

IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여[†]

최진욱¹ · 신동원² · 이한준³

¹고려대학교 스마트미디어 서비스 연구센터 · ²명지대학교 산학협력단 · ³명지대학교 경영정보학과

접수 2021년 8월 27일, 수정 2021년 9월 7일, 게재확정 2021년 9월 7일

요약

IT 산업에서의 이직은 기술 축적과 개발 및 관리에 있어 기업들에게 큰 어려움이 되고 있어 이에 대한 연구가 긴요한 실정이다. 하지만 아직까지 IT 분야의 이직 및 인력 관리와 관련된 연구는 많지 않다. 이에 본 연구에서는 이직률에 영향을 미치는 직원들의 만족 요인과 불만족 요인들을 살펴보고 이를 활용한 IT 기업의 이직률 예측 모델을 제안하고자 한다. 이를 위해 온라인 기업 리뷰 사이트인 잡플래닛에서 국내 유가증권시장에 상장된 129개 IT 기업에 종사하는 직원들의 리뷰 21,589건을 수집하고 토픽모델링을 실시하였다. 추출된 토픽들을 활용하여 이직률에 대한 머신러닝 기반의 예측 모델을 제안하였다. 또한 변수 중요도 평가를 통하여 이직률에 영향을 미치는 직원들의 만족 및 불만족 요인별 영향 정도를 분석하여 제시하였다.

주요용어: 기업 리뷰, 머신러닝, 이직률, 직원 만족, 토픽모델링.

1. 서론

인공지능을 비롯한 IT 기술이 4차산업혁명을 선도하면서 IT 인력의 중요성이 부각되고 있다. 우수한 IT 인력은 IT 기업의 주요 자산이자 국가 경쟁력의 핵심이 되기 때문이다. 정보통신정책연구원의 발표 정보통신정책연구원, ICT 통계분석 2021-1호, 2021.4. 에 따르면 2021년 2월 기준 국내 IT 산업 종사자 수는 105만여 명에 이른다. IT 산업의 규모와 종사자 수는 지속적으로 증가하고 있으나 IT 산업은 타 산업 대비 직무 수명이 짧고, 높은 이직률로 인해 여전히 상당수 IT 기업들은 인력 부족과 직원들의 이직 문제에 시달리고 있는 것이 현실이다.

일반적으로 이직은 유능한 인력과의 단절로 인해 기업 내 업무 균형을 일시적으로 무너뜨려 생산성의 저하를 가져오는 등 기업 운영에 부정적인 결과를 야기할 소지가 높은 것으로 여겨져왔다 (Staw, 1980). 더욱이 IT 산업에서의 이직은 IT가 갖는 특수성으로 인해 기업의 기술 축적과 개발 및 관리에 있어 타 산업 대비 더 큰 어려움을 초래할 수 있다 (Jung과 Son, 2017). 그리고 이는 국내 IT 산업 발전에 악영향으로 이어지질 가능성이 크다. 따라서 IT 인력의 이직은 IT 기업의 안정적 성장과 IT 산업 발전에 있어서 매우 중요한 문제라고 할 수 있다. 이직은 인사관리 분야에서 1970년대부터 연구되어 온 오랜 연

[†] 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (No. NRF-2019R1G1A1084863).

¹ (02841) 서울특별시 성북구 안암로 145, 고려대학교 스마트미디어 서비스 연구센터. 연구교수.

² (17058) 경기도 용인시 처인구 명지로 116 명지대학교 산학협력단. 연구원.

³ 교신저자: (03674) 서울특별시 서대문구 거북골로 34, 명지대학교 경영정보학과. 조교수. E-mail: hjlee1609@gmail.com

구주제이지만 아직까지 IT 분야의 이직에 대한 논의는 충분하지 못하다 (Bludedorn, 1978). IT 분야 이직 관련 선행 연구에서는 주로 해당 연구자의 학문적 관점에서 소수의 선행 변수와 이직의도에 초점을 두고 진행되어 왔다 (Kim과 Bae, 2018; Kwon과 Park, 2010). 일반적으로 이직은 특정 요인보다 여러 요인들의 복합적인 작용에 의해 발생하기 때문에 이직을 유발하거나, 혹은 이직을 억제하는 데 관련이 있는 다차원적인 속성에 대한 보다 폭넓은 고찰이 필요한 것으로 판단된다 (Jung과 Son, 2017).

한편, 사회 전반에 걸쳐 온라인 리뷰 작성과 검색이 일상화됨에 따라 여러 분야에서 온라인 리뷰를 활용한 연구가 진행되고 있다. 그 중 온라인 기업 리뷰는 자신이 근무한 이력이 있거나 현재 근무 중인 기업에 대한 직원들의 평가를 담고 있는데 최근 기업 리뷰를 이용한 연구가 활발하다 (Choi와 Lee, 2018; Symitsi, 2018; An 등, 2020). 온라인 기업 리뷰는 설문조사에 비해 용이하게 대량의 데이터를 수집할 수 있는 장점이 있으며 리뷰 작성자의 익명성이 비교적 더 잘 보장되므로 기업에 대한 솔직한 의견을 수집할 수 있는 장점이 있다 (Kim과 Lee, 2021). 또한 리뷰에는 비정형 데이터인 텍스트를 포함하고 있는데 텍스트 분석 기법을 활용하여 보다 다양한 정량적 분석이 가능하다 (Kim과 Lee, 2021).

이러한 배경 하에 본 연구에서는 온라인 기업 리뷰 사이트인 잡플래닛 (<http://jobplanet.com>)에 등록된 국내 IT 기업들에 대한 리뷰를 수집하여 IT 기업 직원들의 다양한 만족 및 불만족 속성을 도출하고 이러한 속성들이 이직에 어떠한 영향력을 갖는지 파악하고자 한다. 그리고 이러한 속성들을 정량적으로 활용하여 IT 기업의 이직률을 예측하는 예측모델을 제안하고자 한다. 본 연구의 결과는 IT 기업 직원의 이직을 다차원적으로 분석하고 실무 응용이 가능한 이직 예측 방법론을 제시한다는 점에서 의미가 있을 것으로 판단된다.

본 연구의 구성은 다음과 같다. 이어지는 2절에서는 IT 분야 이직과 온라인 리뷰 및 토픽모델링 관련 선행연구를 살펴보도록 한다. 3절에서는 본 연구의 연구 접근 방법론에 대해 소개하고 4절에서는 연구 결과와 이에 대한 해석을 제시한다. 마지막으로 5절 결론에서는 연구 결과를 요약하고 시사점을 제시하며 본 연구의 한계점과 향후 연구에 대하여 논의한다.

2. 관련 연구

2.1. IT 기업 이직

이직이란 일반적으로 조직의 구성원이 자발적인 동기로 조직을 이탈하는 것을 의미한다 (Bludedorn, 1978). 이직 요인을 설명하기 위한 다양한 이론들이 제시되어왔다. 대표적인 이론에는 조직균형이론 (Theory of Organizational Equilibrium) (March와 Simon, 1958), 기대충족모형 (Met Expectations Model) (Porter와 Steers, 1973), 연결모형 (Linkage Model) (Mobley, 1977), 이직표출모형 (Unfolding Model of Turnover) (Lee와 Mitchell, 1994), 직무임베디드니스 이론 (Job Embeddedness Theory) (Mitchell과 Lee, 2001) 등에 이른다. 이들 이론에서는 조직에 있어서의 자기 효능감, 조직에 대한 기대와 실제 경험 사이 기대 충족 여부, 직무 불만족의 심화, 조직 내에서 이직을 유발하는 사건으로 인한 충동적인 요인 등 다양한 요인들에 착안하였으며 이러한 논의를 필두로 직무만족, 조직몰입 등의 태도 변수에 초점을 둔 연구들이 이직 분야연구의 주류를 이루어 수행되었다. 2000년 이전에 이루어진 이직 관련 연구들을 메타 분석한 Griffeth 등 (2000)의 연구에 따르면 직무 만족, 조직 몰입, 이직 인식과 이직 의도 등이 이직에 큰 영향을 주는 것으로 나타났으며 그 외 인구통계학적인 속성이나 외부환경 요인 등 다양한 요인들이 이직의 선행요인으로 작용함을 보인 바 있다.

IT 분야의 이직 관련 연구는 기존 이직 분야에서 이어져온 논의를 IT 분야의 특성을 고려하여 확장하는 연구들이 진행되어 왔다. Moore (2000)는 IT 업무에서 비롯된 피로 (work exhaustion)를 이직의 주요 요인으로 판단하고, 직무량, 자율성, 보상 공정성, 직무 모호성, 직무 갈등 등의 이직 결정 요인들과 이직의도 사이를 매개하는 요인으로 제시하였다. Jiang과 Klein (2002)은 직원들의 조직에 대

한 기대와 그 기대에 대한 충족 간의 불일치 정도가 직원 만족과 이직의도를 결정하는 요인으로 영향력을 가짐을 보였다. Ahuja 등 (2007)의 연구에서는 IT 인력의 이직의도에 직간접적으로 영향을 미치는 다양한 요인들을 제시하였는데 업무 피로, 조직몰입, 직무자율성, 보상 공정성, 업무와 가족 간 갈등, 과다업무를 포함하는 이직의도 결정 구조모형을 제안한 바 있다. Joseph 등 (2007)은 IT 인력의 이직의도 관련 연구를 메타 분석하였는데 일반적인 이직 선행요인 외에 바운더리 스패닝 활동 (Boundary Spanning Activities)과 IT 분야 경력, 전문성 몰입 등을 IT 분야의 구별된 이직의도 선행요인으로 제시하였다. 중소 IT 기업을 대상으로 논의한 Woo와 Hwang (2010)의 연구에서는 Mitchell and Lee (2001)의 이론에 기초하여 직무 임베디드니스가 보상, 인정, 임파워먼트 (Empowerment), 정보공유, 기술역량 개발, 일과 생활 균형 등의 여섯가지 인적자원관리 활동과 이직의도를 연계하는 매개 요인으로 도입하였다. Kim (2010)의 연구에서는 IT 인력에 대한 임금수준이 높을수록, 그리고 IT 인력의 규모와 내부 노동시장의 위계가 클수록 이직률이 낮아짐을 확인하였다. 또한 IT 기업의 조직문화는 이직률에 직접효과가 있는 동시에 임금과 이직률의 관계에 대한 조절효과가 있음을 보인 바 있다 (Kim, 2010). Jung과 Son (2017)의 연구에서는 IT 종사자가 받는 고유한 직무스트레스에 착안하고 이를 테크노스트레스 (Technostress)라 정의하였다. 그리고 테크노스트레스를 직무, 개인, 조직, 환경 측면의 이직 결정요인과 이직의도 사이를 매개하는 요인으로 제시하였다 (Jung과 Son, 2017). 결과에 따르면 자율성이 높고 테크노스트레스가 높을 때 이직의도가 배가되며 자기효능감과 임금 수준이 높은 상황에서 테크노스트레스가 높을 때 역시 이직의도가 더욱 촉진되는 것을 확인할 수 있었다 (Jung과 Son, 2017).

한편, IT 인력의 특성에 관한 논의도 참고할 필요가 있다. Crepeau (1992)는 독립적인 자율성, 강한 인내력, 신기술 활용 추구, 업무를 통해 조직에 기여하고자 하는 전성, 성장추구 성향을 언급한 바 있으며, Bartol (1982)은 보상에 대한 중요도는 타 업종 대비 낮으나 업무 성과에 합당한 보상을 원하는 성향이 강한 것을 IT 인력의 특성으로 제시하였다. 이를 볼 때, IT 인력은 타 분야 대비 자율적이고 독립적인 업무 처리 문화를 선호하며 개인의 경력과 직무에 대한 관심이 크고, 인정받는 것을 중요시하는 특성을 지닌 것으로 정리할 수 있다. 이들 연구에서 직접적으로 이직에 대해 다루지는 않았으나 IT 분야 이직을 보다 심층적으로 이해하기 위해 함께 고려해야 할 연구들이라 판단된다.

2.2. 온라인 기업 리뷰

최근 온라인 리뷰 플랫폼은 유형의 제품을 비롯하여 무형의 서비스, 콘텐츠 등 점차 다양한 대상에 대해 영역을 넓혀가고 있다. 이러한 흐름 속에 전·현직 직원들이 작성한 기업에 대한 리뷰를 제공하는 기업 리뷰 플랫폼은 기업 내부의 다양한 정보를 제공하기 때문에 취업과 이직을 준비하는 사용자들에게 뿐 아니라 연구들에게도 관심의 대상이 되고 있다. 국내에서는 잡플래닛, 미국에서는 글래스도어 (<http://glassdoor.com>)가 대표적이다. 두 플랫폼은 유사한 형태로 기업을 평가하는데, 크게 숫자 형태의 평점과 텍스트 형태의 리뷰로 구분될 수 있다.

직원들이 기업의 항목별로 평점을 입력하는 수치형 데이터의 경우, 데이터로 활용하기에 용이하고 해석이 직관적이라는 이점 때문에 관련 연구에서 꾸준히 이용되었다. 글래스도어의 경우, Symitsi 등 (2018)은 평점 데이터를 활용해서 기업의 각 항목에 대한 직원들의 만족은 기업성과와 양의 관계를 갖는다는 것을 회귀분석을 통해서 설명했다. Huang 등 (2015)은 가족기업을 대상으로 직원 만족이 기업성과에 미치는 영향을 평점 데이터로 분석했다. 그 결과, 창업자와 함께 근무하는 직원들은 회사를 더 높게 평가하는 반면, 상속자가 운영하는 회사의 경우 직원 만족도가 낮다고 밝혔다. Huang 등 (2020)은 기존 연구에서는 상대적으로 덜 사용된 직원들의 기업 전망에 관한 평가가 기업의 성과를 예측하는 데 도움이 된다고 밝혔으며, 특히 기업에 좋은 소식보다는 나쁜 소식을 더욱 잘 예측한다고 분석했다. 잡플래닛 데이터를 사용한 예시로는, Seo와 Kim (2016)이 평점의 각 항목이 직무만족도에 미치는 영향을

대기업과 중소기업을 구분하여 분석하였다.

비정형 텍스트 데이터 분석기법이 발달함에 따라, 기업 리뷰 플랫폼의 텍스트 리뷰 데이터 역시 최근 들어 연구자로부터 많은 관심을 받고 있다고 할 수 있다. Jung과 Suh (2019)는 잡플래닛의 리뷰 데이터를 바탕으로 토픽모델링을 통해 얻은 토픽들과 리뷰의 감정을 통해서 직무 만족도에 미치는 영향 등을 분석했다. Choi와 Lee (2018) 역시 잡플래닛 리뷰를 바탕으로 내부마케팅 영역별로 포함되는 단어 수의 변화와 기업의 시가총액 변동률 간의 관계를 확인하였다. 결과로써 내부마케팅의 효과는 제한적으로만 기업 성과에 양의 관계를 갖는다고 밝혔다. An 등 (2020)은 잡플래닛 리뷰 데이터를 Word2Vec 모델에 적용하여 기업의 개성을 설명하는 단어들을 추출하였다. 이 단어들을 기반으로 각 기업의 개성을 규명하고 각 기업의 개성과 이직률과의 관계를 설명하였다.

이처럼 기업 리뷰 플랫폼에서 수집한 데이터를 활용한 연구가 점차 발전하고 있는 상황에서 본 연구는 잡플래닛 데이터를 활용하여 IT 기업의 이직 문제에 대하여 고찰해보고자 한다.

2.3. 토픽모델링

토픽모델링은 다수의 문서 속에 있는 토픽들을 계량적으로 추론하는 방법론을 일컫는다. 여러 방법론 중에서 Blei 등 (2003)이 제안한 Latent Dirichlet Allocation (LDA)는 가장 대표적인 토픽모델링 방법론이다 (Cho 등, 2015; Bang과 Moon, 2019). LDA는 문서에서 사용된 단어들의 확률 분포를 기반으로 문서의 생성 과정을 추론하고자 하는 생성 모델이다. LDA에서 각 문서는 문서 집합 전체에 내재하여 있는 각 토픽의 조합이며, 각 토픽은 확률 분포를 가지는 각 단어로 구성된다고 가정한다. 즉, LDA는 하나의 문서를 토픽들의 분포로 표현하며, 각 토픽은 단어들의 분포로 표현한다. LDA는 비지도 학습으로 분류되는 방법론으로써 학습을 위해 다수의 문서 집합만을 필요로 하고, 비정형 텍스트인 문서를 수치 연산과 직관적인 해석이 가능한 토픽들의 벡터 형태로 전환 시켜준다는 특징 때문에 다양한 연구에서 사용되어 왔다.

토픽모델링을 적용하는 연구들은 크게 다음의 3가지 범주로 정리될 수 있다.

첫째, 토픽모델링을 통해 추출한 토픽들의 분포 및 변화를 통해 해당 분야의 트렌드를 파악하고자 하는 연구이다. 대표적으로, Park과 Song (2013)은 학회지에 발표된 논문의 초록을 수집한 뒤 토픽 모델링을 적용하여 국내 문헌정보학계의 연구 동향을 토픽의 변화를 통해서 설명하였다. Kim과 Yun (2016)은 호텔을 이용한 고객들의 리뷰로부터 추출한 토픽들을 통해서 분석 기간의 전기와 후기 사이의 고객 인식 변화를 추정했다.

둘째, 토픽모델링을 통해 얻은 토픽들을 기반으로 새로운 변수를 생성하여 추가적인 분석에 사용하는 연구이다. 예를 들어, Shi 등 (2016)은 기업의 사업 설명서를 수집하여 토픽을 추출한 뒤, 토픽의 비중으로 이뤄진 벡터들 사이의 거리를 계산하여 사업의 유사성을 나타내는 변수를 생성하였다. 이 변수를 이용하여 사업의 유사성이 직원 이동, 투자, 인수합병에 미치는 영향을 분석했다. Lee 등 (2020)은 스마트폰 앱의 기능의 다양성을 측정하기 위해서 앱을 설명하는 텍스트들로부터 추출한 토픽들을 엔트로피 공식에 적용하여 기능의 다양성을 측정하였고, 이를 기반으로 앱 사이의 유사성을 분석하여 유사한 앱들 사이 프로모션 (promotion)의 유효성을 검증하였다.

셋째, 토픽의 분포로 이루어진 벡터를 머신러닝 모델의 입력값으로 사용하여 예측모델을 설계하는 연구이다. Wang과 Xu (2018)는 자동차 보험사기에 관한 딥러닝 기반 예측모델을 개발하면서, 수치형 변수에 더해 자동차 사고 기록문서에서 추출한 토픽에 관한 벡터를 함께 사용하였다. 토픽 벡터는 예측 모델의 정확도를 높이는 데 기여하였다. Yuan 등 (2016) 역시 크라우드펀딩 플랫폼에서의 프로젝트 모금 성공 여부를 예측하기 위해서 프로젝트 기술서로부터 추출한 토픽 벡터를 머신러닝 알고리즘의 입력값으로 사용하였다. 특히 토픽을 추출하는 데 있어서 해당 도메인의 특성을 파악하도록 토픽모델링 모델

을 개선함으로써 예측 정확도를 높였다. Atkins 등 (2018)은 뉴스 기사로부터 추출한 토픽이 주식 시장의 변동성을 예측에 효과적임을 보인 바 있다. 이처럼 토픽모델링에서 추출한 토픽 벡터는 머신러닝 모델의 효과적인 입력값으로써 활용될 수 있다는 것이 다양한 문제에서 확인되었다. 따라서 본 연구에서는 기업 리뷰 데이터로부터 추출하는 토픽 벡터들을 머신러닝 알고리즘의 입력 데이터로 활용한 이직률 예측 모델을 제안하고자 한다.

3. 연구 방법

3.1. 연구 흐름

본 연구는 크게 데이터 수집 및 전처리, 토픽모델링, 머신러닝 모델 구축, 성능평가의 순으로 진행되었다. 기업 리뷰를 포함하여 개별 기업의 재무 데이터와 이직률 데이터를 함께 수집하여 연구에 필요한 데이터 세트를 구축한다. 기업 리뷰에 대해서 전처리한 뒤 토픽모델링을 수행한다. 토픽모델링을 통해 추출된 토픽에 대한 정보와 재무 데이터를 활용하여 다양한 머신러닝 모델을 통해 이직률 예측 모형을 구축한다. 예측하려는 목표변수는 기업의 차년도의 이직률이며 설명변수는 해당 기업에 대해 당해년도에 작성된 리뷰로부터 추출한 토픽 정보와 당해년도 재무 데이터이다. 이를 패널데이터로 구축하여 모델의 학습에 활용한다. 구축한 모델 중 성능 평가를 통해 최적의 모델을 선정하고 이 과정에서 이직에 영향을 주는 만족 및 불만족 요인을 도출하여 제시한다. 이 과정을 도식화하면 Figure 3.1과 같다. 세부적인 절차는 이어지는 절에서 단계별로 설명하도록 한다.

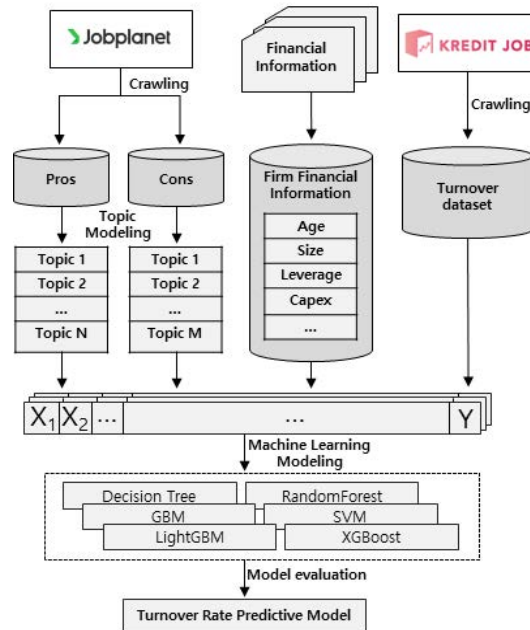


Figure 3.1 Overall process

3.2. 데이터 수집 및 전처리

본 연구에서 수집한 데이터는 모두 세 가지 종류이다.

첫 번째 데이터는 기업 리뷰 데이터로서 잡플래닛 사이트로부터 수집하였다. 잡플래닛은 전·현직원이 자신이 근무 중이거나 근무했던 기업에 대한 리뷰를 남기는 사이트로서 2014년 4월 개설 이래 현재까지 국내 16,834개 기업 및 공공기관에 대한 리뷰를 보유하고 있다. 잡플래닛에 게시된 기업 리뷰 예시는 Figure 3.2에서 보는 바와 같다. 리뷰에는 해당 기업에 대한 5개 분야 별점 및 전체 만족도 별점, 한 줄평과 장점, 단점, 경영지에 바라는 점과 해당 조직의 1년 뒤 전망 및 추천 여부를 포함하고 있다. 본 연구는 IT 기업을 대상으로 하고 있으므로 IT 분야 상장기업에 대한 리뷰를 수집하였는데 2015년부터 2018년까지의 리뷰 21,589건이 해당된다. 본 연구에서는 플랫폼에 의해서 사전에 정의된 기업 평가 항목보다 유연하게 이직에 영향을 미치는 요인들을 파악하기 위해서, 사용자들에 의해서 해당 기업의 장점과 단점에 대해 자유롭게 기술한 텍스트를 연구에 활용하였다. 수집된 리뷰에 대해서는 KoNLPy 라이브러리를 활용하여 전처리를 수행하였다.



Figure 3.2 A review example from Jobplanet

두 번째 데이터는 기업 재무 정보로서 Dataguide DB로부터 수집하였다. 수집 대상이 되는 변수들은 2015년부터 2018년까지에 해당하는 IT 분야 상장기업의 재무 변수 중 기업의 이직률에 영향을 미칠 수 있는 업력, 규모, 부채비율, 대주주 지분, 연구개발, 자본적 지출, 매출액 성장률, 기업가치 등 총 8개 변수를 활용하였다.

마지막 데이터는 각 기업의 이직 정보 데이터로서 크레딧잡 (<http://kreditjob.com>) 사이트로부터 수집하였다. 해당 사이트에서는 국내 기업들의 월별 입사자 및 퇴사자 수에 대한 정보를 제공하고 있는데 이를 활용하여 2016년부터 2019년까지 각 기업의 연도별 이직률을 계산하여 연구에 활용하였다.

3.3. 토픽모델링

잡플래닛에서 수집된 장점과 단점 리뷰로부터 기업에 대한 직원의 만족 요인과 불만족 요인을 추출하기 위하여 토픽모델링을 진행하였다. gensim 라이브러리를 활용하여 LDA 모델을 학습시켰다. 우선 적정 토픽 개수를 판단하기 위하여 Figure 3.3과 같이 토픽 개수별 응집도 (coherence)를 산출한 결과 만족 요인의 경우 8개, 불만족의 경우 7개로 나타났다.

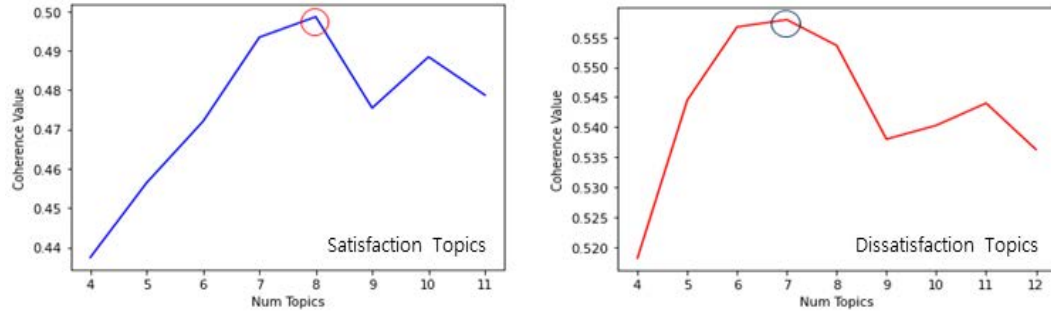


Figure 3.3 Coherence score

Table 3.1과 Table 3.2는 각각 토픽모델링을 통해 추출된 주요 키워드들과 해당 키워드의 토픽 내 비율을 보여준다. 연구진에서는 토픽별 키워드들의 구성과 함께 각 토픽에 해당하는 리뷰들을 검토하여 각 토픽에 적합한 토픽명을 부여하였다.

Table 3.1 Satisfaction topics

Probability words (Ratio)	Topic
분위기 (atmosphere) (15.4%), 업무 (task) (11.5%), 자유로움 (freedom) (6.5%), 팀 (team) (5.5%), 부서 (department) (4.8%)	Flexible work atmosphere
일 (work) (8.4%), 사람 (person) (7.5%), 개발자 (developer) (1.8%), 능력 (ability) (1.5%), 경력 (career) (1.5%), 실력 (skill) (1.1%), 인정 (recognition) (1.0%)	Fairness of rewards and recognition
복지 (welfare) (15.7%), 직원 (employee) (6.4%), 지원 (support) (4.6%), 제도 (policy) (1.9%), 혜택 (benefit) (1.6%)	Well-established welfare system
회사 (company) (10.8%), 연봉 (annual salary) (6.7%), 장점 (pros) (5.1%), 대기업 (big company) (3.8%), 수준 (level) (2.9%), 급여 (salary) (2.1%)	Satisfactory salary
다양 (variety) (5.0%), 경험 (experience) (3.7%), 기회 (opportunity) (3.2%), 개발 (develop) (2.8%), 교육 (training) (2.4%), 사업 (business) (2.2%)	Opportunities for growth
연차 (leave) (10.1%), 가능 (possible) (8.2%), 사용 (use) (6.0%), 야근 (overtime work) (5.4%), 휴가 (vacation) (4.2%), 눈치 (pressure to get permission which is already permitted) (3.0%), 칼퇴 (leaving work on time) (2.3%)	Free atmosphere to work
기업 (company) (8.2%), 문화 (culture) (6.2%), 수평 (horizontality) (3.2%), 조직 (organization) (1.3%), 구조 (structure) (1.1%)	Horizontal organizational culture
근무 (work) (5.7%), 환경 (environment) (4.2%), 제공 (offer) (4.1%), 카페 (cafe) (2.0%), 자율 (autonomy) (1.6%), 복장 (dress code) (1.6%), 카페테리아 (cafeteria, 1.5%)	Employee-friendly work environment

만족 토픽은 8개로 분류되었으며 유연한 업무 분위기, 능력에 따른 공정한 보상과 인정, 우수한 복지 제도, 만족할만한 연봉, 성장을 위한 기회 제공, 눈치 안보는 자유로운 분위기 (자유로운 연차 사용 및 정시 퇴근 등), 수평적인 조직문화, 직원 친화적인 근무환경 (자율출퇴근, 복장 규정 없음, 직원을 위한 음료 제공 등) 등이었다. 불만족 토픽은 7개로 분류되었으며 불확실한 성장 전망, 열악한 업무 프로세스와 조직구조, 사내 정치와 불공정한 보상, 낮은 연봉과 미흡한 복리후생, 끈대 문화, 불합리한 업무 배분, 보수적인 조직문화 등으로 나타났다.

Table 3.2 Dissatisfaction topics

Probability words (Ratio)	Topic
개발 (development) (3.1%), 단점 (cons) (2.9%), 사업 (business) (2.8%), 개발자 (developer) (2.1%), 프로젝트 (project) (1.8%), 발전 (progress) (1.4%)	Unclear growth outlook
체계 (system) (2.6%), 업무 (task) (2.6%), 조직 (organization) (2.4%), 구조 (structure) (2.3%), 보고 (report) (2.0%), 부족 (lack) (1.9%)	Poorly implemented business process and organizational structure
직원 (employee) (6.5%), 급 (position) (2.1%), 능력 (ability) (1.7%), 정치 (politics) (1.4%), 승진 (promotion) (1.2%), 라인 (connection) (0.7%)	Politics at work and unfair reward
연봉 (salary) (10.9%), 야근 (overtime work) (6.8%), 복지 (welfare) (5.3%), 시간 (time) (2.8%), 근무 (work) (2.1%), 퇴근 (leaving work) (1.6%)	Low salary and poor welfare
사람 (person) (8.9%), 일 (work) (7.8%), 생각 (opinion) (1.3%), 상사 (senior employee) (1.0%), 끈대 (senior employees who believe they are always right) (1.0%)	Uncomfortable relationship with stubborn senior employees
업무 (task) (9.7%), 팀 (team) (7.1%), 부서 (department) (6.1%), 분위기 (atmosphere) (5.5%), 강도 (2.1%)	Unreasonable work allocation
회사 (company) (8.6%), 문화 (culture) (4.0%), 보수 (conservatism) (2.3%), 대기업 (big company) (2.0%), 경영진 (board of directors) (1.3%)	Hierarchical work culture

3.4. 머신러닝 모델 적용

기업의 이직률 예측을 위한 머신러닝 모델을 구축하기 위하여 데이터를 학습 가능한 형태로 바꿀 필요가 있다. 우선 예측하고자 하는 이직률은 기업의 차년도 이직률 4개년치값(2016년 2019년)을 이용하였다. 학습에 사용하는 데이터는 토픽모델링 결과 데이터와 재무 변수들을 연도별로 구성하여 패널데이터(2015년 2018년)로 구성하였다. 당해연도 리뷰 데이터와 재무 변수들로 해당 기업의 차년도 이직률을 예측하는 것이다. 토픽모델링 결과 데이터의 경우, 3.3절에서 도출된 리뷰 당 만족 및 불만족 토픽들의 비중을 기업별로 평균값을 구하여 기업당 만족 토픽 8개와 불만족 토픽 7개에 대한 평균값을 변수로 구성하였다. 재무 변수들은 3.2절에서 언급한 8개 변수를 사용하였다. 잡플래닛에 IT 기업으로 분류된 전체 기업 중 재무 데이터 수집이 어려운 비상장사를 제외하고 총 129개 기업이 분석 대상이 되었다. 소수 의견이 과대표되는 것을 줄이기 위해서 리뷰 수가 연 30개 미만인 기업의 해당 연도 데이터를 제외하였으며, 이직률이 50% 이상인 특수한 경우에 대해서는 이상치로 판단하여 제거하였다. 최종적으로 남은 총 489건의 데이터를 학습에 사용하였다.

이직률은 선형적인 값이므로 회귀 모형을 위한 알고리즘을 적용하였다. 적용한 알고리즘에는 Decision Tree, Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (GBM), Random Forest, Support Vector Machine (SVM), eXtreme Boosting (XGBoost) 등 총 여섯 가지였으며 scikit-learn, lightgbm 및 xgboost 파이썬 라이브러리를 활용하였다.

4. 연구결과

4.1. 모델 성능평가

제안한 여섯 개 예측 모형에 대해 MAE와 RMSE를 기준으로 성능을 평가하였다. scikit-learn 라이브러리의 k-Fold 교차검증 (cross validation)을 사용해서 5개의 fold로 검증하였다. 성능 비교를 위하여 토픽모델링에서 추출한 변수를 사용하지 않고 재무 변수만으로 학습시킨 모델에 대해서도 결과를 제시하였다. Table 4.1과 Figure 4.1에서 보는 바와 같이 재무변수만을 사용한 모델 대비 본 연구에서 제안한 모델들이 알고리즘의 종류와 상관없이 보다 나은 성능을 보이는 것을 확인할 수 있었다.

모델 중에서는 Light GBM 기반의 모델의 MAE가 0.04690으로 가장 우수한 성능을 보였다. 달리 말하면 본 연구에서 제시한 모델을 통해 4.69%의 오차율로 IT 기업 이직률을 예측할 수 있는 것으로 해석할 수 있다. 참고로 분석 대상 IT 기업들의 이직률 평균은 21.08%였으며 표준편차는 8.8625%였다.

Table 4.1 Performance evaluation

Model	Financial Features only		Financial Features with Topic Features	
	MAE	RMSE	MAE	RMSE
Decision Tree	0.06274	0.08143	0.05277	0.07046
GBM	0.06070	0.07750	0.05456	0.07346
Light GBM	0.05995	0.07721	0.04690	0.06577
Random Forest	0.05958	0.07710	0.04883	0.06570
SVM	0.06131	0.07850	0.05936	0.07527
XGBoost	0.06165	0.07845	0.05143	0.06916

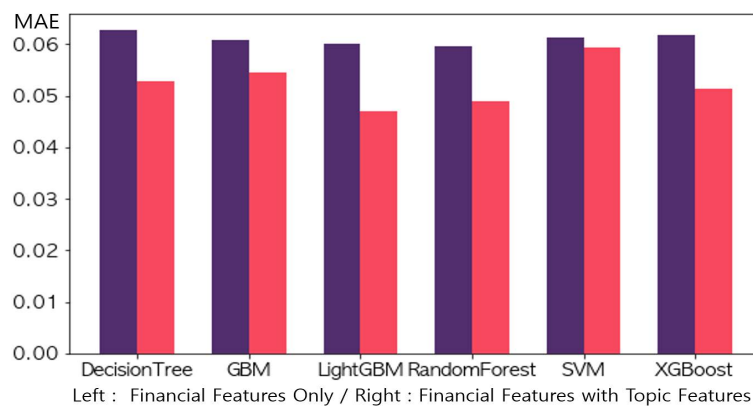


Figure 4.1 Model performance comparison

4.2. 변수 중요도 평가

트리 기반 모델에서 변수 중요도는 트리가 분할 (split)될 때 정보이득 값을 합산하여 구할 수 있으며 값이 클수록 중요도가 높다. 4.1의 모델 성능평가에서 가장 우수한 성능을 보였던 모델에 대하여 lightgbm 라이브러리에 있는 feature importance 모듈을 이용하여 변수 중요도를 계산하였다. Figure 4.2에서 보는 바와 같이 이직률에 영향을 미치는 만족 요인과 불만족 요인을 영향력의 크기순으로 추출할 수 있었다. 결과에 따르면 이직을 억제하는 요인은 만족할만한 연봉, 능력에 따른 공정한 보상과 인

정, 성장을 위한 기회 제공, 우수한 복지 제도 등의 순이었으며, 이와 반대로 이직을 유발하는 불만족 요인은 사내 정치와 불공정한 보상, 열악한 업무 프로세스와 조직구조, 불투명한 성장 전망, 낮은 연봉과 미흡한 복리후생 순으로 나타났다.

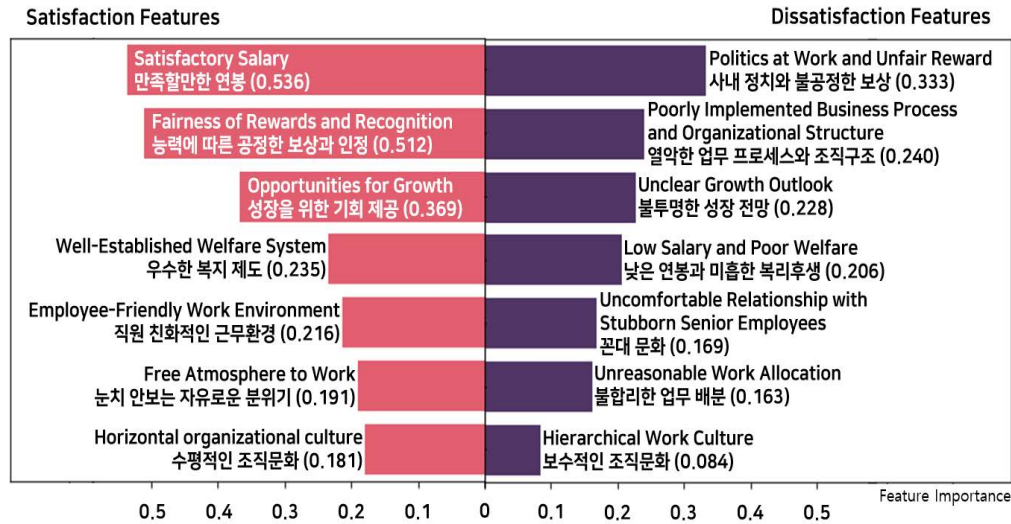


Figure 4.2 Feature importances

이직에 영향을 미치는 만족 요인과 불만족 요인의 영역은 상당수 유사한 영역들이 있었는데 연봉, 공정한 보상, 성장 기회 및 전망 등의 요인은 각각 만족과 불만족 영역에서 이직에 대한 영향력이 큰 것으로 나타났다. 반면 업무 프로세스와 조직구조의 적절성 혹은 업무 배분의 공정성과 같은 요인은 불만족 요인에서만 식별되었다. 또한 변수 중요도의 전반적인 크기를 비교할 때 만족 요인으로 인한 이직 억제 효과가 불만족 요인으로 인한 이직 유발 효과보다 대체로 더 큰 것을 확인할 수 있었다.

5. 시사점 및 결론

본 연구에서는 IT 기업 리뷰 데이터에 토픽모델링 기법을 적용하고 이를 기반으로 기업의 차년도 이직률을 예측하는 머신러닝 모형을 제안하였다. 또한, 이 과정에서 IT 인력의 이직률에 영향력을 갖는 직원 만족 및 불만족 요인을 각각 도출하여 제시하였다. 본 연구는 온라인에서 수집되는 대량의 리뷰 데이터를 활용한 시의성 있는 연구로서 비정형 데이터 분석기법을 인사 분야에 적용함으로써 지금까지의 관련 연구와는 다른 접근 방법으로 IT 분야 이직에 대한 논의의 폭을 확장시켰다는 점에서 의미를 갖는다. 또한 이직의 선행 요인을 만족과 불만족으로 구분하여 이직 유발 및 이직 억제 요인을 입체적으로 제시했다는 점에서도 시사점을 갖는다. 실무적인 관점에서 볼 때 IT 기업에서 이직률을 사전 예측할 수 있다면 기업 입장에서 인력 수급에 대한 계획과 운영에 의사결정 지원도구로서 유용하게 활용할 수 있다. 또한 이직을 예측치를 기업의 현 인력 관리 체계에 대한 하나의 진단 도구로 활용하여 이직으로 인한 기업 경쟁력 저하를 최소화할 수 있는 기업 운영의 가이드를 얻을 수 있을 것으로 판단된다.

본 연구의 한계점과 향후 연구에 대한 제안은 다음과 같다. 선행 연구들이 소규모의 피설문자를 대상으로 일회성의 설문을 이용하여 연구를 진행한 반면, 본 연구에서는 여러 해에 걸쳐 작성된 대량의 리뷰를 시계열 데이터로 분석하였다는 점에서 차별점이 있었으나 수집 데이터의 한계로 장기간의 데이터를

연구에 활용하지는 못하였다. 후속 연구에서 추가적인 데이터를 수집한다면 보다 더 정교한 모델을 구현할 수 있을 것으로 보인다. 또한 본 연구에서는 IT 기업에 종사하는 직원 전체를 대상으로 분석을 진행하였으나 IT 기업 내에도 다양한 직무가 존재하고 직무별 특성이 다르므로, 향후 연구에서는 IT 인력의 세부 직무를 고려하여 분석할 것을 제안한다. 끝으로 IT 분야 이직에 관한 전문가들의 의견과 본 연구결과의 비교 분석이 가능하다면 의미있을 것이라 판단된다.

References

- Ahuja, M. K., Chudoba, K. M., Kacmar, C. J., McKnight, D. H. and George, J. F. (2007). IT road warriors: Balancing work-family conflict, job autonomy, and work overload to mitigate turnover intentions. *MIS Quarterly*, **31**, 1-17.
- An, B., Choi, J. and Shu, Y. (2020). Impact of corporate personality on the relationship between job satisfaction and turnover rate: based on the corporate review of Job-planet. *Journal of Information Technology Services*, **19**, 35-56.
- Atkins, A., Niranjana, M. and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, **4**, 120-137.
- Bang, H. S. and Moon, H. S. (2019). A study on the methodology to express the main topics of text in time series using text mining. *Journal of the Korean Data & Information Science Society*, **30**, 1259-1276.
- Bartol, K. M. and Martin, D. C. (1982). Managing information systems personnel : A review of the literature and managerial implications. *MIS Quarterly*, Special Issue, 49-70.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993-1022.
- Bludedorn, A. C. (1978). A taxonomy of turnover. *Academy of Management Review*, **3**, 647-651.
- Cho, A., Lee, K. and Cho, W. (2015). Latent mobility pattern analysis of bus passengers with LDA. *Journal of the Korean Data & Information Science Society*, **26**, 1061-1069.
- Choi, K. C. and Lee, S. Y. (2018). An analysis of the internal marketing impact on the market capitalization fluctuation rate based on the online company reviews from Jobplanet. *Information Systems Review*, **20**, 39-62.
- Couger, J. D., Adelsberger, H., Borovits, I., Zviran, M. and Motiwalla, J. (1990). Commonalities in motivating environments for programmer/analysts in Austria, Israel, Singapore, and the U.S.A. *Information & Management*, **18**, 41-46.
- Crepeau, R. G., Crook, C. W., Goslar, M. D. and McMurtrey, M. E. (1992). Career anchors of information systems personnel. *Journal of Management Information Systems*, **9**, 145-160.
- Griffeth, R. W., Hom, P. W. and Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover : Update, moderator tests, and research implications for the next millennium. *Journal of Management*, **26**, 463-488.
- Huang, K., Li, M. and Markov, S. (2020). What do employees know? Evidence from a social media platform. *The Accounting Review*, **95**, 199-226.
- Huang, M., Li, P., Meschke, F. and Guthrie, J. P. (2015). Family firms, employee satisfaction, and corporate performance. *Journal of Corporate Finance*, **34**, 108-127.
- Jiang, J. and Klein, G. (2002). A discrepancy model of information systems personnel turnover. *Journal of Management Information Systems*, **19**, 249-272.
- Joseph, D., Ng, K., Koh, C. and Ang, S. (2007). Turnover of IT professionals : A narrative review, meta-analytic structural equation modeling, and model development. *MIS Quarterly*, **31**, 1-31.
- Jung, D. M. and Son, D. H. (2017). Comprehensive approach to job turnover of it workforce - perspectives of technostress control model. *The Journal of Internet Electronic Commerce Research*, **17**, 127-142.
- Jung, Y. and Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems*, **123**, 1-12.
- Kim, B. S. (2010). The determinant of turnover behavior of it professionals - Focusing on the direct and moderating effects of pay level and trust culture. *Journal of Human Resource Management Research*, **17**, 177-196.
- Kim, C. and Bae, B. (2018). An analysis of factors influencing the turnover intension of it service employees in SMEs : Focusing on the moderating effect of congruity between leadership-organizational culture. *Journal of Information Technology Services*, **17**, 1-16.

- Kim, G. and Yun, H. (2016). Topic modeling approach to understand changes in customer perceptions on hotel services in Seoul. *Journal of Korea Service Management Society*, **17**, 217-231.
- Kim, H. E. and Lee, H. (2021). Online employee reviews and firm value. *Journal of Management and Economics*, **43**, 27-52.
- Kwon, M. J. and Park, S. C. (2010). Examination of the determinants of SW manpower' turnover intention : Testing the mediating role of job satisfaction. *Journal of Information Technology Services*, **9**, 73-90.
- Lee, G. M., He, S., Lee, J. and Whinston, A. B. (2020). Matching mobile applications for cross-promotion. *Information Systems Research*, **31**, 865-891.
- Lee, T. W. and Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary employee turnover. *Academy of Management Review*, **19**, 51-89.
- March, J. G. and Simon, H. A. (1958). *Organizations*, Wiley, New York.
- Mitchell, T. R. and Lee, T. W. (2001). The unfolding model of voluntary turnover and job embeddedness: Foundations for a comprehensive theory of attachment. *Research in Organizational Behavior*, **23**, 189-246.
- Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, **62**, 237-240.
- Moore, J. E. (2000). One road to turnover: An examination of work exhaustion in technology professionals. *MIS Quarterly*, **24**, 141-148.
- Park, J. H. and Song, M. (2013). A study on the research trends in library & information science in Korea using topic modeling. *Journal of the Korean Society for Information Management*, **30**, 7-32.
- Porter, L. W. and Steers, R. M. (1973). Organizational work and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, **80**, 151-176.
- Seo, W. C. and Kim, H. J. (2016). A study on the internal reputation factors affecting the job satisfaction: Focusing on big data analysis in the social media for corporation reputation. *Journal of Digital Contents Society*, **17**, 295-305.
- Shi, Z., Lee, G. M. and Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, **40**, 1035-1056.
- Staw, B. M. (1980). The consequence of turnover. *Journal of Occupational Behavior*, **1**, 253-273.
- Symitsi, E., Stamolampros, P. and Daskalakis, G. (2018). Employees' online reviews and equity prices. *Economics Letters*, **162**, 53-55.
- Wang, Y. and Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, **105**, 87-95.
- Woo, H. J. and Hwang, K. T. (2010). A study on the human resource practices and turnover intention of it professionals software developers in small and medium it enterprises. *Journal of Information Technology Applications & Management*, **17**, 47-69.
- Yuan, H., Lau, R. Y. and Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, **91**, 67-76.

Turnover rate prediction among IT firms according to job satisfaction and dissatisfaction factors: Using topic modeling and machine learning[†]

Jinwook Choi¹ · Dongwon Shin² · Hanjun Lee³

¹Smart media Service Research Center, Korea University

²Industry Academic Cooperation Foundation, Myongji University

³Department of Management Information Systems, Myongji University

Received 27 August 2021, revised 7 September 2021, accepted 7 September 2021

Abstract

Job turnover in the IT industry is a major challenge for companies in technology accumulation, development, and management, and thus, research on this is essential. However, there are still few studies related to turnover and personnel management in the IT field. This study examines the satisfaction and dissatisfaction factors of employees that affect the turnover rate and proposes a model for predicting the turnover rate of IT companies using these factors. To this end, we collected 21,589 reviews from employees of 129 IT companies listed on the domestic stock market on Jobplanet, an online company review site, and conducted topic modeling. Using topics extracted, machine learning-based predictive models for turnover rate were proposed. In addition, this study analyzed the degree of influence of each employee satisfaction and dissatisfaction factor on the turnover rate through variable importance evaluation.

Keywords: Employee satisfaction, firm review, machine learning, topic modeling, turnover rate.

[†] This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2019R1G1A1084863).

¹ Research professor, Smart Media Service Research Center, Korea University, Seoul 02841, Korea.

² Researcher, Industry Academic Cooperation Foundation, Myongji University, Yongin 17058, Korea.

³ Corresponding author: Assistant professor, Department of Management Information Systems, Myongji University, Seoul 03674, Korea. E-mail: hjlee1609@gmail.com