웹 검색 환경에 적용할 추론 망 기반 검색모델

최익규°, 김민구
아주대학교 정보통신전문대학원
(ikchoi, minkoo)@madang.ajou.ac.kr

Inference Network-Based Retrieval Model for Web Search Environment

Ik-Kyu Choi°, Min-Koo Kim
Graduate School of Information and Communication, Ajou University

요 약

대다수의 사용자는 웹 검색에서 자신이 찾고자 하는 것을 표현할 때, 평균 2, 3개의 단어를 사용하고 있다. 벡터 모델이나 추론 망 모델에서 이런 질의 정보를 이용하여 좋은 결과를 얻기에는 몇 가지 어려움이 있다. 특히 추론 망 모델에서 많이 사용되는 유사도 계산식인 weighted-sum방법은 질의에 나타나는 단어의 수가 적고 많은 문서들이 이 단어들을 모두 가지고 있을 경우에 좋지 않은 검색결과를 보여주고 있다. 본 논문은 추론 망 모델에 적용되는 유사도 계산식인 weighted-sum방법을 개선하였고, 이를 기반으로 Web Trec 9의 자료를 검색하여 좋은 결과를 얻었다.

1. 서 론

정보 검색모델은 불린 모델에서 출발하여 벡터모델과 확률모델 그리고 각각의 확장 모델들까지 다양한 모델들이 연구되고 있다. 베이지안 망을 기반으로 하는 추론 망 모델은 확률모델의 확장 모델로서 좋은 결과를 보여주고 있다[1]. 하나의 추론 망 안에 여러 가지의 질의 형태를 동시에 지원할 수 있으며, 이들 결과들을 취합하여 좀더 좋은 결과를 얻을 수 있는 모델이다.

정보 검색분야의 연구는 인터넷의 발달을 계기로 활발하게 전개 되고 있다. 인터넷상에 존재하는 문서의 수는 기하급수적으로 증가하여 그 수가 이미 10억 개를 넘었다. 이렇게 끝없이 쌓여가고, 생겨나는 문서들 속에서 자신이 원하는 정보를 찾는 것은 점점 더 어려운 일이 되어버렸다. 또한 사용자의 질의에 평균 2, 3개의 단어가 사용되고 있어 벡터모델이나 확률모델에서 좋은 결과를 얻기가 어려운 현실이다.

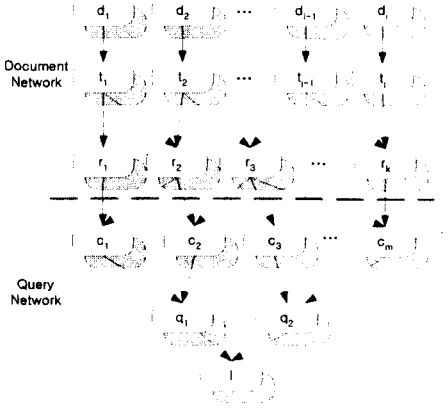
추론 망 모델에서도 질의에 출현하는 단어가 적을 경우에 같은 유사도 값을 갖는 문서가 많아져 문서들간에 순위를 결정하기 어려운 문제를 가지고 있다. 본 논문에서는 추론 망 모델에서 사용하는 유사도 계산식인 weighted-sum방법의 문제점을 파악하여 이를 개선하여 웹 환경에서 좀더 좋은 결과를 얻을 수 있도록 하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 추론 망을 소개하고, 3장에서는 웹 검색환경의 특징을 기술하고 4장에서는 웹 검색환경에서의 추론 망 모델의 문제점을 제기하고 이를 개선할 수 있는 변경된 weighted-sum 방법을 제시한다. 그리고 5장에서는 이를 기반으로 Web TREC9 자료를 이용하여 실험한 결과를 개선하기 전 모델과 비교하여 제시한다. 6장에서는 본 논문의 결론 및 향후 연구방향을 제시한다.

2. 추론 망을 기반으로 하는 검색 모델

추론 망 모델은 Inquery 시스템[2]에서 작동중인 모델로서 베이지안 망 모델에 기반을 두고 있다. 베이지안 망[3]은 방향성 비순환 그래프(Directed Acyclic Graphs ; DAGs)로서 노드는 랜덤 변수(random variable)를 나타내며, 아크는 이 변수들 사이의 인과 관계를 나타내고, 이 인과 관계 영향의 정도는 조건 확률로 표시된다.

추론 망 모델은 아래 [그림 1]처럼 문서 망과 질의 망으로 구성된다.



[그림 1] 기본 추론 망

2.1 Document Network

문서 망은 색인된 문서들의 정보를 표현하는 망으로 문서 노드(Document node)들 계층, 텍스트 표현 노드(Text representation node)들 계층과 내용 표현 노드

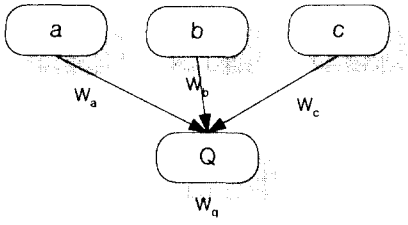
(Content representation node or Representation node)들 계층으로 이루어 진다. 여기에서 문서 노드는 색인된 문서들의 추상적인 표현이고, 텍스트 표현 노드는 문서의 내용들이 공유되는 환경에서 유용하게 적용되는 노드로서 공유되는 부분내용이 없을 경우에는 문서 노드와 일대일 관계로 형성된다. 내용 표현 노드는 내용을 표현하는 정보들로서 일반적으로는 문서에 존재하는 단어들로 구성된다. 임의의 단어는 자신이 출현하는 텍스트 표현 노드와 멀티로 연결되어 진다. 문서 망은 문서를 색인 할 때 한번 구축되는 망으로 색인이 추가될 때만 망의 변경이 이루어 진다.

2.2 Query Network

질의 망은 사용자의 요구사항을 표현하는 망으로 정보요구 노드(Information need node) 계층, 질의 노드(Query node)들 계층과 질의 개념 노드(Query concept node)들 계층으로 이루어 진다. 정보 요구 노드는 사용자의 요구를 추상적으로 표현한 노드이고, 질의 노드는 이러한 사용자의 요구를 질의로 변경하여 표현하는 노드이다. [그림 1]처럼 하나의 사용자의 요구가 여러 개의 질의 노드로 표현될 수 있다. 예를 들면, 사용자의 요구를 불린 질의와 확률 질의로 표현하고 각각의 결과를 합하는 것을 지원하는 모델이다. 질의 망은 사용자의 요구가 있을 때 마다 새롭게 구성되는 노드로서 질의 개념 노드에 시소러스 확장 같은 질의 확장 기능을 추가할 수 있다.

2.3 유사도 계산

추론 망 모델에서 사용하는 유사도 계산은 weighted-sum link matrix를 이용하여 높은 가중치를 가진 부모에게 많은 영향을 받는다. 망의 구조가 다음 [그림 2]와 같다고 가정하자.



[그림 2] 추론 망 예제

위의 그림처럼 노드 Q는 부모 노드로서 a, b와 c 노드를 가지고 있고, 각각의 링크 가중치가 w_a, w_b와 w_c 이고, Q 노드의 가중치를 w_q라 한다. 여기서 w_a, w_b, w_c >= 0, 0 <= w_q <= 1 이고, $t = w_a + w_b + w_c$ 라 한다면, weighted-sum link matrix 에서 다음과 같은 식을 얻을 수 있다.

$$P(Q = true) = \frac{(w_a a + w_b b + w_c c) w_q}{t} \quad \text{.....(1)}$$

위의 식을 이용하여 [그림 1]에서 각각의 문서에 대한 질의의 유사도를 계산 할 수 있다.

3. 웹 환경의 특징

인터넷의 발달로 인하여 수많은 사람들이 인터넷을 활용하고 있으며, 수 많은 정보들이 인터넷상에서 서비스되고 있다. 인터넷의 웹 문서의 개수는 벌써 10억 개를 넘었으며, 기하 급수적으로 늘고 있는 추세이다. 이러한 문서들에서 자신이 원하는 정보를 찾고자 하는 요구는 점차 커지고 있으며, 이를 지원해줄 수 있는 정보검색시스템이 절실히 필요한 실정이다.

웹 환경의 특성을 정리해 보면 다음과 같다. 첫째, 문서의 수가 방대하다. 둘째, 문서의 생명력이 짧다. 셋째, 빠른 검색을 지원해야 한다. 넷째, 짧은 질의에서 사용자의 요구를 정확하게 찾아야 한다. 이와 같은 특성을 모두 만족하는 시스템을 만들고자 많은 곳에서 연구가 진행되고 있다.

본 논문에서는 좋은 검색 결과를 보이고 있는 추론 망 모델에서 질의에 출현하는 단어가 적을 경우에 발생되는 문제를 발견하고 이를 수정하고자 한다.

4. 추론 망 모델의 문제점 제기 및 해결 방안 제시

추론 망 모델은 질의에 출현하는 단어가 적을 경우에 같은 유사도 값을 갖는 문서들이 많이 나와 랭킹에 어려움을 안고 있다.

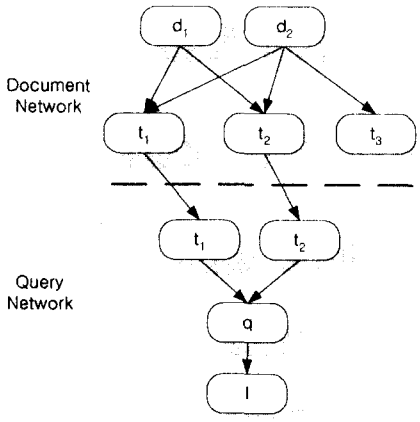
4.1 문제점 제기

3개의 문서가 있고 각각의 문서가 다음과 같이 단어와 가중치를 포함하고 있고, 질의에 출현하는 단어가 다음과 같다면, [그림 3]같은 추론 망이 구성될 것이다.

문서 정보 :

D1 = (t1 : 0.8, t2 : 0.7), D2 = (t1 : 0.6, t2 : 0.5, t3 : 0.6)

질의 정보 : **Q = (t1 : 0.6, t 2: 0.5)**



[그림 3] 가상 추론 망

[그림 3]과 같은 망에서 유사도를 계산해보면, 질의에 나타난 단어 t1에 대하여 문서 d1,d2의 weighted-sum 값은 0.48(0.8*0.6), 0.36(0.6*0.6)의 값을 가지고, 질의 단어 t2에 대하여서는 0.35(0.7*0.5), 0.25(0.5*0.5)의 값을 갖는다. 질의 노드 q에 대한 문서 d1, d2의 유사도 계산에서 (0.48+ 0.35)/(0.48+ 0.35), (0.36+ 0.25)/(0.36+ 0.25)으로 각각 유사도 1의 값이 나오게 된다. 여기서 문서 d1은 d2에 비하여 단어 t1, t2에 높은 가중치가 있어도 같은 유사도의 값을 갖는 문제점이 있다. 단, 질의에 출현하는 단어가 한 개일 경우에는 weighted-sum 방식으로 처리되지 않고 부모의 값이 바로 다음 자식에게 전달됨으로 이러한 현상이 발생되지 않는다.

4.2 해결 방안 제시

위의 문제는 질의에 나오는 모든 단어를 다 가지고 있는 문서에 대하여 나올 수 있는 문제로서 weighted-sum

에서 $t = w_a + w_b + w_c$의 값과 $w_a a + w_b b + w_c c$의 값이 같을 경우에 대하여 예외처리를 함으로 해결하고 자 한다. 수정된 유사도 식은 다음과 같다.

$$P(Q = true) = \quad \ldots\ldots (2)$$

$$if\,(t != (w_a a + w_b b + w_c c))\,then = \frac{(w_a a + w_b b + w_c c) w_q}{t}$$

$$else = (1 + w_a a + w_b b + w_c c) w_q$$

위의 식을 다시 적용하면, 문서 d1, d2의 유사도 값은 각각 1.83과 1.61로 계산되고 문서 d1이 좀더 유사한 문서로 랭킹이 매겨진다.

5. 실험 및 결과

실험은 Web TREC 9의 10G 자료를 사용하여 구축된 문서 망을 기본으로 이루어 졌으며, 질의는 451번부터 470번까지 20개를 가지고 실험을 하였다. Web TREC 문서들은 인터넷상에서 서비스 되고 있는 HTML 문서들이고, 질의는 질의 문서에 있는 TITLE 필드만 가지고 진행하였다.

20개의 질의 중에 본 연구에 적용되는 질의는 5개로서 451번, 454, 461번, 462번 과 466번이다. 다른 질의는 질의에 출현하는 단어가 하나이거나 질의에 출현하는 모든 단어를 포함하는 문서가 적은 경우이다. 본 연구에 적용되지 않는 경우에는 적용 전의 재현율, 정확율이 적용후의 결과와 일치하였다.

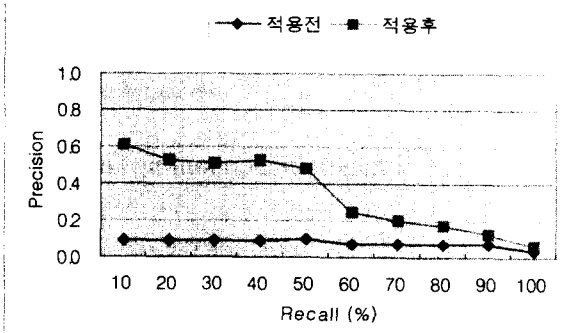
실험의 결과는 다음 표와 같다.

질의 번호 / 재 현 율	451		454		461	
	적용전	적용후	적용전	적용후	적용전	적용후
10%	0.081	1.000	0.161	0.700	0.022	0.250
20%	0.109	0.714	0.181	0.636	0.022	0.250
30%	0.137	0.636	0.160	0.600	0.022	0.250
40%	0.115	0.692	0.159	0.514	0.042	0.400
50%	0.134	0.524	0.162	0.455	0.042	0.400
60%	0.141	0.467	0.172	0.384	0.012	0.333
70%	0.148	0.320	0.174	0.296	0.012	0.333
80%	0.149	0.310	0.185	0.218	0.012	0.333
90%	0.155	0.313	0.179	0.181	0.016	0.129
100%	0.153	0.164	0.004	0.004	0.016	0.129
평균 정확율	**0.132**	**0.514**	**0.154**	**0.399**	**0.022**	**0.281**

질의 번호 / 재 현 율	462		466	
	적용전	적용후	적용전	적용후
10%	0.048	0.077	0.143	1.000
20%	0.014	0.047	0.143	1.000
30%	0.011	0.055	0.143	1.000
40%	0.012	0.050	0.143	1.000
50%	0.013	0.027	0.143	1.000
60%	0.013	0.017	0.019	0.019
70%	0.012	0.012	0.019	0.019
80%	0.004	0.004	0.019	0.019
90%	0.003	0.003	0.019	0.019
100%	0.002	0.002	0.019	0.019
평균 정확율	**0.013**	**0.029**	**0.081**	**0.510**

질의 461번과 질의 466번은 연관성 있는 문서의 개수가 각각 4개와 2개라서 실험에 애로사항이 되었다.

5개 질의에 실험결과를 취합한 결과는 다음[그림4]와 같다.



[그림 4 Average Recall-Precision]

위의 그림에서 볼 수 있듯이 재현율의 앞부분(50%)까지는 높은 향상을 얻었으나 뒷부분에서는 적용전의 결과에 수렴하는 결과를 얻었다.

6. 결론 및 향후과제

좋은 검색결과를 보이고 있는 추론 망 검색 모델[1]이 질의에 나타나는 단어가 적고, 그 단어들을 모두 포함하고 있는 문서가 많을 경우에 같은 유사도 값으로 인하여 검색결과의 랭킹에 어려움이 있었다. 본 논문에서는 이런 문제를 유사도 계산식인 **weighted-sum** 방법을 변형하여 검색결과를 좋게 하는 연구를 수행하였다.

20개의 질의 중 본 연구에 적용되지 않은 **15**개의 질의는 적용전과 적용후가 동일한 결과를 얻었지만, 본 연구에 적용된 **5**개의 질의에서 평균 정확율이 약 **333%**정도 증가 하는 결과를 얻었다.

향후과제로는 웹 문서에 존재하는 링크정보와 현재 나온 검색 결과를 결합하는 방법을 연구하자고 한다. 이미 비슷한 연구[4]가 이루어지고 있으나 좀더 좋은 결과를 얻고자 한다.

7. 참고문헌

[1] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, July 1991.

[2] J. Broglio, J. P. Callan, W. B. Croft, and D.W. Nachbar. Document retrieval and routing using the INQUERY system. In D. K. Harman, editor, Overview of the Third Retrieval Conference, NIST Special Publication 500-225, 1995

[3] Judea Pearl. Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference. Morgan Kaufmann Publishers, Inc., 1988.

[4] Ilmerio Silva, Berthier Ribeiro-Neto, Pavel Calado, Edleno Moura, Nivio Ziviani. Link-Based and Content-Based Evidential Information in a Belief Network Model. ACM SIGIR Conference on Research and Development in Information Retrieval, July 2000