Answer set 자동 구축을 위한 문서 필터링

정용교*0 신승은* 오효정** 장명길** 서영훈*
*충북대학교 컴퓨터공학과
**한국전자통신연구원, 휴먼정보처리연구부, 휴먼정보처리연구팀
*{kkabi810^0, seshin}@dcenlp.chungbuk.ac.kr , **{ohj, mgjang}@etri.re.kr , *yhseo@cbucc.chungbuk.ac.kr

Document filtering for automatic construction of Answer Set

Yong-Kyo Jeong*0 Seung-Eun Shin* Hyo-Jung, Oh** Myung-Gil, Jang** Young-Hoon Seo*
*Dept. of Computer Engineering, Chungbuk National University
**Member of Engineering Staff , Human Information Retrieval Technology Research Team ,
Human Information Processing Department, ETRI-Computer & Software Technology Lab.
Electronics Telecommunications Research Institute (ETRI)

요 약

본 논문은 의미기반 정보검색 소프트웨어 기술에서 정답 문서 자동 구축을 위한 문서 필터링기법을 제안한다. 문서 필터링은 1차 질의어와 문서간의 유사도와 2차 질의어와 문서간의 유사도를 이용하여 이루어지며, 1차 질의어와 문서간의 유사도를 구하기 위하여 개념망과 백과사전 정보를 이용한 1차 질의어 확장 과정을 수행하고, 확장된 질의어와 문서와의 유사도를 계산한다. 1차 확장 질의어를 이용해 얻어진 결과 중 유사도가 상위 10%에 속하는 문서를 이용하여 2차 질의어 확장을 한다. 2차 질의어 확장은 상위 10% 문서에 출현하는 명사중 문서 출현 빈도가 임계치 이상인 명사를 선택하여 이루어지고, 그것을 이용하여 문서의 유사도를 계산한다. 이렇게 얻어진 두 가지의 유사도를 결합하여 문서들을 순위화하고 Accept Point를 이용하여 문서를 필터링한다.

1. 서론

인터넷이 등장하고 발전하면서 우리는 수많은 정보속에서 살아가고 있다. 이러한 많은 정보속에서 사용자의 요구를 만족시키는 정보 검색을 하기 위해 자연어처리 기술이 많이 응용되고 있다.

기존의 정보 검색 방법은 단순히 질의어가 포함된 문서만을 제공했기 때문에 질의어와 관련이 없는 문서들도 대량으로 검색 결과에 포함되었고 그로 인해 사용자는 검색결과에서 또 다시 문서를 찾아야만 하는 어려움이 있었다. 그러나 자연어처리 기술을 응용한 정보 검색방법은 자연어 질의어에 대해 그와 연관된 양질의 결과를 제공해줌으로써 기존의 검색 방법 보다 사용자의 노력을 줄일 수 있도록 하였다.

의미기반 정보검색 소프트웨어는 그러한 자연어 검색 방법을 기반으로 사용자의 요구를 충족시킨다. 이 기술은 사용자의 질의에 미리 대비하여 정답 문서 집합을 구축해 놓은 후, 사용자의 질의를 파악하여 원하는 정답이 포함된 문서를 제공하는 기술이다[1].

여기서 정답 문서 집합이란 웹 문서를 명사 개념별로 분류하고, 사용자의 관심 영역에 따라서 세분해서 나누어 놓은 문서 집합을 의미한다. 이러한 정답 문서는 개념망을 기반으로 하기 때문에 해당 분야의 대부분의 질의에 대한 정답을 제공할 수 있는 특징이 있다[2]. 하지만 그것을 사람이 일일이 구축한다는 것은 많은 비용과 시간을 필요로 하기 때문에 그것을 자동화하는 방안이 요구된

본 논문은 한국전자통신연구원 지원의 "의미기반 정보검색 소프트웨어" 과제의 일부로 연구되었음

다.

본 논문에서는 그러한 정답 문서 자동화 구축을 위한 문서 필터링에 대한 방법을 제안한다. 여기서 제안하는 문서 필터링 기법은 기존에 연구되어진 기법과는 다소 차이가 있다.

[3]은 스팸메일을 필터링하기 위하여 스팸메일 사전과 비스팸메일 사전을 생성하여 그것을 학습문서로 이용하여 필터링을 한다. [4]또한 Unlabeled data를 학습문서로 사용하여 필터링을 한다. 이처럼 기존의 연구들은 학습문서를 가지고 있고 학습을 통해서 얻어진 특성으로 문서들을 분류하지만 본 연구에서는 메타 검색에 대한 결과물에 대해 가지고 있는 정보는 단지 개념어 하나이므로 별도의 과정을 거쳐 필터링을 구현하게 된다.

2. 정답 문서 집합 구축 과정 및 개념망

정답 문서 집합은 사용자의 질의에 대비하여 미리 구축해 놓는 자료로서 그림 1과 같은 구축 과정을 거치게 된다.

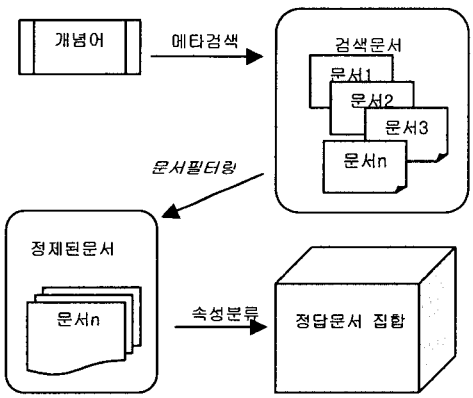

그림 1. 정답 문서 자동 구축 과정

개념어에 대해서 메타 검색을 하고, 검색된 결과에 대하여 필터링을 수행한다. 필터링을 통해 개념어와 관련이 없는 문서는 제거되고, 관련이 있는 문서는 개념어의 속성에 따라 분류하여 정답 문서를 구축한다.

개념망은 ETRI의 지식 베이스의 일부분이며, 지식베이스는 사용자 질의의 의미적인 분석을 토대로 사용자의 요구에 근접한 문서를 제공하기 위하여 지식 구조와 웹문서를 연결해 놓은 데이터 베이스다. 여기서 지식 구조에 해당하는 개념망은 한국어 명사 어휘로 표현되는 개념을 정확하게 파악하기 위하여 개념들간의 다양한 관계를 연결시켜 놓은 어휘 데이터베이스를 말하고, 이 개념망은 국어학적인 의미관계를 이용하여 상하관계를 기본축으로 하고 있으며, 상하관계의 보완적 측면에서 동의-유의관계, 부분-전체관계, 반의관계 등을 추가로 정의하고 있다[5,6,7]. 그림 2는 지식베이스의 구조를 보여준다.

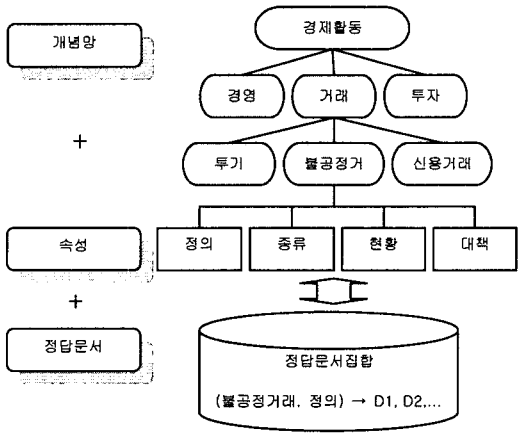

그림 2. 지식베이스 구조

이러한 개념망은 질의어와 문서간의 유사도를 구하기위해 질의어를 확장할 때 사용한다. 그 사용방법에 대해서는 3절에서 기술하도록 한다.

3. 문서 필터링

3.1. 시스템 설계

초기 질의어인 개념어에 대하여 개념망과 백과사전 정보를 이용하여 1차 질의 확장을 수행한다. 확장된 질의어를 이용하여 문서들과의 유사도를 계산하고, 그것을 이용하여 2차 질의 확장을 수행한다. 2차 질의 확장은 1차 유사도에서 얻어진 상위 10%의 문서들의 명사를 이용하여 수행한다. 2차 확장된 질의어와 문서들과의 유사도를 계산하고, 1차 유사도와 2차 유사도를 결합하여 문서들을 순위화하며, Accept point를 정하여 문서를 필터링한다. 자세한 수행과정은 다음절에서 기술하도록 한다. 그림 3은 시스템의 진행과정을 나타낸다.

3.1. 1차 질의어 확장

본 연구는 개념망의 개념어를 필터링 주제로 사용하기 때문에 문서 필터링에 활용할 수 있는

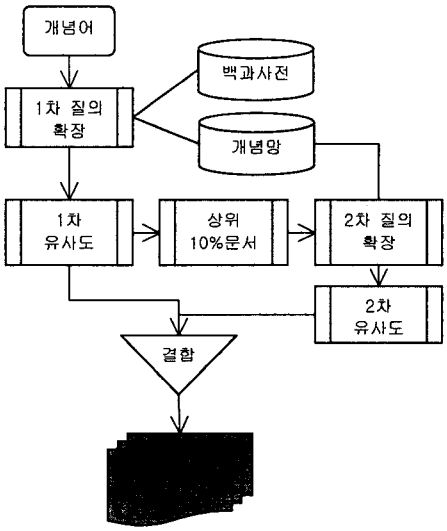

그림 3. 시스템 구성도

정보는 메타검색에 사용된 개념어뿐이다. 초기 질의어(initial query)의 양이 매우 적기 때문에 질의어의 확장이 필요하며, 1차 질의어 확장을 위해 개념망과 백과사전 정보를 이용하였다. 이 중 상위어 정보와 백과사전 정보는 가공하여 사용하게 되는고, 가공 과정은 다음과 같다.

개념망을 이용하여 질의어(개념어)의 상위어와 하위어를 찾고, 상위어를 이용하여 질의어를 분할한다. 만약 개념어가 "품질관리" 일 경우 상위어는 "관리" 가 된다. 이 때 개념어에서 상위어를 패턴 매칭하여 제거한 나머지 단어(품질)와 상위어(관리)가 확장 질의어에 포함된다. 표 1은 개념망을 이용한 질의어 확장의 예이다.

백과사전 정보를 이용한 질의 확장에서는 개념어에 대한 백과사전 정의를 색인하여 나온 결과중 개념망에 존재하고, 개념망을 이용한 확장 질의에 포함되지 않는 단어만이 확장 질의어에 포함된다. 여기서 개념망에 존재하는 단어만 선택하는 이유는 필터링할 문서가 경제관련 문서이고, 개념망 또한 경제 용어 관련 개념망이기 때문에 확장 질의어에서 경제 분야와 관련이 없는 단어를 제거하고 양질의 확장 질의어를 얻기 위함이다. 표 2는 백과사전을 이용한 질의어 확장의 예이다

3.2. 1차 질의어-문서간의 유사도

질의어 확장 과정을 통해 얻어진 질의어와 메타검색에 의해 얻어진 문서와의 유사도를 계산하는데 방법은 식(1)과 같다.

$$S_{d_i} = \frac{\sum (q_w \times q_{tf})}{\sqrt{TF_{d_i}}} \quad (1)$$

표 1. 개념망을 이용한 질의 확장 예

질의어	상위어	하위어	질의어 분할
품질관리	관리	티큐시	품질, 관리
재산권	권리	채권, 무제재산권, 지능권, 산업재산권	재산, 권리
확장 질의어			
품질관리	*품질관리, 티큐시, 품질, 관리*		
재산권	*재산권, 채권, 무제재산권, 지능권, 산업재산권, 복제권, 재산, 권리*		

표 2. 백과사전을 이용한 질의 확장 예

품질관리	백과사전 정의
	과학적 원리를 응용하여 제품품질의 유지,향상을 기하기 위한 관리
	색인 결과
	과학, 원리, 응용, 제품, 품질, 유지, 향상, 관리
	확장 질의어
	과학, 원리, 제품
재산권	백과사전 정의
	재산적 가치를 지니는 권리
	색인 결과
	재산, 가치, 권리
	확장 질의어
	가치

1차 질의어와 문서간의 유사도는 각 문서에 나타나는 색인어 가중치의 합을 그 문서의 총 색인어 개수의 제곱근으로 나누어 유사도를 계산한다.

이렇게 얻어진 유사도에 따라 하위 10%의 문서는 제거하고, 상위 10%의 문서는 2차 질의어 확장에 이용한다.

3.3. 2차 질의어 확장

1차 질의어와 문서간의 유사도를 이용하여 상위 10%의 문서를 추출한 후, 이를 2차 질의 확장에 사용한다. 본 연구와 기존의 문서 필터링의 큰 차이는 학습 문서가 없다는 것이다. 따라서 1차 질의어와 문서간의 유사도에서 상위 10%문서를 질의어에 대한 관련 문서라 가정하고 2차 질의를 확장 한다. 질의 확장 방법은 상위 10%문서에 출현하는 명사중 출현 빈도가 임계치 이상인 명사를 선택하고, 그 중 1차 확장 질의에 사용되었던 명사를 제거한 나머지 명사에서 개념망에 존재하는 것만을 2차 질의어로 선택한다. 표 3은 2차 질의어 확장의 예이다.

표 3. 상위 10%문서를 이용한 2차 질의 확장 예

질의어	상위 10% 문서의 색인어	2차 확장 질의어
품질관리	가격, 고객, 공사, 계획, 도구, 방법, 인증, 산업, 시험, 전기, 품질관리, …	*계획, 방법, 산업, 시험*
재산권	가격, 감시, 강화, 기술, 분야, 사건, 산업, 소개, 위원, 저작권, …	*기술, 분야, 산업, 상표, 저작권, 특허*

이렇게 얻어진 2차 확장 질의어와 문서간의 유사도는 식(1)을 이용하여 계산하고 1차 질의어를 이용하여 계산된 유사도와 결합한다.

$$sim_{final} = sim_{1st} + sim_{2nd}$$

1차 질의어와 문서간의 유사도와 2차 질의어와 문서간의 유사도를 결합하여 문서들을 재순위화 하고 Accept point를 이용하여 대상 문서를 필터링한다.

4. 실험 및 결과

본 연구에서 제안한 문서 필터링의 성능을 검증하기 위해 다음과 같은 실험을 실시하였다.

실험은 개념망에서 경제 분야의 27개의 개념어, 전체 문서 2511개에 대하여 수행하였다.

다음은 개념어 및 하위어 등의 1차 확장 질의어에 대한 가중치를 달리하며 F-Measure를 계산한 결과이고, F-Measure를 계산하는 방법은 다음과 같다.

$$F = \frac{(\beta^2+1)pr}{\beta^2 p + r} \quad (2)$$

여기서 β는 재현률(r)과 정확률(p)의 비중을 선택 할 수 있게 하는 변수로 β>1이면 정확도의 비중을 β<1이면 재현도의 비중을 높게 두는 의미이다. 실험에서는 β = 1로 하여 재현률와 정확률의 비중을 같게 두었고 이를 Break-Even point라 한다. 표 4는 개념망을 이용한 확장 질의어 가중치별 Break-Even point에서의 F-Measure를 나타낸다.

이 결과에서 1차 확장 질의는 <질의어, 하위어, 상위어를 제외한 질의어 분할 결과, 상위어>의 가중치 비율이 <6:2:1.5:1>에서 가장 높은 값을 보이고 있다.

표4. 개념망을 이용한 확장 질의어 가중치별 Break-Even point에서의 F-Measure

	상위어를 제외한 질의어 분할 결과/상위어			
질의어/하위어	1/1	1.5/1	1.5/0.5	1/0.5
8/1	0.695307	0.705271	0.699726	0.685916
8/2	0.6973	0.705271	0.699726	0.685012
8/4	0.695307	0.704181	0.699726	0.685012
6/1	0.707702	0.711835	0.699233	0.697055
6/2	0.707702	0.711836	0.700322	0.698144
6/3	0.70571	0.710932	0.700322	0.697055
4/1	0.687216	0.687619	0.696785	0.705045
4/2	0.687216	0.687619	0.696785	0.705045

표 4의 결과를 기반으로 하여 가중치 비율 <6:2:1.5:1>로 하여 실험을 진행하였다.

표 5는 개념망을 이용한 질의 확장의 결과이고, 표 6은 개념망을 이용한 질의 확장과 백과 사전 정보를 결합하여 사용한 결과이다. 표 7은 개념망을 이용한 질의 확장과 2차 질의 확장을 결합하여 나타낸 결과이고, 표 8은 개념망을 이용한 질의 확장과 백과 사전, 그리고 2차 질의 확장을 이용한 결과이다. 각각은 break-even point와 accept point별 평균 정확률과 평균 재현률, 그리고 필터링되어 남은 문서의 비율을 나타낸다.

표 5. 개념망을 이용한 질의 확장 이용

Accept point	평균정확률 / 평균재현률	남은 문서 비율
Break even point	0.711836 / 0.711836	33.04%
1.5	0.537637 / 0.835826	51.69 %
2	0.619083 / 0.765496	41.41 %
2.5	0.706337 / 0.685802	31.86 %
3	0.734095 / 0.615565	26.93 %
3.5	0.772634 / 0.532510	21.66 %

표 6. 개념망을 이용한 질의 확장 + 백과사전 정보

Accept point	평균정확률 / 평균재현률	남은 문서 비율
Break even point	0.699961 / 0.699961	33.04%
1.5	0.502095 / 0.884251	58.16%
2	0.582532 / 0.797527	46.23%
2.5	0.641379 / 0.725493	37.26%
3	0.706265 / 0.653823	30.02%
3.5	0.735357 / 0.574081	24.76%

표 7. 개념망을 이용한 질의확장 + 2차 질의 확장

Accept point	평균정확률 / 평균재현률	남은 문서 비율
break even point	0.705075 / 0.705075	33.04%
1.5	0.477653 / 0.905287	62.27%
2	0.537758 / 0.840757	51.64%
2.5	0.611505 / 0.758354	41.59%
3	0.655501 / 0.692022	34.68%
3.5	0.739113 / 0.566228	24.67%

표 8. 개념망을 이용한 질의확장 + 백과사전 정보 + 2차 질의 확장

Accept point	평균정확률 / 평균재현률	남은 문서 비율
break even point	0.659797 / 0.659797	33.04%
1.5	0.459618 / 0.923004	0.657883
2	0.525290 / 0.860237	0.547687
2.5	0.572216 / 0.777881	0.455526
3	0.626423 / 0.716680	0.376106
3.5	0.665504 / 0.642418	0.312164

본 연구에서는 실험 문서의 초기 정확률을 높이고, 개념어에 해당하는 문서들에 대해 1차 확장 질의어의 가중치를 다르게 함으로써 각각의 F-Measure를 구하여 최적의 가중치 비율를 구하였다.

그리고 그 가중치 비율을 이용하여 개념망을 이용한 질의 확장과 백과사전 정보를 이용한 질의 확장, 상위 10%문서를 이용한 질의 확장등에 대하여 각각 평균 정확률과 평균 재현률, 그리고 필터링후 남은 문서의 비율을 구하였고, 이를 이용하여 Accept point를 결정하여 문서를 필터링한다.

5. 결론 및 향후 연구 과제

본 논문에서는 개념망과 백과사전 정보, 1차 유사도 결과를 이용한 질의어 확장 기법을 통해 개념어에 대한 메타검색 결과를 필터링하는 방법을 제안했다. 개념망을 이용하여 초기 질의어(개념어)에 대한 상위어와 하위어 정보를 얻고, 상위어를 이용하여 질의어를 분할하여 1차 질의 확장을 한다. 그리고 개념어에 대한 백과 사전 내용을 색인한 후, 개념망을 이용한 확장 질의에 없는 명사와 개념망에 존재하는 명사 또한 1차 확장 질의에 포함한다. 1차 확장 질의어를 이용하여 문서들과의 유사도를 구하고 문서를 순위화한다.

그 중 상위 10% 문서는 2차 질의 확장에 사용하게 되는데, 상위 10% 문서에 출현하는 명사중 출현 빈도가 임계치 이상의 명사를 선택하고, 그 중 1차 확장 질의에 사용되었던 명사를 제거한 나머지 명사에서 개념망에 존재하는 것만을 2차 질의어로 선택한다. 여기서 개념망의 단어만을 선택하는 이유는 양질의 확장 질의어를 얻기 위함이며, 이렇게 확장된 2차 질의어를 이용해 문서들과의 유사도를 계산하고, 1차 질의어와 문서간의 유사도와 2차 질의어와 문서간의 유사도를 결합하여 문서를 재순위화 하고 accept point에 따라 문서를 필터링한다. 향후 과제로는 질의어와의 높은 유사도를 갖는 문서중 "거짓" 인 문서를 제거해야하고, 낮은 유사도를 갖는 문서중 "참" 인 문서를 선택해야 한다. 질의어와 높은 유사도를 갖는 문서중 "거짓" 인 문서는 대부분 질의어에 관련된 내용이 아님에도 불구하고 질의어가 빈번하게 쓰이는 경우이고, 반대로 질의어와의 낮은 유사도를 갖는 문서중 "참" 인 문서는 질의어와 관련된 문서이지만 질의어의 출현 빈도는 적고, 내용은 표의 수치로 나타나는 경우이다. 필터링의 성능을 높이기 위하여 이러한 문제점을 해결하기 위한 연구가 필요하다.

6. 참고 문헌

[1] 장명길, 김현지, 장문수, 최재훈, 오효정, 이충희, 허정, "의미기반 정보검색", 정보과학회지 10월호 한글정보처리 특집, 2001.

[2] 최호섭, 옥철영, 장문수, 장명길, "사전을 기반으로 한 한국어 의미망 구축과 활용", 제 29회정보과학회 춘계학술발표회, 2002.

[3] 김호성, 정경호, 황도삼, "단어 가중치를 이용한 스팸메일 필터링", 제 13회 한글 및 한국

어 정보처리 학술대회, 2001.
[4] Park, S.-B., Zhang, B.-T. "Document Filtering Boosted by Unlabeled Data", Proceedings of the 2001 IEEE International Symposium on Industrial Electronics (ISIE2001), vol. 1, pp. 328-333, 2001.
[5] 장문수, 장명길, 김현진, 오효정, 이재성, "인터넷 질의/응답을 위한 지식베이스 구축", 제 12회 한글 및 한국어 정보처리 학술대회, 2000.
[6] 장문수 외, 경제개념망 구축 결과보고서, TDP, ETRI, 2001.
[7] 장문수 외, 의미관계연구회 결과보고서, TDP, ETRI, 2001.
[8] 오효정, 임정묵, 이만호, 맹성현, "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", 제 11회 한글 및 한국어 정보처리 학술대회, 1999.
[9] 김영택, "자연언어처리", 생능출판사, 2001
[10] 장문수, 오효정, 장명길, "자동분류를 이용한 정답문서집합 구축", 제 13회 한글 및 한국어 정보처리 학술대회, 2001.
[11] 정성화, 이종혁, "문서 구조 정보에 기반한 웹 페이지 범주화 모델", 제 10회 한글 및 한국어 정보처리 학술대회, 1998.
[12] Kyung-Soon Lee, Young-Chan Park, Key-Sun Choi, "Re-ranking Model Based on Document Clusters", Information Processing and Management, vol 37, 2001.
[13] Joon Ho Lee and Jeong Soo Ahn, Using n-Gramsfor Korean Text Retrieval. In Proc. 19th ACMSIGIR International Conference on Research and Development in Information Retrieval, 1996.
[14] Joon Ho Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes", In Proc. 18th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1995.