개념적 거리와 밀도를 고려한 온톨로지 기반의 코스웨어 분류

*조미영, *최창, **김판구

*조선대학교 전자계산학과
**조선대학교 컴퓨터공학과
irune@chosun.ac.kr, enduranceaura@gmail.com, pkkim@choun.ac.kr

Courseware Classification using Conceptual Distance and Density based on Ontology

*Miyoung Cho, *Chang Choi, **Pankoo Kim
*Dept. of computer science, Chosun university
**Dept. of computer engineering, Chosun university

요 약

인터넷의 비약적인 발전으로 많은 강의 자료가 존재하게 되었으며, 어느 누구나 손쉽게 강의 자료를 구할 수 있게 되었다. 하지만 사용자는 단순히 많은 정보만을 원하는 것이 아니라 정확한 정보를 얻기를 원한다. 이에 본 논문에서는 기존의 단어 빈도수 기반의 분류 방식이 아닌 개념적 분류 방식으로 온톨로지를 이용하여 코스웨어를 분류해보고자 한다. 온톨로지로는 어휘적 온톨로지의 일종인 WordNet의 과목에 대한 계층적 구조를 활용하였다. 실험 데이터로는 강의 자료 중 파워포인트로 작성된 코스웨어를 이용하였으며, 코스웨어의 메타데이터들과 과목들간의 개념적 거리 및 밀도를 측정하여 코스웨어를 분류하였다. 또한 WordNet상의 어휘 확장을 통하여 분류과목 확장이 가능함을 보였다.

1. 서 론

인터넷의 비약적인 발전으로 많은 강의 자료가 존재하게 되었으며, 어느 누구나 손쉽게 강의 자료를 구할 수 있게 되었다. 하지만 검색된 정보 중 사용자가 진정으로 원하는 정보를 찾기란 쉽지 않다. 즉 사용자는 단순히 많은 정보만을 원하는 것이 아니라 의미적으로 정확한 정보를 얻기를 원한다. 이에 본 논문에서는 의미 기반 및 개념 기반 검색으로 개념간의 관계들로 표현된 온톨로지를 이용하고자 한다.

언어 온톨로지(Linguistic Ontology)의 일종인 WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년도 중반부터 프린스톤 대학 인지과학연구실이 구축해온 영어어휘 데이터베이스이다. WordNet은 인간의 어휘지식을 모방하여 다의성과 동의관계를 이용하였으며, 의미를 최대한 정확히 표현하고 있을 뿐만 아니라 개념간의 관계 표현 등을 통해 개념을 계층적으로 표현하고 있다. 또한 각 단어들은 synset이라는 동의어의 집합을 기준으로 노드들간의 관계를 이용하여 정의하고 있다. 이에 본 논문에서는 WordNet상의 과목 분류 체계를 기반으로 코스웨어를 과목별로 분류해 보고자 한다.

이를 위해, WordNet의 계층적인 특성을 활용한 개념적 거리 측정으로 코스웨어의 메타데이터와 분류하고자 하는 과목(영어, 수학 등)들의 유사성을 측정하였다. 그리고 측정된 유사성 값을 이용하여 개념적 밀도를 구해 적합한 과목으로 분류하였다. 특히, 본 논문에서는 코스웨어 데이터로 웹 문서가 아닌 파워포인트 파일을 이용하였고, 메타데이터는 각 문서의 목차와 제목을 통해 구성하였다.

논문의 구성은 다음과 같다. 2장에서는 코스웨어의 개념 및 분류 방법에 대해 간략히 소개하고, 3장에서는 코스웨어 분류를 위한 개념적 거리 및 밀도의 정의와 실제 코스웨어를 어떻게 분류할 수 있는지에 대해 기술하였으며. 4장에서는 제기한 문제점 해결방안으로 개념 추가 등을 통한 WordNet 확장으로 다른 여타 과목 분류와 비교하고, 마지막 5장에서 결론을 맺는다.

2. 관련연구

코스웨어(courseware)란 컴퓨터를 활용한 각종 교육과정 시스템에 사용되는 프로그램과 데이터를 통틀어 일컫는 용어로 본 논문에서는 정의된 과목들에 이러한 코스

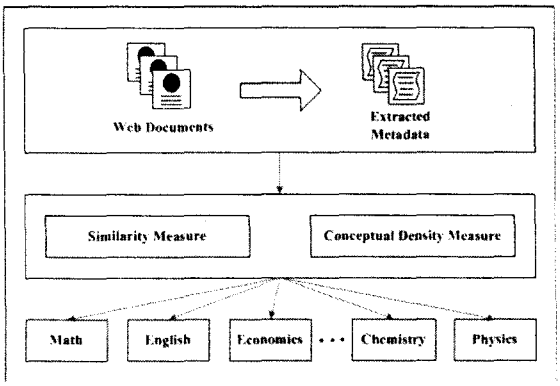
웨어를 분류하는 문제에 대해 다루어보았다. 기존의 웹상에서 문서분류 등의 코스웨어 분류를 위한 시도는 많이 있었다. 그 중 문서분류(document classification)는 정의된 클래스나 카테고리에 따라 문서를 분류하는 것을 의미한다. 이때 문서는 문자열의 집합으로 이루어져있고, 일정 수의 문자열로 구성된 단어들은 이 문서 기본 표현 단위로 사용하는 것이 적절하다. 따라서 문서분류를 위해서 대부분 문서들은 단어들의 집합(bag of words)으로 표현된다고 할 수 있다. 문서분류를 위해 문서의 모든 단어들을 사용하기에 어려움이 있으므로, 단어들 중에서 키워드를 정하고, 키워드들을 벡터로 구성하는 방법들이 있는데, 단어의 존재 유무를 이진벡터(binary vector)로 나타내는 방법과 가중치를 두어서 가중치 벡터(weight vector)표현하는 방법 단어의 빈도수와 역문서빈도수를 이용한 TFIDF 표현법 등이 있다[1].

또한 특징어 선택(feature selection)은 문서를 표현하는 특징벡터의 차수(dimentionality)를 줄임으로써 분류작업의 효율성을 높이기 위한 처리과정이다. 특징어를 선택하는 방법은 일정 척도(measure)를 기준으로 단어의 중요도나 관련성을 평가하여 불필요하거나 중요도가 낮은 특징어를 필터링(filtering)하여 중요도가 높은 단어들을 모으는 방법 등이 있으며, 여과방법으로 TF-IDF weighting 나 LSI(Latent Semantic Indexing) 등이 있다[2,3]. 기존 문서분류에 가장 많이 이용되는 대표적인 학습법으로는 Bayesian algorithm, k Nearest Neighbor algorithm and Decision Tree algorithm 등이 있다[4].

이렇듯 대부분 문서의 단어 빈도수 등을 이용한 특징어 선택을 통해 대표 단어를 추출하고, 이를 학습함으로써 분류가 이루어졌다. 그러나 이는 단순 학습에 의한 분류로 개념적인 접근이 필요하다. 이에 본 논문에서는 개념적 관계 표현으로 이루어진 WordNet을 이용하여 개념적 접근을 시도하였다.

3. 제안한 방법

2장에서 보듯이 기존의 방법이 문서내의 단어들의 빈도수 등을 고려한 문서 분류라면 본 논문에서 제안한 방법은 WordNet을 이용한 개념적 접근이라 할 수 있겠다. WordNet을 이용한 개념간 거리, 밀도, 정보량, 깊이 정보 등을 이용할 수 있으나, WordNet상의 과목 도메인 특징으로 인하여 본 논문에서는 개념적 거리와 밀도만을 고려하며, 제안한 코스웨어 분류 방법은 [그림 1]과 같다.

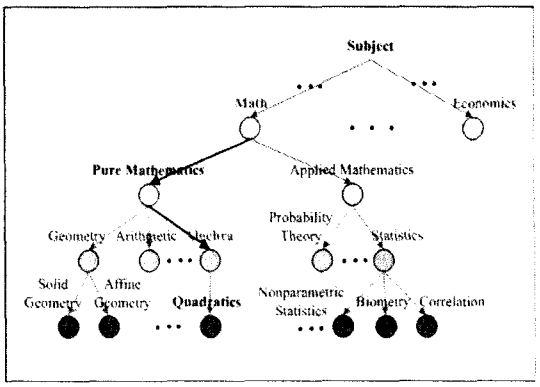


[그림 1] 제안한 방법

여기서 메타데이터 추출은 파워포인트 강의 자료의 목차와 제목만을 고려하도록 한다. 추출된 메타데이터와 분류하고자 하는 과목들 간의 개념적 거리 측정 및 밀도 측정을 통해 적합한 과목으로 분류한다. 여기서 주제 분류를 위해 우리는 목차만을 고려하였다. 본 논문에서 제안한 방법은 WordNet이라는 개념적 데이터베이스를 이용하므로, 기존의 방법에 비해 좀 더 개념적 접근법이라 할 수 있다.

3.1 개념적 거리

개념적 거리란 측정하고자 하는 두 개념간의 최단 거리라고 정의할 수 있다. 특히 WordNet은 어휘간 계층적 구조뿐만 아니라 반의어, 유의어 관계 등 단어들 간의 관계를 표현을 이용한 개념적 거리 측정이 가능하다. 예를 들어, [그림 2]는 WordNet상에서 과목에 대한 개념들간의 계층구조이다.



[그림 2] WordNet상의 과목 계층구조

계층적 구조를 이용한 개념간 거리 측정 즉, 유사성 측정은 노드의 정보량을 이용한 노드 기반의 방법과 최단거리를 이용한 에지 기반의 방법으로 나눈다. [그림 2]에서 보듯이 WordNet상에서 과목 도메인의 계층구조에서 개념별로 깊이 정보는 큰 차이를 보이지 않으므로 정보량을 이용한 개념간의 유사성 측정은 어려움이 있다. 이에 본 논문에서는 최단거리를 이용하는 후자의 방법을 택하였다. [그림 3]은 실제 WordNet을 이용한 과목분류를 위한 개념간 거리측정 과정을 보여주고 있다.

```
HyperTree:  *Root*#n#1  entity#n#1  abstract_entity#n#1
abstraction#n#6  psychological_feature#n#1  cognition#n#1
content#n#5  knowledge_dom-ain#n#1  discipline#n#1
science#n#1  mathematics#n#1  pure_mathemati-cs#n#1
algebra#n#1

HyperTree:  *Root*#n#1  entity#n#1  abstract_entity#n#1
abstraction#n#6  psychological_feature#n#1  cognition#n#1
content#n#5  knowledge_dom-ain#n#1  discipline#n#1
science#n#1  mathematics#n#1

LCS: mathematics#n#1  Path length: 3.

algebra  math  0.333333333333333
```

[그림 3] 'algebra'와 'math'간 유사성 측정

개념적 거리를 측정하고자 하는 두 개념('algebra'와 'math')의 각각 계층 구조(IS_A 관계)를 고려하여 최상위 노드부터 해당 노드까지의 경로를 찾는다. 그리고 두 단어가 동시에 포함하고 있는 공통 부모 중 가장 하위의 단어를 찾아 이를 중심으로 두 단어와의 거리를 고려하여 최단거리를 측정한다. 측정된 거리 값은 두 개념간의 유사도가 아니기 때문에 유사성 값으로 변환해야 한다. 거리는 유사성과 반비례 관계이므로 이를 고려하여 유사성 값을 구한다.

[표 1]은 개념적 거리 측정을 이용한 'algebra'와 각 과목들과의 유사성 측정값이다. 즉, 각 코스웨어에 대한 메타데이터의 단어와 각 과목 단어들과의 유사성을 측정한다. 측정결과 'algebra'는 'math' 과목과 가장 유사한 값을 보였다.

[표 1] algebra와 과목들간 유사성 측정값

keyword	subject	Similarity value
Algebra#n#1	Math#n#1	**0.333333333333333**
Algebra#n#1	English#n#3	0.142857142857143
Algebra#n#1	Physics#n#1	0.166666666666667
Algebra#n#1	Chemistry#n#1	0.166666666666667
Algebra#n#1	Economics#n#1	0.166666666666667

[표 2]는 개념적 거리를 이용한 실제 파워포인트 자료의 과목 분류를 위한 유사성 측정값이다.

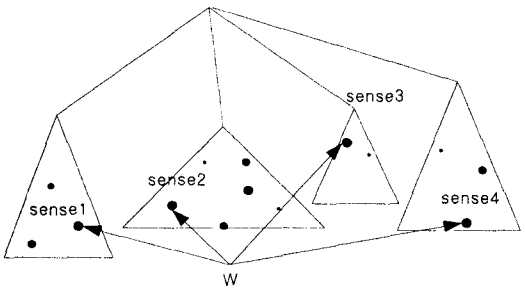
[표 2] 메타데이터와 과목들간 유사성 측정값

Word	Subject				
	Chemistry	Physics	Economics	Math	English
Algebra	0.06667	0.166667	0.166667	**0.333333**	0.142857
Connections	0.25	0.090909	0.090909	0.1	0.125
Professional	0.1	0.071429	0.071429	0.076923	0.1
Development	0.125	0.083333	0.083333	0.090909	0.125
Model	0.125	0.125	0.125	0.142857	0.142857
Links	0.090909	0.066667	0.066667	0.071429	0.090909
Middle	0.142857	0.1	0.1	0.111111	0.111111
School	0.125	0.083333	0.083333	0.090909	0.142857
Teachers	0.1	0.1	0.1	0.111111	0.111111

좌측의 단어들은 실제 코스웨어의 목차에서 추출된 메타데이터 정보이다. 이와 같이 개념적 거리를 이용하여 유사성 값을 구한결과 'math'와 'algebra'가 0.3333이라는 최대값을 가지므로, 이 코스웨어는 'math' 과목에 가깝다는 것을 알 수 있다. 여기에 정확도를 높이기 위하여 단순히 최대값만이 아닌 다른 메타데이터 고려한 개념적 밀도 개념이 필요하다. 또한, 개념적 분류가 가능하지만 WordNet이라는 모든 도메인에 대해 다룬 일반적인 링귀스틱 온톨로지의 경우 동음이의어라는 문제점이 있다. 특히 WordNet의 기본 단위는 단어가 아닌 synset으로써 의미가 기본단위로 사용된다. 즉 동음이의어 처리를 위해 같은 의미의 단어들을 synset이라는 집합으로 동일하게 처리하고 있다. 다음 3.2절에서는 이러한 문제점을 해결하기 위한 개념적 밀도에 대해 살펴보도록 하겠다

3.2 개념적 밀도

본 절에서는 개념적 거리 측정값을 바탕으로 개념적 밀도를 통한 동음이의어 문제 해결에 대해 논하고자 한다. 예를 들어, [그림 4]에서 단어 W는 4가지의 의미를 가지고 있으며 각각은 WordNet의 Sub-hierarchy에 포함되며 이는 각 의미(sense)별 영역이 된다. 영역내의 각 점은 문맥내 단어들과 단어 W의 의미이다. 그리고 영역내 존재하는 노드의 수를 더하는 방법으로 개념적 밀도를 구한다. 개념적 밀도가 클수록 즉, 영역이 클수록 각 문맥에서 정확한 의미를 표현할 가능성이 크다. [그림 4]의 경우 sense 2가 찾고자 하는 문맥에 가장 알맞은 의미이다.



[그림 4] 개념적 밀도 정의

WordNet상에서 각 W_s의 의미를 따라 영역 R을 생성한다. 그리고 각 의미별 R_s를 기반으로 한 D_s는 다음과 같다.

$$D_s = \sum_{i=1}^{n} N_i \text{---------- (1)}$$

여기서 n은 영역 내 노드의 총 수이다. 위에서 언급했듯이 밀도는 영역(R)에 비례한다. [표 3]은 개념적 밀도를 고려한 과목분류이다.

[표 3] 개념적 밀도를 고려한 분류

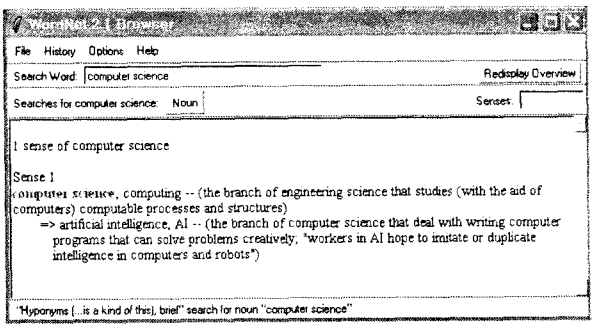
Keyword		Subject							
		A		B	C	D	E		
		#1	#2	#1	#1	#1	#1	#2	#3
제목	Unit	0.00	0.20	0.10	0.10	0.11		0.14	
	basics		0.14	0.09	0.09	0.10			0.10
Sum		0.00	0.34	0.19	0.19	0.21	0.00	0.14	0.10
목차	Basics		0.14	0.09	0.09	0.10			0.10
	Geometry	0.17		0.17	0.17	0.33			0.14
	Basics		0.14	0.09	0.09	0.10			0.10
	Definitions		0.14	0.07	0.07	0.07		0.09	
	Theorem		0.14	0.13	0.13	0.14			0.14
	Proofs		0.17	0.10	0.10	0.11		0.13	
	Reminde		0.20	0.13	0.13	0.14			0.14
	Theorem		0.14	0.13	0.13	0.14			0.14
	Solving		0.08	0.06	0.06	0.07		0.08	
	Triangles		0.14	0.08	0.08	0.08		0.11	
	Pythagoras		0.09	0.07	0.07	0.07		0.09	
	Triples		0.08	0.08	0.08	0.09			0.09
	Fermat		0.09	0.07	0.07	0.07		0.09	
	Theorem		0.14	0.13	0.13	0.14			0.14
	Proofs		0.17	0.10	0.10	0.11		0.13	
	Points		0.20	0.11	0.11	0.13		0.14	
	Segments		0.11	0.08	0.08	0.08		0.11	
	Rays		0.13	0.06	0.06	0.07		0.08	
	Lines		0.17	0.08	0.08	0.09		0.14	
	Formulas		0.17	0.11	0.11	0.13			0.13
	Circles		0.14	0.08	0.08	0.09		0.17	
	Strategy	0.17		0.17	0.17	0.20			0.20
	Composite	0.10		0.10	0.10	0.11			0.11
	Figures		0.14	0.11	0.11	0.13			0.13
Sum		0.43	2.94	2.38	2.38	2.80	0.00	1.36	1.57

(A-chemistry, B-physics, C-economics D-math E-english)

[표 3]에서 측정값은 메타데이터와 과목 단어들간의 유사성 측정으로 과목 단어는 'English'처럼 과목만을 의미할 뿐만 아니라 언어 등 다른 의미를 가지고 있다. [표 3]에서는 chemistry의 #1의미와 english은 #3의미가 과목 의미로 과목의 의미들과 키워드들을 비교한다면 가장 큰 밀도값을 가지는 것은 'math'과목이다.

4. 온톨로지 확장

4장에서는 3장의 실험 결과를 토대로 WordNet의 어휘 확장을 통한 분류 과목의 확장에 대해 논하고자 한다. 3장의 실험에서 다른 과목의 경우 특정 분야에서 존재하는 단어가 WordNet에서 존재하지 않는 단어로 인하여 효율성이 떨어지는 결과가 나타났다. [그림 5]는 WordNet에서의 'Computer Science'를 검색했을 때, 하위개념으로 단지 'AI' 하나만 존재하는 것을 알 수 있다. 이와 같이 WordNet에 존재하지 않는 특정 단어 대한 보완의 필요성이 존재하고 이를 위하여 개념과 단어를 추가하여 어휘를 확장하였다.



[그림 5] WordNet상에서 'computer science'

```
<Synset.N rdf:ID="computer_science">
                    ...
        <rdfs:subClassOf rdf:resource="#engineering_science"/>
        <part_of rdf:resource="#information_science"/>
        <category rdf:resource="#computer"/>
</Synset.N>

<Synset.N rdf:ID="artificial_intelligence">
        <gloss>the branch of computer science that deal with writing
            computer programs that can solve problems creatively;
            "workers in AI hope to imitate or duplicate
            intelligence in computers and robots"</gloss>
        <word rdf:resource="&wen:artificial_intelligence.n" />
        <word rdf:resource="&wen:AI.n" />
        <rdfs:subClassOf rdf:resource="#computer_science"/>
</Synset.N>
```

[그림 6] WordNet OWL 파일

[표 3] 확장된 온톨로지를 이용한 분류 결과

Word	Subject				
	Chemistry	...	Computer_Science	Math	English
Database#n#1	0.166667	...	0.066667	0.066667	0.1
Database#n#2	0.142857	...	0.5	0.166667	0.166667

기존 WordNet에서의 'Database'는 Database#n#1로 놓고 'Computer_Science' 하위에 'Database'과목을 Database#n#2로 생성하였다. 그 결과 기존은 'Database'가 'Chemistry'에 가장 가까운 값을 나타내었으나 확장결과 'Computer_Science'에 가장 가까운 값을 나타내는 것을 볼 수 있다. 이와 같이 온톨로지에서 어휘 확장으로 인하여 다른 과목의 분류도 가능하다.

5. 결론

본 논문에서는 개념적 코스웨어 분류를 위해 WordNet의 과목 분류 체계를 기반으로 개념적 거리와 밀도를 이용하여 웹 데이터(강의자료)를 과목별로 분류하였다. 하지만 WordNet을 그대로 이용할 경우 특정 분야에서 쓰는 단어들이 포함되어 있지 않기 때문에 과목에 제한이 있었다. 이에 어휘 및 개념 확장을 통해 다른 세부적인 과목도 분류 가능함을 보였다.

Acknowledgement

"본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2006-C1090-0603-0040)

참고문헌

[1] Sahami, Mehran, "Using Machine Learning to Improve Information Access", a Dissertation Stanford : Dept. of Computer Science, Stanford University, 1998.

[2] Han, Jiawei and Micheline Kamber, "Data Mining : Concepts and Techniques," New York : Morgan Kaufmann, 2001.

[3] R. Dolin, J. Pierre, M. Butler, and R. Avedon. Practical Evaluation of IR within Automated Classification Systems. Eighth International Conference of Information and Knowledge Management, 1999.

[4] Y. Yang and X. Liu. A re-examination of text categorization methods. In Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 42-49, 1999.

[5] John M. Pierre, "On the Automated Classification of Web Sites"

[6] http://wordnet.princeton.edu

[7] http://www.keris.or.kr/datafiles/main/

[8] William D. Lewis "Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity", The University of Arizona Working Papers in Linguistics, 2002

[9] Eneko Agirre, German Rigau, "Word Sense Disambiguation using Conceptual Density", International Conference On Computational Linguistics, 1996

[10] Terje Brasethvik and Jon Atle Gulla , "A Conceptual Modeling Approach to Semantic Document Retrieval", LNCS Volume 2348, 2002