온톨로지 기반 점진적 클러스터링 기법에 관한 연구

김제민° 박영택
숭실대학교 컴퓨터학과
d5f4g3h2@hanmail.net, Park@computing.ssu.ac.kr

A Study of Incremental Clustering Technique based on Ontology

Je-Min Kim° Young-Tack Park
Soongsil University

요 약

클러스터링은 무질서한 데이터들의 상호 연관 관계를 정의하고, 이를 통하여 보다 체계적으로 데이터를 군집화하는 것이다. 클러스터링을 적용한 웹 서비스 시스템은 비슷한 내용을 묶어 제공하기 때문에 사용자는 보다 효율적으로 정보를 제공받을 수 있다. 시멘틱 웹의 기반이 되는 온톨로지는 클러스터링을 위한 완벽한 입력 데이터를 제공한다. 본 논문은 온톨로지를 기반의 메타 데이터를 클러스터링 하기 위한 기법을 제안한다. 본 논문의 목적은 온톨로지 기반의 메타 데이터들의 유사성을 측정하기 위한 평가함수를 정의하고, 이러한 평가함수를 적용한 계층적 클러스터링 알고리즘을 연구하는 것이다.

1. 서 론

1989년에 처음 제안된 월드와이드웹은 간단한 HTML을 사용하여 사용자가 쉽게 정보를 접근하거나 게시할 수 있게 되었다. 그러나 HTML은 문서의 의미와 시멘틱 정보를 표현하기 어렵기 때문에 소프트웨어 에이전트가 문서의 의미를 추출하기는 어렵다. 시멘틱 웹은 웹에 올라오는 정보에 의미를 부여하고, 이를 통해서 사람과 소프트웨어 에이전트의 협동적인 작업이 가능하게 하는 차세대 웹 관련 패러다임이다. 온톨로지는 각각의 도메인 영역을 대표하는 개념들을 계층화하고 관계를 모델링한 일종의 스키마로서 시멘틱 웹의 중추적인 역할을 한다. 즉, 시멘틱 웹 상에서는 온톨로지를 가반으로 웹 페이지마다 의미 정보를 메타 데이터로 덧붙이면, 소프트웨어 에이전트가 이 웹 페이지의 의미를 이해할 수 있다.

현재 시멘틱 웹을 기반으로 사용자에게 웹 정보를 효율적으로 제공하기 위한 웹 에이전트들이 개발되고 있다. 시멘틱 웹 에이전트는 각 웹 페이지에 덧붙여진 메타 데이터를 기반으로 사용자의 인터넷 정보 브라우징 행위를 모니터하고, 모니터된 메타 데이터를 기반으로 사용자의 관심 정보를 학습하여, 사용자가 필요한 정보를 자동으로 제공해 준다. 이러한 웹 에이전트의 성능은 사용자의 기호 파악에 좌우 되며, 사용자의 기호를 파악하기 위한 방법으로 클러스터링을 이용한다.

클러스터링은 무질서한 데이터들의 상호 연관 관계를 정의하고, 이를 통하여 보다 체계적으로 데이터를 군집화하는 것이다.[1] 웹 문서를 클러스터링 하기위해서는 웹 문서에 나오는 모든 단어들 중에서 클러스터링을 하기 위한 속성으로 유용하게 사용될 단어를 선택하는 특징 추출(Feature Selection)이 중요하다. 그러나 기존의 웹 환경에서는 이러한 특징 추출이 웹 컨텐츠에 명시되어 있는 키워드를 기반으로 실행되기 때문에 정확한 클러스터 속성을 파악하기가 힘들다. 반면에 시멘틱 웹의 기반이 되는 온톨로지는 웹 컨텐츠를 설명해주는 여러 가지 대표적인 속성들이 명시된 메타 데이터를 제공하기 때문에 클러스터링을 위한 완벽한 속성과 입력 데이터들을 제공한다.

본 논문은 시멘틱 웹 기반의 웹 정보를 효과적으로 클러스터링 하기 위해, 온톨로지 기반의 메타 데이터를 클러스터링 하기 위한 기법을 제안한다. 본 논문의 목적은 온톨로지 기반의 메타 데이터들의 유사성을 측정하기 위한 평가함수를 정의하고, 이러한 평가함수를 적용한 계층적 클러스터링 알고리즘을 연구하는 것이다.

2. 온톨로지와 메타데이터 모델

본 절에서는 온톨로지 기반의 메타데이터(metadata)를 클러스터링 위해서 가장 기본이 되는 온톨로지와 메타데이터에 대한 모델을 정의한다. 온톨로지는 일종의 스키마로서 "공유가 가능하도록 정형적이고 명시적으로 표현된 개념"으로 정의할 수 있다. 보다 간단하게 표현하면 특정 도메인을 대표하는 단어들을 정의하고, 이들이 가지는 속성을 정의하며 이들 간의 관계를 설정한다. 다음 식은 온톨로지의 구성에 대한 모델이다.

$$O := \{C, P, A, S^c, prop\}$$

C와 P는 개념과 개념들 간의 관계(속성)를 나타내며 S^c 는 개념간의 계층구조를 표시한다. S^c는 C_1이 C_2의 하위개념 이라는 상세 정보를 $S^c=(C_1,C_2)$로 나타낼 수 있으며, P 역시 C_1과 C_2가 P라는 관계를 맺고 있다는 상세 정보를 prop(P) = (C_1,C_2)로 나타낸다. 즉, prop(P) = (C_1,C_2)는 P의 도메인이 C_1이고 범위(Range)가 C_2이라는 의미를 갖는데, 이는 domain:P→C_1, range:P→C_2로 나타낸다. 단, 범위(range)가 일반적인 데이터 값(string, integer,...)을 가질 경우 range(A):=STRING 으로 나타낸다.

$$MD := \{O, I, L, inst, instr, instl\} \quad (2)$$

위의 식은 메타데이터의 구성에 대한 모델이다. O와 I는 특정 온톨로지와 온톨로지의 메타데이터 집합(instance set)을 나타내며, L은 인스턴스가 가질 수 있는 리터럴들의 집합을 의미한다. 특정 개념(Class)에 속하는 인스턴스는 inst(I) = C로 표현되고, 인스턴스 간의 관계는 $P(I_1,I_2)$로 표현되는데 관계를 형성하는 범위(Range)가 특정 개념(Class)이라면, $P \rightarrow instr^{I \times I}$로 나타내며, 범위(Range)가 일반적인 데이터 값일 경우 $P \rightarrow instl^{I \times L}$로 나타낸다.

3. 온톨로지 기반 메타데이터들의 유사도 계산

데이터를 클러스터링 하기 위한 대부분의 기법들은 각각의 데이터 간의 유사도를 측정한 후, 유사도가 높은 데이터들끼리 클러스터링 하는 알고리즘을 사용한다. 본 논문에서 제안한 클러스터링 기법 역시 이와 같은 단계로 진행되는데, 먼저 시멘틱 웹 기반의 문서들의 유사도를 측정하기 위해서, 각 문서의 계층구조와 문서의 속성 관계를 명시한 온톨로지를 기반으로 개념 유사도(Class Similarity)와 관계 유사도(Relation Similarity) 구한다.

온톨로지는 상위 개념과 상위 개념의 속성들을 상속받는 하위 개념으로 구성되는데, 개념 유사도와 관계 유사도는 온톨로지의 이러한 특징을 이용한 유사도 계산방법이다. 즉 서로 다른 하위 개념들이 같은 상위 개념에 속한다면 개념적으로 유사한 부분을 공유하게 되는 것이며, 이들의 특정 속성 값의 범위가 같은 개념이라면 역시 어느 정도의 유사성을 공유하게 되는 것이다. 본 논문에서 제안하는 개념 유사도와 관계 유사도의 정의는 다음과 같다.

- 개념 유사도 - 문서와 문서, 문서와 문서군집이 위치하는 개념(Class)간의 거리를 계산
- 관계 유사도 - 문서와 문서, 문서와 문서군집의 속성 값들이 위치하는 개념간의 전체 유사도의 합

3.1 개념유사도

개념 유사도는 문서와 문서, 문서와 문서군집이 위치하는 개념(Class)간의 거리를 의미한다. 온톨로지의 개념체계(Class Hierarchy)에서는 각 개념들은 하위 개념을 가지며, 이들 하위 개념들은 또 다른 하위 개념을 가질 수 있다. 이러한 개념체계를 바탕으로 두 문서간의 개념 관계를 설명하기 위한 거리(Distance)를 측정함으로써 두 문서가 얼마나 유사한 개념을 공유하는지 알 수 있다. 바로 이 거리를 "개념 유사도"라고 부른다. 다음 식은 두 문서간의 개념 유사도를 구하는 방법을 정의하고 있다. (C, H^c)는 개념 C에 속한 문서의 계념계층 H^c를 의미한다.

$$CM(C_1,C_2) := \frac{|(C_1,H^c) \cap (C_2,H^c)|}{|(C_1,H^c) \cup (C_2,H^c)|} \quad (3)$$

3.2 관계 유사도

관계 유사도는 문서와 문서, 문서와 문서군집의 속성 값들이 위치하는 개념간의 전체 유사도를 의미한다. 온톨로지의 개념들은 각 개념에 속하는 개체(instance)의 특징을 설명해주는 프로퍼티(Property)를 가지며, 이중 객체 프로퍼티(Object Property)는 프로퍼티 값으로 다른 개념에 속하는 개체를 택할 수 있다. 이러한 객체 프로퍼티를 바탕으로 두 문서가 가지는 속성들의 유사성를 설명하기 위해 속성 값들의 유사도를 측정함으로써 두 문서가 얼마나 유사한지 알 수 있다. 따라서, 관계 유사도는 두 문서의 속성 값들의 개념간의 거리가 근간을 이루는데, 이는 두 문서의 속성 값들의 전체 유사도로서 정의된다. 전체 유사도는 3.3에서 설명하도록 한다. 다음 식(4)는 각 개체들이 공통된 객체 프로퍼티를 가지지 않는 경우를 정의한 것이다. 이는, 서로 다른 개념 내의 개체들이 더 이상 다른 개체들과 특정한 관계를 갖지 않음을 의미하며 이때의 속성 값들의 전체 유사도(Min_Similarity)는 0이 된다. 반면에 식(5)는 각 개체들이 공통된 객체 프로퍼티로 공통된 개체를 가지는 경우를 정의한 것이다. 이는, 서로 다른 개념 내의 개체들이 같은 개체와 관계를 갖는 것을 의미하며, 이때의 속성 값들의 전체 유사도(Mex_Similarity)는 1이 된다.

$$\begin{aligned} &if\ A_s(P,I_1)=0 \vee A_s=(P,I_2) \\ &then\ \ MinSim_{(P)} := 0 \end{aligned} \quad (4)$$

$$\begin{aligned} &if\ A_s(P,I_1)=A_s(P,I_2) \\ &then\ \ MaxSim_{(P)} := 1 \end{aligned} \quad (5)$$

다음 식은 두 문서간의 관계 유사도를 구하는 방법을 정의하고 있다. $Sim(A_s(P, I_1),A_s(P, I_2))$는 객체 I_1에 대한 프로퍼티 P의 속성 값과 객체 I_2에 대한 프로퍼티 P의 속성 값에 전체 유사도를 의미한다.

$$RS(I_1,I_2) := \frac{\sum_{Rcount=1}^{n} Sim(A_s(R,I_1),A_s(R,I_2))}{Rn} \quad (6)$$

3.3 전체유사도

전체 유사도는 문서와 문서, 문서와 문서군집에 대한 유사도를 나타내며, 개념유사도와 관계 유사도가 적용된다. 이렇게 구해진 유사도는 시멘틱 웹 기반의 문서들을 클러스터링 하는데 사용된다. 본 논문에서는 문서간의 유사도를 구할 때 가중치를 고려하지 않고 개념 유사도와 관계 유사도를 절반씩 적용하였다. 다음 식은 두 문서간의 유사도를 구하는 방법을 정의하고 있다. $Sim(I_1,I_2)$는 두 문서 I_1,I_2의 계념계층 유사도를 의미한다.

$$sim(I_1,I_2) = \frac{CS(I_1,I_2)+RS(I_1,I_2)}{2} \quad (7)$$

4. 온톨로지 기반의 점진적 클러스터링 알고리즘

여러 클러스터링 알고리즘 중 Cobweb은 점진적으로 개념 형성(Incremental Concept Formation)의 모델을 구축한다. 이는 범주 정보가 주어지지 않은 학습 예제 집합을 입력받아 각 단위 예제의 내용에 따라서 계층적인 범주분류 정보를 도출해내는 알고리즘이다. 예제에 대한 분류작업이 진행되면서 생성되는 계층적 개념 모델은 새로운 학습 예제를 입력받음에 따라서 현재까지 구

성된 개념 모델의 계층을 동적으로 변경한다. 새로운 학습 예제는 전체 개념을 표현하는 트리 구조의 정점에서부터 시작하여, 단말 노드에 이르기까지 재귀적으로 평가 함수에 의해 노드간의 유사도가 평가되며, 평가 결과에 따라 트리를 재구성하기 위한 연산을 실행한다. Cobweb의 트리 구성 연산자로는 각 하위 노드에 새로운 예제를 포함시키는 Incorporation, 새로운 노드를 생성하는 Create-new-disjunct 및 이미 구성된 하위 트리 모델의 구조 자체에 변경을 가하는 Merge와 Split이 있으며 평가 함수에 따른 결과에 따라 가장 적합한 연산자를 선택하여 적용한다.

시멘틱 웹 기반의 웹 문서들은 명확하게 문서의 정보의 계층 구조와 속성과 속성 값들을 메타 데이터로써 제공하며, 속성 값은 대부분 명사 형태(Nominal)다. 온톨로지 기반 시멘틱 웹의 이러한 특징과 본 논문에서 제안하는 유사도 측정법을 고려해 볼 때, Cobweb과 같은 클러스터링 알고리즘을 적용하는 것이 적절하다고 판단되어, 본 논문에서는 기존의 Cobweb 알고리즘을 기반으로 시멘틱 웹 기반의 웹 문서를 계층적으로 클러스터링하는 알고리즘인 Onto-Cobweb을 제안한다.

Onto-Cobweb이 기존의 Cobweb과 다른 점은 새로운 노드를 생성하는 연산자를 적용할 때이다. Cobweb은 새로운 예제가 다른 노드들과 독립적으로 떨어져 있을 때의 평가 함수가 가장 높으면 새로운 노드를 생성하는 연산자를 적용한다. Cobweb에 사용되는 Category Utility는 특정 군집 내부의 속성이 특정 속성 값을 가질 확률과 특정 속성 값이 특정한 군집의 속성에 속할 확률을 동시에 고려하기 때문에 독립된 노드를 생성하는 연산이 항상 높은 평가 결과를 갖지 않는다. 반면에 본 논문에서 제안하는 평가 함수는 항상 두 노드간의 유사도를 고려하기 때문에, 독립된 노드를 생성하는 연산이 매번 높은 평가 결과를 갖게 된다. 따라서, Onto-Cobweb에 적용되는 연산자 중에서 새로운 노드를 생성하는 연산자는 특정 임계값(Threshold)을 적용한다. 즉, 적용된 임계값이 다른 연산자에 적용된 평가 함수의 결과 값들 보다 높아야 새로운 노드를 생성하는 연산을 실행하게 된다. 본 논문에서의 임계값은 새로 생성된 노드가 같은 트리 레벨에서 존재할 확률(1 / (같은 트리 레벨에 존재하는 노드들의 총합+1))을 적용하였으며, 보다 정확하게 임계값을 구하는 것은 차후에 연구한다.

```
Onto-Cobweb (N, i)
    If N is a terminal node,
        Then Create-new-terminals (N, i) (similarity value < similarity threshold)
            Join(N,i) (similarity value   similarity threshold)
    Else Incorporate (N, i).
        For each child CN of node N, Compute the score for placing i in CN.
            Let Hs be the node with the highest score FN.
            Let Ss be the node with the second highest score SN.
            Let Ts be the threshold score for placing i in a new node.
            Let Ms be the score for merging Hs and Ss into one node.
            Let Ps be the score for splitting into its children Node (CN).
            If FN is the best evaluation,
                Then Onto-Cobweb (N, i) - place i in category N.
            Else if Ts is the best evaluation,
                Then - place i by itself in the new category Q.
            Else if (Hs + Ss) is the best evaluation,
                Then let M be Merge (FN, SN, N).
                Onto-Cobweb (M, i).
            Else if Ps is the best evaluation
                Then Split (CN, N).
                Onto-Cobweb (N, i).
```

Onto-Cobweb Algorithm

5. 결론 및 향후 연구

본 논문의 목적은 온톨로지 기반의 메타 데이터들의 유사성을 측정하기 위한 평가함수를 정의하고, 이러한 평가함수를 적용한 계층적 클러스터링 알고리즘을 연구하는 것이다. 시멘틱 웹의 기반이 되는 온톨로지는 웹 컨텐츠를 설명해주는 여러 가지 대표적인 속성들이 명시된 메타 데이터를 제공하기 때문에 클러스터링을 위한 완벽한 속성과 입력 데이터들을 제공한다. 이러한 특징을 적용하여 본 논문에서는 시멘틱 웹 기반의 웹 정보를 효과적으로 클러스터링 하기 위해, 온톨로지 기반의 메타 데이터를 클러스터링 하기 위한 기법을 제안하였다.

본 논문에서는 입력 데이터들 간의 유사도를 측정하기 위해서, 각 문서의 계층구조와 문서의 속성 관계를 명시한 온톨로지를 기반으로 개념 유사도(Concept Similarity)와 관계 유사도(Relation Similarity)를 적용한 평가 함수와, 이러한 평가 함수의 결과를 바탕으로 계층적으로 클러스터링하는 알고리즘인 Onto-Cobweb을 정의하였다. 본 논문에서 제안하는 클러스터링 기법은 시멘틱 웹상에 존재하는 웹 문서들을 각각의 특성에 따라 분류함으로써 콘텐츠 추천 에이전트가 사용자의 기호에 맞게 웹 문서를 추천해주는데 유용하게 사용될 수 있다.

향후 연구에서는 Onto-Cobweb에 전체 크기와 깊이를 고려한 유사도 임계값을 구하기 위한 연구를 계속 수행해 나갈 것 이다. 그리고 개념간의 가중치 조절과 개념-관계 유사도 간의 가중치 조절 역시 응용 시스템에서 자동적으로 조절이 가능하도록 연구할 것 이다.

OWL Web Ontology Language Overview" http://www.w3.org/TR/owl-features/, W3C Recommendation 10 February 200