신경망을 이용한 적응형 시소러스
(Adaptive Thesaurus using a Neural Network)

최 종 필 † 최 명 복 †† 김 민 구 †††
(JongPill Choi) (Myeongbok Choi) (Minkoo Kim)

요 약 정보검색 분야에서 시소러스는 용어와 용어 사이의 관계를 나타내어, 질의어와 검색될 정보 사이에 존재하는 용어적 차이를 줄이는데 사용될 수 있다. 시소러스를 사용하는 방법 중 진보된 것은 용어 사이의 관계에 가중치를 주어, 소위 스프레딩 엑티베이션 방법을 이용하여 주어진 용어에서 다른 용어들 사이의 유사성을 측정하여 이를 검색에 이용한다. 그러나, 이러한 방법은 가중치를 어떻게 할당하느냐에 따라 그 결과가 달라지는 문제점이 발생한다. 본 논문에서는 시소러스의 가중치를 사용자의 검색된 정보에 대한 적합성 반응에 근거하여 조절할 수 있는 신경망 기반 시소러스를 제안한다. 제안된 시소러스의 타당성을 위하여 프로토타입의 시소러스를 WordNet으로부터 추출하여 실험하였으며, 그 결과로 recall-precision 값이 향상됨을 보였다.

Abstract In the field of information retrieval, thesaurus can be used to bridge the gap between the terminology used in defining queries and the terminology used in representing documents. In many approaches using thesauri, they compute the similarity between terms using so called spreading activation process with the term weights in thesauri. One difficulty with these approaches is that the resulting performance is sensitive to the weight assignments. In this paper, we propose a thesaurus model using a neural network in which the term weights can be adjusted by users' relevance feedback. Experiment are performed on a small document collection using a proto-typed thesaurus extracted from WordNet. The result shows excellent performance in terms of recall-precision values.

1. 서 론

일반적인 검색시스템에서 사용자는 자기가 원하는 정보를 얻기 위해 시스템에 존재하는 색인어들로 이루어진 질의어를 정보검색 시스템에 입력함으로써 시스템에서 원하는 정보, 즉 문서를 얻는다. 하지만 이러한 시스템은 사용자가 질의어를 생성할 때 사용된 용어가 색인어이어야 하는데 이것은 사용자가 정보검색 시스템의 전문가가 아닌 이상 어려움이 있다. 또한 사용자마다 용어들에 대하여 자신만의 배경지식 그리고 시간적, 공간적 상황에 따라 서로 다른 의미로 사용한다[1]. 이러한 어려움을 해결하기 위한 방법으로 많은 검색 시스템들은 시소러스를 채택하고 있다.

시소러스란 용어들과 용어들 사이의 관계 집합으로 구성된 일종의 용어사전으로 정의할 수 있다. 시소러스는 문헌에 대한 색인과 검색을 위해 정확하고 통제된 용어를 제공해 주어 색인과 검색 작업 시 가장 적절한 용어를 선정할 수 있도록 해준다. 일반적으로 진보된 시소러스는 용어와 용어 사이에 관계를 가중치로 나타내고, 소위 스프레딩 엑티베이션 (spreading activation) [2] 방법을 이용하여 용어간의 유사도를 측정하여 용어간의 관련도를 나타낸다. 그러나, 이러한 방법은 용어간의 가중치가 적절하지 못하면 오히려 검색의 효과가 없거나 나빠질 수 있다. 본 논문에서는 적절한 가중치를 찾는 방법으로 사용자의 검색된 정보에 대한 적합성을 근거로 학습을 통하여 가중치를 찾는 신경망 기반 시소러스 모델을 제안한다. 제안된 논문의 검증을 위하여 프로토 타입의 시소러스를 WordNet[3]에서 수동 추출하

· 본 논문은 한국과학재단의 연구비 지원(과제번호 971-0901-007-2)을 받아 수행함.
† 비 회 원 : 아주대학교 컴퓨터공학과
cjp@ccai.ajou.ac.kr
†† 정 회 원 : 원주대학 행정전산과 교수
cmb@web.wonju.ac.kr
††† 종신회원 : 아주대학교 컴퓨터공학과 교수
minkoo@madang.ajou.ac.kr
논문접수 : 2000년 5월 25일
심사완료 : 2000년 9월 29일

여 실험하였다. 학습된 시소러스를 이용한 검색 결과는 recall-precision 값의 향상을 보여 주고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 시소러스를 위한 기존 연구 분석으로 심볼릭 시소러스의 표준적인 모형과 신경망 기반 시소러스의 모형을 분석한다. 3장에서는 본 연구에서 제안하는 신경망 기반 시소러스의 모형을 위한 이론적 배경, 신경망 구조, 활성화 함수 및 학습 방법을 소개한다. 4장에서는 실험을 통하여 제안된 모형의 타당성을 제시한다. 5장에서는 결론과 향후 연구에 대하여 설명한다.

2. 기존의 시소러스

대부분의 시소러스는 심볼릭 시소러스로 전문적 분야의 정보검색을 위한 시소러스가 대부분이다. 비록 그 수는 적지만 일반적인 의미해석을 지향하는 단어 기반 시소러스, 예를 들면 Roget's [4]와 WordNet [3] 등이 있다. 전문적 분야의 시소러스는 대부분 ISO 표준[5]에 입각하여 용어들 사이의 관계를 나타내고 있다. 예를 들어, BT(Broad Term), NT(Narrow Term), UF(Used For), 그리고 RT(Related Term)과 같은 시소러스 용어들 사이의 관계를 나타낸다. 이런 시소러스를 이용하는 방법으로 단순히 한 용어에 대한 관계만을 찾는 단순한 방법과 관계에 근접도를 가중치로 하여 두 용어간의 유사도를 측정하는 진보된 방법이 있을 수 있다.

유사도는 검색 분야에서 질의어와 문서 사이의 유사도를 측정하는 연구로부터 기인한다. 유사도를 측정하는 방법은 수학적 이론을 배경으로 하는 벡터 모델 [6]과 인지과학을 배경으로 하는 스프레딩 엑티베이션 기반 방식으로 크게 나눌 수 있다. 벡터 모델은 질의어, 문서 등을 용어들의 가중치 벡터로 표현하여 그들 사이의 유사도를 벡터의 곱으로 (일반적으로 코사인(cosine) 유사도를 사용하여) 나타낸다. 단순 벡터 모델에서는 용어간의 독립성을 가정하여 계산하므로 그 값을 간단히 구할 수 있지만, 용어간의 독립성이라는 가정은 현실성이 없으므로 용어간의 관계를 첨가하는 연구가 지속되고 있다. 해결책으로 일반화된 벡터 (generalized vector space) 모델이 [7, 8] 제안되었으나, 사람들이 생각하는 용어간의 관계와의 차이를 극복하지 못하고 있다. 이를 극복하기 위하여 몇 연구자들은[9, 10] 벡터 모델을 기반으로 용어와 용어의 관계를 행렬로 나타내는 가상 시소러스를 구축하고 이를 신경망을 이용하여 학습시키는 방법을 제안하고 있다.

벡터 모델을 기반으로 하는 방법은 비록 이론적으로 잘 연구되어 있지만, 사용되는 용어의 수가 많아지면, 비록 사용 공간을 줄이는 연구가 되고 있지만, 현실적으로 적용하기 어려운 점이 있다. 한편, 스프레딩 엑티베이션 방법은 인간의 뇌를 모델로 하여 자연스러우며 기존 심볼릭 시소러스하에서 작동 가능하므로 공간의 효율성을 그대로 유지할 수 있다.[1] 그러나, 스프레딩 엑티베이션 모델은 수학적 모델에 근거하지 못하므로 휴리스틱을 필요로 하는 경우가 생긴다.

본 연구에서는 인지과학적 모델인 스프레딩 엑티베이션 기반의 유사도 측정 방법을 기반으로 신경망을 이용하여 학습 가능한 시소러스를 제안한다. 본 연구에서 제안하는 신경망 모델은 기존에 정보 검색을 위하여 제안된 신경망[9, 10, 11]과 다음과 같은 관점에서 차이가 있다. 첫째, 기존의 신경망은 대부분 벡터검색 모델의 효율을 높이기 위한 것인 반면, 본 연구에서 제안하는 신경망은 스프레딩 엑티베이션 기반 시소러스의 학습을 위한 방식이라는 점이다. 둘째, 심볼릭 시소러스를 신경망 시소러스로 변환하고, 학습된 가중치를 심볼릭 시소러스를 위한 가중치로 제변환이 가능하다는 점이다. 세째, 기존 방법에서는 각 연구의 고유 신경망 구조 및 학습 방법을 사용하는데 비하여, 본 연구에서 제안하는 신경망은 이미 그 성능이 인정된 다계층 퍼셉트론 (multi-layered perceptron) 구조와 역전파 (back-propagation) [12] 학습 방법을 사용한다는 점이다.

3. 제안된 신경망 기반 시소러스

본 연구에서 제안하려는 신경망 시소러스는 일반적 표준을 갖는 심볼릭 시소러스로부터 출발하여 구축할 수 있다. 예를 들어 그림 1에서 보는 것과 같이 관계 BT, NT, RT, SYN(Synonym)을 이용하여 개념 Violent-act에 연관된 용어들의 관계를 표현하고 있는 심볼릭 시소러스가[2] 주어졌다고 하자. 주어진 시소러스에서 용어들간의 유사도를 측정하기 위하여, 각 관계에 일정한 초기 가중치를 부여한다. 예를 들어, NT=0.6, BT=0.7, RT=0.8, SYN=0.9. 이를 이용하여 스프레딩 엑티베이션을 다음과 같은 제한 조건을 두어 시행할 수 있다 [13]. 스프레딩 엑티베이션 패스의 길이에 제한을 둔다 (본 연구에서는 5로 하였음). 유사도 값이 어느 특정 값 이하로 오면 멈춘다 (본 연구에서는 0.1로 제한하였음). 유사도의 값은 단순히 패스를 따라 곱하여 계산

1) 기존 심볼릭 시소러스는 계층구조로 정보 저장의 효율성을 꾀할 수 있다.

2) 이 시소러스는 WordNet[3]으로부터 Violent-act에 관련 된 용어들을 수동으로 추출하여 만든 것이다.

하는 방법을 이용하였다. 예를 들어 그림 1에서 weapon으로부터 firearm 까지 유사도는 0.6 (NT) * 0.6 (NT) = 0.36을 계산할 수 있다. 만약 스프레딩 엑티베이션 과정 중에 한 노드에 여러 개의 값이 들어오면 최대치를 갖도록 한다.

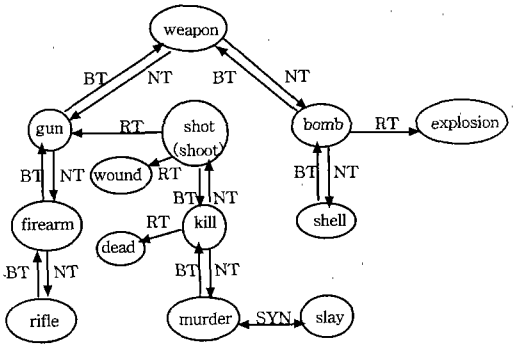

그림 1 개념 Violent-act에 관련된 용어들의 관계의 예.

앞에서 설명한 스프레딩 엑티베이션 방법을 이용하여 용어간의 연관성을 찾아 질의어의 확장 등을 통하여 검색의 정확도를 높힐 수 있겠다. 그러나, 이러한 방법은 그 가중치 할당 방법에 따라 민감하게 결과가 달라질 수 있다. 본 논문에서는 이러한 일반적 스프레딩 엑티베이션의 모형을 신경망 모형, 특별히 다계층 퍼셉트론으로 전환하고 사용자의 검색된 문서의 적합성에 따른 학습을 통하여 가중치를 조정하려 한다. 이러한 변환이 가능하기 위해서는 다음과 같은 두 문제를 해결해야 한다.

(1) 심볼릭 시소러스의 구조를 다계층 퍼셉트론으로 변환하는 문제

(2) 스프레딩 엑티베이션 작업을 어떻게 신경망 구조에서 가능케 할 것인가 하는 문제 앞에서 설명한 스프레딩 엑티베이션 과정에서 최대치 값을 구하는 것은 신경망을 이용해서 가능하나, 단순히 최대치만을 구하는 것이 아니라 스프레딩 엑티베이션에서 사용하는 가중치를 학습시켜야 하므로 간단하지 않다. 우리는 두 번째 문제를 다음에 있는 3.1에서 첫 번째 문제는 3.2과 3.3에서 그 해결 방법을 제시하고자 한다.

3.1 P-Norm 기반 확장 불리언 모델

질의어나 문서를 표현할 때 단순 불리언 모델[14]에서는 AND/OR 접속자와 가중치가 없는 용어를 사용하여 표시한다. 그러나, 가중치를 사용하지 못하므로 그 표현이 떨어져 정확한 검색을 꾀할 수 없다. 이러한 문제를 해결하기 위하여 많은 확장 불리언 모델이 연구되었다 [15, 16, 17, 18]. 그 중에서 본 논문은 p-norm을 이용한 확장 불리언 모델[15]을 이용하여 앞에서 제기한 첫 번째 문제를 해결하고자 한다.

P-norm 확장 불리언 모델에서 제안하는 AND/OR 논리 연결자는 가중치를 갖는 문서와 질의어간의 유사도를 잘 정의하고 있다. 만약 어떤 시스템에서 사용되는 용어들을 T1, T2, ... , Tn라하고, 문서 D를 용어 Ti에 대한 가중치 a_i, $1 \le i \le n$, 로 표시하면 D는 $(a_1, a_2, \ldots, a_n)$으로 나타낼 수 있겠다. 또 OR 질의어와 AND 질의어를 QOR(p) = ORp(q1, q2, ..., qn)와 QAND(p) = ANDp(q1, q2, ..., qn)의 형태로 각각 주어진다고 하자, 단 qi는, $1 \le i \le n$, 질의어 안에 있는 용어 Ti의 가중치이다. 그러면, 문서 D와 이 두 질의어 사이의 유사도는 *p-norm* 확장 불리언 모델에서는 다음과 같이 주어진다.

$$\mathrm{sim}(D, Q_{OR(p)}) = \left[\frac{q_1^p a_1^p + q_2^p a_2^p + \ldots + q_n^p a_n^p}{q_1^p + q_2^p + \ldots + q_n^p} \right]^{1/p} \quad (1)$$

$$\mathrm{sim}(D, Q_{AND(p)}) = 1 - \left[\frac{q_1^p (1 - a_1^p) + q_2^p (1 - a_2^p) + \ldots + q_n^p (1 - a_n^p)}{q_1^p + q_2^p + \ldots + q_n^p} \right]^{1/p} \quad (2)$$

이 모델에서, p = 1이면, AND와 OR의 구분이 없어진다. 다시 말하면, 이 경우는 단순 벡터 모델 [6]에서 유사도를 질의어 벡터와 문서 벡터의 내적(inner product)으로 계산하는 경우와 일치한다. 만약 $p = \infty$ 이고 모든 질의어 용어의 가중치가 1이면, sim(D, QOR(p)) = $\max(a_1, a_2, \ldots, a_n)$와 sim(D, QAND(p)) = $\min(a_1, a_2, \ldots, a_n)$이 된다. 이것은 p의 값을 조정하면 단순 벡터 모델 (p = 0)과 일반적인 불리언 모델(p = ∞)의 중간 형태를 취할 수 있음을 의미한다.[3)]

본 연구에서는 앞에서 설명한 스프레딩 엑티베이션 방법에서 사용한 최대치를 구하는 방법 대신 OR연산자를 이용하여 계산하므로 이 연산을 합의 형태로 표현하여 신경망에서의 연산이 가능하도록 하였다. 이에 대한 자세한 내용은 다음 두 장에서 설명하기로 하겠다.

3.2 신경망 구조

본 장에서는 심볼릭 시소러스를 신경망 구조의 시소러스로 변환하는 방법을 제안하려고 한다. 본 연구에서 제안하는 신경망 구조는 입력층, 히든층, 출력층을 갖는 다계층 퍼셉트론과 같은 구조를 갖는다. 이러한 변환은 시소러스를 이용한 스프레딩 엑티베이션을 두 단계로 나누어 생각하므로 가능하다. 어떤 용어와 용어간의 관

3) 자세한 내용은 참고문헌 [15]를 참고하기 바람.

련도를 찾는 스프레딩 엑티베이션의 과정은 마치 용어를 노드로 용어와 용어간의 관계를 링크로 하는 그래프에서 비순환(acyclic) 패스를 찾는 것과 같다. 다만, 한 노드에 여러 개의 값이 엑티베이션 되면 그 값 중 최대값(본 연구에서는 OR 연산자를 이용하여 얻는 값)을 구한다. 그러나, 우리는 신경망을 이용하여 시소러스에서 주어진 용어간의 가중치를 학습을 통하여 조정하는 목표가 있으므로 이 학습된 가중치를 알 수 있는 구조이어야 한다. 따라서, 본 연구에서는 스프레딩 엑티베이션을 위한 패스를 그림 2에서 보는 것과 같이 두 개의 패스로 나누어 생각한다. 첫 번째는 패스의 첫 노드와 두 번째 사이의 패스 (t1, t2), 패스-1이라고 부르는 패스이고, 둘째는 그 나머지 패스(t2, t3, ..., tn), 패스-2라 부르는 패스이다. 두 번째 패스는 다시 두 번째 패스의 첫 번째 노드와 마지막 노드로 된 단위로 패스(t2, tn)로 줄이고 그 사이의 가중치는 그 패스를 거치면서 계산되는 엑티베이션 값, 즉 그 사이에 있는 가중치의 곱 (w2*w3* ..... *wn-1)을 부여한다.

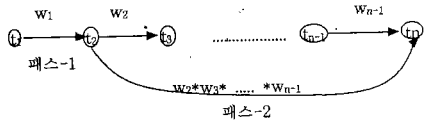

그림 2 두 노드 t1과 tn간의 패스를 2개의 패스로 나누어 생각한다.

그림 2에서 패스-1은 심볼릭 시소러스에 표현된 용어들간의 가중치로 한 개의 값을 갖으나, 패스-2의 가중치 값은 다중 패스가 존재할 경우 여러 개를 갖을 수 있다. 여러 개의 값을 갖을 경우, 그 중에서 최대치를 그 패스의 가중치라고 한다. 물론, 노드들 사이의 패스 길이가 1이면 패스-2는 없고 패스-1만 존재한다.

이렇게 정의한 패스-1과 패스-2를 이용하여 신경망을 다음과 같이 구축한다. 모든 용어와 용어 사이에 대하여, 패스-1은 입력층과 히든층을 연결하는데 사용하고, 패스-2는 히든층과 출력층을 연결하는데 사용한다. 다만, 패스의 길이가 1인 경우, 즉 패스-1만 있는 경우는 입력층에서 히든층으로 연결하지 않고 출력층으로 직접 연결하여 사용한다. 사실, 입력층과 출력층 사이의 연결은 심볼릭 시소러스에 있는 용어와 용어 사이의 관계를 나타낸다. 왜냐하면, 심볼릭 시소러스에서 표현된 용어의 관계는 모두 패스가 1이기 때문이다. 예를 들어, 그림 1에 있는 심볼릭 시소러스를 신경망 구조로 바꾼다고 할 때, 그림 3은 용어 "bomb"에 해당하는 입력층 노드에 출발하여 히든층과 출력층에 연결된 노드들의 모양과 연결 가중치들이다. 그림 3에서 입력층에 있는 용어 "bomb"에 대응되는 노드에서 연결되는 용어 "explosion"에 대응되는 노드가 출력층에는 있는데 히든층에 없는 것에 주의하기 바란다. 이는 "explosion"에 대응되는 노드로부터 나가는 연결이 없으므로 이 노드들을 처음 통과하여 패스의 길이가 2 이상 될 수 없기 때문이다.

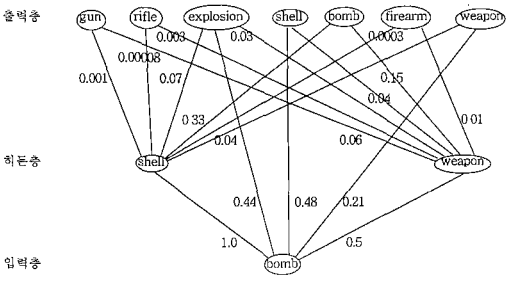

그림 3 그림 1에 있는 심볼릭 시소러스를 신경망 구조 시소러스로 바꿀 때에

용어 "bomb"에 해당하는 입력층 노드로부터 시작되는 신경망 모습이다.

그림 3의 연결 가중치는 다음 방법으로 구한다. 제안한 신경망의 히든층과 출력층 노드의 활성화 값은 그 노드로 입력된 값과 가중치를 곱하여 계산될 것이다. 만약, 활성화 값을 계산하는 노드에 여러 개의 입력이 있으면 이를 가중치와 곱한 것을 합하여 그 노드의 순 입력 값으로 계산하게 된다. 그러나, 스프레딩 엑티베이션 방법에서는 최대치를 이용하여 계산하므로, 이를 합의 형태로 바꾸기 위하여 확장된 방법인 ORp (식 (1) 참조) 연산자를 이용하려고 한다. 식 (1)의 값을 활성화 값으로 생각하고, 활성화 함수를 매개변수에 1/p 승한 것으로 간주하면 순 입력 값 hi는 식 (3)과 같이 주어진다. 여기서, 가중치 w_{ij}와 곱하는 것이 a_j(입력 값 혹은 전 층에서 활성화 된 값)에 p 승한 것에 주의하기 바란다.

$h_i = \sum_j w_{ij} a_j^p$, 단 wij는 노드 j에서 노드 i 사이의 가중치, (3)

a_j는 만약 j가 히든층에 있는 노드이면 노드 j의 활

성화 값, 입력층에 있는 노드이면 입력 값.

이 식에서, 입력으로 들어오는 값들을 p-norm 확장 모델에서 어떤 문서 D의 용어 T1, T2, ... , Tn에 대한 가중치 $a_1, a_2, \ldots, a_n$라고 생각하고, 가중치를 질의어의 용어 T1, T2, ... , Tn에 대한 가중치 q1, q2, ..., qn라고 하면 노드 i에 대한 가중치 wij는 식 (1)로부터 식 (4)와 같이 구해진다.

$$w_{ij} = \frac{q_j^p}{q_1^p + q_2^p + \ldots + q_n^p} \quad \text{for } 1 \le j \le n \qquad (4)$$

예를 들어, 히든층에 용어 "shell"에 해당하는 노드로 오는 입력층의 노드는 용어 "bomb"에 해당하는 것밖에 없으므로 그 가중치는 1이되고, 용어 "weapon"에 해당하는 히든층의 노드로 오는 입력층의 노드는 용어 "gun"과 "bomb"에 해당하는 것인데 둘 다 NT관계(0.6)이므로 이들 신경망에서의 연결 가중치는 0.6p/(0.6p + 0.6p) = 0.5이다.

앞에서 가정한 것처럼 활성화 함수는 순 입력 값 hi에 1/p 승한 것을 사용하는 것이 식 (1)과 일치하지만, 학습과정에 적당하지 못한 요소가 있어 학습과정에 적합한 근사 함수를 사용하려고 한다. 이에 대한 자세한 내용은 다음 장에서 설명하려고 한다.

3.3 활성화 함수 및 학습 방법

활성화 함수 hi에 1/p 승한 것을 사용하면 식 (1)의 값과 일치하는 값을 구할 수 있지만, 이 함수의 일차 미분 값이 0 근처에서 극도로 커지므로 역전파 학습 방법(back-propagation)[12]을 이용하는 과정에서 수렴하지 못하는 결과를 자주 초래한다. 이러한 문제를 해결하는 방법의 하나로 본 논문에서는 식 (5)와 같이 2p-sigmoid 함수를 활성화 함수로 사용하려고 한다. 본 논문에서는 $\theta = 0.5$을 사용하였다. 실제로 이 함수는 원래 함수와 많은 hi값에 대하여 매우 근사한 값을 나타낸다.

$$a_i = F(h_i) = \frac{1}{1 + e^{-2p(h_i - \theta)}} \qquad (5)$$

학습 방법으로는 역전파 학습 방법을 사용한다. 오차의 정도 E는 희망하는 출력치(d_i)와 실제로 계산된 출력치(a_i)의 차이를 제곱한 것을 합한 형태의 함수로 다음과 같이 표현한다.

$$E = \frac{1}{2}\sum (d_i - a_i)^2$$

그러면, 신경망 연결 가중치의 변화량은 다음과 같이 표현된다. 이때, γ는 학습율이며 0 보다 크다.

$$\triangle w_{ij} = -\gamma \frac{dE}{dw_{ij}}$$

이를 다시 쓰면 다음과 같이 변형할 수 있다.

$$\frac{dE}{dw_{ij}} = \frac{dE}{dh_i}\frac{dh_i}{dw_{ij}}$$

식 (3)로부터 다음 값을 구할 수 있다.

$$\frac{dh_i}{dw_{ij}} = a_j^p$$

이 결과를 가중치 변화량에 대입하면 그 결과는 다음과 같다.

$$\triangle w_{ij} = \gamma \delta_i a_j^p, \quad \text{단} \quad \delta_i = -\frac{dE}{dh_i}.$$

만약 노드 i가 출력층에 있는 노드이면, δ_i 값은 다음과 같이 구할 수 있다.

$$\delta_i = (d_i - a_i)F'(h_i)$$

이때 F의 미분 함수는 아래와 같이 주어진다.

$$F'(h_i) = \frac{2p}{(1 + e^{-2p(h_i - \theta)})^2} e^{-2p(h_i - \theta)} = 2pa_i(1 - a_i)$$

만약 노드 i가 히든층에 있는 노드이면, δ_i는 노드 j와 연결된 노드들의 δ값을 이용하여 다음과 같이 구할 수 있다.

$\delta_i = F'(h_i)\sum_k w_{ki}$, 단 k는 노드 i 위에 있는 층의 모든 노드에 대하여 변한다.

4. 실 험

본 연구에서 제안한 신경망 시소러스가 사용자의 검색 문서에 대한 적합성 판단에 의하여 잘 학습되는지를 알기 위하여 다음과 같은 실험을 수행하였다. 그림 1에서 언급한 개념 Violent-act에 관련된 문서를 찾기 위하여 그림 4와 같은 AND/OR 트리로 표현된 질의어를 생각하였다. 이 그림에서 링크 사이에 수평선이 있으면 AND 연산자이고 없으면 OR 연산자를 의미한다. 그림 4의 내용은 표 1에 있는 12개의 질의어를 OR한 것과 같다[4]. 이러한 질의어를 인터넷 검색 엔진 Google [20]을 이용하여 검색된 문서 중 순위 20위까지의 문서를 1차로 모으고, 그 중 중복된 것 등을 제외하여 960 개의 문서를 모았다. 이 중 절반을 임의로 뽑아, 훈련 문서 집합으로 사용하고 나머지를 평가 실험을 위하여 사용하였다. 모아진 문서가 개념 Violent-act에 적합한지에 대한 판단은 일반 대학원 학생이 문서를 보고 판단하여 정하였다.

앞에서 선정한 480개의 훈련 문서에 대한 사용자의

4) 이와 같은 질의어는 다른 참고 문헌[19]에서도 사용하고 있음.

적합 판정에 따라 시소러스 신경망을 학습시키기 위하여 각 문서의 색인어로부터 시소러스 신경망을 통하여 확장된 용어들의 가중치를 질의어에서 사용되는 용어의 활성화 값으로 사용하려고 한다. 이를 위하여 P-norm 확장 불리언의 AND/OR 연산자를 사용하면 질의어를 표현하는 그림 4의 AND/OR 트리를 그림 5의 신경망으로 변환할 수 있다. 이에 대한 자세한 내용은 참고문헌 [19]를 참조하기 바란다. 이렇게 변환된 두 신경망을 연결하기 위하여, 시소러스 신경망의 출력층은 질의어를 위한 입력층이 되도록 한다. 이렇게 만들어진 복합 신경망 입력층의 노드에 대응되는 용어가 각 문서에 있는 색인어로 존재하면 그 입력 값이 1이 되고 아니면 0이 색 상태 값(Retrieval Status Value)이 된다. 따라서 된다. 복합 신경망 출력층은 노드 하나로 그 활성화 값은 문서와 질의어(개념 Violent-act)와의 유사도, 즉 검그 문서가 적합하면(Relevant) 희망 값(Desired Value)를 0.7 이상으로, 부적합하면(Nonrelevant) 0.4 이하가 되도록 학습을 진행시켰다. 학습은 오차가 줄어지면 계속하였는데 보통 20 에포크(epoch)를 넘지 않았다.

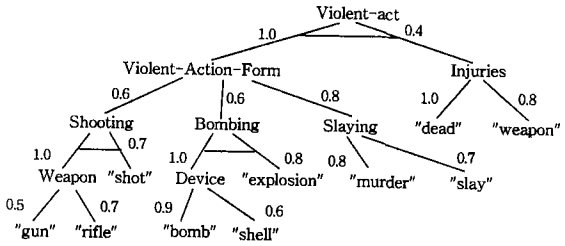

그림 4 개념 Violent-act를 위한 AND/OR 트리

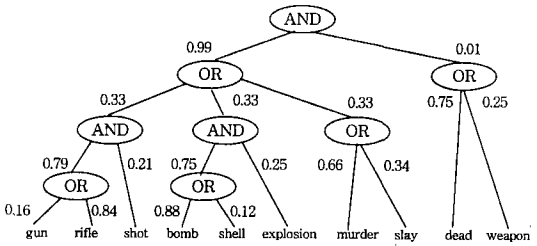

그림 5 그림 4에 있는 AND/OR 트리로부터 변환된 신경망

표 2는 p=5, 4 에포크 후에 얻은 결과이다. 본 논문에서는 시소러스를 사용하는 것과 사용하지 않는 것과의 차이, 또한 학습 받은 시소러스를 사용한 것과 아닌

표 1 실험을 위한 질의어

질의어 번호	질 의 어
1	gun(0.3) AND shot(0.42) AND dead(0.4)
2	gun(0.3) AND shot(0.42) AND wound(0.32)
3	rifle(0.42) AND shot(0.42) AND dead(0.4)
4	rifle(0.42) AND shot(0.42) AND wound(0.32)
5	bomb(0.54) AND explosion(0.48) AND dead(0.4)
6	bomb(0.54) AND explosion(0.48) AND wound(0.32)
7	shell(0.36) AND explosion(0.48) AND dead(0.4)
8	shell(0.36) AND explosion(0.48) AND wound(0.32)
9	murder(0.64) AND dead(0.4)
10	murder(0.64) AND wound(0.32)
11	slay(0.56) AND dead(0.4)
12	slay(0.56) AND wound(0.32)

표 2 신경망 시소러스 모델에서 학습 전과 후 향상된 recall-precision 비율(가중치 없는 시소러스란 모든 관계의 가중치를 1로 하는 일반적인 시소러스를 의미한다.)

Recall	Precision			
	시소러스를 사용하지 않은 경우	가중치 없는 시소러스를 사용한 경우	가중치 있는 시소러스를 사용한 경우	
			학습 전	학습 후
0.1	0.5000	0.5000	0.5000	1.0000
0.2	0.6667	0.5882	0.6250	0.9091
0.3	0.6818	0.6522	0.6522	0.6818
0.4	0.6667	0.6897	0.6667	0.6897
0.5	0.7143	0.7143	0.6944	0.7353
0.6	0.7317	0.7143	0.7317	0.7692
0.7	0.6538	0.6538	0.6415	0.7234
0.8	0.6333	0.6333	0.6441	0.6552
0.9	0.5250	0.5250	0.5250	0.6176
1.0	0.4792	0.4792	0.4792	0.4792
시소러스를 사용하지 않을 경우 보다 향상된 평균 비율		-1.5%	-1.3%	18.8%

것을 비교할 수 있도록 실험하였다. 실험을 위해서 남겨둔 480개의 평가용 문서에 대해 학습된 시소러스와 표 1에 있는 질의어를 이용하여 앞에서 설명한 방법으로 검색 상태 값을 구하여 recall-precision 값을 구하였다. 표 2에서 보는 것과 같이, 시소러스의 가중치가 적절치 못하면 시소러스를 사용하는 것이 심지어 recall-precision을 나쁘게 할 수도 있다. 그러나, 학습 후 그 비율은 18.8%나 향상됨을 보였다.

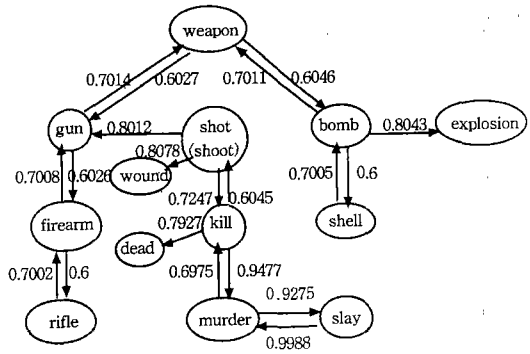

그림 6 학습 후 변경된 심볼릭 시소러스의 모습

마지막 단계로, 초기의 심볼릭 시소러스에 부여한 가중치를 학습 후에 얻은 신경망 가중치로부터 계산하려 한다. 이를 위하여 신경망 가중치 w_{ij}가 식 (4)에 있는 것처럼 주어졌고, 심볼릭 시소러스의 가중치가 q_i라고 하자. 또한, 학습 후 신경망 가중치를 $w_{ij}^{'}$, 심볼릭 시소러스 가중치를 r_i할 때 다음과 같은 가정을 하여 γ_i를 구하려 한다.[5]

$q_1^p + q_2^p + \ldots + q_n^p = r_1^p + r_2^p + \ldots + r_n^p = C$, 단 C는 상수.

이런 가정하에 w_{ij}와 는 다음과 같이 표현된다.

$$w_{ij} = \frac{q_j^p}{C}, \qquad w_{ij}^{p} = \frac{r_j^p}{C}$$

그러므로, 학습 후 변경된 심볼릭 시소러스의 가중치는 다음과 같이 구할 수 있다.

$$r_j^p = q_j^p \frac{w_{ij}^{'}}{w_{ij}}$$

그러나, 입력층에 있는 노드 중에서 히든층과 출력층에 같은 용어에 대응되는 노드와 동시에 연결되는 경우가 있다. 예를 들어, 그림 3에서 "shell"과 "weapon"에 대응되는 노드는 히든층과 출력층에 모두 있으며 용어 "bomb"에 대응되는 입력층 노드와 연결되어 있다. 용어와 용어 사이의 근접도는 하나인데 이용하는 방법을 패스-1과 패스-2의 방법으로 나누어 사용하고 학습도 각각 시키므로 조절된 가중치는 다르게 나타날 수 있다. 이런 경우 본 연구에서는 조정된 가중치로 두 가중치의 평균을 취하였다. 그림 6은 학습 후 달라진 용어간의 가중치를 보여 주고 있다.

5) 사실, 학습 후 심볼릭 시소러스의 가중치를 수학적으로 정확하게 구 할 수 있는 방법은 없다.

5. 결 론

본 연구에서는 사용자의 문서에 대한 적합성을 이용하여 시소러스의 가중치를 효과적으로 조절할 수 있는 신경망 기반 시소러스를 제안하였다. 제안된 시소러스는 다계층 퍼셉트론과 역전파 학습 방법을 이용하여 효율적으로 검색의 효율을 높힐 수 있음을 제한된 검색 문서를 이용하여 입증하였다. 추후 연구로 보다 일반적인 검증을 위하여 더 많은 문서를 기반으로 한 실험이 필요하며, 이를 위한 한 방법으로 WordNet으로부터 표준화된 형태의 시소러스를 자동으로 추출하는 것도 추후 연구과제라 할 수 있겠다. 또한, 제안된 모델의 활성화 함수 보다 적절한 함수를 찾는 것과, 보다 직접적인 실험 방식(질의어의 가중치에 영향을 받지 않는 실험 방식)에 대한 연구도 필요하겠다.

참 고 문 헌

[1] Chen, H., Schatz, B., Yim, T., and Fye, D. "Automatic Thesaurus Generation for an Electronic Community System," Journal of the American Society for Information Science, 1994.

[2] Collins, A. M. and Loftus, E. F. "A Spreading-Activation Theory of Semantic Processing," Psychological Review, Vol. 82, pp. 407-425, 1975.

[3] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. "Introduction to WordNet: An On-line Lexical Database," Princeton University, 1993.

[4] Editors of The American Heritage Dictionary, "*Roget's II : the new thesaurus*," *Boston* : Houghton Mifflin Co., 1995.

[5] ISO 2788-1986(E) : Documentation Guidelines for the Establishment and Development of Monolingual Thesauri. 2nd.

[6] Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.

[7] Wong, S.K.M., Ziarko, W., Raghavan,V., and Wong, P. C. N. On Modeling of Information Retrieval Concepts in Vector Spaces, ACM Transaction on Database System, Vol. 12, No. 2, June 1987. pp. 299-321.

[8] Wong, S.K.M., Ziarko, W., and Wong, P. C. N. Generalized Vector Space Model in Information Retrieval. In Proceedings of the 8th Annual International *ACM-SIGIR* Conference, 1985, New York, pp. 18-25.

[9] Gwang S. Jung, "Connectionist Domain Knowledge

Acquisition And Its Evaluation In Information Retrieval," PhD thesis, The Center for Advanced Computer studies, University of Louisiana at Lafayette, 1991.
[10] Wong, S.K.M. and Cai, Y.J. Computation of term associations by a neural network. In Proceedings of ACM-SIGIR Conference, 1993, pp. 107-115.
[11] *Wilkinson, R. and Hingston, P. Using the cosine* measure in a neural network for document retrieval. In Proceedings of ACM-SIGIR Conference, 1991, pp. 202-210.
[12] Lippmann, R. P. An introduction to computing with neural nets. *IEEE ASSP Magazine*, Vol. 3, No. 4, pp. 4-22.
[13] Salton, G. and Buckley, C. On the use of spreading activation methods in automatic information retrieval. In 11th Annual International ACM SIGIR Conference on Research and Development in Retrieval (1988), pp. 147-160.10.
[14] Lancaster, F. W. Information retrieval systems: *characteristics, testing and evaluation, 2nd Ed.*, John Wiely and Sons, New York, 1979.
[15] Salton, G., Fox, E. A., and Wu, H. Extnded Boolean Information Retrieval, Vol. 36, No. 11, December 1983, Communication of the ACM, pp. 1022-1036.
[16] Bookstein, A. Fussy requests: An approach to weighted Boolean searches, *J. ASIS*, Vol 31, No. 4, July, 1980, pp. 275-279
[17] Waller, W. G. and Kraft, D. H. A mathematical model for a weighted Boolean retrieval system. Information Processing and Management, Vol 15, No. 5, 1979, pp. 235-245.
[18] *Wong, S.K.M., Ziarko, W., Raghavan,V., and* Wong, P. C. N. Extended Boolean Query Processing in the Generalized Vector Space Model, *Information Systems* Vol. 14, No. 1, pp. 47-63, 1989.
[19] Kim, M., and Raghavan, V. V. Adaptive concept-based retrieval using a neural network. In Proceedings of the ACM SIGIR 2000 Workshop on Mathematical/Formal Methods in Information Retrieval, Athens, Greece, July 28, 2000.
[20] Google Search Engine, http://www.google.com.

최 종 필
1994년 아주대학교 컴퓨터공학과(학사). 1999년 아주대학교 컴퓨터공학과(석사). 1999년 ~ 현재 아주대학교 컴퓨터공학과(박사과정). 관심분야는 지능형 정보검색 시스템, 신경망, 기계학습

최 명 복
1992년 호서대학교 전자계산학과(학사). 1994년 아주대학교 컴퓨터공학과(석사). 1994년 ~ 현재 아주대학교 컴퓨터공학과(박사과정). 1997년 ~ 현재 원주대학 행정전산과(조교수). 관심분야는 지능형 정보검색, 퍼지 응용, 지식표현, 의사결정 시스템

김 민 구
1977년 서울대학교 계산통계학과(이학사). 1979년 한국과학기술원 전산학과(공학석사). 1989년 Pennsylvania 주립대(박사). 1999년 ~ 2000년 Louisiana 대학 연구과학자. 1981년 ~ 현재 아주대학교 컴퓨터공학과(교수). 관심분야는 지능형 정보검색 시스템, 지능형 교수 시스템, 지능형 캐릭터 에이전트