확장된 파스 트리 커널을 이용한 온톨로지 표절 검출

손정우 박성배 김상수O
경북대학교 전자전기컴퓨터학부
기계 학습 연구실
{jwson, sbpark, sskimO}@sejong.knu.ac.kr

Detection of Ontology Plagiarism Using an Augmented Parse Tree Kernel

JeongWoo Son Seong-Bae Park Sang-Soo KimO
Kyungpook National University
School of Electrical Engineering and Computer Science
Machine Learning Lab.

1. 서 론

온톨로지는 특정 영역에 대한 사람의 지식을 표현하고 있는 데이터 모델이다. 온톨로지 구축의 가장 큰 문제점은 많은 시간과 노력이 필요하다는 것이다. 이를 해결하기 위해 기존의 온톨로지를 확장하여 만드는 방법이 제시되었으나, 이는 전문가의 감수를 필요로 하기에 완전한 해결책이라 보기 어렵다. 온톨로지를 구축할 때, 가장 쉬운 방법은 기존의 온톨로지로부터 필요한 부분을 차용하여 쓰는 것이다. 하지만 이는 온톨로지를 구축한 기관의 지적 재산권을 침해하는 행위일 수 있지만 이를 검출할 수 있는 효과적인 방법이 없어 방지하거나 적발하기 힘들다. 본 논문에서는 컨볼루션 커널[1]을 이용한 허가없이 온톨로지를 차용하여 만든 표절 온톨로지를 검출할 수 있는 방법을 제안한다. 제안한 방법에서는 컨볼루션 커널의 한 종류인 파스 트리 커널[2]을 이용하여 두 온톨로지의 구조를 비교하고 온톨로지의 어휘가 가지는 유사도를 커널에 반영하여 더 정확한 표절 검출이 가능하도록 하였다.

2. 온톨로지 표절 검출

온톨로지는 클래스와 프로퍼티, 인스턴스들로 이루어진 하나의 그래프로 볼 수 있다. 그래프 상의 노드들은 클래스와 인스턴스들을 나타내며, 노드들을 연결하고 있는 연결선은 프로퍼티를 나타낸다. 파스트리 커널을 온톨로지에 적용하기 위해서는 온톨로지를 트리구조로 변환해야 한다. 그렇기 때문에 제안한 방법에서는 먼저 온톨로지를 트리로 변환한 후, 두 온톨로지의 유사도를 변환된 트리를 파스 트리 커널에 적용함으로써 계산한다.

2.1. 온톨로지 변환

제안한 방법은 먼저 온톨로지를 트리로 변환한다. 그래프 상의 연결선은 프로퍼티를 나타내기 때문에 트리로 변환하기 위해서는 특정 프로퍼티의 속성을 바꿀 필요가 있다. 온톨로지 상의 프로퍼티는 데이터 타입의 프로퍼티와 오브젝트 타입의 프로퍼티로 나눌 수 있다. 이 중 오브젝트 타입의 프로퍼티가 온톨로지를 그래프형태로 만드는 요인이다. 이와 같은 오브젝트 타입의 프로퍼티 중 subClassOf는 트리로 변환되더라도 유지된다. 이는 subClassOf 프로퍼티가 구조적 정보를 가지기 때문이다. 그 외의 오브젝트 프로퍼티는 연결된 클래스 혹은 인스턴스의 대표 어휘만을 가지는 데이터 타입의 프로퍼티로 변환된다. 그리고 각 인스턴스는 해당 클래스의 자식 노드로 위치하게 되며 클래스 간의 계층은 트리로 변환 되더라도 유지하게 된다.

2.2 파스 트리 커널

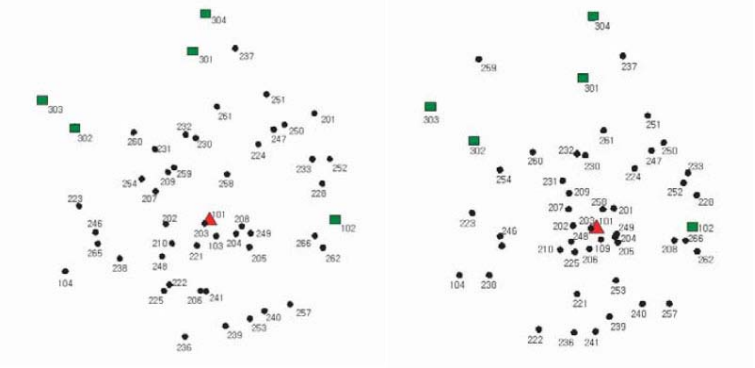
파스 트리 커널[2]은 컨볼루션 커널[1]의 하나로 파스 트리들을 다루는데 특화된 커널이다. 파스 트리 커널에서 벡터의 자질는 각 파스 트리에 나타날 수 있는 모든 subtree로 이루어진다. 이때, 각 자질의 값은 subtree의 빈도수로 할당된다. 두 파스 트리 간의 유사도는 이들 파스 트리 벡터 간의 내적을 이용하여 계산할 수 있다. 하지만 파스 트리 벡터를 생성하기 위해 모든 자질을 나열하는 것은 불가능하다. 이를 명시적으로 나열하지 않고 계산하기 위해 다이나믹 알고리듬을 이용한다.

온톨로지를 트리형태로 변환하여 커널을 이용하여 비교하는 것은 온톨로지의 구조에 중점을 두고 있다. 온톨로지가 가지는 의미 정보는 이러한 구조에 반영되고 있지만 구조를 비교하는 것만으로 의미정보를 충분히 반영할 수는 없다. 이와 같은 문제를 해결하기 위해 본 논문에서는 파스 트리 커널을 이용하여 두 노드를 비교할 때, 정확한 스트링 비교 대신 근사 스트링 비교(approximate string matching)을 이용한다. 스트링의 비교는 일반적으로 두 스트링 사이의 거리를 정의함으로써 구현되어진다. 본 논문에서는 이들 알고리듬 중, Resnik's distance[3], Lin's distance[4], Jiang-Conrath's distance[5]를 사용한다.

3. 실 험

제안한 방법을 검증하기 위해 두가지 데이터를 이용하였다. 첫번째 데이터는 웹상에서 구할 수 있는 3개의 온톨로지인 amino-acid, Biblio, BibTex를 이용하여 구축하였다. 이들 온톨로지는 크기가 비교적 작고, 간단한 구조를 이루고 있다. 모든 온톨로지는 약 60개 이하의 클래스와 150개 이하의 인스턴스로 이루어져 있다. 세가지 온톨로지 중, Biblio와 BibTex는 같은 개념들로 이루어져 있으며 amino-acid는 완전히 다른 온톨로지이다. 나머지 데이터는 OAEI(Ontology Alignment Evaluation Initiative) 2005에서 사용된 것이다. 이 데이터는 51개의 온톨로지로 구성되어있다.

온톨로지를 이루고 있는 클래스, 인스턴스, 프로퍼티 중 어떤 부분이 표절 검출에 있어 중요한지를 알아보기 위해



(a) 정확한 비교 방법 (b) Jiang-Conrath's distance

그림 1. 비교 방법에 따른 OAEI 2005 데이터의 분포

4가지 형태(원본 온톨로지, 서브클래스정보만 있는 온톨로지, 프로퍼티 정보를 뺀 온톨로지, 인스턴스 정보를 뺀 온톨로지)로 하나의 온톨로지를 변형하여 원본 온톨로지와 비교하였다. 실험에서는 서브 트리의 깊이를 3으로 제한하였다. 실험 결과, 온톨로지 스크마(schema)만을 이용할 경우 표절을 검출하는 것이 어려움을 알 수 있었다.

OAEI 2005 데이터를 이용한 실험결과 구조를 수정하지 않고 노드 상에 나타난 레이블 값만을 변화시킨 경우, 유사도 값이 상대적으로 높게 나타났으며, 구조를 수정하더라도 유사도 값이 완전히 다른 온톨로지인 102 온톨로지나 독립적으로 만들어진 3xx 온톨로지에 비해 높게 나타났다. 이는 정확한 스트링 비교 방법을 이용했을 때와 Jiang-Conrath's distance를 이용했을 경우 모두에서 나타나는 결과이다. 하지만 Jiang-Conrath's distance를 이용할 경우 유사도 값이 상대적으로 높게 나타났다. 그림 1은 정확한 비교 방법과 Jiang-Conrath's distance를 이용했을 경우의 유사도 값에 따른 온톨로지들의 분포를 보여준다. 사각형으로 표시된 온톨로지는 102 온톨로지와 3xx 온톨로지로 다른 데이터와 달리 101 온톨로지로부터 멀리 떨어져 있음을 알 수 있다.

4. 결 론

본 논문에서는 온톨로지 표절 검출을 위한 새로운 방법을 제안했다. 제안한 방법은 두 개의 복잡한 데이터 구조를 비교하기 위해 파스 트리 커널을 이용하였다. 먼저, 온톨로지를 트리 형태로 변환한 후, 두 온톨로지 사이의 유사도를 파스 트리 커널을 이용하여 계산하였다. 제안한 방법을 검증하기 위해 두가지 데이터 집단에 대한 실험을 하였다. 실험 결과, 온톨로지 표절 검출 시, 스키마 만을 이용한 표절 검출은 힘듬을 알 수 있었다. OAEI 2005 데이터를 이용한 실험 결과 원본 온톨로지로부터 변환된 온톨로지가 상대적으로 높은 순위의 유사도를 보였다. 이와 같은 결과는 제안한 방법이 온톨로지 표절을 검출할 수 있음을 의미한다. 제안한 방법은 온톨로지 사이의 유사도를 측정할 수 있는 방법으로 볼 수 있다. 이를 활용할 수 있는 분야로 컨셉 매칭(Concept matching)을 둘 수 있다. 제안한 방법을 이용하여 작은 단위의 온톨로지 사이의 유사도를 측정하여 컨셉 매칭에 적용할 수 있는 방법을 향후 연구로 두고 있다.

참 고 문 헌

[1] D. Haussler, Convolution Kernels on discrete structures. Technical report, UCS-CRL-99-10, UC Santa Cruz, 1999.

[2] M. Collins and N. Duffy, Convolution Kernels for natural language. In *Procedings of the 15th Neural Information Processing Systems*, pages 625-632, 2001.

[3] P. Resnik, Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* pages 448-453, 1995.

[4] D. Lin, An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* pages 296-304, 1998.

[5] J. Jiang and D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics.* pages 19-33, 1997.