질의어의 근접성 정보와 객체 그래프 모델링을 이용한 반구조 문서 검색 기법

박준영○, 한기준, 이문용
한국과학기술원 지식서비스공학과
{j.park89, keejun.han, munyi}@kaist.ac.kr

Semi-Structure Search Using Query Proximity and Entity Graph Modeling

Juneyoung Park○, Keejun Han, Mun Y. Yi
Department of Knowledge Service Engineering, KAIST

요 약

본 연구는 반구조적인 형태를 지니고 있는 객체를 대상으로, 질의어의 근접성 정보와 각 객체의 그래프 모델링을 통해 유사도를 도출하는 형식의 반구조적 객체검색에 최적화된 검색 모델을 제시한다. 기존의 질의어에 대한 통계 모델에 기반한 검색의 한계를 지적하며, 실제 이용자들의 질의어로 형성된 데이터를 이용하여 수행된 객관적인 성능 평가 실험을 통해 제안하는 질의어 근접성 정보 및 그래프 모델링을 이용한 검색의 우수성을 검증하였다.

1. 서 론

최근 인터넷의 발전과 스마트폰과 같은 휴대기기의 확산으로 온라인상의 정보와 각종 매체들의 검색과 이용이 급속하게 늘고 있다. 영화 혹은 E-book과 같은 자료들을 제공하는 서비스들이 늘어감에 따라 이용자가 원하는 미디어 자료들을 찾을 수 있는 검색 기술의 중요성이 대두되고 있다[1]. 일반적인 온라인 검색에서 이용되는 문서 자료들과 달리 미디어 자료들은 반구조(Semi-structure)적인 형태를 지니고 있는데 반구조적인 자료란 유연하고 정립되지 않은 형태를 지니고 있는 자료를 의미하며 SQL 데이터베이스와 같이 확립된 형태를 지니고 있는 구조적 자료 또는 형태를 아예 지니고 있지 않은 인터넷 문서와 같은 비구조적 자료들과도 그 형태에 있어서 차이를 두고 있다[2]. 이러한 반구조적 자료들에 대한 검색은 비구조적 혹은 구조적인 자료들과 그 접근에 있어서 차이를 두고 있는데, 정형화된 질의를 통해 검색하는 구조적 자료들과 단순 질의어를 이용하는 비구조적 자료들에 반해 반구조적 자료들은 추가적인 정보를 필요로 한다. 간략한 예로 하나의 영화에는 다양한 정보가 집합되어 있는데, 영화의 내용 뿐만 아니라 출연진, 제작진 혹은 장르 등 하나의 객체에 여러 종류의 정보가 유연하게 밀집되어있고 이러한 종류의 정보들 또한 검색에 사용된다. 그러나 일반적인 사용자들이 이러한 추가적인 정보를 반영하는 검색을 이용하게끔 하는 방법은 한계가 있으며, 추가적인 단계들을 거쳐야 하는 경우가 많아 사용자들의 긍정적인 반응을 얻지 못한다. 대신에, 사용자들에게 익숙한 형태인 질의어를 입력했을 때 추가적인 단계 없이 검색결과가 바로 제공되는 Ad-hoc 검색 방식을 유지하면서도 반구조적 자료에 대해 적합한 검색 방법은 사용자의 만족도를 향상시킬 수 있다[1,2,4].

*이 논문은 2014년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0024560).

본 연구에서는 반구조적 자료들을 질의어만을 통해 효율적으로 검색할 수 있는 방법을 제안한다. 온라인 상에 제공되고 있는 영화 자료들의 분석을 통하여 1) 반구조적인 자료를 표현하는 그래프 기반의 모델을 제안하고 2) 질의어의 근접성 점수(Proximity Score)를 이용하여 유사도를 평가하며 3) 종합적인 검색 결과를 제공한다.

2. 반구조 문서 특화 검색 기법

반구조적 자료들의 특징은 흔히 필드로 불리는 정형화되지 않은 구조를 지니고 있다는 점이다. 자료의 필드들을 검색에 활용하는 가장 대표적인 방식으로 language model을 개선하는 Mixed Language Model(MLM)[3]을 꼽을 수 있는데, 이는 질의어가 각각의 필드에서 대응될 확률을 통계학적으로 계산하여 통합하는 방식이다. MLM에서는 각각의 필드가 고정된 가중치를 갖고 있으며 가중치는 일괄적으로 계산에 활용되고 그에 따른 결과값을 반환한다. 그러나 고정된 가중치는 다양한 질의어와 많은 수의 필드에 적합하게 이용되기 어려우며 이를 개선하기 위한 노력으로 각 필드에서의 단어의 출현빈도를 통해 계산되는 확률로 가중치를 대신하는 Probabilistic Retrieval Model for Semi-structured Data(PRMS)[4]와 같은 방식도 제시되고 있다.

하지만 위와 같은 방식들은 모두 통계적인 방법을 통하여 질의어의 각 필드에 대한 가중치를 계산하여 이용하고자 한다는 점에서 그 한계가 있다. 반구조적 자료의 특징 중 하나인 유연한 구조로 인해 사용되는 용어 혹은 단어들이 시시각각 변할 수 있고 이러한 변화는 각 필드에 사용되는 단어들이 통계적으로 정형화되기 어려워진다. 본 연구에서는 기존 기술의 필드별 단어의 출현 확률을 이용한 통계학적 검색이 아닌 구조적인 필드를 그래프 기반의 형태로 취합하여 검색에 이용한다.

3. 제안하는 방법

본 연구에서 제안하는 방법은 크게 세 가지의 단계로 이루어져 있다: 1) 사용자의 질의어에 관련된 초기 검색결과를 반환한다, 2) 데이터 셋의 자료들을 그래프 기반의 형태로 변환한다, 3) 초기 검색결과 내 자료들의 그래프 형태에서 사용자의 질의어간의 근접성 정보를 통해 유사도를 계산하고 초기 검색결과와 융합하여 최종 검색 결과를 제공한다.

3.2 초기 검색결과 반환

제안하는 방법의 성능을 평가하기 위해서는 사용자의 질의 Q와 관련된 자료들을 찾아 초기 검색결과를 얻어와 제안하는 방법을 통해 순위를 재정렬하여야 한다. 그렇기 때문에 초기 검색결과를 얻기 위해서 기존 연구에의해 제시된 방법 중 가장 보편적으로 이용되는 BM25F[9] 기법을 사용한다. BM25F 기법은 BM25 기법을 반구조문서 검색에 최적화시킨 형태로써 필드 가중치를 사용하는 형태의 기법이다. 이렇게 얻어진 초기 검색결과는 Original Score을 통해 순위가 정렬되며 Original Score를 구하는 공식은 아래와 같다.

$$\text{BM25F} = \sum_{t \in q \cap d} \frac{tf(t,d)}{k1 + tf(t,d)} \times idf(t) \quad (1)$$

이때 사용자의 질의 q 에 속한 각각의 질의어 t를 문서 d 에 대하여 계산하는 경우, tf(t,d) 함수는 질의어 t가 문서 d에서의 출현빈도를 BM25F 기법으로 계산하는 함수이며 idf 함수는 문서셋 내에서 질의어 t의 출현빈도의 역수를 의미한다. 또한 k1 변수는 tf 함수의 증가치를 조절하기 위한 매게 변수이다. Original Score를 통해 얻어진 순위는 제안되는 방법을 통해 얻어지는 근접성 점수를 통해 재정렬되어 최종검색결과로써 반환된다.

3.2 그래프 기반 모델링

반구조적 자료의 형태를 그래프 상에서 표현하기 위해 각 자료들은 필드를 의미하는 클러스터들로 이루어진 1단계 그래프와 그 클러스터 내부에 존재하는 핵심개념들로 이루어진 2단계 그래프로 변환되는데, 1단계 그래프 내의 모든 클러스터들은 각각 한 개의 하위 2단계의 그래프를 포함하게 된다. 예를 들어 영화의 경우 감독, 출연진, 내용이라는 3개의 클러스터의 연관관계를 표현하는 1단계와 출연진 클러스터 내부의 배우들의 연관관계를 표현하는 2단계로 이루어져있다. 본 연구에서 제안하는 기법은 2단계의 그래프들로 이루어진 모델에서 질의어의 근접성 점수를 기반으로 유사도 계산이 이루어지게 된다. 사용자의 질의어의 근접성 점수를 계산하기 위해 2단계 그래프에서 질의어에 해당되는 객체를 찾게 되며 찾은 객체들간의 관계를 사용하게 된다. 이때 객체들이 같은 필드에 속해 있지 않을 경우 1단계 그래프의 근접성 정보를 사용하고 같은 필드에 속해 있을 때 2단계 그래프의 근접성 정보를 사용하게 된다. 이러한 계산을 위한 1단계와 2단계 그래프의 구성은 다음과 같다.

첫 단계인 상위 그래프에서 각 자료 u에 대해 그래프 G_u = (V,E)로 표현되며 V는 각각의 필드를 의미하는 클러스터들의 집합이며 E는 클러스터들간의 연결을 의미하는 선(Edge)을 의미하고 있다. V에 속하는 클러스터 v_1 과 v_2 의 관계를 의미하는 e 의 근접성 정보 $W_{v1,v2}$는 고정 점수로써 반구조적 자료에 있어서 각각의 부문들은 일정한 연관관계를 갖고 있다는 가정하에 지정되었다.

하위 단계인 2단계에서 각 클러스터 c는 그래프 G_c = (N,E)로 표현했을 때 N은 각 자료 내의 핵심 정보라고 판단되는 객체를 표현하는 노드(Node)들의 집합이며 E는 노드들간의 연결을 표현하는 선(Edge)이다. 이때 노드들을 연결하는 선은 각각의 근접성 정보를 갖게 된다. N에 속하는 노드 n1과 n2의 관계를 의미하는 선의 근접성 정보 $W_{n1,n2}$는 다음과 같다.

$$W_{n1,n2} = \text{co-occur}(n1,n2) \quad (2)$$

Co-occur은 각 부분내의 핵심정보 n1과 n2와 함께 사용되는 횟수를 나타내며 각 자료의 부문들을 문서화하여 분석함으로 형성되며 그 수가 낮을수록 높은 근접성을 갖고 있음을 의미한다.

3.3 질의어와 자료간의 유사도 계산

위에서 설명한 것과 같이 제안하는 방법의 최종 결과물로 반환되는 검색순위를 계산하기 위하여 사용자의 질의 Q에 대하여 초기 검색결과들과의 근접성 점수(Proximity score)를 계산하게 된다. 근접성 점수는 질의어간의 근접성 정보를 의미하며 그래프에서 노드들을 연결하는 선들의 가중치(weight)로 표현되고 두 질의어의 의미가 가깝고 연관관계가 높을수록 근접성 점수가 낮다. 근접성 점수를 구하는 공식은 다음과 같다.

$$\text{Proximity Score} = \frac{\sum_{n1,n2 \in N, n1 \neq n2} W_{n1,n2} \times 2}{MaxDistance \times Q_n} \quad (3)$$

이때 $W_{n1,n2}$의 경우 n_1과 n_2를 연결해주는 직접적인 선이 없을 때 Dijkstra's shortest path[6]를 이용하여 가장 근접한 path를 지정하여 $W_{n1,n2}$를 지정해주었으며 질의의 길이와 각 그래프가 갖고 있는 특성들을 고려하여 정규화하였다.

최종적으로 질의어가 그래프 기반 모델에서 갖는 근접성 점수 Proximity Score와 초기 검색결과에서 갖는 Original Score을 융합하여 최종 점수 Final Score을 계산하여 순위를 정렬하며 Final Score의 공식은 다음과 같다.

$$\text{FinalScore} = \text{EXP}(-\text{ProximityScore} \times \alpha) \times OriginalScore \quad (4)$$

여기서 지수 함수와 매개변수 α 는 Proximity Score 의 영향을 일정화 시킴으로써 그 크기에 상관없이 균일하게 최종결과 값에 영향을 미칠 수 있도록 조정하고 있다.

4. 실험

4.1 실험 설계

온라인상의 유명한 영화 매체인 IMDB.com 에서 선정한 매출 상위 1000개의 영화들을 실험 자료로 이용하여 본 연구의 성능을 검증하였다. 지정된 1000개의 영화들에 대하여 사용자의 질의를 얻기 위하여 크라우드 소싱 업체인 Amazon Mechanical Turk[10]를 이용하였으며, 총 355명의 사용자들이 1000개의 영화에 대해 제공한 6101개의 질의를 이용하여 본 연구의 성능을 검증하였다.

평가척도로써 Precision@N(1,2,5,10)과 Mean Reciprocal Rank(MRR)가 사용되었다. Precision@N의 경우 상위 N개의 결과값 내에 적합한 문서가 포함되었을 때 1의 수치로 표현되며, MRR의 경우 검색하고자 하는 문서의 순위를 r 이라고 했을 때 1/r 의 값을 갖게 된다.

4.2 실험 결과

앞서 설명한 바와 같이 1000개의 영화들에 대하여 6101개의 질의에 대해 두 가지 평가척도를 갖고 실험하였을 때 평가척도에 관계없이 본 연구에서 제시하는 방법이 기존의 BM25F 보다 우수한 성능을 보였으며 그 해당 실험결과를 표1에 정리하였다.

표1 두 반구조 특화 검색 기법간의 성능 차이

	P@10	P@5	P@2	P@1	MRR
BM25F	0.901	0.802	0.640	0.512	0.641
제안 기법	**0.945**	**0.865**	**0.708**	**0.568**	**0.697**

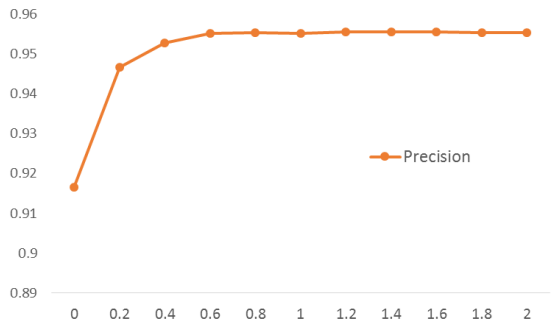
모든 평가척도에서 제안한 기법의 성능은 BM25F 기법과 비교하였을 때 그 성능이 통계학적으로 유의미한 향상(Wilcoxon Test, $p < 0.001$)을 보였고 평균적으로 약 8%의 향상을 보였다(Precision@N=8.57%, MRR=8.7%). 이러한 성능의 향상을 토대로 반구조 특화 검색에 대해 다음과 같은 분석 결과를 도출해낼 수 있다.

- 앞서 설명하였듯이 반구조 자료의 특성인 구조의 유연함으로 인하여 통계적 계산을 통해 특정 단어의 필드를 유추해내는 과정은 어려움이 많다. 그에 비교하였을 때 유연한 구조적인 특색을 반영한 그래프 기반의 검색 기법은 통계적 계산에서 생기는 오류들을 최소화할 수 있고 결과적으로 더 좋은 검색 성능을 보여준다.
- 또한 근접성 정보를 이용하는 제안된 방식이, 단어의 출현빈도를 이용하는 기존의 방식 대비, 더 정밀하게 검색하고자 하는 자료를 찾아내며 단어의 유무뿐만 아니라 단어들간의 연관관계까지 고려함으로써 아주 세밀한 평가척도에서도(P@1) 높은 성능 향상을 보여준다.

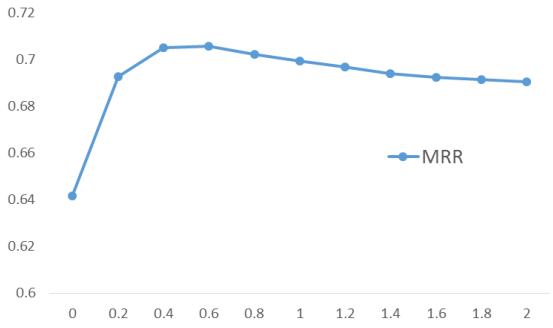
또한 그림 1의 a) 와 b)는 각각 Precision@N과 MRR에서 파라미터들의 영향을 기록한 것이다. 두 평가 척도에서 동일하게 관찰 할 수 있듯이 파라미터의 값의 변화에 의한 영향이 안정화 되어있으며 큰 변화 없이 결과값이 균일함을 의미한다. 안정화 되어있음은 즉 제안된 방법을 이용하여 더 많은 수의 질의 혹은 더 많은 수의 자료들을 검색하는 실험으로 확장을 하여도 같은 결과를 얻을 수 있음을 의미하고 있고 제안된 방법의 안정성 및 확장성의 우수함을 증명하고 있다.

그림1 파라미터 값 α의 변화에 따른 Precision@N과 MRR 척도 값의 변화

a) Precision@N



b) MRR



5. 결론

본 논문에서는 반구조 자료에 최적화된 그래프 모델링과 그래프 모델 안에서 질의어간의 근접성 정보를 이용한 반구조 자료 검색 기법을 제안하였다. 기존의 필드 가중치를 이용한 검색 기법과의 객관적인 성능 비교를 통하여 제안하는 방법의 우수성을 확인하였다. 제안하는 기법은 사용자들에게 추가적인 필드 입력 과정 없이 정확한 검색 결과를 제공하는 장점을 통해 급속도로 확산되고 있는 미디어 자료의 검색에 유용하게 활용될 수 있음을 실제 사용자들의 질의어로 형성된 데이터를 이용하여 본 연구는 분명하게 보여준다. 향후 연구에서는 제안하는 기법의 고도화에 대한 연구 및 실제 미디어 자료 서비스 환경에서의 실험이 필요하다.

참고문헌

[1] Liu, Jingjing, et al. "A Conversational Movie Search System Based on Conditional Random Fields." *INTERSPEECH*. 2012.

[2] Balog, Krisztian. "Semistructured Data Search." *Bridging Between Information Retrieval and Databases*. Springer Berlin Heidelberg, 2014. 74-96.

[3] Ogilvie, Paul, and Jamie Callan. "Combining document representations for known-item search." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003.

[4] Kim, Jinyoung, Xiaobing Xue, and W. Bruce Croft. "A probabilistic retrieval model for semistructured data." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2009. 228-239.

[5] Büttcher, Stefan, Charles LA Clarke, and Brad Lushman. "Term proximity scoring for ad-hoc retrieval on very large text collections." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006.

[6] Dijkstra, Edsger W. "A note on two problems in connexion with graphs." *Numerische mathematik* 1.1 (1959): 269-271.

[7] Elbassuoni, Shady, and Roi Blanco. "Keyword search over RDF graphs." *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011.

[8] Broschart, Andreas, and Ralf Schenkel. "Proximity-aware scoring for XML retrieval." *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008.

[9] Pérez-Agüera, José R., et al. "Using BM25F for semantic search." *Proceedings of the 3rd international semantic search workshop*. ACM, 2010.

[10] Amazon Mechanical Turk, https://www.mturk.com/