에 대한 검색 결과의 정확률 향상을 목적으로 한다. 문서 필터링을 위한 질의어 확장은 개념망, 백과사전, 유사도 상위 10% 문서를 이용하며, 각각의 확장 질의어에 가중치를 부여하여 질의어와 문서들간의 유사도를 계산한다. 첫 번째 단계에서 개념망과 백과사전을 이용하여 초기 질의어에 대한 1차 확장 질의어를 생성하고, 1차 확장 질의어에 가중치를 부여하여 질의어와 문서들간의 유사도를 계산한다. 다음 단계에서는 높은 유사도를 갖는 상위 10% 문서들을 이용하여 2차 확장 질의어를 생성하고, 2차 확장 질의어에 가중치를 부여하여 질의어와 문서들간의 유사도를 계산한다. 다음으로 1차 유사도와 2차 유사도를 결합하여 문서들을 재순위화하고, 임계치보다 낮은 유사도를 갖는 문서들을 필터링함으로써 웹 검색 엔진들의 검색 결과 정확률을 향상시킨다. 실험에서 이러한 문서 필터링을 위한 질의어 확장과 가중치 부여 기법은 정확률-재현율과 F-measure를 이용하여 성능 평가를 할 때 정보 검색 효율성에서 주목할 만한 성능 향상을 보였다.

Query Expansion and Term Weighting Method for Document Filtering

Seung-Eun Shin[†]·Yu-Hwan Kang[†]·Hyo-Jung Oh[††]·Myung-Gil Jang[††]
Sang-Kyu Park[††]·Jae Sung Lee[†††]·Young-Hoon Seo[††††]

ABSTRACT

In this paper, we propose a query expansion and term weighting method for document filtering to increase precision of the result of Web search engines. Query expansion for document filtering uses ConceptNet, encyclopedia and documents of top 10% high similarity. Term weighting method is used for calculation of query-documents similarity. In the first step, we expand an initial query into the first expanded query using ConceptNet and encyclopedia. And then we weight the first expanded query and calculate the first expanded query-documents similarity. Next, we create the second expanded query using documents of top 10% high similarity and calculate the second expanded query-documents similarity. We combine two similarities from the first and the second step. And then we re-rank the documents according to the combined similarities and filter off non-relevant documents with the lower similarity than the threshold. Our experiments showed that our document filtering method results in a notable improvement in the retrieval effectiveness when measured using both precision-recall and F-Measure.

키워드 : 문서필터링(Document Filtering), 질의어 확장(Query Expansion), 가중치 부여(Weighting Method)

1. 서 론

정보검색 기술은 1990년대 후반부터 인터넷의 발전과 더불어 상업적 응용이 확대되면서 급속히 발전하고 있다. 최근에는 웹 문서의 양이 급격히 증가하면서 세계적으로는 7억 페이지 이상을 색인하는 대용량 문서 색인 기술과 함께 수만에서 수십만의 검색 결과 중에서 사용자가 원하는 의도에 맞는 정보를 정확하게 찾아주는 효과적인 검색 랭킹 기술이 요구되고 있다. 특히 웹과 같은 영역에서의 정보검색은 다양한 분야의 정보들이 서로 연결되어 있는 상황에서 빠르고 정확하게 찾아주는 점에 초점을 맞추어 기술 개발이 집중적으로 이루어지고 있다[1].

최근에는 정보검색에 자연언어처리 기술을 적용하여 검색 효과를 높이려는 연구가 국내외적으로 활발히 이루어지고 있다. 이러한 자연언어처리 기술의 하나인 문서 필터링(document filtering)은 사용자의 관심분야를 지정해 놓고 연속적으로 입력되는 문서들에서 관심분야 문서만을 추출하는 것으로 정의할 수 있다[2]. 또한 이러한 문서 필터링을 확대 정의하면 특정 질의어와 관련 있는 문서들 중에서 관련도가 높은 문서를 추출하는 것으로 정의할 수 있으며

† 준 회 원 : 충북대학교 대학원 컴퓨터공학과
†† 정 회 원 : 한국전자통신연구원
††† 종신회원 : 충북대 컴퓨터교육과 교수
†††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수
논문접수 : 2003년 2월 24일, 심사완료 : 2003년 11월 15일

[1], 관련 범주(relevant)와 비관련 범주(not relevant)로 분류하는 이진 문서 분류(binary text classification)라 할 수 있다[3].

일반적으로 사람이 질의어와 문서와의 관련도를 측정할 때는 어떠한 문서가 사용자의 질의어에 포함된 단어들을 적게 포함하더라도 질의어와 관련된 단어들이 문서에 많이 포함되어 있다면, 그 문서는 사용자의 요구에 적합한 문서가 된다. 정보검색기법에서도 질의어확장, 은닉의미색인을 이용한 검색 등과 같이 질의어 뿐만 아니라 질의어가 내포하고 있는 의미를 고려하고자 하는 연구들이 많이 진행되고 있다[4, 5].

질의어 확장(query expansion)은 사용자가 제시한 질의어에 이와 관련된 단어들을 추가해서 문서를 검색함으로써 보다 연관된 문서들을 검색하고자 하는 것이다. 시소러스를 이용해서 질의어를 확장하거나 코퍼스에 나타나는 단어들의 형태에 따라 단어들의 상호관계를 분석해서 질의어를 확장하는 방법 등이 있다. 자동적인 질의어 확장 방법은 재현율은 높일 수 있으나 높은 순위의 문서들에서 정확률은 일반적으로 낮아지기 때문에 실용적이지 않다[6]. 이에 대한 보완기법으로 사용자의 적합성 피드백을 이용하여 새로운 질의어를 형성하는 방법은 검색된 문서를 기반으로 해서 사용자가 직접 관련 문서와 비관련 문서를 판단해야 하고 사용자의 판단의 질에 매우 종속적이다. 최근에는 사용자가 제시한 질의어에 대한 검색된 문서들을 분석해서 질의어를 자동으로 확장하는 방법 등이 연구되고 있다[7].

은닉의미색인(latent semantic indexing)은 벡터공간 검색 기법의 확장으로, 개념기반 검색 기법이다. 단어들 사이의 의존 관계가 문서와 질의의 표현에서 고려되고, 검색에서 이를 활용하기 때문에, 문서가 질의어에 포함된 단어를 포함하지 않더라도 관련있는 문서로 검색될 수 있다[4, 5, 8].

전통적인 웹 문서 검색 시스템들은 질의어에 대해 많은 양의 순위화된 문서 목록을 결과로 보여준다. 오늘날의 웹 문서 검색 엔진들(Naver, Empas, Yahoo 등)의 대부분은 이러한 시스템들의 전형이며, 따라서 매우 낮은 정확률을 보여준다. 순위화된 검색 결과 목록을 보여주는 웹 검색 엔진이 낮은 정확률을 보여주기 때문에, 사용자들은 원하는 정보를 정확하게 찾는 것이 어렵다[9]. 본 연구에서는 문서 필터링에 의한 웹 검색 엔진의 정확률 향상을 목적으로 하며, 문서 필터링을 위한 질의어 확장과 가중치 부여 기법을 제안한다. 먼저 초기 질의어를 개념망과 백과사전을 이용하여 1차 질의어 확장을 하고 웹 검색 엔진의 결과와 1차 확장 질의어의 유사도를 계산한 후, 높은 유사도를 갖는 상위 10% 문서들을 이용하여 2차 질의어 확장을 한다. 각각의 확장 질의어에 가중치를 부여하고, 각 단계에서 계산된 1차 유사도와 2차 유사도를 결합하여 임계치 이하의 유사도를 갖는 문서를 필터링한다. 이러한 문서 필터링을 통해 웹 검색 엔진들의 정확률과 정보 검색 효율성을 향상시킬 수 있다.

2. 1차 질의어 확장

개념망과 백과사전을 이용하여 1차 질의어 확장을 한다. 먼저 개념망을 이용하여 질의어 확장을 하고, 다음으로 백과사전을 이용하여 질의어 확장을 한다. 1차 질의어 확장은 두 단계에서 확장된 질의어들을 통합함으로써 이루어진다.

2.1 개념망을 이용한 질의어 확장

1차 질의어 확장의 첫 단계로, 개념망을 이용하여 초기 질의어를 확장한다. 개념망은 한국어 명사에 대해서 사전적인 의미의 상하관계를 정의한 한국어 명사 개념망을 말한다. 한국어 명사 개념망은 단어들의 관계를 설정한다는 점에서 기존의 시소러스와 유사하지만, 시소러스가 단어간의 관계에 대한 기준이 명확하지 않은 반면, 개념망은 사전의 뜻풀이를 중심으로 개념어들 간의 국어학적 의미관계를 연결하므로 명백한 차이가 있다[1]. 그리고, 단어의 의미관계를 표현하는 리소스로 잘 알려진 워드넷(WordNet)은 의미가 유사한 단어들의 집합(SynSet)간의 연결로써, 단어 하나하나의 개념관계를 표현하는 개념망과는 다르다. 개념망은 경제분야의 약 2만개의 사전 엔트리를 가지며, 1만 5천여 개의 상하관계와 1천여 개의 동의와 유의 관계를 포함하고 있다[1]. 본 연구에서는 초기 질의어에 대한 상하관계(상위어와 하위어)를 이용하여 질의어 확장을 한다.

〈표 1〉 개념망을 이용한 질의어 확장 예

초기 질의어	상위어	하위어	상위어를 이용한 초기 질의어 분할
품질관리	관 리	티큐시	품질, 관리
재 산 권	권 리	채권 무체재산권 지능권 산업재산권	재산, 권리
개념망을 이용한 질의어 확장 결과			
품질관리	티큐시, 품질, 관리		
재 산 권	채권, 무체재산권, 지능권, 산업재산권, 재산, 권리		

<표 1>은 개념망의 상하관계를 이용하여 초기 질의어를 확장한 예이다. 상위어는 개념망의 상하관계에서 초기 질의어보다 한 단계 높은 단어이며, 하위어는 한 단계 낮은 단어이다. 상위어는 초기 질의어를 분할하기 위해 사용되며, 또한 '재산권'에서 '권리'를 유추하는 것과 같이 축약된 형

1) 개념망의 상하관계는 국어학적 의미의 kind-of 관계를 주로 하며, 일부 부분-전체 및 is-a 관계를 사용한다. 그러나, 시소러스에서의 상하관계는 사전적인 의미가 아닌 필요에 따른 용례에 의한 관계가 혼재함으로써, 둘 이상의 층위를 넘어가면 의미관계가 성립하지 않는 경우가 많다.

태의 단어를 유추하여 초기 질의어를 확장하기 위해 사용된다. 하위어는 그 자체를 확장 질의어로 사용한다. 개념망을 이용한 확장 질의어는 <표 1>과 같이 초기 질의어, 하위어, 상위어에 의한 초기 질의어 분할로 구성한다.

2.2 백과사전을 이용한 질의어 확장

개념망을 이용하여 초기 질의어를 확장한 후, 백과사전을 이용하여 다시 초기 질의어를 확장한다. 어떤 단어에 대한 백과사전의 정의는 그 단어를 설명하기 위해 매우 관계 깊은 단어들로 구성되어있다. 따라서 질의어 확장을 위해 초기 질의어에 대한 백과사전 정의를 이용한다. 먼저 초기 질의어에 대한 백과사전 정의로부터 명사들을 추출하고, 그 중 개념망에 존재하는 명사들만을 선택한다. 이때, 개념망을 이용한 질의어 확장에 포함된 명사는 선택하지 않는다. 개념망에 존재하는 명사들만을 선택하는 이유는 질의어 확장을 위해 의미있는 단어들만을 선택하는 것이 필요하기 때문이다. 질의어 확장에 사용한 백과사전은 두산세계대백과 엔싸이버(DOOSAN EnCyber)이다.

<표 2>는 백과사전 정의를 이용하여 초기 질의어를 확장한 예이다.

〈표 2〉 백과사전을 이용한 질의어 확장 예

초 기 질의어	백과사전 정의	명사 추출 결과	질의어 확장 결과
품질관리	과학적 원리를 응용하여 제품품질의 유지 향상을 기하기 위한 관리	과학, 원리, 응용, 제품, 품질, 유지, 향상, 관리	과학 원리 제품
재 산 권	재산적 가치를 지니는 권리	재산, 가치, 권리	가치

2.3 1차 확장 질의어-문서 유사도 계산

1차 확장 질의어는 <표 1>에서의 개념망을 이용한 질의어 확장 결과와 <표 2>에서의 백과사전을 이용한 질의어 확장 결과를 통합한 질의어들이며, <표 3>은 1차 확장 질의어의 예이다. 1차 확장 질의어와 문서간의 유사도(S_{First})는 식 (1)과 같이 계산한다.

〈표 3〉 1차 확장 질의어 예

초기 질의어	1차 확장 질의어
품질관리	티큐시, 품질, 관리, 과학, 원리, 제품
재 산 권	채권, 무제재산권, 지능권, 산업재산권, 재산, 권리, 가치

$$S_{First} = \frac{\sum (Q_{tf} \times Q_w)}{\sqrt{N}} \quad (1)$$

여기에서 Q_{tf}는 질의어의 출현 빈도수(term frequency), Q_w는 질의어의 용어 가중치(term weight), N은 문서의 전체 명사 수를 나타낸다. 이 단계에서 계산된 유사도에 따라 모든 문서들을 순위화한다.

식 (1)의 일반적인 벡터 모델에서 코사인 유사도(cosine coefficient similarity)와의 차이점은 문서 벡터의 크기(|D|)를 사용하는 것 대신 문서의 전체 명사 수($\sqrt{N}$)를 사용하여 정규화를 한다는 것이다. 확장 질의어의 수가 많지 않고, 또한 확장 질의어가 전체 문서를 반영한다고 보기 어렵기 때문에 문서 벡터의 크기(|D|)를 이용한 정규화를 하지 않고, 전체 문서를 반영할 수 있는 문서 전체 명사 수($\sqrt{N}$)를 사용하여 정규화를 한다. 이것은 가장 좋은 실험결과를 보인 루트 정규화를 적용한 것이다. 식 (1)의 사용은 일반적인 벡터 공간 모델에서의 코사인 유사도보다 F-measure (+0.0583)의 성능 향상을 보인다[10].

3. 2차 질의어 확장

3.1 높은 유사도를 갖는 문서를 이용한 질의어 확장

2차 질의어 확장을 위해 1차 확장 질의어와 문서간의 유사도를 이용한다. 계산된 유사도가 높은 상위 10% 문서를 문서 필터링의 학습문서로 사용한다. 다시 말해, 상위 10% 문서를 초기 질의어에 대한 관련 문서로 판단하여, 이를 이용하여 질의어를 확장한다. 먼저, 상위 10% 문서로부터 명사들을 추출하고, 해당 명사를 포함하고 있는 문서의 수(document frequency)가 임계치보다 큰 명사만을 선택한다. 많은 문서에서 출현하는 명사만을 선택하는 이유는 상위 10% 문서를 초기 질의어에 대한 관련 문서로 판단하기 때문이며, 명사의 출현 빈도수(term frequency)를 고려하지 않는 이유는 하나의 관련 문서에 편중되어 나타나는 명사는 상위 10% 문서 전체와 관련있는 명사가 아니기 때문이다. 즉, 2차 질의어 확장에서 초기 질의어와 관련있고, 의미있는 질의어를 선택하기 위함이다.

〈표 4〉 2차 확장 질의어 예

초기 질의어	유사도 상위 10% 문서에서 추출한 명사	2차 확장 질의어
품질관리	가격, 고객, 공사, 계획, 도구 산업, 방법, 인증, 시험, 전기 품질, 관리, …	계획, 방법 산업, 시험
재 산 권	가격, 감시, 강화, 기술, 분야 산업, 사건, 소개, 위원, 상표 저작권, 특허, …	기술, 분야 산업, 상표 저작권, 특허

<표 4>는 유사도 상위 10% 문서들을 이용하여 질의어를 확장한 예이다. <표 3>에서 2차 확장 질의어는 임계치 0.9에서 선택한 확장 질의어들이다. 임계치는 상위 10% 문서 중 해당 명사가 나타난 문서의 비율을 의미한다.

3.2 2차 확장 질의어-문서 유사도 계산

2차 확장 질의어는 <표 4>에서와 같이 유사도 상위 10% 문서의 명사들 중 문서 출현 비율이 임계치 이상인 명사만을 선택하여 확장한다. 2차 확장 질의어와 문서간의 유사도(S_{Second})는 식 (2)와 같이 계산한다.

$$S_{Second} = \frac{\sum (Q_{tf} \times Q_w)}{\sqrt{N}} \quad (2)$$

식 (1)과 같이 Q_{tf}는 질의어의 출현 빈도수(term frequency), Q_w는 질의어의 용어 가중치(term weight), N은 문서의 전체 명사 수를 나타낸다.

4. 확장 질의어 가중치 부여와 유사도 결합

4.1 확장 질의어 가중치 부여

Rocchio에 의해 개발된 피드백 질의 생성 알고리즘(feedback query creation algorithm)은 가장 성능이 뛰어난 렐러번스 피드백 알고리즘들(relevance feedback algorithms) 중의 하나로 증명되었다. 식 (1)과 식 (2)에서 Q_w는 확장 질의어의 용어 가중치를 의미하며, 확장 질의어의 가중치는 Rocchio의 최적 질의(Rocchio's optimal query)를 만들기 위한 Rocchio 공식을 이용하여 구한다. Rocchio 공식은 식 (3)과 같다[3, 11-13].

$$\vec{Q} = \alpha \vec{Q}_{init} + \beta \frac{1}{R} \sum_{d \in Rel} \vec{Q}_d - \gamma \frac{1}{N-R} \sum_{d \notin Rel} \vec{Q}_d \quad (3)$$

식(3)에서 $\beta = \gamma$라는 것이 적합하다는 것이 증명[14]되었기 때문에 식 (3)을 식 (4)로 변환하여 사용한다.

$$\vec{Q} = \alpha \vec{Q}_{in} + \frac{1}{R} \sum_{d \in Rel} \vec{Q}_d - \frac{1}{N-R} \sum_{d \notin Rel} \vec{Q}_d \quad (4)$$

식 (4)에서 Q_d는 문서 d에서 확장 질의어의 가중치 벡터이며, R은 관련 문서의 수, 그리고 N은 전체 문서의 수이다. 실험 문서의 초기 질의어별 확장 질의어를 분류하여 각각의 가중치를 계산한다. 확장 질의어 벡터는(q_1, q_2, q_3, q_4, q_5)이며, q_1은 초기 질의어, q_2는 하위어, q_3는 초기 질의어 분할, q_4는 백과사전을 이용한 확장 질의어, q_5는 유사도 상위 10% 문서를 이용한 확장 질의어이다. 확장 질의어를 q_1-q_5까지 세분화하고 식 (4)에서 $\alpha = 0$로 설정한 후, 이를 이용하여 각각의 확장 질의어에 대한 가중치를 계산한다. 먼저 확장 질의어의 가중치를 구하고, 질의어-문서간의 유사도를 계산할 때는 식 (4)와 같이 초기 질의어(Q_{init})와 확장 질의어를 결합하여 유사도를 계산한다. 식 (4)에서 α의 의미는 초기 질의어와 확장 질의어의 비중을 선택할 수 있게 하는 변수로써, 실험에서 α의 값은 1, 2, 3, 4, 5, 6을 사용한다.

확장 질의어 가중치 부여는 먼저 2차 질의어 확장을 위한 유사도 상위 10% 문서를 구하기 위해 1차 확장 질의어(q_5를 제외한 확장 질의어들)에 대해 이루어진다. 1차 확장 질의어와 초기 질의어를 식 (4)에 따라 결합하여 질의어-문서간의 1차 유사도를 구하고, 유사도 상위 10% 문서를 이용하여 2차 질의어 확장을 한다. 확장된 2차 질의어에 대해 다시 가중치를 구하고, 질의어-문서간의 2차 유사도를 계산한다. 이렇게 구해진 1차 유사도와 2차 유사도를 결합하여 질의어-문서간의 최종 유사도를 구한다.

실험 문서 집합에 대해 위와 같이 초기 질의어와 확장 질의어의 가중치를 구하고, 이를 문서 필터링 시스템의 초기 질의어와 확장 질의어의 가중치로 사용한다.

4.2 두 유사도의 결합

이제까지 개념망과 백과사전을 이용한 1차 질의어 확장과 유사도 상위 10% 문서를 이용한 2차 질의어 확장을 했다. 이 과정에서 각각의 확장 질의어에 가중치를 부여하여 계산된 2개의 유사도(S_{First}, S_{Second}) 값을 식 (5)와 같이 결합한다.

$$S_{combined} = \beta S_{First} + (1-\beta) S_{Second} \quad (5)$$

여기서 β는 각 단계에서 계산된 유사도의 비중을 조절하는 의미를 가진다. β를 이용해 1차 유사도와 2차 유사도의 비중을 조절하는 이유는 2차 질의어 확장을 위해 사용한 상위 10% 문서의 정확률이 약 84% 정도이기 때문이다. 이렇게 결합된 유사도와 Accept Point(관련 문서와 비관련 문서를 구분하는 유사도 값)와 비교하여 작은 유사도를 갖는 문서들을 필터링하여 웹 검색 엔진 결과의 정확률과 검색 효율성을 향상시킨다.

5. 실험 및 평가

제안된 문서 필터링의 성능을 평가하기 위해 5개의 검색 엔진들(Naver, Empas, Yahoo, Altavista, Google)의 검색 결과를 이용하였다. <표 5>는 실험 문서 집합에 대한 통계 정보이다.

〈표 5〉 실험 문서 집합의 통계정보

	실험 문서 집합 A	실험 문서 집합 B
초기 질의어 수	27	2
전체 문서 수	2510	181
검색 결과의 정확률	0.3750	0.3267

실험 문서 집합 A는 경제 분야의 초기 질의어에 대한 검

색 엔진들의 결과이며, 중복된 문서는 제거된 문서 집합이다. 실험 문서 집합 A가 경제 분야 문서인 이유는 현재 개념망이 경제 분야에 한해서 구축되었기 때문이며, 실험 문서 집합 B는 경제분야가 아닌 일반 영역의 문서집합이다. 실험 문서 집합 B에 대한 실험은 본 연구가 경제분야뿐만 아니라 일반 영역으로의 확장 가능성을 의미한다.

문서 필터링에 대한 평가 방법은 일반적으로 검색 시스템의 성능을 평가하는 기준인 정확률, 재현율과 F-measure를 이용하였다. 정확률은 문서 필터링 결과의 전체 문서 중 질의어 관련 문서의 비율을 의미하고, 재현율은 전체 관련 문서에 대한 문서 필터링 결과의 질의어 관련 문서의 비율을 의미한다. 또한 F-measure는 정확률과 재현율 모두를 이용하여 시스템의 성능을 하나의 척도로 평가하는 방법이다. F-measure를 계산하는 방법은 식 (6)과 같다.

$$F = \frac{(\gamma^2+1)PR}{\gamma^2 P+R} \qquad (6)$$

여기서 γ의 의미는 정확률(P)과 재현율(R)의 비중을 선택할 수 있게 하는 변수로써 $\gamma > 1$이면 정확률의 비중을, $\gamma < 1$이면 재현율의 비중을 높게 둔다는 의미이다. 일반적으로 기술적인 성능평가를 위한 γ값으로 1, 2, 5를 사용한다[14].

5.1 초기 질의어 비중(α) 실험

1차 확장 질의어에 대해 가중치를 부여하여, 초기 질의어의 비중을 조절하며, 실험 문서 A에 대하여 유사도 실험을 하였다. <표 6>은 초기 질의어 비중(α)에 따른 실험의 결과로, F-Measure값은 식 (6)에서 γ를 1로 하여 정확률과 재현율의 비중을 같게 하였을 때의 값이다.

q1 : 초기 질의어 q3 : 초기 질의어 분할
q2 : 하위어 q4 : 백과사전을 이용한 확장 질의어

〈표 6〉 초기 질의어 비중(α)에 따른 F-Measure

가중치 / 초기질의어 비중(α)	q1	q2	q3	q4
	0.126464	0	0.076225	0.01756
1	0.668841			
2	0.682556			
3	0.702845			
4	0.714398			
5	0.704055			
6	0.693402			

<표 6>의 실험결과에서 가장 높은 F-Measure값을 갖는 α값 4를 이용하여 1차 유사도를 계산한다. 이렇게 계산된 1차 유사도는 2차 질의어 확장을 위한 유사도 상위 10% 문서 선택을 위해 사용되며, 최종 유사도를 구하기 위해 사용된다.

5.2 유사도 비중(β) 실험

1차 유사도와 2차 유사도 결합 과정에서 1차 유사도의 비중(β)을 조절하며, 실험 문서 A에 대하여 실험하였다. 그 결과는 <표 7>과 같다.

〈표 7〉 유사도 비중(β)에 따른 F-Measure

가중치 / 유사도 비중(β)	q5 : 유사도 상위 10% 문서를 이용한 확장 질의어
	0.029902
1/2	0.70419
2/3	0.71646
3/4	0.71576
4/5	0.71576
5/6	0.71479

<표 7>의 결과에서 가장 높은 F-Measure값을 갖는 β값 2를 문서 필터링 시스템의 β값으로 정한다. β의 값을 조절하는 이유는 q_5(유사도 상위 10% 문서를 이용한 확장 질의어)를 확장하기 위해 사용한 유사도 상위 10% 문서의 정확률이 약 84%이기 때문이다.

5.3 Accept Point에 따른 문서 필터링 실험

앞의 실험에서 정해진 α와 β를 이용해 문서 필터링 시스템을 구현하고, 실험 문서 A에 대해 Accept Point에 따른 문서 필터링 시스템의 성능을 실험하였다. 그 결과는 <표 8>과 같다.

〈표 8〉 Accept Point에 따른 문서 필터링 시스템의 성능

가중치 & 비중 / Accept Point	가중치 q1 = 0.126464 q2 = 0 q3 = 0.076225 q4 = 0.01756 q5 = 0.029902 비중 α = 4, β = 2		
	정확률	재현율	F
Break Even Point	0.7165	0.7165	0.7165
0.2	0.5042	0.8763	0.6401
0.3	0.6245	0.7803	0.6937
0.4	0.7136	0.6772	0.6949
0.5	0.7715	0.5798	0.6621
0.6	0.8131	0.4957	0.6159

Accept Point는 문서 필터링 시스템에서 문서를 제거하는 유사도 기준 값을 의미하고, Break Even Point는 정확

률과 재현율이 같은 점을 의미한다. 문서 필터링 시스템에서 Accept Point의 값이 증가함에 따라 정확률은 증가하고, 재현율은 감소한다. <표 9>는 실험 문서 A에서의 11-Point 재현율에 대한 정확률 실험의 결과이며, <표 10>은 웹 검색 엔진 결과의 정확률 향상을 위해 제안된 질의어 확장 기법과 가중치 부여 기법을 이용한 문서 필터링의 성능 평가를 나타낸 표이다.

〈표 9〉 11-Point 재현율에 대한 정확률

가중치 & 비중 / 재현율	가중치 q1 = 0.126464 q2 = 0 q3 = 0.076225 q4 = 0.01756 q5 = 0.029902 비중 α = 4, β = 2
	정확률
0	0.8907
0.1	0.8913
0.2	0.8905
0.3	0.8752
0.4	0.8542
0.5	0.8121
0.6	0.7595
0.7	0.6946
0.8	0.5998
0.9	0.4818
1.0	0.3750
Average	0.7387

〈표 10〉 문서 필터링의 성능

문서 필터링	실험 문서 A			실험 문서 B		
	정확률	재현율	F	정확률	재현율	F
전	0.3750	1	0.5455	0.3267	1	0.4925
후	0.7165	0.7165	0.7165	0.6259	0.6259	0.6259
비고	+0.3415	-0.2835	+0.1710	+0.2992	-0.3741	+0.1334

문서 필터링 전의 정확률과 재현율은 웹 검색 엔진의 정확률과 재현율을 의미한다. 본 실험에서는 웹 검색 엔진 결과의 재현율은 1로 설정하고 실험하였다. 실험 문서 A는 학습 문서 집합(training corpus)이며, 실험 문서 B는 테스트 문서 집합(test corpus)이다. 먼저 실험 문서 A에서 F-measure에서 +0.1710(정확률 : +0.3415, 재현율 : -0.2835)의 성능 향상을 보였다. 이러한 질의어 확장 기법과 가중치 부여 기법을 실험 문서 B에 적용한 결과 F-measure에서 +0.1334(정확률 : +0.2992, 재현율 : -0.3741)의 성능 향상을 보였다. 문서 필터링 후 실험 문서 A보다 실험 문서 B에서 향상된 F-measure가 작은 이유는 2차 확장을 위한 상위 10% 문서의 정확률이 약 79%로 실험 문서 A보다 약 5% 정도 낮으며, 실험 문서 집합 B의 초기 질의어의 수가 2개 이므로 초기 질의어의 특성에 민감하기 때문이다. 또 다른 이유는 현재 개념망의 구축이 경제분야에서 이루어졌기 때문에 경제분야 이외의 초기 질의어를 포함하는 실험 문서 B에서 비교적 낮은 성능 향상을 보였다.

〈표 11〉 질의어 확장 단계별 F-measure 분석표

	메타검색 결과	개념망 이용	백과사전이용	상위 10% 문서 이용
F-measure	0.5455	0.6903	0.7119	0.7165
비 고		+0.1448	+0.0216	+0.0046

<표 11>은 실험 문서 A에서의 질의어 확장 방법에 따른 각 단계별 F-measure 분석표이다. 백과사전이용 단계는 개념망을 이용한 질의확장 다음 단계로 개념망과 백과사전을 이용해 확장 질의어를 생성한 단계이다. 개념망을 이용한 단계에서 F-measure가 크게 증가됨을 볼 수 있으며, 상위 10% 문서를 이용한 단계에서는 F-measure의 향상 정도가 크게 떨어짐을 볼 수 있다. 이러한 결과는 웹 검색엔진과 같이 빠른 서비스 시간을 요구하는 시스템에서는 2차 질의어 확장(상위 10% 문서 이용)을 제외한 1차 질의어 확장만을 적용하여 성능 향상을 보일 수 있음을 의미한다.

6. 결 론

본 논문에서는 웹 검색 엔진 결과의 정확률 향상을 위한 문서 필터링을 위해 질의어 확장 기법과 가중치 부여 기법을 제안하였다. 문서 필터링을 위한 질의어 확장은 개념망, 백과사전, 유사도 상위 10% 문서를 이용하며, 각각의 확장 질의어에 가중치를 부여하여 질의어와 문서들간의 유사도를 계산하여 문서를 필터링한다. 이 과정에서 초기 질의어의 비중 α와 1차 유사도 비중 β를 최적화하여 설정한다.

실험을 통해 제안한 문서 필터링의 결과는 웹 검색 엔진 결과보다 정확률에서 평균 +0.3204, F-Measure에서 평균 +0.1522의 성능 향상을 보였다. 이것은 문서 필터링을 위해 제안한 질의어 확장 기법과 가중치 부여 기법이 웹 검색 엔진들의 정보 검색 효율성에서 주목할 만한 성능 향상을 보임을 의미한다.

참 고 문 헌

[1] 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정, "의미

기반 정보검색", 정보과학회지, 제19권 제10호, pp.7-18, 2001.
[2] David A. Hull, Stephen Robertson, "The TREC-8 Filtering Track Final Report," The Eighth Text REtrieval Conference(TREC-8), pp.35-56, 2000.
[3] Robert Schapire, Yoram Singer, Amit Singhal, "Boosting and Rocchio Applied to Text Filtering," In Proc. 21'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.215-223, 1998.
[4] 이경순, 박영찬, 최기선, "문서 클러스터를 이용한 재순위화 모델", 제10회 한글 및 한국어정보처리학회, pp.81-87, 1998.
[5] Kyung-Soon Lee, Young-Chan Park and Key-Sun Choi, "Re-ranking model based on document clusters," Information Processing and Management, 37, pp.1-14, 2001.
[6] Larry Fitzpatrick and Mei Dent, "Automatic Feedback Using Past Queries : Social Searching?," In Proc. 20'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.306-313, 1997.
[7] Chris Buckley and Gerard Salton and J. Allan, "The effect of adding relevance information in a relevance feedback environment," In Proc. 17'th ACM SIGIR International Conference on Research and Development in Inforamtion Retrieval, pp.292-298, 1994.
[8] Scott Deerwester and Susan T. Dumais and Richard Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, 41(6), pp. 391-407, 1990.
[9] Oren Zamir and Oren Etzioni, "Web Document Clustering : A Feasibility Demonstration," In Proc. 21'th annual international ACM SIGIR conference on Research and development in information retrieval, pp.46-54, 1998.
[10] 김영택, "자연언어처리", 생능출판사, 2001.
[11] J. J. Rocchio, "Document Retrieval Systems- Optimization and Evaluation," PhD thesis, Harvard Computational Laboratory, Cambridge, MA, 1966.
[12] J. J. Rocchio, "Relevance feedback in information retrieval," In The SMART Retrieval System-Experiments in Automatic Document Processing, Prentice Hall, pp.313-323, 1971.
[13] Gerard Salton and Chris Buckley, "Improving retrieval performance by relevance feedback," Journal of the American Society for Information Science, 41(4), pp.288-297, 1990.
[14] Amit Singhal, Mandar Mitra and Chris Buckley, "Learning routing queries in a query zone," In Proc. 20'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.25-32, 1997.
[15] E. Hovy and C. Y. Lin, "Automated Text Summarization in SUMARIST," Proc. of a Workshop on Intelligent Scalable Text Summarization, pp.18-24, 1997.

신 승 은

e-mail : seshin@dcenlp.chungbuk.ac.kr
1999년 충북대학교 컴퓨터공학과(학사)
2001년 충북대학교 컴퓨터공학과(공학석사)
현재 충북대학교 컴퓨터공학과 박사과정
관심분야 : 정보검색, 자연언어처리, 인공지능 등

강 유 환

e-mail : eric@nlp.chungbuk.ac.kr
1998년 충북대학교 컴퓨터공학과(학사)
2000년 충북대학교 컴퓨터공학과(공학석사)
현재 충북대학교 컴퓨터공학과 박사과정
관심분야 : 자연언어처리, 구문분석 등

오 효 정

e-mail : ohj@etri.re.kr
1998년 충남대학교 컴퓨터과학과(학사)
2000년 충남대학교 컴퓨터과학과(석사)
2000년~현재 한국전자통신연구원 휴먼정보검색연구팀 연구원
관심분야 : 문서자동분류, 정보검색, 자연어처리, 기계학습

장 명 길

e-mail : mgjang@etri.re.kr
1988년 부산대학교 계산통계학과(학사)
1990년 부산대학교 계산통계학과(석사)
2000년 충남대학교 컴퓨터과학과(박사)
1990년~1998년 시스템공학연구소 선임연구원
1998년~현재 한국전자통신연구원 휴먼정보검색연구팀 팀장
관심분야 : 자연어처리, 정보검색, 생물정보학

박 상 규

e-mail : parksk@etri.re.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 KAIST 전산학과(공학석사)
1997년 KAIST 전산학과(공학박사)
1984년~1987년 대림산업 전산실
1987년~현재 ETRI 책임연구원
관심분야 : 언어처리, 자동번역, 정보검색, HCI, 지능형 에이전트

이 재 성

e-mail : jasonl@cbu.ac.kr

1983년 서울대학교 컴퓨터공학과(학사)

1985년 한국과학기술원 전산학과(석사)

1999년 한국과학기술원 전산학과(박사)

1985년~1988년 큐닉스컴퓨터 개발부 과장

1988년~1989년 미국 마이크로소프트 S/W 설계자

1988년~1993년 마이크로소프트 개발부 차장

1999년~2000년 한국전자통신연구소 선임연구원/팀장

2000년~현재 충북대 컴퓨터교육과 조교수

주관심분야 : 정보검색, 자연언어처리, 컴퓨터교육

서 영 훈

e-mail : yhseo@chungbuk.ac.kr

1983년 서울대학교 컴퓨터공학과(학사)

1985년 서울대학교 컴퓨터공학과(공학 석사)

1991년 서울대학교 컴퓨터공학과(공학 박사)

1994년~1995년 미국 Carnegie-Mellon 대학 기계번역 센타 객원교수

1988년~현재 충북대학교 전기전자컴퓨터공학부 교수

관심분야 : 정보검색, 자연언어처리, 음성언어처리, 기계번역 등