DOI: 10.3745/KIPSTB.2009.16-B.3.247

중첩 클러스터를 이용한 피드백 문서의 재샘플링 기법

이 경 순†

요 약

대부분의 잠정적 적합피드백기법들은 질의에 대해 검색된 상위검색문서들이 적합하다고 가정하고, 그 문서들을 질의 확장을 위한 피드백 문서로 이용하고 있다. 그러나 초기검색결과에는 상당한 양의 부적합 문서를 포함하고 있는 것이 현실이다. 이 논문에서는 보다 좋은 피드백 문서를 선택하기 위해서 중첩클러스터를 이용한 피드백문서의 재샘플링 기법을 제안한다. 주요 아이디어는 질의 중심적인 초기검색문서집합에 대해서 중첩이 허용된 문서클러스터를 이용하여 문서들 사이의 관계를 반영하여 질의에 핵심역할을 하는 지배적 문서를 찾고, 이 문서들을 반복적으로 피드백 하여 질의가 내포하는 핵심 주제를 강조하는 것이다. 대규모 실험집합인 TREC GOV2와 WT10g에 대한 실험비교에서, 최근 잠정적 적합피드백 기법들 중에서 가장 좋은 성능을 보이고 있는 적합모델보다 재샘플링기법이 우수한 성능향상을 보였다. 제안기법에 대한 검증을 위해서 피드백문서에 포함된 적합문서의 정도를 나타내는 적합밀도를 측정하였다. 재샘플링 기법이 TREC 실험집합에 대해서 적합모델에 비해 높은 적합밀도를 보였고, 이 결과 적합피드백에서 검색성능을 향상시키게 되었다. 이는 제안 기법이 잠정적 적합피드백에서 유효한 방법임을 알 수 있다.

키워드 : 정보검색, 잠정적 적합피드백, 재샘플링, 중첩 클러스터링, 지배적 문서, 질의 확장, 언어기반 검색모델, 적합모델

Resampling Feedback Documents Using Overlapping Clusters

Kyung Soon Lee†

ABSTRACT

Typical pseudo-relevance feedback methods assume the top-retrieved documents are relevant and use these pseudo-relevant documents to expand terms. The initial retrieval set can, however, contain a great deal of noise. In this paper, we present a cluster-based resampling method to select better pseudo-relevant documents based on the relevance model. The main idea is to use document clusters to find dominant documents for the initial retrieval set, and to repeatedly feed the documents to emphasize the core topics of a query. Experimental results on large-scale web TREC collections show significant improvements over the relevance model. For justification of the resampling approach, we examine relevance density of feedback documents. The resampling approach shows higher relevance density than the baseline relevance model on all collections, resulting in better retrieval accuracy in pseudo-relevance feedback. This result indicates that the proposed method is effective for pseudo-relevance feedback.

Keywords : Information Retrieval, Pseudo-Relevance Feedback, Resampling, Overlapping Clustering, Dominant Document, Query Expansion, Language-Based Retrieval Model, Relevance Model

1. 서 론

정보검색에서 검색된 결과를 이용해서 질의를 확장하거나 질의 가중치를 조정하는 잠정적 적합 피드백(Pseudo-Relevance Feedback) 기법은 원래 질의를 사용한 검색결과 보다 성능 향상을 보여오고 있다. 대부분의 잠정적 적합피드백 기법[15, 23, 22]은 원래 질의에 대해 검색된 상위 검색문서집합이 잠정적으로 질의에 대해 적합하다고 가정하고, 그 문서들을 이용해서 질의를 확장하거나 원래 질의의 가중치를 조정한다. 이는 적합피드백에서 실제로 적합한 문서를 사용하는 것과 비슷한 처리 과정이다[26].

그러나 일반적으로 상위검색결과는 부적합한 문서들을 포함하고 있다. 검색결과 상위 10개의 문서에 대한 정확률이 0.5일 때 (P@10), 5개의 문서는 질의에 적합한 문서이지만, 5개의 문서는 부적합한 문서들이다. 이는 현재의 정보검색 성능 수준이고, 심지어 현재의 모든 정보검색 모델에서 공통적으로 나타나고 있는 행태이다. 이러한 부적합문서들이 상당수 포함된 문서들로부터의 질의 확장은 원래 질의로부터 질의 표현을 동떨어지게 만드는 결과를 초래하게 된다.

본 논문은 잠정적 적합피드백을 위해 적합한 문서들을 더 많이 포함시키기 위해 클러스터를 바탕으로 한 재샘플링 기

※ 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-611- D00025).
† 정 회 원 : 전북대학교 전기전자컴퓨터공학부/영상정보신기술연구센터 조교수
논문접수 : 2008년 8월 8일
수 정 일 : 1차 2008년 10월 20일
심사완료 : 2008년 10월 21일

법을 제안한다. 특히 웹 문서와 같은 대량의 문서집합에 대한 초기 검색집합은 다양한 세부주제를 포함할 수 있기 때문에, 초기 검색결과집합에 대해 생성된 문서 클러스터는 질의에 대한 문서의 행태를 잘 표현할 수 있다. 초기 검색집합에 대한 클러스터링에서 적합 문서들로만 구성된 하나의 최적의 클러스터를 생성할 수 있다면 그것을 이용해서 피드백을 하면 이상적일 것이나, 하나의 최적의 클러스터를 찾는 것은 클러스터링에서 여전히 어려운 문제로 남아있다. 따라서 본 연구에서는 생성된 클러스터들 중에서 최상위로 선택된 여러 클러스터들을 적합 클러스터로 가정하고 피드백을 위해 사용한다. 이때 클러스터들이 상위 검색문서들에 대해서 같은 문서를 중복해서 포함할 수 있도록 허용한다. 클러스터 생성시 지배적 역할을 하는 문서(dominant document)는 여러 클러스터에 반복적으로 나타나고 질의와 클러스터의 관련도에 따른 클러스터 순위에서도 높은 우선순위로 나타난다고 본다. 이러한 클러스터 형성에서 지배적 역할을 하는 문서를 반복해서 피드백 함으로써 확장된 질의는 원래 질의에 대한 핵심주제를 강조할 수 있을 것으로 기대한다. 클러스터에 중복으로 나타나는 문서를 반복해서 이용하는 것이 재샘플링 효과이다.

적합피드백 연구에서 클러스터링을 이용한 연구는 이 논문이 처음은 아니다. 사실 클러스터링은 잠정적 적합피드백에 관한 초기 연구 [1]에서 언급이 되었으나, 초기의 클러스터를 이용한 시도는 성능향상을 보이지 못했다. 본 논문에서는 클러스터를 이용하되 지배적 문서를 재샘플링 함으로써 검색 성능을 상당히 개선시킨 새로운 연구로써 의미가 있다.

본 연구에서 클러스터와 재샘플링을 이용한 동기는 다음과 같다: (1) 초기 검색문서집합은 질의 중심적인 순위화를 나타내는 것이지 문서와 문서들 사이의 관계를 고려하지는 못한다. (2) 잠정적 적합피드백을 위한 질의 확장 문제는 학습예제들에 의존해서 질의와 밀접한 관련이 있는 어휘를 학습하는 문제로, 이는 학습예제들에 의존해서 정확한 결정경계를 학습하는 분류(classification) 문제와 비슷하다고 본다. 따라서 본 연구에서는 이 문제를 기계학습연구의 부스팅 (Boosting) 기법 [27, 10]에서 학습하기 어려운 예제들을 반복적으로 학습예제로 선택함으로써 약학습자(weak learner)를 학습이 어려운 예제들 쪽으로 결정경계(decision boundary)를 바꾸도록 학습을 시키는 것에 착안하여, 질의 확장을 위해서 초기 검색집합의 지배적인 문서들을 반복적으로 선택해서 반영함으로써 지배적인 문서들 쪽으로 질의 확장을 할 수 있도록 접근한다. 문서들 사이의 관계를 반영하기 위해 클러스터링을 하면서 중복 문서클러스터를 허용한 것에 대한 기본 가정은 질의를 잘 표현하고 있는 문서는 질의와 관련이 높은 여러 개의 가까운 이웃들을 가질 것이고, 또한 세부주제를 표현하고 있어서 여러 클러스터를 형성하는데 중심 역할을 하는 문서들이 있다고 가정한 것이다. 반복적으로 그러한 지배적 문서들을 피드백 함으로써 질의가 표현하고자 하는 핵심 주제들을 강조할 수 있을 것이다.

다양한 TREC 컬렉션들에 대한 실험을 통해서 본 논문에서 제안한 클러스터를 이용한 재샘플링 기법이 피드백을 위한 문서들에 대해 높은 적합 밀도(relevance density)를 갖도록 기여하는 것을 보인다. TREC WT10g와 GOV2 컬렉션과 같은 대규모의 웹 컬렉션에 대해서 현재 가장 우수한 성능을 보이는 기법인 적합모델(relevance model) [15]과의 비교실험에서 상당한 성능향상을 나타냄을 보인다.

본 논문의 구성은 다음과 같다. 2절에서는 선택적 재샘플링 기법의 가정에 대해 설명하고, 3절에서 중첩 클러스터를 이용한 피드백 문서의 재샘플링 기법을 소개한다. 4절에서 다양한 TREC 컬렉션들에 대한 실험 결과 및 분석을 하고, 5절에서 적합밀도에 의한 검증을 한다. 6절에서 관련 연구를 소개하고, 7절에서 결론을 맺는다.

2. 선택적 재샘플링 기법

잠정적 적합피드백의 주요 이슈는 상위 검색문서집합에서 어떻게 적합한 문서를 선택하고, 그 문서들에서 어떻게 확장 어휘를 선택하는가 하는 것이다. 본 연구에서는 보다 적합한 피드백 문서를 선택하는 문제를 다룬다. 전형적인 잠정적 적합피드백 기법의 문제는 정확률이 낮은 상위 검색문서집합에서 확장 어휘를 선택한다는 것이다. 보다 적합한 문서집합에서 확장 어휘를 선택할 수 있다면 보다 좋은 질의 표현이 될 것이다.

통계학에서 재샘플링 (부트스트래핑) [8]은 통계적 추론과정으로, 원래 샘플로부터 반환을 하면서 무작위로 샘플링함으로써 보다 견고한 추정을 하면서 샘플 통계(중간값, 분산 등)의 정확도를 추정하고자 하는 기법이다. 원래 샘플 공간으로부터 보다 좋은 예제들을 선택할 수 있는 방법이 있다면, 무작위 샘플링(random sampling)을 하는 것 보다는 선택적 샘플링 (selective sampling)을 하는 것이 보다 더 좋은 결과를 낼 수 있을 것이다. 기계학습 연구분야에서 부스팅 (Boosting) [27, 10] 기법은 선택적 재샘플링의 하나이다. 이는 약 학습자 (weak learner)가 이전 약 학습자가 잘못 분류한 예제들에 중점을 두도록 하기 위해서 학습 예제들의 분포를 적응적으로 변화시키는 데 사용하는 반복적 절차이다.

잠정적 적합피드백에서의 초기 검색집합은 샘플링 분포를 추정하기 위한 질의확장 어휘의 샘플공간으로 볼 수 있다. 질의는 여러 세부주제 (subtopics)를 포함할 수 있기 때문에, 검색집합은 여러 세부주제 그룹으로 나뉘어질 수 있다. 어떤 문서가 여러 세부주제를 표현하고 있으면, 클러스터링에서 각 세부주제에 따른 클러스터에 나타날 것이다. 그래서 이러한 문서를 지배 문서라고 한다. 잠정 적합피드백 문제에서 분포를 변화시킬 어떤 방향을 찾기 위해, 질의에 대해 지배적 문서 (dominant document)를 질의가 내포하는 주제들을 잘 표현한 문서로서, 문서들 사이에서도 여러 가까운 이웃 문서들과 높은 유사도의 관계를 가질 것이라고 가정을 하였다. 따라서, 중첩 클러스터에서 지배 문서는 질의에 대

해 높은 우선순위를 갖는 여러 클러스터에 나타날 것이다. 이러한 지배 문서로부터 확장 어휘를 선택한다면 모든 세부 주제와 관련된 문서를 검색할 수 있는 어휘를 선택할 수 있을 것이다.

이러한 가정에 기초하여, 초기 검색문서집합 공간으로부터 최근접이웃 (k-Nearest Neighbors) 클러스터링 기법을 이용하여 중첩된 클러스터를 생성하고, 문서를 선택적으로 다시 샘플링하여 피드백을 한다.

3. 중첩 클러스터를 이용한 피드백 문서의 재샘플링 기법

본 연구에서 제안한 클러스터를 이용한 재샘플링 기법은 정보검색을 위한 언어모델(language model)[21]과 적합 모델(relevance model) [15] 틀 안에서 피드백을 위해 보다 적합한 문서를 획득하기 위한 방법이다. 적합 모델은 가장 첨단의 잠정적 적합피드백 기법으로서, 상위 검색문서들로부터 질의 모델을 구축하는 가장 강력한 방법으로 알려져 있다.[15, 7]

제안 기법의 핵심은 질의와 높은 관련도를 갖는 클러스터들에 중복되어 나타나는 지배적 문서는 다른 문서들에 비해 질의 확장에서 질의 어휘 표현에 보다 더 기여할 것이라는 것이다. 잠정적 적합피드백을 위해 중첩 클러스터를 이용한 재샘플링 과정은 다음과 같다.

우선, 주어진 질의에 대해 언어모델 [21]에 기반해서 문서들을 검색한다. 이때 언어모델에서 문서에 나타나지 않은 질의 어휘를 다루기 위해 디리슈레 평활 (Dirichlet smoothing) [35]을 이용하였다.

통계적 언어모델은 텍스트의 일부를 생성하는 가능한 모든 단어 열에 대한 확률 분포 [22]를 나타낸다. 정보검색에서 언어모델은 문서 그 자체를 모델로 다루고, 질의는 문서 모델로부터 생성된 텍스트의 열로써 다룬다. 질의확률 검색모델(query-likelihood retrieval model)은 최대확률추정을 이용하여 문서언어모델을 추정한다. 문서는 문서언어모델로부터 질의를 생성 또는 샘플링 할 확률 $P(Q|D)$에 의해서 순위화된다.

$$P(Q|D)=\prod_{i=1}^{m}P(q_i|D) \tag{1}$$

여기서 q_i는 i 번째 질의 어휘, m은 질의 Q의 어휘 개수이고, D는 문서 모델을 나타낸다.

디리슈레 평활은 문서에 나타나지 않은 질의 어휘에 대해 0이 아닌 값으로 추정하는데 사용된다. 질의확률 언어모델에 적용된 것은 다음과 같다:

$$P(w|D)=\frac{|D|}{|D|+\mu}P_{ML}(w|D)+\frac{\mu}{|D|+\mu}P_{ML}(w|Coll) \tag{2}$$

$$P_{ML}(w|D)=\frac{freq(w,D)}{|D|},\quad P_{ML}(w|Coll)=\frac{freq(w,Coll)}{|Coll|} \tag{3}$$

여기서 $P_{ML}(w|D)$은 문서 D에서의 어휘 w의 최대확률추정을 나타내고, $Coll$은 전체문서집합, μ은 평활 매개변수(smoothing parameter)를 나타낸다. $|D|$와 $|Coll|$은 문서 D와 전체문서집합 $Coll$의 길이를 나타낸다. $freq(w,D)$와 $freq(w,Coll)$은 문서 D와 전체문서집합 $Coll$에서의 어휘 w의 빈도수를 나타낸다. 평활 매개변수는 실험에서 각 전체 문서집합에 대한 학습질의를 이용해서 학습을 하였다.

그 다음으로, 검색된 N개의 문서집합에 대해서 지배적 문서들을 찾기 위해서 최근접이웃 (k-nearest neighbors; k-NN) 클러스터링 [9]을 하였다. 최근접이웃 클러스터링 방법은 전체 문서들 사이의 유사도를 모두 계산해서, 각 문서에 대해서 유사도가 높은 순서대로 k개의 가장 가까운 문서(이웃)을 선택해서 그 문서에 대한 클러스터를 형성하도록 한다. (실험에서 N은 100으로 설정하였고, k는 5로 설정하였다.) 클러스터를 생성하는데 드는 시간복잡도는 $O(N^2)$이다. 이때, 하나의 문서는 여러 개의 클러스터에 속할 수 있다. 즉, k-NN 클러스터링은 중첩클러스터를 생성한다.

문서 유사도(document similarity) 계산을 위해서 문서는 각 어휘를 tfidf 가중치로 계산한 후 코사인 정규화를 하여 표현하였다. 유사도 측정은 코사인 계수(cosine coefficient)를 이용하여 최상위 검색문서들에 대해서 계산하였다.

본 연구에서의 가정은 질의를 잘 표현한 지배적 문서는 높은 유사도로 여러 개의 이웃과 관계를 맺고, 또한 여러 클러스터를 형성하는데 핵심 역할을 할 것이라는 것이다. 반면에 부적합한 문서는 이상적으로는 높은 유사도의 이웃을 갖지 않아서 그 자신으로만 된 클러스터 (singleton)을 형성해야 하지만, 실제로는 일반 어휘나 여러 개의 뜻을 갖는 어휘 등으로 인해서 이웃들을 가져서 클러스터를 형성할 수도 있다. 문서 클러스터는 유사도 계산에서 문서와 어휘들 사이의 관계를 반영할 수 있다. 본 연구에서는 어떤 문서가 여러 클러스터의 멤버이고, 이 클러스터가 질의와 밀접한 관련이 높은 것이라면, 이러한 문서를 지배적 문서라고 가정하였고, 이 지배적 문서를 질의 확장을 위해서 반복적으로 피드백을 한 것이다.

클러스터를 생성한 후, 클러스터기반 질의확률 언어모델(cluster-based query-likelihood language model) [18]을 이용하여 클러스터를 순위화 한다. 최상위로 순위화된 클러스터들에 속하는 문서들이 피드백을 위해 사용된다. 여기서 클러스터는 피드백 문서를 선택하기 위해서만 사용된다.

클러스터기반 언어모델에서 클러스터는 자신의 멤버로 속한 모든 문서를 연결해서 하나의 큰 문서처럼 표현한 후 언어모델에 적용한 것이다.

$$P(Q|Clu)=\prod_{i=1}^{m}P(q_i|Clu) \tag{4}$$

$$P(w|Clu)=\frac{|Clu|}{|Clu|+\lambda}P_{ML}(w|Clu)+\frac{\lambda}{|Clu|+\lambda}P_{ML}(w|Coll) \tag{5}$$

$$P_{ML}(w|Clu)=\frac{freq(w,Clu)}{|Clu|},\quad P_{ML}(w|Coll)=\frac{freq(w,Coll)}{|Coll|} \tag{6}$$

여기서 *Clu*는 클러스터, *freq(w,Clu)*는 클러스터 *Clu*에 속하는 문서 D의 *freq(w, D)*를 합한 것이다.

질의 확장 어휘는 최상위로 순위화된 클러스터에 속하는 각 문서에 대해서 적합모델(relevance model)을 그대로 이용해서 선택한다. 여기서, 최상위 클러스터에서 선택한 피드백 문서들은 각 문서들의 초기 질의확률로 적합모델을 추정하는데 사용되는 것이다. 즉, 클러스터 표현이나 질의-클러스터 유사도가 사용되는 것이 아니다.

적합모델은 언어모델의 틀에 기반한 질의확장 기법이다. 적합 모델 [15]은 질의 Q가 주어졌을 때 어휘 w의 확률을 추정하는 다항분포이다. 이 모델에서 질의 어휘는 $q_1 \cdots q_m$ 이고, 적합 문서에서의 어휘 w는 분포 R에서 동시에 독립적으로 샘플링 한 것이다. 분포 R에서 어휘의 확률은 다음과 같이 추정된다.

$$P(w|R)=\sum_{D\in R} P(D)P(w|D)P(Q|D) \qquad (7)$$

여기서 R은 질의 Q에 대해 잠정적으로 적합(pseudo-relevant)하다고 가정한 문서들의 집합이다. 그리고 $P(D)$ 는 전체 집합에 대해 균일하다고 가정하였다. 결국 질의 확장을 위한 어휘 선택은 처음 질의에 대한 언어모델의 확률 $P(Q|D)$와 적합성 피드백 문서집합의 각 문서에서의 단어의 확률 $P(w|D)$를 곱한 것을 피드백문서들 전체에 대해서 누적된 값이 높은 것 순서대로 선택하게 된다.

이렇게 추정을 한 후에, 가장 확률 $P(w|R)$이 높은 e개의 어휘를 질의확장을 위해 선택한다. 최종 확장된 질의에는 원래 질의와 확장 어휘를 매개변수 λ로 가중치를 주어서 선형보간으로 합한다. 이때 λ도 실험에서 각 전체문서집합의 학습주제에 대해서 학습하여 얻는다.

적합모델 및 전형적인 잠정-적합 피드백 기법들은 초기 검색집합에 대해서 첫 단계 이후에 바로 질의확장 어휘를 얻는다. 이때의 문제점은 질의 중심적으로 순위화된 초기검색문서집합이 부적합한 문서들을 많이 포함하기 때문에 확장 어휘에도 부적합한 어휘가 포함되기 쉽다는 것이다. 이를 다루기 위해 본 연구에서는 문서들 사이의 관련도를 반영하면서, 중복된 클러스터를 생성해서 질의에 대해 핵심적인 역할을 하는 지배적 문서를 찾아서 반복적으로 피드백한 것이다. 제안 기법도 여전히 부적한 문서들을 포함할 수 있으나, 기존의 질의 중심적으로 선택된 기존 방법의 피드백 문서보다는 덜 포함하고 있음을 실험결과 분석에서 보일 것이다.

4. 실험 및 검증

본 논문에서 제안한 재샘플링 기법이 잠정적 적합피드백에 유효한지를 보기 위해, 다양한 TREC 실험집합 5개에 대해서 기본 검색모델, 기본 잠정적 피드백모델과 실제 피드백모델과 비교 실험을 하였다.

4.1 실험 환경 설정

4.1.1 실험 집합

다양한 평가를 위해 실험문서집합으로 뉴스기사집합인 TREC ROBUST, AP과 WSJ 컬렉션과 다양한 행태의 문서들로 구성된 대량의 웹 문서집합인 GOV2와 WT10g를 선택하였다. 모든 실험집합에 대해 질의는 TREC 질의형태에서 짧은 질의인 topic 필드를 사용하였다. 테스트컬렉션에 대한 보다 자세한 정보는 <표 1>과 같다:

색인과 검색을 위해서 인드리(Indri) 시스템 [28] 버전 2.3을 사용하였다. 모든 컬렉션에 대해서 포터 스태머 (Porter stemmer)로 어근처리 (stemming)를 하였다. 불용어 (stop word list)는 흔히 사용되고 있는 418개 리스트에 대해서 검색 시에 제거하였다.

4.1.2 학습 및 평가 방법

모든 컬렉션 각각은 학습 질의와 테스트 질의로 나누고, 각 모델의 파라미터 추정을 위해서 학습 질의를 이용하였고, 테스트 질의에 대해서 평가를 하였다.

언어모델에서 평활(smoothing) 파라미터 설정을 위해 다음과 같은 값($\mu \in$ {500, 750, 1000, 1500, 2000, …, 5000})에 대해서 최적의 파라미터를 찾았다. 적합모델에서 피드백 문서의 개수 ($|R| \in$ {5, 10, 25, 50, 75, 100}), 확장 어휘의 개수 ($e \in$ {10, 25, 50, 75, 100}), 원래 질의에 대한 가중치 ($\lambda \in$ {0.1, 0.2, …, 0.9})로 실험하였다. 제안 모델에서 피드백 클러스터의 개수 ($|C| \in$ {1, 2, 5, 10, 15, 20})로 이는 피드백 문서의 개수에 상응하는 값을 갖는다. k-NN 클러스터링에서 (k=5) 하나의 클러스터가 5개의 문서를 소속멤버로 가질 수 있기 때문이다. 클러스터링에서 무조건 5개의 멤버로 구성하지 않고, 문서 사이의 유사도가 0.25 이상의 값을 가지도록 제한하였다. (학습질의에 대해 0.15~0.35 사이의 값에서 비슷하게 좋은 성능을 보였다).

〈표 1〉 학습 및 테스트 실험집합들

테스트컬렉션	설명	문서의 수	학습 질의		테스트 질의	
			질의번호	개수	질의번호	개수
GOV2	2004 .gov 도메인	25,205,179	701-750	50	751-800	50
WT10g	TREC 웹 컬렉션	1,692,096	451-500	50	501-550	50
ROBUST	Robust 2004 컬렉션	528,155	301-450	150	601-700	100
AP	AP 88-90년 기사	242,918	51-150	100	151-200	50
WSJ	월스트리트 저널 87-92년 기사	173,252	51-150	100	151-200	50

확장 어휘는 다음과 같이 인드리 검색시스템의 형식으로 표현한다:

$$\#weight(\lambda\ \#combine(q_1 \ldots q_m) \quad (1-\lambda)\#weight(p_1\ t_1 \ldots p_e\ t_e))$$

여기서 $q_1 \cdots q_m$ 는 원래 질의 어휘를 나타내고, $t_1 \cdots t_e$ 는 확장 어휘로, e 개의 어휘를 나타낸다. 각 어휘는 확장 확률로 $p_1 \cdots p_e$,를 갖는다. λ는 원래 질의와 확장질의를 결합할 때의 파라미터로, 원래 질의에 어느 정도의 가중치를 부여하는지를 나타낸다.

학습 집합에 대해 모든 비교 기법은 평균정확률에 대한 평균값 (mean average precision; MAP)으로 최적화 시켰다.

$$MAP = \frac{1}{|Q|}\sum_{q\in Q} ap(q) \tag{8}$$

여기서 $ap(q)$는 질의 집합 Q에 있는 질의 q에 대한 평균 정확률을 나타낸다.

각 컬렉션의 학습집합에 대해 학습한 최적의 파라미터는 테스트 질의에 대한 실험 평가에서 이용된다.

4.1.3 비교 실험 방법들

비교 실험 (Baselines)으로 기준 검색모델로 언어모델과 기준 잠정적 적합피드백모델을 설정하였다.

- 언어모델 (*LM*): 기준이 되는 검색모델로 질의확률 언어모델의 성능과 비교한다. 적합모델과 제안모델은 이 언어모델의 틀에서 적합피드백을 하고 있어, 기본적으로 언어모델에 비해 우수한 성능을 보이는지 비교해야 한다.
- 적합모델(*RM*): 기준이 되는 잠정적 적합피드백 모델로 적합모델의 성능과 비교한다. 확장된 질의는 원래 질의와 결합시킨다 (RM3라고도 알려져 있다). 제안모델은 적합모델 틀에서 피드백 문서를 다른 것을 사용하고 있으므로, 기본적으로 적합모델에 비해 우수한 성능을 보이는지가 본 연구에서의 핵심 비교 실험이 된다.
  제안 모델의 상한선을 보기 위해, 실제로 적합한 문서로 피드백을 했을 때 (true relevance feedback)의 성능을 비교하였다.
- 실제 적합피드백(*TrueRF*): 잠정적 적합피드백 문서를 사용하지 않고, 사용자가 초기검색결과 100개에 대해 실제로 적합문서만을 선택해서 피드백 했을 때의 성능과 비교한다. 이는 적합모델을 이용했을 때의 상한선이 된다. 제안모델을 사용했을 때의 상한선이기도 하다.
  초기 검색결과집합에 대한 클러스터의 효과를 보기 위해, 클러스터 기반 재순위화 기법과 비교하였다.
- 클러스터 기반 재순위화 (*Rerank*): 질의중심적으로 검색된 상위 N개의 문서들에 대해서 k-NN 클러스터링 기법으로 생성된 클러스터와 문서에 대한 질의확률을 결합해서 재순위화 한 방법 [17]과 비교한다. 이때 N은 1,000으로, k는 5로 설정했다.

$$P'(Q \mid D) = P(Q \mid D)\cdot MAX_{D\in Clu_i} P(Q \mid Clu_i) \tag{9}$$

여기서 문서는 여러 클러스터의 멤버가 될 수 있기 때문에, 문서 D가 속하는 클러스터 Clu 중에서 질의확률을 최대로 갖는 값으로 선택한다.

4.2 실험 결과

각 비교방법들에 대해서 모든 테스트컬렉션에 대한 실험결과가 <표 2>에 나타나 있다. 기본 검색모델과의 비교실험에서, 재샘플링 기법 (*Resampling*)은 모든 테스트컬렉션에 대해서 질의확률 언어모델 (*LM*) 보다 현저히 성능향상을 보이고 있다.

기본 잠정적 적합피드백 모델과의 비교실험에서, GOV2와 WT10g와 같이 다양한 형태의 웹 문서들로 구성된 대량의 컬렉션에 대해서 재샘플링 기법은 적합모델에 비해 현저히 성능향상을 보이고 있다.

RM에 비해 Resampling의 상대적인 성능향상 수준은 GOV2에 대해서는 6.28%, WT10g에 대해서는 19.63%이다. 뉴스기사들로 구성된 ROBUST 컬렉션에 대해서는 Resampling은 RM에 비해 조금 낮은 성능을 보이고 있고, AP와 WSJ 컬렉션에 대해서는 Resampling이 RM 보다 조금 높은 성능을 보이고 있으나 현저한 수준은 아니다.

상위문서 5개에 대한 정확률 (P@5)로 평가했을 때, 재샘플링기법은 기본 검색모델인 언어모델에 비해 각 컬렉션 GOV2, WT10g, ROBUST, AP, 그리고 WSJ에 대해 각각

〈표 2〉 모든 테스트컬렉션에 대해 테스트 질의에 대한 성능 비교 (MAP). 윗첨자로 표시한 α, β, γ 와 δ는 각각 LM, Rerank, RM, 그리고 Resampling에 대해 통계적으로 우수한 수준으로 성능향상이 되었음을 나타낸다. 즉, α는 언어모델(LM)에 대해 성능향상에 차별성이 있음을 나타내고, β는 Rerank에 비해, γ는 RM에 비해, δ는 Resampling에 비해 보다 성능향상이 차별성이 있음을 표시함. 대응표본 t-검증을 $p < 0.05$ 수준에서 한 것이다

	언어모델 (LM)	재순위화 (Rerank)	적합모델 (RM)	재샘플링 (Resampling)	적합모델 상한선 (TrueRF)
GOV2	0.3258	0.3406 $^{\alpha}$	0.3581 $^{\alpha\beta}$	0.3806 $^{\alpha\beta\gamma}$	0.4315 $^{\alpha\beta\gamma\delta}$
WT10g	0.1861	0.2044 $^{\alpha}$	0.1966	0.2352 $^{\alpha\beta\gamma}$	0.4030 $^{\alpha\beta\gamma\delta}$
ROBUST	0.2920	0.3206 $^{\alpha}$	0.3591 $^{\alpha\beta}$	0.3515 $^{\alpha\beta}$	0.5351 $^{\alpha\beta\gamma\delta}$
AP	0.2077	0.2361 $^{\alpha}$	0.2803 $^{\alpha\beta}$	0.2906 $^{\alpha\beta}$	0.4253 $^{\alpha\beta\gamma\delta}$
WSJ	0.3258	0.3611 $^{\alpha}$	0.3967 $^{\alpha\beta}$	0.4033 $^{\alpha\beta}$	0.5306 $^{\alpha\beta\gamma\delta}$

〈표 3〉 상위5개 문서에서의 정확률 (P@5)에 대한 성능 비교

	언어모델	재순위화	(변화율)	적합모델	(변화율)	재샘플링	(변화율)	상한선
GOV2	0.6200	0.6720	(8.39%)	0.5760	(-7.10%)	0.7120	(14.84%)	0.8960
WT10g	0.3306	0.4041	(22.23%)	0.3551	(7.41%)	0.4122	(24.68%)	0.7306
ROBUST	0.5152	0.5434	(5.47%)	0.5232	(1.55%)	0.5354	(3.92%)	0.8384
AP	0.3200	0.3920	(22.50%)	0.3800	(18.75%)	0.3840	(20.00%)	0.7640
WSJ	0.5400	0.5760	(6.67%)	0.5800	(7.41%)	0.6040	(11.85%)	0.8560

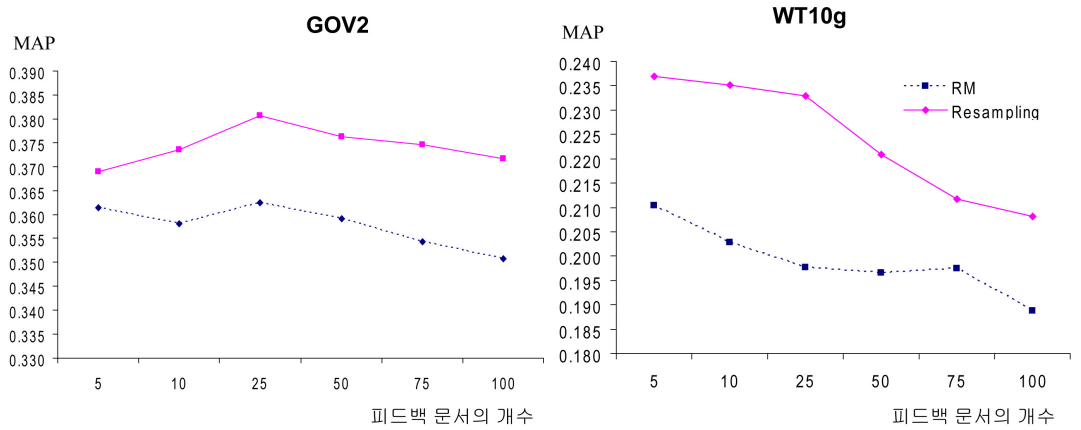
14.84%, 24.68%, 3.92%, 20.0%, 그리고 11.85%의 성능향상을 보였다. 반면에 적합모델은 언어모델에 비해 해당 각 컬렉션에 대해 -7.1%, 7.4%, 1.6%, 18.8% 그리고 7.4%의 변화를 나타냈다. 재샘플링 기법이 적합모델에 비해 P@5에서 성능 향상률이 높았다.

클러스터를 이용한 재순위화 기법은 모든 테스트컬렉션에서 MAP 평가에서 LM에 비해 현저한 성능향상을 보였다. 사실, Rerank 기법은 WT10g에 대해서 RM보다 높은 성능을 보이고 있다. P@5 평가에서는 대부분의 컬렉션에 대해 RM보다 높은 정확률을 나타내고 있다. 이러한 결과는 문서 클러스터링이 초기검색결과에 대해 적합문서 그룹을 찾고, 질의에 대한 문서의 잠정적 문맥을 제공하는데 도움이 된다는 것을 보여주는 것이다.

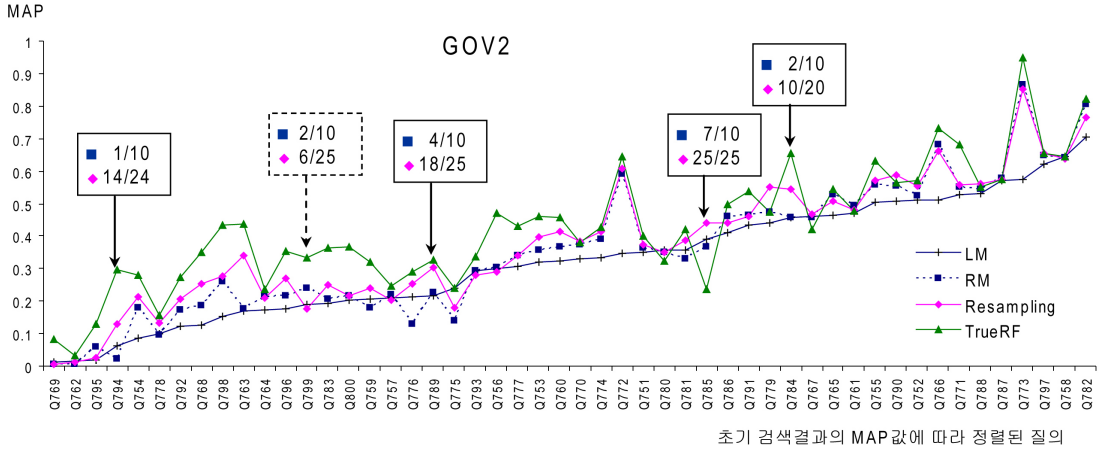
제안모델 및 적합모델의 성능 상한선 (upper-bound performance)을 보기 위한 실험으로, 실제 적합피드백 (*TrueRF*)은 모든 컬렉션에 대해 현저한 수준의 성능향상을 보이고 있다. 적합모델의 틀에서 보다 좋은 잠정적 적합피드백 문서를 선택했을 때 예측할 수 있는 성능 수준이다. MAP 평가에서 0.4와 0.5의 수준이므로, 보다 좋은 확장 어휘와 어휘에 대한 가중치를 적용하는 문제는 여전히 남아있다.

위의 실험결과는 각 컬렉션에 대해 학습질의를 이용하여 최상의 성능을 보이는 파라미터들을 학습하여 실험한 것이다. 피드백 문서의 개수가 달라짐에 따라 성능이 어떻게 변하는지를 분석하였다. (그림 1)에서 보는 것과 같이, 재샘플링 기법이 피드백 문서의 수에 상관없이 적합모델에 비해 높은 성능을 보였다.

(그림 2)는 각 질의에 대해 피드백으로 사용된 문서가 포함하는 적합한 문서와 부적합한 문서의 수를 분석하였다.



(그림 1) GOV2와 WT10g 컬렉션에 대한 피드백 문서의 개수에 따른 성능 비교



(그림 2) 피드백에 사용된 문서에 대한 분석 (적합문서개수/전체피드백문서개수). GOV2 컬렉션

GOV2 컬렉션에서 대부분의 경우 재샘플링기법이 적합모델보다 피드백에 사용된 적합문서의 비율이 높고, 그 결과로 성능 또한 높게 나타났다. 점선으로 표시된 부분은 그 반대의 경우를 보이는데 이러한 행태는 많지 않다. 전체 50개의 질의에 대해, 재샘플링기법이 41개의 질의에 대해서 LM에 비해 성능향상을 보였고, 9개의 질의에 대해 성능이 낮아졌다. 적합모델은 37개의 질의에 대해서 LM에 비해 성능향상을 보였고, 13개의 질의에 대해 성능이 저조했다.

5. 적합밀도를 이용한 검증

본 연구의 기본 가정인 지배적 문서는 질의 적합한 문서이고, 상위로 순위화된 클러스터에 반복적으로 나타난다는 것을 정당화하기 위해 적합밀도를 측정하였다.

적합 밀도(relevance density)는 피드백 문서 수에 비해서 실제 적합 문서의 포함 비율로 정의하였다.

$$\text{적합밀도} = \frac{\text{적합한 피드백 문서의 개수}}{\text{피드백 문서의 개수}} \qquad (10)$$

적합밀도가 높다는 것은 포함하고 있는 적합문서의 수가 많기 때문에 검색 정확도가 더 클 것이라는 것을 의미한다. 적합밀도가 1이 되면 사람에 의해 판별한 적합한 문서만으로 피드백하는 실제 적합피드백 결과에 접근하게 된다. 제안하는 기법이 효과적이라면 기존 방법보다 높은 적합밀도를 보일 것이다.

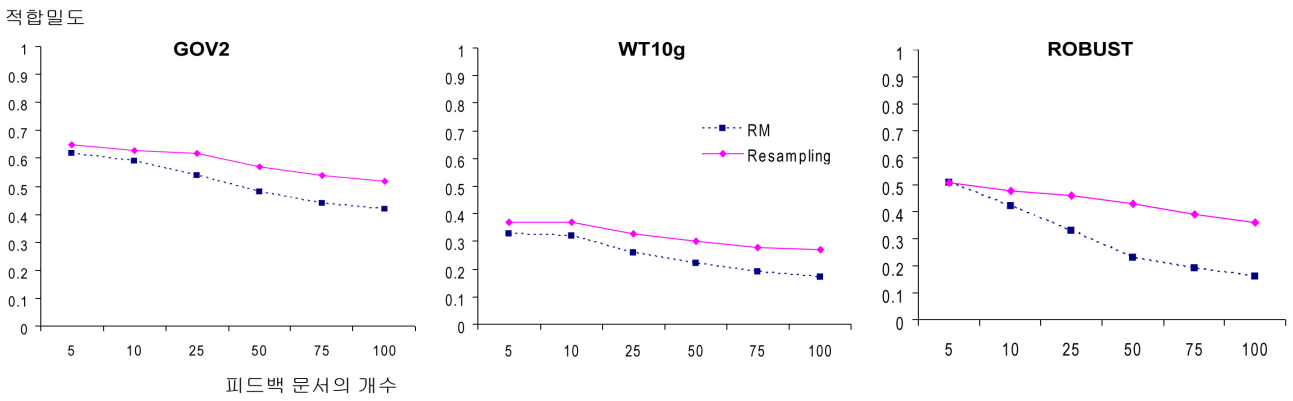
제안 방법을 정당화 하기 위해 피드백 문서에 대한 적합밀도를 기존의 방법과 비교하고, 지배적 문서의 효과를 보기 위해 중복해서 나타나는 문서를 반복적으로 피드백을 하지 않았을 때의 성능을 비교 분석한다.

(그림 3)에서 보는 것처럼, 제안 방법이 최신의 잠정적 피드백모델에서 가장 우수한 성능을 보이고 있는 적합모델에 비해 모든 테스트컬렉션에 대해 높은 적합밀도를 나타내고 있다. (AP와 WSJ 컬렉션에 대한 결과는 그림에는 나타나 있지 않으나 ROBUST와 같은 행태를 보였다.) <표 4>에서 보는 바와 같이 피드백 문서의 수가 100일 때, 모든 컬렉션에서 RM보다 높은 적합밀도를 보이고 있다. 이는 피드백 문서의 수를 100으로 고정했을 때, 제안 방법이 적합모델의 성능을 능가할 것이라는 것을 예측할 수 있다. 이러한 결과는 적합밀도가 성능향상과 관련을 지지하고 있음을 보인 것이다.

질의에 대한 적합밀도의 구체적인 분석을 위해 WT10g의 한 질의에 대해서 피드백에 사용된 상위 문서들 5, 10, 25, 50, 75, 100개에 대해서 적합밀도를 살펴보았다. 적합모델기법에 대해서는 적합밀도가 각각 0.6, 0.5, 0.36, 0.3, 0.25를 나타낸 데 비해, 제안기법에 대해서는 거의 완벽한 수준으로 각각 1.0, 1.0, 0.96, 0.98, 0.97, 0.89를 나타냈다. 질의에 적합한 문서들이 상위의 클러스터에 반복적으로 나타나는 것을 확인할 수 있다. 이러한 적합밀도에의 변화가 앞서 실험에서 제안기법이 성능을 향상시킬 수 있었던 바탕이 된 것이라고 확신할 수 있다.

6. 관련 연구

적합피드백 (Relevance feedback)와 잠정적 적합피드백



(그림 3) 피드백 문서의 개수에 따른 적합모델(RM)과 제안 방법 (Resampling)의 적합밀도 비교

〈표 4〉 피드백 문서의 개수를 고정시켰을 때의 성능 비교. 문서와 어휘의 개수를 100으로 고정시켰다. 윗첨자 α와 β는 각각 LM과 RM에 비해 통계적으로 상당히 성능향상을 보였음 (대응표본 t-검증 $p < 0.05$)을 나타낸다.

	언어모델(LM)	적합모델(RM)	(변화율)	재샘플링(Resampling)	(변화율%)
GOV2	0.3258	0.3519 $^{\alpha}$	(8.01%)	0.3764 $^{\alpha\beta}$	(15.53%)
WT10G	0.1861	0.1886	(1.34%)	0.2072 $^{\alpha}$	(11.34%)
ROBUST	0.2920	0.3262 $^{\alpha}$	(11.71%)	0.3549 $^{\alpha\beta}$	(21.54%)
AP	0.2077	0.2758 $^{\alpha}$	(32.79%)	0.2853 $^{\alpha}$	(37.36%)
WSJ	0.3258	0.3785 $^{\alpha}$	(16.18%)	0.4009 $^{\alpha\beta}$	(23.05%)

(pseudo-relevance feedback)은 처음 검색결과를 이용해서 적합문서나 잠정적 적합문서를 이용해서 원래 질의를 다시 작성함으로써 검색의 정확도를 향상시키는데 효율적인 방법으로 알려져 있다. TREC 2008 [30]의 적합피드백 트랙(relevance feedback track)을 새로 시작하면서 적합피드백에 관한 새롭게 관심을 받게 되었다.

전형적인 잠정적 적합피드백 알고리즘인 Okapi BM25 [22]와 적합모델 [15] 등은 검색된 초기검색결과의 상위 문서들은 질의에 적합하다고 가정을 하고 있다. 이 방법의 성능을 향상시키기 위한 연구는 문서 대신 문단(passage)을 이용하는 방법 [33], 지역문맥분석 기법 [31], 질의에 대해 정규화된 추정 기법 [29], 잠정적 개념을 이용하는 기법 [20] 등이 있다. 이러한 연구들도 기본적인 가정은 상위검색결과가 질의에 대해 적합하다는 것이다.

반면에 최근에 초기검색결과의 상위문서를 그대로 피드백으로 사용하지 않고, 샘플링과 재샘플링을 통해서 피드백을 하는 연구가 있다. Sakai 등에 의해 제안된 선택적 샘플링 기법 [25]에서는 클러스터링 기준에 따라 상위 검색된 문서들 중에서 일부를 피드백에 사용하지 않도록 한다. 이때 클러스터는 문서들 사이의 유사도에 의해 생성된 것이 아니라 같은 질의 어휘 집합을 갖느냐에 따른 것이다. 샘플링의 목적은 보다 다양하고 새로운 문서집합을 피드백에 사용하려는 것이다. 이는 상위 문서들은 서로 비슷하거나 중복될 것이라는 기본 가정에 따른 것이다. 본 연구에서 지배적 문서를 반복적으로 사용하는 기본 가정과는 서로 다르다.

Collins-Tompson와 Callan에 의해 제안된 재샘플링 기법 [5]은 질의에 대해 검색된 상위 문서에 대한 부트스트랩 샘플링과 질의 어휘에서 하나의 어휘를 제거해서 만든 질의 변이를 사용한다. 질의 변이를 사용한 가정은 질의 어휘들 중의 하나는 쓸데없는 어휘일 것이라는 것이다. 그들의 실험분석에서 언급한 것처럼, 성능향상은 문서 재샘플링이 아니라 질의 변이를 사용한 것의 효과이다. 본 연구에서는 주로 상위검색결과에 대한 재샘플링의 효과에 초점을 맞춘 것이다.

정보검색의 다른 측면에서는, 많은 검색기법이 성능향상을 위해서 클러스터 가설(cluster hypothesis)을 채택해오고 있다. 클러스터 가설은 아주 가깝게 관련된 문서들은 같은 질의에 대해 적합하다 [11]고 보고 있다. 클러스터를 이용하여 순위를 다시 매긴 기법 [16, 17]은 벡터공간검색 모델에서 성공적인 결과를 보였다. 클러스터 기반 언어모델 검색기법 [18]은 질의를 생성하는 확률로 클러스터를 순위화 한 것으로 질의확률검색모델에 비해 성능향상을 보였다. 지역점수 정규화 (local score regularization) 기법 [6]은 주제별 관련문서들이 유사한 점수를 받도록 문서유사도 행렬을 이용해서 초기검색 점수를 적응시킨 것으로, 언어모델에서 좋은 결과를 보였다. 또한, 클러스터 정보를 문서기반 언어모델에 통합한 기법 [12], 그래프구조 틀 안에서 클러스터기반 언어모델을 이용한 재순위화 기법 [14], 유사 그래프 구조를 이용한 재순위화 기법 [34] 등 클러스터 정보를 이용하여 검색 결과를 향상시키는 연구가 지금도 계속 연구되고 있다.

7. 결 론

본 연구에서 제안한 방법인 중첩 클러스터를 이용해서 상위검색 문서들을 재샘플링하여 피드백 하는 것은 잠정적 적합피드백에서 유효한 방법이다. 거의 모든 실험집합에서 일관적으로 성능향상을 보였고, 특히 대규모의 다양한 문서를 포함하는 TREC GOV2와 TREC WT10g 컬렉션에서는 상당히 우수한 성능향상을 보였다. 정보검색모델에서 우수한 성능을 보이는 언어모델 (LM)과 잠정적 적합피드백에서 최고의 성능을 보이는 적합모델 (RM)과의 비교실험에서 GOV2 집합에 대한 상대적인 성능향상은 각각 16.82%와 6.28%를 보였다. WT10g 집합에 대해서는 각각 19.63%와 26.38%의 성능향상을 보였다. 클러스터에 기반한 재샘플링에 의한 질의 확장이 어떻게 성능향상을 도왔는지에 대한 검증을 위해서 적합밀도를 측정했는데, 모든 실험집합에 대해서 비교 피드백모델 보다 높은 적합밀도를 보였음을 확인하였다. 이러한 실험결과에서 중첩 클러스터가 질의에 대해 핵심 문서를 찾는데 도움이 되었음을 알 수 있다.

향후 연구에서는 질의 특성에 따른 클러스터링 결과의 적용 방법과 클러스터링에 질의 변이를 채택하는 방법에 대한 연구가 필요하다. 또한 클러스터의 표현으로 단순히 문서들을 연결하지 않고, 보다 나은 클러스터 표현 방법과 클러스터 순위화 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] Attar, R. and Fraenkel, A. S. 1977. Local Feedback in Full-Text Retrieval Systems. *Journal of the ACM* 24, 3 (Jul. 1977), pp.397-417.

[2] Buckley, C. and Harman, D. 2004. Reliable information access final workshop report. http://nrrc.mitre.org/NRRC/publications.htm

[3] Buckley, C., Mitra, M., Walz, J., and Cardie, C. 1998. Using clustering and superconcepts within SMART: TREC 6. In Proc. 6th Text REtrieval Conference (TREC-6).

[4] Buckley, C. and Robertson, S. 2008. Proposal for relevance feedback 2008 track. http://groups.google.com/group/trec-relfeed

[5] Collins-Thompson, K., and Callan, J. 2007. Estimation and use of uncertainty in pseudo-relevance feedback. In Proc. 30th ACM SIGIR conference on Research and Development in Information Retrieval, pp.303-310.

[6] Diaz, F. 2005. Regularizing ad hoc retrieval scores. In Proc. 14th ACM international conference on Information and knowledge management (CIKM), pp.672-679.

[7] Diaz, F., and Metzler, D. 2006. Improving the Estimation of Relevance Models Using Large External Corpora, In Proc.

29th ACM SIGIR conference on Research and Development in Information Retrieval, pp.154-161.

[8] Efron, B. 1979. Bootstrap methods: Another look at the jackknife, The Annals of Statistics, 7, pp.1-26.

[9] Fix, E. and Hodges, L. 1951. Discriminatory analysis: non-parametric discrimination: consistency properties. Technical Report, USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004.

[10] Freund, Y. 1990. Boosting a weak learning algorithm by majority. In Proc. 3rd Annual Workshop on Computational Learning Theory.

[11] Jardine. N. and Rijsbergen, C.J.V. 1971. The use of hierarchic clustering in information retrieval. Information Storage and Retrieval, 7, pp.217-240.

[12] O. Kurland and L. Lee, 2004. Corpus structure, language models, and ad hoc information retrieval. In Proc. 27th ACM SIGIR conference on Research and Development in Information Retrieval, pp.194-201.

[13] Kurland, O., and Lee, L. 2005. Better than the real thing? Iterative pseudo-query processing using cluster-based language models. In Proc. 28th ACM SIGIR conference on Research and Development in Information Retrieval, pp.19-26.

[14] Kurland, O., and Lee, L. 2006. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In Proc. 29th ACM SIGIR conference on Research and Development in Information Retrieval, pp.83-90.

[15] Lavrenko, V. and Croft, W.B. 2001. Relevance-based language models. In Proc. 24th ACM SIGIR conference on Research and Development in Information Retrieval, pp.120-127.

[16] Lee, K.S., Park, Y.C., and Choi, K.S. 2001. Re-ranking model based on document clusters. Information Processing and Management, 37, pp.1-14.

[17] Lee, K.S., Kageura, K., and Choi, K.S. 2004. Implicit ambiguity resolution based on cluster analysis in cross-language information retrieval, Information Processing and Management, 40, pp.145-159.

[18] Liu, X., and Croft, W.B. 2004. Cluster-based retrieval using language models. In Proc. 27th ACM SIGIR conference on Research and Development in Information Retrieval, pp. 186-193.

[19] Lynam, T., Buckley, C., Clarke, C., and Cormack, G. 2004. A multi-system analysis of document and term selection for blind feedback. In Proc. 13th ACM international conference on Information and knowledge management (CIKM), pp. 261-269.

[20] Metzler, D., and Croft, W. B. 2007. Latent Concept Expansion Using Markov Random Fields, In Proc. 30th ACM SIGIR conference on Research and Development in Information Retrieval, pp.311-318.

[21] Ponte, J.M., and Croft, W.B. 1998. A language modeling approach to information retrieval. In Proc. 21st ACM SIGIR conference on Research and Development in Information Retrieval, pp.275-281.

[22] Robertson, S.E., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. 1996. Okapi at TREC-4. In Proc. 4th Text REtrieval Conference (TREC).

[23] Rocchio, J.J. 1971. Relevance feedback in information retrieval. The SMART retrieval system, Prentice-Hall, pp.316-321.

[24] Rosenfeld, R. 2000. Two decades of statistical language modeling: where do we go from here? In Proc. of the *IEEE*, 88(8), pp.1270-1278.

[25] Sakai, T., Manabe, T. and Koyama, M. 2005. Flexible pseudo-relevance feedback via selective sampling. ACM Transactions on Asian Language Information Processing (TALIP), 4(2), pp.111-135.

[26] Salton, G., and Buckley, C. 1990. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), pp.288-297.

[27] Schapire, R. 1990. Strength of weak learnability. Journal of Machine Learning, 5, pp.197-227.

[28] Strohman, T., Metzler, D., Turtle, H., and Croft, W.B. 2005. Indri: A language model-based search engine for complex queries. In Proc. International Conference on Intelligence Analysis.

[29] Tao, T., and Zhai, C. 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. In Proc. 29th ACM SIGIR conference on Research and Development in Information Retrieval, pp.162-169.

[30] TREC. 20008. Call for participation. http://trec.nist.gov/call08.html

[31] Xu, J and Croft, W.B. 1996. Query expansion using local and global document analysis. In Proc. 19th ACM SIGIR conference on Research and Development in Information Retrieval, pp.4-11.

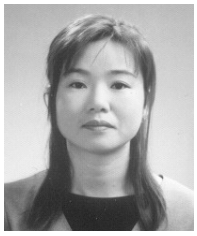
[32] Yang, L., Ji, D., Zhou, G., Nie, Y., and Xiao, G. 2006. Document re-ranking using cluster validation and label propagation. In Proc. 15th ACM international conference on Information and knowledge management CIKM), pp.690-697.

[33] Yeung, D.L., Clarke, C.L.A., Cormack, G.V., Lynam, T.R., and Terra, E.L. 2004. Task-specific query expansion. In Proc. 12th Text REtrieval Conference (TREC), pp.810-819.

[34] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. 2005. Improving web search results using affinity graph. In Proc. 28th ACM SIGIR conference on Research and Development in Information Retrieval,

pp.504-511.

[35] Zhai, C., and Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22(2), pp.179-214.



이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과 학사

1997년 한국과학기술원 전자전산학 석사

2001년 한국과학기술원 전자전산학 박사

2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현 재 전북대학교 전기전자컴퓨터공학부 조교수

관심분야 : 정보검색, 정보 마이닝, 자연언어처리