추론망을 이용한 시소러스의 자동구축*

박영찬, 한영석, 최기선
한국과학기술원 전산학과

Automatic Thesaurus Construction using Inference Networks

YoungChan Park, YoungSeok Han, Key-Sun Choi
KAIST Computer Scienct Dept.

요 약

정보 검색의 효율은 정보검색 시스템에서 사용되는 지식의 질에 상당한 영향을 받는다. 이러한 지식 표현의 한 가지로 널리 사용되고 있는 것이 시소러스이다. 이러한 시소러스의 구축은 지식을 얼마만큼 잘 구성하는가에 있다. 따라서 시소러스의 자동 구축은 상당한 효용을 지니게 된다. 시소러스의 자동구축시에 대량의 말뭉치로부터 지식을 추출하는 방법론이 많이 연구되어 오고 있다. 그러나 이러한 방법은 단어의 통계적인 행태에 크게 의존하고 있기때문에 자료 희귀(data sparseness)의 문제가 큰 장애 요인이 되고 있다. 본 연구에서는 이러한 자료희귀문제를 해결하기 위해 추론망을 사용하고자 하는 모델을 제시하고자 한다.

1. 서론

정보검색의 목표는 검색의 효율을 증가시키는 것이다. 이러한 검색 효율은 사용자의 질의에 가장 적합한 문헌만을 골라서 제공하여 주는 것을 말한다. 이러한 검색의 효율은 시스템에 의해 사용되는 지식에 크게 의존한다. 전문가의 지식표현 중에 가장 널리 쓰여지고, 가장 효율증대에 기여가 크다고 여겨지고 있는 것이 시소러스이다[Salton89]. 잘 구성된 시소러스가 정보 검색에 끼치는 효과는 잘 알려져 있지만, 이러한 시소러스의 자동 구축법은 그다지 크게 확립되어지지 않았다.

수작업에 의한 시소러스의 구축은 막대한 노력과 시간을 들이는 작업이지만 단어의 유동적인 특성으로 인해 계속적인 유지 보수가 행해져야만 한다. 따라서 이러한 시소러스를 자동으로 구축하는 기법에 대한 연구가 활발히 진행되고 있다[Crouch90, Rada86, Soergel74]. 이러한 시소러스의 자동구축에 있어서 Soergel(1974)이 언급한 바와 같이 단어들 간의 통계적인 관계는 그 단어의 의미적 해석과도 관련이 있다는 말은 상당히 중요한 의미를 가진다. 즉 단어의 문헌내에서의 통계적 정보로부터 그 단어의 의미를 해석할 수 있다는 것은 대량의 말뭉치(corpus)로부터 단어와 의미 구조를 얻어낼 수 있다는 것을 의미한다.

단어의 통계적인 정보로 부터 단어들의 관계를 결정하고자 하는 연구가 많이 진행되어 왔다[Hindle93, Pereira93]. 그러나 이러한 방법들은 단어 간의 거리를 대체로 단어의 통계적인 행태에 크게 의존하고 있다. 이러한 통계적인 접근 방법은 자료 희귀(Data Sparseness)가 가장 큰 문제로 부각되고 있다[Pereira93].

Hindle(1993)은 이러한 자료 희귀문제를 해결하기 위해 말뭉치에 나타나지 않은 자료를 그와 비슷한 자료들로 부터 얻어내는 방법을 제한하였다. 구체적으로 말하면, 동사와 중심명사의 관계를 얻어내는데 있어서, 말뭉치에 나타나지 않은 동사와 중심명사의 관계를 그 동사와 비슷한 동사로 부터 얻어 낸다는 방법이다. 또한 Crouch(1990)는 벡터 공간 모델에 기반하여 문헌을 계층 구분하고, 이러한 문헌구분으로부터 단어의 클래스를 얻어내는 모델을 제안하였다. 그러나 Hindle의 접근 방법은 동사의 유사도를 제한하는 범위가 명확하지 않으며, 유사도 비교의 방법도 명확하지가 않은 단점이 있다[Pereira93]. Crouch의 제안은 자료희귀문제를 해결하는데 있어서 Hindle과 같은 비슷한 행태가 아닌 그 단어가 속한 문헌의 행태로써 해결하고자 하는 방법이다. 단어의 자료희귀를 보다 단위가 큰 문헌이라는 단위로 해결하고자 하는 방법은 우리의 직관에 일치하기는 하나, 단어 클래스를 얻어내는데 있어 단어가 문헌에서 차지하는 역할이 명확하지 않은 단점이 있다.

이러한 자료의 희귀를 다루기 위해, 본 논문에서는 지역적인 의존관계를 이용하여 전체적인 의존 관계를 계산할 수 있는 추론망을 사용하고자 한다. 추론망은 베이지언망(Bayesian network), 인과망(causal net)등의 다양한 이름들로 불리고 있으며, 그 응용에 따라 다양한 구조와 사용방법들이 제안되어 왔다. 전통적으로 전문가 시스템에서 많이 쓰여져 왔는데, 하나의 노드가 전문적인 지식을 나타내며, 그 노드를 잇는 에지는 두 지식간의 논리적 의존관계를 나타낸다. 추론망은 무작위적인 노드집합 간의 조건 확률을 구할 수 있는 특성을 갖는다. 이러한 추론망에 자동 학습의 개념을 추가한 것이 Neal(1992)에 의해 제안되었다. 본 논문에서는 추론망의 이러한 특성을 이용하여 자료희귀문제를 해결한 자동 시소러스 구축을 시도하고자 한다.

2장에서는 추론망의 기본 개념에 대해 설명하고, 3장에서는 이를 이용한 시소러스의 자동구축법에 대해 알아보고자 한다.

*본 연구는 (주)한국통신의 장기 기초 연구 "지능형 정보 검색 연구"의 지원을 받아 수행하였슴.

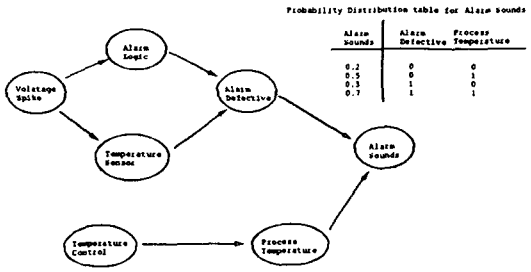

그림 1: 신념망의 예

이어 4장에서는 실험을 통한 모델의 검증과 5장에서는 결론을 맺는다.

2. 추론망(Inference Networks)

추론망은 베이지언망(Bayesian Net),인과망(causal net),영향 다이어그램(influence diagram), 신념망(belief net)등으로 불리우는 한 집합의 원소들간의 확률적인 분포를 나타내는 모델을 말한다. 이러한 망구조들은 원래 전문가의 지식을 표현하고 이를 이용하여 어떠한 결정을 내리고자 하는 전문가 시스템에 주로 이용되어 왔다. 추론망은 각각의 에지에 국부지역간의 의존도를 조건 확률로 표현한다. 이러한 조건확률의 부여는 전문가에 의해 주어지는 이른바 수동구축방법을 기반으로 한다. 수동구축방법의 제한으로 인해서 그 동안 추론망을 사용한 응용이 제한되어 왔다. Neal(1992)이 제안한 시그모이드 신념망은 이러한 기존의 프레임에 신경망적인 요소를 결합하여, 자동학습기법을 제안하여 추론망 자동 구축의 방법을 제시하였다. 이 장에서는 추론망을 신념망의 관점을 통하여 설명하고, Neal이 제안한 시그모이드 신념망에 대해서 기술하고자 한다.

2.1 신념망(Belief Network)

신념망은 노드와 방향성을 갖는 에지로 구성되며, DAG (Directed Acyclic Graph) 형태를 갖는 그래프를 말한다. 그래프의 노드는 표현하고자 하는 모델의 변수를 의미하며, 에지는 이들 사이의 관계를 나타낸다. 노드는 상수, 불확실한 양 또는 목적, 그리고 행해질 결정 등을 표현한다.

그래프 내의 노드는 $N = 1, ..., n$로 나타내고, 각각의 노드에 대응하는 모델의 변수들은 $X_1, ..., X_n$으로 나타낸다. 각각의 변수 X_i에 대해서는 유한개의 출력값인 Ω_i로 정의 되고 각각의 노드에는 π_i 의 확률 분포 테이블 (probabilistic distribution table)이 존재한다. 또한 노드 i의 조건 노드는 $C(i)$로 나타내어 진다. 일반적으로 I, J등의 대문자는 노드의 집합을 나타내며, 소문자는 특정 노드 한 개를 가리킨다. 만약 J가 노드의 집합이며 $N \supseteq J$를 만족할때, X_J는 J에 대응하는 노드이고 이 노드들의 교환곱(Cross Product)의 출력은 Ω_J로 나타낸다. 예를 들어 X_j의 조건노드들은 $X_{C(j)}$가 되고 이들의 출력은 $\Omega_{C(j)}$으로 나타낸다. 그림 1은 일반적인 신념망의 예를 보여주고 있다.

신념망에서의 일반적인 확률추론 문제는 두 노드 집합I, J간의 조건 확률, $Pr\{X_J|X_I\}$을 구하는 것이다. 신념망에서는 베이지언 정리를 이용하여 임의의 조건 확률은 접합 확률(joint probability or marginal probability)로 변환될수가 있다.

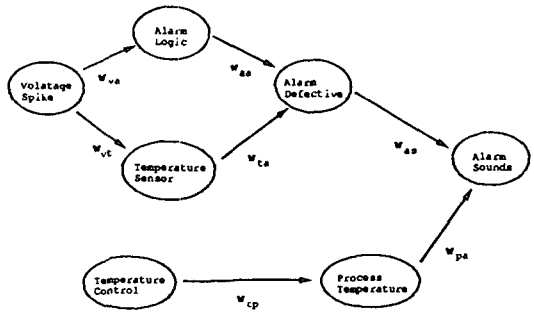

그림 2: 시그모이드 신념망의 예

일반적으로 접합확률은 다음과 같이 개개의 조건 확률로 분리될 수 있다.

$$\begin{aligned} Pr(X_K) &= Pr(X_{i_1}) \cdot Pr(X_{i_2}|X_{i_1}) \cdot \\ &\quad \ldots \cdot Pr(X_{i_k}|X_{i_1}, X_{i_2}, \ldots, X_{i_k}) \\ &= \pi_{i_1}(X_{i_1}|X_{C(i_1)}) \cdot \pi_{i_2}(X_{i_2}|X_{C(i_2)}) \cdot \\ &\quad \ldots \cdot \pi_{i_k}(X_{i_k}|X_{C(i_k)}) \end{aligned}$$

여기서 π_{i_1}는 X_{i_1}의 가장자리(marginal)확률분포를 나타내므로 $C(i_1)$는 널(null)이 된다.

위의 계산을 통하여 구하고자 하는 문제의 확률을 계산할수가 있다. 신념망에서의 정확한 확률값의 계산에 관한 여러가지 방법들이 제안되어 왔다[Neal92, Lau88, Metropolis53, Ackley85]. 그러나 이러한 방법은 그래프의 구조를 제한하거나 또는 최악의 경우 지수승의 시간 복잡도를 가진다.

2.2 시그모이드 신념망(Sigmoid Belief Network)

앞에서는 신념망에 대해서 알아보았다. 이러한 신념망은 전문가의 지식을 표현하는 것이 주된 목적이었다. 따라서 각각의 지역 조건 확률 분포를 모두가 사람이 작성하는 것이었다. 이에 반해 Neal(1992)이 제안한 시그모이드 신념망은 실험적인 자료로부터 자동적으로 신념망을 자동구축할 수 있는 신념망의 일종이다.

시그모이드 신념망은 신념망과는 달리 각 에지마다 가중치가 주어져 있고, 각 조건 확률분포는 관련된 에지가중치의 시그모이드 함수로써 표현된다. 노드 j로 부터 노드 i로의 에지의 가중치는 w_{ij}로 나타낸다. 그림 2는 시그모이드 신념망의 예를 보여주고 있다.

각각의 노드 i의 조건 확률분포, π_i는 그의 조건 선행자(conditional predecessor) $C(i)$에 대해 다음과 같이 표현된다.

$$\pi_i(x_i|x_{Ci}) = \sigma\left(i^* \sum_{j \in Ci} j^* \cdot w_{ij}\right)$$

시그모이드 함수 $\sigma(t)$는 $\sigma(t) = \frac{1}{1+e^{-t}}$로 표현된다.

시그모이드 신념망에서의 학습은 주어진 학습 자료를 통하여 각각의 에지의 가중치를 바꾸어가는 과정이라고 할 수 있다. 이러한 학습은 많이 사용되고 있는 최대 근사법(Maximum-Likelihood Estimation)이다.

시그모이드 신념망에서의 추론은 원래의 추론망과 같은 $Pr\{X_J|X_I\}$를 구하는 문제이다. 그러나 내부적인 표현의 차이로 인하여 이러한 조건 확률의 정확한 값이 적절히 얻어질 수 없다. 따라서 이를 근사하는 방법으로 깁스 샘플링(Gibbs sampling)[Gelfand90]을 사용한다.

시그모이드 신념망에서 노드 i에 대한 조건 확률 분포는 다음과 같이 나타내어 진다. S_i는 시그모이드 신념망의 각 노드들이 가지는 상태를 나타내는 상태벡터이다.

$$
\begin{aligned}
&P(S_i = x|S_j = s_j : j \neq i)\\
&\propto P(S_j = x|S_j = s_j : j < i) \cdot\\
&\prod_{j>i} P(S_j = s_j|S_i = x, S_k = s_k : k < j, k \neq i)\\
&\propto \sigma(x \sum_{j<i} s_j w_{ij}) \cdot \prod_{j>i} \sigma(s_j(xw_{ij} + \sum_{k<j,k\neq i} s_k w_{jk})).
\end{aligned}
$$

여기서 "$S_j = s_j : j \neq i$"라는 식은 $j \neq i$를 만족하는 모든 j 에 대한 $S_j = s_j$의 결합 확률을 의미한다.

3. 추론망을 이용한 시소러스 구축

3.1 추론망으로서의 공기지도

우리가 확률적 추론의 근간으로 삼는 기본틀은 Neal(1992)에서 설명한 시그모이드 신념망이다. 시그모이드 신념망은 앞에서 언급한 바와 같이 확률 테이블이 없이 가중치를 갖는 에지와 각 노드에서 정의된 시그모이드 함수를 갖는 신념망의 일종이다. 이러한 신념망의 모델의 변형은 통계학에서의 논리적 회귀모델(logistic regression model)[Tamas86, Harter75]의 일반화로 볼 수 있다. 이러한 모델의 변형을 단어의 계층 구조를 얻어내기 위한 확률 근사의 도구로 사용하고자 한다. 즉 노드의 지식내용을 단어로 하고, 각각의 에지의 가중치를 단어간의 2진 공기 분포를 가지도록 모델한다.

정의 1 (공기지도(Collocation Map)) 공기지도란 용어의 확률분포를 부호화하는 신념망이고 각 노드는 하나의 용어에 대응된다.

공기지도는 대량의 문헌 집합으로부터 단어들의 사용 패턴을 찾고 이것을 신념망의 구조로 변화함으로써 만들어 진다. 단어들의 사용패턴을 고정된 윈도우내의 공기정보로 뽑아내고 이를 공기지도에 학습시킴으로써 공기지도가 단어의 공기확률을 가지도록 하는 것이다. 이러한 구축되어진 신념망에서의 추론은 대량의 문헌에 내포된 단어간의 의존관계를 규명하는 일이라 할 수 있다.

구체적으로 공기지도의 구축은 다음과 같은 과정으로 이루어 진다.

1. 문헌을 입력받는다.
2. 명사열을 추출한다.
3. 명사열로부터 μ 형태를 만들어 낸다.
4. 공기지도에 기록한다.

μ 형태란 특정 윈도우내의 2진공기관계를 나타낸다. 가령 단어의 열이 다음과 같다면

(개념, 거리, 정보, 시소러스)

윈도우크기가 3인 μ 형태는

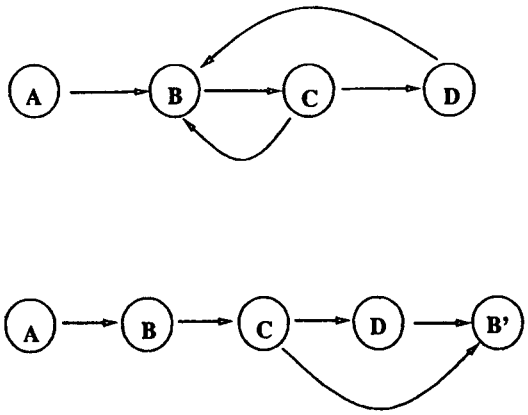

그림 3: DAG로의 변환

(개념,거리), (개념,정보), (거리,정보), (거리, 시소러스), (정보,시소러스) 가 된다.

공기지도의 각 에지의 값은 앞에서 말한 일종의 상호공기정보의 식을 이용하는 방법이 있다. 노드 i와 j를 연결하는 에지의 가중치는 다음과 같이 정의될수 있다.

$$w_{ij} = P(S_j|S_i) \simeq \frac{frequency(i,j)}{frequency(i)}$$

이 때 공기지도는 DAG(Directed acyclic graph)가 되어야 한다. 확률결정을 계산하는데 있어서 DAG의 성질은 상당히 유용한 방법이다. 따라서 DAG를 유지하기 위해서는 그림3과 같이 닫힌구조(cycle)를 제거해야 한다.

이러한 공기지도로 부터 우리가 원하는 두 임의의 집합사이의 의존관계인 조건 확률을 구하는 것은 시그모이드 신념망에서의 일반적인 추론 문제인 $Pr\{f(X_J)|X_K\}$를 구하는 것이다. 이러한 문제는 그 동안 연구되어온 신념망에서의 근사추론 방법의 하나인 깁스 샘플링(Gibbs sampling)을 사용한다.

깁스 샘플링 추론 방법은 한 상태벡터로 부터 상태전이를 통하여 계속적인 샘플을 추출하여, 이렇게 추출된 샘플에 근거하여 구하고자 하는 추론문제인 조건 확률의 값을 구하는 방법이다. 깁스샘플링은 임의의 공기지도 상태로 부터 시작한다. 다음 상태벡터를 구하기위해 임의의 노드를 선택해서 그 상태를 바꾼다. 그리고 이 작업을 계속해서 오랫동안 반복한다. 이러한 반복을 '충분히' 한 후 샘플링을 중단한다. 이렇게해서 얻어진 샘플을 관찰해서 우리가 원하는 확률정보를 추출할 수가 있다. 여기서 '충분히'라는 말는 수렴상태를 의미한다.이는 샘플링을 중단하는 시점을 결정한다. 즉 상태들의 확률값간의 분산을 계산해서 일정한도에 도달하면 즉 수렴을 하면 중단하는 방법도 수렴상태를 고려한 샘플링 중단의 한 방법이다. 본 논문에서 이러한 수렴상태까지의 샘플링을 위해 시뮬레이티드 어닐링(simulated annealing)을 사용한다. 원래 시뮬레이티드 어닐링은 최적의 상태벡터를 구하는데 주로 쓰이는데 여기서는 수렴을 통한 종료점을 찾기 위함이다. 또한 확률이 높은 상태벡터를 샘플로 하기 위함도 있다. 무작위 샘플보다는 양질의 샘플을 얻을 수 있다고 보고되고 있다(Otten,1989).

위에서 언급한 공기지도는 대량의 문헌으로부터 단어의 공기 정보를 얻어내어 이를 단어간의 의존성 계산에 이용하고자 세워진 모델이다. 이러한 모델에 있어서 단어간의 의존성은 시소러스 구축에 있어서 단어간의 계층구조 확립, 즉 단어간의 의미적 유사도를 판별하는데 좋은 도구가 된다. 시소러스가 가지는 계층구조의 연결에 있어서 기존의 연구에서 보듯이, 상대적으로 낮은 출현 빈도를 갖는 단어들의 상관관계를 구하기가 어렵다는 것이다. 이러한 낮은 빈도 단어들간의 상대적인 유사도를 공기지도가 제공하는 조건확률의 추론값으로 대신할수가 있다. 이러한 방법은 자료 희귀를 극복할수 있는 한 방법이 된다.

3.2 공기지도를 이용한 시소러스의 자동구축

시소러스의 자동구축은 시소러스 클래스의 생성과 생성된 클래스간의 계층성 부여의 두가지로 나뉠 수가 있다. 시소러스 클래스란 비슷한 의미를 공유하는 단어의 집합을 의미한다. 이러한 시소러스 클래스가 하나의 개념을 이루게 되며, 하나의 독립된 의미를 가진다고 볼 수 있다. 이러한 개념간의 관계를 이어주는 것이 시소러스의 구축이라고 할수 있다.

3.2.1 계층 클래스(Hierachical class)와 시소러스 클래스(Thesaurus class)

계층 클래스(hierachical class)는 시소러스 상의 동일한 레벨, 즉 동일한 개념적인 위치를 갖는 단어들의 클래스를 의미한다. 이러한 계층클래스는 시소러스의 클래스를 생성하는 근간이 된다. 임의의 n개의 계층 집합으로 단어들을 나눈다고 할때, 이러한 계층집합은 단어들의 응집도를 고려해야 한다. 즉 개개의 단어들이 계층 집합을 이룰 때 단어들이 나타내고자 하는 의미의 분할로 이루어져야 한다는 것이다. 한 단어의 의미의 계층상의 정도는 그 단어의 출현빈도와 관련이 있다는 Forsyth and Rada(1986)의 개념을 이용한다. 따라서 각 단어의 출현빈도로 부터 계층집합의 엔트로피(Entropy)를 얻어내어 엔트로피를 최소화하는 방법이 그 한 가지가 될수 있다.

한 계층집합의 엔트로피양은 다음과 같이 정의 될수 있다.

$$H_m = \sum_{i=1}^{k} -P_i log P_i$$

여기서 $P_i = \frac{Frequency(i)}{\sum_{j=1}^{|m|} Frequency(j)}$ 이며 $|m|$은 클래스 m의 크기를 나타낸다.

단어를 총 n개의 계층 클래스로 나누고자 할때의 전체 엔트로피 H_{total}는 다음과 같이 정의된다.

$$H_{total} = \sum_{i=1}^{n} H_i$$

최종적인 단어계층 집합은 위의 총 엔트로피를 최소화하는 계층 집합을 찾으면 된다. 이렇게 얻어진 계층 클래스는 비슷한 의미 영역을 갖는 단어들로 이루어진다. 따라서 계층 클래스 내에서의 시소러스 클래스 생성이 이루어지면 시소러스 클래스가 계층클래스의 계층을 이어받게 된다. 계층 클래스 내에서의 시소러스 클래스의 생성은 공기지도로 부터 얻어 낸다. 하나의 계층 클래스는 각각의 시소러스 클래스로 이루어지게 된다. 계층 클래스 C_i를 이루는 단어들을 $w_1, ..., w_n$이라고 할때, 시소러스 클래스 T_{i_j}는 다음과 같이 구성된다.

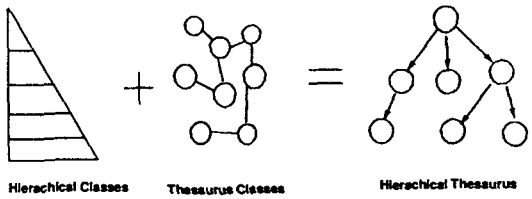

그림 4: 시소러스의 생성

$$C_i = \bigcup T_{i_j}, 1 \leq j \leq m$$

각 시소러스 클래스는 CM에서 제공하는 다음의 조건 확률을 이용하여 구해질수가 있다.

$$T_{i_j} = \{W_k | Pr(W_k | W_j) \geq \Theta_1\}$$

여기서 Θ_1은 임계값을 의미한다. 위의 임계값은 각각의 시소러스 클래스의 응집도를 결정하는 매개변수가 된다. 즉 임계값이 클수록 각각의 시소러스 클래스는 높은 응집도를 가진 여러개의 시소러스를 생성하게 되며, 반대로 값이 작을수록 응집도가 낮은 큰 시소러스 클래스를 생성하게 된다.

3.2.2 시소러스 클래스간의 연결

위에서 얻어진 시소러스 클래스는 계층 클래스로부터 얻어진 클래스이기 때문에 각각의 의미 레벨을 갖게 된다. 이러한 의미레벨은 시소러스의 계층성의 기반이 된다. 그림 3은 시소러스 클래스간을 연결함으로써 계층적 시소러스의 생성을 보여주고 있다. 이러한 시소러스 클래스간에 관계를 지어주는 것이 시소러스 구축의 마지막이 된다. 시소러스는 시소러스 클래스간에 상하위 관계는 계층구조로 부터 얻어진 계층을 사용하며, 시소러스 클래스 C_{i_j}는 C_{i+1_k}간의 의미 관계 is_a는 다음과 같이 구해질수 있다.

$$C_{i_j}\ is_a\ C_{i+1_k}, if Pr(C_{i_j} | C_{i+1} k) \geq \Theta_2$$

위의 임계값 Θ_2는 시소러스의 의미관계인 is_a 관계의 응집도를 결정한다. 이 값이 크면 클수록 시소러스는 적은수의 자식노드를 갖게 되며, 값이 클수록 많은 수의 자식노드를 갖게 된다.

4. 평가

4.1 실험방법

위에서 제시된 시소러스 구축방법에 의해 구축된 시소러스의 유용성을 입증하기 위하여 널리 알려진 CACM IR 실험 문헌 집합을 사용하였다. 먼저 기존에 제시되어온 단어간의 유사성 비교로 사용되는 문헌 벡터(document vector)를 사용하여 단어의 유사도를 비교하는 방법과 공기지도를 통하여 구축된 시소러스를 질의어 확장에 사용함으로써 모델의 유용성을 보이고자 한다.

실험 방법은 CACM IR 실험 문헌집합 내의 질의 64개중에서 낮은 출현빈도를 갖는 단어들을 가지고 있는 26개의 질의를 질의 확장에 사용하였다. 검색은 벡터 검색 모델을 사용하였다.

결과는 일정 재현율(recall)에서의 정확률(precision)을 계산함으로써 평가 하였다.

Recall level	DOC. vector 시소러스	공기지도 시소러스
Level 1	12	21
Level 2	5.48	12
Level 3	10.8	16.98
Level 4	9.11	14.60
Level 5	8.01	12.87

표 1: 문헌벡터 시소러스와 공기지도 시소러스의 성능향상 비교(%)

4.2 실험결과

실험결과로는 공기지도를 이용하여 구축된 시소러스가, 문헌 벡터를 사용한 시소러스보다 평균 18재현율에서의 정확률의 비교이다.

위의 실험결과에서 보듯이 낮은 빈도수를 보이는 단어의 질의 확장에서 공기지도를 이용한 시소러스 구축이 유용함을 볼 수 있다.

5. 결론

정보검색의 발전은 지식의 구축에 있다. 이러한 지식베이스의 자동구축은 상당히 중요한 의미를 가진다. 시소러스는 정형화된 지식베이스의 하나로서 그 효용가치가 매우 높은 지식 표현 방법의 하나이다. 본 논문에서는 기존에 제시되어 왔던 시소러스 구축방법들의 단점인 낮은 출현빈도를 갖는 단어의 통계적인 정보의 부족을 개선한 모델을 제시하였다. 또한 단어간의 연관성을 추론 빗을 이용하여 정형화된 확률 도구로 모델링 하였으며, 이 모델의 유용성을 실험으로 증명하였다.

참고문헌

[Baker] Baker, J. K. 1979. Trainable grammars for speech recognition. Proceedings of Spring Conference of the Acoustical Society of America, 547-550. Boston, MA.

[Ackley85] Ackley, G.E. Hinton and T.J. Sejnowski. (1985). A Learning Algorithm for Boltzmann machines, *Cognitive Science*. 9. 147-169.

[Crouch90] C. J. Crouch. (1990). An Approach to the Automatic Construction of Global Thesaurus, *Information Processing and Management* Vol26, No 5. pp.629-640

[Dempster] A.P. Dempster, N.M. Laird and D.B. Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc.* B 39, 1-38.

[Rada86] Richard Forsyth and Roy Rada, (1986) Machine Learning : application in expert systems and information retrieval, Ellis Horwood.

[Gelfand90] A.E.Gelfand and A.F.M.Smith. (1990). Sampling-based approaches to calculating marginal densities, *J. Am. Stat. Assoc* 85. 398-409.

[Han93] YoungSeok Han, YoungChan Park, Key-Sun Choi (1992). Relevancy Computation of words using Bayesian Net: Application to Automatic Indexing, *Korea Information Science Society SIGAI*, Spring.

[Hindle93] Donald Hindle, 1993. A parser for text corpora. In B.T.S Atkins and A.Zampoli, editors, *Computational Approaches to the Lexicon*. Oxford Universiti Press, Oxford, England

[Lau88] S.L. Lauritzen and D.J. Spiegelhalter. (1988). Local computation with probabilities on graphical structures and their application to expert systems. *J. Roy. Stat. Soc.* 50. 157-224.

[Metropolis53] N. Metropolis, A. W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller. (1953). Equation of state calculations by fast computing machines. *J. Chem. PHys.* 21. 1087-1092.

[Neal92] R.M. Neal. (1992). Connectionist learning of belief network. *Artificial Intelligence* 56. 71-113.

[Pearl88] J. Pearl. (1988). Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference. Morgan Kaufman, San Mateo.

[Pereira93] Fernando Pereira, Naftali Tishby, Lillian Lee. (1993) Distributional Clustering of English Words, *ACL-93*. pp.183-190.

[Salton89] Salton, G. (1989) Automatic Text Processing, Addison-Wesley

[Soergel74] Soergel, D. (1974) Automatic and Semi-Automatic Methods as an Aid in the construction of Indexing language and Thesauri", *Intern. Classif.*, 1,1,pp. 34-39

[Harter75] Harter, Stephen. P. A probabilistic approach to automatic keyword indexing, *JASIS*, 26, 197-206.

[Tamas86] Doszkocs, Tamas. (1986) Natural Language processing in information retrieval *JASIS*, 37, 4, pp. 191-196.