「정보시스템연구」 제21권 제3호
한국정보시스템학회
2012년 9월, pp. 137~161
http://dx.doi.org/10.5859/KAIS.2012.21.3.137

시맨틱 텍스트 마이닝을 위한 온톨로지 활용 방안*

유은지** · 김정철*** · 이춘열**** · 김남규*****

〈목 차〉

Ⅰ. 서론
Ⅱ. 관련 연구
Ⅲ. 시맨틱 텍스트 마이닝 절차
3.1 연구 범위 및 모형
3.2 온톨로지를 이용한 의미적 모호성 해결
Ⅳ. 온라인 쇼핑몰의 이탈 고객 예측 실험
4.1 실험 개요 및 데이터 소개
4.2 실험 설계
4.3 실험 결과 및 해석
Ⅴ. 결론
참고문헌
<Abstract>

Ⅰ. 서 론

최근 빅데이터(Big Data) 관련 기술은 IT 분야의 핵심 키워드로 자리 잡고 있다. Gartner 그룹은 향후 유망 기술을 분석한 2011년 보고서(Gartner, 2011)에서 빅데이터 관련 기술이 향후 2~5년 내에 IT 분야의 주요 기술로 자리 잡을 것으로 예상하였으며, 많은 기업 및 교육기관에서 빅데이터 분석 전문가인 데이터 과학자(Data Scientist)를 양성하기 위한 준비에 노력을 기울이고 있다. 빅데이터란 데이터의 양이 너무 방대해서 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형 데이터를 의미하며(Mckinsey, 2011). 빅데이터의 특성은 일반적으로 데이터의 용량(Volume), 속도(Velocity), 다양성(Variety)에 의해 규정된다. 이러한 빅데이터 기술에 대해 관심이 급증한 원인은 크게 다음의 세 가지 현상(유지연, 2012)에서 찾을 수 있다. 우선 스마트폰을 비롯한 다양한 모바일 기기로부터 카메라, RFID 리더 등을 통해 수많은 비정형데이터가 생성되고 있는 현상과 클라우드 서비스의 확산을 통해 개인과 조직의 데이터가 한 곳으로 축적되는 현상을 그 원인으로 들 수 있다. 마지막 세 번째 원인은

* 본 연구는 (주)신세계의 지원으로 수행됨
** 국민대학교 비즈니스IT전문대학원 석사과정, nlj0123@naver.com
*** 국민대학교 비즈니스IT전문대학원 박사과정, kjcacc@mnd.go.kr
**** 국민대학교 경영정보학부 교수, cylee@kookmin.ac.kr
***** 국민대학교 경영정보학부 조교수(교신저자), ngkim@kookmin.ac.kr

Twitter, Facebook 등의 소셜미디어를 통해 유통되는 비정형데이터의 양이 급증한 것에서 찾을 수 있다. 즉 이미지, 사운드, 동영상, 텍스트 등 급증한 비정형데이터를 효율적으로 저장하고 분석하기 위한 새로운 도구 및 방법론이 필요하게 되었다는 것이다. 이러한 비정형데이터에 대한 분석 중 텍스트 데이터에 대한 분석은 소셜미디어 메시지 분석, 쇼핑몰 Q&A 분석, 뉴스 분석 등 다양한 분야에서 활용되고 있다.

텍스트 데이터 분석은 주로 텍스트 마이닝(Text Mining), 문서 클러스터링(Document Clustering), 문서 요약(Text Summarization) 등의 다양한 용어로 불려왔으나, 이들 중 텍스트 마이닝이 다른 용어들을 포함하는 가장 대표적인 용어로 최근 널리 사용되고 있다. 텍스트 마이닝과 데이터 마이닝은 그 용어의 유사성에도 불구하고 본질적인 특성 면에서 상당한 차이점을 갖는다. 데이터 마이닝은 방대한 양의 데이터로부터 이전에 알려져 있지 않은, 잠재적으로 유용한 정보를 발견해내는 일련의 과정을 의미한다(손윤호 외, 2009; 홍태호와 김진완, 2006; Han and Kamber, 2006). 하지만 텍스트 마이닝의 경우 획득하고자 하는 정보는 이미 입력 텍스트에 명확하게 기술되어 있다(Witten, 2004). 즉 전통적인 데이터 마이닝과 달리, 텍스트 마이닝의 주요 관심은 이미 기술된 정보를 최대한 손실 없이 활용하여 분석에 적용하기 위한 과정에 집중되어 있다고 할 수 있다.

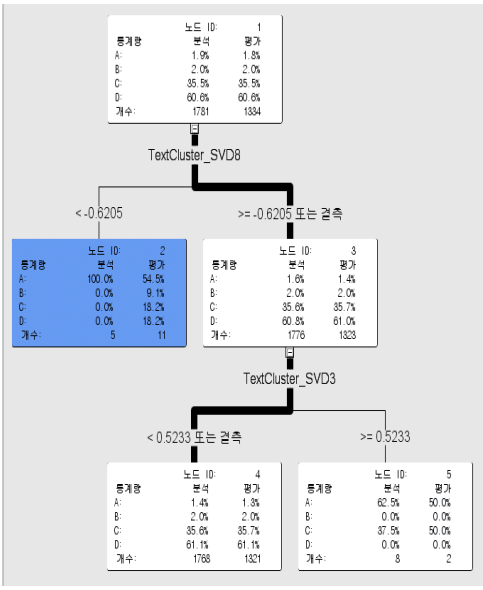
이처럼 데이터 마이닝과 텍스트 마이닝이 갖는 본질적인 차이점에도 불구하고, 많은 경우에 텍스트 마이닝은 넓은 의미의 데이터 마이닝의 주요 부분집합으로 인식되기도 한다. 즉 데이터 마이닝은 구간변수, 명목변수 등의 다양한 유형의 데이터를 분석에 활용할 수 있으며, 텍스트 마이닝은 분석 대상 데이터의 유형이 텍스트인 경우에 지나지 않는다는 것이다. 따라서 SAS 사를 비롯한 기존의 주요 데이터 마이닝 도구 제작 업체들은 기존의 데이터 마이닝 도구에 텍스트 마이닝 기능을 추가하여 배포하고 있다. 즉 기존의 도구에서는 입력 데이터에 텍스트 변수가 있는 경우 이를 분석 과정에서 사용할 수 없었지만, 텍스트 마이닝 기능이 포함된 마이닝 도구의 경우 텍스트 변수도 구간변수, 명목변수 등과 함께 분석에 활용할 수 있게 되었다. 예를 들면 SAS 사의 Enterprise Miner 최근 버전의 경우, 텍스트 입력에 대해 클러스터링 분석을 수행함으로써 각 단위 텍스트를 다수의 SVD(Singular Value Decomposition) 값으로 설명한다(그림 1). 이렇게 SVD 값으로 수치화된 텍스트 변수는 다른 변수들과 마찬가지로 의사결정트리, 신

클러스터

클러스터 ID	기술 용어	빈도	비율	조합 1	조합 2	조합 3	조합 4
1	'check engine light' +code +light +stay charged check cleared ...	169	9%	0.685312	0.216171	0.034954	-0.00238
2	'at all' 'oil leak' +leak +line +vibration +work mph oil remote s...	177	10%	0.075263	-0.14264	-0.00318	-0.14103
3	'cruise control' 'power outlet' 'rear wiper inop' +fuse +remote ...	95	5%	0.019913	-0.04397	0.08916	-0.23036
4	'rear spoiler' +replace bulb locked noise rear replaced rotors s...	369	21%	0.032793	-0.07046	0.095911	-0.17528
5	+align +clean +seat seats track row middle cable +move 3rd +...	55	3%	0.013258	-0.01272	0.045826	-0.07353
6	'third brake light' +brake +light leaking test third cracked shoc...	211	12%	0.256008	-.000701	0.072921	-0.0419
7	'fuel gauge' ac cd fuel gauge player work working guage +doo...	352	20%	0.026253	-0.02182	0.03784	-0.09372
8	'center console' +rattle +side center clean clips console conta...	251	14%	0.007781	-0.01361	0.017138	-0.03778
9	+balance +start car jumped pulling pulls right rotated tires ve...	102	6%	0.073868	-0.09066	0.09523	-0.21385

<그림 1> SAS Enterprise Miner의 텍스트 클러스터링 결과 예

경망분석 등 후속 모델의 입력으로 사용된다 <그림 2>.



<그림 2> 텍스트 입력의 SVD 값을 사용한 의사결정나무 예

하지만 텍스트 파싱, 필터링, 클러스터링 등의 기능은 주로 빈도수에 기반하여 동작하므로, 텍스트 분석 과정에서 용어의 시맨틱(Semantic)을 파악하지는 못한다는 한계를 갖는다. 이러한 문제는 초창기의 자연어 처리(Natural Language Process) 연구에서부터 지적되어 온 것으로, 결국 동음이의어(Homonym)과 이음동의어(Synonym)의 처리 문제로 귀결된다. 이음동의어의 경우를 예로 들면, 한 문서에서 Purchase라는 용어가 사용되고 다른 문서에서 Buy라는 용어가 사용되었을 경우 이 두 용어는 서로 전혀 다른 의미를 갖는 것으로 해석된다는 것이다. 물론 대부분의 텍스트 마이닝 도구는 이러한 한계를 극복하기 위해 유의어 사전을 제공하고 있다. 즉 유의어 사전에 (Purchase, Buy)의 쌍을 등록함으로써 두 용어가 동일한 의미로 사용된다는 정보를 분석 과정에서 이용할 수 있는 것이다.

하지만 이러한 수준의 유의어 사전은 매우 단순한 형태의 이음동의어 문제는 해결할 수 있지만, 다음과 같은 대부분의 경우에 적용되기는 어렵다는 한계를 갖는다. 예를 들어 *Address*와 *Speech*라는 두 용어가 사용되는 경우를 살펴보자. 우선 *Address*라는 용어는 주소라는 의미(Case 1)로 사용될 수도 있고, 연설이라는 의미(Case 2)로 사용될 수도 있다. 또한 *Speech*라는 용어도 연설이라는 의미(Case 3)로 사용될 수도 있고 언어능력이라는 의미(Case 4)로 사용될 수도 있다. 즉 두 용어 *Address*와 *Speech*는 동일한 의미로 사용될 수도 있지만 (Case 2 와 Case 3), 서로 다른 의미로 사용될 수도 있기 때문에 이들 용어를 단순하게 유의어 사전에 등록하는 것은 바람직하지 않다. 게다가 유의어 사전은 동음이의어 문제의 해결과는 아무 관계가 없다. 따라서 동음이의어와 이음동의어의 문제의 해결을 위해서는 단순한 수준의 유의어 사전이 아닌 문장 내에서 용어의 시맨틱을 파악할 수 있는 장치가 마련되어야 한다.

따라서 본 연구에서는 온톨로지(Ontology)를 활용하여 의미적 모호성을 해결함으로써 텍스트 마이닝 결과의 품질을 향상시킬 수 있는 방안을 제시하고자 한다. 온톨로지는 의미의 삼각관계(Meaning Triangle)(Maedche et al., 2001)에서 하나의 기호(용어)가 여러 개념에 대응되는 기호작용의 오류를 해결하기에 적합한 체계라는 측면에서, 본 연구에서 다루고 있는 텍스트 문서의 의미적 모호성 해결에 매우 적합한 도구가 될 것으로 기대한다. 정보시스템 분야에서 온톨로지를 다루는 대부분의 연구는 BWW

프레임워크(Bunge, 1977, 1979; Wand and Weber, 1993, 1995)의 구성요소(Constructs) 중 주로 사물(Thing)과 프로퍼티(Property)에 대한 논의에 집중하고 있다. 하지만 본 연구에서는 온톨로지 설계 단계에서 제약조건(Restriction)이 갖는 역할을 매우 중요하게 인식하고, 텍스트 마이닝 입력의 의미적 모호성 해결을 위해 제약조건을 활용하는 방안을 제안하고자 한다.

본 논문의 이후 구성은 다음과 같다. 다음 절인 2절에서는 본 연구의 이론적 배경이 되는 텍스트 마이닝, 온톨로지, 그리고 텍스트 마이닝과 온톨로지를 접목시킨 기존 연구 성과를 요약하여 제시한다. 그리고 3절에서는 온톨로지 기반의 시맨틱 텍스트 마이닝을 위한 연구 모형과 절차를 소개한다. 다음 절인 4절에서는 연구에서 제안하는 방안의 실무 적용 가능성을 평가하기 위해 국내 한 대형 온라인 쇼핑몰의 고객, 주문, 문의게시판 데이터에 대해 텍스트 마이닝 분석을 수행하고, 이를 통해 해당 쇼핑몰 고객의 이탈을 예측하고자 한다. 마지막 절인 5절에서는 본 연구의 기여 및 한계, 그리고 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

텍스트는 현실 세계에서 정보를 교환하거나 표현하는 방법으로 가장 널리 사용되는 수단이다(Witten, 2004). 따라서 많은 연구자들은 풍부한 정보를 담고 있는 텍스트에 대한 분석을 통해 의미 있는 지식을 발견하기 위한 노력을 기울여왔다. 텍스트 마이닝은 대용량의 텍스트로부터 구문 분석을 통해 유용한 정보를 추출하는 과정(Hearst, 1999; Sebastiani, 2002)으로 이해될 수 있다. 텍스트 마이닝의 활용 분야는 텍스트 형태로 된 정보를 사용하는 모든 분야를 아우를 정도로 매우 다양하다. 예를 들면, 특정 기사(Article)의 원문(Source)를 파악하기 위한 연구(Metzler et al., 2005), 특정 범죄와 다른 범죄들 간의 유사성 측정을 통해 새로운 범죄를 발견하기 위한 연구(Fan et al., 2006), 텍스트 범주화(Categorization)를 통해 비구조적 저장소(Repository)를 구조화하기 위한 연구(Sebastiani, 2006) 등을 들 수 있다. 특히 최근에는 사용자의 성향이 다양한 소셜 미디어를 통해 텍스트로 표현되고 저장됨에 따라, 소셜 미디어 데이터에 대한 텍스트 마이닝을 통해 기존의 데이터 분석에서는 찾을 수 없었던 새로운 유형의 지식을 찾기 위한 시도들이 활발하게 이루어지고 있다(김인현, 2012; 최광선, 2012).

텍스트 마이닝은 데이터 마이닝(Data Mining), 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적(Topic Tracking) 등의 분야의 기술을 종합적으로 활용한다(Mooney and Bunescu, 2006; Rijsbergen, 1979). 특히 자연어 처리 기술은 텍스트 마이닝의 성패를 좌우하는 핵심 기술이라고 할 수 있으며, 자연어 처리의 대상이 되는 텍스트는 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현된다<표 1>.

SAS Enterprise Miner, IBM SPSS Modeler, R, Linguamatics I2E 등 대부분의 데이터 마이닝 소프트웨어는 텍스트 마이닝 기능을 지원하고 있다. SAS Enterprise Miner의 경우 분석 단위는 각 문서가 되는데, 여기서 문서란 제목, 요약, 본문, 문서전체 등 텍스트로 기술된 모든 데이터를 일컫는 폭넓은 개념으로 사용된다. 기본

<표 1> 텍스트 표현을 위한 접근법(수정 인용)(Stanvrianou et al., 2007)

대상	표현 형태	목적
Sentences	term-by-sentence matrices	Text mining
words, phrases, concepts	association rules	Representation of medical texts
words and concepts	combination of bag-of-words and concept hierarchy	Text clustering and classification
concepts	Hierarchy	Feature extraction
phrases	n-grams	Text categorization
concepts	concept hierarchy	Automatic acquisition of a taxonomy
word senses	sense-based vector	Text categorization
phrases	n-grams	Text learning
words and compounds	Vector	IR
noun phrases	Tree	Book indexing
words	Vector	IR
words	Tree	Semantic similarity for IR

적으로 각 문서는 벡터공간모델(Vector Space Model)(Albright, 2006; Salton et al., 1975)을 이용하여 표현되며, 각 문서에 사용된 용어(Term)의 빈도에 따라 해당 문서의 주제 및 특성이 요약된다. 대부분의 경우 용어의 단순 빈도수보다는 TF-IDF(Term Frequency-Inverse Document Frequency)(Han and Kamber, 2006)에 근거한 분석을 주로 수행한다. 이 개념은 어떤 문서에서 용어 A와 B가 동일한 빈도수로 발생하였을 때, A가 다른 전체 문서에서도 일반적으로 자주 발생하는 용어라면 그 문서에서 더욱 중요하게 사용되는 용어는 A가 아니라 B라는 인식에 기초한다. 빈도수에 기반한 분석에서 각 문서는 용어 수만큼의 차원을 갖게 되며, "문서 X 용어" 로 표현된 행렬의 각 셀에 각 문서에서 해당 텍스트가 나타난 빈도수를 기재함으로써 모든 문서를 행렬화할 수 있다. 하지만 문서에 포함된 용어의 수는 일반적으로 매우 많기 때문에, 문서 간 유사성 측정을 위해 각 문서는 SVD(Singular Value Decomposition) 기법을 통해 차원이 축소된다(Albright, 2006). 상용 텍스트 마이닝 도구(SAS, 2010)는 이러한 이론을 기본으로 하여 파싱, 필터링, 클러스터링 등의 작업을 수행하게 된다. 이러한 작업의 결과는 그 자체로도 문서 분류, 토픽 추출 등의 분야에 사용될 수 있을 뿐 아니라, 기존의 마이닝 분석 모델인 의사결정나무, 인공신경망 등 후속 분석의 입력으로 사용될 수도 있다.

하지만 현재까지 논의된 텍스트 마이닝 이론 및 관련 상용화 도구들은 대부분 문장의 구조적(Syntactic) 분석에 의존하고 있으며, 문장 내에서 각 용어가 갖는 의미적 차원의 분석은 아직

효과적으로 이루어지지 못하고 있다(Stanvrianou et al., 2007). 의미 차원의 텍스트 마이닝을 위한 실마리는 온톨로지(Bunge, 1977, 1979), 토픽맵(Topic Map)(정윤수 외, 2009) 등의 지식표현 모델에서 찾을 수 있다. 온톨로지에 대한 연구의 기원은 고대 그리스 철학자들의 주된 관심이었던 존재론까지 올라갈 정도로 역사가 깊지만, 최근 시맨틱 웹(Antoniou and Harmelen, 2008)에 대한 관심의 증가로 온톨로지의 이론 및 활용분야가 새롭게 주목받고 있다. 고대 존재론에서는 "분류"를 인간이 정보를 체계화하고 그로부터 추론을 통해 결론을 도출하는 가장 기본적인 행위로 파악하였으며(노상규와 박진수, 2007), 온톨로지 구축 과정은 바로 이러한 분류를 수행하는 개념화 과정으로 볼 수 있다.

정보시스템 분야에서 활용되는 대표적인 온톨로지로는 번지(Bunge) 온톨로지(Bunge, 1977, 1979)를 들 수 있다. Wand and Weber(1993, 1995)는 번지의 온톨로지에서 정보시스템 모델링에 필요한 기본 개념을 차용하여 BWW (Bunge-Wand-Weber) 온톨로지 프레임워크를 설계하였으며, 정보시스템 분야에서 이루어지는 온톨로지 관련 연구(Wand et al., 1995)는 주로 BWW 프레임워크에 기반하고 있다. 번지 온톨로지에 의하면 이 세계는 고유한 프로퍼티를 갖고 있는 사물들로 구성되며, 모든 사물은 하나 이상의 프로퍼티를 갖는다. 둘 이상의 사물의 조합을 통해 합성물(Composite Thing)을 형성할 수 있으며, 이 경우 합성물은 반드시 창발적 프로퍼티(Emergent Property)을 가져야 한다. 프로퍼티는 프로퍼티를 가질 수 없으며, 프로퍼티는 하나의 사물에 연관되는 본질적(Intrinsic) 프로퍼티와 두 개의 사물에 연관되는 상호(Mutual) 프로퍼티로 구분된다.

온톨로지를 기술하기 위한 모델 및 언어는 매우 다양하며, 이 가운데 가장 널리 사용되는 모델로 RDF/OWL를 들 수 있다. RDF/OWL의 구문 구조인 RDF/XML에 대한 소개는 Masahide (2008)에서 자세히 다루고 있다. 또한 RDF(S) (RDF & RDF Schema)를 활용한 온톨로지 기술법, RDF/OWL 문법, 다양한 규칙과 질의, 그리고 온톨로지 추론의 술어 논리에 대한 소개는 Hitzler et al.(2009)에서 찾을 수 있다. 이미 이러한 이론적 성과를 반영한 상용 온톨로지 설계 도구가 다수 제공되고 있으며, 간단한 수준의 설계는 공개 소프트웨어인 Protege(Horridge, 2011)를 통해 수행할 수 있다.

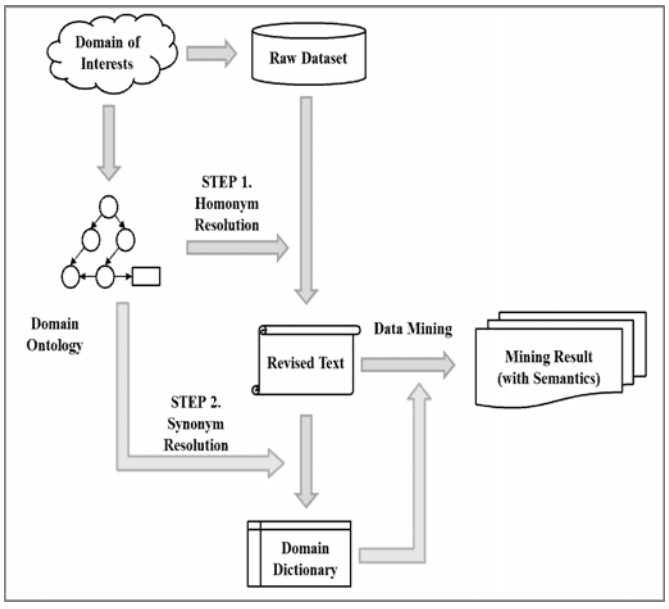
이러한 온톨로지의 활용 분야는 매우 다양하며, 그만큼 다양한 관점에서 이에 대한 연구가 이루어지고 있다. 개념적 모델링 도구로서의 온톨로지에 대한 연구가 이동훈 외(2011), Gemino and Wand(2005), Jones and Weber(1999), Shanks et al.(2010), Shanks et al.(2008), Spyns et al.(2002), 그리고 Want et al.(1995)에서 수행된 바 있으며, 시맨틱 웹에서의 추론(홍준석, 2008), 기업 간 비즈니스 프로세스의 메타데이터 표현(김형도와 김종우, 2006)에 대한 연구도 많은 성과를 내고 있다. 또한 응답률과 정확도를 고려하여 다양한 온톨로지 매핑 방법론의 성능을 평가하기 위한 연구는 안성준 외(2007)에서 이루어진 바 있다. 온톨로지에 관한 대부분의 연구는 번지 온톨로지에서 사물에 해당되는 개념의 식별에 초점을 두고 수행되었으며, 상대적으로 소수의 연구(Storey, 2005)가 프로퍼티에 해당되는 개념의 식별을 주제로 다루고 있다.

온톨로지 기반의 텍스트 마이닝을 시도한 가장 대표적인 연구로는 생체임상의학(Biomedicine) 분야의 온톨로지를 활용한 Spasic et al.(2005)를 들 수 있다. 이 연구에서는 생체임상의학 분야에서 수행된 온톨로지 및 텍스트 마이닝 관련 연구를 상세히 요약하고 있다. 하지만 용어의 정의가 비교적 명확하고 사용되는 용어의 집합이 어느 정도 한정되어 있는 생체임상의학 분야의 특성상, 이 연구의 성과를 일반화하여 다른 분야로 확산시키기에는 어려움이 있다. 또한 이 연구 역시 주로 온톨로지의 사물에만 집중하여 수행되었다 한계를 갖는다. 문장의 구문뿐 아니라 의미까지 반영한 시맨틱 텍스트 마이닝의 성공적 수행을 위해서는, 사물뿐 아니라 프로퍼티까지 식별할 수 있는 온톨로지가 제공되어야 한다. 또한 사물과 프로퍼티를 효과적으로 식별하기 위해서는 지금까지의 연구에서 간과되어온 제약조건을 활용하는 방안이 모색되어야 한다.

Ⅲ. 시맨틱 텍스트 마이닝 절차

3.1. 연구 범위 및 모형

본 절에서는 온톨로지를 활용하여 시맨틱 텍스트 마이닝을 수행하는 방안을 제시한다. 물론 이미 알려진 바와 같이 온톨로지를 성공적으로 구축하고 효과적으로 현실에 적용하는 과정에는 많은 어려움이 존재한다. 이러한 어려움의 가장 큰 원인은 온톨로지를 구축하기 위해 막대한 시간과 노력이 필요하다는 현실적인 상황에서 찾을 수 있다. 이와 관련해서는 현재도 많은 연구들이 활발하게 수행되고 있으므로, 본 논문에



<그림 3> 연구 범위 및 모형

서는 온톨로지 구축 자체에 대한 이슈는 다루지 않도록 한다. 이미 구축되어 있는 온톨로지를 전제로, 이를 활용하여 텍스트 데이터의 의미를 보다 명확하게 정제하는 전체 과정이 <그림 3>에 제시되어 있다.

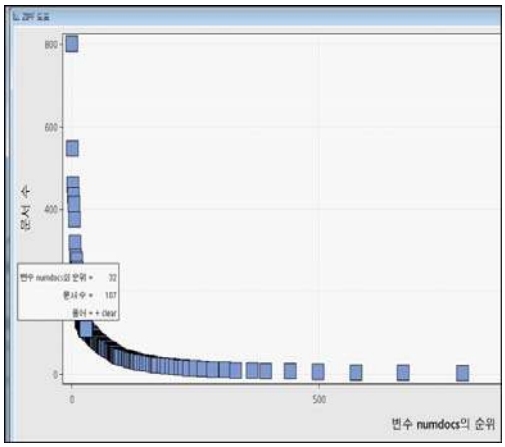
<그림 3>에서 관심 분야(Domain of Interest)의 분석 원본 데이터(Raw Dataset)와 이 분야의 어휘에 대한 온톨로지가 준비된 것으로 가정하자. 분석 대상 데이터에는 텍스트 데이터뿐 아니라 다양한 유형의 데이터가 포함되어 있을 수 있지만, <그림 3>에서 이루어지는 모든 정제 과정은 텍스트 데이터에만 적용된다. 초기 텍스트 데이터는 의미적으로 모호한 단어를 다수 포함하고 있을 수 있으며, 이러한 의미적 모호성 중 동음이의어 문제와 이음동의어 문제는 텍스트 마이닝의 결과에 직접적인 영향을 미치게 된다. 동음이의어와 이음동의어는 각각 상이한 개념을 동일한 개념으로 인식하는 오류와 동일한 개념을 상이한 개념으로 인식하는 오류를 야기한다. 이러한 오류는 단어와 개념간의 정확한 대응관계 파악을 어렵게 만들며, 이는 기본적으로 단어 간 동시 출현 빈도수에 기반하여 수행되는 텍스트 마이닝의 분석 결과를 왜곡시키는 원인이 된다. 따라서 본 연구에서는 위의 두 가지 문제 해결에 가장 중점을 두고자 하며, 이 과정은 각각 <그림 3>의 [STEP 1. Homonym Resolution]과 [STEP 2. Synonym Resolution] 단계에서 수행된다.

[STEP 1]은 온톨로지에 나타난 각 개념의 제약조건에 근거하여 동음이의어 문제를 해결하는 과정을 나타낸다. 즉 특정 단어가 어떤 개념으로 사용되었는지를 찾고, 해당 단어를 이에 대응되는 개념의 URN으로 보다 명확히 표기하게 된다. [STEP 2]는 온톨로지에 나타난 동등성 정보에 기반하여 이음동의어 문제를 해결하는 과정으로, 이 과정에서 파악된 이음동의어는 유의어 사전에 등록되어 추후 분석에 활용된다. 최종 마이닝 분석은 [STEP 1]에서 수정된 텍스트 입력에 대해 수행되며, 이 과정에서 [STEP 2]에서 생성된 유의어 사전을 활용한다.

3.2. 온톨로지를 이용한 의미적 모호성 해결

3.2.1. 온톨로지 구축 범위의 선택

텍스트 데이터의 수정을 위해 우선 관심 분야의 용어에 대한 온톨로지가 구축되어야 한다. 본 연구는 온톨로지의 구축이 아닌 활용 방안에 초점을 두고 있으므로, 관심 분야의 용어 중 텍스트 마이닝에 영향을 미치는 부분만을 온톨로지로 구축하여 사용하고자 한다. 본 연구의 관점에서 온톨로지의 구축 범위는 크게 다음의 네 가지로 구분된다: (1) 관심 영역의 용어 전체에 대한 온톨로지 구축 (2) 원본 데이터에 포함된 용어 전체에 대한 온톨로지 구축 (3) 원본 데이터에 포함된 용어 중 특정 빈도수 이상 사용된 용어에 대한 온톨로지 구축 (4) 특정 빈도수 이상 사용된 용어 중 동음이의어 또는 이음동의어와 관련된 용어에 대한 온톨로지 구축. 본 절에서는 언급한 네 가지 수준 중 가장 규모가 작은 (4) 번의 범위에 대한 온톨로지 구축 예를 통해 제안 방안을 설명하고자 한다. 예를 들어 <그림 4>의 ZIPF 도표는 약 800여개의 문서 중 용어 *Clear*가 나타난 문서가 107개이며, 출현 문서 수 기준으로 볼 때 용어 *Clear*의 순위가 32위임을 보여주고 있다. 용어 *Clear*를 기준으로 부분 온톨로지를 구축한 예가 <그림 5>에 나타나있다.

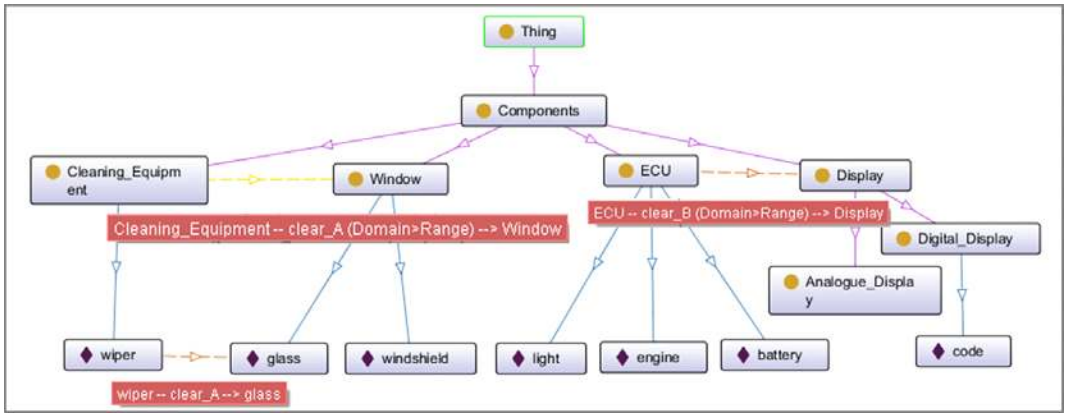


<그림 4> 온톨로지 구축 대상 용어 선정을 위한 ZIPF 도표 분석

<그림 5>는 자동차 수리 요청글에 사용된 용어 중 몇 개를 발췌하여 온톨로지로 구축한 것으로, Protege 4.2.0 에서 구축한 내용을 OntoGraf 1.0.1 Plug-in 에서 도식화한 것이다. 자동차의 구성 요소를 나타내는 *Cleaning_Equipment*, *Window*, *ECU*, *Display*가 모두 클래스(Class)로 선언되어 있으며, 이들은 각각 1 ~ 3개의 인스턴스(Instance)를 포함하고 있다. 또한 이들 네 개의 클래스는 모두 *Components*의 서브클래스(Subclass)로 선언되어 있다. 주목해야 할 부분은 동음이의어의 표현 방법이다. 실제 온톨로지에서는 개념을 명확히 기술하기 위해 접두어로 네임스페이스(Namespace)를 사용하며, 그 결과 온톨로지에서 하나의 용어는 정확하게 하나의 개념을 나타내게 된다. 하지만 텍스트 입력으로 사용된 모든 용어에 대해 접두어를 부여하는 것은 매우 번거로운 작업이므로, 본 연구에서는 둘 이상의 개념을 갖는 용어를 접두어 대신 언더스코어(_) 를 사용하여 구분하는 약식 방법을 사용하였다. 예를 들어 <그림 5>에서 *clear* 라는 용어는 "청소하다"를 의미하는 *clear_A*와 "초기화하다"를 의미하는 *clear_B*로 구분되어 나타나고 있다. <그림 5>는 프로퍼티 간에 나타난 동음이의어를 보여주고 있지만, 동음이의어는 클래스 간에도 존재할 수 있고 프로퍼티와 클래스 간에도 존재할 수 있다.

3.2.2. 동음이의어 문제의 해결

<그림 3>의 [STEP 1]은 <그림 5>와 같이 간략하게 구축된 부분 온톨로지를 활용하여 동음



<그림 5> Protege로 구축한 부분 온톨로지 예(clear_A & clear_B)

이의어 문제를 해결하는 단계를 나타낸다. 기본적인 방향은 특정 단어의 개념적 모호성이 존재할 경우, 주어진 텍스트 문장(Assertion)이 추론에 의해 참이 될 수 있는 개념을 해당 단어에 대응시킨다는 것이다. 대부분의 추론 엔진들은 RDF(S), OWL 문법에 근거한 다양한 추론 규칙들(Hitzler et al., 2009)을 내장하고 있으며, 본 연구에서는 이러한 규칙들 중 다음의 일부를 도입하여 사용하고자 한다. 단, 다음의 추론 규칙에서 a, b, x, y, u, v 는 임의의 URI를 나타낸다.

[규칙 1] (a rdfs:domain x) and (u a y) →u rdf:type x
[규칙 2] (a rdfs:range x) and (u a v) →v rdf:type x
[규칙 3] (u rdfs:subClassOf x) and (v rdf:type u) → v rdf:type x

전체 추론 규칙 및 이에 대한 증명, 그리고 자세한 표기법은 Hitzler et al.(2009)에서 찾아볼 수 있으며, 여기에서는 위의 세 가지 규칙의 의미만을 간략히 살펴보도록 한다. 우선 [규칙 1]은 임의의 프로퍼티 a의 정의역(Domain)이 x로 주어질 때, a를 사용하여 기술된 트리플(Triple)의 주어에는 항상 x의 인스턴스만 올 수 있다는 것이다. 이와 유사하게 [규칙 2]는 임의의 프로퍼티 a의 치역(Range)이 x로 주어질 때, a를 사용하여 기술된 트리플의 목적어에는 항상 x의 인스턴스만 올 수 있다는 것이다. [규칙 3]은 [규칙 1]과 [규칙 2]를 확장 적용하기 위한 것으로 임의의 클래스 u가 다른 클래스 x의 서브클래스일 때, u의 모든 인스턴스는 동시에 x의 인스턴스가 된다는 것이다.

위의 규칙을 <그림 5>의 예에 적용하는 경우를 살펴보자. 입력 텍스트에 *clear*라는 용어가 나타날 경우, 이 용어가 <그림 5>의 온톨로지에 프로퍼티로 나타난 *clear_A*와 *clear_B* 중 어떤 개념으로 사용되었는지 파악해야 한다. 그림에서 *clear_A*는 정의역과 치역으로 각각 *Cleaning_Equipment*와 *Window*를 가질 수 있으며, *clear_B*는 정의역과 치역으로 각각 *ECU*와 *Display*를 가질 수 있음을 알 수 있다. 따라서 "Wipers peeling do not clear glass"라는 문장이 입력으로 들어온 경우, 문장에 함께 나타난 *wiper*와 *glass*가 각각 *Cleaning_Equipment*와 *Window*의 인스턴스이기 때문에 [규칙 1]과 [규

	ID	Text	RepairDate	SalesDate	RepairAmount	Mileage
620	092598B66N	Power sliding door s do not work and dtc 25 is stored.	12NOV2009	20JUN2009	$105.23	5263
621	092628B66N	Right rear vent is noisy.	07OCT2009	10FEB2009	$125.30	589
622	092847B30U	Throttle sticking.	03FEB2009	04OCT2008	$32.99	7360
623	092878B32N	Car will not start.	12OCT2009	11AUG2009	$105.40	894
624	417659B11N	Wipers are not clearing_A the windshield.	18JUN2009	10SEP2008	$15.21	9164
625	683897B11J	Wipers peeling do not clear_A glass.	09FEB2009	25SEP2008	$13.79	8102
982	107371B56U	Garnish falling off.	01OCT2009	08JUN2009	$85.16	5738
983	107432B42U	Check grinding noise in front end.	19AUG2009	15JUN2009	$130.34	8717
984	107478B42N	Left frontwindow noisy.	23OCT2009	17SEP2008	$182.32	17845
985	107554H61J	Car will not start.	03OCT2009	26AUG2009	$98.46	1186
986	153204B72J	Check engine light evap system cleared_B codes.	25SEP2009	09FEB2009	$42.64	5440
987	238117B72X	Engine l on codes operating data find and clear_B.	25SEP2008	24SEP2008	$46.19	91
988	107555B32N	SRS light on battery was dead clear_B codes.	20MAY2009	14MAY2009	$42.64	157
989	107605B32N	TCS light on replace brake system control unit.	27MAY2009	15DEC2008	$912.96	6857
990	107664B32U	Right side rocker clips need replaced.	22MAY2009	03FEB2009	$28.12	6005

<그림 6> Homonym Resolution의 예(*clear_A & clear_B*)

칙 2]에 의해 이 문장에서 사용된 *clear*는 <그림 5>의 *clear_A*에 대응됨을 알 수 있다. 이와 유사하게 만약 "SRS light on battery was dead clear codes"라는 문장이 입력으로 들어온 경우, [규칙 1] ~ [규칙 3]에 의해 이 문장에서 사용된 *clear*는 <그림 5>의 *clear_B*에 대응됨을 알 수 있다. 이와 같은 과정을 통해 수정된 텍스트 입력의 일부가 <그림 6>에 나타나있다. 이와 같이 [규칙 1] ~ [규칙 3]에 의해 동음이의어 문제를 해결하는 과정은 클래스 또는 프로퍼티 중 둘 이상의 URI로 표시되는 모든 용어에 대해 적용된다. 이러한 과정을 통해 동음이의어 문제가 해결된 상태의 텍스트 데이터가 최종 마이닝 분석의 입력으로 사용되게 된다.

3.2.3. 이음동의어 문제의 해결

[STEP 1]에서의 Homonym Resolution 단계에서는 향후 분석을 위해 입력 텍스트 자체가 수정되는 반면, [STEP 2]의 Synonym Resolution 단계에서는 텍스트 데이터의 수정 없이 별도의 유의어 사전이 생성된다. 텍스트 마이닝 과정에서 이음동의어는 동음이의어에 비해 상대적으로 수월하게 해결될 수 있는데, 이는 대부분의 상용 텍스트 마이닝 도구가 유의어 사전 편집 기능을 지원하고 있기 때문이다. 또한 이들 도구들은 어느 정도 수준의 이음동의어 자동 식별 기능을 지원하고 있다. 예를 들어 SAS Enterprise Miner 7.1의 경우 형태소 분석에 근거하여 동일 개념을 나타내는 다양한 형태의 단어들을 자동으로 식별하고, 이들을 하나로 묶는 작업을 수행한다. 예를 들어 <그림 7>은 SAS Enterprise Miner 7.1의 필터 뷰어의 한 화면으로, *clear, cleared, clearing*이 모두 같은 의미로 사용됨을 나타내고 있다.

	용어	빈도 ▼	문서 수	유지	가중	역할
⊟	wiper	61	52	☑	0.477	Noun
	wipers	11	10			Noun
	wiper	49	42			Noun
	wper	1	1			Noun
	rack	59	57	☑	0.46	Noun
⊞	power	59	58	☑	0.457	Noun
⊞	fuel	58	52	☑	0.475	Noun
	car	57	57	☑	0.458	Noun
⊞	code	57	56	☑	0.462	Noun
⊞	off	56	56	☐	0.0	Adv
⊞	gauge	54	52	☑	0.474	Noun
	rear	50	49	☑	0.48	Noun
⊟	clear	49	49	☑	0.479	Verb
	clear	3	3			Verb
	cleared	43	43			Verb
	clearing	3	3			Verb

<그림 7> 형태소 분석 기반의 이음동의어 자동 식별 예

하지만 이러한 방식의 이음동의어 식별은 용어 간 형태론적 유사성에 근거하여 수행되므로, *Purchase*와 *Buy*의 예와 같이 형태가 전혀 다르지만 개념적으로 유사성을 갖는 두 용어는 유의어로 식별하지 못한다는 한계가 있다. 이러한 한계를 극복하기 위해 대부분의 상용 텍스트 마이닝 도구는 사용자가 임의로 유의어 사전을 추가 및 수정할 수 있도록 허용하며, 본 연구에서는 이음동의어 문제의 해결 과정에 이러한 기능을 활용하고자 한다. 구체적으로 이 과정에서 탐색되는 RDF/OWL 온톨로지의 프로퍼티 요소는 다음과 같다.

1) rdfs:subClassOf - 하나의 클래스가 다른 클래스의 서브클래스임을 명시할 때 사용된다. 만약 클래스 A가 클래스 B의 서브클래스라면, 유의어 사전에 클래스 A를 하위 용어로, 클래스 B를 상위 용어로 추가함으로써 이 관계를 명시

할 수 있다. 예를 들어 *Car*를 하위 용어로, *Vehicle*을 상위 용어로 추가하면, 입력 문장에 사용된 모든 *Car*는 *Vehicle*로 해석된다.

2) rdfs:subPropertyOf - 하나의 프로퍼티가 다른 프로퍼티의 서브프로퍼티(Subproperty)임을 명시할 때 사용된다. 만약 프로퍼티 A가 프로퍼티 B의 서브프로퍼티라면, 유의어 사전에 프로퍼티 A를 하위 용어로, 프로퍼티 B를 상위 용어로 추가함으로써 이 관계를 명시할 수 있다. 예를 들어 *Write*를 하위 용어로, *Create*를 상위 용어로 추가하면, 입력 문장에 사용된 모든 *Write*는 *Create*로 해석된다.

3) owl:equivalentClass - 두 클래스가 서로 동등한 경우를 나타낸다. 즉 클래스 A가 클래스 B의 서브클래스이고 이와 동시에 클래스 B가 클래스 A의 서브클래스일 때, 클래스 A와 클래스 B 는 서로 동등한(Equivalent) 위치에 있다. 이 경우 두 용어의 상위/하위 구분에 대한 명확한 기준은 없지만, 가급적 빈번하게 사용되는 용어를 상위 용어로 두는 것이 보다 바람직하다.

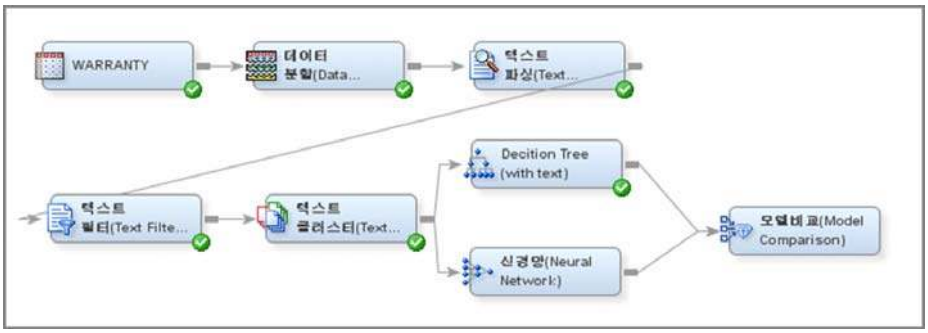
4) owl:equivalentProperty - 두 프로퍼티가 서로 동등한 경우를 나타낸다. 즉 프로퍼티 A가 프로퍼티 B의 서브프로퍼티이고 이와 동시에 프로퍼티 B가 프로퍼티 A의 서브프로퍼티일 때, 프로퍼티 A와 프로퍼티 B 는 서로 동등한 위치에 있다. 이 경우 역시 두 용어의 상위/하위 구분에 대한 명확한 기준은 없지만, 가급적 빈번하게 사용되는 용어를 상위 용어로 두는 것이 보다 바람직하다.

위의 네 가지 경우 중 마지막 두 경우, 즉 클래스 또는 프로퍼티 간의 동등관계를 설정할 경우 보다 세심한 주의가 요구된다. 두 개념이 동등함을 선언하기 위해서는, 두 개념이 서로 다른 뜻으로 사용되는 경우는 전혀 없음을 확신할 수 있어야 한다. 예를 들어 Clear라는 용어가 Clean이라는 용어와 동일한 의미를 갖는 경우가 존재한다고 해서 두 용어가 동등하다고 선언하는 것은 무리가 있다. 왜냐하면 Clear는 Clean 뿐 아니라 Reset의 의미로 사용되는 경우도 있기 때문이다. 이러한 경우는 동음이의어 문제와 이음동의어 문제가 혼합된 경우로 이해할 수 있으며, 이러한 경우가 발생하지 않도록 하기 위해 본 연구에서 제안하는 방안은 [STEP 1]을 [STEP 2]보다 반드시 먼저 수행할 것을 규정하고 있다. 즉, 하나의 용어가 둘 이상의 개념을 나타내는 데에 사용될 수 있다면, 우선 이 용어를 각각의 개념에 대응되는 세부 용어로 분할한 뒤 각각 세부 용어를 기준으로 이음동의어를 식별해야 한다는 것이다. 예를 들면 용어 *Clear*는 [STEP 1]에서 *Clear_A, Clear_B*로 세분화되며, [STEP 2]에서 이들 각각은 *Clean, Reset*과 동등관계가 설정된다. 이러한 과정을 통해 추가된 유의어 사전의 예가 <그림 8>에 나타나있다.


EMWS1.TextParsing2_synonymDS

하위 용어	상위 용어	TERMROLE	PARENTROLE
Car	Vehicle		
Work	Operate		
Difficult	Hard		
Clear_A	Clean		
Clear_B	Reset		
Delete	Reset		
Crack	Break_A		

<그림 8> 의미기반 유의어 사전의 예



<그림 9> SAS Enterprise Miner를 활용한 텍스트 마이닝 분석 예

3.2.4. 시맨틱 텍스트 마이닝

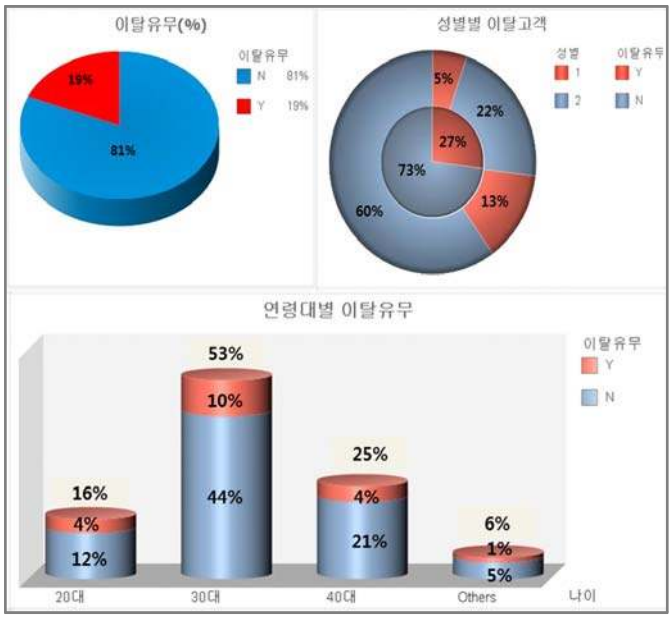
[STEP 1]과 [STEP 2]의 과정을 통해 도출된 결과물에 대해 텍스트 마이닝을 수행하는 방법은 일반적인 텍스트 마이닝 분석 과정과 크게 다르지 않다. 입력 문장에 사용된 용어의 의미를 보다 명확하게 파악하여 수정하는 작업은 크게 동음이의어 처리와 이음동의어 처리로 구성되며, 전자의 결과물은 입력 데이터를 직접 수정함으로써 반영되고 후자의 결과물은 유의어 사전의 추가에 반영된다. 이러한 결과물에 대해 SAS Enterprise Miner를 사용하여 텍스트 마이닝을 수행하는 전체 흐름의 예가 <그림 9>에 나타나있다.

Ⅳ. 온라인 쇼핑몰의 이탈 고객 예측 실험

4.1. 실험 개요 및 데이터 소개

본 절에서는 국내 "S" 온라인 쇼핑몰의 실제 거래 데이터 및 게시물에 대한 분석을 통해 본 연구에서 제안하는 방안의 실무 적용 가능성을 살펴보도록 한다. "S" 온라인 쇼핑몰은 약 1,100만명의 회원을 보유하고 있으며, 일평균 약 20,000 건의 상품을 판매하고 있는 대형 쇼핑몰이다. 본 연구에서는 2011년 1월부터 3월 사이에 Q&A 게시판에 한 번이라도 글을 남긴 고객을 대상으로 분석을 수행하였다. 구체적으로는 해당 기간에 게시물을 남긴 고객 139,952 명의 기본 정보(성별, 연령, 가입경과월수 등), 해당 고객들의 2010년 9월부터 2012년 3월까지의 거래 정보(총 구매액, 총 구매 회수 등), 그리고 해당 고객들이 마지막으로 등록한 게시물의 제목을 입력 데이터로 활용하여 고객의 이탈 가능성을 예측하고자 한다.

예측 대상이 되는 목표 변수는 각 고객별 이탈 가능성이다. "S" 쇼핑몰 고객의 주문별 평균 간격 즉 평균 재구매 간격이 약 132일인 점을 감안하여, 최종 게시물을 남긴 시점을 기준으로 132일 이내에 다시 구매할 가능성이 낮은 고객을 잠재 이탈 고객으로 정의하고 이를 예측하고자 한다. 전체 고객 139,952명 중 잠재 이탈 고객으로 분류된 고객이 20,046 명으로 그렇지 않은 고객 119,906 명에 비해 매우 적게 나타난다. 따라서 과표본추출(Oversampling)을 통해 잠재 이탈 고객 20,046 명 모두를 표본에 포함시키고,

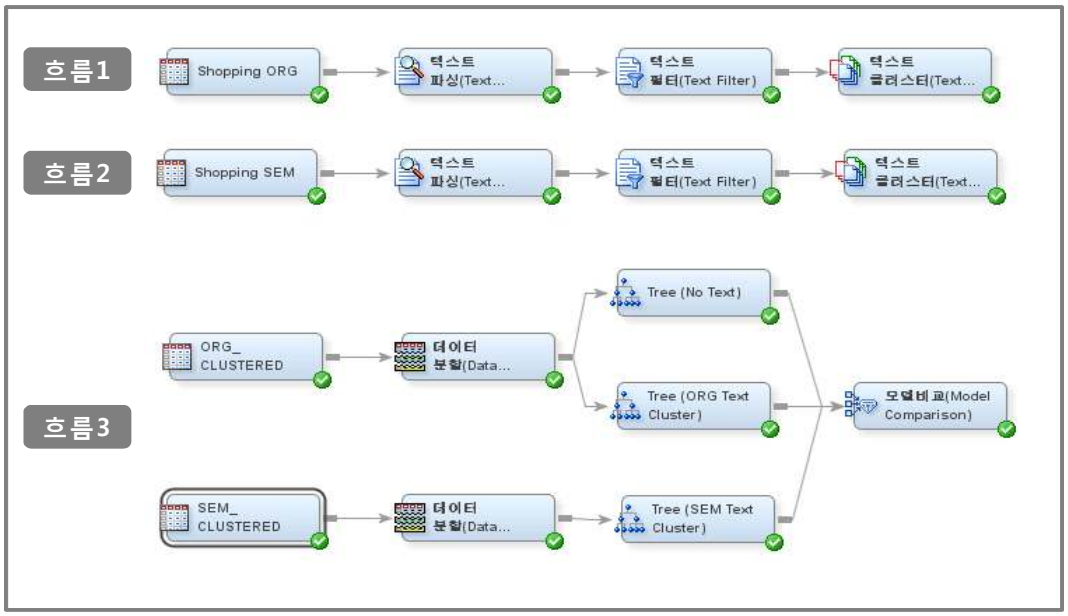


<그림 10> 전체 고객의 성별, 연령별 이탈 가능성 분포(QlikView 사용)

그렇지 않은 고객 중 20,227 명을 추출하여 표본에 포함시키도록 하였다. <그림 10>은 QlikTech 사의 데이터분석 소프트웨어인 QlikView를 이용하여, 표본추출 이전의 전체 고객에 대한 성별, 연령별 이탈 가능성의 분포를 나타낸 그림이다. 우측 상단의 그래프에서 안쪽 원은 성별(남자:1, 여자:2)을 나타내고, 바깥 원은 성별별 이탈 여부를 나타낸다. 전체 고객 중 약 19%가 잠재 이탈 고객으로 분류되었으며, 남자 이탈 고객은 전체 고객 중 약 5%, 여자 이탈 고객은 전체 고객 중 약 13%를 차지하고 있음을 알 수 있다. 연령대별로 살펴보면, 20대의 경우 전체 20대 고객 중 약 25%가 이탈 고객으로 분류되어 다른 연령대에 비해 이탈 비율이 다소 높게 나타남을 알 수 있다.

4.2. 실험 설계

본 실험은 <그림 11>의 세 가지 흐름에 의해 수행된다. 우선 흐름1은 표본 데이터에 대해 텍스트 마이닝 즉 파싱, 필터링, 그리고 클러스터링을 수행하는 과정이다. <그림 12>에 나타난 소스 데이터의 변수들 중 게시물의 제목을 나타내는 QNA_Title 만이 분석 대상으로 사용되며, 이 흐름을 통해 파악된 게시물의 클러스터 값이 추후 분석에 활용된다. 흐름2는 흐름1과 동일하나 분석 과정에서 용어의 시맨틱에 대한 추가 고려가 포함되었다는 점에서 차이가 있다. 즉 단순 텍스트 마이닝을 통해 얻을 수 있는 결과가 흐름1이라면, 흐름2는 시맨틱 텍스트 마이닝을 통해 얻을 수 있는 결과이다. 구체적으로, 흐름2의 입력 데이터는 Homonym Resolution을 거친 텍스트를 의미하고, 파싱 및 필터링 과정



<그림 11> 예측 모형 평가를 위한 분석 흐름도

번호 ▲	QNA_TITLE	GENDER...	AGE_...	CHILD...	REG_MONTH...	MARRIED...	ACTIVE...	All_Order_Sum	All_Oder_Count	Target...
1	환불요청	2	31	F	36	F	T	510048	7	Y
2	[기타][루이까또즈시계]...	1	21	F	16	F	T	183755	4	Y
3	[주문오류(고객)][코오...	1	34	F	39	F	T	510842	8	Y
4	[기타][마인엠옴므]EV코...	2	28	F	74	F	F	355925	8	Y
5	1+1	2	29	F	34	F	T	64470	2	Y
6	[기타][Dr.Martens(닥터...	1	31	F	84	F	F	44100	3	Y
7	상품교환확인	2	36	F	68	F	T	151300	5	Y
8	적립금	1	34	F	63	F	T	685480	6	Y
9	[결제수단 변경][삼성]P...	1	29	F	44	F	F	983670	2	Y
10	[배송지연][현대의료기]...	2	35	F	35	F	F	20800	4	Y
11	1월4일	1	29	F	36	F	T	168720	3	Y
12	[기타][series(시리즈)]5...	1	42	F	62	F	T	587305	6	Y
13	[기타][마인드브릿지] ...	2	32	F	72	T	F	83900	2	Y
14	[상품 불만][코디아이]8...	2	31	F	28	F	F	237070	7	Y

<그림 12> 입력 변수 및 데이터(일부)

에서 새로 추가된 유의어 사전을 활용해야 한다는 것이다. 하지만 언급한 바와 같이 해당 도메인에 대해 전체 온톨로지를 구축하고 이 온톨로지를 통해 입력 데이터를 직접 수정하는 과정은, 시맨틱 텍스트 마이닝 방안의 제안을 목표로 하는 본 연구에서 다루기에는 무리가 있다. 따라서 본 실험에서는 <그림 3>의 단계 중 Homonym Resolution 과정은 제외하고 Synonym Resolution만 간략히 수행하도록 한다.


EMWS1.TextParsing2_synonymDS

하위 용어	상위 용어	용어 역할	상위 용어 역할
물건	상품		
컬러	색상		
취소	반품		
환불	반품		
에러	오류		
증정용	사은품		
유아	아동		
배달	배송		
크기	사이즈		
재질	소재		
특가	할인		
파손	불량		

<그림 13> 동음이의어 문제 해결을 위한 간단한 유의어 사전

용어

용어	빈도 ▼	문서 수	유지	가중	역할	속성
요청	26919	26895	☑	1.567	Noun	알파
⊟ 반품	23906	23694	☑	1.75	Noun	알파
반품	15008	14967			Noun	알파
취소	7952	7924			Noun	알파
환불	946	939			Noun	알파
⊞ 반품#요청	13923	13923	☑	2.517	명사 그룹	혼합
⊟ 상품	11544	11306	☑	2.818	Noun	알파
상품	11318	11082			Noun	알파
물건	226	225			Noun	알파
불만	9474	9474	☑	3.073	Noun	알파

<그림 14> 유의어 사전이 반영된 텍스트 필터링 결과

클러스터

클러스터 ID	기술 용어	빈도	비율
1	결품품절 취소 취소요청 품 품절 해외 나이키 nike 티셔츠 +black 사이즈 니트 +free...	1035	3%
2	'상품 불만' nike s 공용 교환 교환요청 그레이 기타 나이키 남성 넌 데님 반품 반품...	5725	14%
3	+오류 고객 교환 교환요청 주문 주문오류 데님 블루 팬츠 공용 번호 네이비 요청 여...	1661	4%
4	'사은품 파손' '상품 불만' +네이 교환요청 네이비 더블 버튼 본품 불량 비 사은품 자...	2705	7%
5	+세트 개 너무 무료 무료배송 문의 배송 번호 부탁 상품권 선물 신세계 적립 제품 ...	4223	10%
6	'사은품 파손' 교환 교환요청 본품 불량 사은품 파손 ennee 의자 누락 기 베이비 반...	1029	3%
7	+되 +언제 날짜 너무 답변 드려요 문의 물건 배 배송 배송완료 부탁 부탁드립니다 ...	4746	12%
8	+emma +short +w608 8인치 bear bearpaw meadow paw smoke 베어 숏 스모크 엠...	156	0%
9	'결제수단 변경' '변경 요청' 결제 배송지 변경 수단 연 주소 취소 취소요청 소다 요...	2568	6%
10	+몰 hd kb 보호 불필요 삼성 삼성직배송설 송 송설 신세계 임직원 정품 증정 직배 ...	823	2%
11	'구성품 정보' '부분 출고' '사은품 파손' '상품 불만' +오 +오류 교환 교환요청 구성 ...	5085	13%
12	+l +m 가디건 사이즈 사이즈불만 색상 여성용 위즈 지오다노 취소요청 칼라 코튼 ...	1471	4%
13	'쿠폰 미적용' 가격 가격불만 기구 기구입 단순 단순변심 미적 변심 불필요 선물받...	4048	10%
14	+free 니트 르샵 반품 반품요청 배색 베이지 불만 브라운 블랙 상품 상품불만 여성 ...	4998	12%

(a) 흐름1의 클러스터링 결과

클러스터

클러스터 ID	기술 용어	빈도	비율
1	'부분 출고' '사은품 파손' '쿠폰 미적용' +데 +배송 +불량 +오 교환 교환요청 무...	3220	8%
2	+세트 기타 나이키 단순 단순변심 무료 무료배송 변심 불필요 신세계 정품 증 ...	7622	19%
3	배송지 소다 연 요청 취소 취소요청 펌프스 +free 해외 블랙 남성 무료배송 형 ...	1681	4%
4	'타미 힐' +m +w +거 +네이 네이비 데님 미 비 여성 타 타미힐 피 피거 피거데...	742	2%
5	'상품 불만' +네이 +오류 고객 교환 교환요청 기타 네이비 비 상품정보오류 요...	9948	25%
6	+free +반품 +상품 bk 르샵 반품요청 베이지 불만 블랙 상품불만 여성 요청 헌...	6285	16%
7	'결제수단 변경' '변경 부탁드립니다' '변경 요청' '주소 변경' 결제 드립니다 배...	1190	3%
8	+되 +배송 +언제 건 관련 너무 답변 되나요 드려요 드립니다 문의 배 배송일 부...	5349	13%
9	가격 가격불만 기구 기구입 불필요 선물받음 요청 입 입고 입고지연 재주문 지...	1885	5%
10	결 결품품절 취소 취소요청 품 품절 해외 티셔츠 +black p 건 니트 +l 티 +free ...	919	2%
11	'타미 힐' +거 +사이즈 +색상 p 및 위즈 지오다노 칼라 코튼 타 피 피거데님 하...	1432	4%

(b) 흐름2의 클러스터링 결과

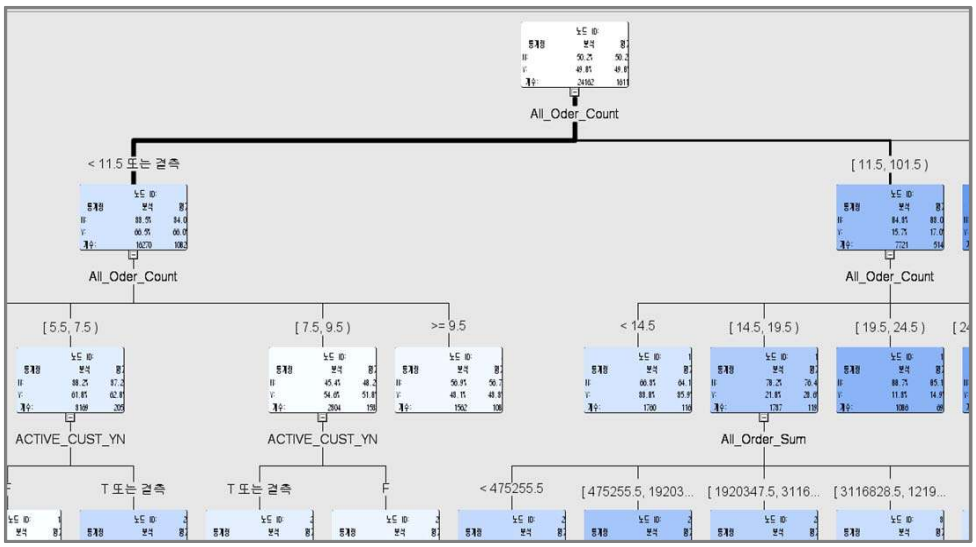
<그림 15> 흐름1과 흐름2의 클러스터링 결과 비교

<그림 13>은 흐름2에서 사용된 간단한 유의어 사전을 보여준다. <그림 13>의 유의어 사전은 흐름1의 텍스트 필터링 분석 결과로 파악된 빈발 용어 중 일부에 대해 작성되었으며, 그 결과는 흐름2의 텍스트 필터링 결과에 <그림 14>와 같이 반영되어 나타난다.

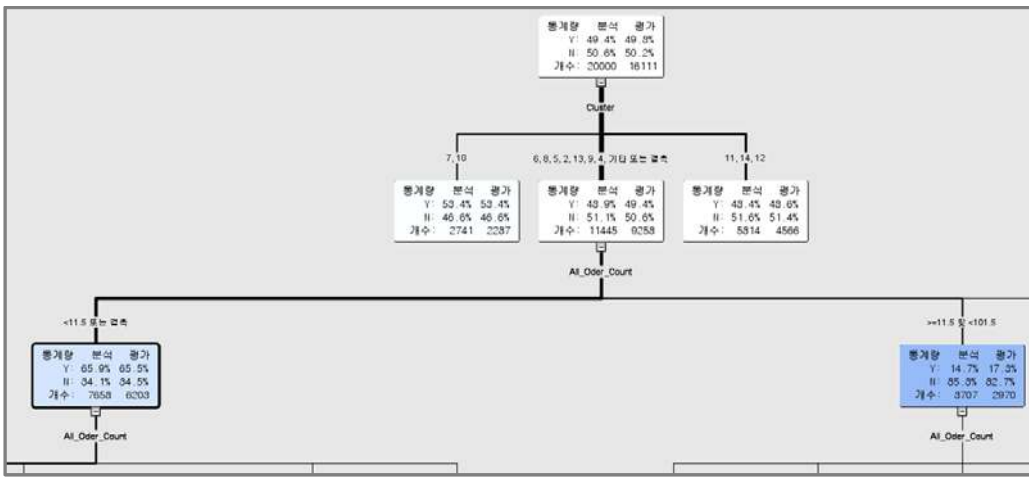
4.3. 실험 결과 및 해석

<그림 11>의 흐름1과 흐름2의 결과, 즉 단순 텍스트 클러스터링 결과와 시맨틱 텍스트 클러스터링 결과가 각각 흐름3의 *ORG_CLUSTERED*와 *SEM_CLUSTERED*의 입력으로 활용된다. 흐름1의 결과 14개의 클러스터가 생성되고 흐름2의 결과 11개의 클러스터가 생성된 것으로 보아 최소한 두 흐름의 결과가 서로 다르게 나타나고 있음을 알 수 있다<그림 15>.

흐름3에서는 세 개의 의사결정나무기반 예측 모델을 비교하고 있다. 맨 위의 모델은 텍스트 입력을 전혀 고려하지 않은 모델로, 입력 변수 중 텍스트 관련 입력을 모두 사용하지 않음으로써 수행 가능하다. 두 번째와 세 번째 모델은 각각 흐름1과 흐름2의 텍스트 클러스터링 결과를 반영한 모델로 대화식의사결정나무를 활용하여 도출되었다. 세 가지 모델 중 *Tree(No Text)* 모델과 *Tree(ORG Text Cluster)* 모델의 결과 트리 일부가 <그림 16>과 <그림 17>에 나타나있다. T*ree(ORG Text Cluster)* 모델은 최상위 노드의 최초 분기 기준을 대화식으로 선택한 것으로, 흐름1의 결과로 부여된 텍스트 클러스터의 번호가 최초 분기 기준으로 사용되었다. 즉 클러스터의 번호에 따라 최상위 노드는 세 개의 가지로 나뉘게 되며, 그 이후 분기 과정은 일반적인 의사결정나무 분석 기법에 따라 자동으로 수행된다. 다만 <그림 17>에서는 세 개의 클러스터 분기 결과 중 좌측과 우측 노드는 편의상 생략하여 보여주고 있다. *Tree(No Text)* 모델의 결과 트리에서는 최상위 분기 기준으로 총



<그림 16> *Tree(No Text)* 노드의 결과 트리 일부



<그림 17> *Tree(ORG Text Cluster)* 노드의 결과 트리 일부

주문 회수인 *All_Order_Count*가 채택되었으며, 이 변수는 <그림 17>의 결과에서도 클러스터 번호 다음의 분기 기준으로 사용되었다. 참고로 세 가지 모델에서 계산된 각 변수의 중요도는 <그림 18>에 나타나있다.

<그림 18>에 나타난 세 가지 의사결정나무 모델에 의해 잠재 이탈 고객을 예측한 각각의 결과에 대한 비교가 <그림 11>의 모델비교 (Model Comparison) 노드에서 수행된다. <그림 18>의 변수 중요도는 세 가지 모델에서 서로 다르게 나타나며, <그림 18(b)>와 <그림 18(c)>에서 *Cluster*로 나타난 변수명은 텍스트 분석 결과로 생성된 파생 변수이다. 즉 특정 용어들을 포함하는 문장은 특정 클러스터에 속하게 된다. 모형 개발자의 관점에서는 예측을 위한 의미있는 변수의 발굴이 매우 중요하다는 점을 감안할 때, 텍스트 분석을 통해 집합으로 규정되는 새로운 변수 *Cluster*는 이후 분석에서 활용 가치가 매우 높다고 할 수 있다. 모델 비교를 위해 반응률, 반응검출율, 향상도 등의 예측력 평가 지표가 사용되었으며(그림 19), 그 중 검증 (Validation) 데이터에 대한 누적 반응률

변수 중요도

변수 이름	중요도
All_Oder_Count	1
ACTIVE_CUST_YN	0.191529
All_Order_Sum	0.071275
AGE_CODE	0
REG_MONTH_CNT	0
GENDER_CODE	0
CHILD_EXISTS	0
MARRIED_YN	0

(a) *Tree(No Text)*의 변수 중요도

변수 중요도

변수 이름	중요도
All_Oder_Count	1
ACTIVE_CUST_YN	0.192931
All_Order_Sum	0.080747
Cluster	0.06134
REG_MONTH_CNT	0
MARRIED_YN	0
AGE_CODE	0
CHILD_EXISTS	0
GENDER_CODE	0

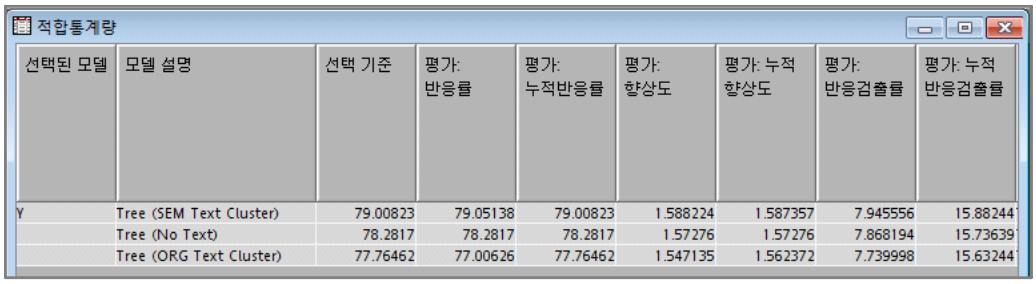
(b) *Tree(ORG Text Cluster)*의 변수 중요도

변수 중요도

변수 이름	중요도
All_Oder_Count	1
ACTIVE_CUST_YN	0.191952
All_Order_Sum	0.076821
Cluster	0.063345
REG_MONTH_CNT	0
MARRIED_YN	0
AGE_CODE	0
CHILD_EXISTS	0
GENDER_CODE	0

(c) *Tree(SEM Text Cluster)*의 변수 중요도

<그림 18> 세 가지 모델의 변수 중요도 비교

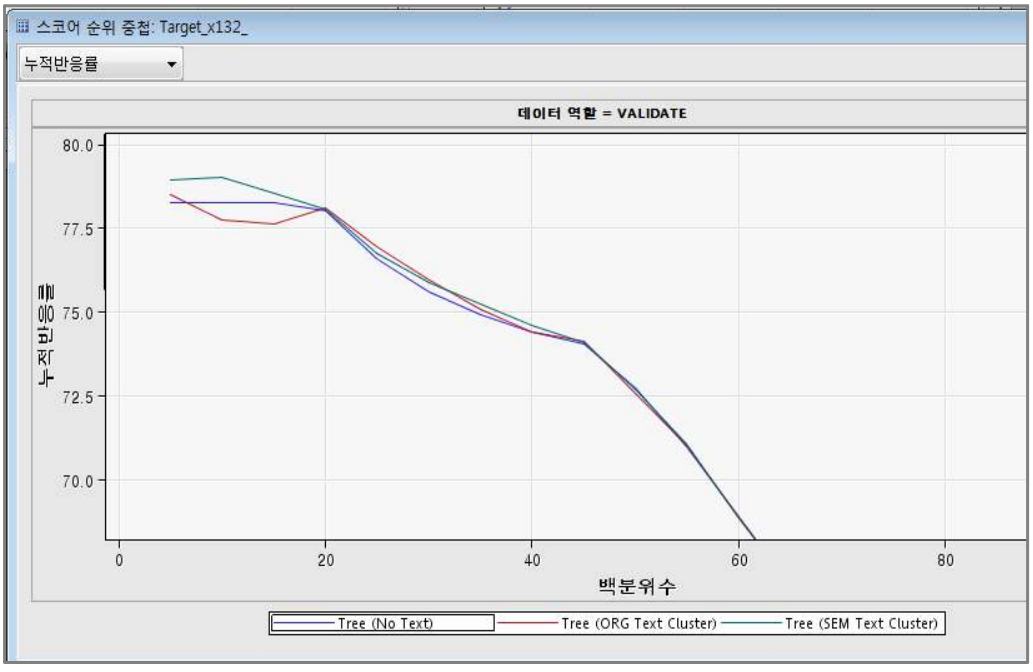

적합통계량

선택된 모델	모델 설명	선택 기준	평가: 반응률	평가: 누적반응률	평가: 향상도	평가: 누적 향상도	평가: 반응검출률	평가: 누적 반응검출률
Y	Tree (SEM Text Cluster)	79.00823	79.05138	79.00823	1.588224	1.587357	7.945556	15.88244
	Tree (No Text)	78.2817	78.2817	78.2817	1.57276	1.57276	7.868194	15.73639
	Tree (ORG Text Cluster)	77.76462	77.00626	77.76462	1.547135	1.562372	7.739998	15.63244

<그림 19> 세 모델의 예측력 비교

(Cumulative Response)을 비교한 그래프가 <그림 20>에 나타나있다. 분석 결과에 따르면 백분위수 상위 20% 구간에서 *Tree(SEM Text Cluster)*의 결과가 다른 두 모델의 결과에 비해 우수하게 나타남을 알 수 있다. 동일한 구간에서 *Tree(ORG Text Cluster)*의 결과는 텍스트 입력을 전혀 고려하지 않은 *Tree(No Text)*의 결과보다도 오히려 저조하게 나타남을 알 수 있다. 한편 상위 20% 이후의 구간에서는 세 모델의 예측 정확도는 거의 차이가 없는 것으로 나타나고 있다.

<그림 19>와 <그림 20>의 분석 결과에 따르면 본 실험에서의 모델별 예측력은 시맨틱 텍스트 마이닝, 일반 데이터 마이닝, 그리고 일반 텍스트 마이닝의 순서로 높게 나타남을 알 수 있다. 하지만 <그림 19>의 수치에서도 알 수 있듯



<그림 20> 세 모델의 누적 반응률 비교

이 이들 세 모델의 예측력이 텍스트 입력의 사용 여부, 그리고 시맨틱 정보의 활용 여부에 따라 매우 작은 차이만을 보이는 것으로 나타났다. 이는 입력 텍스트에 사용된 방대한 양의 용어 중 극히 일부에 대해서만 유의어 사전을 구축한 상태에서 시맨틱 텍스트 마이닝을 수행했기 때문에 나타난 당연한 결과라고 할 수 있다. 하지만 본 실험 결과는 텍스트 마이닝 분석 결과의 향상을 위해 시맨틱 텍스트 마이닝이 수행되어야 한다는 본 연구의 주장을 뒷받침했다는 측면에서 의미가 있다고 할 수 있다.

Ⅴ. 결 론

다양한 데이터 마이닝 기법을 활용한 빅데이터 분석에 대한 관심이 높아짐으로 인해, 많은 상용 데이터 마이닝 도구들이 텍스트 분석 기능을 제공하게 되었다. 인간이 지식을 표현하고 저장하고 공유하는 가장 기본적인 형태가 텍스트라는 점에서, 텍스트 데이터 분석 결과는 그 활용 가치가 높고 파급 효과도 상당할 것으로 예상된다. 하지만 텍스트 문서를 정확하게 분석하기 위해서는 각 용어가 해당 문서에서 갖는 시맨틱에 대한 정확한 이해가 필요하며, 이는 자연어처리 분야의 오랜 연구 주제인 동음이의어 및 이음동의어 해결의 문제로 귀결된다. 따라서 본 논문에서는 해당 도메인의 지식을 온톨로지로 구조화하고, 이를 활용하여 의미적 모호성을 해결함으로써 텍스트 마이닝 결과의 품질을 향상시킬 수 있는 방안을 제시하였다.

본 연구의 가장 큰 기여는 크게 다음의 세 가지에서 찾을 수 있다. 우선 상용 마이닝 도구에서도 지원이 될 정도로 최근 수요가 높아지고 있는 텍스트 마이닝과 온톨로지를 접목시킴으로써, 온톨로지의 활용 가능성 및 연구 가치를 다시 한 번 강조하였다. 또한 온톨로지를 통해 텍스트 마이닝 과정에서의 의미적 모호성을 줄일 수 있는 시맨틱 텍스트 마이닝의 개념을 소개하고, 이 과정을 통해 동음이의어와 이음동의어의 문제를 해결할 수 있는 방안을 제시하였다. 마지막으로 국내 "S" 온라인 쇼핑몰의 실제 거래 정보 및 게시글 데이터에 대한 분석을 통해 제안하는 방안의 실무 적용 가능성을 살펴보았다. 실험 결과 아주 작은 규모의 시맨틱 정보 반영을 통해서도 일반 텍스트 마이닝에 비해 우수한 예측력을 갖는 모델을 도출할 수 있음을 알 수 있었다.

본 연구의 가장 큰 한계는 실험 및 구현의 측면에서 찾을 수 있다. 본 논문에서 제안하는 방안이 실무에 적용되기 위해서는 우선 해당 도메인에 대한 온톨로지가 구축되어야 한다. 하지만 온톨로지의 구축 자체가 현재도 연구가 활발하게 이루어지고 있는 도전적인 분야이기 때문에, 본 연구에서 제안하는 방안이 실무에서 즉시 적용될 것을 기대하기란 매우 어렵다. 또한 본 연구에서는 온톨로지가 구축되어 있는 상황에서 추론 규칙에 기반하여 개념을 식별하기 위한 방안을 제안하였지만, 이를 실제로 자동화하는 과정에는 또 다른 많은 어려움이 따를 것으로 예상된다. 마지막으로 본 연구에서 수행된 실 데이터 대상 실험의 경우, 앞에서 언급한 두 가지 한계로 인해 사용된 텍스트 데이터 전체 범위에 대한 시맨틱 식별 작업이 충분히 수행되지 못했다는 한계를 갖는다. 즉 일부 데이터에 대해서만 시맨틱 식별이 이루어졌기 때문에, 현

재의 분석 결과는 예측력 차이의 통계적 유의성을 논하기엔 매우 부족하다고 할 수 있다. 본 논문에서의 실험은 제안 방안에 따른 분석 과정을 소개하고 그 결과로 시맨틱 텍스트 마이닝이 단순 텍스트 마이닝에 비해 더 좋은 예측력을 가질 수 있다는 가능성을 보였다는 측면에서 제한적인 의의를 가지며, 추후 시맨틱 텍스트 마이닝을 통해 예측력의 향상이 통계적으로 유의한 수준으로 나타나는지 여부를 엄밀하게 살펴볼 필요가 있다. 또한 향후에는 보다 넓은 범위의 온톨로지를 구축하고, 이에 기반한 시맨틱 텍스트 마이닝 수행 과정의 자동화 수준을 높이기 위한 시도가 반드시 필요하다.

참고문헌

김인현, "빅데이터 가치와 도입 전략," 2012 Big Data 검색 분석 기술 Insight, 보고서, 2012.

김형도, 김종우, "기업간 비즈니스 프로세스 메타데이터 온톨로지 설계," 한국IT서비스학회 2006년 추계학술대회, 2006.

노상규, 박진수, 인터넷 진화의 열쇠 온톨로지, 가즈토이, 2007.

손윤호, 김인규, 김남규, "연관규칙 마이닝을 활용한 개념적 데이터베이스 설계 자동화 기법," 정보시스템연구, 제18권, 제4호, 2009, pp.59-86.

안성준, 김우주, 박상언, "최적 온톨로지 매핑 방법론에 관한 연구," 한국지능정보시스템학회 2007년 추계학술대회 논문집, 2007. pp.457-462.

유지연, "세계경제포럼(WEF)을 통해 본 빅데이터 논의 동향과 함의," 정보통신정책연구원 방송통신정책, 제24권, 제4호, 2012.

이동훈, 김남규, 정인환, "온톨로지와 개체관계 모델의 상호운용성에 대한 연구," *Journal of Information Technology Applications and Management,* 제18권, 제4호, 2011. pp.95-118.

정윤수, 이춘열, 김남규, "토픽맵의 다중역할 토픽 보존을 위한 관계형 데이터베이스 구조," 정보시스템연구, 제18권, 제3호, 2009, pp.327-349.

최광선, "SNS 시대의 하이브리드 빅데이터 분석 기술 및 사례," 2012 Big Data 검색 분석 기술 Insight, 보고서, 2012.

홍준석, "시맨틱 웹에서의 효율적인 온톨로지 추론을 위한 개선방법에 관한 연구," 한국전자거래학회지, 제13권, 제3호, 2008, pp.85-101.

홍태호, 김진완, "데이터 마이닝의 비대칭 오류비용을 이용한 지능형 침입탐지시스템 개발," 정보시스템연구, 제15권, 제4호, 2006, pp.211-224.

Albright, R., *Taming Text with the SVD*, SAS Institute Inc., 2006.

Antoniou, G., and Harmelen, F. V. V., *A Semantic Web Primer*, 2nd edition, The MIT Press, 2008.

Bunge, M. A., *Treatise on Basic Philosophy (Volume 3): Ontology I*, The Future of the World, D. Reidel Publishing Company, Boston, 1977.

Bunge, M. A., *Treatise on Basic Philosophy (Volume 4): Ontology II*, A World of Systems, D. Reidel Publishing Company, Boston, 1979.

Fan, W., Wallace, W., Rich, S., and Zhang, Z., "Tapping the Power of Text Mining," *Communications of the ACM*, Vol.49, No.9, 2006. pp.76-82.

Gartner, *Hype Cycle for Emerging Technologies*, 2011, Gartner, 2011.

Gemino, A., and Wand, Y., "Complexity and Clarity in Conceptual Modeling: Comparison of Mandatory and Optional Properties," *Data & Knowledge Engineering*, Vol.55, No.3, 2005, pp.301-326.

Han, J., and Kamber, M., *Data Mining: Concepts and Techniques*, 2nd, Morgan Kaufmann Publishers, 2006.

Hearst, M. A., "Untangling Text Data Mining," *In Proceedings of ACL*, 1999, pp.3-10.

Hitzler, P., Krotzsch, M., and Rudolph, S., *Foundations of Semantic Web Technologies*, CRC Press, 2009.

Horridge, M., *A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools,* The University of Manchester, 2011.

Jones, A. B., and Weber, R., "Understanding Relationships with Attributes in Entity-Relatioship Diagrams," *in Proceedings of the 20th International Conference on Information Systems(ICIS)*, 1999, pp.241-228.

Maedche, A., Staab, S., Stojanovic, N., Studer, R., and Sure, Y., "SEAL-A Framework for Developing Semantic Web PortALs," *in Proceedings of British National Conference on Databases*, Vol.2097, 2001, pp.1-22.

Masahide, K., 시맨틱 웹을 위한 RDF/OWL 입문, 홍릉과학출판사, 2008.

Mckinsey, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Mckinsey Global Institute, 2011.

Metzler, D., Bernstein, Y., Crofit, W. B., Moffat, A., and Zobel, J., "Similarity Measures for Tracking Information Flow," *in Proceedings of CIKM,* 2005, pp.517-524.

Mooney, R. J., and Bunescu, R., "Mining Knowledge from Text using Information Extraction," *ACM SIGKDD Explorations,* Vol.7, No.1, 2006, pp.3-10.

Rijsbergen, C. J. V., *Information Retrieval*, 2nd edition, Butterworth, London, 1979.

Salton, G., Wong, A., and Yang, C. S., "A Vector Space Model for Automatic Indexing," *Communications of the ACM,* Vol.18, No.11, pp. 613 - 620, 1975.

SAS, *Text Analytics with SAS Text Miner Course Note*s, SAS Institute Inc., 2010.

Sebastiani, F., "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol.34, No.1, 2002, pp.1-47.

Sebastiani, F., *Classification of Text, Automatic*, The Encyclopedia of Language and

Linguistics 14, 2nd edition, Elsevier Science Pub., 2006.

Shanks, G., Nuredini, J., Tobin, D., Moody, D., and Weber, R., "Representing Things and Properties in Conceptual Modelling: An Empirical Evaluation," *Journal of Database Management,* Vol.21, No.2, 2010, pp.1-25.

Shanks, G., Tansley, E., Nuredini, J., Tobin, D., and Weber, R., "Representing Part-Whole Relations in Conceptual Modeling: An Empricial Evaluation," *MIS Quarterly,* Vol.32, No.3, 2008, pp.553-573.

Spasic, I., Ananiadou, S., Mcnaught, J., and Kumar, A., "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text," *Briefing in Bioinformatics,* Vol.6, No.3, 2005, pp.239-251.

Spyns, P., Meersman, R., and Jarrar, M., "Data Modelling versus Ontology Engineering," *ACM SIGMOD Record,* Vol.31, No.4, 2002, pp.12-17.

Stanvrianou, A., Andritsos, P., and Nicoloyannis, N., "Overview and Semantic Issues of Text Mining," *ACM SIGMOD Record,* Vol.36, No.3, 2007, pp.23-34,

Storey, V. C., "Comparing Relationships in Conceptual Modeling: Mapping to Semantic Classifications," *IEEE Transactions on Knowledge and Data Engineering,* Vol.17, No.11, 2005, pp.1478-1489.

Wand, Y., Monarchi, D. E., Parsons, J., and Woo, C. C., "Theoretical Foundations for Conceptual Modelling in Information Systems Development," *Decision Support Systems,* Vol.15, No.4, 1995, pp.285-304.

Wand, Y., and Weber, R., "On the Ontological Expressiveness of Information System Analysis and Design Grammars," *Journal of Information Systems,* Vol.3, No.4, 1993, pp.217-237.

Wand, Y., and Weber, R., "On the Deep Structure of Information Systems," *Information System Journal*, Vol.5, No.3, 1995, pp.203-223.

Witten, I. H., *Text Mining*, Practical Handbook of Internet Computing, edited by M. P. Singh, CRC Press, 2004.

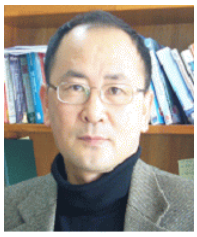
유은지(Yu, Eun-Ji)

원광대학교 정보·전자상거래학부와 소방행정학부에서 복수 학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원 정보미디어전공 석사 과정에 재학 중이다. 주요 관심분야는 CRM, 데이터 관리, 데이터 마이닝 등이다.

김정철(Kim, Jung-Chul)

성균관대학교 회계학과 학사, 국민대학교 비즈니스IT전문대학원 석사를 마치고 동대학원 박사과정에 재학 중이다. 국방부 회계관리과장, 정보화정책과장, 시설기획과장 등 주요 보직을 역임하였으며 현재 시설기획환경과장으로 재직 중이다. 주요 관심분야는 회계 및 자산 관리, 데이터 자산 및 마스터 데이터 관리 등이다.

이춘열(Lee, Choon-Youl)



1979년 서울대학교 산업공학과를 졸업하고, 1983년 서울대학교 대학원 경영학과에서 경영학석사 학위를 취득하였으며, 1990년 미시간대학교에서 경영정보학박사 (Computer and Information Systems) 학위를 취득하였다. 또한, 1779년부터 1984년까지 국방정보체계연구원에서 연구원으로 근무하였으며, 1991년부터 1993년까지 한국통신 소프트웨어연구소에서 근무하였다. 국민대학교 비즈니스 IT 전문대학원 원장을 역임하였으며, 현재 국민대학교 비즈니스IT학부 교수로 재직하고 있다. 한국경영정보학회, 학국경영학회, 대한산업공학회, 데이타베이스학회 등의 정회원이며, 주요 관심분야는 데이터베이스, 데이터웨어하우징, 정보 자원 계획 및 관리 등이다.

김남규(Kim, Nam-Gyu)

현재 국민대학교 경영정보학부에서 조교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, JITAM 편집위원을 역임하였으며, 한국경영정보학회, 한국지능정보시스템학회, 한국정보시스템학회 종신회원 및 한국생산성본부 자문위원으로 활동 중이다. 주요 관심분야는 시맨틱 데이터 관리, 데이터베이스 설계 및 데이터 마이닝 등이다.

<Abstract>

Using Ontologies for Semantic Text Mining

Yu, Eun-Ji · Kim, Jung-Chul · Lee, Choon-Youl · Kim, Nam-Gyu

The increasing interest in big data analysis using various data mining techniques indicates that many commercial data mining tools now need to be equipped with fundamental text analysis modules. The most essential prerequisite for accurate analysis of text documents is an understanding of the exact semantics of each term in a document. The main difficulties in understanding the exact semantics of terms are mainly attributable to homonym and synonym problems, which is a traditional problem in the natural language processing field. Some major text mining tools provide a thesaurus to solve these problems, but a thesaurus cannot be used to resolve complex synonym problems. Furthermore, the use of a thesaurus is irrelevant to the issue of homonym problems and hence cannot solve them. In this paper, we propose a semantic text mining methodology that uses ontologies to improve the quality of text mining results by resolving the semantic ambiguity caused by homonym and synonym problems. We evaluate the practical applicability of the proposed methodology by performing a classification analysis to predict customer churn using real transactional data and Q&A articles from the "S" online shopping mall in Korea. The experiments revealed that the prediction model produced by our proposed semantic text mining method outperformed the model produced by traditional text mining in terms of prediction accuracy such as the response, captured response, and lift.

Keywords: Classification, Data Mining, Ontology, Semantic, Text Mining

* 이 논문은 2012년 8월 9일 접수되어 1차수정(2012년 8월 31일)을 거쳐 2012년 9월 6일 게재 확정되었습니다.