유전자 알고리즘을 이용한 통합의학언어시스템(UMLS)의 의미망 자동 군집 방법

지영신[O]* 김태준* 전혜경* 정헌만* 이정현**
인하대학교 전자계산공학과* 인하대학교 컴퓨터공학부**
{ys_ji[O], hope673, jhk0214, hmjung}@nlsun.inha.ac.kr* jhlee@inha.ac.kr**

Semantic Network Automatic Clustering Method of the Unified Medical Language System Using Genetic Algorithm

Youngshin Ji[O]* Taejun Kim* Hyekyoung Jeon* Heonman Jeong* Junghyun Lee**
{Dept.*, School**} of Computer Science & Engineering, INHA University

요 약

UMLS 의미망은 크기가 방대하고 복잡하여 사용자가 이해하기가 어렵고 화면상에 모든 의미망을 모두 표현할 수 없다는 단점을 가지고 있다. 이 문제를 해결하기 위해 의미망을 효율적으로 분할하기 위한 규칙들이 소개되고 있지만 이것은 UMLS 의미망이 수정될 때마다 규칙을 적용하여 수작업으로 분류를 해야한다는 단점이 있다. 이 문제점을 해결하기 위해 유전자 알고리즘을 이용한 UMLS 의미망의 자동 군집화 방법을 제안한다. 제안한 방법은 각각의 의미유형 간의 연결된 의미관계를 사용하여 의미망을 구조적으로 유사한 의미유형 집합들로 군집화하고 규칙에 의한 군집 방법의 결과 비교 평가한다.

1. 서 론

UMLS(Unified Medical Language System)는 서로 다른 의학전문용어 사전들간의 차이점에 의해서 야기되는 문제를 극복할 수 있도록 해주지만, 약 190만개의 개념(concept)명과 73만개의 개념이 포함된 메타시소러스의 방대한 크기와 복잡성은 사용자의 이해를 어렵게 할수도 있다. 그래서 UMLS는 사용자의 이해를 돕기위해 방대한 지식을 다이어그램의 형태로 제공하는 의미망을 제공하고 있지만 이것 또한 크기가 크고 이해하기 어렵다[1]. 또한, 서로 복잡하게 연결된 의미유형들을 하나의 컴퓨터 화면에 표현하는 것이 불가능하다는 문제점이 대두되고 있다. 최근 이런 문제점을 해결하기 위해 의미망을 구조적으로 유사한 군집들로 분할하기 위한 여러가지 규칙들이 소개되고 있다.

본 논문에서는 의미유형과 군집들 사이의 구조적 유사성과 결합적합도의 계산 방법을 정의하였다. 또한, 유전자 알고리즘의 전역적 탐색 기법을 이용하여 의미망을 구조적으로 자동분할하는 방법을 소개하고 기존의 규칙에 의한 분류결과와 비교 평가한다.

2. Unified Medial Language System의 의미망

2.1 정의 및 범위

UMLS는 미국 NLM(National Library of Medicine)에서 장기과제로 개발하고 있는 통합용어시스템이다[2]. 향후 의료분야에 컴퓨터를 이용한 지능적인 자료처리를 위해서는 컴퓨터가 이해할 수 있는 형태의 의학용어모델을 활용하는 핵심기술에 대한 연구가 필요하다. 이에 맞추어 광범위하게 생명과학분야의 전문 검색 시스템을 개발하고 지원하기 위한 연구들이 세계각국에서 진행되고 있다. 그 중 미국국립의학도서관이 운영하는 UMLS가 가장 활발히 진행되고 있는 연구분야 중 하나이다.

의미망은 UMLS의 일부분으로 지식요소의 하나이며, 다른 지식요소들로는 메타시소러스, SPECIALIST 어휘사전, 그리고 정보요소 지도가 있다. 의미망은 메타시소러스에서 개념에 해당하는 의미유형에 대한 일종의 전거로서, 계층관계로 표현되는 정보의 형태를 말한다.

의미망의 목적은 UMLS 메타시소러스에서 표현된 모든 개념들에 대한 일관성 있는 범주화와 유용한 관계집합을 제공하는데 있다. 즉, 의미망은 기본적인 의미형태나 개념에 할당될 수 있는 범주, 그리고 의미유형 사이의 관계집합을 정의하는데 필요한 정보를 제공한다. 2003AA edition의 의미망은 135개의 의미유형과 55개의 의미관계가 정의되었다.

2.2 구조 및 관계정의

메타시소러스는 정보자료 어휘들에서 추출하여 만든 통제된 용어로 구성된다. 각각의 용어의 의미는 정의나 주석에 의해, 또는 계층관계에서 상하위의 관계등을 통해 명확하게 정의된다. 각각의 메타시소러스 용어의 개념은 적어도 하나 이상의 의미유형(semantic type)이 할당된다. 즉, 계층구조로 이루어진 의미망의 노드인 특수한 의미유형이 개념으로 할당되는 것이다.

의미망은 *Entity*와 *Event*의 두개의 의미유형을 루트로 갖으며, 각각의 의미유형은 isa관계로 계층을 결정짓고 계층관계 이외에도 54개의 의미관계(semantic relationship)로 의미유형 사이의 관계를 정의한다. 의미유형 사이의 isa 관계는 의미관계들의 상속을 의미한다. 즉, 자식 의미유형은 부모 의미유형의 모든 의미관계 정의를 상속받으며, 상속받는 의미관계 이외에도 의미유형들은 새로운 의미관계를 정의하기도 한다. UMLS는 의미관계의 상속에 영향을 미치는 두개의 특별한 모델링 특징을 제공한다. 첫번째는 정의는 되지만 계승되지 않는 관계로 이를 DNI라고하며, 두번째는 계승된 관계를 무효로 하는 blocking을 말한다.

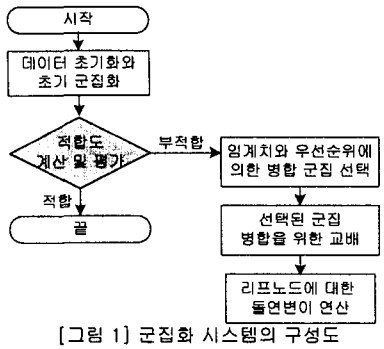
3. 의미망의 자동 분할 방법

3.1 시스템의 구조

본 논문에서는 의미유형들의 초기 군집을 구성하고 적합도 평가에 의해 각 군집이 상위 군집과 통합될 수 있는지를 평가하여 선택된 군집들을 통합한다.

시스템의 구성은 초기군집으로부터 선택, 교배, 돌연변이 등의 연산들을 사용하는 진화 과정을 통해 다음 세대의 군집을 생성하는 유전자 알고리즘과 유사한 흐름으로 진행되며, 시스템 구성도는 그림 1과 같다.

- 1단계. 데이터 초기화와 초기 군집화 : 각 의미유형의 의미망 구조와 정보들을 데이터베이스화하고, is-a관계에 있는 의미관계가 동일한 의미유형들을 하나로 군집화한다
- 2단계. 적합도 평가 : 적합도 함수에 의해 군집들 사이의 결합적 합성을 판단한다.
- 3단계. 선택 : 적합도 평가의 결과에 의해 유사한 군집들을 선택한다.
- 4단계. 교배 : 선택한 군집들을 군집화한다.
- 5단계. 돌연변이 : 리프노드는 DNI가 선언되어도 하위 계층에 영향을 주지 않으므로 데이터 조작에 의해 별도의 군집으로 분류되지 않도록한다.

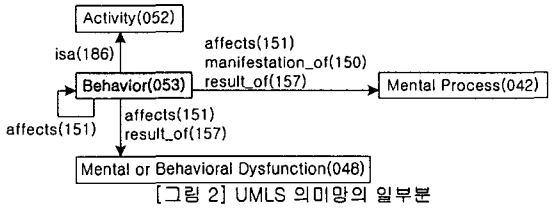


[그림 1] 군집화 시스템의 구성도

3.2 데이터 초기화와 초기 클러스터링

본 논문에서는 의미유형과 의미관계에 십진수 세자리 숫자의 고유번호를 할당하여 이를 데이터베이스화하고 각 의미유형은 다른 의미관계의 종류와 개수를 가지므로 각각의 개체는 가변길이를 갖도록한다.

그림 2는 의미망의 일부를 보여준다. Activity, Behavior, Mental Process, Mental or Behavioral Dysfunction이 의미유형이고, affects, manifestation_of, reuslt_of가 의미관계로 화살표를 이용하여 방향을 표현하고 있다. 이들 각각에 052, 053, 042, 048, 151, 150, 157와 같은 세자리 숫자의 고유번호를 부여한다.



[그림 2] UMLS 의미망의 일부분

부여된 고유번호로 각 의미유형의 정보와 의미관계들을 표현한다. 예를들면, result_of 관계로 의미유형 Mental Process에 연결되어 있는 의미관계는 151042로 각각의 고유번호를 나열하여 표현한다. 또한 해당 의미유형의 의미망에서의 위치를 나타내는 TID, 하위 의미유형의 개수, 상위 의미유형 그리고, DNI 관계를 포함하는지의 여부를 포함한다. 그림 2와 같은 Behavior 의미유형은 {053|B1.1|2|186052|N|150048|151042|151048|151053|157042|157048}로 표현한다.

데이터베이스 구성 후, isa 관계에 있는 의미유형들 중 동일한 의미관계를 가지는 의미유형들을 하나로 군집화하여 초기 군집을 구성한다.

3.3 적합도 함수

본 논문에서는 유사한 두개의 군집을 발견하여 병합하는 과정을 반복적으로 수행한다. 이 때 우수한 두 개의 군집을 찾아내기 위해 식(1)과 같은 UMLS 의미망에 적합한 적합도 측정방법을 고안했다.

적합도는 군집 i와 군집 j의 유사도와 결합적합도에 의해 계산된다.

$$Fit_{ij} = \frac{UFit_{ij}}{Sim_{ij}} \qquad \text{식(1)}$$

$UFit_{ij}$: 군집 i와 군집 j의 결합적합도

Sim_{ij} : 군집 i와 군집 j의 유사도

3.3.1 결합적합도

여러 의미유형들이 하나의 군집을 이루면 그 군집에서의 루트노드가 해당 군집을 대표하는 의미유형이 되므로, 결합적합도를 계산하여 두 군집이 병합되었을 때 병합된 군집의 루트노드가 해당 군집을 얼마나 잘 대표할 수 있는지를 평가한다. 두 군집의 결합적합도 계산 방법은 식(2)와 같다.

$$UFIT_{ij} = \frac{UCoij}{\frac{ECoi + ECoj}{2}}$$

$$UCo_{ij} = \frac{ECo_i \times n_i + \sum_{a=1}^{n_j} \frac{M(R_i, I_a)}{\max(|R_i|,|I_a|) - M(R_i, I_a) + 1}}{n_i + n_j} \qquad \text{식(2)}$$

$$ECo_i = \frac{\sum_{a=1}^{n_i} \frac{M(R_i, I_a)}{\max(|R_i|,|I_a|) - M(R_i, I_a) + 1}}{n_i}$$

n_i : 군집 i에 속하는 의미유형의 개수

R_i : 군집 i의 루트 의미유형

$M(R_i, I_a)$: 의미유형 R_i와 의미유형 I_a사이에 일치하는 의미관계의 개수

$\max(|R_i|,|I_a|)$: R_i와 I_a중 의미관계 개수 중 큰 숫자

군집 i는 군집 j의 루트 의미유형의 부모 의미유형이 속한 군집으로 한다. 즉, 부모군집과 자식군집의 적합도를 계산하여 병합여부를 결정하는 것이다.

ECo_i는 군집 i의 루트와 그 외 의미유형들과의 의미관계를 비교하여 일치하는 의미관계의 개수를 비일치하는 의미관계의 개수로 정규화한 것의 평균이다. UCo_{ij}는 군집 i와 군집 j가 병합했을 때의 루트가 군집을 얼마나 잘 대표하는지를 나타내는 것을 말한다.

각 군집의 결합적합도를 나타내는 ECo_i와 ECo_j의 평균에 비해 두 군집을 병합했을때의 결합적합도를 나타내는 $UFit_{ij}$가 크다면 두 군집의 결합적합도는 높은 것이다.

3.3.2 군집간의 유사도

본 논문에서는 군집내의 모든 의미유형들이 공통적으로 가지고 있는 공통 의미관계와 그 외의 비공통 의미관계들을 분류하여 유사도를 계산함으로써 계산량을 줄이고자 한다.

먼저, 두 군집에서 공통 의미관계를 추출하여 그림 3과 같은 비교 테이블을 구성한다. 2행은 군집 i의 의미유형이 정의하고 있는 것이면 1, 아니면 0이라고 표기하고, 3행은 같은 방법으로 군집 j에 대하여 나타내고 있다

군집 i의 공통 의미관계 : 150048 151042 151048 151053
군집 j의 공통 의미관계 : 150048 151042 157042 157048

의미관계	150048	151042	151048	151053	157042	157048
군집i	1	1	1	1	0	0
군집j	1	1	0	0	1	1

(111100) XOR (110011) = 001111

[그림 3] 공통 의미관계 유사도 계산

이 테이블의 2행과 3행을 행단위로 추출하여 XOR연산을 수행하면 두 군집의 모든 의미유형이 공통적으로 가지고 있는 의미관계의 개수와 그렇지 않은 것의 개수를 구할 수 있다. XOR연산의 결과를 사용하여 식(3)과 같이 유사도를 계산한다.

$$SimA = \frac{(n_i^2 \times D_n) + (n_j^2 \times D_n)}{2} \qquad \text{식(3)}$$

: 군집 i의 의미유형 개수

D_n : XOR 연산 결과 1의 개수

군집내의 비공통 의미관계를 사용한 유사도 계산을 하기위해 그림 4와 같은 별도의 테이블을 구성한다.

군집 i의 비공통 의미관계 : 150048 151042 151048 151053157042
군집 j의 비공통 의미관계 : 150048 151042 157042 157048

의미관계	150048	151042	151048	151053	157042	157048
군집 i	3	2	1	4	6	0
군집 j	4	1	0	0	3	1

[그림 4] 비공통 의미관계 유사도 계산

2행은 군집 i에서 나타난 해당 의미관계의 빈도수이고, 3행은 군집 j에서 나타난 의미관계의 빈도수이다. 이 테이블을 이용한 유사도 계산법은 식(4)와 같다.

$$SimB = \frac{\sum_{a=1}^{row}(w_i \times T_{2a} - T_{3a})^2 + \sum_{a=1}^{row}(T_{2a} - w_j \times T_{3a})^2}{2}$$

식(4)

$w_i = \frac{|n_j|}{|n_i|}$ $w_j = \frac{|n_i|}{|n_j|}$

n_i : 군집 i 의 의미유형의 개수

row: 행의 수

T_{2a} : 테이블의 2행 a열의 빈도수

위의 식(3)과 식(4)의 결과로 식(5)과 같이 두 군집간의 유사도를 계산한다.

$$Sim = \sqrt{SimA + SimB}$$ 식(5)

3.4 선택, 교배와 돌연변이

선택단계에서는 우수한 성질의 염색체에 보다 많은 선택의 기회를 주는 방법인 룻렛휠(roulette wheel) 방법을 사용하여 적합도 값이 크고 의미관계의 개수가 적은 군집에 상대적으로 많은 병합 기회를 부여한다.

교배단계에서 군집 j의 의미관계를 군집 i에 통합하고 군집 j를 삭제하는 과정을 수행한다. 먼저 $SimA$ 계산 과정에서 만들어진 테이블을 이용하여 군집 i와 군집 j에 모두 있는 공통 의미관계를 군집 i의 새로운 공통 의미관계로 생성한다. 군집 i의 의미유형의 비공통 의미관계에 기존 군집 i의 공통 의미관계에서 탈락한 의미관계들을 추가하고 군집 j의 의미유형의 비공통 의미관계에 기존 군집 j의 공통 의미관계에서 탈락한 의미관계들을 추가한다. 교배단계의 마지막으로 군집 i의 의미유형에 추가하고 군집 j를 삭제한다.

돌연변이단계에서는 다른 의미유형들과는 별도로 처리되어야 하는 의미유형들을 처리하는 연산이다. 예를 들어, 리프노드인 의미유형이 DNI 선언이 되었을 경우 리프노드는 상속될 노드가 없으므로 별도로 새 군집을 생성하지 않아도 된다. 그러므로 DNI선언 정의를 삭제하여 군집의 효율성을 높인다.

4. 실험 및 평가

본 논문에서는 [3]에서 제공하는 2003AA edition의 의미망 데이터를 사용하여 실험하였다. 초기 군집 단계는 isa 관계를 제외한 다른 의미관계들이 모두 같은 의미유형들을 하나의 군집으로 통합하는 과정으로 82개의 군집이 형성되었다. 각 군집에 포함된 의미유형의 개수는 표 1와 같다. 첫번째 행은 군집내 의미유형의 개수이고, 두번째 행은 해당 군집의 개수이다.

[표 1] 초기 군집의 결과

의미유형 개수	1	2	3	4	6	8	14
군집의 개수	62	9	5	3	1	1	1

각 군집과 그 상위 그룹을 병합후보들이라고 한다. 각 병합후보들에 대해 결합적합도, 유사도, 적합도를 구하여 모든 병합후보들의 유사도가 임계값 32를 초과하면 분류과정을 마친다.

선택단계에서는 유사도가 임계값 12을 초과하지 않는 병합후보들 사이에 서로 연관관계가 있는 후보들은 우선순위에 의한 경쟁을 통해 병합할 군집을 선택한다. 10번의 적합도 판정 과정을 거쳐 최종 요약된 의미망을 구성하였으며, 각 반복때마다 선택된 병합후보의 개수는 각각 17, 13, 13, 9, 5, 3, 1, 1, 0개이다.

교배 연산은 선택 단계에서 선택된 후보군집을 대상으로 수행되었으며 앞에서 언급한 대로 하위 군집을 상위 군집에 합병하고 하위 군집을 삭제하였다.

돌연변이 연산은 10개의 의미유형에서 발생했고, DNI선언을 별도 처리한 결과 모두 상위 그룹과 통합되었다.

최종 분할은 23개의 군집으로 분할되었으며 결과는 표 2와 같다. 첫번째 행은 군집내 의미유형의 개수이고, 두번째 행은 해당 군집의 개수이다.

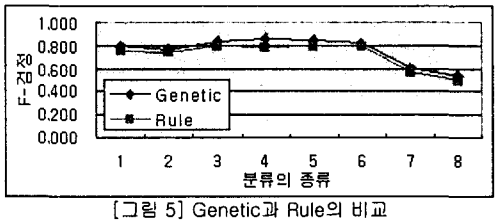
[표 2] 23개의 군집

의미유형 개수	1	2	3	4	5	6	7	8	9	15	16	21
군집의 개수	4	3	3	2	2	1	3	1	1	1	1	1

본 논문의 평가는 전문가들과 비전문가들이 규칙에 의해 수작업으로 분류한 결과와 비교하여 재현율, 정확율, F-검정을 계산하여 평가하였다. 본 논문에서의 결과를 Genetic이라하고, [4]의 결과를 Rule이라하며, 비교결과는 표 3과 그림 5에서 보이고 있다.

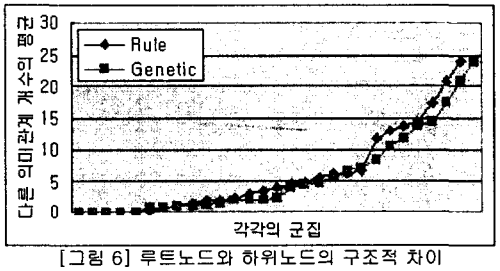
[표 3] 평가결과 (재현율/정확율/F-검정)

비교대상	Genetic	Rule
	재현율/정확율/F-검정	재현율/정확율/F-검정
1	0.966 / 0.683 / 0.800	0.928 / 0.634 / 0.753
2	0.897 / 0.667 / 0.765	0.892 / 0.641 / 0.745
3	0.931 / 0.771 / 0.844	0.892 / 0.714 / 0.793
4	0.897 / 0.839 / 0.867	0.821 / 0.741 / 0.778
5	0.862 / 0.833 / 0.847	0.821 / 0.766 / 0.792
6	0.724 / 0.955 / 0.824	0.714 / 0.909 / 0.799
7	0.448 / 0.929 / 0.605	0.428 / 0.857 / 0.570
8	0.379 / 0.917 / 0.537	0.357 / 0.833 / 0.499



[그림 5] Genetic과 Rule의 비교

의미망의 분할은 군집을 대표하는 루트와 하위노드와의 구조적 유사성이 중요하므로 각 군집의 루트와 하위노드와의 의미관계를 비교하여 일치하지 않는 의미관계 개수의 평균을 구하여 비교하였으며, 그 비교 결과는 그림 6과 같다.



[그림 6] 루트노드와 하위노드의 구조적 차이

루트와 하위노드와의 일치하지 않는 의미관계 개수 평균의 총합은 Genetic은 166.967, Rule은 167.339이고 군집의 개수로 평균화한 값은 각각 5.757, 5.967로 Genetic 방식이 군집의 루트와 하위노드간의 유사성이 더 높다.

위와 같이 제안한 방법은 [4]보다 높은 평가결과를 보였다. 유전자 알고리즘을 이용한 의미망의 분할은 임계값에 따라 군집의 개수가 결정된다. 효율적인 분할을 위해 좋은 결과를 기대할 수 있는 임계값을 결정하는 방법을 고안하여 보완해야 할 것으로 보인다.

5. 결론

본 논문에서는 UMLS 의미망을 사용자들이 쉽게 이해할 수 있도록 하기 위한 자동 군집화 방법을 제안하였다. 유사도와 적합도를 이용한 적합도함수를 제안하고 유전자 알고리즘을 이용하여 분류의 효율성을 향상시켰다. 향후 과제로는 의미망의 내부구조외에 메타시소러스에서의 개념들과 의미유형과의 관련성을 함께 고려한다면 좋은 결과를 기대할 수 있을 것이다.

참고 문헌

[1] James Geller, Yehoshua Perl, Michael Halper, Zone Chen, and Hyanying Gu, " Evaluation and Application of a Semantic Network Partition," IEEE Transaction on Information Technology Biomedicine Vol. 6, no. 2, 109-115, June 2002.

[2] Soumeya L, " A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development," JAMIA Vol. 8, No. 4, 351-360, Jul./Aug. 2001.

[3] NLM, UMLS Knowledge Source Server, URL: http://umlsks.nlm.nih.gov/.

[4] Zong Chen, Yehoshua Perl, Michael Halper, James Geller, and Huanying Gu " Partitioning the UMLS Semantic Network" IEEE Transaction on Information Technology Biomedicine Vol. 6, no. 2, 102-108, June 2002.