KIPS Tr. Software and Data Eng.
Vol.5, No.9 pp.451~464 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.9.451

Mining Search Keywords for Improving the Accuracy of Entity Search

Lee Sun Ku[†] · On Byung-Won[††] · Jung Soo-Mok[†††]

ABSTRACT

Nowadays, entity search such as Google Product Search and Yahoo Pipes has been in the spotlight. The entity search engines have been used to retrieve web pages relevant with a particular entity. However, if an entity (e.g., Chinatown movie) has various meanings (e.g., Chinatown movies, Chinatown restaurants, and Incheon Chinatown), then the accuracy of the search result will be decreased significantly. To address this problem, in this article, we propose a novel method that quantifies the importance of search queries and then offers the best query for the entity search, based on Frequent Pattern (FP)-Tree, considering the correlation between the entity relevance and the frequency of web pages. According to the experimental results presented in this paper, the proposed method (59% in the average precision) improved the accuracy five times, compared to the traditional query terms (less than 10% in the average precision).

Keywords : Entity Search, FP-Tree, Query, Frequency

엔터티 검색의 정확성을 높이기 위한 검색 키워드 마이닝

이 선 구[†] · 온 병 원[††] · 정 수 목[†††]

요 약

최근 Google Product Search와 Yahoo Pipes와 같은 엔터티 검색이 각광을 받고 있다. 특정 엔터티와 관련 있는 웹 페이지를 검색하기 위해 엔터티 검색이 사용된다. 그러나 엔터티(예를 들면, 차이나타운 영화)가 다양한 의미(예를 들면, 차이나타운 영화, 차이나타운 음식점, 인천 차이나타운 등)을 포함하고 있다면 엔터티 검색의 정확성은 크게 떨어진다. 이러한 문제를 해결하기 위해, 본 논문에서는 웹 페이지의 빈도수와 엔터티 관련성 간의 상관관계를 고려하여, Frequent Pattern (FP)—Tree에 기반을 둔 질의어의 중요도를 측정하고 베스트 질의어를 제안하는 새로운 방안을 제안한다. 본 논문의 실험 결과에 의하면, 기존 방안의 정확도가 10% 미만인데 비해, 제안 방안의 평균 정확도는 59%로, 약 5배 향상시킨다.

키워드 : 엔터티 검색, FP-트리, 질의어, 빈도

1. 서 론

최근 엔터티 검색(entity search)에 대한 연구가 활발히 진행되고 있다. 네이버 또는 구글 검색 엔진에서 '암살'과 같은 영화 엔터티를 검색하면, 그 엔터티(entity)[1]와 관련 있는 웹 페이지들(relevant web pages)이 검색된다. 그러나 엔터티의 의미가 모호(ambiguous)하거나 여러 의미를 가진 경우에는 검색 결과의 정확성은 현저히 떨어진다. Fig. 1은 네이버 검색 엔진에서 '차이나타운'이라는 영화 엔터티를 검색하여 얻은 상위 6개의 웹 페이지 결과이다. '차이나타운'은 영화명, 음식점명, 인천의 차이나타운 등 여러 의미를 동시에 지니고 있기 때문에 원래 찾고자하는 '차이나타운' 영화 엔터티에 대한 검색 결과의 정확성은 크게 떨어지게 된다.

이러한 엔터티 검색에서 발생하는 문제를 해결하기 위해, 본 연구에서는 엔터티와 관련 있는 웹 페이지들을 가장 잘 찾을 수 있는 질의어를 자동으로 추출하는 알고리즘을 제안한다. 제안방안은 엔터티가 입력으로 주어지면, 가능한 모든 질의어(all possible queries)들을 생성하고, 일반 검색 엔진을 통해서 검색 결과를 얻는다. 본 연구의 실험에 의하면, 웹 페이지의 빈도수(frequency)가 높을수록 그 웹 페이지는 해당 엔터티와 관련성이 높은 상관관계(correlation)를 관찰하였다. 따라서 웹 페이지들의 빈도수를 고려하여 Frequent Pattern (FP)-Tree를 생성하고, 이를 바탕으로 각 질의어의 중요도를 측정하는 새로운 알고리즘을 제안한다. 그리고 중

† 비 회 원 : 다음소프트 마이닝랩 연구원
†† 정 회 원 : 군산대학교 통계컴퓨터과학과 조교수
††† 종신회원 : 삼육대학교 컴퓨터학부 교수
Manuscript Received : February 1, 2016
First Revision : April 8, 2016
Accepted : May 11, 2016
* Corresponding Author : On Byung-Won(bwon@kunsan.ac.kr)

1) 사람, 장소, 물건, 사건, 개념과 같은 사물의 실체를 의미한다.

Fig. 1. An example of the entity search to the Chinatown movie

요도가 가장 높은 질의어를 베스트 질의어로 출력하게 된다. 영화와 휴대폰 엔터티들을 사용하여 제안방안을 실험한 결과, 기존방안에 비해 평균 5배의 정확도를 향상시켰다.

다음 장에서는 최근 엔터티 검색의 관련연구를 토의한다. 그리고 3장에서는 본 연구의 제안방안을 자세히 설명하고 4장에서 실험 환경과 결과를 토의한다. 마지막으로 결론과 향후 연구에 대해 설명한다.

2. 관련 연구

2007년에 T. Cheng et al.은 엔터티 검색 엔진에 대한 연구의 필요성을 제기하였다. 기존의 검색 엔진은 검색 키워드를 통해 관련있는 웹 페이지들을 검색 결과로 출력하였다. 기존 검색 엔진에서는 동일한 엔터티를 여러 웹 페이지에서 부분적으로 다루고 있는데 이를 개선하여 여러 웹 페이지의 내용을 통합하여 한 페이지로 출력할 수 있도록 웹에 있는 정보를 통합(integration)하는 기술을 소개하였다[4]. 또한 전통적인 검색 엔진이 문서의 중요도를 측정하기 위해 페이지랭크(PageRank) 또는 HITS 알고리즘을 사용하였다면, T. Cheng et al.은 엔터티의 랭킹(ranking)을 측정할 수 있는 확률 모델을 제안하였다[5]. 또한 대용량(large-scale) 웹 데이터로부터 엔터티 검색을 할 수 있는 프로토타입 검색 엔진을 개발하였다[3]. K. Balog et al.은 엔터티 검색을 위한 개선된 확률 모델을 제안하였다. 질의어의 유사도에 따라 엔터티 랭킹이 수행되고, 질의어의 유사도는 확률 분포의 유사도에 의해 결정된다[1]. 그러나 이러한 확률 모델을 만들기 위해서는 잘 정리된 학습 데이터(training set)가 필요하다. G. Hu et al.은 엔터티 검색을 위해 지도 학습(supervised learning) 방법을 사용하였다[8]. 단어(word), 문단(passage), 위치(position), 메타데이터(metadata), 문서에 포함된 표(table)과 같은 특징(feature)들을 고안하여 기계학습(machine learning)에 적용하였다. 2012년에 S. Endrullis et al.은 다중 질의어 생성기(multiple query generators)를 사용하여 질의어를 생성하고 랭킹을 계산하기 위해 Apriori 알고리즘을 사용하였다. 그리고 상위 랭크된 질의어를 사용하여 엔터티 검색 엔진으로부터 웹 페이지를 검색하고 해당 엔터티와 매칭 하는 알고리즘을 제안하였다[6]. 이 방안은 다중 질의어를 고려하는 점에서 제안방안과 유사한 접근 방식을 취했지만, 실험을 통해 본 연구의 제안방안이 보다 우수함을 입증하였다. R. Blanco et al.은 Resource Description Framework (RDF) 데이터에서 엔터티 검색을 지원하기 위해 효율적이고 효과적인 방안을 제안하였다[2]. 특히 BM25F 랭킹 알고리즘이 RDF 데이터에 적용되었다. 원래 엔터티 검색은 웹상에 있는 전문가의 정보를 찾기 위해 처음 시작되었다. M. Ikeda et al.은 2단계 클러스터링 알고리즘을 사용하여 person name disambiguation 문제에 대한 해결 방안을 제안하였다[9]. E. Elmacioglu et al.은 tokens, named entities, host names and domains, URL 등의 다양한 특징들을 사용하는 클러스터링 알고리즘을 제안하고, person name disambiguation 문제에 적용하였다[6]. J. Lee와 S. Cheon은 다양성을 가진 방대한 양의 정보에서 원하는 정보를 검색하기 위해서는 많은 시간과 노력이 필요하며, 이를 해결하기 위해 검색엔진을 효과적으로 사용할 수 있는 검색어 질의 확장 방법을 제안하였다[11]. S. Yoon은 정보검색 시스템의 질의어와 색인에 기반을 둔 검색 과정에서 나타나는 중의성 해소를 위해 질의어 의미정보와 사용자 피드백을 사용하여 검색 성능을 향상시켰다[12].

3. 제안 방안

3.1 문제 정의

하나의 엔터티가 입력으로 주어지면, 그 엔터티와 관련된 모든 가능한 질의어들(all possible queries)을 생성한다. 제안방안은 모든 가능한 질의어 중에서 관련된 문서들(relevant web pages)을 가장 잘 찾아주는 상위 질의어들(top-k queries)를 출력한다. 영화와 같은 특정 도메인은 엔터티들의 집합으로 구성된다. 엔터티는 식별자(identifier)와 엔터티의 특성을 설명하는 애트리뷰트 세트(a set of attributes)로 구성된다. 예를 들면, '7번방의 선물' 엔터티는 식별자로 영화명(7번방의 선물)과 그 엔터티를 설명하는 영화명(7번방의 선물), 감독(이환경), 배우(류승룡), 국적(한국), 장르(코미디), 등급(15세 관람가) 등의 애트리뷰트 세트로 이루어진다. 본 연구에서 엔터티는 미리 주어졌다고 가정한다.

3.2 질의어 생성 및 웹 페이지 검색

하나의 엔터티에 대해 모든 가능한 질의어는 엔터티의 애트리뷰트들을 조합(combination)하여 생성한다. 이와 같이 k

개의 애트리뷰트들을 가진 엔터티의 생성 가능한 모든 질의어의 수는 2^k-1이다. 각 질의어를 사용하여 네이버 검색엔진을 통해 상위 랭크된 10개의 웹 페이지들을 검색 결과로 얻는다.[2)]

3.3 제안방안의 가설 및 검정

제안방안의 목표는 입력으로 주어진 엔터티의 베스트 질의어를 자동으로 찾는 것이다. 베스트 질의어를 사람의 개입 없이 자동으로 찾는 알고리즘을 개발하기 위해서 다음과 같은 가설을 제안한다.

Table 1. An example of computing the frequency of web pages

Query	Search Results	
q_1	w_1	w_2
q_2		w_2
q_3	w_1	w_2
$freq(w_i)$	2	3

1) 만일 웹 페이지 w_i가 m개의 서로 다른 질의어들을 사용하여 검색된 top-10 웹 페이지 리스트에 포함된다면, w_i의 빈도수는 m이다. 본 논문에서는 $freq(w_i)=m$이라고 표기한다. 예를 들면, Table 1에서 보는 것처럼, 3개의 다른 질의어들(q_1,q_2,q_3)과 각 질의어에 대한 검색 결과에 대해, $freq(w_1)=2$와 $freq(w_2)=3$이 된다.
2) 만일 $freq(w_1)<freq(w_2)$이면, w_2는 w_1보다 엔터티와 좀 더 관련성이 있는(relevant) 웹 페이지라고 가정한다.

제안된 가설을 검정하기 위해 전체 웹 페이지 중에서 임의 추출(random sampling) 방식으로 100개의 웹 페이지를 추출하고, 수작업으로 각 웹 페이지가 관련성이 있는지를 조사하였다. Table 2에서 보는 것처럼 관련성이 있으면 집단 1(group 1)로 하고, 그렇지 않으면 집단 2(group 2)로 분류하여, 두 집단에 속하는 웹 페이지들의 빈도수(frequency)의 평균(μ) 차이를 비교하기 위해 t-Test[3)]를 수행한다. t-Test에는 크게 단일표본, 독립표본, 대응표본 방식이 있지만, 본 실험에서는 두 집단에서 공통적으로 가지는 독립변

Table 2. Sample data set for t-Test

Group	Frequency	Group	Frequency	Group	Frequency	Group	Frequency
2	1	2	4	1	44	1	84
2	1	2	4	1	45	1	86
2	1	2	5	1	54	1	87
1	1	2	5	1	56	1	87
1	1	2	5	1	56	1	88
1	1	2	6	1	56	1	88
2	1	2	6	1	57	1	89
2	1	2	7	1	63	1	89
2	1	2	7	1	64	1	89
2	1	2	7	1	65	1	89
2	1	2	8	1	67	1	90
2	1	2	11	1	69	1	90
2	1	2	11	2	69	1	93
2	1	2	11	1	70	1	97
2	2	2	12	1	70	1	99
2	2	2	13	1	73	1	109
2	2	2	15	1	74	1	115
2	2	2	15	1	75	1	116
2	3	2	21	2	76	1	123
2	3	1	22	1	78	1	123
2	3	1	24	1	78	1	143
2	3	1	32	1	79	1	212
2	4	1	33	1	79	1	221
2	4	1	34	1	79	1	225
1	4	1	40	1	84	1	111

2) [10]에서 대부분의 웹 이용자는 검색 엔진을 사용할 경우에 상위 랭크된 10개의 웹 페이지만을 고려한다는 통계에 관한 연구이다. 이러한 연구를 바탕으로 본 논문에서는 상위 랭크된 10개의 웹 페이지만을 고려한다.

3) t-Test는 단일 집단 및 두 집단 간의 평균 또는 표준편차의 차이가 통계적으로 유의한지를 파악하는 통계적 검정기법임

수인 빈도수에 대한 평균 차이를 검증하는 것이기 때문에 독립표본 t-Test를 사용한다. t-Test를 시행하기에 앞서 몇 가지 조건을 고려하는데, 첫 번째 조건은 전체 웹 페이지는 정규분포를 따른다고 가정하고, 표본 개수가 두 집단 별로 50개에 근접한 것을 고려하여 표본은 t-분포를 따른다. 두 번째 조건은 집단 1과 2의 정확한 분산을 알지 못하기 때문에 Levene의 등분산 검정을 통해 분산의 동일성 결과에 따라 t-Test 결과를 도출한다. 마지막으로 귀무가설(H_0)과 대립가설(H_1)은 아래와 같이 설정한다.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \ (\text{두 집단간의 평균의 차이가 없음}) \\ H_1 : \mu_1 \neq \mu_2 \ (\text{두 집단간의 평균의 차이가 있음}) \end{cases}$$

이러한 가설을 세우고 통계 소프트웨어인 SPSS 21을 사용하여 표본으로부터 t-Test를 실시하였다(유의수준 $\alpha = 0.05$).

Table 3. Sample statistics

Group	N	Mean	Standard Deviation	Standard Error of Mean
frequency 1	58	78.74	45.492	5.973
frequency 2	42	8.57	15.579	2.404

Table 3과 Table 4는 집단통계량과 독립표본 검정을 나타내는 실험 결과이다. 집단통계량은 표본에서 집단 1에 속하는 웹 페이지 수가 58건, 집단 2에 속하는 웹 페이지의 수가 42건을 나타낸다. 그리고 각 집단의 빈도수 평균은 78.74와 8.57이기 때문에 집단 1에 비해 집단 2의 평균이 통계적으로 유의하게 낮다고 할 수 있다. 또한 두 집단의 빈도수의 크기는 집단 1의 빈도수 평균이 훨씬 높음을 알 수 있다. 하지만 단순한 표본의 집단통계량을 통해 전체 집단의 차이를 판단할 수 없기 때문에 독립표본 검정표를 분석한다.

유의확률의 Levene의 등분산 검정은 집단 1과 집단 2의 분산이 동일한지를 검증하는 것이고, Levene의 등분산 검정을 실시하는 이유는 두 집단의 분산이 동일한지, 않은지에 따라 t-Test의 검정 결과가 달라지기 때문이다. t값은 평균의 차이를 판단하기 위해 사용되며, 값이 클수록 신뢰성 있는 결과를 얻을 수 있다. 유의확률(양쪽)은 Levene의 등분산 검정 결과에 맞는 유의확률을 나타내고 유의확률의 결과가 낮을수록 신뢰할 수 있다. 본 실험 결과는 유의확률이 0.0000미만이기 때문에 분산이 같지 않다고 결론내릴 수 있다. 이처럼 등분산이 가정되지 않는 상황에서 t값은 10.898이며, 유의확률은 0.0000미만이기 때문에 독립표본 t-Test의 귀무가설을 기각하고, 대립가설을 채택한다.

이러한 가설을 바탕으로 제안 알고리즘은 웹 페이지의 빈도수를 고려하여 각 질의어의 중요도를 0~1 사이의 값으로 계산하고, 가장 높은 중요도를 가진 질의어를 베스트 질의어로 출력한다.

3.4 제안 알고리즘

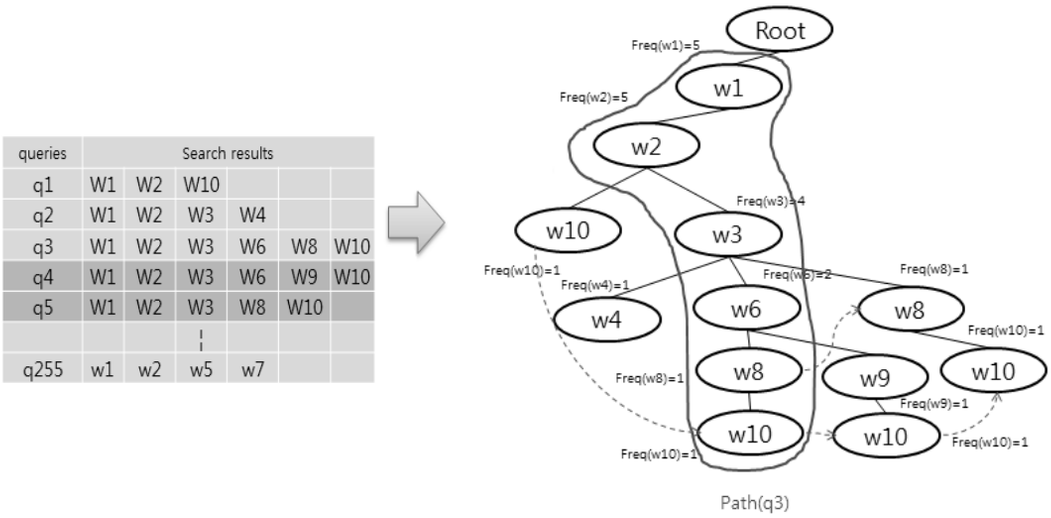
제안방안은 크게 4단계로 구성된다. (1) 검색 결과의 전처리 단계, (2) Frequent Pattern (FP)—Tree 생성, (3) 질의어의 중요도 측정, (4) 베스트 질의어 출력.

검색 결과의 전처리 단계에서는 Fig. 2A와 같이 각 질의어에 대한 검색 결과를 포함하는 행렬(matrix)를 생성한다. 만일 한 엔터티가 8개의 애트리뷰트들을 가진다면, 255개의 서로 다른 질의어가 생성된다. 그리고 네이버 검색 엔진을 통해 각 질의어에 대한 상위 랭크된 10개의 웹 페이지들을 얻는다. 각 웹 페이지는 URL로 나타나고, 하나의 웹 페이지는 서로 다른 질의어들에 의해 검색될 수 있다. 전처리 단계에서는 웹 페이지의 빈도수를 고려하여 웹 페이지들을 내림차순으로 정렬한다. Fig. 2A에서 q_1 질의어를 사용하여 w_1, w_2, w_{10}의 웹 페이지들을 얻었다. 또한 w_1와 w_2의 웹 페이지의 빈도수는 w_{10}의 빈도수보다 높다.

Fig. 2A 행렬을 이용하여 Fig. 2B에서 보는 FP-Tree를 생성한다. 트리는 루트 노드(root node)부터 시작하여 각 질의어에 대해 위에서 아래로 패스(path)를 생성한다. 예를 들면, q_1 질의어의 검색 결과는 w_1, w_2, w_{10} 웹 페이지들이다. 따라서 루트 노드의 자식 노드는 w_1이 되고, w_1의 자식 노드는 w_2가 되며, 차례로 w_2의 자식 노드는 w_{10}이 된다. w_{10} 웹 페이지는 FP-Tree의 단말 노드(leaf node)가 된다. 동일한 방식으로 q_2 질의어의 검색 결과를 이용하여 FP-Tree에 (root)—(w_1)—(w_2)—(w_3)—(w_4) 패스를 추가한다. 이때 FP-Tree의 w_2 노드는 2개의 자식 노드(w_3, w_{10})를 가진다. 행렬에 있는 255개의 질의어에 대한 패스를 추가하여 최종적인 FP-Tree를 생성한다. 이때 FP-Tree의 각 노드는 웹 페이지의 빈도수를 노드의 가중치(weight) 값으로 가진다. 예를 들면, w_1 노드는 $freq(w_1) = 5$를 가진다. w_{10} 노드는 다른 패스에 있는 w_{10} 노드에 점선 화살표로 연결된다. 이러한 화

Table 4. t-Test in independent samples

	Levene의 등분산 검정		평균의 동일성에 대한 t-검정						
	F	유의확률	t	자유도	유의확률 (양쪽)	평균차	차이의 표준 오차	차이의 95% 신뢰구간	
								하한	상한
frequency 등분산이 가정됨	16.429	.000	9.586	98	.000	70.170	7.320	55.644	84.696
frequency 등분산이 가정되지 않음			10.898	74.249	.000	70.170	6.439	57.341	82.999



(A) Query-web page matrix (B) FP-Tree

Fig. 2. An example of the proposed method

살표는 다른 패스에 존재하는 동일 노드를 빠르게 찾기 위한 포인터 역할을 한다.

이러한 FP-Tree를 이용하여 각 질의어의 중요도를 계산하기 위해 아래와 같은 수식을 제안한다.

$$weight(path|e) = \alpha \times T_1 + (1-\alpha) \times T_2 \quad (1)$$

$$T_1 = \frac{\sum_{j=1}^{n} freq(w_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} freq(w_{ij})} \quad (2)$$

$$T_2 = \frac{|D(e) \cap C(path_l)|}{|D(e)|} \quad (3)$$

$weight(path|e)$는 주어진 엔터티 e에 대한 패스 path(질의어)의 중요도를 0~1 사이의 값으로 나타낸다. Equation (1)은 Equation (2)와 Equation (3)으로 나누어지고, α 값을 사용하여 Equation (2)와 Equation (3)의 중요도를 결정한다. 이러한 α 값은 실험을 통해 정해진다.[4] α 값이 클수록 Equation (1)이 Equation (2)보다 상대적으로 중요하다. 반대로 α 값이 작으면 Equation (2)가 Equation (1)보다 중요하다. $\alpha = 0.5$인 경우에는 Equation (1)과 Equation (2)는 동등한 중요성을 가진다. Equation (2)에서 m은 FP-Tree에 있는 모든 패스들의 수, n은 한 패스에 있는 웹 페이지들의 수, w_{ij}는 i번째 패스의 j번째 웹 페이지를 나타낸다. 따라서 Equation (2)는 한 패스에 존재하는 모든 노드의 빈도수를 합하여 정규화(normalize)한 값이다. 3.3절의 제안방안의 가설에 의하면, 한 노드(웹 페이지)의 빈도수가 높을수록, 엔터티 e와 관련성(relevance)가 높으며, 패스에 있는 모든 노드의 빈도수의 합이 클수록 그 패스에 해당하는 질의어는 베스트 질의어일 확률이 높다. 반면, FP-Tree의 단말 노드들은 적은 빈도수를 가진다. 본 연구의 가설에 의하면 엔터티 e와의 관련성이 높지 않을 것이다. 따라서 적은 빈도수를 가지는 단말 노드가 실제로 엔터티와 관련성이 있는지를 확인하기 위해 단말 노드와 엔터티의 유사도(similarity)를 측정한다. Equation (3)에서 $D(e)$는 엔터티 e의 모든 애트리뷰트 값을 하나의 스트링(string)으로 나타낸 것이다. 또한 $path_l$은 한 패스의 단말 노드를 의미하고, $C(path_l)$는 단말 노드(예를 들면, q_2에 있는 w_4)가 포함하고 있는 단어 리스트를 나타낸다. 따라서 Equation (3)은 $D(e)$와 단말 노드에 모두 포함되는 단어들의 개수를 정규화한 값으로, 엔터티의 애트리뷰트들과 단말 노드의 내용이 얼마나 유사한지를 측정한다. Equation (3)의 값이 클수록 엔터티와 단말 노드는 관련성이 매우 높다는 것을 알 수 있다.

Equation (1)을 사용하여 FP-Tree의 각 패스(질의어)에 대한 중요도인 $weight(path|e)$를 계산한다. 그리고 중요도에 따라 255개의 질의어들을 내림차순으로 정렬한다. 그리고 $weight(path|e)$ 값이 큰 상위 top-k 질의어들을 추출하여 베스트 질의어로 출력한다.

4. 실 험

4.1 실험 환경

제안 알고리즘의 정확성을 측정하기 위하여 영화와 휴대폰 등 2개의 다른 데이터 세트를 사용하였다. Table 5에서 보는 것처럼, 영화 데이터 세트에는 50개의 최신 영화 엔터티들이 있고, 하나의 엔터티는 8개의 애트리뷰트들을 가지고 있기 때문에, 모든 가능한 질의어의 수는 255개이다. 또한 Table 6에 정리된 휴대폰 데이터 세트는 8개의 엔터티들

4) 다양한 α 값을 실험하였으나 $\alpha = 0.5$일 때 가장 좋은 성능을 보인다.

Table 5. Movie data set

No	Movie name	Director	Actor	Nationality	Genre	Rating level	Running time	Year
1	7번방의 선물	이환경	류승룡	한국	코미디	15세 관람가	127분	2013
2	감기	김성수	장혁	한국	드라마	15세 관람가	121분	2013
3	감시자들	조의석 김병서	설경구	한국	범죄	15세 관람가	119분	2013
4	강남	유하	이민호	한국	액션	청소년 관람불가	135분	2015
5	겨울왕국	크리스 벅 제니퍼 리	크리스틴 벨	미국	애니메이션	전체 관람가	108분	2014
6	공범	국동석	손예진	한국	스릴러	15세 관람가	95분	2013
7	관상	한재림	송강호	한국	드라마	15세 관람가	139분	2013
8	국제시장	윤제균	황정민	한국	드라마	12세 관람가	126분	2014
9	군도	윤종빈	하정우	한국	액션	15세 관람가	137분	2014
10	그래비티	알폰소 쿠아론	산드라 블록	미국	SF	12세 관람가	90분	2013
11	기술자들	김홍선	김우빈	한국	액션	15세 관람가	116분	2014
12	남자가 사랑할 때	한동욱	황정민	한국	드라마	15세 관람가	120분	2014
13	노아	대런 아로노프스키	알렉스 카터	미국	모험	15세 관람가	139분	2014
14	논스톱	자움 콜렛 세라	리암 니슨	미국	액션	15세 관람가	106분	2014
15	더 테러 라이브	김병우	하정우	한국	스릴러	15세 관람가	97분	2013
16	맨 오브 스틸	잭 스나이더	헨리 카빌	미국	액션	12세 관람가	143분	2013
17	명량	김한민	최민식	한국	액션	15세 관람가	128분	2014
18	몽타주	정근섭	엄정화	한국	스릴러	15세 관람가	120분	2013
19	박수건달	조진규	박신양	한국	코미디	15세 관람가	127분	2013
20	베를린	류승완	하정우	한국	액션	15세 관람가	120분	2013
21	변호인	양우석	송강호	한국	드라마	15세 관람가	127분	2013
22	빅매치	최호	이정재	한국	액션	15세 관람가	112분	2014
23	설국열차	봉준호	송강호	한국	SF	15세 관람가	125분	2013
24	소원	이준익	설경국	한국	드라마	12세 관람가	122분	2013
25	수상한 그녀	황동혁	심은경	한국	코미디	15세 관람가	124분	2014
26	숨바꼭질	허정	손현주	한국	스릴러	15세 관람가	107분	2013
27	스파이	이승준	설경구	한국	코미디	15세 관람가	121분	2013
28	신세계	박훈정	이정재	한국	범죄	청소년 관람불가	134분	2013
29	신의 한 수	조범구	정우성	한국	액션	청소년 관람불가	118분	2014
30	쎄시봉	김현석	정우	한국	멜로	15세 관람가	122분	2015
31	아이언맨 3	쉐인 블랙	로버트 다우니 주니어	미국	액션	12세 관람가	129분	2013
32	어메이징 스파이더맨 2	마크 웹	앤드류 가필드	미국	액션	12세 관람가	142분	2014
33	어바웃 타임	리차드 커티스	돔놀 글리슨	영국	멜로	15세 관람가	123분	2013
34	역린	이재규	현빈	한국	드라마	15세 관람가	135분	2014
35	연애의 온도	노덕	이민기	한국	멜로	청소년 관람불가	112분	2013
36	오늘의 연애	박진표	이승기	한국	멜로	15세 관람가	118분	2015
37	용의자	원신연	공유	한국	액션	15세 관람가	137분	2013
38	월드 워 Z	마크 포스터	브래드 피트	미국	드라마	15세 관람가	115분	2013
39	은밀하게 위대하게	장철수	김수연	한국	액션	15세 관람가	123분	2013
40	전설의 주먹	강우석	황정민	한국	액션	청소년 관람불가	153분	2013
41	조선명탐정	김석윤	김명민	한국	코미디	12세 관람가	125분	2015
42	집으로 가는 길	방은진	전도연	한국	드라마	15세 관람가	131분	2013
43	친구 2	곽경택	유오성	한국	느와르	청소년 관람불가	124분	2013
44	컨저링	제임스 완	베라 파미가	미국	공포	15세 관람가	112분	2013
45	터보	데이빗 소렌	라이언 레이놀즈	미국	애니메이션	전체 관람가	95분	2013
46	패션왕	오기환	주원	한국	드라마	15세 관람가	114분	2014
47	퍼시픽 림	길예르모 델 토로	찰리 헌냄	미국	액션	12세 관람가	131분	2013
48	표적	창	류승룡	한국	액션	15세 관람가	98분	2014
49	해적	이석훈	김남길	한국	모험	12세 관람가	130분	2014
50	허삼관	하정우	하정우	한국	드라마	12세 관람가	124분	2015

Table 6. Cellphone data set

No	Model name	Maker	Year	RAM	OS	Weight	Screen size	Battery
1	갤럭시S6	삼성전자	2015년4월	3GB	안드로이드5.0	138g	12.95cm	2550mAh
2	갤럭시S6엣지	삼성전자	2015년5월	3GB	안드로이드5.0	132g	12.95cm	2600mAh
3	G4	LG전자	2015년4월	3GB	안드로이드5.1	155g	13.97cm	3000mAh
4	G3	LG전자	2014년5월	3GB	안드로이드4.4	149g	13.88cm	3000mAh
5	갤럭시S5	삼성전자	2014년3월	2GB	안드로이드4.4	145g	12.95cm	2800mAh
6	갤럭시노트4	삼성전자	2014년9월	3GB	안드로이드4.4	176g	14.48cm	3220mAh
7	iPhone6	애플	2014년1월	1GB	iOS8	112g	11.9cm	2915mAh
8	iPhone5	애플	2012년9월	1GB	iOS8	112g	10.16cm	1440mAh

이 있고 휴대폰 엔터티는 모델명, 기업, 출시연도, RAM, 운영체제, 무게, 화면크기, 배터리 등 총 8개의 애트리뷰트들을 포함하고 있다. 예를 들면, 갤럭시S6은 모델명="갤럭시S6", 기업="삼성전자", 출시연도="2015년 4월", RAM="3GB", 운영체제="안드로이드5.0", 무게="138g", 화면크기="12.95cm", 배터리="2550mAh" 등의 애트리뷰트 값을 가진다. 마찬가지로 하나의 휴대폰 엔터티로부터 255개의 질의어들이 생성된다.

Fig. 3에서 보는 것처럼, '베를린' 영화 엔터티의 3개의 애트리뷰트는 영화이름="베를린", 영화배우="하정우", 장르="액션"이라고 가정하면, 3개의 애트리뷰트를 조합하여 다양한 질의를 아래와 같이 생성할 수 있다.

- 질의어1: <베를린>,
- 질의어2: <하정우>,
- 질의어3: <액션>,
- 질의어4: <베를린+하정우>,
- 질의어5: <베를린+액션>,
- 질의어6: <하정우+액션>,
- 질의어7: <베를린+하정우+액션>

총 $2^3-1=7$개의 서로 다른 질의어가 생성된다.

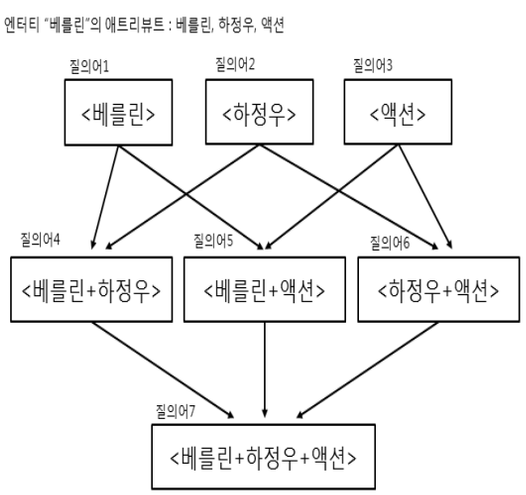


Fig. 3. Attribute value lattice

이와 같은 방식으로 Table 5와 Table 6에 있는 각 엔터티의 8개의 애트리뷰트 값을 조합하여 $2^8-1=255$개의 모든 가능한 질의어들을 생성한다.

생성된 모든 질의어들은 파싱 라이브러리인 JSoup과 네이버 OpenAPI를 사용하여 네이버 검색 엔진으로부터 각 질의어 별로 상위 10개의 웹 페이지들을 수집한다. 이런 방식으로 수집된 모든 웹 페이지들을 매뉴얼하게 조사하여 각 웹 페이지가 해당 엔터티와 관련이 있는지를 조사한다. 다음과 같은 기준을 적용하여, 각 웹 페이지가 엔터티와 관련이 있는지를 판단한다.

1) 웹 페이지에 엔터티와 애트리뷰트, 그리고 질의어와 같은 키워드가 있으면, 그 웹 페이지는 엔터티와 관련 있는 것으로 간주한다. 예를 들면, '베를린' 영화 엔터티일 경우에 웹 페이지에 '베를린', '하정우', '액션' 등과 같은 키워드가 있어야 한다.
2) 웹 페이지에 엔터티에 관한 설명이 있으면, 그 웹 페이지는 엔터티와 관련 있는 것으로 한다. 예를 들면, 영화 엔터티인 경우에 영화에 관한 줄거리, 영화에 관한 인터뷰 등이 웹 페이지에 나와야 한다.
3) 웹 페이지에 다른 엔터티를 설명하는 경우에는 그 웹 페이지는 해당 엔터티와 관련이 없는 것으로 간주한다. 예를 들면, '베를린' 영화 엔터티에서 독일의 수도인 베를린의 정보가 있는 웹 페이지는 영화 엔터티와 관련이 없다. 또한 질의어에 영화배우 하정우가 있을 경우에 하정우가 출연한 다른 영화들을 설명하는 웹 페이지는 해당 엔터티와 관련이 없다.
4) 광고성 웹 페이지는 엔터티와 관련이 없다.

관련 있는 웹 페이지를 수집하여 '골드 스탠더드 세트(gold standard set; GSS)'로 정의한다. GSS는 해당 엔터티와 관련 있는 모든 웹 페이지들을 포함하는 실제 정답 세트이다. 질의어마다 검색된 상위 10개의 웹 페이지들을 '검색된 문서 집합(retrieved document set; RDS)'라고 하면, 질의어들의 정확성을 평가하기 위해 정밀도(precision), 재현율(recall), 조화평균(F-measure) 값들을 측정할 수 있다. 정밀도는 검색된 웹 페이지들 중 관련 있는 웹 페이지들의 비율이고, 재현율은 관련 있는 웹 페이지들 중 실제로 검색된

웹 페이지들의 비율이다. 조화평균은 정밀도와 재현율의 정확도를 판단하는 기준으로 정밀도와 재현율의 값의 차이가 크면 조화평균의 값은 낮아지고, 정밀도와 재현율의 값이 유사하면 조화평균의 값은 크다. 어떤 질의어 q에 대한 정밀도, 재현율, 조화평균은 다음과 같이 계산할 수 있다.

$$\text{정밀도}(q) = \frac{|GSS \cap RDS|}{|RDS|} \tag{4}$$

$$\text{재현율}(q) = \frac{|GSS \cap RDS|}{|GSS|} \tag{5}$$

$$\text{조화평균}(q) = \frac{2 \times \text{정밀도}(q) \times \text{재현율}(q)}{\{\text{정밀도}(q) + \text{재현율}(q)\}} \tag{6}$$

제안방안으로 Algorithm 1을 구현하고 실험하였다.

4.2 실험 결과

영화 엔터티에 관한 정보를 얻기 위해 일반 사용자들은 <영화명>, <배우명>, <'영화'+영화명>, <영화명+배우명>과 같은 질의어를 사용하는 것이 일반적이다. 이처럼 사용자들이 주로 사용하는 질의어들과 제안방안을 통해 얻어진 베스트 질의어에 대한 정밀도, 재현율, 조화평균 값을 측정하였다. Fig. 4는 사용자들이 흔히 사용하는 질의어들과 제안방안으로부터 얻어진 베스트 질의어에 대해서 50개의 엔터티들의 정밀도, 재현율, 조화평균 값들의 평균을 보여준다. <영화명+배우명>과 <'영화'+영화명>의 질의어들은 평균 정밀도, 재현율, 조화평균 값은 0에 가깝다. 이것은 네이버 검색 엔진에서 <'영화'+영화명>의 질의어를 사용하여 상위 랭크된 10개의 웹 페이지 중에 찾고자 하는 영화 엔터티와 관련 있는 웹 페이지가 거의 없다는 것을 의미한다. 다른 질의어인 <영화명>과 <배우명>을 사용한 경우에도 평균적으로 0.1을 넘지 않는다. 반면에 제안방안으로부터 도출된 베스트 질의어의 정밀도, 재현율, 조화평균의 평균값은 0.59, 0.74, 0.61이다. 이러한 결과는 <영화명> 질의어와 비교할 경우에 약 6배 향상된 결과이고, <배우명> 질의어와 비교할 때, 약 4배 향상된 수치이다.

사용자들이 주로 사용하는 일반적인 질의어들의 정확성이 낮은 이유는 크게 2가지로 요약할 수 있다. 첫 번째 이유는 엔터티의 의미가 모호하기 때문이다. 예를 들면, 베를린 영화 엔터티에 대한 정보를 찾기 위해 <베를린> 질의어를 사용할 경우에는 네이버나 구글 검색 엔진은 영화 베를린의 정보가 아닌 독일의 수도 베를린에 대한 정보, 그리고 베를린 여행 정보 등 영화 엔터티와 관련 없는 정보들을 담고 있는 웹 페이지들을 상위 랭크 시킨다. 두 번째 이유는 <영화명> 또는 <배우명>과 같은 질의어를 통해 검색된 웹 페이지는 찾고자 하는 영화 엔터티와 관련 없는 사건이나 사고를 설명하는 것이 주를 이루는 경우이다.

Algorithm 1. Suggesting the top-k best queries, given a target entity

알고리즘	Top_k_Best_Queries()	
입력	엔터티 e($a_1, a_2, ..., a_k$)	// e : 엔터티 // a_i : i번째 애트리뷰트
출력	Top-k 베스트 질의어들	
1:	$Q = 2^k - 1$ 질의어 생성;	// Q : e의 애트리뷰트들을 조합하여 생성된 질의어의 집합
2:	$W = \varnothing$;	// W : 웹 페이지들의 집합
3:	for each query $q \in Q$ search_engine(q) $\rightarrow$ W_q; $W = W \cup W_q$;	// W_q : q에 의해서 검색된 상위 10개 웹 페이지들
4:	for each webpage $w \in W$ $freq(w)$값 계산;	// w : 웹 페이지 // $freq(w)$: w 웹 페이지의 빈도수
5:	FP-Tree = empty;	
6:	for each W_q $freq(w)$값에 의해 W_q에 있는 웹 페이지들을 내림차순으로 정렬; FP-Tree에 추가 (Fig. 2 참조);	
7:	for each query q Equation (1)을 사용하여 $weight(q)$값 측정;	// $weight(q)$: q의 중요도
8:	$weight(q)$값에 의해 $2^k - 1$ 쿼리들을 내림차순으로 정렬;	
9:	상위 k개의 쿼리 선정;	

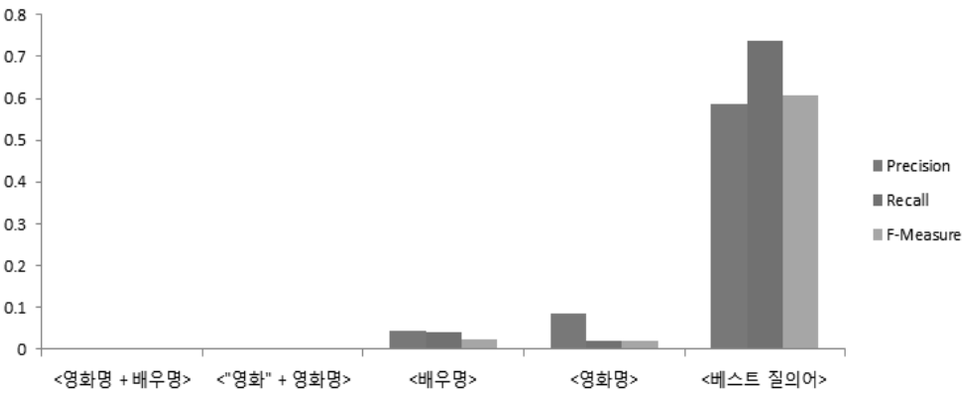


Fig. 4. Precision/Recall/F-measure of entity search queries (Movie data set)

류승룡 '7번방의 선물' 러닝개런티 10억...감독은 18억

입력 2013-02-02 11:57:00

사진 류승룡, 동아외전DB.

류승룡이 영화 '7번방의 선물'로 받은 러닝개런티가 10억6000만원에 이르는 것으로 드러났다.

연출자인 이환경 감독은 18억 원을 받았고, 또 다른 배우 정진영은 5억2000만원의 러닝개런티를 받았다.

영화 흥행에 따라 배우들이 따로 받는 '보너스' 개념인 러닝개런티가 구체적으로 공개되는 건 이례적이다.

이번 '7번방의 선물'의 경우 공동제작사간 진행되고 있는 법적 분쟁 과정을 통해 드러났다.

2013년 1월 개봉해 누적관객 1281만 명을 모은 '7번방의 선물'은 한국영화로는 역대 흥행 순위 3위에 올라있다.

개봉 당시 누구도 예상하지 못했던 흥행 성적으로 화제를 모았고, 주인공 류승룡은 이 영화를 통해 흥행 배우로

Fig. 5. An example of an irrelevant web page with the movie entity entitled 'Miracle in Cell No.7'

Fig. 5에서 보는 것처럼, <영화명> 또는 <배우명> 질의어를 통해 얻은 웹 페이지 가운데 하나로 영화 엔터티 자체에 대한 설명이라기보다는 류승룡 배우가 받는 러닝개런티에 대한 설명이 주를 이루고 있다. 이에 더하여 휴대폰의 경우에는 아이폰4와 아이폰5의 경우처럼 한 휴대폰 엔터티에 대해 시리즈 별로 각각의 엔터티가 존재하는데 이것은 검색을 어렵게 하는 요소이다.

Fig. 6은 <영화명>, <배우명>, <베스트 질의어>에 대한 산점도(scatter plot)를 보여준다. x축과 y축은 정밀도와 재현율을 가리킨다. 그림에서 A, B, C는 각각 <영화명>, <배우명>, <베스트 질의어>를 사용한 결과이다. A 산점도에서는 전체적으로 모든 영화 엔터티의 정밀도와 재현율은 낮지만, <조선명탐정> 질의어의 경우에는 정밀도와 재현율은 매우 높음을 알 수 있다. 반면에 <감기>와 <국제시장>과 같은 질의어의 정밀도와 재현율은 매우 낮다. 감기와 국제시장이라는 단어는 영화 제목이라기보다는 다른 의미로 널리 통용되는 용어이기 때문에 이러한 영화명을 질의어로 사용할 경우에는 정밀도와 재현율이 상당히 낮음을 알 수 있다. 일반적으로 영화명은 대중에게 친숙하게 다가올 수 있도록 익숙한 용어를 사용하며 긴 문장보다는 단문을 사용하여 대중의 뇌리에 쉽게 기억될 수 있도록 하는 것이 특징이다. 따라서 영화명을 질의어로 사용하여 엔터티 검색을 수행할 경우에는 찾고자 하는 엔터티와 관련 있는 웹 페이지를 찾기가 쉽지 않음을 산점도를 통해 쉽게 알 수 있다. Fig. 6B 산점도는 50개의 영화 엔터티에 대해 <배우명> 질의어를 통해 웹 페이지를 검색하고 그 질의어의 정밀도와 재현율을 산점도로 나타내었다. A 산점도와 유사하게, '관상' 엔터티의 경우에 정밀도와 재현율이 크게 낮다. 그 이유는 관상의 주연배우인 송강호는 이미 수많은 다른 영화에 출연하였기 때문에 <송강호>라는 질의어로 웹 페이지를 검색할 경우에 다른 웹 페이지들이 검색될 확률이 높다. 반면에 '조선명탐정' 영화의 경우에는 주연배우인 김명민이 다른 영화에 출연한 빈도가 낮기 때문에 배우명을 질의어를 사용하더라도 높은 정밀도와 재현율을 보이는 것을 알 수 있다. 주목할 점은 A와 B 산점도에서 50개 엔터티의 대부분은 0값을 가진다. C 산점도의 경우에는 50개의 영화 엔터티에 대해서 제안방안으로부터 도출된 베스트 질의어를 사용하여 얻은 정확도와 재현율을 산점도로 나타낸 것이다. 대부분의 영화 엔터티들의 정밀도와 재현율이 향상되었다. 또한 감기와 기술자들과 같은 영화 엔터티들의 경우에는 <영화명>과 <배우명>을 사용할 경우에는 정밀도와 재현율이 낮았지만, 제안방안에서는 높은 정밀도와 재현율을 보이는 것을 알 수 있다. 이러한 영화 엔터티들은 모호한 의미를 가지고 있기 때문에 <영화명>이나 <배우명>을 사용하면 해당 엔터티와

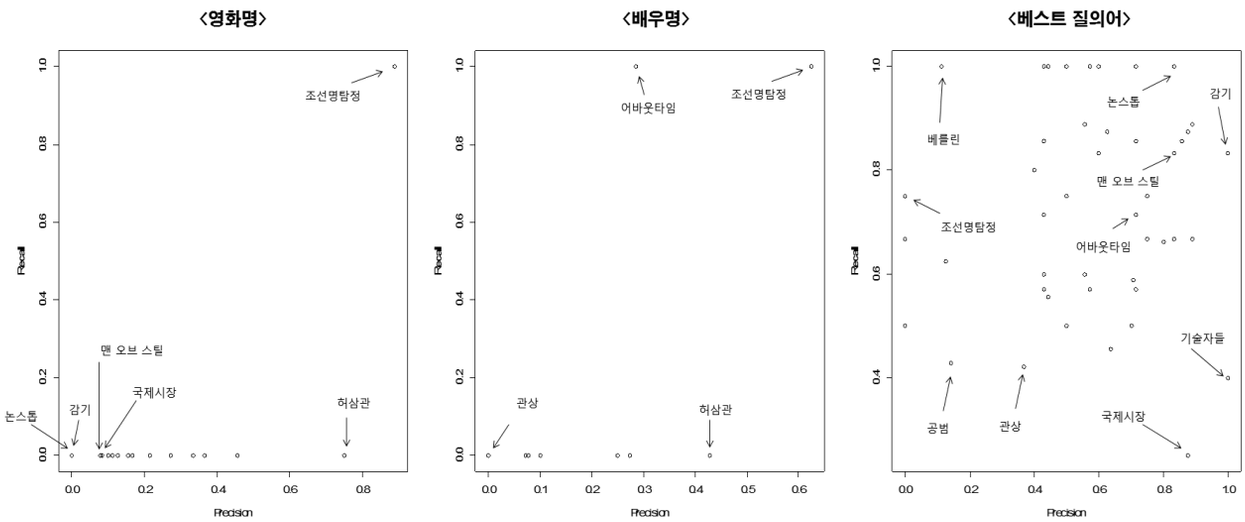


Fig. 6. Scatter plots of entity search queries (A: 〈Movie name〉, B: 〈Actor name〉, C: 〈Best query〉)

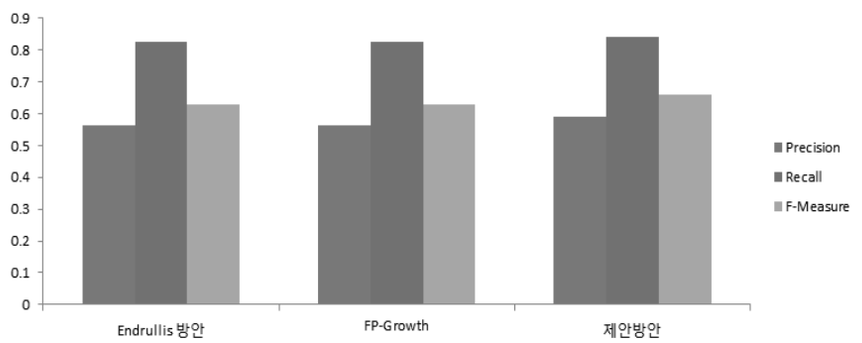


Fig. 7. Precision/Recall/F-measure of entity search methods (Movie data set)

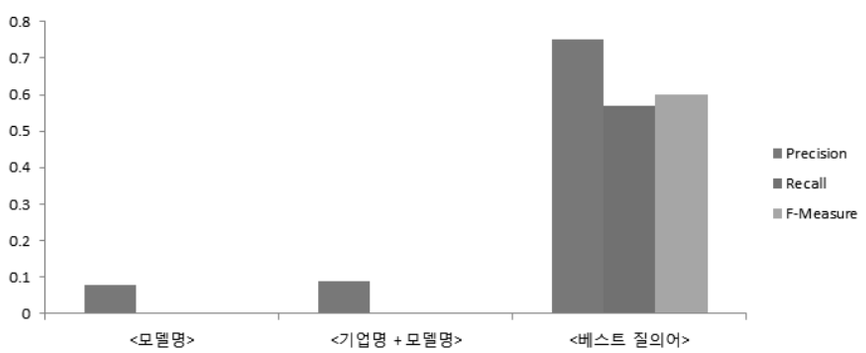


Fig. 8. Precision/Recall/F-measure of entity search queries (Cellphone data set)

관련 없는 웹 페이지를 검색하게 된다. 하지만 제안방안에 의해서 추출된 베스트 질의어를 사용할 경우에는 높은 정확도를 보이게 된다. C 산점도에서 베를린 영화의 경우에는 재현율은 높지만 상대적인 정밀도는 높지 않았다. 그 이유는 베를린 영화 엔터티와 관련 있는 웹 페이지보다는 베를린 수도 엔터티 등 다른 엔터티에 관련 있는 웹 페이지들이 웹상에 많이 존재하기 때문이다. 따라서 상위 랭크된 10개의 웹 페이지 중에서 영화 엔터티가 아닌 다른 엔터티들이 검색됨을 알 수 있다. 그러나 중요한 점은 재현율이 높다는 것이다. 이것은 제안된 베스트 질의어를 사용할 경우에 웹상에 존재하는 베를린 영화 엔터티와 관련 있는 거의 모든 웹 페이지를 모두 찾을 수 있다는 것을 의미한다. 반대로 국제시장의 경우에는 재현율은 낮지만 상대적으로 정밀도는 높았다. 그 이유는 앞서 설명한 베를린 영화 엔터티와는 다르게

Table 7. Precision/Recall/F-measures of the best query to each movie entity

No	Movie name	Best query	Precision	Recall	F-measure
1	7번방의 선물	<이환경+류승룡+15세 관람가+2013>	0.43	0.60	0.50
2	감기	<감기+김성수+장혁+한국+15세 관람가>	1.00	0.83	0.91
3	감시자들	<감시자들+설경구+15세 관람가+2013>	0.71	0.59	0.64
4	강남	<이민호+액션+청소년 관람불가+135분>	0.80	0.66	0.72
5	겨울왕국	<크리스틴 벨+전체 관람가>	0.57	1.00	0.73
6	공범	<국동석+손예진+한국+스릴러+2013>	0.14	0.43	0.21
7	관상	<송강호+드라마+139분>	0.37	0.42	0.39
8	국제시장	<윤제균+황정민+12세 관람가+2014>	0.88	0.25	0.39
9	군도	<군도+하정우+한국+액션+15세+관람가>	0.89	0.67	0.76
10	그래비티	<그래비티+산드라 블록+미국+SF+2013>	0.64	0.45	0.53
11	기술자들	<김홍선+김우빈+15세 관람가+116분>	1.00	0.40	0.57
12	남자가 사랑할 때	<남자가 사랑할 때+한동욱+한국+15세 관람가+120분>	0.50	0.75	0.60
13	노아	<대런+아로노프스키+모험>	0.43	0.57	0.49
14	논스톱	<논스톱+15세 관람가+106분>	0.83	1.00	0.91
15	더 테러 라이브	<더 테러 라이브+김병우+하정우+스릴러+15세 관람가+97분>	0.71	1.00	0.83
16	맨 오브 스틸	<맨 오브 스틸+잭 스나이더+12세 관람가+143분>	0.83	0.83	0.83
17	명량	<명량+김한민+한국+15세 관람가+2014>	0.50	0.50	0.50
18	몽타주	<엄정화+15세 관람가+2013>	0.44	1.00	0.62
19	박수건달	<박수건달+15세 관람가>	0.75	0.67	0.71
20	베를린	<베를린+류승완+15세 관람가+2013>	0.11	1.00	0.20
21	변호인	<변호인+송강호+한국+127분>	0.60	0.83	0.70
22	빅매치	<최호+액션+15세 관람가+112분>	0.56	0.60	0.58
23	설국열차	<송강호+SF+15세 관람가+125분>	0.71	1.00	0.83
24	소원	<소원+이준익+드라마+122분>	0.75	0.75	0.75
25	수상한 그녀	<황동혁+심은경+코미디+15세 관람가+124분>	0.63	0.88	0.73
26	숨바꼭질	<허정+손현주+스릴러+15세 관람가>	0.70	0.50	0.58
27	스파이	<이승준+설경구+한국+코미디+15세 관람가+121분>	0.43	1.00	0.60
28	신세계	<박훈정+이정재+한국+범죄+134분>	0.60	1.00	0.75
29	신의 한 수	<조범구+청소년+관람불가+118분>	0.75	0.75	0.75
30	쎄시봉	<한국+멜로+15세 관람가+122분>	0.50	1.00	0.67
31	아이언맨 3	<로버트 다우니 주니어+12세 관람가+129분>	0.43	0.86	0.57
32	어메이징 스파이더맨 2	<어메이징 스파이더맨 2+마크 웹+12세 관람가>	0.89	0.89	0.89
33	어바웃 타임	<돔놀 글리슨+123분>	0.71	0.71	0.71
34	역린	<역린+현빈+한국+15세 관람가+2014>	0.86	0.86	0.86
35	연애의 온도	<연애의 온도+이민기+청소년 관람불가+2013>	0.43	0.71	0.54
36	오늘의 연애	<박진표+이승기+멜로+15세 관람가+2015>	0.88	0.88	0.88
37	용의자	<용의자+원신연+한국+15세 관람가+137분>	0.56	0.89	0.68
38	월드 워 Z	<월드 워 Z+브래드 피트+드라마+15세 관람가>	0.71	0.57	0.63
39	은밀하게 위대하게	<은밀하게 위대하게+장철수+한국+15세 관람가+123분>	0.83	0.67	0.74
40	전설의 주먹	<강우석+황정민+청소년 관람불가+153분>	0.57	0.57	0.57
41	조선명탐정	<한국+12세 관람가+125분>	0.00	0.75	0.00
42	집으로 가는 길	<집으로 가는 길+드라마+15세 관람가>	0.44	0.56	0.49
43	친구 2	<친구 2+유오성+한국+느와르+청소년 관람불가>	0.71	0.86	0.78
44	컨저링	<컨저링+미국+공포+15세 관람가+2013>	0.00	0.67	0.00
45	터보	<터보+미국+애니메이션+전체 관람가>	0.00	0.50	0.00
46	패션왕	<주원+드라마+15세 관람가+2014>	0.88	0.88	0.88
47	퍼시픽 림	<퍼시픽 림+길예르모 델 토로+찰리 헌냄+12세 관람가+131분>	0.71	1.00	0.83
48	표적	<창+류승룡+액션+15세 관람가+98분>	0.50	0.75	0.60
49	해적	<한국+모험+130분>	0.13	0.63	0.21
50	허삼관	<하정우+드라마+12세 관람가+124분>	0.40	0.80	0.53
평균			0.59	0.74	0.61

Table 8. Precision/Recall/F-measures of the best query to each cellphone entity

No	Cellphone name	Best query	Precision	Recall	F-Measure
1	갤럭시S6	<갤럭시S6+삼성전자+안드로이드5.0+2550mAh>	0.88	0.71	0.79
2	갤럭시S6엣지	<갤럭시S6엣지+삼성전자+3GB+안드로이드5.0>	0.80	0.63	0.70
3	G4	<G4+안드로이드5.1+3000mAh>	0.67	1.00	0.80
4	G3	<G3+3GB+149g+3000mAh>	0.90	0.44	0.60
5	갤럭시S5	<2014년3월+2GB+안드로이드4.4+145g>	1.00	0.67	0.80
6	갤럭시노트4	<갤럭시노트4+안드로이드4.4+176g>	0.44	0.75	0.56
7	iPhone6	<iPhone6+1GB+2915mAh>	0.78	0.14	0.24
8	iPhone5	<iPhone5+2012년9월+1GB>	0.50	0.25	0.33
평균			0.75	0.57	0.60

웹상에 거의 대부분의 웹 페이지들은 국제시장 영화 엔터티와 관련 있다. 따라서 정밀도는 매우 높다. 반면에 재현율이 떨어진 이유는 본 연구에서 상위 랭크된 10개의 웹 페이지만 찾았기 때문에 다른 관련 있는 웹 페이지들은 제외되었기 때문이다. 이러한 결과를 통해 베스트 질의어의 우수성을 간접적으로 증명할 수 있다.

Fig. 7은 기존방안과 제안방안의 정확도를 비교한 그림이다. 기존방안으로는 Endrullis et al.이 Apriori 알고리즘을 사용하여 엔터티 검색을 수행하였다. 또한 연관 규칙 마이닝(association rule mining)에서 Apriori 알고리즘보다 우수한 FP-Growth 알고리즘이 많이 사용된다. 이러한 기존의 알고리즘들과 본 논문에서 제안한 FP-Tree 기반의 알고리즘의 정확도를 비교하였다. 그림에서 보는 것처럼, Endrullis et al. 알고리즘과 FP-Growth 알고리즘의 정확도는 크게 차이가 없으며, 제안방안은 기존방안에 비해 5%의 정확도 향상을 보였다. Apriori 알고리즘을 사용한 Endrullis 방안과 FP-Growth 알고리즘과 같은 기존 방안의 경우에는 그 알고리즘의 특성상 가장 높은 빈도를 가진 소수의 웹 페이지들을 이용하지만, 제안방안의 알고리즘은 높은 빈도를 가진 대부분의 웹 페이지들을 이용하여 질의어의 중요도를 측정하기 때문에 기존 방안보다 높은 성능을 보였다.

제안방안이 다른 데이터 세트에서도 유용한지를 검사하기 위해 휴대폰 데이터 세트에서 제안방안을 실행하고 그 결과를 Fig. 8에서 정리하였다. 영화 데이터 세트에서처럼, 사용자들이 일반적으로 사용하는 <모델명>과 <기업명+모델명> 등의 질의어와 제안방안에서 추천된 베스트 질의어에 대한 정밀도, 재현율, 조화평균 값의 평균을 측정하였다. 영화 데이터 세트에서처럼, 제안방안의 정확도는 다른 일반 질의어에 비해 높은 성능을 보였다. 이것은 휴대폰 엔터티도 영화 엔터티와 비슷한 문제를 가지고 있으며 제안방안을 사용하여 엔터티 검색의 성능을 높일 수 있다. 특이한 점은 휴대폰에 대한 정밀도와 재현율은 영화 엔터티에서는 반대로 나타난다. 즉 휴대폰에서는 정밀도가 낮고 재현율이 높은 반면, 영화 엔터티에서는 정밀도가 높고, 재현율이 상대적으로 낮다. 그 이유는 영화 엔터티의 경우에는 질의어들이 다양한 엔터티와 관련 있는 경우가 많다. 예를 들면, 베를린, 감기 등과 같은 영화 엔터티이외에 다른 엔터티와 관련 있는 웹 페이지가 다수 존재하지만, 휴대폰의 경우에는 대부분의 웹 페이지가 휴대폰 엔터티와 관련 있기 때문에, 정밀도와 재현율이 두 데이터 세트에서 상이하게 측정되었다.

참고로 Table 7과 Table 8은 제안방안에서 도출한 베스트 질의어의 정확도와 Table 9와 Table 10은 top-3 베스트 질의어에 대한 결과를 정리하여 보여준다.

5. 결 론

엔터티 검색에서 한 엔터티가 다양한 의미를 지니고 있다면 엔터티 검색의 정확성은 크게 떨어진다. 이러한 문제를 해결하기 위해, 본 논문에서는 애트리뷰트의 조합을 통해 다양한 질의어를 생성하고, 각 질의어를 사용하여 일반 검색 엔진으로부터 웹 페이지들을 수집한다. 본 연구에 따르면, 웹 페이지의 빈도수가 높을수록 해당 엔터티와 관련성이 많다는 사실을 관찰하였고, 이를 바탕으로 FP-Tree 기반의 각 질의어의 중요도를 측정하는 새로운 모형을 제안하고 베스트 질의어를 추출하였다. 그리고 영화와 휴대폰 데이터 세트에서 기존 방안과 제안방안의 성능을 측정하여, 제안방안의 우수성을 입증하였고, 제안방안의 정확도가 높은 이유를 자세히 살펴보았다.

향후 연구로는 질의어의 중요도를 측정하는데 좀 더 정교한 알고리즘을 개발하는 것이 필요하다. 예를 들면, FP-Tree에서 확률 기반의 모형 또는 다이내믹 프로그래밍 방법을 사용하여 좀 더 세밀하게 질의어의 중요도를 수치화하는 알고리즘을 개발할 예정이다. 또한 단어 중의성 해소(word sense disambiguation) 기법을 적용하거나 제안방안에서 얻은 결과를 피드백으로 사용하여 검색 키워드의 정확성을 높이는 연구를 수행할 것이다.

Table 9. A list of the highly ranked queries by the proposed method (Movie data set)

No	Entity	Top-1 query	Top-2 query	Top-3 query
1	7번방의 선물	<이환경+류승룡+15세 관람가+2013>	<7번방의 선물+류승룡+코미디+15세 관람가+2013>	<이환경+코미디+15세 관람가+2013>
2	감기	<감기+김성수+장혁+한국+15세 관람가>	<감기+김성수+한국+15세 관람가>	<감기+김성수+드라마+15세 관람가>
3	감시자들	<감시자들+설경구+15세 관람가+2013>	<감시자들+범죄+15세 관람가+2013>	<감시자들+119분>
4	강남	<이민호+액션+청소년 관람불가+135분>	<강남+유하+청소년 관람불가+135분>	<강남+유하+이민호+액션+135분>
5	겨울왕국	<크리스틴 벨+전체 관람가>	<크리스틴 벨+애니메이션+전체 관람가>	<겨울왕국+크리스 벅+제니퍼 리+108분>
6	공범	<국동석+손예진+한국+스릴러+2013>	<공범+국동석+한국+2013>	<공범+국동석+손예진+스릴러+2013>
7	관상	<송강호+드라마+139분>	<한재림+송강호+드라마+139분>	<관상+한국+드라마+139분>
8	국제시장	<윤제균+황정민+12세 관람가+2014>	<윤제균+황정민+한국+12세 관람가>	<윤제균+한국+드라마+2014>
9	군도	<군도+하정우+한국+액션+15세 관람가>	<군도+윤종빈+하정우+한국+액션+15세 관람가>	<하정우+한국+액션+137분>
10	그래비티	<그래비티+산드라 블록+미국+SF+2013>	<알폰소 쿠아론+미국+SF>	<알폰소 쿠아론+산드라 블록+2013>
11	기술자들	<김홍선+김우빈+15세 관람가+116분>	<기술자들+김홍선+액션+15세 관람가+116분>	<기술자들+김홍선+김우빈+15세 관람가+116분>
12	남자가 사랑할 때	<남자가 사랑할 때+한동욱+한국+15세 관람가+120분>	<남자가 사랑할 때+한동욱+한국 120분>	<남자가 사랑할 때+황정민+한국+15세 관람가+120분>
13	노아	<대런+아로노프스키+모험>	<모험+15세 관람가+139분>	<노아+모험+15세 관람가>
14	논스톱	<논스톱+15세 관람가+106분>	<논스톱+자움 콜렛 세라+액션>	<논스톱+액션+15세 관람가+106분>
15	더 테러 라이브	<더 테러 라이브+김병우+하정우+스릴러 15세 관람가+97분>	<더 테러 라이브+스릴러+15세 관람가+97분>	<더 테러 라이브+김병우+스릴러+15세 관람가+97분>
16	맨 오브 스틸	<맨 오브 스틸+잭 스나이더+12세 관람가+143분>	<맨 오브 스틸+헨리 카빌+12세 관람가+143분>	<맨 오브 스틸+잭 스나이더+헨리 카빌+12세 관람가+143분>
17	명량	<명량+김한민+한국+15세 관람가+2014>	<명량+김한민+최민식+한국+액션+15세 관람가+2014>	<김한민+최민식+한국+15세 관람가+2014>
18	몽타주	<엄정화+15세 관람가+2013>	<몽타주+15세 관람가+2013>	<정근섭+스릴러+2013>
19	박수건달	<박수건달+15세 관람가>	<박수건달+박신양+코미디+15세 관람가>	<박수건달+조진규+박신양+15세 관람가>
20	베를린	<베를린+류승완+15세 관람가+2013>	<류승완+하정우+한국+15세 관람가+120분>	<류승완+하정우+한국+액션+15세 관람가+120분>
21	변호인	<변호인+송강호+한국+127분>	<변호인+양우석+한국+127분>	<변호인+양우석+송강호+한국+127분>
22	빅매치	<최호+액션+15세 관람가+112분>	<빅매치+최호+이정재+15세 관람가+112분>	<빅매치+이정재+15세 관람가+112분>
23	설국열차	<송강호+SF+15세 관람가+125분>	<봉준호+SF+15세 관람가+125분>	<봉준호+송강호+SF+125분>
24	소원	<소원+이준익+드라마+122분>	<소원+이준익+드라마+12세 관람가>	<소원+드라마+122분>
25	수상한 그녀	<황동혁+심은경+코미디+15세 관람가+124분>	<황동혁+심은경+15세 관람가+124분>	<황동혁+심은경+코미디+15세 관람가>
26	숨바꼭질	<허정+손현주+스릴러+15세 관람가>	<숨바꼭질+허정+손현주+스릴러+107분>	<허정+손현주+107분>
27	스파이	<이승준+설경구+한국+코미디+15세 관람가+121분>	<이승준+설경구+코미디+15세 관람가+121분>	<스파이+이승준+설경구+한국+15세 관람가+121분>
28	신세계	<박훈정+이정재+한국+범죄+134분>	<신세계+한국+134분>	<신세계+한국+범죄+134분>
29	신의 한 수	<조범구+청소년 관람불가+118분>	<조범구+118분>	<신의 한 수+한국+118분>
30	쎄시봉	<한국+멜로+15세 관람가+122분>	<쎄시봉+멜로+15세 관람가+122분>	<쎄시봉+김현석+멜로+15세 관람가+122분>
31	아이언맨 3	<로버트 다우니 주니어+12세 관람가+129분>	<아이언맨 3+12세 관람가+129분>	<아이언맨 3+로버트 다우니 주니어+12세 관람가+129분>
32	어메이징 스파이더맨 2	<어메이징 스파이더맨 2+마크 웹+12세 관람가>	<어메이징 스파이더맨 2+미국+액션>	<어메이징 스파이더맨 2+마크 웹+앤드류 가필드+2014>
33	어바웃 타임	<돔놀 글리슨+123분>	<돔놀 글리슨+영국+15세 관람가>	<돔놀 글리슨+15세 관람가+123분>
34	역린	<역린+현빈+한국+15세 관람가+2014>	<역린+드라마+15세 관람가+2014>	<이재규+15세 관람가+2014>
35	연애의 온도	<연애의 온도+이민기+청소년 관람불가+2013>	<연애의 온도+한국+청소년 관람불가+2013>	<연애의 온도+노덕+한국+청소년 관람불가+2013>
36	오늘의 연애	<박진표+이승기+멜로+15세 관람가+2015>	<오늘의 연애+이승기+멜로+15세 관람가>	<오늘의 연애+이승기+멜로+15세 관람가+2015>
37	용의자	<용의자+원신연+한국+15세 관람가+137분>	<용의자+원신연+한국+액션+15세 관람가+137분>	<용의자+원신연+공유+액션+15세 관람가>
38	월드 워 Z	<월드 워 Z+브래드 피트+드라마+15세 관람가>	<월드 워 Z+브래드 피트+미국+15세 관람가>	<월드 워 Z+미국+15세 관람가>
39	은밀하게 위대하게	<은밀하게 위대하게+장철수+한국+15세 관람가+123분>	<장철수+한국+액션+15세 관람가+123분>	<은밀하게 위대하게+15세 관람가+123분>
40	진설의 주먹	<강우석+황정민+청소년 관람불가+153분>	<강우석+액션+청소년 관람불가+153분>	<강우석+황정민+한국+청소년 관람불가>
41	조선명탐정	<한국+12세 관람가+125분>	<한국+코미디+12세 관람가+125분>	<조선명탐정+한국+125분>
42	집으로 가는 길	<집으로 가는 길+드라마+15세 관람가>	<방은진+드라마+15세 관람가>	<집으로 가는 길+드라마+131분>
43	친구 2	<친구 2+유오성+한국+느와르+청소년 관람불가>	<유오성+느와르+124분>	<유오성+느와르+청소년 관람불가+124분>
44	컨저링	<컨저링+미국+공포+15세 관람가+2013>	<컨저링+미국+15세 관람가+2013>	<제임스 완+미국+15세 관람가+2013>
45	터보	<터보+미국+애니메이션+전체 관람가>	<터보+애니메이션+전체 관람가>	<터보+데이빗 소렌+미국>
46	패션왕	<주원+드라마+15세 관람가+2014>	<오기환+드라마+2014>	<패션왕+15세 관람가+2014>
47	퍼시픽 림	<퍼시픽 림+길예르모 델 토로+찰리 헌냄+12세 관람가+131분>	<퍼시픽 림+길예르모 델 토로+찰리 헌냄+131분>	<길예르모 델 토로+찰리 헌냄+미국+12세 관람가>
48	표적	<창+류승룡+액션+15세 관람가+98분>	<창+류승룡+15세 관람가>	<창+류승룡+15세 관람가+98분>
49	해적	<한국+모험+130분>	<해적+이석훈+모험+130분>	<해적+김남길+한국+12세 관람가+130분>
50	허삼관	<하정우+드라마+12세 관람가+124분>	<허삼관+124분>	<허삼관+하정우+하정우+124분>

Table 10. A list of the highly ranked queries by the proposed method (Cellphone data set)

No	Entity	Top-1 query	Top-2 query	Top-3 query
1	갤럭시S6	<갤럭시S6+삼성전자+안드로이드5.0+2550mAh>	<갤럭시S6+삼성전자+안드로이드5.0>	<삼성전자+안드로이드5.0+12.95cm>
2	갤럭시S6엣지	<갤럭시S6엣지+삼성전자+3GB+안드로이드5.0>	<갤럭시S6엣지+삼성전자+안드로이드5.0>	<갤럭시S6엣지+안드로이드5.0+12.95cm>
3	G4	<G4+안드로이드5.1+3000mAh>	<G4+3GB+안드로이드5.1>	<G4+LG전자+3GB+안드로이드5.1>
4	G3	<G3+3GB+149g+3000mAh>	<G3+LG전자+3GB+149g+3000mAh>	<G3+LG전자+3GB+3000mAh>
5	갤럭시S5	<2014년3월+2GB+안드로이드4.4+145g>	<갤럭시S5+2GB+안드로이드4.4+145g>	<갤럭시S5+삼성전자+145g>
6	갤럭시노트4	<갤럭시노트4+안드로이드4.4+176g>	<갤럭시노트4+삼성전자+176g>	<갤럭시노트4+2014년9월+안드로이드4.4+176g>
7	iPhone6	<iPhone6+1GB+2915mAh>	<iPhone6+1GB+112g+2915mAh>	<iPhone6+2915mAh>
8	iPhone5	<iPhone5+2012년9월+1GB>	<iPhone5+애플+2012년9월+1GB>	<iPhone5+애플+2012년9월+iOS8+1GB>

References

[1] K. Balog, M. Bron, and M. Rijke, "Query modeling for entity search based on terms, categories, and examples," *The ACM Transactions on Information Systems*, Vol.29, No.4, pp.22, 2011.

[2] R. Blanco, P. Mika, and S. Vigna, "Effective and efficient entity search in RDF data," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011.

[3] T. Cheng, X. Yan, and K. Chang, "Supporting entity search: A large-scale prototype search engine," in *Proceedings of ACM SIGMOD/PODS Conference*, Beijing, China, 2007.

[4] T. Cheng and K. Chang, "Entity search engine: Towards agile best-effort information integration over the web," in *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research*, CA, USA, 2007.

[5] T. Cheng, X. Yan, and K. Chang, "EntityRank: Searching entities directly and holistically," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, Vienna, Austria, 2007.

[6] S. Endrullis, A. Thor, and E. Rahm, "Entity search strategies for mashup applications," in *Proceedings of IEEE 28th International Conference on Data Engineering*, Washington DC, USA, 2012.

[7] E. Elmacioglu, Y. Tan, S. Yan, M. Kan, and D. Lee, "PSNUS: Web people name disambiguation by simple clustering with rich features," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech, 2007.

[8] G. Hu, J. Liu, H. Li, Y. Cao, J. Nie, and J. Gao, "A supervised learning approach to entity search," *Information Retrieval Technology*, Vol.4182, pp.54-66, 2006.

[9] M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa, "Person name disambiguation on the web by two-stage clustering," in *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, 2009.

[10] B. Jansen and A. Spink, "An analysis of web documents retrieved and viewed," in *Proceedings of the 16th International Conference on Internet Computing and Big Data*, NV, USA, 2003.

[11] J. Lee and S. Cheon, "Recommendation query ranking system for the search query expansion," *Journal of KIISE*, Vol.36, No.2(c), 2009.

[12] S. Yoon, "Using query word senses and user feedback to improve precision of search engine," *Journal of Korea Society for Information Management*, Vol.26, No.4, pp.81-91, 2009.

이 선 구

e-mail : devleesk@daumsoft.com

2016년 군산대학교 통계컴퓨터과학과(학사)

2016년~현 재 다음소프트 마이닝랩 연구원

관심분야 : 데이터 마이닝, 텍스트 마이닝

온 병 원

e-mail : bwon@kunsan.ac.kr

1998년 안양대학교 컴퓨터공학과(학사)

2000년 고려대학교 컴퓨터학과(석사)

2007년 펜실베이니아주립대학교 컴퓨터공학과(박사)

2014년~현 재 군산대학교 통계컴퓨터과학과 조교수

관심분야 : 데이터 마이닝, 정보검색, 데이터베이스, 빅데이터

정 수 목

e-mail : jungsm@syu.ac.kr

1984년 경북대학교 전자공학과(학사)

1986년 경북대학교 컴퓨터공학과(석사)

2002년 고려대학교 컴퓨터학과(박사)

1986년~현 재 삼육대학교 컴퓨터학부 교수

관심분야 : 멀티미디어, 영상처리