주제를 깊이 있게 다루는 블로그 피드 검색을 위한 위키피디아 기반 질의 확장 방법

(A Wikipedia-based Query Expansion Method for In-depth Blog Distillation)

송 우 상 † 이 예 하 ††
(Woosang Song) (Yeha Lee)
이 종 혁 ††† 양 기 주 ††††
(Jong-Hyeok Lee) (Gijoo Yang)

요 약 본 논문에서는 질의로 주어진 주제를 깊이 있게 다루는 블로그 검색을 위한 위키피디아 기반 질의 확장 방법을 제안한다. 제안된 방법은 질의와 연관된 위키피디아 문서를 질의 확장에 사용한다. 실험을 위해 대규모 블로그 실험 데이터인 TREC Blogs08 collection과 영문 위키피디아 데이터를 사용하였다. 실험 결과 제안된 방법은 기존의 블로그 포스트 기반 질의 확장 방법에 비해 MAP을 비롯한 검색 성능을 큰 폭으로 향상시켰다.

키워드 : 블로그 피드 검색, 위키피디아, 정보 검색, 질의 확장

· 본 논문은 2010년도 두뇌한국21사업, 포항공과대학교 정보통신연구소 자체 학술연구과제(선도과제), 그리고 한국과학재단 기초연구사업(No. 2010-0012662)의 지원으로 수행되었습니다.
· 이 논문은 2010 한국컴퓨터종합학술대회에서 '주제를 깊이 있게 다루는 블로그 피드 검색을 위한 위키피디아 기반 질의확장 방법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : POSTECH 컴퓨터공학과
woosang@postech.ac.kr
†† 비 회 원 : POSTECH 컴퓨터공학과
sion@postech.ac.kr
††† 종신회원 : POSTECH 컴퓨터공학과 교수
jhlee@postech.ac.kr
†††† 정 회 원 : 동국대학교 정보통신공학과 교수
gjyang@dgu.edu

논문접수 : 2010년 8월 10일
심사완료 : 2010년 10월 7일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제11호(2010.11)

Abstract This paper proposes a Wikipedia-based feedback method for in-depth blog distillation whose goal is to find blogs that represent in-depth thoughts or analysis on a given query. The proposed method uses Wikipedia articles which are relevant to the query. TREC Blogs08 collection which is a large-scale blog corpus and English Wikipedia dump were used for experiments. The proposed method significantly increased the retrieval performance including MAP over the conventional post-based feedback method.

Key words : Blog Feed Search, Wikipedia, Information Retrieval, Query Expansion

1. 서 론

블로그 피드 검색(blog feed search, blog distillation)은 질의로 주어진 주제를 중점적이며 반복적으로 다루는 블로그를 찾는 것을 목적으로 한다. 블로그 피드 검색은 개별 문서(포스트)의 집합인 블로그를 검색 대상으로 한다는 점과, 주관적이며 표현 및 형식이 자유로운 블로그 문서를 처리한다는 점에서 일반적인 정보 검색과 구분된다. 2007년 Text REtrieval Conference(TREC) [1]에서 처음으로 소개된 이후 블로그 피드 검색을 위한 다양한 방법들이 제안되었다[1,2]. 그러나 현재까지의 블로그 피드 검색은 질의로 주어진 주제와 블로그간의 연관 여부만을 고려해왔으며, 검색된 블로그가 주제에 대해 깊이 있는 수준의 지식을 제공하는지 아니면 일반적 또는 피상적인 수준의 지식만을 제공하는지에 대한 여부는 고려되지 않았다.

이러한 단점을 개선하기 위해 주제를 깊이 있게 다루는 블로그에 대한 검색(in-depth blog feed search, in-depth blog distillation)이 2009년 TREC에서 새로운 연구 주제로 제안되었다[3]. In-depth 블로그 피드 검색은 질의로 주어진 주제와 연관성을 가지며, 동시에 해당 주제에 대한 깊이 있는 지식과 분석을 포함하는 블로그(in-depth blog)를 검색 대상으로 한다. 2009년 TREC에서 이를 위한 다양한 방법들이 제안되었지만, 기존의 블로그 피드 검색에서 사용된 방법들에 비해 in-depth 블로그 피드 검색 성능을 향상시키지 못했다.

본 논문에서는 in-depth 블로그 피드 검색 성능 향상을 위해 위키피디아 문서를 사용한 질의 확장 방법을 제안한다. 위키피디아는 사용자 참여 방식의 인터넷 백과 사전으로 다양한 방면의 주제에 대한 깊이 있고 분석적인 내용을 담고 있다. 질의가 나타내는 주제와 연관된 위키피디아 문서를 질의 확장에 사용할 경우 주제와 연관된 다양한 전문 용어들과 세부 개념을 설명하는 용어들을 중심으로 질의를 확장할 수 있다. 이렇게 확장된 질의를 검색 과정에 적용하여 in-depth 블로그에 대해

높은 연관 점수를 부여할 수 있으며 이를 통해 in-depth 블로그 피드 검색 성능을 향상시킬 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 in-depth 블로그 피드 검색과 관련된 연구를 소개한다. 3장에서는 in-depth 블로그 피드 검색 모델에 대한 내용을 기술하며, 4장에서는 본 논문에서 제안하는 위키피디아 기반 질의 확장 방법을 설명한다. 5장에서는 제안된 방법에 대한 실험 결과를 기술 및 분석하며, 6장에서는 결론과 함께 향후 연구 과제에 대해서 검토한다.

2. 관련 연구

2009년 TREC에서는 in-depth 블로그 피드 검색을 위한 다양한 방법들이 제안되었다. [4]는 블로그 포스트의 길이를 in-depth 블로그에 대한 판단 자질로 사용한 검색 방법을 제안하였다. 이 방법은 길이가 긴 포스트가 주제를 깊이 있게 나타낸다고 가정하였다. [5]는 일반적인 블로그에서 나타나지 않는 단어를 포함하는 블로그를 in-depth 블로그라고 가정하였고, 이를 위해 블로그 포스트와 전체 문서 집합간의 Cross Entropy를 계산하여 랭킹에 반영하였다. [6]은 in-depth 블로그는 객관적 사실을 기술하는 단어들을 주로 포함한다고 가정하였고, 이를 위해 subjective lexicon을 사용한 질의 확장을 통해 추정된 언어 모델을 검색 과정에 적용하는 방법을 제안하였다. [7]은 문서의 길이, 문장의 길이, 인칭 대명사의 수 등 다양한 자질을 사용한 classification 기반 방법을 제안하였다. 하지만 이러한 방법들은 in-depth 블로그 피드 검색에서 기존의 블로그 피드 검색에서 사용되었던 일반적인 방법들보다 낮은 검색 성능을 보여주었다.

3. In-depth 블로그 피드 검색

In-depth 블로그 피드 검색은 2009년 TREC blog track에서 주제와의 연관성과 함께 각 질의마다 제시된 기준을 만족시키는 블로그를 검색 대상으로 하는 faceted blog distillation의 subtask로 처음 소개되었다 [3]. 기존의 블로그 피드 검색은 질의로 주어진 주제와 연관된 모든 블로그를 검색 대상으로 하는 반면, in-depth 블로그 피드 검색은 주제와 연관된 블로그를 "in-depth" 블로그와 "shallow" 블로그로 구분하며 in-depth 블로그만을 검색 대상으로 한다. In-depth 블로그는 주제에 대한 깊이 있는 지식과 분석을 포함하는 블로그를 말한다. Shallow 블로그는 in-depth 블로그와 반대되는 개념으로 주제에 대해 피상적이며 일반적인 블로그에서 제공할 수 있는 수준의 정보만을 제공하는 블로그를 말한다.

3.1 검색 모델

본 논문에서 사용된 in-depth 블로그 피드 검색 모델의 전체적인 검색 과정은 다음과 같다. 초기 질의(initial query)는 질의가 의미하는 주제를 명확하게 나타내지 못하기 때문에 질의 확장 과정을 통해 질의와 연관된 다수의 단어로 확장된다. 일반적인 질의 확장 방법은 초기 질의 검색 결과 상위의 블로그 포스트들을 질의 확장에 사용한다. 본 논문에서는 in-depth 블로그 피드 검색을 위해 질의로 주어진 주제와 연관된 위키피디아 문서를 질의 확장에 사용한다. 이를 통해 주제와 연관된 다양한 전문 용어 및 세부 개념을 설명하는 용어들을 중심으로 질의를 확장한다. 확장된 질의를 통해 각 블로그의 연관 점수가 계산되며 최종적인 블로그 랭킹이 이루어진다. 위키피디아 문서를 통해 확장된 질의는 초기 질의와 블로그 포스트를 통해 확장된 질의에 비해 in-depth 블로그에 높은 연관 점수를 부여할 수 있으며 이를 통해 in-depth 블로그의 검색 순위를 상위에 위치시킬 수 있다. 본 논문의 검색 모델은 언어 모델 방법에 기반하며, 문서 및 질의 언어 모델 추정 방법과 연관 점수 계산 방법은 아래와 같다.

3.2 문서 언어 모델 추정

문서(블로그 포스트 또는 위키피디아 문서) D의 언어 모델 $p(w|\theta_D)$는 maximum likelihood estimation과 Dirichlet prior smoothing[8]을 통해 다음과 같이 계산된다.

$$p(w|\theta_D) = \frac{c(w,D) + \mu p(w|C)}{|D| + \mu} \quad (1)$$

$c(w,D)$는 문서 D 내의 단어 w의 발생 횟수이며 $|D|$는 단어 기준 문서의 길이이다. $p(w|C)$는 $c(w,C)/|C|$로 추정되며 $c(w,C)$는 단어 w가 전체 문서 집합에서 나타나는 횟수, $|C|$는 전체 문서 집합 내에서 나타나는 총 단어 수로 정의된다. μ는 smoothing 파라미터로 본 논문에서는 500으로 설정하였다.

3.3 문서의 연관 점수 계산

질의 Q에 대한 문서 D의 연관 점수는 KL-divergence retrieval framework[9]을 통해 다음과 같이 계산된다.

$$Score(Q,D) = \sum_{w} p(w|\theta_Q) \log p(w|\theta_D) \quad (2)$$

$p(w|\theta_Q)$는 질의 언어 모델로 초기 질의의 경우 maximum likelihood estimation을 통해 $c(w,Q)/|Q|$로 추정된다. 질의 확장을 통한 $p(w|\theta_Q)$의 추정 방법은 아래와 같다.

3.4 질의 확장을 통한 질의 언어 모델 추정

질의 확장을 통한 질의 언어 모델 추정을 위해 [10]에서 제안된 모델 기반 피드백 방법 중 generative model 방법을 적용한다. 확장된 질의 언어 모델은 다음

과 같이 계산된다.

$$p(w|\theta_Q) = (1-\alpha)p_{ml}(w|\theta_Q) + \alpha p(w|\theta_F) \quad (3)$$

$p_{ml}(w|\theta_Q)$는 초기 질의에 대한 maximum likelihood estimation으로 계산된다. $p(w|\theta_F)$는 질의 확장을 위한 피드백 문서 집합 F ={$D_1, \ldots, D_k$}로부터 EM 알고리즘 [11]을 통해 단계적으로 추정된다.

$$z^{(n)}(w) = \frac{(1-\lambda)p^{(n)}(w|\theta_F)}{(1-\lambda)p^{(n)}(w|\theta_F) + \lambda p(w|C)} \quad (4)$$

$$p^{(n+1)}(w|\theta_F) = \frac{\sum_j c(w, D_j) z^{(n)}(w)}{\sum_i \sum_j c(w, D_j) z^{(n)}(w_i)} \quad (5)$$

λ는 피드백 문서 내의 단어가 $p(w|C)$에서 생성될 확률을 나타내며 본 논문에서는 0.8로 설정되었다. α는 $p(w|\theta_F)$의 가중치를 나타내며 본 논문에서는 0.6으로 설정되었다.

3.5 블로그의 연관 점수 계산

질의와 블로그간의 연관 점수를 계산하기 위해 [12]에서 제안된 모델과 설정값을 사용한다. 이 모델은 2008년 TREC 블로그 피드 검색에서 최고의 성능을 기록한 모델이다. 질의 Q에 대한 블로그 B의 연관 점수 $S(Q,B)$는 아래와 같이 계산된다.

$$S(Q,B) = \gamma S_G(Q,B) + (1-\gamma) S_L(Q,B) \quad (6)$$

$S_G(Q,B)$은 블로그 내의 모든 포스트와 질의간의 평균 연관 점수이며, $S_L(Q,B)$은 블로그 내 연관 점수 기준 상위 N개 포스트의 평균 연관 점수로 계산된다. 본 논문의 실험에서는 γ는 0.3, N은 2로 설정하였다. $S_G(Q,B)$를 통해 최종적인 블로그 랭킹이 결정된다.

4. 위키피디아 기반 질의 확장 방법

4.1 위키피디아 문서를 사용한 질의 확장

In-depth 블로그는 어떤 주제에 대한 깊이 있는 정보를 포함하기 때문에 해당 주제에 대한 전문 용어 또는 세부적인 개념을 설명하는 용어들을 포함할 가능성이 높다. 따라서 이러한 용어를 중심으로 질의를 확장하여 검색에 적용할 경우 in-depth 블로그에 높은 연관 점수를 부여하여 검색 성능을 향상시킬 수 있다.

기존의 블로그 피드 검색에서는 초기 질의와의 연관 점수를 기준으로 상위의 블로그 포스트들이 질의 확장을 위한 피드백 문서로 사용되었다. 하지만 상위의 포스트들이 주제와 연관되어 있더라도 해당 주제를 깊이 있게 다루지 않을 경우 in-depth 블로그 피드 검색에 필요한 용어들을 효과적으로 확보하기 어려워진다.

위키피디아 문서를 질의 확장에 사용할 경우 이러한 문제를 해결할 수 있다. 위키피디아는 사용자 참여 방식의 온라인 백과사전으로 다양한 방면의 주제에 대한 깊이 있고 분석적인 정보를 포함한다. 따라서 주제와 연관된 위키피디아 문서를 질의 확장에 사용할 경우 in-depth 블로그 검색에 필요한 용어들을 효과적으로 얻을 수 있으며 이를 검색 과정에 적용하여 in-depth 블로그 피드 검색 성능을 향상시킬 수 있다.

4.2 피드백 문서 선택 방법

본 논문에서는 다음의 두 가지 피드백 문서 선택 방법을 위키피디아 기반 질의 확장 과정에 적용한다.

4.2.1 연관 점수 기준 선택 방법

초기 질의와의 연관 점수 기준 상위 K개의 문서를 선택한다. 이는 정보 검색 분야에서 일반적으로 사용되는 피드백 문서 선택 방법과 동일한 방법이다.

4.2.2 문서 제목 기준 선택 방법

위키피디아 문서의 제목은 블로그 문서의 제목과 달리 문서가 다루고 있는 주제를 명확하게 나타낸다. 따라서 질의와 제목간의 일치 여부를 통해 주제를 다루는 정확한 하나의 문서를 선택할 수 있다. 또한 위키피디아는 하나의 문서를 통해 해당 주제에 대한 모든 내용을 다루기 때문에 선택된 문서를 통해 질의 확장에 필요한 충분한 양의 정보를 얻을 수 있다.

하지만 문서마다 제목은 하나씩만 존재하기 때문에 같은 주제를 나타내지만 형태가 다른 질의에 대해서는 제목 기준 선택 방법을 적용할 수 없다. 이 문제를 해결하기 위해 위키피디아의 redirect 페이지의 정보를 이용할 수 있다[13]. Redirect 페이지는 다양한 형태의 질의와 실제 내용을 다루는 문서의 제목간의 연결 정보를 포함한다. 예를 들어 "친환경 건축물"이라는 주제에 대해 위키피디아는 "Green building"이라는 제목의 문서에서 내용을 기술하며 "Green construction", "Sustainable building", "Green buildings", "Green-building"이라는 제목의 redirect 페이지를 통해 해당 주제에 대한 내용을 "Green building" 문서에서 다루고 있음을 표시한다. Redirect 페이지를 통해 다양한 형태의 질의와 실제 문서 제목간의 대응 테이블을 만들 수 있으며 이를 이용하여 질의와 위키피디아 문서 제목간의 일치 확률을 높일 수 있다.

5. 실험 및 결과

5.1 실험 데이터

본 논문의 실험을 위해 다음과 같은 데이터가 사용되었다.

5.1.1 블로그 데이터

2009년 TREC 블로그 트랙에서 적용된 Blogs08 collection[3]을 사용하였다. 이 데이터 집합은 1,303,520개의 블로그와 28,488,766개의 포스트를 포함하고 있다.

Permalink 페이지를 실험에 사용하였으며 모든 문서는 HTML 태그 제거 후 Porter stemmer로 stemming 되었다.

5.1.2 영문 위키피디아 데이터

위키피디아 다운로드 페이지(http://download.wikipedia.org)에서 제공하는 2010년 3월 12일자 영문 위키피디아 데이터를 사용하였다. 블로그 데이터와 마찬가지로 모든 문서는 XML 태그 제거 후 Porter stemmer로 stemming 되었다.

5.1.3 질의 및 평가 데이터

2009년 TREC 블로그 트랙의 Facet blog distillation 질의 및 평가 데이터 39개 중 in-depth 블로그 피드 검색과 관련된 18개를 사용하였다. 질의 및 평가 데이터를 통해 'in-depth' 블로그와 'shallow' 블로그에 대한 각각의 검색 성능을 평가할 수 있다. 성능 평가 척도로는 Mean Average Precision(MAP), Precision at 10(P@10), normalized Discounted Cumulative Gain(nDCG)[14]이 적용되었다.

5.2 질의 확장 방법에 따른 평가 대상 분류

본 논문에서는 in-depth 블로그 피드 검색에 다음의 질의 확장 방법을 적용하여 검색 성능을 분석한다.

① NO-FEEDBACK : 질의 확장을 사용하지 않는다.

② POST-TOP-K : 초기 질의에 대한 연관 점수 기준 상위 K개의 블로그 포스트를 질의 확장에 사용한다.

③ WIKI-TOP-K : 위키피디아 문서를 질의 확장에 사용하며 연관 점수 기준 선택 방법을 적용하여 K개의 문서를 사용한다.

④ WIKI-1-K : 위키피디아 문서를 질의 확장에 사용하며 문서 제목 기준 선택 방법을 적용한다. 만약 질의와 일치하는 제목을 가진 문서가 없을 경우 연관 점수 기준 선택 방법을 적용하여 K개의 문서를 사용한다.

5.3 실험 결과 및 분석

표 1은 각 질의 확장 방법의 in-depth 블로그 피드 검색 성능을 나타낸다. 본 논문에서 제안한 위키피디아 기반 질의 확장 방법은 기존의 블로그 포스트 기반 질의 확장 방법에 비해 in-depth 블로그 피드 검색 성능을 큰 폭으로 향상시키는 것을 알 수 있다. WIKI-TOP-10은 NO-FEEDBACK과 POST-TOP-10에 비해 MAP을 기준으로 각각 4%와 3%에 가까운 성능 향상을 보여주었다. 이는 위키피디아 문서가 일반 블로그 포스트에 비해 in-depth 블로그 피드 검색을 위한 질의 확장에 효과적임을 의미한다.

WIKI-1-10은 MAP을 기준으로 WIKI-TOP-10의 성능을 다시 3% 가까이 향상시킨다. 이를 통해 위키피디아 문서를 질의 확장에 사용할 경우, 연관 점수를 기준으로 다수의 문서를 선택하는 것 보다 질의와 일치하는 제목을 지닌 단 하나의 문서만을 선택하는 것이 in-depth 블로그 피드 검색에 효과적이라는 것을 알 수 있다. 블로그 포스트와 달리 위키피디아 문서의 제목은 해당 문서가 나타내는 주제를 정확하게 표현하기 때문에 피드백 문서 선택을 위한 중요한 자질로 사용될 수 있다. 또한 위키피디아 문서는 하나의 문서 내에서 해당 주제와 관련된 모든 내용을 다루기 때문에 제목을 통해 선택된 하나의 문서를 통해 질의 확장에 필요한 충분한 양의 정보를 효과적으로 얻을 수 있다.

표 2는 각 질의 확장 방법의 shallow 블로그에 대한 검색 성능을 나타내며, 표 3은 NO-FEEDBACK 기준 shallow MAP 대비 in-depth MAP의 상승 비율을 나타낸다. 블로그 포스트 기반 질의 확장 방법인 POST-TOP-10은 in-depth 블로그보다 shallow 블로그에 대한 검색 성능을 우선적으로 향상시키는 것을 볼 수 있다. 위키피디아 기반 질의 확장 방법의 경우 WIKI-1-10의 상승 비율이 WIKI-TOP-10보다 더 높은 것을 볼 수 있다. 이는 질의 확장에 정확하게 연관되는 하나의 문서만을 사용함으로써 주제와 연관성이 없는 용어를 배제하고 주제와 연관된 더 많은 전문 용어 및 세부 개념을 설명하는 용어들을 질의 확장에 포함하여 in-depth 블로그의 검색 순위를 상승시키고 shallow 블로그의 검색 순위 상승을 상대적으로 억제시켰다고 볼 수 있다.

표 1 In-depth 블로그 기준 검색 성능 비교 (K=10)

Method	MAP	P@10	nDCG
NO-FEEDBACK	0.3354	0.2333	0.5727
POST-TOP-10	0.3446	0.2278	0.5753
WIKI-TOP-10	0.3734	0.2944	0.6033
WIKI-1-10	0.4022	0.3000	0.6266

표 2 Shallow 블로그 기준 검색 성능 비교 (K=10)

Method	MAP	P@10	nDCG
NO-FEEDBACK	0.1241	0.1333	0.3438
POST-TOP-10	0.1420	0.1333	0.3634
WIKI-TOP-10	0.1683	0.1444	0.3930
WIKI-1-10	0.1637	0.1500	0.3920

표 3 Shallow MAP 상승 대비 In-depth MAP 상승 비율 비교 (NO-FEEDBACK 기준, K=10)

Method	ΔIn-depth MAP/ΔShallow MAP
POST-TOP-10	0.5139
WIKI-TOP-10	0.8597
WIKI-1-10	1.6868

6. 결 론

본 논문에서는 질의로 주어진 주제를 깊이 있게 다루는 블로그 검색을 위한 위키피디아 기반 질의 확장 방법을 제안하였다. 실험 결과 위키피디아 기반 질의 확장 방법은 기존의 블로그 포스트 기반 질의 확장 방법에 비해 MAP을 기준으로 한 검색 성능을 최대 6% 가까이 향상시켰다. 제안된 방법은 검색 성능을 큰 폭으로 향상시킬 수 있다는 점과 구현이 비교적 간단하다는 점에서 주제를 깊이 있게 다루는 블로그 검색을 위해 유용하게 사용될 수 있을 것으로 생각된다. 앞으로의 연구에서는 질의 확장 외에 검색 성능을 향상시킬 수 있는 다른 방법 및 자질에 대한 연구와 더 많은 질의 및 평가 데이터를 통한 실험이 이루어져야 할 것이다.

참 고 문 헌

[1] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of TREC-2007 Blog track," *in Proc. of TREC-2007*, 2008.

[2] I. Ounis, C. Macdonald, and I. Soboroff, "Overview of TREC-2008 Blog track," *in Proc. of TREC-2008*, 2009.

[3] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of TREC-2009 Blog track," *in Proc. of TREC-2009*, 2010.

[4] S. LI, H. Gao, H. Sun, F. Chen, O. Feng, S. Gao, H. Zhang, X. Li, C. Tan, W. Xu, G. Chen, and J. Guo, "A Study of Faceted Blog Disilltation - PRIS at TREC 2009 Blog Track," *in Proc. of TREC-2009*, 2010.

[5] M. Keikha, M. Carman, R. Gwadera, S. Gerani, I. Markov, G. Inches, A. A. Alidin, and F. Crestani, "University of Lugano at TREC 2009 Blog Track," *in Proc. of TREC-2009*, 2010.

[6] P. Jiang, Q. Yang, C. Zhang, and Z. Niu, "BIT at TREC 2009 Faceted Blog Distillation Task," *in Proc. of TREC-2009*, 2010.

[7] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, R. L. T. Santos, "University of Glasgow at TREC 2009: Experiments with Terrier," *in Proc. of TREC-2009*, 2010.

[8] C. Zhai, J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. on Inf. Syst.*, vol.22, no.2, pp.179-214, April, 2004.

[9] J. Lafferty, C. Zhai, "Document language models, query models, and risk minimization for information retrieval," *in Proc. of the 24th ACM Annl. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2001, pp.111-119.

[10] C. Zhai, J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," *in Proc. of the 10th ACM Conf. on Information and knowledge management*, 2001, pp.403-410.

[11] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of Royal Statist. Soc. B*, vol.39, no.1, pp.1-38, 1977.

[12] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, J.-H. Lee, "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval," *in Proc. of TREC-2008*, 2009.

[13] Y. Xu, G. J. F. Jones, B. Wang, "Query Dependent Pseudo-Relevance Feedback based on Wikipedia," *in Proc. of the 32nd ACM Annl. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2009, pp.59-66.

[14] K. Järvelin, J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. on Inf. Syst.*, vol.20, no.4, pp.422-446, Oct, 2002.