시소러스 자동 구조화

Automatic Structuralization of Thesaurus

○김 해 수 이 남 경 이 원 규
한국문화예술진흥원

HaeSoo Kim NamKyung Lee WonGyu Lee
{hskim,nklee,lee}@caibs.kcaf.or.kr
The Korean Culture & Arts Foundation, Seoul, KOREA

요 약

정보과학 분야에서 필요로 하는 의미해석 기술 개발에 중요한 역할을 하는 것이 시소러스이다. 정확한 관계 정의 및 대량정보 수용의 필요성은 시소러스 구축의 커다란 장애 요인이다. 시소러스 구축에은 다방면의 전문지식 활용과 막대한 비용 및 시간 투자가 요구된다. 기계적으로 관리 운영이 가능한 시소러스내의 관계는 BT, NT로 표현되는 계층관계와 USE, UF로 표현되는 등가관계로 이루어 진다. 본 연구는 개념적인 관계를 정의하는 두 관계를 기계적으로 추출하고, 기존의 평면적인 시소러스 구조를 실세계의 정보 구조에 적합하게 조직화하여 시소러스 구축에 있어서의 문제점을 개선하는데 그 목적이 있다. 제시되는 알고리즘은 단일 언어내의 시소러스 구축 뿐만 아니라, 구축된 시소러스들의 융합 및 다국어 시소러스 구축에도 적용된다.

1. 서 론

시소러스(Thesaurus)란 「어떤 두 단어간의 개념적인 계층관계를 정의한 관련어 사전」이라 정의된다. 시소러스의 시초라 일컬어 지는 P. M. Roget의 "Thesaurus of English words and phrases"는 문장 작성시에 표현하고 싶은 의미를 명확히 나타내는 단어를 선택하기 위해서 이용된 유사어 사전이었다. 도서관에서 방대한 도서를 정리하기 위해 사용된 주제표목표(subject headings list)도 시소러스의 일종이라 하겠다. 기계에 의한 자연언어처리에서는 정보검색시에 원하는 정보를 추출하기 위하여 사용되는 검색용어집으로서의 역할을 수행한다.

기존의 시소러스로는 Thesaurus of Engineering and Scientific Terms(TEST), BSI[1] ROOT 시소러스, CAB 시소러스, Food: multilingual thesaurus, INSPEC 시소러스, Thesaurus of ERIC descriptors, UNESCO:IBE 교육 시소러스, UNBIS 시소러스, IRRD[2] 시소러스, JICST 과학기술용어 시소러스, 뉴스 시소러스[3] 등이 있다. 표현방법이나 범위 등에 따라 각각 특성을 갖고 있지만, 기본적으로는 등가관계, 계층관계, 연관관계와 같은 기본적인 관계들로 구성되어 있다.

[1] British Standard Institute
[2] Intenational Road Research Documentation
[3] 일본 중일신문사

1. 등가관계

BS 5723 및 ISO 2788에서는, 색인작업시 복수개의 단어가 같은 개념을 나타내고 있다고 인정되는 경우, 우선어 및 비우선어의 관계라고 정의하고 있다. 우선어란 그 개념을 표현하기 위하여 색인 작업시 이용되는 단어를 말하며, 비우선어란 이용되지 않는 용어를 말한다. 우선어에 대한 접두기호는 USE, 비우선어에 대한 접두기호로는 UF(use for)가 사용된다.

2. 계층관계

개념상 상위어 및 하위어의 관계를 정의한 것이다. 상위어는 클래스 또는 전체를 나타내며, 하위어는 그 한 요소 또는 일부분을 나타낸다. 계층관계는 상위 및 하위 개념을 논리적으로 전개하는 순서로 위치시키기 위해 이용된다. 상위어 표시기호로는 BT(Broader term), 하위어 표시기호로는 NT(Narrower term)가 사용된다.

3. 연관관계

연관관계란 색인작성과 검색에 이용될 지도 모르는 대체용어를 제시하기 위해 용어간의 관계가 심리적으로 연상된다면 시소러스 내에 명시하는 것이다. 일반적으로 이 관계를 나타내기 위한 기호로는 RT(related term)를 이용한다.[1]

개념간의 관계는 관계명으로 관계가 설정되며, 두가지 개념은 서로 관계명으로 링크된다. 이 관계명은 방향성(양방향성)을 가지고 있으며 하나의 관계류가 하나의 층(layer)을 형성한다.

2. 문제점 분석

기존의 시소러스의 구축 및 유지에 있어서 중요한 문제점으로 고려되야 할 사항으로는 다음과 같은 것이 있다.

1. **많은 시간과 비용을 필요로 한다.**
분야에 따라 차이는 있겠지만 한 전문분야의 시소러스를 구축하는 데는 5 ~ 10년이 걸린다. 시소러스 구축에 선행되야 할 것이 용어 선정이다. 전문용어 사전구축이 여기에 해당하는 선행작업이라 하겠다. 일본 문부성이 주관한 "과학기술전문용어집"에 포함된 전문용어 하나당 40,000원 상당의 비용이 요구되었고, 5년간의 프로젝트로 진행된 EDR 사전구축에는 11조2천억원이라는 비용이 소요되었다. 더욱 어려운 점은 힘들여 수집한 용어가 시간의 흐름에 따라 변한다는 것이다. 결국 시소러스의 완성은 참여자들의 포기와 때를 같이 한다고 하겠다.

2. **대상 전문분야별로 많은 전문가를 필요로 한다.**
구축된 시소러스의 관리 및 이용 전문가는 있을 수 있어도, 시소러스 구축 전문가란 존재할 수 없다. 시소러스를 구성하고 있는 용어의 의미는 단일 전문분야내에서도 반드시 일치하는 것은 아니어서, 용어의 개념정립 등에 전문가들의 지식이 필수적이다. 그러나 이들 전문가도 반복적인 시소러스 구축작업에는 의욕을 보이지 않는 경우를 흔히 접할 수 있다.

3. **여러 분야에 걸친 전문가를 필요로 한다.**
지식 표현의 대표적인 수단이 언어이다. 언어는 생성, 성장, 소멸의 과정과 더불어 변화한다. 새로운 지식이나 수정된 지식을 표현하기 위해 기존의 용어를 이용하는 경우가 대부분이다. 그러나 기존의 용어는 타분야의 지식을 표현하기 위해 또는 다른 개념을 나타내기 위해 이미 이용되었던 것이다. 따라서 기존의 시소러스의 틀에 맞추기 위해서는 여러 분야의 전문가가 모여서 타협할 수 밖에 없다. 이 과정은 '표준화'와 자주 혼동되는 경우가 많으며 소집단의 주관적 관점에서 구축될 위험이 있다.

4. **다언어 시소러스 구축이 어렵다.**
언어간 개념차이 및 어감차이로 인하여 용어의 의미가 일치하지 않는다. 외국의 시소러스를 번역한 사례에 자주 등장하듯이, 언어간에 일대일 대응이 이루어지지 않는 경우가 빈번하다. 통신기술의 발달에 따라 국제적인 정보 유통이 활발해 지고 서로의 정보를 공유하고자 하는 움직임이 대두되면서 번역이 불가능한 부분은 그대로 반영하면서 다언어 시소러스 융합에 관한 제의도 찾아 볼 수 있다.[5]

이상의 문제점을 분석한 결과는 다음과 같다.

1. **실세계의 정보구조에 적합한 정보모델이 필요하다.**
정보(용어)의 특성은 중의성, 방향성, 재귀성으로 대표될 수 있다. 그러나 기존의 모델은 이러한 정보의 특성을 수용할 수 없다. 따라서 n차원의 정보공간에 표현할 수 있는 정보모델과 관리시스템이 요구된다. 동시에 고려해야 할 사항이 유연한 지식표현방법이다. 전문(Fulltext)에서 인식하는 것이 가장 정확하지만, 전문인식을 위해서도 기계가독형의 지식표현이 유용하다. 시스템 구축과정도 순환적으로 정의된다.[3]

2. **시소러스의 자동구조화가 필요하다.**
다언어(Multilingual) 다층(Multilayer)의 시소러스 구축도 순환적으로 이루어 진다. 구축에서 뿐만 아니라 유지, 관리에 있어서도 자동화는 필수적이다. 자동구축된 것을 포함해서 모든 시소러스의 평가는 정보검색에 있어서 시소러스의 이용결과에 근거한다. 따라서 새로운 정보를 바탕으로 학습하고 전문가의 참여를 최소화하면서 일관성 있는 시소러스의 유지, 관리가 이루어 져야 한다.

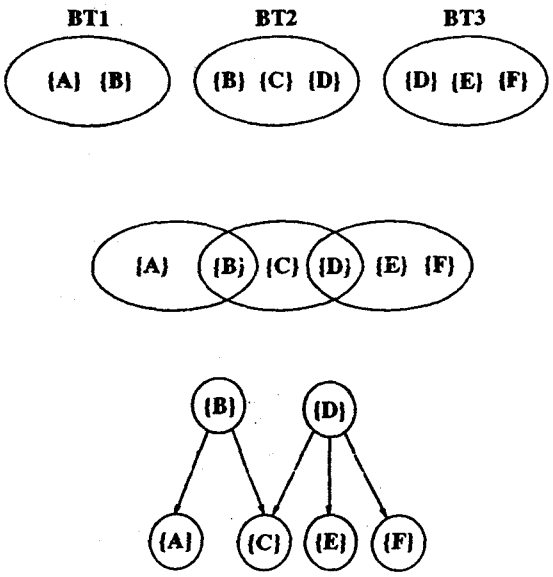

그림 1: 시소러스 자동생성 기본개념

3. 시소러스 자동구조화

3.1 기본 개념

통제언어(Control Word) 시스템에서 유래한 전거통제(Authority Control)는 표준화라는 전제하에 인위적인 색인시스템을 구축한다. 다음과 같은 이유에서 전거통제는 시소러스에 흡수된다.

1. 표현의 다양성은 동의어 관계로 흡수된다.
2. 연상되는 모든 표현을 관리할 수 없다.
3. 이용자의 요구가 없는 한 새로운 표현은 관리, 검색되지 않는다.

동의어집합을 생성하여 상위어 또는 상위어집합과 연결하여 전거처리의 과정은 시소러스 구축 및 참조과정내에 포함된다. 동의어집합의 생성은 용어집내에 표현된 동의어관계를 이용하여 기계적으로 구축할 수 있다. 그러나, 용어의 중의성에 의해 동의어집합이 무한대로 커지는 것이 단점이라 하겠다.

그림 1과 같이 구조화하기 위해서는 상위어를 추출해야 한다. 이상의 문제점을 바탕으로 시소러스를 자동구조화하기 위한 과정은 그림 2와 같다.[2]

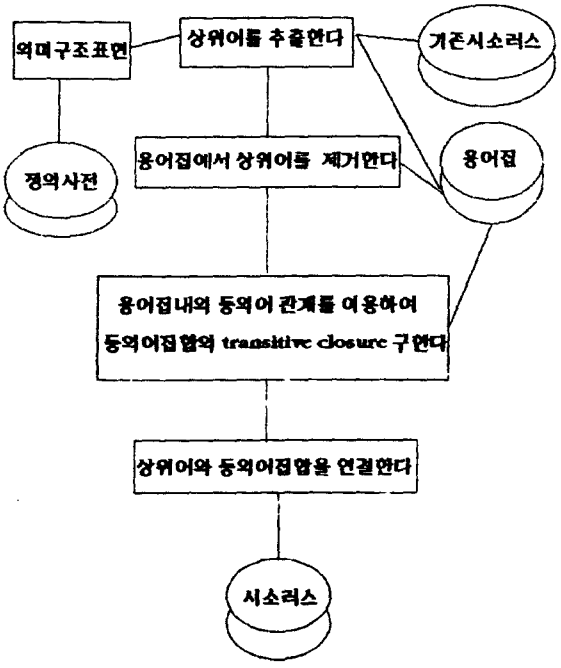

그림 2: 시소러스 자동생성 작업흐름도

3.2 상위어 추출방법

기본적으로 시소러스에서의 상위어는 전문가에 의해 추출되지만, 이를 위해서는 많은 시간과 높은 비용을 요구하며, 또한 시소러스를 유지하기 위해서는 자동적으로 상위어를 추출해야만 하는 문제점을 안고 있다. 다음은 상위어 추출을 위한 4가지 방법이다.

3.2.1 조어의 규칙(Coinage rule) 사용

일반적으로 용어는 대부분 복합어로 되어 있다. 통계분석에 의하면 전체 용어에서 복합어가 차지하는 비율이 약 70%가 된다고 한다. (표 1). 조어의 규칙이란 복합어들 중에서 공통된 단어를 포함한 용어들의 집합을 구한 후 그 단어를 상위어로 추출하는 방법을 말한다.

표 1: 영어에서의 복합어 비율

단어수	용어결과	비율
1	36,337	31.0
2	66,085	56.4
3	11,861	10.1
4	2,320	2.0
5	405	0.3
6	105	0.1
7	26	0.0
8	8	0.0
9	4	0.0
10	6	0.0

조어의 규칙을 이용한 상위어 추출방법의 하나로 SS-KWIC(Semantically Structured Key Word Element Index in Terminology Context)은 복합어의 특성을 이용하여 계층관계를 추출하는 새로운 타입의 인덱스이다.

- SS-KWIC의 여섯가지 규칙[6]
  1. 수식 원소들과 수식된 원소들의 위치
  2. 원소들의 결합
  3. 구성 원소들의 수
  4. 접두사, 접미사, 수사 등과 같은 특수 원소
  5. 명사 또는 복합어의 복수형
  6. 그외(생략, 단순화 등등)

3.2.2 용어의 정의(definition) 사용

용어의 정의는 관계된 많은 용어들을 수록한 사전에 정의되어 있다. 새로운 용어는 잘 정의된(well-defined) 용어에 의해 새로이 정의됨으로써 새로운 용어와 기존에 있던 용어들 간의 관계가 정의된다. 예를 들어, "~ 의 일종이다." 라고 하면, 그 용어에 대한 계층 관계를 나타내는 말로 사용되기 때문에 이러한 방법에 의해 계층관계를 생성 시킬 수 있다.

3.2.3 기존의 시소러스 사용

기존의 시소러스는 대부분 전문가들에 의해 수동으로 만들어 졌다. 비록 기존의 시소러스들이 구축된 오래되고, 새로운 용어들의 갱신이 이루어지지 않았다 하더라도, 분류용어에 가까운 기본적인 용어들간의 관계를 정의하고 있으므로 시소러스 자동구조화에 폭 넓게 활용된다. 타 시소러스와의 융합 및 세분화된 분야의 하위어 집합과의 연계에도 효율적이다.

3.2.4 용어의 외미관계 표현에서 추출

만약 외미구조가 구축되었다면, 이 외미구조안에서 표현된 외미관계에서 상위어를 추출할 수 있다. 즉, 외미구조 모델이 새로운 용어를 정의하는데 사용되면, 새로운 용어와 기존의 용어 사이의 관계는 외미구조안에 묘사된다. 위의 세가지 방법은 전문가에 의해 결과를 최적화해야 할 필요가 있지만, 이 방법은 시소러스를 자동 구조화시키는데 가장 좋은 방법이라 할 수 있다.

3.3 동의어집합 생성방법

용어들 사이의 동의어 관계를 이용하여 자동으로 동의어 집합을 생성하여 구조화시키는 방법으로는 C-TRAN(Constrained Transitive Closure)를 이용한다. Transitive Closure(TC)는 용어에 대한 동의어 관계를 이용하여 동의어 집합을 구하는 방법으로서, 시작 용어를 선택하여 그 용어에 대한 동의어를 구하고, 시작용어와 동의어들을 포함한 동의어 집합에 대해 더 이상 동의어 요소(element)가 증가하지 않을 때까지 동의어 관계를 이용하여 동의어를 찾아 동의어 집합에 추가시킨다. 동의어 집합의 TC를 구하기 위한 알고리즘은 다음과 같다.[4]

$\mathcal{M}, \mathcal{N}$: 각 용어의 집합
$\mathcal{E}$: 동의어 관계의 집합
S_m, S_n : 각각 $\mathcal{M}, \mathcal{N}$ 의 동의어 집합

동의어 집합의 초기 상태($m \in \mathcal{M}, n \in \mathcal{N}$) :

$$S_m^0 = \{m_0\}, S_n^0 = \{\}$$

동의어 집합 :

$$S_n' = \{n_k | (m_0, n_k) \in \mathcal{E}\}$$

동의어 관계를 이용한 $\mathcal{M}, \mathcal{N}$ 의 동의어 집합 :

$$S_n' = \bigcup_{m_p \in S_m^i} \{n_q | (m_p, n_q) \in \mathcal{E}\}$$

$$S_n^{i+1} = S_n^i \cup S_n'$$

$$S_m' = \bigcup_{n_s \in S_n^{i+1}} \{m_r | (m_r, n_s) \in \mathcal{E}\}$$

$$S_m^{i+1} = S_m^i \cup S_m'$$

$S_n^+ \subseteq \mathcal{M}, S_m^+ \subseteq \mathcal{N}$ 이 TC일 조건 :

$$S_n^i = S_n^{i+1} = S_n^{i+2} = \ldots = S_n^+$$
$$S_m^i = S_m^{i+1} = S_m^{i+2} = \ldots = S_m^+$$

4. 결론

본 연구는 기존의 시소러스 구축 및 관리에 수반되는 제반 문제점들을 파악하여, N 차원 공간에서 표현되는 시소러스 자동구축 방법을 제시하였다. 제시된 방법은 기존의 시소러스의 융합 및 타언어로 구축된 시소러스의 흡수와 언어의 장벽을 초월한 다국어 시소러스 구축의 가능성을 보인다. 그러나 모든 관계를 허용하는 N 차원의 시소러스를 구축하는 데는, 첫째, 용어들 사이의 관계들이 무한정 늘어나고, 둘째, 개념적인 관계가 명확히 표현되지 않기 때문에, 전문가의 주관적인 생각이 개입되므로, 도메인에 따라서 관계명을 설정하는 것이 본 연구에 있어서 해결해야 할 가장 큰 과제라 할 수 있다. 따라서 본 연구는 앞으로 방대한 데이타를 연구 결과에 적용시켜 구축 가능성을 확인한 후, 본 연구 결과가 적용된 다국어 시소러스 자동 구축 및 멀티미디어 데이타를 수용한 다국어 다매체 시소러스 자동구축 방법 및 그 활용에 대한 연구를 추진할 예정이다.

참고 문헌

[1] J. Aitichison and A. Gilchrist. *"Thesaurus Construction: A Practical Manual"*. Aslib, 2nd edition, 1987.

[2] Y. Fujiwara, W. G. Lee, Y. Ishikawa, T. Yamaguchi, A. Nishioka, K. Katada, N. Ohbo, and S. Fujiwara. "A Dynamic Thesaurus for Intelligent Access to Research Databases". In *44th FID Congress*, Helsinki, August 1988.

[3] YunHi Kang, SungHo Cho, and WonGyu Lee. "The Information Model based on Semantic Structures". In *KSIM*, 1994.

[4] W. G. Lee. *"Construction of Semantic Structures in the Self-Organizing Information-Base System"*. PhD thesis, The University of Tsukuba, February 1993.

[5] Gerhard Rahmstorf. "A New Thesaurus Structure for Semantic Retrieval". In S. Fujiwara, editor, *47th FID Conference and Congress*, pages 114–121, October 1994.

[6] Norihiko Uda. *"Information Analysis for Modeling and Representation of Meaning"*. PhD thesis, The University of Tsukuba, February 1994.