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

송우상○, 이예하, 이종혁
포항공과대학교 컴퓨터공학과
woosang@postech.ac.kr, sion@postech.ac.kr, jhlee@postech.ac.kr

A Wikipedia-based Feedback Method for In-depth Blog Distillation

Woosang Song○, Yeha Lee, Jong-Hyeok Lee
POSTECH, Department of Computer Science and Engineering

1. 서론

블로그 피드 검색(blog distillation)은 질의로 주어진 특정 주제를 중점적이며 반복적으로 다루는 블로그를 찾는 것을 목적으로 한다. 블로그 피드 검색에 관한 기존의 연구에서는 질의와 블로그 간의 연관 여부만을 고려하였으나, 최근에는 단순한 연관 여부를 넘어 질의로 주어진 주제를 깊이 있게 다루는 블로그(in-depth blog)를 찾기 위한 연구(in-depth 블로그 피드 검색, in-depth blog distillation)가 진행되고 있다[1].

본 논문에서는 위키피디아 문서를 질의 확장을 위한 피드백 문서로 사용하여 in-depth 블로그 피드 검색 성능을 향상시키는 방법을 제안한다. 위키피디아는 다양한 방면의 주제에 대해 전문적이며 분석적인 내용을 담고 있다. 따라서 질의와 연관된 위키피디아 문서를 질의 확장에 사용할 경우 질의에 해당하는 주제와 연관된 세부적이며 전문적인 용어를 중심으로 질의를 확장할 수 있으며, 이렇게 확장된 질의를 통해 in-depth 블로그에 높은 연관 점수(relevance score)를 부여하여 검색 성능의 향상을 기대할 수 있다.

2. 본론

사용자를 통해 입력된 질의는 사용자가 검색하고자 하는 주제에 대한 자세한 내용을 담고 있지 않기 때문에, 질의와 관련성이 높은 문서를 통해 질의를 확장하여 검색 성능을 향상시키는 방법이 사용된다. 일반적인 블로그 피드 검색에서는 최초 검색 결과 상위 N개의 포스트들이 질의 확장을 위해 사용된다.

In-depth 블로그는 어떤 주제에 대한 깊이 있는 정보를 기술하고 있기 때문에 이러한 블로그를 효과적으로 검색하기 위해서는 주제와 연관된 세부적이며 전문적인 용어를 중심으로 질의를 확장할 필요가 있다. 하지만 최초 검색 결과 상위 포스트들은 주제에 대한 깊이 있는 정보를 기술하고 있다는 보장이 없으며 이를 통한 질의 확장은 주제에 대한 세부적이며 전문적인 용어를 효과적으로 포함할 수 없다.

본 논문에서는 in-depth 블로그 피드 검색 성능 향상을 위해 위키피디아 문서를 사용한 질의 확장 방법을 제안한다. 위키피디아는 사용자 참여 방식의 인터넷 백과사전으로 다양한 방면의 주제에 대한 깊이 있고 분석적인 내용을 담고 있다. 따라서 질의와 일치하는 또는 관련성이 높은 위키피디아 문서를 질의 확장에 사용할 경우 주제와 밀접한 관련이 있는 세부적이며 전문적인 용어를 중심으로 질의를 확장할 수 있다. 이렇게 확장된 질의를 블로그 피드 검색 모델에 적용시 in-depth 블로그에 대해 높은 점수를 부여하여 상위 랭크에 위치시킬 수 있다.

위키피디아 문서 중 질의 확장을 위해 사용될 피드백 문서를 선택하는 방법은 두 가지로 나뉠 수 있다. 첫 번째 방법은 사용자가 입력한 질의를 통해 위키피디아 문서를 검색한 후 연관도 기준 상위 K개의 문서를 선택하는 것이다. 이는 일반적인 질의 확장 방법과 동일한 방법이다. 두 번째 방법은 질의와 위키피디아 문서의 제목을 비교하여 일치하는 하나의 문서를 질의 확장에 사용하는 방법이다. 위키피디아는 어떤 주제에 대해 하나의 문서를 할당하여 해당 주제에 대한 내용을 집중적으로 다루며 제목을 통해 문서가 다루는 주제를 나타낸다. 따라서, 제목과의 일치 여부로 선택된 하나의 문서는 질의 확장을 위한 충분한 정보를 담고 있다고 볼 수 있다. 또한, 다수의 문서를 질의 확장에 사용할 경우 발생하는 검색 성능 저하 문제도 해결할 수 있다.

사용자로부터 입력되는 다양한 형태의 질의와 위키피디아 문서 제목간의 일치 확률을 높이기 위해 redirect 페이지의 정보를 이용한다. 위키피디아는 다양한 형태의 질의에 대해 각각 redirect 페이지를 형성하여 해당 주제를 실제로 다루는 페이지와 대응시킨다. 예를 들어 “Podcast"라는 주제에 대해 위키피디아는 "Podcast”라는 제목의 페이지에 해당 주제와 관련된 내용을 기술하며 "Podcasting", "Podcasts", "Pod cast"라는 제목의 redirect 페이지를 만들어 “Podcast” 페이지에서 관련 내용을 다루고 있음을 표시한다. Redirect 페이지를 통해 다양한 형태의 질의와 관련 내용을 다루는 페이지의 제목간의 대응 테이블을 만들 수 있으며 이를 이용하여 질의와 위키피디아 문서 제목간의 일치 확률을 높일 수 있다.

본 논문에서 사용된 블로그 피드 검색 시스템은 [2]에서 제안된 모델을 블로그 피드 검색 모델로 사용한다. 개별 문서(블로그 포스트 또는 위키피디아 문서)와 질의 간의 연관도는 [3]에서 제안된 언어 모델의 KL Divergence Framework을 통해 계산되며, 문서 언어 모델은 최대 우도 추정법(Maximum Likelihood Estimation)과 Dirichlet Prior Smoothing[4]을 통해 추정된다. 질의 언어 모델 추정에는 [5]에서 제안된 모델 기반 피드백 방법 중 생성 모델 방법이 적용된다.

본 논문에서 제안한 방법을 실험하기 위해 2009년 TREC 블로그 트랙에서 적용된 BLOGS08 COLLECTION[1]과 2010년 3월 12일자 영문 위키피디아 데이터가 사용되었다. 또한, in-depth 블로그 피드 검색의 성능 평가를 위해 2009년 TREC 블로그 트랙의 Facet Blog Distillation[1]의 질의 및 평가 데이터 39개 중 In-depth 블로그 피드 검색과 관련된 18개를 사용하였다. 성능 평가 척도로는 MAP(Mean Average Precision), P@10(Precision at 10), NDCG(Normalized Discounted Cumulative Gain)[6]를 사용하였다.

표 1. In-depth 블로그 피드 검색 성능

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 1을 통해 각 질의 확장 방법이 in-depth 블로그 피드 검색 성능을 얼마나 향상시키는지 비교할 수 있다. NO-FEEDBACK은 질의 확장을 사용하지 않은 경우이며 POST-TOP-10은 초기 검색 결과 상위 10개의 포스트를 사용한 질의 확장 방법이다. WIKI-TOP-10은 본 논문에서 제안한 위키피디아 기반 질의 확장 방법 중 첫 번째 방법을 사용하며, 질의와의 연관도 기준 상위 10개의 위키피디아 문서를 사용한 질의 확장 방법이다. WIKI-TOP-10은 MAP을 기준으로 NO-FEEDBACK에 비해 4%, POST-TOP-10에 비해 3%에 가까운 큰 폭의 성능 향상을 기록하며 위키피디아 문서를 이용한 질의 확장 방법이 in-depth 블로그 피드 검색에 효과적임을 보여 준다.

WIKI-1-10은 본 논문에서 제안한 두 번째 방법을 사용하며 질의와 제목간의 일치여부를 통해 질의 확장에 사용될 문서를 선택한다. 만약, 일치하는 문서가 없을 경우 WIKI-TOP-10과 마찬가지로 연관도 기준 상위 10개의 위키피디아 문서를 선택한다. WIKI-1-10은 MAP을 기준으로 WIKI-TOP-10의 성능에서 다시 3%에 가까운 성능을 향상시킨다. 이를 통해 위키피디아 문서를 질의 확장에 사용할 때, 연관도를 기준으로 다수의 문서를 선택하는 것 보다 질의와 일치하는 제목을 지닌 단 하나의 문서만을 피드백 문서로 사용하는 것이 In-depth 블로그 피드 검색에 효과적이라는 사실을 알 수 있다.

3. 결론

본 논문에서는 in-depth 블로그 피드 검색 성능 향상을 위해 위키피디아 문서를 사용한 질의 확장 방법을 제안하였다. 위키피디아 문서를 질의 확장에 사용할 경우 주제와 밀접히 연관된 전문 용어들을 확장된 질의에 많이 포함시킬 수 있으며, 이렇게 확장된 질의를 통해 in-depth 블로그에 높은 연관 점수를 부여하여 검색 성능을 향상시킬 수 있다. 실험 결과를 통해 본 논문에서 제안한 방법이 in-depth 블로그 피드 검색 성능을 큰 폭으로 향상시키는 것을 보였다.

앞으로의 연구에서는 질의 확장 방법 외에 in-depth 블로그 피드 검색 성능을 향상시키기 위한 다양한 방법 및 자질에 대한 연구가 필요하다. 블로그 포스트의 길이나 블로그 저자에 대한 정보 등이 좋은 자질이 될 수 있을 것이다. 또한, 지금보다 더 많은 질의 및 평가 데이터를 통한 실험이 이루어져야 할 것이다.

감사의 글

본 논문은 2010년도 두뇌한국21사업, 포항공과대학교 정보통신연구소 자체 학술연구과제(선도과제), 그리고 한국과학재단 기초연구사업(No. 2010-0012662)의 지원으로 수행되었습니다.

참고문헌

[1] Craig Macdonald, Iadh Ounis, Ian Soboroff, "Overview of the TREC-2009 Blog Track", http://www.dcs.gla.ac.uk/~craigm/publications/blogOverview2009.pdf.

[2] Yeha Lee, Seung-hoon Na, Jungi Kim, Sang-hyob Nam, Hun-young Jung, Jong-hyeok Lee, "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval", *Proceedings of TREC 2008*, 2008.

[3] John Lafferty, Chengxiang Zhai, "Document language models, query models, and risk minimization for information retrieval", *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, Pages 111-119, 2001.

[4] Chengxiang Zhai, John Lafferty, "A study of smoothing methods for language models applied to information retrieval", *ACM Transactions on Information Systems*, 22(2):179-214, 2004.

[5] Chengxiang Zhai, John Lafferty, "Model-based feedback in the language modeling approach to information retrieval", *Proceedings of the tenth international conference on Information and knowledge management*, Pages 403-410, 2001.

[6] Kalervo Järvelin, Jaana Kekäläinen, "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems*, 20(4):422-446, 2002.