시소러스 작성을 위한 개념 획득 도구

김 명철, 이 운재, 최 기선, 김 길창
한국과학기술원 전산학과

A Concept Acquisition Tool for Thesaurus Construction

Myoung-cheol Kim, Woon-jae Lee, Key-Sun Choi, Gil Chang Kim
Dept. of Computer Science, KAIST

요 약

시소러스를 작성하기 위해 시소러스 작성자가 고려해야 하는 문제는 단어간의 개념 관계이다. 단어간의 관계는 계층구조에 정의된 개념을 기반으로 분석하여 하향식으로 시소러스를 작성하는 것이 일반적이다. 이러한 방식은 작성자에 의존적이므로 시소러스의 정확도를 보장할 수 없고 주관적인 성향을 가진다. 그래서 Corpus에서 자동으로 개념 및 개념 관계를 추출하는 상향식 방법들이 다양하게 시도되고 있다.

본 논문에서는 시소러스 작성을 위한 자동 개념 획득 도구를 설계, 구현하였다. Mutual Information이라는 방법을 이용하여 공기 정보(Collocation)를 정량화하고 이를 통하여 단어간의 개념관계의 크기를 측정한 후 개념 관계의 크기(MI 값)가 큰 값을 선택하여 개념 화일을 작성한다. 실험 결과로 얻은 개념 화일은 두 개념간의 밀접도를 나타내므로 시소러스 작성에 매우 유용하다.

I. 서 론

최근에, 정보 검색의 효율 및 정확성을 높이기 위하여 지식에 바탕을 둔 지적 정보검색 시스템들이 등장하고 있다. 이러한 시스템들은 그 지식구조로서 실세계의 개념과 그 관계를 표현하는 지식베이스를 이용하고 있다. 이런 지식베이스의 종류로는 개념 공간(Concept Space), 의미망(Semantic Network), 시소러스 (Thesaurus)등이 있으며 이들은 모두 개념과 그 개념들간

의 관계를 나타내주고 있다. 이 중 개념 공간이나 의미망등은 시소러스를 효율적으로 컴퓨터에 표현한 또 다른 형태이기도 하다.

시소러스(Thesaurus)란 희랍어에서 파생된 용어로 '지식의 보고'또는 '백과사전'이라는 의미를 갖고 있다. 이 용어는 Roget의 'Thesaurus of English Word and Phrases'에서 처음으로 기술되었는데 그 의미는 '어떤 개념을 가장 적절히 표현할 수 있는 표목을 선정하기 위하여 만들어진 어구의 집대성'이라고 되어 있다. 따라서 일반 사전은 용어의 의미를 모를 때 사용하는데 비하여 Roget의 시소러스는 그 반대로 개념은 알고 있으나 그에 해당하는 용어를 모를 때 사용하는 것이다. 즉, 시소러스의 정의는 협의로는 '자연언어를 통제언어로 변환할 때 도움이 되는 수단'이며 광의로는 '용어간의 관계를 표시하는 표'라고 할 수 있다.

정보검색에서 시소러스는 색인작업시에는 적절한 색인어의 선택과 색인어의 통제를 위해 필요하며, 검색시에는 적절한 탐색 용어의 선택을 지원한다. 이 외에 시소러스는 용어통제 및 탐색어의 확장이나 축소를 통하여 검색 효율을 조절하는 데에도 사용된다. 즉, 용어간의 계층관계 및 연관관계를 이용하여 포괄적인 탐색을 하거나 특정한 용어를 사용하여 보다 한정된 탐색을 함으로써 검색문헌의 수를 적절하게 조절할 수 있다.

이와 같이 시소러스는 정보검색에 있어서 시스템과 검색자간의 교량적인 역할을 하는 도구로서 검색효율을 높이는데 중요한 역할을 한다. 그러나 시소러스안의 개념 자체 및 그 개념들간의 관계를 명확히 정의하기는 매우 힘들며 각 시스템간에도 많은 차이가 존재한다. 또한 이런 개념을 획득하는 문제 및 각 개념관계의 거리를 구하는 문제도 매우 어렵다.

본 논문에서는 시소러스를 얻기 위한 전단계로서 Corpus로부터 개념관계를 구하는 CACA(Computer Aided Concept Acquisition)시스템의 설계 및 구현에 대하여 설명한다.

II. 관련 연구

1 개념 표현

정보 검색에서 일반적으로 많이 사용되는 개념 표현 방식은 다음과 같다.

1. Classification : 개념들을 계층적으로 구조화
   예) UDC (Universal Decimal Class) : 모든 주제를 UDC로 표현

2. Thesauri : 각 개념간의 관계 정의

- 일반화,세분화등의 관계 (polyhierarchical 관계)
  : NT (Narrow Term), BT (Broad Term)
- Asymmetric, Symmetric 관계 : RT (Related Term)
- Equivalence 관계 : SYN (Synonym)

3. Semantic Network

4. Concept Space Model : 가중치를 갖는 에지를 이용한 그래프 표현

정보검색에서의 개념표현은 검색 성능 향상에 매우 중요한 요소이다. 그러나 현재까지의 개념 표현은 부분적으로만 실험되어 왔다. 이는 자연언어 처리 기술과 융합되어 Corpus로부터 자동개념획득이 된다면 더 좋은 결과를 얻을 수 있을 것이다.

2 Mutual Information

문장에 쓰인 모든 단어는 서로 유기적인 관계를 가진다. 단어는 그 자체로서 의미를 가지기 보다는 다른 단어와 함께 쓰일 때 더욱 명확한 의미를 전달하게 되는 것이다. 따라서 한 단어의 의미는 그 자체에 내재되었다기 보다는 다른 단어에 의하여 정의된다고 볼 수 있는 것이다.

⟨ 단어간의 결합관계 ⟩

문서들을 보면 단어와 단어 사이에는 의미적, 구문적 관계가 있음을 알 수 있다. 예를 들어 '의사'라는 단어는

병원, 간호원, 환자, 진료

등의 단어들과 함께 쓰이는데 이러한 단어의 연관성을 측정하는 한가지 방법으로서 Mutual Information을 사용할 수 있다.

MI는 다음과 같은 식에 의해 구해진다.

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$p(x,y) = freq(x,y)/N$$

$$p(x) = freq(x)/N$$

$N = totalword$

MI의 값의 의미는 다음과 같다.

$MI(x,y) \gg 0$	x와 y는 밀접한 상관 관계를 가지고 있다. x와 y가 함께 사용되는 경우가 많다. 예를 들면 유의어, 반의어, 관련 용어, 또는 복합 명사 등이다.
$MI(x,y) \simeq 0$	x와 y는 아무 관계가 없다.
$MI(x,y) \ll 0$	x와 y는 대용어 관계에 있다. 따라서 함께 쓰이기 보다는 x가 쓰이면 y는 쓰이지 않고 y가 쓰이면 x는 쓰이지 않는다.

적당한 MI를 구하기 위해서는 p(x,y)를 구하기 위한 window size를 얼마로 하는 것이 좋은가 하는 문제가 있다. 즉, 이는 x와 y가 몇개 단어 거리 이내에 있을 때 서로에게 영향을 미치는가의 문제이다. 실제 문장에서 x와 y는 단어 거리 보다는 문장의 구조에 의해 영향을 받는다. 그러나 문장의 구조를 알기 위해서는 parsing을 해야하는 문제가 발생하므로 단어간의 거리에 제한을 두는 것이 타당하다. 한 단어가 가장 영향을 많이 미치는 것은 바로 앞 뒤의 단어이다. 그러나 두번째 혹은 세번째 단어도 영향을 미치므로 그것을 배제할 수는 없다. 또한 너무 멀리 떨어져 있는 단어간에 상관 관계를 따지는 것은 무의미하다. 따라서 window size 5 ~ 10 정도의 단어가 가장 밀접한 관계를 가진다고 보여진다.

한편 정보검색에서 사용되는 시소러스를 작성하기 위해 MI를 구하고 이를 적절하게 처리하면 실용적인 시소러스를 얻을 수 있다. MI가 높은 값들은 BT, NT, RT 등을 모두 포함하며, 손으로 작성된 시소러스가 의미에 치중해 실제로 잘 사용하지 않는 단어들도 모두 기술해야 하고 더 많이 사용되는 단어 관계를 구분하는 방법이 모호한 반면, MI는 실제로 사용되는 단어간의 관계만을 보임으로써 실용성을 가진다.

III. Corpus로부터 자동 개념 획득

본 시스템은 크게 2개의 모듈로 구성되어 있다. 앞 부분은 원문 Corpus를 입력으로 받아들여 명사에 해당하는 용어들만을 추출하는 용어 추출 모듈이다. 이 모듈은 기존의 색인어 추출 시스템인 KAIS를 부분적으로 이용하였으며 형태소 해석,애매성 제거, 구문해석,색인어 추출, 불용어 제거등을 수행한다. 이와 같이 하여 얻은 색인 화일은 Corpus에서 개념으로서 중요하다고 생각되어지는 개념어(용어)들의 리스트가 된다. 즉,이 중간결과인 개념어 화일은 Corpus의 또 다른 표현이 되며 Corpus가 담고 있는 개념만을 추출한 골격 화일이 된다. 시스템의 후반부는 이 개념어 화일을 다시 입력으로 받아 각 개념간의 Mutual Information을 구하는 부분이다. 이 때 각 개념의 빈도수, 두 개념간의 공기 빈도수등의 정보를 이용하여 두 개념간의

상호 관련성을 나타내주는 Mutual Information을 계산한다.

여기서 개념어란 개념을 이루는 요소중의 하나로써 개념은 개념어와 그 개념어들간의 Mutual Information으로 이루어진다고 가정한다.

이를 그림으로 보면 다음과 같다.

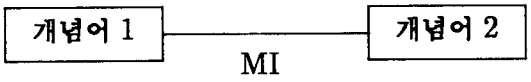

그림 1: CACA에서의 개념 표현

어떤 Corpus에 나타나는 개념어가 n개 존재한다고 할 경우, N x N 행렬이 가능하다. 그러나 대부분의 행렬값이 0이 되는 Sparse Matrix가 되므로 마지막 결과는 MI값이 일정한 기준 값을 넘는 각 개념어 1에 대한 타 개념어들의 리스트 형식을 갖게 한다. 또한 각 개념어쌍은 탐색의 효율을 높이기 위해 역순서쌍도 표현해준다.

본 연구에서는 시소러스를 얻기위한 전단계로서 Corpus로부터 개념을 획득하고 각 개념간의 거리인 가중치를 나타내는 MI값을 구하고자 한다. 이 CACA(Computer Aided Concept Acquisition) 시스템 구성은 다음과 같다.

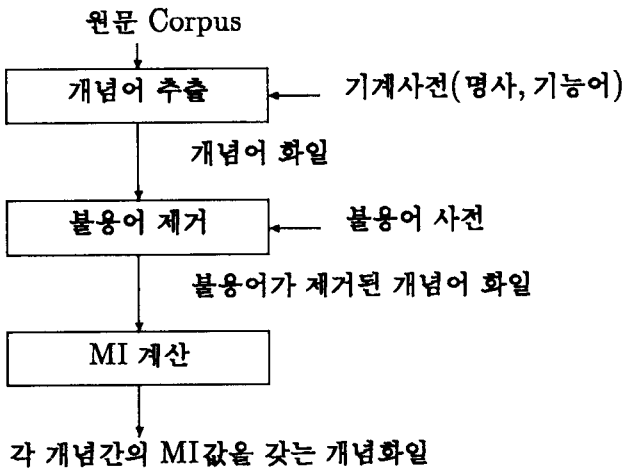

그림 2: CACA 시스템의 구성

1 개념어 추출

이 모듈은 대상 Corpus의 각 문장을 형태소 해석과 구문해석한 결과로부터 개념어를 추출한

다. 구문해석은 한국어 용언이 가질 수 있는 격틀 (Case Frame)을 이용하여 각 문장의 필수격을 찾아 낸다. 그리고 필수격이 되는 명사,명사 상당어구가 전체 Corpus에서 중요한 역할을 할 수 있다는 가정에서 필수격에 해당하는 명사,명사구를 개념어(Concept Word) 후보들로 선택한다. 개념어 후보로 인식된 명사,명사구들중에서 개념어로써 가능성이 없는 단어는 불용어 사전을 이용하여 제거한 다음 적절한 개념어를 추출한다. 이 모듈은 크게 3부분으로 나누어져 있다. 처음은 대상 Corpus의 각 문장을 입력으로 하는 형태소 해석기,형태소 해석기 결과를 가지고 구문분석을 행하는 구문분석기, 불용어 제거를 행하는 불용어 처리 부분으로 구성되어 있다.

2 Mutual Information 계산

MI를 계산하기 위해서는 적당한 크기의 window 내에 있는 단어 pair를 조사하는 일이 필요하다. 일단 windwos size가 결정되면 각 단어에 대하여 window size 이내의 거리에 있는 단어와 pair를 만든다. 최종적으로 각 단어의 사용 빈도와 각 단어 pair의 사용 빈도를 찾아내어 각 단어 pair에 대한 MI를 구하게 된다.

본 시스템에서 MI를 계산하는 알고리즘은 다음과 같다.

1. 적당한 window size를 사람이 선택한다.
2. 대상 text내에 있는 모든 문장에 대하여
    2.1 문장 내의 모든 단어에 대하여
        2.1.1 window에 기억된 단어와 결합하여
            pair를 만든다.
        2.1.2 이미 존재한는 pair인 경우에는
            pair counter를 1 증가시키고
            아니면
            데이타베이스에 pair를 등록한다.
        2.1.3 2.1.2와 같은 방법으로 단어를
            데이타베이스에 등록한다.
        2.1.4 word counter N을 1 증가 시킨다.
        2.1.5 문장의 끝이면 window에 기억된 단어들을
            clear한다.

아니면 2.1.1로 가서 계속한다.

3. 데이타베이스에 있는 모든 단어 pair에 대하여

3.1 단어 x와 y에 대한 freq(x)와 freq(y), freq(x,y)를 데이타베이스에서 구한다.($freq = counter$)

3.2 $p(x) = freq(x)/N$

$p(y) = freq(y)/N$

$p(x,y) = freq(x,y)/N$

3.3 MI(x,y)를 구한다.

3.4 결과를 출력한다.

4. 끝

IV. 실험 결과 및 분석

1 실험 결과

본 연구에서 채택한 대상 corpus는 "정보 검색론"이라는 책자이다. 이 corpus의 크기는 4,164 line, 29,922 어절이다. 이 corpus에서 추출한 명사 화일은 24,795 단어이고, 불용어를 제거한 명사 화일은 15,768 단어이며 전체 단어수는 4266개 이다.

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

X Y [freq(x,y), freq(x), freq(y), p(x,y), p(x), p(y), MI(x,y)]

가공 방법 [5, 6, 160, 0.000393205, 0.000471846, 0.0125826, 4.04939]

가공 색인언어 [5, 6, 60, 0.000393205, 0.000471846, 0.00471846, 5.46443]

가공 주제분석 [9, 6, 38, 0.00070777, 0.000471846, 0.00298836, 6.97139]

가능성 색인어 [9, 6, 172, 0.00070777, 0.000471846, 0.0135263, 4.79306]

가능성 용어 [5, 6, 220, 0.000393205, 0.000471846, 0.017301, 3.58996]

가변장 고정장부분 [8, 5, 5, 0.000629129, 0.000393205, 0.000393205, 9.9905]

가변장 논리레코드 [10, 5, 25, 0.000786411, 0.000393205, 0.00196603, 7.9905]

위의 예에서 나타난 MI 정보는 다음과 같은 개념 관계를 나타낸다.

1. 복합명사: 가공 — 방법

2. 반의어, 유의어 : 가변장 — 고정장부분(반의어)

3. Related Term : 가중치 — 기준치

4. Broader Term : 가변장레코드 — 레코드

5. Narrower Term : 레코드 — 가변장레코드 ...

위 개념 쌍들 간의 MI값은 두 개념어 간의 거리를 나타내 줌으로써 시소러스 작성자에게 도움을 줄 수 있다. 시소러스는 체계적이고 확실한 관계만을 포함하지만 개념 관계의 밀접도를 나타내지 못하고, 직관적으로 정의하기 어려운 관계에 대해서는 많은 노력이 필요하다. 반면 corpus에서 얻은 개념 화일은 비체계적이나 개념 관계의 밀접도를 표시할 수 있고 정의되지 않은 관계를 가진 개념들의 개념관계도 포함한다. 따라서 본 연구의 성과로 볼 수 있는 것은 시소러스의 작성 도구로서, 또는 시소러스와 상호 보완되는 기능을 지닌 개념 획득 도구를 구현하였다는 것이다.

2 결과 분석

MI을 추출하는 경우 window size와 collocation 정보의 두 인수가 매우 중요한 작용을 한다. MI 값을 이용하는 목적에 따라 이 두 인수를 조정해 주어야만 한다. Church의 경우 사전 작성에 필요한 의미 분류를 위하여 단어의 개념을 MI를 이용하여 추출하였다. 이 때 적절한 window size는 5이었고 이것은 단어 간의 의미를 한정하는 데 중점을 두었기 때문이다. Magerman의 경우는 태깅된 corpus로 부터 품사간의 MI를 구하고 이를 이용하여 품사열에서의 disjunction을 구하고 이를 parser의 preprocessor로 이용하였다. 이때의 적절한 window size는 10이었다.

V. 결 론

본 연구에서는 시소러스 작성 도구로서의 개념획득도구를 설계,구현하였다. 개념 화일은 Corpus로부터 자동으로 추출된 개념쌍간의 MI값으로 구성된다. 여기서 개념은 두 개념어 쌍과 그들간의 MI값으로 정의한다. 이 때의 MI값은 두 개념간의 관련도를 나타내 주고 있으며 향후 개념간의 거리를 계산하는데 근간이 될 수 있다. 즉, MI값이 크면 두 개념간의 밀접한 관계가

있음을 나타내고 작으면 별 관계가 없음을 나타낸다. 이는 또한 한 개념과 관련이 있는 개념들 간의 관련 밀접도가 될 수 있다.

실험은 '정보 검색론'이라는 책자 4164 line, 29922어절에 대하여 수행하였으며 시소러스 작성을 위한 개념 획득을 수행할 경우 적절한 Window크기와 최소 공기횟수 (w,c) 는 (10,5) 임을 실험을 통하여 얻었다. 이 때의 결과로서 얻은 개념쌍은 5514개이며 계산된 MI 값은 개념간의 거리를 계산할 때 유용하게 사용될 수 있다.

향후 연구과제로는 Corpus 크기를 늘려서 다양한 종류의 Corpus에 대해서 더 많은 실험을 하면 더 정확한 개념쌍 및 MI값을 추출할 수 있을 것이다. 이 외에도 Corpus에 나오는 자연언어 문장에 대하여 구문해석 및 의미해석을 수행하면 주어-동사, 목적어-동사 관계 등의 관계의 구분을 더 명확히 하여 좋은 결과를 얻을 수 있을 것이다.

〈 부록 A. 시소러스의 예 〉

검색		
	USE	정보검색
정보검색		
	UF	검색
	UF	문헌검색
	NT	데이터검색
	NT	서지편집
	NT	참고업무
	NT	탐색
	NT	탐색결과
	NT	탐색과정
	NT	탐색기법
	NT	탐색질문
	NT	탐색프로파일
	RT	문헌기술
	RT	색인언어
	RT	정보/도서관망
	RT	정보축적장치
	RT	참고봉사

〈 부록 B. MI 값을 갖는 개념 화일의 예 〉

* 는 시소러스에 등록 가능한 개념어

X Y [freq(x,y), freq(x), freq(y), p(x,y), p(x), p(y), MI(x,y)]

검색 가능성 [5, 170, 6, 0.000393205, 0.013369, 0.000471846, 3.96193]

검색 가중치 [27, 170, 52, 0.00212331, 0.013369, 0.00408934, 3.27941]

검색 개발 [10, 170, 44, 0.000786411, 0.013369, 0.00346021, 2.08746]

검색 갱신 [9, 170, 12, 0.00070777, 0.013369, 0.000943693, 3.80993]

검색 *검색결과 [17, 170, 9, 0.0013369, 0.013369, 0.00070777, 5.1425]

검색 *검색시스템 [17, 170, 17, 0.0013369, 0.013369, 0.0013369, 4.22497]

검색 *검색조건 [6, 170, 2, 0.000471846, 0.013369, 0.000157282, 5.80993]

검색 *검색효율척도 [6, 170, 4, 0.000471846, 0.013369, 0.000314564, 4.80993]

검색 계산처리 [6, 170, 2, 0.000471846, 0.013369, 0.000157282, 5.80993]

검색 공식 [16, 170, 10, 0.00125826, 0.013369, 0.000786411, 4.90304]

검색 기억장소 [5, 170, 13, 0.000393205, 0.013369, 0.00102233, 2.84645]

검색 기준치 [9, 170, 15, 0.00070777, 0.013369, 0.00117962, 3.488]

검색 누락률 [12, 170, 5, 0.000943693, 0.013369, 0.000393205, 5.488]

검색 능력 [12, 170, 7, 0.000943693, 0.013369, 0.000550488, 5.00257]

검색 대상 [11, 170, 36, 0.000865052, 0.013369, 0.00283108, 2.51447]

검색 *데이터베이스 [11, 170, 93, 0.000865052, 0.013369, 0.00731362, 1.14524]

검색 도서관 [13, 170, 33, 0.00102233, 0.013369, 0.00259516, 2.88101]

검색 도치파일구조 [7, 170, 4, 0.000550488, 0.013369, 0.000314564, 5.03232]

검색 등록번호 [20, 170, 16, 0.00157282, 0.013369, 0.00125826, 4.54689]

검색 레코드 [35, 170, 149, 0.00275244, 0.013369, 0.0117175, 2.13508]

검색 목적 [13, 170, 29, 0.00102233, 0.013369, 0.00228059, 3.06743]

검색 문장 [5, 170, 26, 0.000393205, 0.013369, 0.00204467, 1.84645]

검색 *문헌수 [8, 170, 4, 0.000629129, 0.013369, 0.000314564, 5.22497]

검색 *문헌총수 [8, 170, 2, 0.000629129, 0.013369, 0.000157282, 6.22497]

검색 *문헌파일 [10, 170, 16, 0.000786411, 0.013369, 0.00125826, 3.54689]

검색 방법 [33, 170, 160, 0.00259516, 0.013369, 0.0125826, 1.94743]

검색 방지 [10, 170, 9, 0.000786411, 0.013369, 0.00070777, 4.37697]

참고 문헌

[1] M. Bartschi, *"Overview of Information Retrieval Subjects"* , IEEE Computer, May 1985

[2] P.R.Cohen and R. Kjeldsen, *"Retrieval by Constrainted spreading Activation in Semantic Network"* , IP&M Vol. 23, No. 4, 1987

[3] W.B.Croft and R.H.Thompson, *"I^3R: A New Approach to the Design of Document Retrieval Systems"* , J. of ASIS, Vol. 38, No. 6, 1987

[4] H.P. Giger, *"Concept based Retrieval in Classical IR Systems"* , Procedding of ACM SIGIR Conference, 1988

[5] M.Dillon and Ann S.Gray, *"FASIT: a Fully Automatic Syntactically Based Indexing System"* , J. of ASIS, Vol. 34, No. 2, 1983

[6] E.A.Fox and J.T.Nutter, *"Building a large Thesaurus for Information Retrieval"* , Procedding of ACL Conference, 1988

[7] D. D. Lewis and W.B.Croft, *"Term Clustering of Syntactic Phrases"* , Proceeding of ACM SIGIR Conference, 1990

[8] K. W. Church, P. Hanks, *"Word Association Norms, Mutual Information, and Lexicography"* , Computational Linguistics Vol. 16, No. 1, 1990.3.

[9] D. M. Magerman, M. P. Marcus, *"Parsing a Natural Language Using Mutual Information Statistics"* , AAAI, 1990

[10] 한국과학기술원, "다국어 *DB*를 위한 *Keyword* 및 *Index* 생성 시스템 개발", 과학기술처, 1991

[11] 한국과학기술원, "지능형 정보검색에 관한 연구", 한국통신, 1991