문서로부터 계층적 개념 트리 자동 구축

김희수[O] 조용석[1] 최익규[2] 김민구[3]
아주대학교 정보통신전문대학원[02]
아주대학교 정보통신대학[13]
{heemanz[O], laconic, ikchoi, minkoo}@ajou.ac.kr

Automatic Construction of Concept Hierarchy from Text

Hee-soo Kim[0], Yongsuk Cho[1], Ikkyu Choi[2], Minkoo Kim[3]
Graduate School of Information and Communication, Ajou University[02]
College of Information Technology, Ajou University[13]

요 약

계층적 개념 트리는 개념의 전체적인 구조를 제공하여 사용자의 이해를 돕는다. 이러한 계층적 개념 트리는 특정 분야의 전문가나 지식 공학자, 혹은 개인에 의해서 제공되어 왔다. 하지만 계층적 개념 트리를 구축하기 위해서는 많은 시간과 노력이 요구된다. 따라서 계층적 개념 트리를 자동으로 구축하기 위한 시스템 필요하게 되었다. 이 논문에서는 정보 검색 및 온톨로지 연구에 있어서 계층적 개념 트리를 자동으로 구축하기 위한 기존 연구에 대해서 알아보고, 개념적 클러스터링 방법인 FCA(Formal Concept Analysis)를 사용하여 문서로부터 계층적 개념 트리를 구축하는 방법을 제안하고자 한다.

1. 서 론

계층적 개념 트리(**concept hierarchy**)는 문서집합의 구조화와 요약을 제공하고, 필요한 정보의 수월한 접근을 제공하기 위해 사용되었으며[1], 또한 지식을 표현하는데 있어서 간결하면서도 효과적으로 개념간의 구조를 설명하기 위해 사용되었다. 이러한 개념 트리의 대표적인 예로는, 사용자의 지식을 이용하여 정보에 접근하는 경로를 제공하는 **'Yahoo!'**의 디렉터리 서비스, 또는 지식을 나타내거나 **WordNet**과 같이 개념간의 **is-a** 관계(**hyponym, hypernym**)를 나타내기 위해 사용되는 트리 형태의 그래프 등이 있다. 그러나 지금까지 대부분의 계층적 개념 트리는 특정 분야의 전문가나 지식 공학자들에 의해서 만들어졌다. 하지만, 계층적 개념 트리를 구축하는 작업은 많은 시간과 노력을 필요로 한다. 또한, 오늘날과 같이 급변하는 환경에서 이미 구축된 계층적 개념 트리는 지속적인 수정을 통해 새롭게 생성되는 개념 및 현상, 사물 등을 반영해야 할 필요가 있다. 따라서 본 논문에서는 계층적 개념 트리의 구축을 돕기 위한 방법으로 문서로부터 계층적 개념 트리를 구축하는 기존 방법들에 대해 알아보고, 자동화된 계층적 개념 트리의 구축 방법을 제안한다.

본 논문은 다음과 같이 구성된다. **2**장에서는 관련 연구와 문제정에 대해서 알아보고, **3**장에서는 정형화된 계층적 개념 트리 구축 방법에 대해 설명한다. 그리고 **4**장에서는 실험 결과를 보여주고, **5**장에서는 결론과 추후 과제에 대해서 언급한다.

* 본 논문은 과학기술부의 국가지정연구실 사업(과제명:차세대 인터넷을 위한 지능형 온톨로지 자동생성 시스템 개발, 과제번호:M10302000087-03J0000-04400) 지원으로 수행되었음

2. 관련 연구

문서로부터 계층적 개념 트리를 구축하는 방법으로는 개념의 정의에 따라 크게 두 가지로 분류할 수 있다. 첫 번째 방법은 개념을 단어들의 집합으로 간주하여 계층적 개념 트리를 구축하는 방법으로 계층적 군집화(**clustering**)를 이용하는 것이다. 하지만 계층적 군집화를 사용한 방법은 형성된 클러스터에 대한 명명 문제 등과 같이 사용자에게 직관적인 정보를 제공하기 어렵고[1], 클러스터링의 결과, 클러스터가 편중되는 현상을 보이는 문제를 가지고 있다. 다른 방법은 문서에 존재하는 어휘를 개념으로 간주하고 통계적 방법을 이용하여 계층적 개념 트리를 구축하는 것이다[2][3][4]. 이러한 방법들은 일반적으로 검색된 문서 집합들로부터 정보의 수월한 접근을 제공하기 위해 사용되지만, 개념간의 관계가 엄밀한 의미의 **is-a** 관계가 아닌 연구자가 제안한 방법에 의해 부여된 개념간의 관계에 의해 구축된 것이다. 이에 반하여 개념적 클러스터링 기법인 **FCA (Formal Concept Analysis)**는 개념에 대한 속성을 정의하고 속성의 포함관계를 사용하여 계층적 개념 트리를 생성하는 수학적 방법을 제시한다[5]. **Gu Tao**는 온톨로지의 계층적 개념 트리의 구축을 돕기위해 **FCA**를 사용하는 방법을 제안하였다[6]. 또한 **FCA**를 사용하여 문서로부터 온톨로지의 개념 트리를 자동으로 구축하기 위해 문장에서의 동사와 목적어를 추출하여, 목적어를 개념으로, 동사를 개념의 속성으로 간주하여 **FCA**에 적용하는 방법이 연구되었다[7]. 그러나, 이 방법은 개념의 특징을 설명하기에 부적합한 동사, 특히 일반적으로 자주 사용되는 동사가 개념의 속성으로 추가됨으로써, 합리적인 **is-a** 관계를 형성하는 것을 방해한다.

이 논문에서는 FCA를 사용하여 계층적 개념 트리를 구성하는 방법의 타당성과 발생하는 문제에 대한 해결 방법에 대한 실험 결과를 보여준다.

3. 계층적 개념 트리의 생성

어휘의 통계적 방법을 통한 계층적 개념 트리의 구축은 단지 문서 집합에서 그 어휘의 빈도수에 기반하기 때문에 이러한 방법에 의해 추출된 개념간의 관계가 **is-a** 관계를 의미하지 않을 수 있다. 인공지능의 과거 연구들은 지식을 표현하는 방법으로 개념을 다음과 같이 정의했다. 개념은 속성들의 집합으로 정의되며, 각 개념과 관련된 속성(attribute)은 저마다의 제약(role restrictions)을 갖는다[8]. 이러한 정의는 계층적 개념 트리를 구축하기 위한 중요한 정보를 제공한다. 계층적 개념 트리를 구축하기 위해 개념이 가지고 있는 속성이 중요하게 작용한다는 것이다. 예를 들면, '동물'과 '사람', '개'의 개념을 분류하는 기준은 각각의 개념이 가지고 있는 속성에 기반한다. '동물'의 속성은 모두 '사람'과 '개'의 개념에 포함되며, '사람'과 '개'는 '동물'이 가지는 속성에 '젓을 먹인다'라는 속성을 추가적으로 갖는다. 또한 '사람'과 '개'를 구별하는 속성은 '말한다'와 '짖는다'로 생각할 수 있다. FCA는 개념이 가지고 있는 속성을 기반으로 계층적 개념 트리를 구축하기 위해 사용된다.

3.1 계층적 개념 트리 구축 시스템의 구조

시스템의 전체 구조는 다음과 같다.

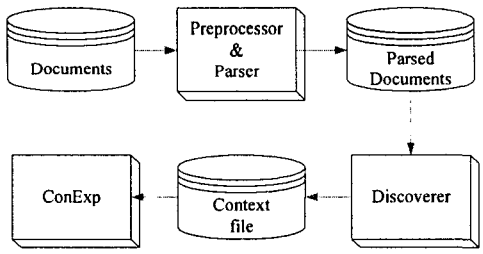

그림 1 시스템의 구조도

먼저, 전처리 과정으로 문서들을 문장 단위로 나누고 Parser를 통해 문법적인 태깅을 한다. 그 다음, Discoverer를 사용하여 문장에서의 목적어-동사를 찾아 formal context를 나타내는 context 파일을 생성한다. Formal context에 대해서는 3.3절에서 설명한다. 마지막으로 ConExp(Concept Explorer)를 사용하여 context 파일로부터 계층적 개념 트리를 구축한다.

시스템의 구성 요소 중 Parser와 ConExp는 각각 Apple Pie Parser (http://nlp.cs.nyu.edu/app/)와 ConExp(http://www.mathematik.tu-darmstadt.de/ags/ag1/fag1/Software/ConExp/)를 사용했다.

3.2 개념의 추출

문서로부터 개념을 추출하기에 앞서, 개념을 속성들의 집합으로 정의한다. 개념과 속성은 각각 문장에서 목적어와 동사의 형태로 문서에 출현한다는 가정하에 개념을 정의한다[7]. 하지만, 동사는 의미를 전달하기 위해 다양한 목적어를 취할 수 있다. 바꾸어 말하면, 문서에서 다양한 특정 동사는 목적어와 일 대 다의 관계를 맺으며, 이러한 관계는 FCA에 의해 구축된 계층적 개념 트리의 정확도를 떨어뜨리는 요인으로 작용한다. 이러한 문제를 해결하기 위해 개념에 포함되는 속성의 필터링이 필요하다. 본 연구에서는 개념과 속성간의 포함관계를 판단하기 위한 척도로써 ***KL (Kullback-Leibler) divergence score***를 사용하며, 이에 대한 정의는 다음과 같다.

$$KL\,contribution_c(a) = P_c(a)\log_2\frac{P_c(a)}{P_C(a)}$$

$$P_c(a) = \frac{freq(c \wedge a)}{\sum_{a' \in c} freq(c \wedge a')} \qquad (1)$$

$$P_C(a) = \frac{\sum_{c \in C} freq(c \wedge a)}{\sum_{c \in C}\sum_{a' \in c} freq(c \wedge a')}$$

C는 전체 개념의 집합을 나타낸다. 개념 c에 포함된 속성 a의 중요도는 KL divergence score으로 나타낼 수 있으며, 속성이 개념을 설명하기 위해 자주 사용되면, 그 속성은 개념을 표현하는 중요한 정보로 사용될 수 있음을 의미한다.

3.3 FCA를 사용한 계층적 개념 트리의 구축

FCA는 주로 데이터의 분석을 위해 사용되며, 이러한 데이터는 개념의 형식적인 추상화(formal abstractions)의 단위로 구조화된다. FCA를 이해하기 위한 정의는 다음과 같다.

정의 1. *Formal context* (G, M, I)는 두 개의 집합 G와 M, 그리고 G와 M사이의 관계 I로 구성된다. G와 M의 원소들은 각각 그 *context*의 *objects*와 *attributes*를 의미한다. 또한 object g가 attribute m과 관계가 있을 때, gIm 또는 $(g, m) \in I$로 나타내며, g는 m을 갖는다는 것을 의미한다.

$A \subseteq G, B \subseteq M$을 만족하는 집합 A와 B에 대하여,

$$A' = \{m \in M \mid (g,m) \in I, \text{ for all } g \in A\},$$
$$B' = \{g \in G \mid (g,m) \in I, \text{ for all } m \in B\}$$

A'과 B'을 각각 정의하면, 다음과 같이 *formal concept*과 상위-하위개념을 정의할 수 있다.

정의 2. Formal context (G, M, I)에서 $A \subseteq G, B \subseteq M$일 때, $A'=B, B'=A$를 만족하는 (A,B)를 *formal concept*이라고 한다.

정의 3. (A_1, B_1)와 (A_2, B_2)가 formal context의 formal concept일 때, concept간의 $(A_1, B_1) \leq (A_2, B_2)$가 성립하기 위한 필요충분 조건은 $A_1 \subseteq A_2$ 또는 $B_1 \supseteq B_2$이다. ($\leq$는 concept의 포함관계를 나타낸다.)

간단한 예로 '생물', '동물', '사람', '개'의 개념이 있고, 이것을 formal context로 나타내면 표 1과 같다고 가정하자.

objects	Attributes			
	살아있다	움직인다	운다	짖다
생 물	O			
동 물	O	O		
고양이	O	O	O	
개	O	O		O

표 1. Formal Context의 예

위 예에서 네 개의 objects는 정의 2를 만족하는 formal concept이다. 그리고 정의 3에 의해 그림 1과 같은 계층적

개념 트리를 얻을 수 있으며, 속성이 개념을 분류하는 기준으로 사용되는 것을 볼 수 있다.

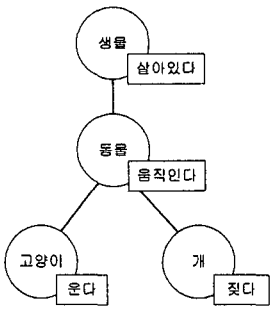

그림 2. FCA로 구성된 계층적 개념 트리의 예

본 논문에서는 **FCA**를 수행하기 위해 **ConExp** 프로그램을 사용했다.

4. 실험 결과

이 논문에서 제한된 계층적 개념 트리의 구축을 위해 사용된 문서집합은 **TREC(Text REtrieval Conference)**에서 제공하는 **Text Research Collection Volume 2**의 **Associated Press (1988)**를 사용하였으며, **AP**와 관련된 개념을 선택하기 위해 **'Yahoo!'** 디렉터리 서비스의 **'Business & Economy(BE)'**와 **'Society & Culture(SC)', 'Social Science(SS)', 'Government(GOV)'**에 포함된 카테고리의 제목으로부터 각각 **253, 273, 220, 163**개의 단어를 선택했다. 또한, **KL-divergence 0.25**를 임계값으로 개념을 속성의 집합으로 정의했다. 마지막으로 **FCA**를 사용하여, 네 개의 개념 집합에서 각각의 계층적 개념 트리를 구성하였다. 그림 **3**은 구축된 개념 트리에서 의미 있는 **is-a** 관계를 간추린 것이다. 그림 **3**에서 개념은 '개념:속성'의 형식으로 표현되며, 표 **2**는 실험 결과에 대한 통계적 결과를 보여준다.

Society & Culture	Social Science
:introduce ├recreation:say └exhibit:review :included ├hippie:call └skinhead:see,look :affect ├gay:play,support └lesbian: :enjoy ├romance:attract └wildflower:tell	:say ├sociology:teach ├economics:impose ├anthropology:get └archaeology:take └psychology:avoid,receive :use ├directory:make,consider └reference:become philosophy:attack └marxism:avoid,consider,ban,modify :ban ├marxism:avoid,consider,attack,modify └propaganda:dismiss
Business & Economy	**Government**
:wear ├badge:flash └jewelry:steal	pact:demand └conscription:sought

그림 3 계층적 개념 트리 구축 결과

실험 결과는 ':use' 등과 같이 개념 집합에 명시되지 않은 많은 잠재적 개념이 상위 개념으로 나타나는 것을 볼 수 있으며, 개념을 정의하기 위해 사용된 속성의 수에 비해

	BE	SC	SS	GOV
No. of concepts	80	92	38	22
No. of attributes	65	69	31	21
Total no. of is-a relations	51	63	28	8
No. of exact is-a relations	2	8	10	1
Precision of relations (%)	3.92	12.69	35.71	12.5

표 2 계층적 개념 트리 구축의 평가

개념의 수가 비교적 많음을 볼 수 있으며, 이것은 속성을 효과적으로 구별하지 못함을 의미한다. 또한, 특수한 개념간에 존재하는 is-a관계가 비교적 정확한 결과를 보인다.

5. 결론 및 추후 과제

본 논문에서는 문서로부터 계층적 개념 트리를 구축하기 위해 **FCA**를 사용하는 방법을 제시했다. 실험 결과는 문서에 등장하는 개념간의 **is-a** 관계를 발견할 수 있으나, 그것의 정확도가 현저히 떨어짐을 보여준다. 이것의 이유는 동사는 취하는 목적어와 전치사의 종류에 따라 다양한 의미를 가지는데 반하여 제안한 방법에서는 동일한 동사를 같은 의미의 속성으로 간주하기 때문이다. 계층적 개념 트리 구축 시스템의 정확도를 높이기 위한 추후 과제로써 동사의 의미를 구별하기 위한 방법을 도입하고, 개념-속성간의 관계를 정의하는 다양한 방법에 대한 실험과 상위 개념으로 나타나는 잠재적 개념의 발견에 대한 연구를 진행할 것이다.

6 참고 문헌

[1] Dawn Lawrie and W. Bruce Croft, "Discovering and Comparing Topic Hierarchies, Proceedings of RIAD2000 conference, pp. 314-330, 2000.

[2] Mark Sanderson and Bruce Croft, Deriving Concept Hierarchies from Text, Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 206-213, 1999.

[3] Dawn Lawrie, W. Bruce Croft, and Arnold Rosenberg, "Finding Topic Words for Hierarchical Summarization", Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 349-357, 2001.

[4] Dawn J. Lawrie and W. Bruce Croft, "Generating Hierarchical Summaries for Web Searches", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 457-458, 2003.

[5] Uta Priss, "Formal Concept Analysis", http://www.fcahome.org.uk/

[6] Gu Tao, "Using Formal Concept Analysis (FCA) for Ontology Structuring and Building", http://www.ntu.edu.sg/home5/pg04247176/index.htm

[7] Philipp Cimiano, Steffen Staab and Julien Tane, "Automatic Acquisition of Taxonomies from Text: FCA meets NLP", Proceedings of the GI Workshop Lehren - Lernen - Wissen – Adaptivität (LLWA), 2003.

[8] Natalya F. Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, 2001.