모바일 통합 검색에서 온톨로지를 활용한 컬렉션 순서의 정확도 향상

(Enhancing Accuracy of Collection Order Using Ontology in Mobile Aggregated Search)

송 효 진 † 이 태 휘 † 김 형 주 ††
(Hyojin Song) (Taewhi Lee) (Hyoung-Joo Kim)

요 약 통합 검색은 단일 인터페이스 내에서 여러 정보 소스를 동시에 검색하여 그 결과를 모아서 보여주는 검색 방법이다. 구글, 야후, 네이버, 다음 등 대부분의 국내외 메이저 포털 사이트들은 통합 검색을 도입하여 서비스하고 있다. 통합 검색 결과 내에서 각 컬렉션(뉴스, 블로그, 이미지 등)의 순서는 사용자가 빠르게 원하는 정보를 찾아내는 데 매우 중요한 역할을 한다. 특히 모바일 환경에서는 PC 환경에 비해 작은 화면과 조작의 불편함이 수반되기 때문에 컬렉션 순서의 영향력은 더 증대된다. 검색 결과 내의 컬렉션 순서는 클릭 로그를 기반으로 결정되는데, 클릭 로그가 충분하지 못한 대다수의 롱테일 키워드에 대해서는 효과적으로 컬렉션 순서를 정렬하지 못한다. 본 논문에서는 온톨로지를 활용하여 의미적 관련이 있는 다른 키워드의 클릭 로그를 참조함으로써 검색 결과를 개선하는 프레임워크를 제안한다. 질의 축소기법을 적용하여 복수 키워드 질의의 경우를 효과적으로 처리하고, 추가적으로 사용자 프로파일을 활용하여 컬렉션 순서의 정확도를 높인다. 실험을 통해 제안한 프레임워크가 검색 결과의 개선과 모바일 환경에서 네트워크 트래픽 감소 효과가 있음을 확인한다.

키워드 : 온톨로지, 통합 검색, 모바일 검색, 정보 검색

Abstract Aggregated search is a technique to search and assemble information from a variety of sources, within a single interface. This method has been adopted in most domestic and foreign portal sites such as Google, Yahoo, Naver and Daum. How to order the collections(news, blogs, images, etc.) in the aggregated search result is very important for finding relevant information quickly. In mobile environment, the influence of collection order is considerably enlarged due to small screen and limited interface. The collection order is usually decided based on click logs, but it is not effective for most long-tail keywords which have not enough click logs. In this paper, we propose a new framework to enhance the aggregated search result by referencing the click logs of semantically related keywords using an ontology. We apply query reduction methods into collection ordering to efficiently process multi-keyword queries. In addition, we exploit user profiles to improve the precision of the collection order. The experiments show that our framework enhances aggregated search results and reduces network traffic in mobile environment.

Key words : Ontology, Aggregated Search, Mobile Search, Information Retrieval

· 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20110017480). 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
songhj@idb.snu.ac.kr
twlee@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2011년 7월 19일
심사완료 : 2011년 12월 20일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제18권 제3호(2012.3)

1. 서 론

웹 2.0과 소셜 네트워크 서비스가 널리 사용되면서 사용자들은 다양한 컨텐츠를 생산하게 되었다. 스마트폰, 디지털 카메라, 태블릿 등 각종 휴대기기의 발전으로 사진, 동영상 등 여러 형태의 데이터를 생성하고 공유하기가 용이해졌다. 사용자들은 한 주제에 대한 여러 가지 정보를 동시에 검색하기도 하고, 함께 융합하여 새로운 컨텐츠를 재생산하기도 한다. 상용 검색 서비스들은 이러한 흐름에 맞추어 검색 알고리즘이나 정책 등을 계속 변화시키고 있다. 이러한 변화들 중 한 가지 추세가 바로 통합 검색(aggregated search)이다.

통합 검색은 그림 1과 같이 단일 인터페이스 내에서 여러 정보 소스를 동시에 검색하여 그 결과를 모아서 보여주는 검색 방법으로, 포털 사이트뿐만 아니라 기업 내의 자원 관리, 디지털 도서관 등 다양한 분야에서 활용되고 있다. 통합 검색 결과 내에서 한 구역을 차지하는 동일 분류의 정보 소스들을 컬렉션이라 한다. 2000년도부터 통합 검색을 서비스한 네이버는 현재 50개가 넘는 컬렉션(뉴스, 블로그, 이미지 등)을 제공하고 있다. 통합 검색을 이용하는 경우, 사용자가 다양한 컬렉션을 동시에 검색할 수 있으므로 편리하고 효율적이다. 또한 검색 결과를 컬렉션 별로 분류하여 나타냄으로써 사용자가 원하는 정보를 접근하는 데 유용하다.

한편 스마트폰의 두드러진 약진과 더불어 모바일 환경에서의 검색 기술과 관련한 다양한 연구들이 활발하게 진행 중이다. 유명 IT 리서치 회사인 가트너에서 발표한 보고서[1]에 의하면, 2012년에 각광받을 것으로 예상되는 모바일 기술 부문의 3위로 모바일 검색으로 꼽았다. 실제로, PC에서의 검색 빈도수를 모바일 환경에서의 검색 빈도수가 바짝 추격하고 있는 실정이다.

하지만 모바일 환경에서는 기존의 PC환경에 비해 화면이 작고 조작이 불편하다는 제약이 있다. 그림 1과 동일한 검색 결과가 스마트폰에서는 그림 2와 같이 표시되는데, 한 화면에 많은 결과를 보여주지 못하고 스크롤이 PC보다 훨씬 길어진다. [2]에서 제시한 통계와 같이, 58%의 사용자가 첫 번째 검색 결과 페이지 내의 결과만 선택하기 때문에 모바일 환경에서는 검색 결과의 품질이 더욱 중요해진다.

이러한 맥락에서 통합 검색의 컬렉션 순서는 매우 중요한 요소이다. 보다 정확한 컬렉션 순서는 사용자가 원하는 정보를 더 편하고 빠르게 찾도록 해준다. 검색 결과 내의 컬렉션 순서는 질의 키워드에 따라 변하는데, 대개 클릭 로그를 기반으로 정해지며 해당 질의 키워드에 대해 더 많이 클릭된 컬렉션이 상위에 표시된다. 클릭 로그를 사용하는 이유는 키워드 검색의 한계로 인해 데이터의 형태에 따라 컬렉션 별로 검색의 랭킹 척도 값의 범위가 달라지기 때문이다. 예를 들어, 동영상, 사

그림 1 통합 검색 화면

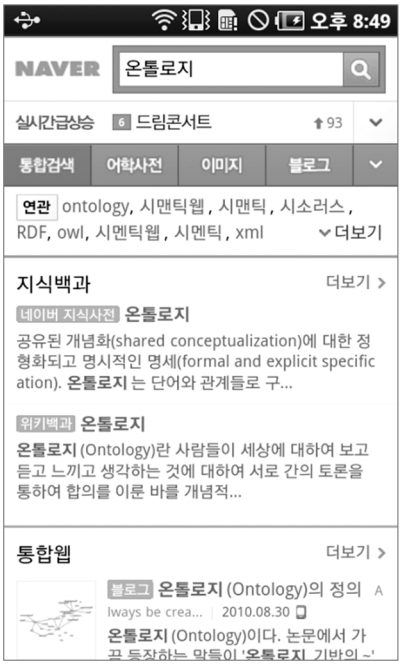


그림 2 모바일 환경에서의 통합 검색 화면

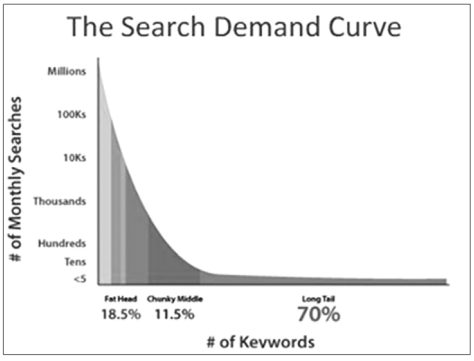


그림 3 검색 키워드의 롱테일 분포

진 등의 데이터에는 제목이나 요약 설명에만 키워드를 매칭할 수 있다. 하지만 이 방법은 클릭 로그가 별로 없는 키워드에 대해서는 효과적으로 컬렉션 순서를 정하지 못하는 단점을 지닌다.

질의 키워드는 그림 3과 같은 롱테일 분포를 나타내는데[3], 그래프의 왼편을 보면 인기 있는 키워드는 한 달에 검색되는 빈도수가 수백만에 달하지만 오른편의 가늘고 긴 부분인 롱테일 부분은 10회 이하인 것을 확인할 수 있다. 그러므로 전체 키워드 중 70% 이상을 차지하는 롱테일 키워드들은 활용할 클릭 로그가 충분치 못하기 때문에, 클릭 로그만을 참조할 경우에는 컬렉션 정렬이 제대로 이루어지지 않게 된다. 이러한 롱테일 키워드들은 검색이 자주 발생하지는 않지만, 그렇다고 무시할 만한 것이 아니다. 롱테일 키워드에 대한 검색 결과의 질이 떨어진다는 것은 각 사용자들이 서로 다른 롱테일 키워드로 나쁜 검색 결과를 경험하게 된다는 의미이기 때문이다. Ziv Bar-Yossef의 연구[4]에 따르면, 실제 질의의 37%정도는 이전에 발생하지 않은 새로운 질의이며, 롱테일 질의의 경우에는 검색 결과를 통해 방문한 페이지 수가 많다.

본 논문에서는 롱테일 키워드 질의가 들어왔을 때, 검색 결과 내 컬렉션 순서의 정확도가 떨어지는 현상에 주목한다. 그리고 클릭 로그를 기반으로 한 기존의 컬렉션 정렬 기법에 온톨로지를 활용하여 의미적 관련이 있는 다른 키워드의 클릭 로그를 참조함으로써 검색 결과를 개선하는 기법을 제안한다. 복수 키워드로 구성된 질의의 경우에는 질의 축소기법을 적용함으로써 핵심 키워드에 온톨로지를 활용하도록 하고, 추가적으로 사용자 프로파일을 이용해 검색 결과의 정확도를 높였다.

본 논문의 구성은 다음과 같다. 2장에서는 통합 검색, 모바일 검색, 의미적 유사성, 질의 축소 등 본 논문과 관련된 연구들을 소개하며, 3장에서는 제안한 프레임워크의 구조와 각 모듈을 실제 질의를 처리하는 과정에 따라 단계별로 서술한다. 4장에서는 실제 구현된 시스템을 통해 성능 평가를 실시하며, 그 결과 제안한 프레임워크를 통해 검색 결과의 개선과 모바일 환경에서 네트워크 트래픽 감소 효과가 있음을 확인한다. 마지막으로 5장에서는 결론과 향후 연구에 대해 설명한다.

2. 관련 연구

2.1 통합 검색

[5]는 대표적인 국내 포털 사이트인 네이버의 통합 검색에 대해 분석하였다. 질의가 수행되면 각 컬렉션을 로드하여 내부적인 알고리즘을 통해 가공하여 사용자에게 보여주는데, 이 과정에서 컬렉션 간의 랭킹과 각 컬렉션 내의 문서들의 랭킹을 계산한다. 구체적인 랭킹 기법은 기업 내부 보안 상 밝히지 않고 있으나, 컬렉션 간의 랭킹에는 문서의 클릭 로그, 각 컬렉션 내에서 유저가 머무른 시간, 컬렉션 내의 문서의 수 등의 정보들을 활용한다는 사실을 저자로부터 확인하였다.

[6]은 2008년에 구글에서 발표한 유니버설 검색(Universal Search)에 대한 연구 내용을 담고 있다. 유니버설 검색은 검색결과가 각각의 컬렉션 별로 구분되지 않고 한 페이지에 통합되어 나타난다. 즉 블로그, 동영상, 뉴스, 이미지, 웹 페이지 등 컬렉션 별로 검색 결과가 나뉘어 보이는 것이 아니라, 각 컬렉션 별로 검색결과가 하나의 페이지 안에 혼합하여 보여주는 방식이다. 구글은 검색 결과 랭킹에 문서 간의 하이퍼링크를 위주로

랭킹을 측정하는 페이지랭크(PageRank) 기법을 사용하므로 서로 다른 형태의 데이터를 통합하여 랭킹을 매기는 것이 가능하다. 서로 다른 컬렉션의 검색 결과를 통합하여 보여줄 것인지, 구분하여 보여줄 것인지는 또 다른 문제인데, 여러 형태의 데이터가 검색 결과에 뒤섞여 보여지면 사용자가 혼란을 겪을 수도 있기 때문이다.

[7]은 실제 통합검색 내에서 각 컬렉션의 사용자 선호도를 정량화 하고, 이를 활용하여 검색결과 내에 중요도에 따라 컬렉션을 배치시키는 다양한 방식을 소개하고 있다. 이 연구에서는 어떤 컬렉션을 선정하여 사용자에게 보여줄 것인지가 아니라, 컬렉션을 어떻게 배치하여 보여줄 것인지에 대한 문제를 다루고 있다. 가로, 세로 등 가능한 결과 배치 방법들 중 사용자와 질의 키워드의 조합에 따라 통계 정보와 확률 모델을 이용하여 최적의 배치를 선택하는 방법을 서술하였다.

2.2 모바일 검색

[8]의 연구에서는 모바일 검색이 가져야 할 성질에 대해 분석하였다. 모바일 검색은 PC와 달리 단말의 특성 및 사용자의 상황 정보를 고려하여 사용자가 언제, 어디서나 원하는 정보를 편리하게 찾을 수 있어야 하며, 다음의 두 가지 핵심 요소를 만족시켜야 한다고 서술하고 있다.

정확한 검색. 제한된 모바일 화면에 보다 정확하고 요약된 검색 결과를 제시해야 하며, 이를 위해서는 모바일 지식화, 지식 마이닝 및 요약, 질의응답(Q/A) 기술이 필요하다.

사용자 맞춤 검색 결과 제공. 모바일 사용자의 정보 및 행동 특성을 분석하여 사용자 맞춤형 검색을 제공해야 하며 이를 위해서는 사용자 최적화 서비스를 위한 모바일 개인화 검색 기술이 필요하다.

제안한 프레임워크에서는 정확한 검색을 위해 클릭 로그에 온톨로지를 활용하고 사용자 맞춤 검색 결과 제공을 위해 사용자 프로파일을 이용한다.

2.3 의미적 유사성

정보 검색, 상품 분류, 생물정보학 등 여러 분야에서 의미적 유사성(semantic similarity, semantic relatedness)에 대한 연구[9]가 이루어지고 있다. 의미적 유사성이란 개념과 개념 사이의 관계를 정의하는 도구인 온톨로지 상의 의미 관계 정보를 이용해 객체간의 유사성을 측정하는 것이다. 의미적 유사성은 어휘 유사성, 구조 유사성, 인스턴스 유사성, 추론 유사성으로 구분되는데[10], 본 연구에서는 질의 키워드와 유사한 의미를 지닌 단어를 찾기 위해 어휘 유사성과 구조 유사성을 이용한다.

어휘 유사성은 어휘의 동의어 집합을 이용하여 측정하는 방법이 대표적으로 WordNet[11]를 많이 활용한다. 구조 유사성은 어휘들간의 관계를 정의한 택소노미(taxonomy)와 온톨로지 상의 개념간의 제약조건을 이용하여 어휘 간의 구조를 비교하여 측정한다.

2.4 질의 축소

그림 4에서 보듯이, 사용자들은 평균 3.5개의 검색 키워드를 입력한다고 한다[12]. 질의를 구성하는 단어에 불필요한 단어가 포함되면 오히려 검색의 정확도가 떨어진다. [13]의 연구는 질의의 길이를 줄이는 질의 축소(query reduction)의 과정을 통해 검색의 정확도를 높이는 과정을 기술하고 있다. 예를 들면, “ideas for breakfast menu for a morning staff meeting”과 같은 긴 질의로 검색하는 것보다 “breakfast meeting menu ideas”와 같이 질의를 축소해서 검색하는 경우에 대부분의 검색 엔진에서 더 좋은 결과를 얻을 수 있다. [14]의 연구에서는 이러한 현상을 정량적으로 분석하여 검색 결과의 정확도가 평균적으로 약 30% 향상됨을 확인하였다.

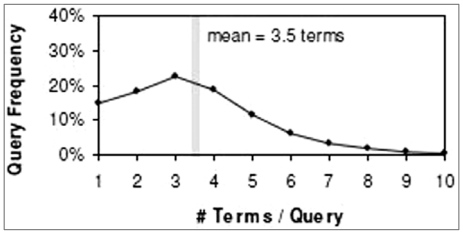


그림 4 질의의 평균 단어 수

이러한 질의 축소를 수행하는 여러 방법들이 있다. Bendersky[15]에서는 긴 질의문 내의 핵심 개념을 식별하기 위한 학습을 질의의 특성을 이용하여 수행한다. Lease의 연구[16]에서는 긴 질의문 내의 모든 단어들 각각에 가중치 재부여를 하는 방식을 통해 회귀 기반의 질의 축소를 수행한다.

대표적인 질의 축소 방법 중 한 가지는 질의 품질 예측기(query quality predictors)의 활용이다. 질의 품질 예측기란 복수의 단어들로 구성된 초기 질의문의 각 단어들의 조합을 구축한 뒤, 이를 이용해 가장 품질이 좋은 키워드 조합을 활용하여 랭킹 함수를 학습하는 기법을 말한다. 이에 대한 다양한 알고리즘이 있는데, 본 연구에서는 그 중 간결하면서도 본 개념에 충실한 상호 정보 척도(mutual information)[14]를 사용하여 복수 키워드로 구성된 질의를 축소하고, 이를 통해 얻은 키워드들의 클릭 로그를 참조하여 컬렉션의 순서 개선에 활용한다.

3. 컬렉션 순서 개선 기법

이 장에서는 여러 기법들을 활용하여 통합 검색의 컬렉션 순서 개선하는 기법을 소개하고자 한다. 먼저 전체

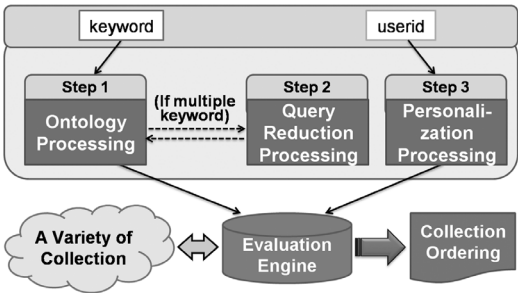


그림 5 전체 프레임워크 구조

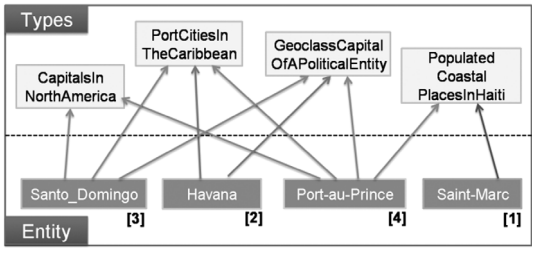


그림 6 "Port-au-Prince" 검색 시 처리되는 온톨로지 상의 개체와 의미 관계

프레임워크 구조를 소개한 뒤, 각 절에서 단계별 내용을 구체적으로 설명한다.

그림 5는 질의 키워드와 사용자 식별자를 입력 받은 후 컬렉션 순서의 정확도를 개선하는 과정을 보여주고 있다. 크게 온톨로지 활용, 질의 축소, 사용자 개인화의 세 부분으로 나뉘며, 모든 처리를 거친 후에 최종적으로 컬렉션 순서를 결정하는 데 반영한다.

3.1 온톨로지를 활용한 질의 키워드의 유사 단어 탐색

첫 단계에서는 질의 키워드가 롱테일 키워드인 경우 온톨로지의 의미 관계를 이용해 질의 키워드와 유사하며 충분한 클릭 로그를 보유한 다른 키워드를 탐색한다. 탐색 결과로 나온 키워드의 클릭 로그를 참조하여 롱테일 키워드의 컬렉션 순서를 결정하는 데 활용한다.

위와 같이 의미적으로 유사한 키워드의 클릭 로그를 참조하는 데 다음의 두 가지를 가정한다.

1) 키워드간의 의미적 유사성이 높을수록 컬렉션 순서가 유사하다.
2) 의미적 유사성 측정에 사용 가능한 온톨로지는 미리 구축되어 있고 계속 최신 정보로 갱신된다.

첫번째 가정은 비슷한 종류의 키워드로 검색하는 경우 컬렉션 순서도 비슷하다는 것이다. 예를 들어, 가수나 연예인을 검색하면 인물 정보나 동영상, 뉴스, 이미지 등이 상위에 표시되고, 지명을 검색하면 지도, 이미지 등이 먼저 상위에 표시된다. 두번째 가정은 활용할 최신 내용의 온톨로지가 미리 존재해야 한다는 것인데, 아직은 온톨로지가 보편적으로 사용되지는 않지만 사용이 활발해지면 각 분야의 전문가 집단이나 기관에서 대중보다 빠르게 온톨로지를 구축하여 배포할 것이다. 또는 검색 제공 사이트에서 검색의 정확도를 높이기 위해 스스로 구축하고 갱신하여 사용하는 방법도 있다.

그림 6은 예제로 아이티 수도의 이름인 "Port-au-Prince"라는 질의 키워드가 들어왔을 때, 이 키워드와 가장 유사한 단어를 찾는 과정에 필요한 온톨로지 상의 개체와 의미 관계의 일부를 나타낸 것이다. 본 연구에서는 대상 온톨로지로 Yago[17]를 사용하였으며 의미 관계로는 타입(Type) 속성을 이용하였다. 그림 6에서 질의 키워드(Port-au-Prince)는 4개의 타입을 가지며, 각 타입들을 공유하는 다른 단어들을 확인할 수 있다. 각 개체 아래에 [n]으로 표시한 부분이 각 단어가 질의 키워드와 공유하고 있는 타입의 수를 나타낸다. 이를 통해 본 예제에서는 가장 많은 3개의 타입을 공유하고 있는 "Santo Domingo"단어가 질의 키워드와 가장 유사하다고 판단할 수 있다.

이와 같은 원리로 질의 키워드와 유사한 단어들을 탐색하되, 다른 단어의 클릭 로그를 참조하는 것이 목적이므로 가장 유사하면서도 롱테일 키워드가 아닌 단어들을 탐색 결과로 선정한다. 결과 키워드들의 공통된 타입의 수가 동일한 경우에는 더 많은 클릭 로그를 지닌 단어를 채택하며, 롱테일 키워드밖에 없는 경우에는 참조하지 않는다.

알고리즘 1은 지금까지 설명한 내용을 바탕으로 온톨로지를 활용한 질의 키워드와 유사한 단어들을 선택하는 과정을 표현한 것이다. 질의 키워드 q와 온톨로지 O를 이용해서 최종적으로 질의 키워드와 유사한 단어들을 정렬하여 유사 키워드와 빈도수를 함께 구한다. 먼저 2-3라인에서는 질의 키워드의 타입, 그리고 각각의 타입들

알고리즘 1. 온톨로지를 활용한 질의 키워드의 유사 단어 탐색 알고리즘

Input: Query keyword q, Ontology O
Output: Ordered words Ws similar to query keyword (2-dim Vector)

```
1: Find the node of query in Ontology
2: L = the number of type of query
3: V = the vector with entities and freq (2-dim vector)
4: For 1≤n≤L
5:     Fetch words e in E using a type relationship
6:     if(the number of logs of e < threshold)
7:         continue    // it is long-tail keyword!
8:     if(e exists in V) then
9:         Save it to V with freq update
10:    else
11:        Save it to V with 0 as freq
12: end for
13: Ordering V along with freq
14: Output top-k e in V according to threshold value
```

Time Complexity: $O(LE+V^2)$
Space Complexity: O(LE)
(Usually E > V >> L ≒ 8)

을 공유하는 단어들과 그 빈도수를 저장할 벡터 변수를 생성한다. 이어서 질의 키워드의 각 타입을 공통으로 가지는 단어들을 가져온다. 6-7라인은 관련도가 높더라도 그 단어가 롱테일 키워드인 경우는 배제하는 부분을 처리한다. 8-11라인에서는 질의 키워드와 연관이 있는 단어들이 질의 키워드와 공유하는 타입의 수를 세어 벡터 변수에 저장한다. 이렇게 얻어낸 단어들을 13라인에서 질의 키워드와 공유하는 타입의 개수에 대해 정렬하고, 14라인에서 최종적으로 k개의 단어를 결과로 출력한다.

3.2 질의 키워드 축소

롱테일 키워드에는 자주 검색되지 않는 단일 키워드 외에 복수 키워드로 질의가 구성된 경우도 속한다. 복수 키워드의 경우에는 질의 키워드를 축소하여 핵심 단어를 추출한 후에 3.1절에서 설명한 온톨로지를 활용하여 의미적으로 유사한 키워드를 탐색한다. 질의를 축소하면 불필요한 단어를 제거함으로써 검색의 정확도를 높일 수 있으며, 온톨로지 상에서 의미적 유사 키워드를 탐색하는 비용도 감소된다.

본 연구에서는 2.4절에서 언급한 대로 상호 정보 척도(MI)를 적용하였다. 상호 정보 척도란 두 대상이 얼마나 서로 의존적인지를 나타내는 지표로써, 질의 키워드의 각 단어들 중에서 가장 연관도가 높은 단어들의 조합을 찾는 데 이용할 수 있다. 복수의 단어들로 구성된 초기 질의문으로부터 생성할 수 있는 각 단어들의 조합의 개수는 2n이다. (n은 질의문 내의 단어의 개수) 이 조합들에 대해 상호 정보 척도 수식을 적용하여 각 MI 값을 계산하여 그 값이 임계값(threshold) 이상인 단어 조합을 추출하여 활용한다.

상호 정보 척도의 수식은 아래와 같이 나타낼 수 있다.

$$MI(x,y) = log \frac{\frac{n(x,y)}{T}}{\frac{n(x)}{T}\frac{n(y)}{T}}$$

여기에서 n(x, y)는 단어 x와 y가 포함된 키워드로 질의했을 때 나오는 문서의 개수이며, n(x), n(y)는 각각의 단어로 질의했을 때 나오는 문서의 개수, T는 컬렉션 내의 모든 문서의 개수를 의미한다. MI 값은 다음과 같이 세 가지로 분류한다.

If MI(x,y) > 0 :	x와 y는 밀접한 상관관계를 가지고 있다 예를 들면 유의어, 반의어, 관련 용어, 또는 복합 명사 등이다.
If MI(x,y) < 0 :	x와 y는 아무 관계가 없다.
If MI(x,y) ≒ 0 :	x와 y는 대용어 관계, 즉 상호 베타적인 관계에 있다. 따라서 함께 쓰이기보다는 x가 쓰이면 y는 쓰이지 않고, y가 쓰이면 x가 쓰이지 않는다.

표 1 상호 정보 척도 계산 결과(질의 키워드: Port-au-Prince, earthquake, Korean, damage)

A subset of keyword	MI
Port-au-Prince+earthquake+damage	1.6099967920123643
Port-au-Prince+earthquake+Korean	1.0164984820072627
Port-au-Prince+Korean+damage	-5.1732773855419305E-5
earthquake+Korean+damage	-1.0819351451359904
Port-au-Prince+earthquake	0.8548662041206355
earthquake+damage	-0.162596283079913 58
Port-au-Prince+damage	-0.33895758364798145
earthquake+Korean	-1.0449396754925742
Korean+damage	-1.191705065124977
Port-au-Prince+Korean	-1.3166359199511886

표 1은 예제로 “Port-au-Prince”, “earthquake”, “Korean”, “damage”의 키워드로 질의하였을 때 MI 값을 계산한 결과이다. 가령, 표 1에서 양수 값이 나온 3개의 항목을 추출하여 각각의 클릭 로그를 참조하여 질의 키워드의 클릭 로그를 보충할 수 있다.

3.3 개인화를 통한 컬렉션 선호도 반영

검색 서비스에 개인화를 적용하여 사용자가 찾는 정보를 보다 정확하게 보여주기 위한 연구가 많이 수행되고 있다. 본 연구에서는 사용자의 컬렉션 선호도를 활용하여 개인화에 적용한다. 사용자 선호도는 다음 식에 따라 적용한다.

$$T = \sum_{i=1}^{n} C_i$$

$$Userpref_i = \alpha\left(1 + \frac{C_i}{T}\right)$$

C_i는 각 컬렉션의 클릭 로그 수, T는 모든 클릭 로그의 수를 나타낸다. 사용자의 컬렉션 선호도 $Userpref_i$는 사용자의 컬렉션 i에 대한 가중치를 의미하며, 클릭 로그를 정규화한 후 더해준다. 상수 α는 3.1절에서 설명한 질의 키워드의 클릭 로그에 비하여 사용자 선호도를 얼마나 반영할 것인지에 대한 비율로, 사용자 클릭 로그의 분포에 따라 α를 조정하여 사용자 선호도의 가중치를 높일 수 있다. 예를 들어 압도적으로 많은 비율로 클릭한 컬렉션의 경우 보다 높은 가중치를 주어 컬렉션 순서의 정확도를 향상시킬 수 있다. 본 논문의 실험에서는 온톨로지의 활용과 질의 축소가 더 큰 비중을 차지하고, 사용자의 컬렉션 선호도의 편차가 심하지 않아 α 값을 1로 두고 계산하였다.

클릭 로그는 표 2와 같은 구조를 지닌다. 가령, 표 2는 songhj라는 사용자가 각 컬렉션을 몇 회나 클릭하였는지를 나타낸다. 클릭 로그 테이블은 데이터베이스에 저장되며, 시스템은 질의 처리 시 이 값을 불러와서 컬렉션 정렬에 반영한다. 클릭 로그 값을 이용해서 위의

표 2 사용자의 컬렉션 선호도에 대한 클릭 로그 테이블

userid	collection	count
songhj	블로그	582
songhj	웹문서	216
songhj	뉴스	109
⋮	⋮	⋮

식으로 각 컬렉션의 사용자 선호도를 계산하여 결과 컬렉션의 순서에 반영한다.

4. 성능 평가

이 장에서는 사용자 실험을 통해 본 논문에서 제안한 기법이 기존 방법에 비하여 컬렉션 순서의 정확도를 얼마나 개선할 수 있는지 살펴본다. 그리고 이를 활용하여 모바일 검색에서 네트워크 트래픽의 절감 효과를 얻을 수 있음을 확인한다. 컬렉션 순서를 개선하기 위해 온톨로지 활용과 질의 축소에 추가적으로 소요되는 비용을 측정하고 분석해 본다.

4.1 실험 환경 및 실험 데이터

실험 환경은 질의를 처리해서 결과를 제공하는 검색 서버와 모바일 환경의 클라이언트의 두 가지로 나누어 진다. 검색 서버는 Intel Core 2 Duo E7400 2.80GHz의 CPU와 4GB의 RAM 사양의 시스템을 사용하였고, 운영체제로는 Microsoft사의 Windows 7 Professional K 32bit 버전을 사용하였다. 모바일 환경의 클라이언트는 ARM Cortex-A8 1GHz의 CPU와 512MB Ram, 운영체제로는 Android 2.2 Froyo 버전을 사용하는 삼성전자의 Galaxy S 제품을 사용하였다.

3장에서 설명한 컬렉션 개선 프레임워크 및 사용자 실험을 위한 웹 페이지가 검색 서버에 구현되었다. 컬렉션 개선 프레임워크는 JAVA 1.6.0_24 버전으로, 사용자 실험을 위한 웹 페이지는 JSP와 Servlet으로 구현되었다. 온톨로지의 저장 및 검색을 위해 Jena 2.6.4 버전을 사용하였으며, 클릭 로그와 사용자 정보를 저장하기 위한 데이터베이스로는 MySQL 5.0.32 버전을 사용하였다.

컬렉션 개선에 사용할 온톨로지로는 WordNet[11]과 DBpedia[18]의 온톨로지를 통합하여 지식 도메인을 확장한 Yago[17]를 선택하였다. Yago 온톨로지는 사람, 장소와 같은 개체(entity) 정보를 170만개 이상, 두 개체가 하나의 관계(relation)를 맺고 있는 트리플 단위의 사실(fact) 정보를 1,500만 개 이상 포함하고 있다.

실험에 사용한 컬렉션은 네이버 검색 엔진의 컬렉션 중 12개(뉴스, 블로그, 웹문서, 카페, 백과사전, 이미지, 책, 지식인, 쇼핑, 영화, 전문자료, 지역정보)를 선정하였고, 실험에 사용할 검색 키워드는 온톨로지의 적용을 위해 온톨로지 상에 존재하는 키워드들의 조합으로 한정하였다. 사용자 컬렉션 로그 데이터는 실험 전에 한 달여 동안 미리 수집하여 실험에 반영하였다.

4.2 성능 평가 방법

성능 평가는 네이버 검색 엔진에서 나타나는 컬렉션 순서를 비교 기준(baseline)으로 설정하고 본 논문에서 제안한 기법의 성능과 비교하였다. 2.1절에 언급한 대로 [5]의 연구에 따르면 네이버 검색 엔진은 클릭 로그와 각 컬렉션 내에서 머무른 시간, 그리고 각 컬렉션 별로 결과의 총 개수 등을 반영한다. 본 논문에서 제안한 기법은 온톨로지를 활용하여 유사 단어의 클릭 로그를 이용한 경우와 사용자의 컬렉션 선호도만을 반영한 경우, 이 두 기법을 함께 사용한 경우의 세 가지로 나누어 성능을 측정하였다.

성능 비교를 위한 척도로는 웹 검색 엔진 알고리즘의 성능 척도로 가장 널리 알려져 있는 NDCG[19]를 사용하였다. NDCG 척도는 검색 알고리즘의 결과가 이상적인 결과 순위에 근접할수록 더 높은 값을 갖게 되며, 상위 랭킹의 결과가 더 정확할수록 더 높은 값을 가진다. NDCG는 다음과 같은 방법으로 계산한다.

$$N_q = M_q \sum_{j=1}^{K} \frac{2^{(r(j))}-1}{log_2(1+j)}$$

질의 q에 대한 NDCG 값 N_q는 첫번째 결과부터 K번째 결과까지 얻는 점수의 합으로 계산된다. r(j)는 j번째 항목의 가중치를 나타내는 함수이다. NDCG에서는 상위 결과일수록 점수의 비중이 크다. 본 논문의 실험에서 r(j)는 K-j로 정의하여 계산하였다. 예를 들어, $M_q = 1$로 가정하고 총 3개의 컬렉션이 존재할 때 검색 결과에서 컬렉션을 이상적인 결과의 1, 2, 3위의 순서대로 정확히 보여준 경우, 1위 결과가 공헌하는 점수는 $(2^{(3-1)}-1)/\log_2 2 = 3$, 2위 결과가 공헌하는 점수는 $(2^{(3-2)}-1)/\log_2 3 = 0.6309$, 3위 결과가 공헌하는 점수는 $2^{(3-3)}-1/\log_2 4 = 0$이 되며, 총합은 3.6309가 된다. 한편, 검색 결과에서 컬렉션을 이상적인 결과의 역순인 3, 2, 1위의 순서대로 보여준 경우에는 1위 결과가 공헌하는 점수는 $(2^{(3-3)}-1)/\log_2 2 = 0$, 2위 결과가 공헌하는 점수는 $(2^{(3-2)}-1)/\log_2 3 = 0.6309$, 3위 결과가 공헌하는 점수는 $(2^{(3-1)}-1)/\log_2 4 = 1.5$가 되며, 총합은 2.1309가 된다. M_q는 정규화 상수로 위와 같이 계산한 척도 값의 범위를 [0, 1]로 조정하는 역할을 한다.

실험은 컴퓨터공학과 대학원생 12명을 대상으로 진행되었다. 4명씩 총 3그룹으로 분류하여 각 그룹별로 검색 키워드로 단일 키워드 10개를 할당했으며, 복수 키워드 검색 실험을 위해 할당 받은 단일 키워드들에 대해 각자 다른 키워드를 추가하여 최대 5개의 키워드를 검색하도록 하였다. 사용자는 검색한 키워드에 대해 기본 검색 결과를 보여주고 컬렉션의 우선 순위를 정하여 기록하게 하였다. 이를 취합하여 각 기법을 통해 얻은 컬렉

션 순서와 비교하였고, NDCG 척도를 통해 정량화하여 개선 여부를 측정하였다.

4.3 실험 결과

그림 7은 각 기법을 이용하여 얻은 단일 키워드 질의에 대한 검색 결과에 대해 상위 k 개 컬렉션의 NDCG 값을 나타낸 그래프이다. 온톨로지를 활용하여 유사 단어의 클릭 로그를 이용한 경우 컬렉션 순서의 정확도가 향상됨을 알 수 있으며, 사용자 선호도를 추가로 반영하였을 때 정확도가 조금 더 향상되는 것을 확인할 수 있다.

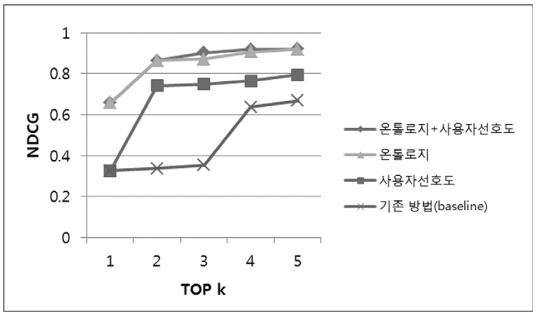
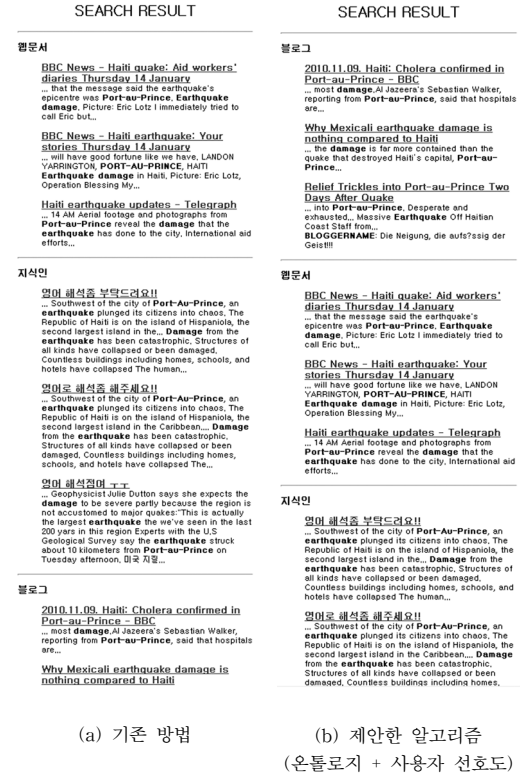


그림 7 상위 k개의 컬렉션에 대한 NDCG(단일 키워드 질의)



(a) 기존 방법 (b) 제안한 알고리즘 (온톨로지 + 사용자 선호도)

그림 8 예제 실험 검색 결과 화면(질의 키워드: Port-au-Prince, earthquake, damage)

표 3 예제 실험 검색 결과의 컬렉션 순서 변화

알고리즘 / 순위	기존 방법	온톨로지	사용자 선호도	온톨로지 + 사용자 선호도
1	웹 문서	웹 문서	블로그 (60%)	블로그
2	지식인	블로그	지식인 (10%)	웹 문서
3	블로그	지식인	카페 (10%)	지식인
4	뉴스	카페	뉴스 (10%)	카페
5	카페	뉴스	웹 문서 (3%)	뉴스

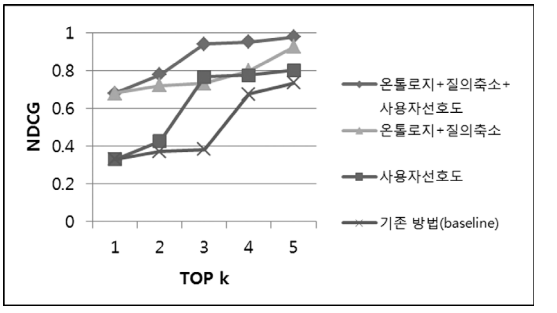


그림 9 상위 k개의 컬렉션에 대한 NDCG (복수 키워드 질의)

그림 8의 실험 검색 결과를 보면 실제로 컬렉션 순서가 어떻게 달라지는지 확인할 수 있다. 그림 8(a)는 기존 방법의 컬렉션 순서로 배치된 화면이고, 8(b)는 제안한 방법의 컬렉션 순서로 배치된 화면이다. 원래는 모바일 환경을 가정하나, 화면이 세로로 길기 때문에 편의상 가로를 모바일 환경 수준으로 좁게 하고 데스크톱 컴퓨터에서 캡쳐하였다. 컬렉션 순서의 단계별 변화 내용을 표 3에 정리하였다. 그림 8의 검색을 수행한 사용자는 블로그 (60%) >> 지식인 = 카페 = 뉴스 (10%) > 웹문서 (3%) 순으로 블로그 컬렉션의 선호도가 매우 높고, 온톨로지로 확장한 키워드에서도 블로그 컬렉션이 원래 순서보다 조금 높기 때문에 최종 순서에서 가장 먼저 나오게 되었다.

그림 9는 그림 7과 동일한 형태의 그래프를 복수 키워드 질의에 대해 나타낸 것이다. 복수 키워드 질의이기 때문에 질의 축소 기법이 적용되었으며, 온톨로지 활용과 질의 축소 기법, 사용자 선호도를 함께 반영한 경우의 정확도가 가장 높음을 확인할 수 있다.

특히, 단일 키워드와 복수 키워드의 경우 모두 기존 방법에 비해 최상위(top 1) 결과의 정확도가 크게 높아지며, 상위 3번째(top 3) 결과까지의 정확도가 월등히 개선됨을 관찰할 수 있다. 높아진 정확도를 바탕으로 하여 검색 결과의 전송량을 줄일 수 있을 것이다.

그림 10은 검색 결과로 상위 k개의 결과 컬렉션만 전달하는 경우의 누적 precision을 측정한 그래프이다. 본 논문에서 제안한 기법이 기존 기법보다 더 높은 누적

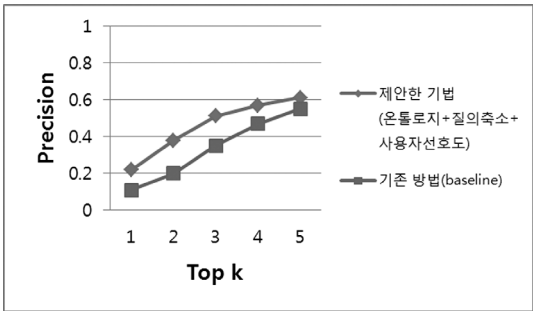


그림 10 상위 k개 컬렉션의 누적 precision

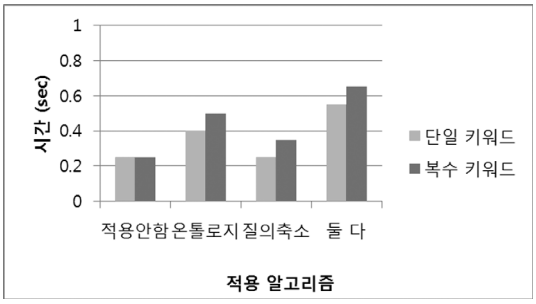


그림 11 컬렉션 정렬에 드는 시간 비용

precision 값을 가지며, 제안한 기법에서는 상위 3개의 컬렉션만 전달해도 사용자가 원하는 정보의 50% 이상이 검색되는 것을 확인할 수 있다. 상위 3개 이후 결과 컬렉션에서는 정확도가 향상되기는 하나 그 정도가 점차 낮아지고 있다. 따라서 네트워크 트래픽에 민감한 모바일 환경에서는 본 논문에서 제안한 기법을 사용하고 상위 3개 컬렉션 정도로 컬렉션을 한정하여 검색 결과를 전달함으로써 네트워크 트래픽을 감소시킬 수 있을 것이다.

하지만 제안한 기법은 높은 정확도에 비해 추가로 처리하는 데 비용이 소요된다. 그림 11은 제안한 기법을 수행하는 데 드는 시간 비용을 나타낸 것이다. 사용자 선호도 로그는 데이터베이스에서 select 질의 한 번으로 쉽게 얻을 수 있으므로 제외하고, 온톨로지를 활용한 유사 단어 탐색과 복수 키워드의 질의 축소 두 가지만 나타내었다. 큰 비용이 드는 부분은 온톨로지를 활용하는 과정으로, 유사 단어를 찾기 위해 SPARQL 질의를 처리하면서 상당한 양의 시간을 소모한다. 본 실험에서는 온톨로지의 규모 상 처리 시간이 최대 0.6초 정도로 오래 걸리지 않아 문제가 되지 않았지만, 더 큰 규모의 온톨로지를 이용하는 경우 처리 시간이 문제가 될 수 있다. 그러한 경우에는 온톨로지 분할 기법 등과 함께 사용하여 문제를 해결해야 할 것으로 보인다. 질의 축소의 경우에는 복수 키워드의 경우에만 해당되는데, 질의 키워드 내의 단어 수는 그림 4의 그래프에서 보듯이 대부분 5개 이하이기 때문에 온톨로지 활용에 비해 상대적으로 많은 비용을 발생시키지는 않는다.

5. 결론 및 향후 연구

본 논문에서는 모바일 통합 검색에서 롱테일 키워드로 인한 문제점을 해결하기 위한 개선 방안을 제안하였다. 이를 통해 검색 결과 화면에서 컬렉션을 보다 효과적으로 정렬함으로써 사용자가 빠르고 편하게 원하는 정보를 찾도록 해준다. 롱테일 키워드의 부족한 클릭 로그를 온톨로지와 질의 축소 기법을 활용하여 유사한 의미의 단어들의 클릭 로그를 이용하여 보충하였다. 또한 사용자의 컬렉션 선호도를 반영하여 검색 결과의 정확도를 높였다. 본 논문에서 제안한 기법들이 계속 증가하는 모바일 검색 사용자들이 작은 화면과 불편한 조작으로 겪는 어려움을 해결하는 데 도움이 되길 기대한다.

향후 연구로는 규모가 큰 온톨로지를 활용할 때의 처리 기법에 대한 연구를 진행하려 한다. 실제로 업계에서 거대한 규모의 온톨로지를 활용하여 제안한 기법을 사용하기 위해서는 효과적인 인덱스 처리나 온톨로지의 분할 등 효율적인 질의 처리를 위한 연구가 추가로 필요하다. 또한 여러 온톨로지를 함께 사용하는 경우에 대한 추가 연구도 필요하다. 어떤 질의 키워드에 대해 어떤 온톨로지를 사용할 것인지 빠르게 매핑하여 처리하는 방법을 모색해야 한다.

참 고 문 헌

[1] Sandy Shen, Tole J. Hart, Nick Ingelbrecht, Annette Zimmermann, Jessica Ekholm, Nick Jones, Jonathan Edwards, and Andrew Frank, "Dataquest Insight: The Top 10 Consumer Mobile Applications in 2012," Gartner Report, 2009.

[2] David J. Brenes and Daniel Gayo-Avello, "Stratified Analysis of AOL Query Log," *Information Sciences*, vol.179, no.12, pp.1844-1858, 2009.

[3] Joe Pulizzi, "Why Long-Tail Search Rules (and what to do about it.)," 2011. http://blog.junta42.com/2011/03/why-long-tail-search-rules-and-what-to-do-about-it

[4] Ziv Bar-Yossef and Naama Kraus, "Context-Sensitive Query Auto-Completion," in *Proc. of the 20th international conference on World wide web (WWW '11)*, pp.107-116, 2011.

[5] Soyeon Park and Joon Ho Lee, "Unified Search Service of NAVER, a Major Korean Search Engine," in *Proc. of the ACM SIGIR 2008 Workshop on Aggregated Search*, 2008.

[6] Danny Sullivan, "Google Universal Search: 2008 Edition," 2008. http://searchengineland.com/google-universal-search-2008-edition-13256.

[7] Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette, "A Methodology for Evaluating Aggregated Search Results," in *Proc. of the 33th European Conference on Information Retrieval (ECIR '11)*, pp.141-152, 2011.

[8] Masaya Murata, Hiroyuki Toda, Yumiko Matsuura, and Ryoji Kataoka, "Access Concentration Detection in Click Logs to Improve Mobile Web-IR," *Information Sciences*, vol.179, no.12, pp.1859-1869, 2009.

[9] Laurent Mazuel and Nicolas Sabouret, "Semantic Relatedness Measure Using Object Properties in an Ontology," in *Proc. of the 7th International Conference on The Semantic Web (ISWC '08)*, pp.681-694, 2008.

[10] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma, "Probabilistic Query Expansion Using Query Logs," in *Proc. of the 11th international conference on World wide web (WWW '02)*, pp.325-332, 2002.

[11] Christine Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[12] Greg Pass, Abdur Chowdhury, and Cayley Torgeson, "A Picture of Search," in *Proc. of the First International Conference on Scalable Information Systems (InfoScale '06)*, Article 1, pp.1-7, 2006.

[13] Giridhar Kumaran and James Allan, "A Case for Shorter Queries, and Helping User Create Them," in *Proc. of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp.220-227, 2006.

[14] Giridhar Kumaran and Vitor R. Carvalho, "Reducing Long Queries Using Query Quality Predictors," in *Proc. of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pp.564-571, 2009.

[15] Michael Bendersky and W. Bruce Croft, "Discovering Key Concepts in Verbose Queries," in *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, pp.491-498, 2008.

[16] Matthew Lease, James Allan, and W. Bruce Croft, "Regression Rank: Learning to Meet the Opportunity of Descriptive Queries," in *Proc. of the 31th European Conference on Information Retrieval (ECIR '09)*, pp.90-101, 2009.

[17] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum, "Yago: a Core of Semantic Knowledge," in *Proc. of the 16th international conference on World wide web (WWW '07)*, pp.697-706, 2007.

[18] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann, "DBpedia - A Crystallization Point for the Web of Data," *Journal of Web Semantics (JWS)*, vol.7, no.3, pp.154-165, 2009.

[19] Kalervo Jarvelin and Jaana Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," in *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*, pp.41-48, 2000.

송 효 진

2009년 2월 한양대학교 컴퓨터공학부 학사. 2011년 8월 서울대학교 컴퓨터공학부 석사. 2011년 7월~현재 LG전자 SW 플랫폼연구소 연구원. 관심분야는 모바일 컴퓨팅, 웹 표준화, 시맨틱 웹, 데이터 마이닝

이 태 휘

2004년 2월 서울대학교 컴퓨터공학부 학사. 2004년 3월~현재 서울대학교 컴퓨터공학부 석박사 통합과정 재학중. 2007년 6월~2010년 4월 티맥스소프트 선임연구원. 관심분야는 텍스트/그래프 데이터 검색, 대규모 데이터 처리, 시맨틱 웹

김 형 주

1982년 서울대학교 컴퓨터공학부 학사 1985년 University of Texas at Austin Computer Science 석사. 1988년 University of Texas at Austin Computer Science 박사. 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이터베이스, XML, 시맨틱 웹, 전자상거래, IT 정책