유사 단어 커뮤니티 기반의 질의 확장

Query Expansion based on Word Sense Community

저자
(Authors)

곽창욱, 윤희근, 박성배
Chang-Uk Kwak, Hee-Geun Yoon, Seong-Bae Park

출처
(Source)

정보과학회논문지 41(12), 2014.12, 1058-1065 (8 pages)
Journal of KIISE 41(12), 2014.12, 1058-1065 (8 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE02505679

APA Style

곽창욱, 윤희근, 박성배 (2014). 유사 단어 커뮤니티 기반의 질의 확장. 정보과학회논문지, 41(12), 1058-1065.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 41, No. 12, pp. 1058-1065, 2014. 12
http://dx.doi.org/10.5626/JOK.2014.41.12.1058

유사 단어 커뮤니티 기반의 질의 확장
(Query Expansion based on Word Sense Community)

곽 창 욱 † 윤 희 근 †† 박 성 배 †††
(Chang-Uk Kwak) (Hee-Geun Yoon) (Seong-Bae Park)

요 약 질의 확장은 입력된 질의와 관련된 키워드를 사용자에게 제시하여 검색 활동에 도움을 주는 방법이다. 최근에는 사용자가 검색한 내용에서 군집화 방법을 이용하여 도메인을 찾고 키워드를 제시하는 연구가 많이 이루어졌다. 하지만 군집화 방법은 군집의 개수를 정해야하기 때문에 다양한 도메인을 나타내는데 적절하지 않다. 따라서 본 논문은 커뮤니티 인지 알고리즘으로 검색 문서에서 질의마다 다양한 수의 도메인을 찾고 키워드로 선택하여 제시하는 방법을 제안한다. 이를 위해 사용자가 검색한 결과 중 상위 30개 문서를 대상으로 단어를 추출하여 그래프 기반의 커뮤니티를 만들고, 각 커뮤니티에서 키워드를 추출하여 이를 질의 확장에 이용하였다. 본 논문에서 제안한 방법은 구글 검색 엔진과 검색된 문서의 tf-idf를 이용한 키워드 추천 방법과 비교하였다. 제안한 방법이 다른 비교 대상들에 비해 더 다양한 키워드를 추천할 수 있었다.

키워드: 질의 확장, 질의 제시, 커뮤니티 인지, Pseudo Relevance Feedback

Abstract In order to assist user's who are in the process of executing a search, a query expansion method suggests keywords that are related to an input query. Recently, several studies have suggested keywords that are identified by finding domains using a clustering method over the documents that are retrieved. However, the clustering method is not relevant when presenting various domains because the number of clusters should be fixed. This paper proposes a method that suggests keywords by finding various domains related to the input queries by using a community detection algorithm. The proposed method extracts words from the top-30 documents of those that are retrieved and builds communities according to the word graph. Then, keywords representing each community are derived, and the represented keywords are used for the query expansion method. In order to evaluate the proposed method, we compared our results to those of two baseline searches performed by the Google search engine and keyword recommendation using TF-IDF in the search results. The results of the evaluation indicate that the proposed method outperforms the baseline with respect to diversity.

Keywords: query expansion, query suggestion, community detection, pseudo relevance feedback

· 본 논문은 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발)의 지원으로 수행되었음
· 이 논문은 2013(2014)학년도 경북대학교 학술연구비에 의하여 연구되었음
· 이 논문은 2014 한국컴퓨터종합학술대회에서 '유사 단어 커뮤니티 기반의 질의 확장'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 경북대학교 컴퓨터학부
cukwak@sejong.knu.ac.kr
†† 비 회 원 : 경북대학교 컴퓨터학부
hkyoon@sejong.knu.ac.kr
††† 종신회원 : 경북대학교 컴퓨터학부 교수(Kyungpook National Univ.)
seongbae@knu.ac.kr
(Corresponding author임)

논문접수 : 2014년 7월 21일
(Received 21 July 2014)
논문수정 : 2014년 9월 29일
(Revised 29 September 2014)
심사완료 : 2014년 9월 29일
(Accepted 29 September 2014)

정보과학회논문지 제41권 제12호(2014. 12)

1. 서 론

정보화 사회에서 사용자들은 필요한 정보를 찾기 위해 네이버, 구글과 같은 웹 검색 엔진을 이용한다. 보통 사용자는 검색 과정에서 한 단어 내외의 짧은 질의를 사용하는데, 이와 같은 단어들은 여러 가지의 의미(Ambiguous)를 지니고 있는 경우가 많다[1]. 이러한 단어를 질의로 하여 검색하는 경우에 사용자는 검색 결과에서 원하는 내용을 찾지 못할 가능성이 높다. 이는 검색 엔진이 사용자의 검색 의도를 정확하게 알지 못하므로, 입력된 검색어가 포함된 모든 문서들을 보여주기 때문이다. 예를 들어, 사용자가 애플사에서 제조한 컴퓨터에 대한 정보를 찾고자 할 때, 보통의 검색에서는 '애플'이라는 짧은 질의를 이용하여 검색 결과에서 관련된 내용을 찾는다. 하지만 검색 엔진은 사용자의 검색 의도를 모르기 때문에 '애플'이 포함된 모든 문서를 결과로서 보여주게 된다. 이와 같은 이유로 검색 시스템이 보여주는 상위 검색 결과에서 사용자가 원하는 컴퓨터에 관련된 내용을 찾기 힘들 수도 있다.

이와 같이 사용자는 상위 검색 결과에서 질의와 관련된 내용을 찾지 못했을 경우에, 모든 검색 결과를 살펴보는 것이 아니라 질의에 새로운 키워드를 입력해서 검색을 진행한다. 이를 돕기 위해 많은 웹 검색 엔진에서는 다수의 사용자들이 과거에 검색한 질의 기록을 시스템에 반영하여 질의와 함께 많이 검색된 키워드를 '연관검색어'라는 이름으로 제공한다. '연관검색어'는 웹 검색 엔진에서 제공하는 질의 확장 방법 중의 하나로, 입력된 질의와 관련된 키워드들을 사용자에게 제시하여 검색에 도움을 준다. 질의 확장에서는 연관된 키워드를 제시하기 위해, 사용자가 만족할만한 도메인 집합을 찾고 이 각각의 집합을 잘 나타내는 키워드들을 찾는 것이 중요하다.

질의 확장에 대한 기존 연구에서는 주로 외부 리소스나 검색된 문서에서 질의와 관련된 도메인을 찾고 각 도메인에서 키워드들을 추출하는 방법이 많이 연구되고 있다. 문서에서 의미에 따라 구분된 도메인을 찾기 위해 군집화 방법을 이용하는데, 사용자의 매개변수에 따라 구분될 군집의 수, 즉 확장될 키워드 수가 정해지게 된다. 하지만 질의마다 가지고 있는 도메인의 수가 다르기 때문에 군집화 방법은 질의 별로 다양한 수의 도메인을 나타내기 적합하지 않다.

본 논문에서는 기존의 문서 군집화 기반의 질의 확장에서 확장될 키워드의 수가 매개변수에 의해 고정되는 문제를 개선하기위해 커뮤니티 인식 알고리즘을 이용한 질의 확장 방법을 제안한다. 커뮤니티 인식 알고리즘은 사용자의 매개변수 없이 그래프 상에서 유사한 의미를 갖는 군집을 찾는 알고리즘이다. 이로써 도메인에 따라 구분된 다양한 토픽들로 질의를 확장할 수 있다. 먼저 사용자가 검색한 결과 문서에서 단어를 추출하여 co-occurrence 그래프를 만든다. 생성된 그래프를 커뮤니티 인식 알고리즘으로 의미상 연관성이 있는 커뮤니티들을 인식한다. 다음으로 TextRank 알고리즘을 사용하여 커뮤니티 내의 단어에서 커뮤니티를 잘 나타내는 단어를 선택하여 키워드로 설정한다. 적은 수의 노드를 가진 커뮤니티는 해당 커뮤니티를 잘 나타내지 못하기 때문에, 이러한 커뮤니티는 정제 과정을 통해 제거한다. 최종적으로 커뮤니티 정제 과정 후에 남겨진 커뮤니티에서 키워드를 추출하여 질의를 확장한다.

실험에서는 영문 위키피디아 덤프 데이터를 검색에 사용하였다. 실험 결과는 두 가지의 모델과 비교하였고, 각 비교 모델에서 생성된 키워드를 사람들이 직접 점수로 평가하였다. 타 비교 모델들과 비교한 결과, 제안한 방법이 다른 비교 모델들에 비해 더 다양한 도메인의 키워드들을 나타냈다.

본 논문의 구성은 다음과 같다. 2절에서는 질의 확장과 관련된 연구들에 대해 설명한다. 3절에서는 커뮤니티 인식 알고리즘을 이용한 질의 확장 방법에 대해서 살펴본다. 4절에서는 위키피디아 데이터를 대상으로 실험한 결과를 비교 분석한다. 마지막으로 5절에서는 결론을 맺는다.

2. 관련 연구

질의 확장은 질의 정제, 질의 제시 등 여러 가지 용어로 다양한 연구들이 이루어지고 있다. 기존 연구에서는 입력된 질의와 관련된 도메인을 찾아 이를 사용자에게 제안하기 위해 다양한 데이터를 이용한 방법들을 제시하였다.

먼저 사용자의 과거 검색 기록을 이용한 방법이다. Cui et al.는 사용자가 과거에 검색한 검색 로그에서 질의와 관련된 도메인을 찾고 이를 질의 확장에 이용하는 방법을 제안하였다[2]. 이 방법은 과거 검색 로그를 분석하여 질의와 사용자가 클릭한 문서의 단어와의 확률적 유사도를 계산한 후, 질의와 가장 유사한 단어를 선택하고 이를 질의 확장하는 방법이다. 하지만 이와 같이 과거 검색 기록을 이용한 방법은 검색 로그 데이터셋에서 개인 정보 유출의 문제가 발생할 수 있어, Google, Microsoft 등 검색 엔진을 운영하고 있는 회사를 중심으로 제한적인 연구가 이루어지고 있다.

두 번째로, 외부 리소스를 이용하여 사용자에게 제안할 키워드 집합을 찾는 방법이다. Bermhard et al.는 도메인을 찾기 위해 외부 사전을 이용하였다[3]. 이 방법은 다양한 도메인을 커버하기 위해 WordNet, GCIDE,

English Wiktionary, Simple English Wiktionary, Omega-Wiki, Wikipedia, Simple Wikipedia, 총 7개의 외부 리소스를 이용하였다. 위의 7개의 외부 리소스에서 단어에 대한 정의(Definition)를 추출하여 그래프를 만든 후, 이를 군집화 방법을 사용하여 도메인에 따라 구분하였다. 구분된 각 군집에서 질의와 가장 유사한 키워드를 선택하여 질의 확장하였다. 또한 Hu et al.는 다양한 질의 제안을 위해 위키피디아 카테고리 정보를 이용하였다[4]. 이를 위해 질의와 관련된 키워드들을 추출하고, 이 키워드들을 위키피디아 카테고리 정보에 매핑한다. 이후에 질의와 카테고리 정보의 유사도 비교를 통해 관련성이 높은 카테고리 이름을 선택하여 다양한 키워드로 확장하였다. 이와 같이 외부 리소스를 이용한 방법은 잘 정리된 데이터에서 도메인을 찾아 질의 확장에 이용하기 때문에 질의와 관련된 키워드들로 효과적으로 확장할 수 있다. 하지만 이 방법은 보통 리소스 크기가 매우 크기 때문에 질의가 입력될 때마다 리소스에서 관련된 도메인을 찾는 비용이 많이 든다.

세 번째로, 사용자가 검색한 결과 문서를 Pseudo Relevance Feedback(PRF)하여 질의 확장하는 방법이다[5, 6,8-10]. PRF를 사용하는 방법은 보통 사용자가 검색한 결과에서 질의와 연관성이 높은 상위 몇 개의 문서를 선택하여 관련된 정보를 찾아 질의 확장에 이용하는 것이다. Xu et al.은 주어진 질의를 위키피디아에서 검색한 후 검색 결과에서 최상위 문서를 선택하고, 선택된 문서 내에서 질의와 연관성 있는 단어를 키워드로 선택하여 질의 확장하였다[5]. Zhao et al.은 검색된 문서들을 요약하여 질의 확장하는 방법을 제안하였다[6]. 이 방법은 문서 내에서 문장과 문장사이의 관계와 문장과 단어의 관계를 이용하여 확장할 단어를 선택하고, 이를 질의 확장하였다. Andrzejewski et al.은 토픽 모델을 이용하여 검색된 문서에서 토픽을 찾고 이를 질의 확장하는 방법을 제안하였다[7]. 이 방법은 LDA라는 토픽 모델을 이용하여 상위 2개의 문서를 대상으로 토픽을 구분한 뒤 키워드를 선택하였다. 이처럼 PRF를 이용하여 질의 확장하는 방법은 정보 검색 분야의 많은 연구에서 효과적인 것으로 증명되었다. 이 방법은 검색된 문서에서 의미에 따라 도메인을 구분하고 이 중에서 질의와 가장 유사한 것을 선택하여 키워드로 확장하는 것이 필요하다.

최근 연구에서는 군집화 방법을 사용하여 의미상 서로 연관된 군집으로 구분하였다[3,8]. Liu et al.는 검색된 문서를 군집화하여 각 군집에서 후보 키워드를 추출하고, 이를 각 스텝마다 질의에 삽입 또는 삭제하여 확장될 질의의 F-measure가 최대화 되는 키워드를 찾는 알고리즘을 제안하였다[8]. 보통 군집화 방법으로 k-means 알고리즘을 사용하는데, 구분될 군집의 수인 K를 정해놓고 분류하게 된다. 이 방법을 이용한 질의 확장은 군집의 수만큼 확장될 질의가 결정되므로, 적절한 군집의 수를 정해야하는 문제가 발생한다. 이는 특정 도메인마다 구분될 군집의 개수, 즉 의미에 따라 구분될 집합의 수가 다를 수 있는데, 각 도메인별로 다양한 토픽들을 나타내기에 적합하지 않다. 이와 같은 이유로 본 논문에서는 커뮤니티 인식 알고리즘을 이용하여 질의마다 다른 개수를 가진 도메인을 다양하게 구분하고, 각 도메인별로 확장될 질의를 선택하여 사용자에게 제시하는 방법을 제안한다.

3. 커뮤니티 인식 알고리즘을 이용한 질의 확장

본 논문에서 제안하는 질의 확장 방법은 그림 1과 같다. 먼저 사용자가 질의로 검색한 결과 문서에서 단어를 추출하여 그래프를 생성한다. 두 번째로, 생성된 그래프에서 의미상으로 유사한 군집을 찾기 위해서, 그래프 기반의 커뮤니티 인식 알고리즘을 사용하여 커뮤니티를 인식한다. 세 번째로, 인식된 각각의 커뮤니티를 대표하는 키워드를 정한다. 네 번째로, 커뮤니티 인식 알고리즘을 이용한 커뮤니티 인식 결과에서 적은 노드 수를 가진 커뮤니티는 해당 커뮤니티를 잘 나타내지 못한다. 이를 위해 노드 수가 적은 커뮤니티는 커뮤니티 정제 과정을 통해 제거한다. 마지막으로 커뮤니티 정제 이후에 남겨진 각각의 커뮤니티에서 키워드를 선택하여 사용자에게 제시한다.

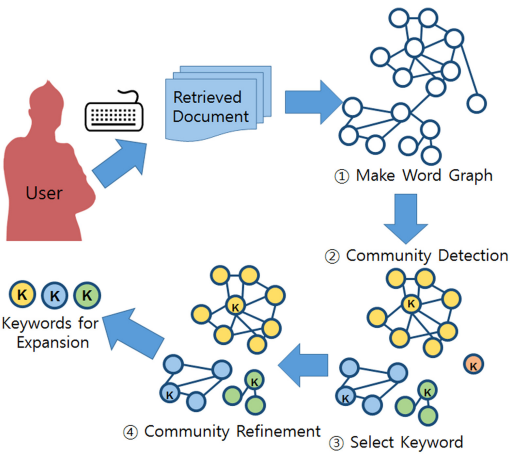

그림 1 질의 확장 흐름도

Fig. 1 A process of query expansion

3.1 단어 그래프

일반적으로 질의에 사용되는 어휘는 단어 단위로 이루어져 있기 때문에 단어 기반의 co-occurrence 그래프

표 1 질의 'Apple' 검색된 문서의 일부
Table 1 An example of retrieved document for 'Apple'

Apple Inc. American ultinational corporation eadquartered Cupertino, California, that designs, ...

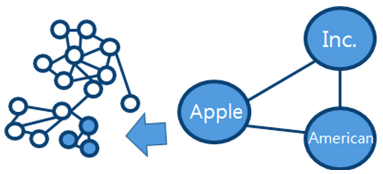

그림 2 단어 그래프 예시
Fig. 2 An example of word graph

를 생성한다. 본 논문에서는 PRF 방법을 이용하여 상위 K개의 문서를 선택한 뒤 이 문서내의 단어들로 단어 그래프를 생성한다[9,10]. 그래프에서 노드는 각 문서에서 추출한 단어이고, 에지는 두 키워드가 주어진 윈도우 범위 내에서 함께 등장하면 생성한다. 표 1은 'Apple'이라는 질의로 검색된 문서 중 일부이다. 예를 들어, 표 1의 문서를 대상으로 그래프를 만들 때, 주어진 윈도우가 3이라면 'Apple'과 'Inc.', 'American'은 윈도우 범위 내에서 함께 등장했기 때문에 그림 2와 같은 그래프가 생성되게 된다. 이와 같은 방법으로 문서내의 모든 단어들을 대상으로 그래프를 구성한다. 식 (1)에서 $W(t_i, t_j)$는 단어 그래프 내에서 두 단어 사이의 에지 가중치를 나타낸다. 에지 가중치 $W(t_i, t_j)$는 두 단어 t_i, t_j가 함께 등장할 확률을 각 단어가 나타날 확률의 곱으로 나눈 값을 설정한다.

$$W(t_i, t_j) = \frac{CoProb(t_i, t_j)}{tfProb(t_i) \times tfProb(t_j)} \quad (1)$$

아래의 식 (2)에서 $CoProb(t_i, t_j)$ 함수는 주어진 윈도우 내에서 두 단어 t_i, t_j가 함께 등장할 확률을 의미하며, $tfProb(t_i)$ 함수는 윈도우 내에서 단어 t_i가 등장할 확률을 나타낸다.

$$CoProb(t_i, t_j) = \frac{tf_{t_i t_j}}{l} \quad (2)$$

$$tfProb(t_i) = \frac{tf_{t_i}}{l}$$

$tf_{t_i t_j}$는 단어 t_i t_j가 문서 내에서 함께 등장한 횟수를 나타낸다. l은 각 문서 내의 단어의 총 개수를 나타낸다. tf_{t_i}는 문서 내에서 단어 t_i가 나타난 횟수를 나타낸다. 위의 식 (1)에 따라 노드 간에 계산된 가중치가 0일 경우에는 에지를 생성하지 않는다. 이와 같은 방법으로 검색된 상위 K개의 문서에 대해 단어 그래프를 생성한다.

3.2 커뮤니티 인식 알고리즘

앞서 생성된 단어 그래프에서 연관된 의미를 가진 도메인 군집을 얻기 위하여 그래프 기반의 커뮤니티 인식 알고리즘을 사용한다. 커뮤니티는 유사한 의미를 가진 단어들이 모여있는 것이다. 커뮤니티 인식 알고리즘은 사용자의 매개변수 없이, 그래프에서 유사한 도메인을 가진 것들끼리 자동으로 군집을 찾을 수 있다. 본 논문에서는 Clauset et al. 이 제안한 커뮤니티 인지 알고리즘을 사용한다[11]. 이 알고리즘은 그래프 내의 모듈성(Modularity)을 기반으로 고안되었다. 모듈성은 커뮤니티 내에는 밀접하게 에지로 연결되어있고, 커뮤니티 사이에는 적은 연결이 있다고 가정할 때, 커뮤니티가 잘 구분되었는지 측정하는 척도이다. 모듈성 Q는 식 (3)를 사용하여 계산할 수 있다. 커뮤니티 인지 알고리즘은 먼저 그래프 내의 모든 노드들을 각각의 커뮤니티에 배정하고, 반복적으로 두 개의 커뮤니티를 비교하여 Q의 값이 가장 높은 두 개의 커뮤니티를 병합한다.

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (3)$$

식 (3)에서 m은 그래프 내의 모든 노드의 개수를 나타낸다. A_{ij}는 노드 i와 j가 서로 에지로 연결되어있으면 1, 연결되어있지 않으면 0으로 나타낸 행렬이다. k_i는 노드 i의 차수를 나타낸다. $\delta(c_i, c_j)$ 함수는 노드 i와 j가 같은 커뮤니티에 속했는지를 나타낸다. c_i는 노드 i가 속한 커뮤니티를 나타내는데, 만약 노드 i와 j가 동일한 커뮤니티에 속하면 1, 서로 다른 커뮤니티에 존재하면 0을 나타낸다. 이 알고리즘은 모듈성 Q가 높아지는 커뮤니티 병합이 있는 한 반복적으로 수행된다.

3.3 키워드 설정

커뮤니티 인식 알고리즘을 통해 인식된 커뮤니티는 의미상으로 유사한 도메인을 가진 단어들이 모여 있는 것이다. 각 커뮤니티가 가지고 있는 도메인을 나타내기 위해, 각 커뮤니티 내에서 해당 커뮤니티를 대표하는 단어를 커뮤니티 키워드로 설정한다. 추후 커뮤니티 정제 후에 커뮤니티 별로 설정된 키워드를 확장될 키워드로 사용자에게 추천한다. 이를 위해 TextRank 알고리즘을 이용하여 커뮤니티 내의 중요한 단어를 찾는다[12].

TextRank 알고리즘은 그래프 내에서 서로 연결된 에지를 이용하여 각 노드의 점수를 계산한 후 중요도에 따라 순위를 정하는 알고리즘이다. TextRank 알고리즘 이용하면 커뮤니티 내의 노드 값을 비교하여 중요한 노드를 찾을 수 있다. 식 (4)에서 $WS(V_i)$는 TextRank 알고리즘 수행 후 그래프 내의 노드 V_i가 가지는 값을 나타낸다.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(v_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (4)$$

d는 현재 노드에서 다른 노드로 이동할 확률을 나타낸다. 여기서 d는 일반적으로 사용되는 0.85로 설정한다. $In(V_i)$는 V_i의 들어오는 에지를 가진 노드의 집합이고, $Out(V_i)$는 V_i로 나가는 에지를 가진 노드의 집합이다. w_{ij}는 노드 V_i와 V_j 사이의 에지 가중치를 나타낸다. TextRank 알고리즘을 이용하여 커뮤니티 내의 모든 노드의 값을 계산한 후 커뮤니티 내에서 가장 큰 값을 가진 노드를 선택하여 키워드로 설정한다. 식 (5)는 커뮤니티 내에서 키워드를 선택하는 것을 나타낸다.

$$Keyword(t_i) = \arg max_{v_i \in C}(WS(V_i)) \quad (5)$$

위의 식 (5)에 따라 커뮤니티 내에서 노드의 값이 가장 큰 것을 선택하여 커뮤니티를 대표하는 키워드로 설정한다. C는 커뮤니티 내의 모든 노드들의 집합을 나타내고, $WS(V_i)$는 TextRank 알고리즘을 수행한 이후에 노드 V_i가 가지는 값을 나타낸다.

3.4 커뮤니티 정제

커뮤니티 인식 알고리즘을 통해 커뮤니티를 인식한 결과에서는 적은 단어로 구성된 노이즈 커뮤니티들이 포함되어 있다. 이러한 노이즈 커뮤니티는 구성된 노드 수가 적어, 도메인을 잘 나타내지 못하기 때문에 커뮤니티 정제 작업을 통해 제거한다. 본 논문에서는 Chen et al.이 제안한 방법을 따라서 전체 키워드의 5%보다 적은 노드 수를 가진 커뮤니티는 제거한다[9].

4. 실험 결과

4.1 실험 설계

실험을 위하여 위키피디아 영문 덤프파일[1]을 이용하였다. 본 논문에서는 리다이렉트 페이지를 제거한 총 500만개의 문서를 대상으로 실험을 진행하였다. 오픈 소스 검색 엔진인 Apache Solr[2]로 이 문서들을 인덱싱하여 검색 시스템을 구축하였다. 여기서 각 질의로 검색된 상위 K개의 문서를 질의 확장에 이용하였다. 본 실험에서는 K의 개수를 30개로 설정하였다. 단어 그래프의 구축을 위해 Stanford POS Tagger[3]를 사용하여 불용어(Stopword)를 제거한 명사, 형용사, 동사를 추출하였다. 노드 사이의 에지 생성에 필요한 윈도우 사이즈는 5로 설정하였다.

본 논문에서 제안한 방법은 Liu et al.이 제시한 비교방법을 이용하여, 두 가지의 질의 확장 모델과 비교한 후 평가하였다[8]. 첫 번째 비교 모델인 구글[4]에서의 검색은 최근의 다수의 사용자가 검색한 키워드를 이용하는 방법으로, 해당 질의에 대해 검색 시스템이 추천한 키워드들을 비교 대상으로 사용하였다. 또한 Data Clouds 모델과 비교하였다[13]. Data Clouds 모델은 문서를 요약하기 위해 제안된 방법으로, 검색 결과에서 tf-idf 값이 높은 단어를 중요한 키워드로 여겨 확장될 질의로 선택한다.

표 2 질의 집합
Table 2 Query Set

Q1	Google	Q6	Microsoft
Q2	Apple	Q7	Android
Q3	Java	Q8	Samsung
Q4	Shell	Q9	Rockets
Q5	Network	Q10	Phone
		Q11	Computer

본 논문에서는 임의로 11개의 질의를 선택하였다. 실험에 사용한 질의 집합은 표 2와 같다. 또한 성능 측정을 위해 7명의 인원이 각 평가 지표에 따라 암맹 평가하였다. 질의 평가에 대한 실험은 아래와 같이 두 가지로 진행하였다.

Part1. 확장될 질의가 사용자에게 도움이 되었는가?(Individual Query Evaluation)

Part2. 확장될 질의가 포괄적이고 다양한가?(Collective Query Evaluation)

4.2 실험 결과

Part1: 개별 질의 점수(Individual Query Score)

Part1에서는 확장될 질의가 사용자에게 도움이 되었는지 여부를 측정하였다. 여기서는 두 가지의 지표에 따라 평가하였다. 첫 번째 지표 평가에서 평가자는 확장될 질의의 집합을 보고 0-5 사이의 점수를 부여하였다. 두 번째 지표 평가에서는 확장될 질의 집합을 보고 입력된 질의와 관련성 척도를 상, 중, 하로 평가하였다. 각 지표에 따른 평가 결과는 그림 3에서 볼 수 있다. 그림 3(a)는 사용자가 확장될 질의를 보고 점수를 부여한 것으로 확장될 질의의 질(Quality)을 측정할 수 있다. 그림은 해당 질문에 대한 모델별 평균 점수를 나타낸다. 그림 3(b)는 확장될 질의가 검색과 관련이 있는지 비교한 것이다. 그림에서 각 지표는 위로부터 '상', '중', '하'의 비율을 의미한다. 또한 표 3은 세 가지 질의를 각 비교 모델에서 실험한 결과를 보여준다. 그림 3(a)를 통해 확인할 수 있듯이, 본 논문에서 제안한 모델과 구글 모델은 Data Clouds 모델에 비해 제안된 키워드의 평균 점수가 더 높았다. 이는 확장될 키워드들이 실용성이 높아 실제로 검색에 사용될 수 있을 만큼 질이 높다는 것을

1) http://dumps.wikimedia.org/enwiki/
2) https://lucene.apache.org/solr/
3) http://nlp.stanford.edu/software/tagger.shtml
4) http://www.google.com

표 3 질의 확장 실험 결과

Table 3 Experimental results on query expansion

	Google	Data Clouds (tf-idf)	Proposed Method
Samsung	q1: “samsung galaxy s5” q2: “samsung mobile” q3: “samsung galaxy s4” q4: “samsung kies” q5: “samsung note 4” q6: “samsung s5”	q1: “samsung touchwiz” q2: “samsung book” q3: “samsung mega” q4: “samsung worldwide” q5: “samsung nature” q6: “samsung rugby”	q1: “samsung kies” q2: “samsung rugby” q3: “samsung language” q4: “samsung galaxy” q5: “samsung gleam” q6: “samsung korea”
Apple	q1: “apple store” q2: “apple id” q3: “apple tv” q4: “applebee’s” q5: “apple iphone 6” q6: “apple trailer” q7: “apple uk”	q1: “apple monitor” q2: “apple sun” q3: “apple grove” q4: “apple store” q5: “chiness apple” q6: “japan apple” q7: “indo apple”	q1: “apple store” q2: “apple specialist” q3: “timeline apple” q4: “apple valley” q5: “apple sun” q6: “cooking apple” q7: “apple caramel”
java	q1: “java download” q2: “javascript” q3: “java 64 bit” q4: “java tutorial” q5: “java jdk” q6: “java javatpoint” q7: “javascript array” q8: “verify java”	q1: “java xml” q2: “java lieftinck” q3: “java kediri” q4: “java earthquake” q5: “java priority” q6: “java jacksonvile” q7: “java calendar q8: “java persistent”	q1: “java kediri” q2: “java sun” q3: “java real-time” q4: “java api” q5: “java card” q6: “java license” q7: “java interface” q8: “java chrysler”
Network	q1: “network rail” q2: “network solution” q3: “network speed test” q4: “networking” q5: “network marketing” q6: “network rail jobs” q7: “network tools” q8: “network topology” q9: “network railcard”	q1: “network dynamic” q2: “asia network” q3: “local network” q4: “network edonkey” q5: “network astro” q6: “radio network” q7: “network operators” q8: “network nhl” q9: “network dynamical”	q1: “network computer” q2: “network university” q3: “network television” q4: “network application” q5: “communication network” q6: “airport network” q7: “network system” q8: “transport network” q9: “asia network”

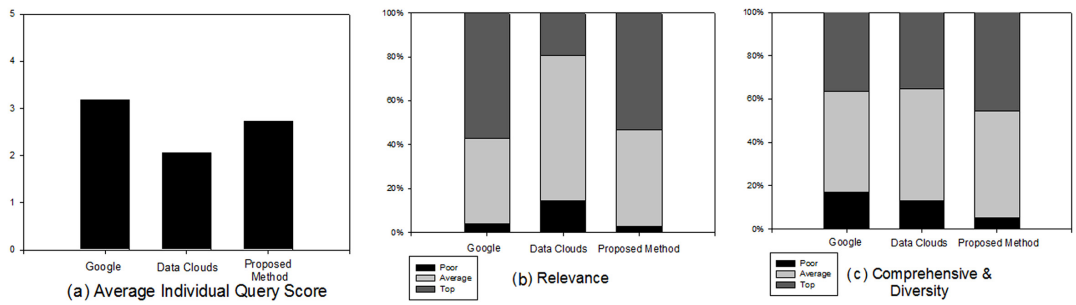

그림 3 확장된 질의의 성능 비교 결과

Fig. 3 Performance for expanded query

의미한다. 예를 들어, ‘Network’에 대한 질의에서 본 논문에서 제안한 모델은 ‘computer’, ‘university’, ‘transport’ 등과 같은 키워드가 추출되었다. 키워드 ‘computer’는 네트워크에 이용되는 컴퓨터에 대한 도메인의 검색이 가능하고, 키워드 ‘university’는 네트워크 관련 대학을 찾는 사용자에게 도움이 될 것이다. 또한 ‘transport’는 교통 네트워크에 대한 도메인을 검색 할 수 있다. 이는 본 논문에서 제시된 키워드들이 질의와 관련된 각 도메인을 효과적으로 나타내는 것을 알 수 있다. ‘Java’라는 질의에서 Data Clouds 방법은 ‘jacksonvile’과 같은 질의가 제시되었다. 이 키워드는 검색 결과 문서 중에 XML을 위한 Java API(JAX)의 종류를 나타내는 글에서 많이 나타났기 때문에 키워드로 제안되었다. 하지만 이 키워드는 ‘Java’라는 도메인과 관련이 없는 키워드로

써 이러한 키워드는 사용자에게 낮은 만족도를 가지게 한다. 이는 전체적으로 보았을 때, 제안한 방법과 구글 모델에서 사용자에게 도움이 되는 키워드가 더 많이 포함된 것을 알 수 있었다.

또한 그림 3(b)를 통해 본 논문에서 제안한 방법의 키워드가 입력된 질의와도 적절한 관련성이 있음을 증명할 수 있었다. 예를 들어, 'Samsung'이라는 질의로 검색했을 때, 제안한 방법은 'voice', 'galaxy', 'gleam'과 같이 삼성 모바일 관련 도메인과 'kies'와 같이 현재 서비스 중인 S/W 관련 키워드들이 제시되었다. 이는 본 논문에서 제안한 방법이 검색 결과에서 질의와 관련된 도메인의 키워드를 효과적으로 선택하여 제시하고 있음을 볼 수 있다. 하지만 그림 3(b)에서 볼 수 있듯이, 제안한 방법은 구글 시스템에 비해 관련성 측면에서 '상'의 비율이 낮았다. 이는 구글 시스템은 최근 다수의 사용자가 검색한 로그 기록을 반영하여 추천할 키워드를 선택하기 때문에 관련성 측면에서 비교적 만족성이 높다는 특징이 있었다.

Part2: 종합 질의 점수(Collective Query Score)

이전 연구의 실험에서 사용자들은 제시될 좋은 질의 집합의 조건에 여러 척도 중 포괄성(Comprehensive)과 다양성(Diversity)이 가장 중요한 것으로 응답하였다[8]. 이는 사용자에게 많은 도메인을 포함한 질의 집합을 제시했을 때 만족도가 높음을 알 수 있다.

Part 2에서는 각 모델별 다양성(Diversity)을 비교하였다. 평가자는 각 비교 모델별 확장될 질의를 보고 '상', '중', '하' 중 하나의 점수를 부여하여 평가하였다. 그림 3(c)는 확장될 질의가 포괄적(Comprehensive)이고 다양한(Diversity) 도메인들을 제시하여 주는지 여부를 비교한 것이다. 각 항목에서 지표는 위부터 '상', '중', '하'의 비율을 나타낸다. 다양성 측면에서 본 논문에서 제시한 방법은 그림 3(c)에서 볼 수 있듯이, 다른 비교 모델들에 비해 높은 지표를 나타냈다. 예를 들어, 'Apple'이라는 질의에서 제안한 방법은 'specialist', 'timeline', 'valley', 'cooking' 등과 같이 질의가 가지고 있는 여러 도메인의 키워드를 제시하였다. 사용자는 'specialist'를 이용하여 애플사의 전문가 직업에 대한 정보 검색이 가능하며, 'timeline'으로 애플사의 역사에 대해 검색이 가능하다. 또한 'cooking'은 사과의 요리 정보에 대해 검색할 수 있다. 이처럼 제안한 방법은 질의가 가지고 있는 여러 도메인을 키워드로 나타내는 것을 볼 수 있다. 이에 반하여 구글 시스템에서의 'Samsung'에 대한 질의 추천 결과에서는 'galaxy s5', 'galaxy note 4', "mobile" 등 최근에 크게 이슈가 되는 핸드폰에 관한 키워드들이 제시되어 해당 질의가 보유하고 있는 다양한 도메인들을 나타내지 않는 것을 볼 수 있었다. 또한 'Java'라는 질의로 검색했을 때, 구글 시스템에서 추천 키워드는 'download, 'jdk', '64 bit'과 같이 대부분이 Sun사에서 제공되는 프로그래밍 언어에 관련된 것이었다. 이에 반하여 제안한 방법에서는 프로그래밍 언어 관련 도메인뿐만 아니라, 'card', 'kediri'와 같은 키워드들도 추천되었다. 키워드 'card'는 Java로 개발된 프로그램이 디바이스 내에서 실행 될 수 있게 하는 플랫폼으로써 이 기술을 찾는 사용자에게 도움이 될 것이다. 또한 'kediri'는 자바 섬에 있는 지역으로써 지명을 찾는 사용자에게 도움이 될 것이다. 이와 같이 구글 시스템에서는 제시된 키워드가 적은 도메인으로 한정된 특성을 보여, 많은 사용자들을 만족시키기 어렵지만 제시한 방법에서는 더 많은 도메인의 키워드를 제시하였다.

또한 Data Clouds 모델도 다양성 측면에서 비교적 낮은 지표를 나타냈다. 검색된 결과에서 tf-idf 값이 높은 단어를 확장할 키워드로 선택하는 Data Clouds 모델은 적은 문서에서 많이 나타난 단어가 키워드로 선택된다. 질의 'Samsung'에 대한 Data Clouds 모델의 추천 결과에서 "touchwiz"란 키워드가 제시되었다. 'touchwiz'는 'Samsung'의 터치 인터페이스 모델로써 검색된 문서 중 'touchwiz'란 글에서 반복적으로 많이 나타났기 때문에 선택되었다. 'nature'는 'touchwiz'의 버전 이름으로써 같은 도메인을 나타낸다. 이는 단순히 적은 문서에서 많이 반복된 단어가 키워드로써 중요한 의미를 가질 수는 있지만, 키워드 추천 결과에서 서로 중복된 도메인을 가진 키워드들이 선택될 수 있음을 보인다. 이처럼 본 실험을 통해서 제안한 방법이 확장될 키워드의 숫자를 정하지 않고도 도메인에 따라 적절한 수의 질의를 효과적으로 추천하는 것을 증명하였다. 또한 다양성 측면에서 다른 비교 모델에 비해서 좋은 성능을 보임을 확인하였다.

5. 결 론

최근의 질의 확장 연구에서는 사용자가 검색한 문서에서 입력된 질의와 관련된 도메인을 찾아 키워드로 제시하는 연구가 많이 이루어지고 있다. 이러한 연구에서는 도메인 구분을 위해 군집화 방법을 이용하는데, 군집의 수를 사용자가 매개변수로 입력하여 결정해야 한다. 하지만 이 방법에서는 질의마다 가지고 있는 도메인의 수가 다르기 때문에 질의마다 다양하게 가지고 있는 도메인들을 나타내기 어렵다.

본 논문에서는 이러한 문제를 해결하기 위해 커뮤니티 인식 알고리즘 기반의 질의 확장 방법을 제안하였다. 제안한 방법에서는 적절한 군집의 수를 정해야하는 문제를 효과적으로 해결하고 질의마다 가지고 있는 다양한 도메인의 군집들을 인식하고 키워드로 제시할 수 있었다.

실험을 통해 다른 모델과 비교하여 다양한 키워드를 사용자에게 제시함을 보였다. 이는 질의 별로 다양하게 가지고 있는 도메인을 사용자에게 제시하고 있음을 의미한다. 실험 결과에 따르면, 실제로 정보 검색에 널리 사용되는 구글에서의 추천 결과와 비교하였을 때, 질의의 실용성과 관련성 측면에서 비슷한 성능을 가지지만 질의의 다양성 측면에서 더 좋은 성능을 나타냈다. 또한 문서의 요약을 위해 tf-idf 방법을 이용한 Data Clouds 모델과의 비교에서도 제시한 방법이 질의와 관련된 다양한 도메인의 키워드를 사용자에게 제시하는 것을 볼 수 있었다. 결론적으로 제시하는 방법이 질의와 관련된 의미 있는 여러 가지의 도메인들을 사용자에게 제시하기 때문에, 질의와 관련된 다양한 정보의 검색 활동에 도움이 될 수 있다.

References

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation using Query Logs in Search Engines," *Proc. of EDBT Workshops*, pp. 588-596, 2004.

[2] H. Cui, J. R. Wen, J. Y. Nie, and W. Y. Ma, "Probabilistic Query Expansion Using Query Logs," *Proc. of the 11th WWW*, pp. 325-332, 2002.

[3] D. Bermhard, "Query Expansion based on Pseudo Relevance Feedback from Definition Clusters," *Proc. of the 23rd Coling*, pp. 54-62, 2010.

[4] H. Hu, M. Zhang, Z. He, P. Wang, and W. Wang, "Diversifying Query Suggestions by using Topics from Wikipedia," *Proc. of WI-IAT*, pp. 139-146, 2013.

[5] Y. Xu, G. JF. Jones, and B. Wang, "Query Dependent Pseudo-Relevance Feedback based on Wikipedia," *Proc. of the 32nd SIGIR*, pp. 59-66, 2009.

[6] L. Zhao, L. Wu, and X Huang, "Using query expansion in graph-based approach for query-focused multi-document summarization," *Journal of Information Processing & Management*, Vol. 45, No. 1, pp. 35-41, 2009.

[7] D. Andrzejewski and D. Buttler, "Latent Topic Feedback for Information Retrieval," *Proc. of the 17th SIGKDD*, pp. 600-608, 2011.

[8] Z. Liu, S. Natarajan, and Y. Chen, "Query Expansion Based on Clustered Results," *Proc. of the 37th VLDB*, Vol. 4, No. 6, pp. 350-361, 2011.

[9] J. Chen, O. R. Zaïane, and R. Goebel, "An Unsupervised Approach to Cluster Web Search Results based on Word Sense Communities," *Proc. of WI-AIT*, Vol. 01, pp. 725-729, 2008.

[10] C.-U. Kwak, H.-G. Yoon, and S.-B. Park, "Query Expansion based on Word Sense Community," *Proc. of KCC*, pp. 656-658, 2014. (in Korean)

[11] A. Clauset, M. EJ. Newman, and C. Moore, "Finding community structure in very large networks," *Journal of Physical review E*, Vol. 70, No. 6, pp. 66-111, 2004.

[12] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts," *Proc. of EMNLP*, pp. 404-411, 2004.

[13] G. Koutrika, Z. M. Zadeh, and H. Garcia-Molina, "Data Clouds: Summarizing Keyword Search Results over Structured Data," *Proc. of EDBT*, pp. 391-402, 2009.

곽 창 욱

2013년 동국대학교 컴퓨터멀티미디어학부 졸업(학사). 2013년~현재 경북대학교 대학원 컴퓨터학부 석사과정. 관심분야는 텍스트마이닝, 기계학습

윤 희 근

2007년 경북대학교 컴퓨터공학과 졸업(학사). 2009년 경북대학교 대학원 컴퓨터공학과(석사). 2009년~현재 경북대학교 대학원 컴퓨터학부 박사과정. 관심분야는 기계학습, 자연어처리

박 성 배

1994년 한국과학기술원 컴퓨터과학과 졸업(학사). 1996년 서울대학교 대학원 컴퓨터공학과 졸업(석사). 2002년 서울대학교 대학원 컴퓨터공학과 졸업(박사). 2004년~현재 경북대학교 IT대학 컴퓨터학부 교수. 관심분야는 기계학습, 자연어처리, 텍스트마이닝, 정보추출, 생명정보학