ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 March: 21(1): 65~82 http://dx.doi.org/10.13088/jiis.2015.21.1.65

항공산업 미래유망분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석*

김현정
이화여자대학교 경영대학
(charitas@empal.com)

조남옥
이화여자대학교 경영대학
(namok.jo@gmail.com)

신경식
이화여자대학교 경영대학
(ksshin@ewha.ac.kr)

최근 경제적·사회적 부가가치를 창출할 수 있는 유망분야를 선정하여 국가 전략 및 정책 수립 시 반영하기 위해 미래 핵심 이슈를 발견하고 트렌드를 분석하는 것에 대한 관심이 급증하고 있다. 기존에는 미래의 핵심 기술이나 이슈를 발견하고 트렌드 분석을 통해 미래유망분야를 선정하는 연구를 위해 문헌 조사 또는 전문가 평가와 같은 정성적 연구방법이 사용되어 왔다. 그러나 이 연구방법은 대량의 정보로부터 결과를 도출하는데 많은 시간과 비용이 소요될 뿐만 아니라 전문가의 주관적인 가치가 반영될 가능성이 존재한다. 이와 같은 한계점을 보완하고자 최근 국토교통, 안전, 정보통신기술 등 다양한 분야에서 미래유망분야를 선정하기 위하여 정성적 연구방법에 텍스트 마이닝과 같은 정량적 연구방법을 상호 보완적으로 활용하는 방식으로 트렌드 분석을 수행하는 연구 방법론의 패러다임 변화가 시도되고 있다.

본 연구는 항공산업 전반적인 분야에 빅데이터 분석 방법인 텍스트 마이닝 기법을 적용하여 항공 분야의 연구동향을 파악하고 미래유망분야를 전망하였다. 텍스트 마이닝 기법 중하나인 토픽 분석을 이용하여 항공산업 전반적인 분야의 문서 집합 내 잠재된 토픽을 추출하고, 연도별로 핵심 토픽의 추이를 분석하였다. 분석 결과 항공산업의 미래유망분야로 항공안전정책, 항공운임(저가항공), 그리고 친환경 고연비 연료가 도출되었다. 본 연구결과는 분석 대상을 논문에 한정하여 수행하였다는 한계점이 존재하나, 항공산업 분야의 핵심 이슈를 도출하기 위하여 텍스트 마이닝 기반의 트렌드 분석에 대한 활용가능성을 제시하고, 미래유망분야를 선정하기 위한 정량적인 분석 방법론의 전형을 마련하였다는 점에서 의의가 있다.

주제어 : 항공, 빅데이터 분석, 텍스트 마이닝, 토픽 분석, 트렌드 분석

논문접수일 : 2014년 11월 12일 논문수정일 : 2014년 12월 16일 게재확정일 : 2014년 12월 18일
투고유형 : 국문급행 교신저자 : 신경식

1. 서론

최근 미래 핵심 이슈를 발견하고 트렌드를 분석하는 것에 대한 관심이 급증하고 있다. 이는 경제적·사회적 부가가치를 창출할 수 있는 유망분야를 선정하여 국가 전략 및 정책 수립 시 반영하기 위한 것이다. 미래 예측에 관한 핵심 유망분야 선정과 같은 연구는 대부분 문헌 조사, 전문가 평가, 델파이(Delphi) 기법(Dalkey and Helmer, 1963)과 같은 정성적(Qualitative) 연구방법이 사용되어 왔다. 이는 대량의 정보로부터 결과를 도출하는데 많은 시간과 비용이 소요될 뿐

* 이 논문은 2014년도 한국교통연구원의 지원을 받아 수행한 '항공교통분야 빅데이터 활용을 위한 기초연구' 연구용역 보고서 일부를 발췌하여 작성하였음.
이 논문은 2013년도 정부재원(교육부)으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013S1A3A2054667).

만 아니라 전문가의 주관적인 가치가 반영될 가능성이 있다는 한계점이 존재한다.

기존 정성적 연구방법의 한계를 보완하고자 국토교통(Korea Agency for Infrastructure Technology Advancement, 2013), 안전(Korea Institute of science and technology Evaluation and Planning, 2014), 정보통신기술(Chung and Lee, 2012), 건설업(Jeong and Kim, 2012; Korea Institute of science and technology Evaluation and Planning, 2010), 철강산업(Min et al., 2014) 등 다양한 분야에서 정성적 연구방법과 함께 텍스트 마이닝(Text Mining) 등과 같은 정량적(Quantitative) 연구방법을 상호보완적으로 연구에 채택하여 활용하는 방식으로 미래의 트렌드 분석을 수행하는 연구 방법론의 패러다임이 변화하고 있다. 이는 대량의 데이터를 기반으로 객관성 및 실효성 높은 트렌드 분석 결과를 확보하기 위한 것이라 할 수 있다. 그러나 기수행되었던 유관 연구들의 경우 단순히 키워드 수준에서 연구를 수행하였기 때문에 단일 키워드로부터 의미를 파악하기에는 트렌드 분석이나 핵심 이슈 선정이 어렵다. 또한, 항공산업과 같은 특수한 분야에서의 핵심 기술이나 이슈를 도출하고 연구동향과 트렌드를 분석한 연구는 전무한 실정이다.

따라서 본 연구에서는 항공산업 전반적인 분야의 미래 핵심 유망분야를 선정하기 위해 빅데이터 분석 방법 중 하나인 텍스트 마이닝을 사용하여 비정형 텍스트 기반의 객관적 데이터에 근거한 정량적 분석을 수행하고자 한다. 키워드 수준에서의 트렌드 분석이 갖는 단일 단어의 의미 파악의 어려움을 보완하기 위해 텍스트 문서 집합 내에 잠재되어있는 전반적 주제를 도출하는 기법인 토픽 분석을 수행하여 키워드들의 집합인 토픽을 도출하고, 이를 통해 트렌드 분석을 실시하고자 한다. 본 연구는 항공산업의 미래유망분야 선정을 위하여 텍스트 마이닝을 도입한 초기 단계의 시도이기 때문에 핵심 키워드 도출 및 트렌드 분석을 위한 연구대상을 전문 학술자료인 논문에 한정하여 분석한다. 향후에는 논문뿐만 아니라 국내외 항공 관련 뉴스 기사, 항공입찰정보 등으로 연구대상을 확대하여 분석 결과의 신뢰성을 고취시키고자 한다.

텍스트 마이닝 결과를 기반으로 항공산업 핵심 키워드 및 토픽을 파악하고, 시계열적으로 어떻게 변화하고 있는지에 대한 추세를 파악하기 위해 연도별로 핵심 토픽에 대한 트렌드 분석을 실시함으로써 미래 항공산업 유망분야 발굴하고, 이에 신속한 대응 체계를 마련하고자 한다. 추가적으로 정부의 정책연구 보고서를 대상으로 텍스트 마이닝을 통해 키워드를 추출하여 실제 정부가 예산을 투자하고 연구분야를 분석하였다. 분석 결과를 토대로 항공산업의 선정된 미래유망분야와 정부가 예산을 실제 투자하고 있는 분야와의 일치 정도를 비교해 보고자 한다. 이와 같은 추가 분석결과는 향후 항공산업 분야의 정책 및 예산 수립 등에 활용 가능할 것으로 기대된다.

2. 텍스트 마이닝

텍스트 마이닝(Feldman and Dagan, 1995)이란 자연어로 구성된 비정형 텍스트 데이터(Unstructured Text Data)에서 숨겨진 패턴 또는 관계를 추출하여 의미 있고 활용 가치가 높은 정보 또는 지식을 찾아내는 분석 기법으로 자연어 처리(Natural Language Processing) 기술을 기반으로 한다. 데이터 마이닝(Data Mining)이 정형 데

이터(Structured Data)에서 패턴을 찾아내는 기술인 반면에 텍스트 마이닝은 비정형 텍스트 데이터에서 의미 있는 지식을 찾아내는 기술이라고 할 수 있다. 대량의 비정형 텍스트 데이터에서 의미 있는 정보를 추출하고, 기존에 정형 데이터로부터 추출한 정보와의 연계성을 파악하며, 개별 텍스트가 속한 범주(Category)를 찾아내는 등 정보 검색(Information Retrieval)의 단순한 기능을 넘어선 분석 기법으로 다양한 분야에서 활용되고 있다.

2.1. 텍스트 데이터 수집

텍스트 마이닝을 수행하기 위한 첫 번째 단계는 분석을 위한 원천(Source)을 선정하여 데이터를 수집하는 것이다. 오피스(Office) 문서, 메일, 게시판, 뉴스, 블로그, 소셜 네트워크 서비스(Social Network Service, SNS) 등 다양한 원천으로부터 텍스트 문서의 수집이 가능하며, 분석 목적에 맞는 데이터를 선택한다.

2.2. 형태소 분석

형태소 분석 단계에서는 문서 내에 표현되어 있는 단어, 구, 절에 해당하는 내용을 형태소 분석 처리 과정으로 데이터를 표현하는 단계이다. 형태소 분리 및 품사 부착(Part-Of-Speech Tagging)을 수행한다. 연구 목적을 고려하여 적합한 품사가 부착된 단어들을 추출한다.

2.3. 의미정보 변환 및 추출

형태소 분석 실시 후에는 텍스트 내에 숨겨져 있는 패턴 및 경향 분석이 가능하도록 분석에 사용될 의미 있는 단어를 선별한다. 단어 필터링(Filtering)을 위하여 분석 대상 최소 문서 수 결정, 불용어(Stopword) 처리, 어간 추출(Stemming), 단어별 가중치 산출 등의 작업을 수행한다. 먼저, 해당 단어를 포함한 문서의 수가 최소 n개 이하인 경우에 제거한다. 해당 단어를 포함한 문서의 수에 대한 정해진 규칙이 없기 때문에 실험을 통하여 탐색적으로 결정한다. 정관사와 같이 의미 파악이 어려운 단어 및 해당 도메인에서 사용되지 않는 단어 등을 불용어로 처리한다. 또한, 어간(Stem) 파악을 통해서 동일한 어간을 가지고 있는 단어는 하나의 단어로 처리하여 텍스트 처리의 효율성을 높인다.

앞서 처리된 텍스트 데이터를 의미 정보로 저장하기 위해 단순히 단어별 빈도를 이용하기보다는 정보 검색과 텍스트 마이닝 관련 연구에서 범용적으로 사용되고 있는 TF-IDF(Term Frequency-Inverse Document Frequency, 단어 빈도-역문서 빈도)를 고려하여 각 단어별 가중치를 산출한다(Salton and McGill, 1983). 이는 여러 문서 집합으로부터 특정 단어가 특정 문서에서 얼마나 중요한가를 판단할 수 있는 값이다. TF(Term Frequency, 단어 빈도)는 특정 단어가 문서 내에 얼마나 빈번하게 등장하는지를 나타내는 값이다. 보통 이 값이 클수록 문서에서 중요한 단어라고 판단할 수 있으나, 문서 집합 내에서 빈번하게 사용되는 것은 그 단어가 흔하다는 것을 의미한다. 따라서 단어 빈도뿐만 아니라 DF(Document Frequency, 문서 빈도)의 역수를 취한 IDF(Inverse Document Frequency, 역문서 빈도)를 고려한다. IDF는 특정 단어가 문서 집합 내에서 얼마나 공통적으로 출현하는지를 나타내는 값으로, 전체 문서 빈도수를 특정 단어를 포함한 문서의 빈도수로 나눈 뒤 로그를 취하여 계산된다. TF-IDF는 TF와 IDF 값을 곱한 값으로 다음과 같

이 계산된다.

$$TF\text{-}IDF = TF \times log(N/DF)$$

TF = 문서 내 특정 단어의 빈도수
N = 전체 문서 빈도수
DF = 여러 문서 내 특정 단어 빈도수
IDF = DF의 역수

형태소 분석 및 의미정보 변환 및 추출과 같은 텍스트 전처리 분석을 통해 비 구조화된 문서 집합은 구조화된 단어 문서 행렬(Term-Document Matrix)로 표현된다.

2.4. 패턴 및 경향 분석

최종 선정된 의미 정보로 표현된 담어 문서 행렬을 기반으로 문서를 분류(Classification)하거나 군집화(Clustering)하는 등을 통해 정보를 재생산하는 단계이다. 비정형 텍스트 문서 집합이 분석 가능한 형태로 구조화되면, 인공신경망(Artificial Neural Network), 의사결정나무(Decision Tree), SVM(Support Vector Machine) 등 데이터 마이닝(Data Mining) 기법 등을 활용하여 문서를 분류하거나 텍스트 군집분석 또는 토픽 분석을 통해 유사한 성격을 갖는 문서들을 군집화한다. 텍스트 군집 분석 및 토픽 분석은 텍스트 문서 집합 내에 숨겨진 군집(Cluster) 또는 주제(Topic)를 도출하는 기법으로 문서 집합 내에서 연관성이 높은 단어들을 기준으로 유사한 문서의 군집화를 수행한다. 단어간 연관성은 문서 집합내 동시 출현 빈도를 기준으로 계산된다. 텍스트 군집분석을 위한 방법으로는 계층적 응집 군집 분석(Hierarchical Agglomerative Clustering), EM 알고리즘(Expectation-Maximization Algorithm) 등이 있다. 토픽 분석을 위한 방법으로는 토픽 분석의 시초가 된 방법인 Deerwester et al.(1990)의 LSA(Latent Semantic Analysis)가 있으며, 이후 확률적 개념을 도입하여 Hofmann(1999)이 PLSA (Probablistic Latent Semantic Analysis)를 제안하였다. 최근에는 Blei et al.(2003)가 제안한 LDA(Latent Dirichlet Allocation)가 다양한 분야에서 사용되고 있다.

기존의 텍스트 군집 분석은 개별 문서가 하나의 주제에만 해당된다는 것을 가정하기 때문에 대량의 문서에 대해 전반적인 주제를 추출할 수 없다는 한계점이 존재하였다. 이에 반해 토픽 분석은 복합적인 주제를 포함한 개별 문서가 여러 토픽을 다룰 수 있다는 점을 가정한다. 하나의 군집 또는 토픽은 몇 개의 키워드의 집합으로 표현되고, 각 군집 또는 토픽에 대한 명명(Naming)은 연구자 또는 도메인 전문가가 결정한다.

최근 다양한 분야에서 대량의 텍스트 데이터 내에 존재하는 주요 이슈를 도출하기 위해 텍스트 마이닝 기법 중에서도 토픽 분석을 활용한 연구가 진행되고 있다. Jeong et al.(2013)은 사회 문제를 다루고 있는 대용량 뉴스기사로부터 LDA 기반의 토픽 분석을 적용하여 사회적 이슈에 관한 키워드를 도출하는 시스템을 제안하였다. Park and Song(2013)은 국내 문헌정보학 관련 연구의 동향을 분석하기 위하여 문헌정보학 분야의 주요 학술지에 게재된 논문을 대상으로 LDA 기반의 토픽 분석을 수행하여 주요 연구 주제들을 규명하였다. Kim et al.(2014)은 사용자의 인터넷 사용 기록을 추출하고, 이들 중 생활문화 카테고리에 해당되는 뉴스기사를 분석 대상으로 하였다. 뉴스기사에 대한 토픽 분석을 수행함으로써 실제 사용자들의 관심 분야를 파악하였다. Bae et al.(2014)는 트위터(Twitter) 데이터를 대상

으로 LDA 기반의 토픽분석을 적용하여 SNS 상에서의 주요 이슈를 추출하는 트위터 이슈 트래킹 시스템을 제안하였다. 그러나 항공산업 전반에 걸쳐 연구동향 및 미래유망분야 전망하기 위하여 텍스트 마이닝 기법을 사용하여 트렌드를 분석한 연구는 수행된바 없다. 본 연구에서는 토픽 분석을 이용하여 항공산업 전반적인 분야의 문서 집합 내 잠재된 토픽을 추출하고, 연도별 토픽의 추이를 분석하고자 한다.

3. 텍스트 마이닝 기반의 트렌드 분석

3.1 실험 데이터

항공산업의 핵심 키워드 도출 및 트렌드 분석을 위해 이용할 수 있는 대상으로 논문, 정책연구보고서, 뉴스 기사 등 다양한 정보 원천이 존재하지만, 본 연구에서는 논문 및 정책연구보고서를 분석 대상으로 활용한다. 뉴스 데이터의 경우 저작권 이슈가 존재하며, 웹상의 다양한 데이터 원천으로부터 데이터를 크롤링(Crawling)하여 통일된 형식으로 표준화하는 것 등의 어려움 등과 같은 한계점이 존재한다. 논문은 뉴스와 같은 인터넷 상의 정보에 비해 빅데이터 속성 중 규모(Volume)와 속도(Velocity) 측면에서 낮은 수준이나, 저자, 연도, 초록, 주제어 등 정형화된 항목으로 구성되어 있고 문서 형식이 통일되어 있어 분석이 비교적 용이한 측면이 있다. 또한, 대량으로 수집된 논문 데이터를 통해 분야별 최근 연구동향 파악이 가능하다. 본 연구에서는 항공 관련 전문 자료인 논문 데이터를 수집하여 논문의 기본적인 정보 및 연구 내용이 요약된 초록과 원문을 분석에 이용한다. 정책연구보고서의 경우에는 기관의 특성에 따라 원문을 비공개하거나 관련 정보에 대한 항목이 상이한 경우가 있어 데이터 수집에 한계가 있었다. 본 연구에서는 제목, 초록 등 정책연구별 요약 정보를 분석에 활용한다.

제목, 초록 등에 '항공'이라는 주제어를 포함한 2000년부터 현재까지 연구된 국내 학술논문 및 정책연구보고서 총 4,104건을 추출하여 분석에 이용하였다. 데이터 추출을 위한 검색 키워드는 해당 도메인의 일반적인 단어를 선정하거나 도메인 전문가가 선택하며, 데이터 분석 목적과 검색된 데이터의 관련성을 고려하여 선정한다. 본 연구는 항공산업 분야의 전반적인 트렌드를 분석하는 것을 분석 목적으로 하였기 때문에 '항공'이라는 검색 키워드를 선정하였으며, 추후 항공 내 세부적인 분야에 대한 트렌드 분석을 위해서는 더욱 더 정교한 방식으로 키워드를 선정하기 위한 사전 검토가 필요하다. 분석 대상 데이터 기간은 2001년 인천국제공항 개항을 시작으로 항공산업의 도약을 위한 기반을 정립하는 시기이고, 국내의 항공 관련 연구가 대부분 2000년도부터 진행되기 시작했다는 점을 고려하여 분석 대상 기간을 2000년부터 2014년 9월까지로 설정하였다. 분석 대상 데이터의 세부적인 내용은 <Table 1>에 제시하였다.

〈Table 1〉 Research Data

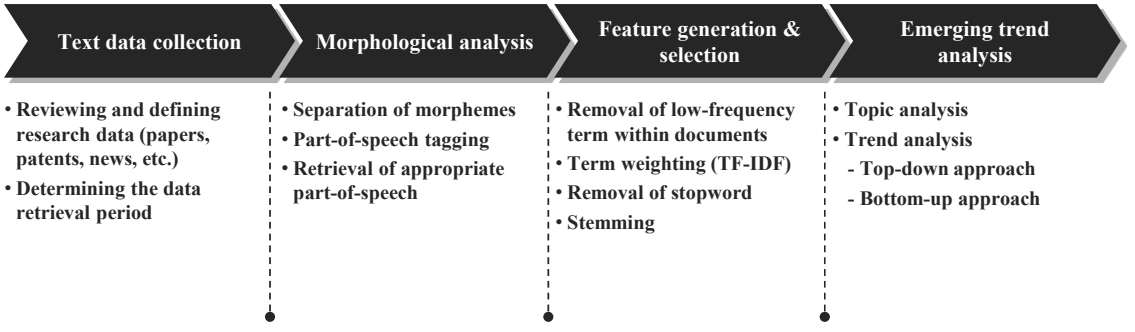
Source	Retrieval period	Frequency
NDSL academic papers	2000.1 ~ 2014.9	2,780
KISS academic papers		1,214
PRISM research reports		110
Total		4,104

NDSL(National Digital Science Library)은 과학기술분야의 논문, 특허, 보고서, 동향, 표준, 사실정보 등을 제공하며, 과학기술분야의 연구의 비중이 높은 지식데이터베이스이다. 따라서 사회과학분야의 연구 비중이 높은 KISS(Koreanstudies Information Service System)를 추가적으로 고려하였다. NDSL에서 항공 관련 연구는 10여년 간 총 2,780건이며 총 415개의 저널에 포함되어 있다. 항공 관련 연구가 다수 포함되어 있는 주요 저널로는 한국항공우주학회지, 한국항공운항학회지, 한국측량학회지, 항공우주기술, 항공우주정책·법학회지 순이다. KISS에서 항공 관련 연구는 총 1,214건이며 총 195개의 저널에 포함되어 있다. 항공 관련 연구가 다수 포함되어 있는 저널로는 한국항행학회논문지, 한국항공경영학회지, 한국항공운항학회지, 항공우주정책·법학회지, 관광연구, 관광경영연구, 호텔경영학연구, 관광연구저널, 한국군사과학기술학회지 순이다. 두 데이터베이스에서 중복적으로 다루고 있는 문서는 하나의 문서만 선택하여 분석하였다. PRISM(Policy Research Information Service and Management)은 중앙부처에서 수행하는 정책연구 과제를 효율적으로 관리하고, 정책연구보고서를 공유하는 시스템이다. PRISM으로부터 항공 관련 정책과제를 검색한 결과, 총 110건의 정책연구가 진행되었다.

3.2 실험 설계

항공산업의 텍스트 마이닝 기반의 핵심 키워드 도출 및 트렌드 분석에 관한 연구는 아직 초기 단계이기 때문에 본 연구를 통하여 항공산업 전반적인 미래유망분야 선정을 위한 핵심 키워드 도출 및 트렌드 분석 프레임워크 구축 및 실제 활용에 대한 가능성을 제시하고, 텍스트 마이닝 기반의 트렌드 분석 방법론에 대한 타당성을 검토해 보고자 한다. 항공산업의 핵심 키워드를 추출하기 위한 방법으로 텍스트 마이닝 기법 중 토픽 분석을 수행한다. 본 연구의 제안 방법론의 분석 단계별 세부 내용은 <Figure 1>에 제시하였다.

분석 대상은 항공 관련 전문 자료인 논문 및 정책연구 보고서이며, 데이터의 형태는 PDF, 워드 프로세스 등 텍스트로 이루어져 있다. 이러한 텍스트 형태의 데이터를 분석하기 위해 먼저, 형태소 분석 및 의미 정보 변환 및 추출 작업을 포함한 텍스트 전처리 단계를 실시한 후, 텍스트 문서 집합 내에 잠재된 주제를 도출하는 기법인



〈Figure 1〉 Stages of Text Mining Based Trend Analysis

토픽 분석과 추출된 토픽에 대하여 시계열적으로 살펴보는 트렌드 분석을 수행한다. 토픽분석은 문서 집합 내에서 동시출현빈도가 높은 단어들을 기준으로 유사한 주제를 가진 문서들을 그룹화한다. 하나의 토픽은 한 개 이상의 키워드의 집합으로 표현되며, 개별 문서가 하나의 주제에만 해당하는 것이 아니라 여러 주제를 다룰 수 있다는 점을 가정하고 있어 대량의 문서에 대해 전반적인 주제를 추출하는데 유용하게 사용된다.

항공 관련 논문으로부터 추출된 토픽의 의미는 키워드의 집합으로 파악할 수 있으며, 문서집합 내에서의 해당 토픽에 대한 출현 빈도는 토픽이 의미하고 있는 개념(Concept)에 대한 관심도를 반영한다고 볼 수 있다. 이러한 관심도를 산출함으로써 해당 개념에 대한 트렌드 분석이 가능해지며, 더 나아가 유망분야의 도출이 가능하다.

항공 트렌드 분석을 위한 분석 접근 방식으로 두 가지가 존재한다. 첫 번째는 상향식 접근 방식(Top-down Approach)으로 최종 토픽을 선정한 후 연도별 추이를 분석하는 것이다. 전체 데이터에서 토픽을 추출한 후, 연도별로 동일한 토픽에 대해 빈도 기반의 추이 분석을 실시하여 증가 또는 감소하는 토픽을 추출할 수 있다. 두 번째는 하향식 접근 방식(Bottom-up Approach)으로 시대별로 토픽을 추출하여 추이를 분석하는 것이다. 이 방법은 시대별로 차별화되는 토픽을 쉽게 파악할 수 있지만, 과거로부터 현재에 이르기까지의 토픽 발생 추이를 살펴보기 어렵다. 따라서 본 연구에서는 두 가지 접근 방식 중 상향식 접근 방식으로 항공산업에 대한 트렌드 분석을 실시하여 연도별로 동일한 토픽에 대한 트렌드를 파악한다.

시각화(Visualization)는 분석 결과를 쉽게 이해할 수 있도록 그래프 등과 같은 시각적인 수단으로 정보를 전달하기 위한 것이다. 본 연구에서는 추출된 토픽들의 연도별 변화를 나타내는 시계열 그래프를 통해 트렌드를 파악한다. 각 토픽에 대한 추이를 시계열 그래프로 표현하여 증가추세를 보이는 토픽을 항공산업 미래유망분야로 선정한다.

4. 실험 결과

2000년부터 2014년 9월까지 연구된 항공 관련 논문 3,994개를 분석에 이용하였다. 텍스트 마이닝 기반의 트렌드 분석에 관한 연구는 아직 초기 단계이기 때문에 비교적 정형화된 항목으로 구성되어 있고 분석이 용이한 논문을 분석 대상으로 선정하였다. 키워드별 가중치로써 TF-IDF를 사용하였고, 토픽 분석을 위해 총 261개의 키워드를 추출하였다. 추출된 키워드별 빈도 및 TF-IDF 값의 예시는 다음의 <Table 2>와 같다.

〈Table 2〉 Frequency and TF-IDF of Derived Keyword Examples

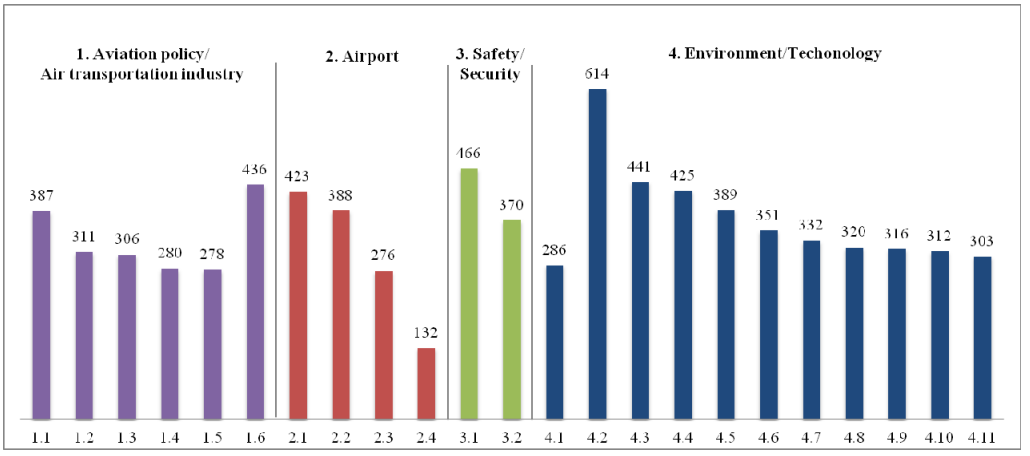
No.	Keyword	Frequency	TF-IDF
1	Hub	76	7.534
2	Threat	82	7.317
3	Lidar	134	7.260
4	Job satisfaction	126	7.260
5	Disaster	86	7.260
6	Aviation safety	87	7.179
7	Airline service	74	7.179
8	Prevention	75	7.153
9	Cell	179	6.983
10	Low cost airline	160	6.960

추출된 토픽의 의미는 키워드의 집합으로 파악할 수 있으며, 토픽 분석 결과와 항공 분야 전문가의 의견을 함께 고려하여 각각의 토픽을 <Table 3>과 같이 4개의 영역으로 구분하였다. 항공정책/항공운송산업에 관한 토픽은 6개, 공항에 관한 토픽은 4개, 안전/보안에 관한 토픽은 2개, 환경/기술에 관한 토픽은 11개의 토픽을 포함하고 있다. 각 영역별 핵심 토픽별 빈도수는 <Figure 2>와 같다.

추출된 토픽 중에서 빈도수가 높은 상위 10개의 토픽을 살펴보면, 항공기 날개/형상 설계 최적화, 항공교통관제(충돌방지), 레이다, 승무원 교육 및 관리, 센서(오류방지), 공항서비스평가, 공간해상도(지형지물) 수준 고도화, 항공 물류, 항공안전 정책, 항공사고 방지 순으로 나타났다. 이들은 항공산업에서 중요하게 연구되는 핵심

〈Table 3〉 Emerging Topics Derived for the Aviation Industry

Categories	Topics	Keywords
1. Aviation policy/ Air transport industry	1.1 Aviation safety policy	Aviation, Safety, Accident, Aviation safety, Operation
	1.2 Airfare (low-cost carriers)	Carrier, Cost, Low cost airline, Operation, Value
	1.3 Distribution channels	Travel, Agency, Airline, Distribution, Channel
	1.4 Job satisfaction	Job, Satisfaction, Job satisfaction, Employee, Commitment
	1.5 Aviation agreements (baggage liability, etc.)	Law, Liability, Convention, State, Damage
	1.6 Flight attendant training/management	Flight, Attendant, Flight attendant, Commitment, Training
2. Airport	2.1 Airport service appraisal	Service, Quality, Passenger, Service quality, Satisfaction
	2.2 Aviation logistics	Airport, Cargo, Passenger, Facility, Logistics
	2.3 Airport hub strategy	Airport, Facility, Passenger, Security, Hub
	2.4 Noise control measures	Noise, Level, Airport, Measurement, Vibration
3. Safety/ Security	3.1 Air traffic control (collision prevention)	Control, Traffic, Controller, Response, Demand
	3.2 Air accident prevention	Accident, Passenger, Damage, Risk, Liability
4. Environment/ Technology	4.1 Eco-friendly high-efficiency fuel	Fuel, Cell, Power, Energy, Density
	4.2 Aircraft wing/shape design optimization	Aircraft, Operation, Wing, Landing, Stability
	4.3 Radar	Radar, Vehicle, Antenna, Traffic, Performance
	4.4 Sensor (error prevention)	Sensor, Error, Accuracy, Camera, Measurement
	4.5 Spatial resolution enhancement	Image, Camera, Resolution, Feature, Photo
	4.6 Unmanned aircraft	Vehicle, Path, Unmanned aerial vehicle, Aerial, Flight
	4.7 Lidar	Building, Lidar data, Lidar, Surface, Height
	4.8 Engine	Engine, Flow, Fuel, Performance, Temperature
	4.9 Composite materials	Material, Property, Composite, Temperature, Strength
	4.10 Digital map	Map, Digital map, Accuracy, Photo, Road
	4.11 Aerial photo	Land, Photo, Aerial photograph, Management, Construction



〈Figure 2〉 Frequency of Emerging Topics by Core Issue Category

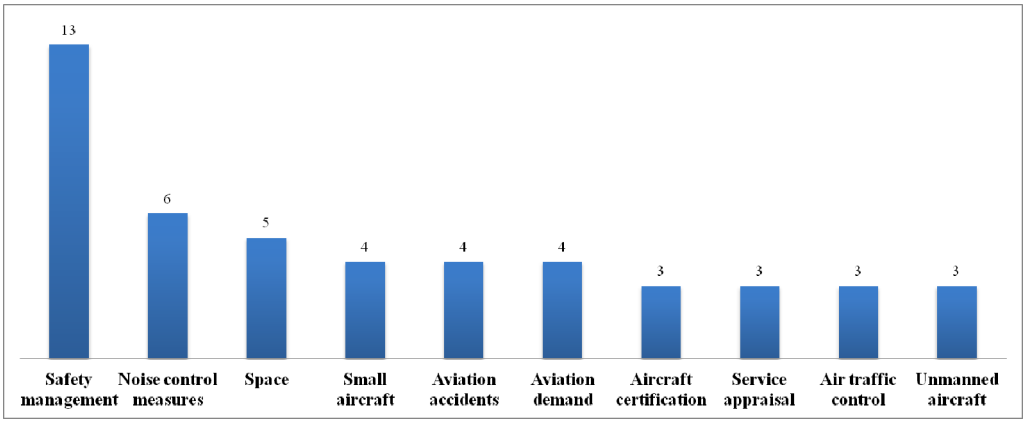
토픽에 해당된다고 볼 수 있다. <Table 4>의 핵심 이슈에 선정된 토픽 및 빈도수를 정리하였다.

추가적으로 향후 항공산업의 정책 및 예산 수립 등에 참고 가능하도록 항공 관련 정책연구보고서를 대상으로 텍스트 마이닝을 수행하여 키워드를 추출하고, 항공 관련 논문으로부터 도출한 분야와 실제 정부가 투자하고 있는 분야를 비교하였다. 정책연구보고서로부터 추출한 10개의 핵심 키워드는 <Figure 3>과 같다.

추출된 키워드 중에서 빈도수가 높은 상위 10개의 키워드를 선정한 결과, 안전관리, 소음대책, 우주, 소형항공기, 항공사고, 항공수요, 항공기인증, 항공교통관제, 무인항공기 순으로 도출되었다. 항공 관련 연구에서의 핵심 토픽과 실제 정부가 투자하고 있는 분야의 일치여부는 <Table 5>에 제시하였다.

〈Table 4〉 Core Emerging Topics of the Aviation Industry

Rank	Categories	Topics	Frequency
1	Environment / Technology	Aircraft wing/shape design optimization	614
2	Safety/Security	Air traffic control (collision prevention)	466
3	Environment/Technology	Radar	441
4	Aviation policy/Air transport industry	Flight attendant training/management	436
5	Environment / Technology	Sensor (error prevention)	425
6	Airport	Airport service appraisal	423
7	Environment / Technology	Spatial resolution enhancement (geographic features)	389
8	Airport	Aviation logistics	388
9	Aviation policy / Air transport industry	Aviation safety policy	387
10	Safety/Security	Air accident prevention	370



〈Figure 3〉 Core Keywords Retrieved from Aviation-Related Policy Research Reports

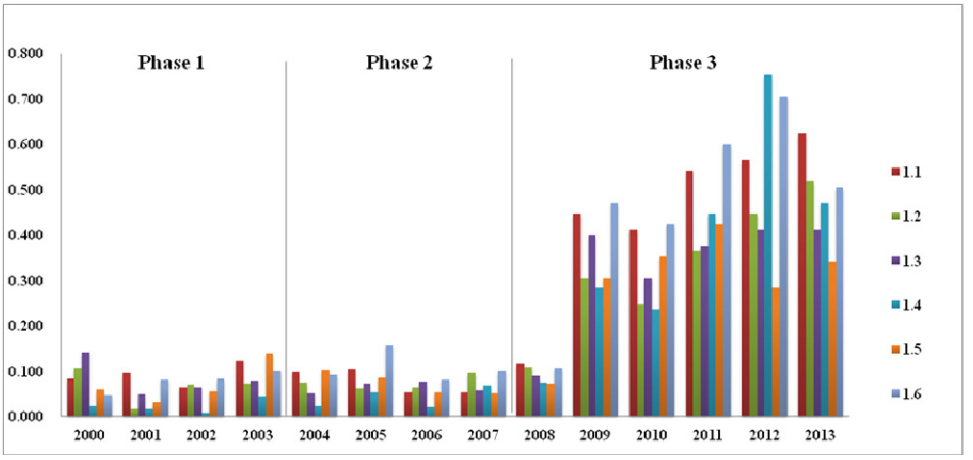
최종 선정된 토픽에 대한 시대별 관심도의 변화를 파악하고 항공산업 미래유망분야를 선정하기 위해 연도별 트렌드 분석을 실시하였다. 2014년도 데이터는 9월까지 확보되었기 때문에 트렌드 분석에서는 제외한다. 핵심 토픽들을 대상으로 상향식 접근 방식을 채택하였다. 이는 전체 데이터에서 토픽을 추출한 후, 연도별로 동일한 토픽에 대해 관심도의 추이를 분석을 실시하여 증가 또는 감소하는 토픽을 추출하는 방식이다. 문서집합 내에서의 해당 토픽에 대한 출현 빈도는 토픽이 의미하고 있는 개념에 대한 관심도를 반영하는 것이다. 개별 토픽에 대한 관심도를 연도별로 다음과 같이 산출하였다.

관심도 = 해당 토픽의 빈도수 / 총 문서수

트렌드 분석은 항공산업의 시대적 흐름을 반영하여 2000년부터 2013년까지의 기간을 3기로 구분하여 각 토픽에 대한 시계열 패턴을 분석하였다. 1기(2000 ~ 2003)는 '안전 성장 및 도약준

〈Table 5〉 Correspondence between Core Emerging Topics and Issues in Aviation Policy Research

Rank	Categories	Keywords	Frequency	Correspondence
1	Safety / Security	Safety management	13	○
2	Airport	Noise control measures	6	○
3	Aviation policy / Air transport industry	Space	5	
4	Environment / Technology	Small aircraft	4	
5	Safety / Security	Aviation accidents	4	○
6	Aviation policy / Air transport industry	Aviation demand	4	
7	Safety / Security	Aircraft certification	3	
8	Airport	Service appraisal	3	○
9	Safety / Security	Air traffic control	3	○
10	Environment / Technology	Unmanned aircraft	3	○



〈Figure 4〉 Trend Analysis for the Issue of Aviation Policy/Air Transport Industry

비기', 2기(2004 ~ 2007)는 '도약 및 격변기', 3기(2008 ~ 현재)는 '안정기 및 제2도약기'로 명명하였다.

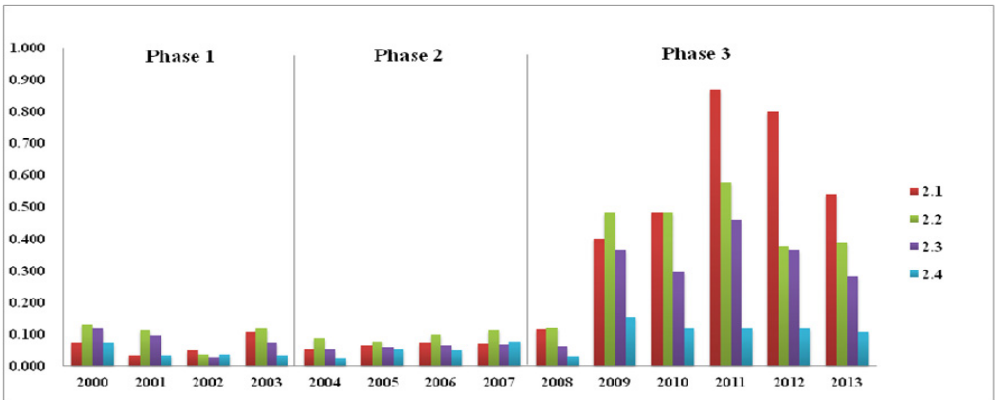
1기에는 항공협정 체결이 가속화되고, 항공안전 정책의 국제기준이 도입되었다. 또한, 준사고 보고제도, 항공안전 감독관제도, 항공운송사업 운항증명, 그리고 정비 조직 승인제도를 도입하여 항공 안전성을 확보하고자 하였다. 2001년 인천국제공항 개항을 시작으로 여러 기준들이 수립됨에 따라 항공산업의 도약을 위한 기반이 마련된 시기라 할 수 있겠다.

2기에는 인천공항 허브화 전략 추진 및 인천공항 2단계 건설 사업이 수행된 시기이다. 안정성 확보를 위해서는 항공안전 관리 시스템이 구축되었다. 특히, 2기에는 대체교통수단 및 경기 침체로 인해 항공수요가 감소함에 따라 국내 실정에 적합한 항공서비스 제공을 위한 저가항공사(Low Cost Carriers)가 출현하기 시작하였다. 국내에서 처음으로 취항한 저가항공사는 2005년 8월에 취항을 시작한 한성항공(현재 티웨이 항공)이다.

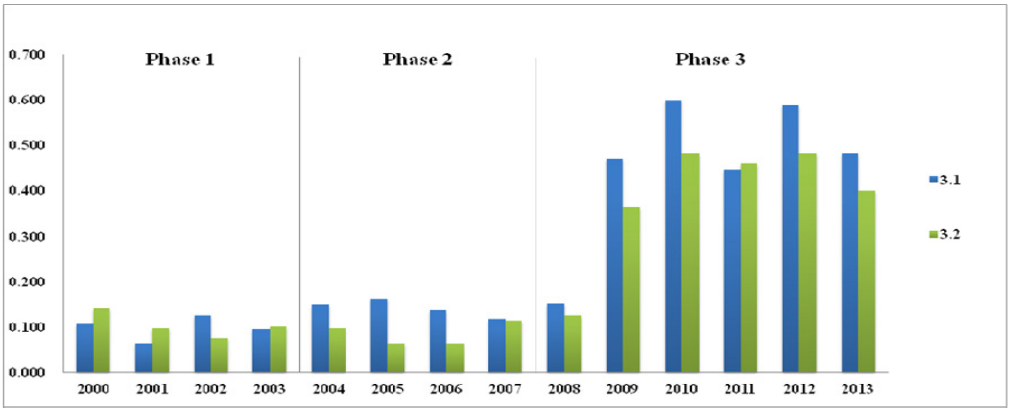
3기에는 인천공항 경쟁력 강화 및 확보를 위한 인천공항 3단계 건설을 착수하였으며, 인천공항 3단계 건설 사업이 2017년까지 수행될 예정이다. 현재, 국내에서는 대형항공사(Full Service Carriers)인 대한항공과 아시아나항공 외에 저가항공사는 현재 5개가 운영되고 있다. 항공안전 자율 보고 규정이 마련되었으며, 항공운송사업자의 시장진입규제 완화가 확대되었다. 또한, 소형항공사업 등이 수행되면서 항공운송사업의 다양화 기반이 마련된 시기라 볼 수 있겠다.

본 연구에서는 트렌드 분석 결과를 통해 증가추세를 보이는 토픽을 항공산업의 미래 핵심 유망분야로 선정하였다. <Figure 4> ~ <Figure 7>과 같이 각 영역별 트렌드 분석 결과를 통해 항공관련 연구가 3기인 '안정기 및 제2도약기'에 활발히 연구되고 있는 것으로 나타났다.

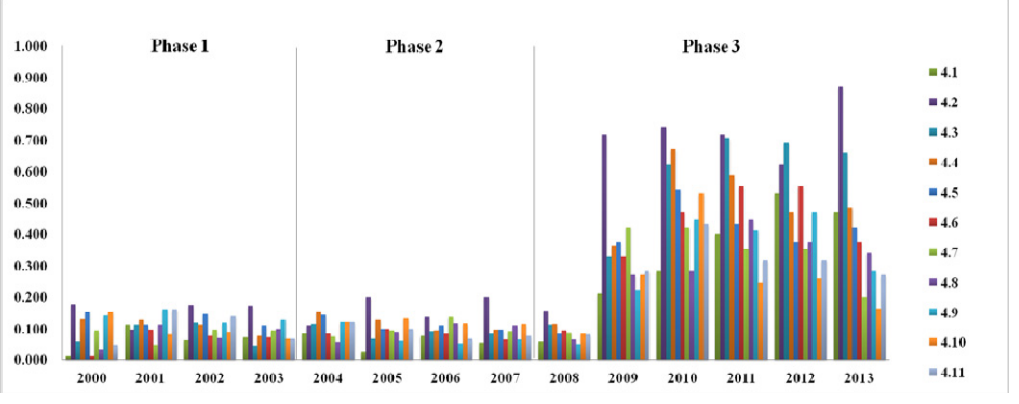
시계열적 패턴 분석을 통해 토픽에 대한 관심도가 증가하는 추세, 감소하는 추세, 그리고 특정한 패턴이 존재하지 않는 유형으로 구분하였다. 특정한 패턴이 존재하지 않는다는 것은 관심도의 증가 또는 감소 추세가 없이 일관적으로 중



〈Figure 5〉 Trend Analysis for the Issue of Airport



〈Figure 6〉 Trend Analysis for the Issue of Safety/Security

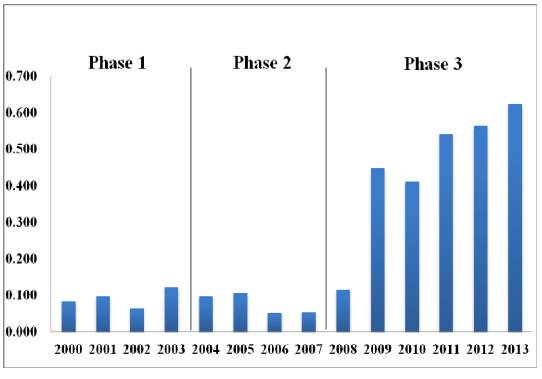


〈Figure 7〉 Trend Analysis for the Issue of Environment/Technology

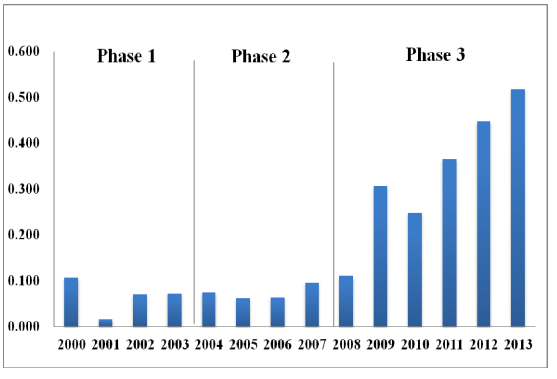
요하게 고려된다는 것을 의미한다. 항공 관련 연구가 최근 들어 활발히 연구되고 있는 점을 고려하여 전체 분석기간 중 관심도가 3회 이상 연속 상승한 횟수와 최근 3기(2008 ~ 현재)에서의 상승 횟수를 고려하여 관심도가 높은 토픽을 항공산업 미래유망분야로 선정하였다.

항공산업에서 증가 추세를 보이는 토픽은 항공안전정책, 항공운임(저가항공), 그리고 친환경 고연비 연료가 도출되었다. 첫째, 항공안전정책은 항공정책/항공운송산업 영역에 해당하는 토픽으로 <Figure 8>과 같이 항공 시대 흐름 중 안정기 및 제2도약기에 해당하는 시점부터 계속적으로 증가하는 추세를 보인다. 안정 성장 및 도약준비기 동안에 항공안전정책의 국제 기준이 도입되었다. 우리나라는 2008년 5월 수검을 실시하였으며, 국제민간항공기구(International Civil Aviation Organization) 평가단으로부터 국제기준 이행률 98.89%를 판정받은 바 있다. 항공 사고는 기차, 버스, 선박 등 다른 운송 수단으로부터 발생한 사고에 비해 사고 횟수는 적으나 한 번 발생했을 경우 인명 피해가 클 뿐만 아니라 기체 손상 및 이용자의 해당 항공사에 대한 불신과 같은 손해가 발생하기 때문에 항공안전정책은 지속적으로 관심도가 높은 이슈일 것으로 보인다. 따라서 지속적인 항공안전에 대한 제도를 개선하여 항공사고, 지연, 결항 등 항공 안전 및 보안 수준을 제고해야 한다.

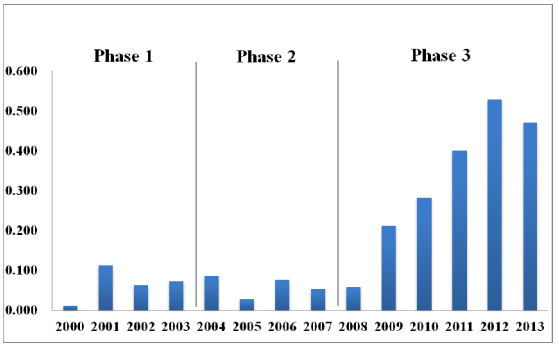
둘째, 항공운임(저가항공)은 항공정책/항공운송산업 영역에 속하는 토픽으로 저가항공에 관한 이슈는 <Figure 9>와 같이 2009년부터 계속적으로 관심도가 높은 추세를 보인다. 대체교통수단 및 경기침체로 항공에 대한 수요가 감소하면서 국내 실정에 적합한 항공서비스 제공을 위한 제도를 개선하기 위해 저가항공사 시장에 참여하기 시작하였다. 2005년 8월 한성항공(현재 티웨이 항공)의 취항을 시작으로 2006년 6월 제주항공, 2008년 7월 진에어, 2008년 10월 에어부산, 2009년 1월 이스타항공이 취항되었다. 국토교통부가 발표한 2014년 국내선 항공사 점유율에 따르면 대한항공 29.4%, 아시아나항공 23.1%, 제주항공 13.2%, 에어부산 11.9%, 이스타항공 7.9%, 티웨이항공 7.3%, 진에어 7.2% 순이다. 국내 저가항공사에서 운항편수를 증가시키고 신규 노선을 계속적으로 개설함에 따라 이용객이 증가했으며, 이에 따라 대형항공사인 대한항공, 아시아나 항공의 점유율이 감소했다. 향후에도 대형항



〈Figure 8〉 Trend Analysis for Aviation Safety Policy



〈Figure 9〉 Trend Analysis for Airfare (Low-Cost Carriers)



〈Figure 10〉 Trend Analysis for Eco-Friendly High-Efficiency Fuel

공사와 저가항공사간 점유율 경쟁이 치열해지고, 저렴한 항공운임에 대한 이슈가 지속될 것으로 보인다.

셋째, 친환경 고연비 연료는 환경/기술 영역에 해당하는 토픽으로 <Figure 10>에서 보는 바와 같이 2013년 다소 관심도가 하락하였으나 2009년부터 계속적으로 증가 추세를 보인다. 항공기에서 발생하는 온실가스의 감축을 위하여 수소연료전지, 2차 전지, 바이오 연료 등과 같은 친환경적이면서 연비가 높은 연료에 대한 관심도가 높다고 볼 수 있다. 친환경적 특성을 갖는 연료를 개발하고 사용하는 것 외에도 복합재료 사용, 연료 효율성이 높은 엔진 사용, 연료절감을 위한 최적의 형상 설계를 하는 등 항공산업의 그린화를 위한 기술에 대한 이슈가 앞으로 지속될 것으로 보인다.

항공산업에서 감소 추세를 보이는 토픽은 공항허브화, 소음대책, 라이다, 그리고 항공사진이 도출되었다. 유통채널, 직무만족, 항공협정(수하물배상 반입제한물품 규정), 승무원 교육 및 관리, 공항서비스평가, 항공물류, 항공교통관제(충돌방지), 항공사고방지, 항공기날개/형상 설계 최적화, 레이다, 센서(오류방지), 공간해상도(지형지물) 수준 고도화, 무인항공기, 엔진, 복합재료, 디지털 지도는 항공산업에서 꾸준히 중요하게 연구되는 토픽이다.

5. 결론

본 연구에서는 항공 관련 전문 학술자료를 대상으로 빅데이터 분석 방법인 텍스트 마이닝 기법을 적용하여 항공산업 전반적인 분야의 핵심 이슈를 추출하고 연구동향 파악 및 미래유망분야를 전망해보고자 하였다. 본 연구결과는 항공산업 분야의 핵심 키워드 및 토픽을 추출하기 위하여 분석 대상으로 항공관련 주제어를 포함한 논문에 한정하여 분석을 시도한 연구라는 한계점이 존재하나, 항공산업의 핵심 이슈를 도출하여 지속적으로 모니터링하고 미래유망분야에 대한 방향을 제시하기 위한 정량적인 분석 방법론의 전형을 마련하였다는 점에서 의의가 있다고 할 수 있다.

본 연구에서 시도한 항공산업의 미래유망분야를 선정하기 위해 빅데이터 분석 기반의 접근 방식을 적용하여 수행하는 텍스트 마이닝 기반의 트렌드 분석은 현재 초기 단계이나, 최근 급증하고 있는 항공 관련 논문, 기사, 특허 등을 종합적으로 반영해 빠른 기술 및 환경변화에 대응하기 위한 방안을 마련하기 위하여 필수불가결할 것으로 보인다. 향후 연구에서는 분석 대상을 선정하기 위한 주제어의 세밀한 선정과 함께 항공 관련 기사, 특허, 입찰정보 등으로 확대 적용함으로써 분석 결과의 신뢰성을 고취시키고자 한다. 또한, 분석 결과의 정확도를 향상시키기 위해 향후 데이터의 저장 형태 표준화 및 대상 정의, 텍스트 마이닝을 위한 문서 수집에 대한 자동화 방

안 마련 등은 향후 심층 연구를 위해 지속적으로 수행해야할 선결과제일 것이다. 이와 같은 연구를 토대로 항공산업의 미래 핵심 이슈 및 유망 분야를 도출하기 위한 객관적인 분석 접근방식은 항공산업의 경쟁력 확보를 위한 중장기 정책 수립의 당위성 및 근간을 마련할 것이라고 기대한다.

참고문헌(References)

Bae, J. -h., N. -g. Han, and M. Song, "Twitter Issue Tracking System by Topic Modeling Techniques," *Journal of Intelligence and Information Systems*, Vol. 20, No. 2(2014), 109~122.

Blei, D. M, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocations," *Journal of Machine Learning Research*, Vol.3(2003), 993~1022.

Chung, J. H. and S. M. Lee, "GSA-based future ICT technology prediction process," *ie Magazine*, Vol. 19, No. 3(2012), 34~40.

Dalkey, N. C. and O. Helmer, "An Experimental Application of the Delphi Method to the Use of Experts," *Management Science*, Vol.9, No.3(1963), 458~467.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science (JASIS)*, Vol. 41, No. 6(1990), 391~407.

Feldman, R. and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *KDD*, Vol. 95(1995), 112~117.`

Hofmann, T., "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), 50~57.

Jeong, C. W. and J. J. Kim, "Analysis of trend in construction using textmining method," *Journal of The Korean Digital Architecture·Interior Association*, Vol. 12, No. 2(2012), 53~60.

Jeong, D., J. Kim, G. -N. Kim, J. -U. Heo, B. -W. On, and M. Kang, "A Proposal of a Keyword Extraction System for Detecting Social Issues," *Journal of Intelligence and Information Systems*, Vol. 19, No. 3(2013), 109~122.

Kim, J., N. Kim, and Y. Cho, "User-Perspective Issue Clustering Using Multi-Layerd Two-Mode Network Analysis," *Journal of Intelligence and Information Systems*, Vol. 20, No. 2(2014), 93~107.

Korea Agency for Infrastructure Technology Advancement, "Technology Forecasting 2040; Land, Infrastructure, and Transport," 2013.

Korea Institute of science and technology Evaluation and Planning, "KISTEP 10 future technologies for next 10 years," *Research Report(2014-059)*, 2014.

Min, K. Y., H. T. Kim, and Y. G. Ji, "A Pilot Study on Applying Text Mining Tools to Analyzing Steel Industry Trends: A Case Study of the Steel Industry for the Company "P"," *Journal of the Society for e-Business Studies*, Vol.19, No.3(2014), 51~64.

Park, J. H. and M. Song, "A Study on the Research Trends in Library and Information Science in Korea using Topic Modeling," *Journal of the Korean Society for Information Management*, Vol. 30, No. 1(2013), 7~32.

Salton, G. and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983.

Abstract

Text Mining-Based Emerging Trend Analysis for the Aviation Industry

Hyun-jung Kim* · Nam-ok Jo** · Kyung-shik Shin***

Recently, there has been a surge of interest in finding core issues and analyzing emerging trends for the future. This represents efforts to devise national strategies and policies based on the selection of promising areas that can create economic and social added value. The existing studies, including those dedicated to the discovery of future promising fields, have mostly been dependent on qualitative research methods such as literature review and expert judgement. Deriving results from large amounts of information under this approach is both costly and time consuming. Efforts have been made to make up for the weaknesses of the conventional qualitative analysis approach designed to select key promising areas through discovery of future core issues and emerging trend analysis in various areas of academic research. There needs to be a paradigm shift in toward implementing qualitative research methods along with quantitative research methods like text mining in a mutually complementary manner.

The change is to ensure objective and practical emerging trend analysis results based on large amounts of data. However, even such studies have had shortcoming related to their dependence on simple keywords for analysis, which makes it difficult to derive meaning from data. Besides, no study has been carried out so far to develop core issues and analyze emerging trends in special domains like the aviation industry. The change used to implement recent studies is being witnessed in various areas such as the steel industry, the information and communications technology industry, the construction industry in architectural engineering and so on.

This study focused on retrieving aviation-related core issues and emerging trends from overall research papers pertaining to aviation through text mining, which is one of the big data analysis techniques. In this manner, the promising future areas for the air transport industry are selected based on objective data from aviation-related research papers. In order to compensate for the difficulties in grasping the

* School of Business, Ewha Womans University
** School of Business, Ewha Womans University
*** Corresponding author: Kyung-shik Shin
52 Ewhayeodae-gil, Seodaemun-Gu, Seoul, 120-750, Korea
Tel: +82-2-3277-2799, Fax: +82-2-3277-2776, E-mail: ksshin@ewha.ac.kr

meaning of single words in emerging trend analysis at keyword levels, this study will adopt topic analysis, which is a technique used to find out general themes latent in text document sets. The analysis will lead to the extraction of topics, which represent keyword sets, thereby discovering core issues and conducting emerging trend analysis. Based on the issues, it identified aviation-related research trends and selected the promising areas for the future.

Research on core issue retrieval and emerging trend analysis for the aviation industry based on big data analysis is still in its incipient stages. So, the analysis targets for this study are restricted to data from aviation-related research papers. However, it has significance in that it prepared a quantitative analysis model for continuously monitoring the derived core issues and presenting directions regarding the areas with good prospects for the future. In the future, the scope is slated to expand to cover relevant domestic or international news articles and bidding information as well, thus increasing the reliability of analysis results.

On the basis of the topic analysis results, core issues for the aviation industry will be determined. Then, emerging trend analysis for the issues will be implemented by year in order to identify the changes they undergo in time series. Through these procedures, this study aims to prepare a system for developing key promising areas for the future aviation industry as well as for ensuring rapid response. Additionally, the promising areas selected based on the aforementioned results and the analysis of pertinent policy research reports will be compared with the areas in which the actual government investments are made. The results from this comparative analysis are expected to make useful reference materials for future policy development and budget establishment.

Key Words : Aviation, Big Data Analysis, Text Mining, Topic Analysis, Emerging Trend Analysis

Received : November 12, 2014 Revised : December 16, 2014 Accepted : December 18, 2014
Type of Submission : Fast Track Corresponding Author : Kyung-shik Shin

저 자 소 개

김 현 정

이화여자대학교 통계학사, 경영학 석/박사 학위를 취득하고, 한국산업은행 리스크관리본부와 딜로이트컨설팅에 재직했으며, 현재 이화여자대학교 경영대학 연구교수로 재직 중이다. 주요 연구분야는 지능형 의사결정지원시스템, 데이터 마이닝과 인공지능 응용, 빅데이터 분석 및 비즈니스 인텔리젼스 등이다.

조 남 옥

현재 이화여자대학교 대학원 경영학과에서 경영정보시스템 전공으로 박사과정에 재학중이다. 주요 연구분야는 지능형 의사결정지원시스템, 데이터 마이닝, 빅데이터 분석 및 응용 등이다.

신 경 식

현재 이화여자대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고 미국 George Washington University에서 MBA, 한국과학기술원(KAIST)에서 인공지능, 지식기반 시스템 등 지능형 기법을 경영분야에 적용하는 연구로 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 데이터 마이닝과 비즈니스 인텔리전스, 빅데이터 분석/비즈니스 애널리틱스, 인공지능 응용과 지식공학 등이다.