김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제11호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 11 JUCHE103(2014).

온톨로지에 기초한 패쎄지검색의 한가지 방법

리 청 한

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술정보사업을 강화하여야 합니다. 과학기술정보사업을 잘하여야 적은 밑천과 품을 들여 과학기술발전에 절실히 요구되는 귀중한 자료들을 얻을수 있습니다.》(《**김정일**선집》 증보판 제15권 501페지)

질문응답체계에서 중요한 문제의 하나는 다량의 분문집합에서 질문에 대한 정답이 들어있는 패쎄지를 검색하는것이다. 그것은 질문에 적합한 응답이 들어있는 패쎄지를 먼저 검색함으로써 제한된 시간안에 패쎄지속에 포함된 정답을 찾는것이 질문응답체계의 성능평가에서 매우 중요하기때문이다. 이로부터 대규모의 분문집합에서 질문에 적합한 패쎄지를 검색하는 방법들이 많이 제안되였다.[1－3]

대부분의 전통적인 질문응답체계에서는 문서로부터 추출한 명사목록을 리용한다. 그러나 명사벡토르들은 명사들사이의 의미적정보를 표현할수 없다.

즉 지난 시기의 방법들은 개별적질문용어들의 의미적관계는 고려함이 없이 질문용어들이 패쎄지안에 얼마나 많이 또 질문용어들이 어느정도 가까이에 있는가에 따라 적합패쎄지를 판정하고있는것으로 하여 질문응답체계의 성능을 높이는데 일정한 제한을 주고있다.[3]

최근년간 언어처리분야에서는 용어들의 의미적관계를 고려한 온톨로지에 기초하여 검색을 진행하는 연구에 많은 주목을 돌리고있다.[4]

이로부터 론문에서는 최근에 언어처리분야에서 주목을 끌고있는 온톨로지를 구축하는 한가지 방법과 구축된 온톨로지에 기초하여 패쎄지를 검색하는 방법을 제기하였다.

1. 온톨로지구축방법

온톨로지를 구축하기 위하여 먼저 중심단어를 찾고 이 중심단어를 리용하여 기초온톨로지를 수동적으로 구축한 다음 기초온톨로지를 리용하여 온톨로지를 확장한다. 론문에서는 경제부문의 경량급온톨로지를 구축하는 방법을 보았다.

일반적으로 온톨로지는 많은 어휘용어들을 포함하는 망의 한 종류로 볼수 있다.

대단히 많은 련결을 가지는 몇개의 중심단어들은 전체 망에서 매우 중요한 역할을 한다.[6]

1) 중심단어추출방법

론문에서는 문서 혹은 분문집합에서 다른 단어들과 련관이 많은 단어들을 중심단어로 정의하며 이 중심단어를 리용하여 기초온톨로지를 구축한다.

다른 한편 본문집합에서는 빈도수가 높은 단어(고빈도단어)들을 찾는다. 왜냐하면 고

빈도단어들이 문서안에서 많은 다른 단어들과 관련이 있다는것을 가정하고있기때문이다. 바로 빈도수가 높은 이 단어를 중심단어라고 한다.

중심단어를 찾기 위하여 먼저 분문집합에서 형태부해석을 진행하여 금지단어를 제외한 모든 명사들을 추출한다.

다음 매 명사의 tf값과 idf값을 계산하여 큰값을 가지는 명사들을 중심단어로 선택한다. 론문에서 취급한 경제학부문 도서 및 잡지에서 600개의 명사들과 70개의 고유명사들이 선택되였다.

600개의 명사들은 총 명사개수의 0.79%로서 총 명사빈도의 55%이상을 차지한다.

사람이름, 나라명, 도시명과 같은 고유명사들이 경제학령역에서 중요하기때문에 고유명사를 일반명사처럼 취급하였다.

2) 기초온톨로지구축방법

선택한 중심단어에 기초하여 기초온톨로지를 수동적으로 구축한다.

우선 한 중심단어를 문맥적으로 그 중심단어와 관련이 있는 다른 중심단어와 련결한다.

동시에 출현하는 단어들이 서로 련결되여있다는 가정에 기초하여 모든 문서들에서 중심단어주위에서 4개의 단어를 추출한다. 그리고 그 단어들의 빈도수를 계산한다.

표 1은 중심단어 《회사》, 《주식》, 《무역》, 《직원》, 《사장》이 호상련결되여있으며 그것들이 다른 단어들과 많이 관련된다는것을 보여준다.

표 1. 중심단어주위단어들

중심단어	동시에 발생하는 단어목록
회사	주식, 시장, 판매, 거래, 보험, 은행, 계획, 무역, 사장, 가격, 봉사, …
무역	시장, 주식, 회사, 무역거래, 발명가, 변화, 공업, 세계, 회의, 은행, …
주식	변화, 시장, 무역, 회사, 가격, 현금, 분석가, 성장, 리윤, 계획, 은행, …
직원	집, 성원, 사장, 회사, 가격, 회의, 판매, 시장, 업무, 재정, 봉사, 대변인, …
사장	실행, 눅거리, 서기, 회사, 회의, 사장, 은행, 집, 시세, 구좌, 관리, 재정, …

표 1의 중심단어를 《매듭》으로 하여 동시에 출현하는 단어들을 《련결》시켜 망을 창조한다.(그림 1)

그림 1에서 중심단어들은 《회사》, 《주식》, 《무역》, 《직원》이다. 그러나 그림 1은 완성된 온톨로지가 아니다. 왜냐하면 이 망에는 의미론적 관계가 반영되여있지 않기때문이다.

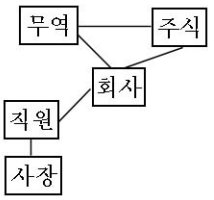

그림 1. 중심단어망

다음 관계를 정의하기 위하여 명사들을 추출할 때와 마찬가지로 본 문집합에서 모든 동사들을 추출한 다음 추출된 동사에서 고빈도동사들을 선택한다. 그리고 이 동사들을 분류하여 50개의 주요관계를 정의한다.

다음 문맥관계를 고려하면서 주요관계주위에서 출현하는 명사들에 대한 목록을 만든다. 표 2는 그 목록의 일부이다.

표 2. 고빈도동사

동사	동시에 발생하는 명사목록
생산하다	백만, 공유, 주식, 시장, 판매, 거래, 보험, 은행, 계획, 분석, 관리, 성장, 리윤, 초청, 자금, 발명, 사장, 가격, 봉사, 재산, 로임, 강철, 자본, 무역
사다	시장, 주식, 미래, 발명가, 무역, 변화, 무역거래, 공업, 세계, 회의, 담화, 국가, 회사, 서기, 사장, 은행, 변화, 시세, 구좌, 관리, 재정, 직원, 무역가
팔다	집, 성원, 회사, 사장, 회의, 가격, 판매, 분야, 보안, 기구, 업무, 재정, 봉사, 대변인, 법률가, 변화, 시장, 무역거래, 가격, 주식, 현금, 선택, 세계, 분석가

기초온톨로지구축의 마지막단계는 망에 관계를 추가하는것이다.

그림 2는 수동적으로 구축한 기초온톨로지의 일부이다.

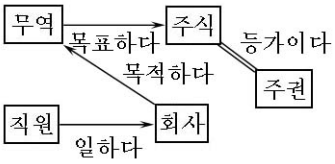

그림 2. 온톨로지에 관계추가

이 망에서 명사들은 개념들을 나타내며 동사들은 그 개념들사이의 관계를 표현한다.

3) 온톨로지확장

기초온톨로지를 구축한 다음 구축된 기초온톨로지를 리용하여 온톨로지를 확장한다.

온톨로지확장에서 기본은 중심단어와 다른 중심단어사이 그리고 한 중심단어와 비중심단어사이관계를 자동적으로 삽입하는것이다.

관계추가방법은 다음과 같은 두 단계를 거쳐 진행된다.

우선 문장추출기술과 구문해석기술을 리용하여 리용중심단어주위에 있는 명사들을 문서로부터 추출한다.

다음 관계추출규칙에 따라 중심단어와 다른 중심단어사이관계를 설정한다.

실례로 《속하다》 혹은 《포함하다》와 같은 동사가 어떤 두 명사사이에 있다면 온톨로지에 《belong to(속하다)》관계를 추가한다.

그림 3은 중심단어 《회사》주위에서의 개념들과 관계들의 결과를 보여준다.

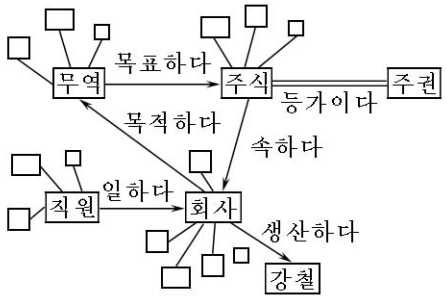

그림 3. 확장된 온톨로지

2. 온톨로지에 기초한 패쎄지검색방법

패쎄지검색을 위해 경제학부문 도서들에서 뽑은 빈도수가 높은 단어 505개의 중심단어들을 자료표현의 중심색인마디로서 선조중심단어로 정의한다.

매 선조중심단어는 그 단어와 련결되여있는 모든 마디 즉 중심단어와 비중심단어들을 자료요소로 가지게 된다. 여기서 중심단어를 후손중심단어, 비중심단어를 후손비중심단어로 정의한다.

선조중심단어 x의 자료구조를 다음과 같이 서술한다.

선조중심단어 x와 관계 $i(i=1, 50)$를 가지는 후손중심단어들을 서술하되 매 후손중심단어는 선조중심단어와 관계 i를 가지게 되는 패쎄지들의 번호렬을 가지게 된다. 이때 이 번호렬은 우선순위로 즉 발생빈도가 높은 순서로 배치한다.

또한 후손비중심단어도 같은 패쎄지번호렬을 가지게 되며 이때 후손비중심단어뒤에는 후손중심단어와의 구별을 위해《－》부호를 붙여준다.

이런 형식으로 50번째 관계까지의 모든 후손단어들을 정의한다.

그리고 선조중심단어와 아무런 관계도 없지만 같은 문장안에서 발생하는 비중심단어에 대하여서도 패쎄지목록을 작성한다. 만일 질문안에 관계에 해당한 내용이 없다 하더라도 용어만으로 탐색을 진행할수 있도록 자료를 구축한다.

구축된 온톨로지에서 매 마디는 한 단어를 표현하며 련관된 패쎄지목록을 가진다. 그러므로 온톨로지의 관계들과 패쎄지목록과 같은 정보를 리용하여 적합패쎄지를 추출할수 있다.

일반적으로 많은 패쎄지들에서 출현하는 단어는 개별적페쎄지에서는 중요한것으로 되지 않는다.

한편 중심단어는 한 색인용어를 다른 색인용어에 련결하는 다리의 역할을 수행한다.

중심단어에 기초한 온톨로지를 리용하여 패쎄지를 검색하는 단계는 다음과 같은 5단계로 되여있다.

① 본문집합에 대한 패쎄지분할단계로서 질문에 따라 패쎄지들을 분할하고 차례대로 패쎄지에 번호를 붙인다.

② 분할된 패쎄지에 대한 자연언어처리를 진행하는 단계이다. 즉 형태부해석과 어근화, 금지단어제거를 진행하여 명사, 고유명사, 동사들을 패쎄지별로 빈도수를 고려하여 추출한다.

③ 앞단계에서 얻은 자료에 기초하여 온톨로지를 구축한다.

④ 질문을 기본관계형태로 표현한다.

⑤ ④에서 진행한 기본관계형태로 표현된 질문을 가지고 ③에서 작성된 온톨로지의 표현에 정합시켜 적합한 패쎄지를 검색한다.

실례로《어느 회사가 강철을 생산하는가?》라는 질문이 제기되였다고 하자. 이 질문을 기본관계형태로 넘기면《생산하다(회사, 강철)》로 된다.

즉《19(회사, 강철)》로 된다.

다음《회사》는 중심단어이므로 선조중심단어로서 해당 자료기지에 가서《생산하다》라는 관계번호 19에 해당한 자료요소에서 우선순위가 제일 높은 페쎄지번호를 찾아 그에 해당한 내용을 추출하게 된다.

이와 같이 먼저 중심단어를 찾아 그에 해당한 자료로 가서 기본관계를 찾은 다음 다른 단어로써 검색을 진행하여 해당 페쎄지를 찾으므로 색인방식으로 보이지만 단어들사이의 관계를 기본관계속에서 표현하므로 단순한 단어색인보다 더 의미적인 정보를 담아 패쎄지를 검색하게 된다. 탐색에 유리한 질문의 형태는《관계 I(중심단어, 비중심단어)》형태로 된다.

그것은 중심단어가 가장 많이 발생하는 단어로서 해당 대답의 중요정보들을 련결하는 다리로서의 역할을 수행하기때문이다. 그러므로 탐색공간을 중심단어의 개수만큼 줄이고 다시 거기에서 의미적정보를 반영한 패쎄지번호(관계가 의미적정보를 반영하므로)를 찾는데 발생빈도수가 적은 비중심단어탐색공정에 의해 최종적인 대답이 들어있는 패쎄지번호를 확정하게 된다. 따라서 의미적이면서도 색인적인 탐색을 진행할수 있게 된다.

3. 실험결과 및 분석

실험을 위해 경제학부문의 문서로부터 627 649개의 명사들을 추출하였다.

실험을 통해 1 000이상의 빈도수를 가진 단어들은 단어빈도수의 54%이상이라는것을 알수 있다. 따라서 기초온톨로지를 위해 1 000번이상 출현하는 505개의 중심단어를 선택하였다. 론문에서는 본문집합에서 1 000번이상 출현하는 505개의 중심단어에 기초하여 온톨로지를 구축하였다. 중심단어의 총빈도수는 1 506 003으로서 모든 단어의 전체 빈도수의 54.18%이다.

그리고 전체 단어의 빈도수분산과 중심단어의 빈도수분산이 아주 류사하다. 그러므로 전체 단어를 놓고 응답추출작업을 진행하는것보다 빈도수가 아주 많고 사용자의 질문에 들어있을 확률이 높은 중심단어를 위주로 하여 온톨로지를 구축하고 검색을 진행하는것이 속도상으로나 자료량상으로 우월하게 된다.

또한 이전의 득점에 의한 패쎄지추출방법이나 베이스리론에 의한 패쎄지추출방법은 복잡하고 많은 수학적계산을 진행해야 하므로 그것을 콤퓨터로 실현하기 위해 복잡한 알고리듬을 리용하였지만 이 방법은 중심단어와 기본관계를 리용하는 새로운 방법으로 원천본문을 보다 론리적으로 표현하고 그 표현도 간단하므로 이전 방법에 비해 우월하다.

맺 는 말

중심단어에 기초한 기초온톨로지를 구축하고 여기에 기초하여 온톨로지를 확장하는 방법을 제기하였다. 그리고 확장된 온톨로지에 기초하여 사용자질문에 적합한 패쎄지검색방법을 새롭게 해결하였으며 제안한 방법에 대한 평가를 진행하였다.

참 고 문 헌

[1] Wei Xu et al.; Proceedings of the 5th International Joint Conference on Natural Language Processing, 11, 1046, 2011.

[2] Dan Moldovan et al.; ACM Transactions on Information Systems, 21, 2, 133, 2003.

[3] Petr Knoth et al.; Proceedings of the 23rd International Conference on Computational Linguistics(Coling 2010), 10, 590, 2010.

[4] Sara Salem et al.; Proceedings of the 23rd International Conference on Computational Linguistics(Coling 2010), 10, 967, 2010.

주체103(2014)년 7월 5일 원고접수

A Method of Passage Retrieval based on Ontology

Ri Chong Han

This paper suggests an ontology construction method on hub words and a method of passage retrieval based on constructed ontology.

Key words: ontology, passage retrieval, question answering