텍스트마이닝을 활용한 건설분야 트랜드 분석

Analysis of trend in construction using textmining method

정 철 우 ▌Jeong, Cheol-woo

정회원, KIST, 선임연구원

김 재 준 ▌Kim, Jae-Jun

정회원, 한양대학교, 교수(교신저자)

Abstracts

In this paper, we present new methods for identifying keywords for foresight topics that utilize the internet and textmining techniques to draw objective and quantified information that support experts' qualitative opinions and evaluations in foresight. Furthermore, by applying this fabricated procedure, we have derived keywords to analyze priorities in architectural engineering.

Not much difference between qualitative methods of experts and quantitative methods such as text mining has been observed from comparison between technologies derived via qualitative method from "The Science Technology Vision" (control group).

Therefore, as a quantitative tool useful for drawing keywords for foresight, textmining can supplement quantitative analysis by experts. In addition, depending on the level and type of raw data, text mining can bring better results in deriving foresight keywords.

For this reason, research activities accommodating Internet search results and the development of textmining methods for analyzing current trends are in demand.

Keywords

Textminign, Foresight, Trend, TF-DI

키워드

텍스트마이닝, 미래예측, 트랜드분석

1. 서 론

1.1 연구의 필요성

최근 국가경쟁력을 강화하기 위하여 국가 과학기술의 미래를 전망하거나 기술을 예측하는 것이 중요한 이슈로 떠오르고 있으며, 이러한 미래전망을 통해 국가의 다양한 전략과 정책을 수립하고 있다. 특히, 미래를 예측하여 중요 과학기술들을 선별 및 발굴하는 작업은 중요한 이슈가 되고 있으며, 이와 관련된 정부부처의 역할과 연구들이 중요한 테마가 되고 있다. 또한, 정부 및 연구주체들은 한정된 자원과 인력을 어떻게 투자 및 관리할 것인가에 대해서 미래예측과 관련하여 많은 관심을 갖고 있다.

이러한 미래를 예측하는 방법들은 대부분 전문가의 정성적인 의견과 주관적인 평가에 의해 이루어지고 있으며, 객관적인 방법론들에 대한 연구와 시도는 아직까지 미흡한 실정이다.

그러나 전문가에 의한 정성적인 미래예측방법은 매우 중요한 요소로 인정되어 왔으나, 전문가들의 경향, 정치적 요소, 인맥 관계 등의 다양한 요소에 의해 일부 편향적인 의견이나 주장이 반영될 가능성이 있고, 또한, 전문가들의 평가도 객관적인 자료가 부족하여 미래를 예측할 경우 많은오류와 어려움을 겪고 있다.

그러므로 과학기술분야의 미래예측 시 전문가의 객관적인 의견과 평가가 이루어지도록 좀 더 구체적이고 객관적인 데이터(data)와 자료를 제공하는 방법이 중요한 이슈(issues)로 떠오르고 있다. 특히, 일본과학기술정책연구소(NISTEP)에서는 논문을 이용한 논문맵(2009)을 활용하여 미래기술에 대한 예측을 국가연구개발에 접목하고 있으며, 국내외 많은 업체들이 특허 맵(patent map) 등을 활용해 특허분석을 하고 있다. 이러한 연구들은 객관적인 자료와 방법 등을 활용한 정량적 분석으로 미래예측을 시도한 것들이다.

본 연구에서는 미래예측 시 전문가들의 정성적인 의견과 평가를 보조할 수 있는 정량적이고 객관적인 자료를 도출하기 위하여, 기존의 논문이나 특허의 자료를 이용하는 방법보다 인터넷(internet)의 데이터를 활용하는 방안을 모색하였다. 또한, 기존에 단순히 논문 및 특허를 검색하여 노이즈(noise)를 제거하는 방법들을 이용하여 논문 및 특허 맵을 활용하는 방법보다는 인터넷을 활용한 중요빈도, 시간과 공간의 정보를 활용한 트랜드(trend) 분석을 포함한 텍스트 마이닝(text mining) 기법에 의해 중요한 키워드를 도출하였다. 이러한 연구방법을 통해 '제3회 과학기술예측조사 수정·보완'(2008)(이하 '제3회 과학기술예측조사'라 함)의 건설분야 기술에 대해 정량적인 분석을 시도하고 미래과 관련된 주요 키워드를 도출하였다.

1.2 선행 연구

일반적으로 수많은 정보들을 크게 2가지의 형태로 구분하면 정형화된 데이터와 비정형화된 데이터로 구분할 수 있다. 정형화된 데이터의 내용은 기존의 데이터를 활용하기 위하여 일정한 형식과 조건을 만족하는 자료로 가공하여 DB(data base)화한 정보이다. 이러한 정보는 전체의 약 20%정도가 자료의 생성, 저장, 재사용하는 정보로 구성되어 있다. 정형화데이터의 정보를 추출하고 가공하는 방법을 데이터 마이닝(data mining)이라고 불리며, 현재 우리가 가장 많이 활용하는 데이터베이스 시스템과 정보분류 체계에 응용되고 있다.

텍스트 마이닝)은 80%를 차지하는 비정형 정보를 어떻게 활용하는 가에 대한 방법을 말하며, 현재 다양한 분야에서 활용되고는 있으나 아직까지는 많은 연구가 필요하다. 텍스트 마이닝은 대용량의 데이터에서 사용자가 관심을 가지는 정보를 키워드의 수준이 아니라 문맥(context) 수준의 의미를 찾아내는 프로세스를 의미한다. 즉, 정보의 폭발적인 증가로 많은 부분을 자동적으로 처리할 수 있는 방법이 필요하게 되었으며, 대용량의 데이터 속에서 숨겨진 패턴을 발견하고 특정 주제와 연관된 데이터를 검색하는 방법으로 발전하고 있다.

텍스트 마이닝으로 인해 과거에 생각할 수 없었던 기술실현 방법들을 예상할 수 있다. 예를 들면, 다양한 자료와 시간 등으로 구성된 범죄 기록들 속에 현재 발생한 유사한 형태의 범죄유형을 찾아냄으로써 범죄자나 테러범을 색출, 웹 게시판에 올라오는 다양하고 비정형화 되어 있는 수천만건의 고객 불만 사항을 특정 카테고리별로 분류하거나 특정문제를 찾아내는 방법, 수많은 환자의 처방 내역서에서 당뇨병에 효과적인 치료 패턴을 자동으로 찾아내는 등 다양한 방향으로 응용이 가능하다.

현재까지 텍스트 마이닝은 크게 인터넷분야와 일반적인 데이터를 마이닝하는 분야로 응용되고 있다. 인터넷을 활용한 데이터 마이닝 기법은 인터넷 검색엔진 등에 활용되고 있다.

텍스트 마이닝의 일반적인 프로세스(process)는 여러 가지로 알려져 있으나 일반적으로 4단계의 프로세스를 거친다. 텍스트 마이닝 과정은 [비정형 정보수집 -> 정보처리 -> 정보추출 -> 정보분석] 등의 일반적인 절차를 따르고 있으며, 정보추출과정에서 수학적 모델이나 알고리즘을 통해 유용한 정보를 도출하는 방법이다. 이를 어떻게 활용할 것인가에 대해 검색엔진에 활용하거나 다른 중요한 키워들을 도출하는 등에 활용하고 있다. 텍스트 마이닝을 위한 정보추출 방법에는 다양한 목적, 조건, 환경 등으로 정보의 추출방법이 다양하며, 정보추출방법은 텍스트 마이닝에서 가장 중요한 부분 중에 하나이다.

특히, 정보추출 방법에는 수많은 수학적 알고리즘과 방법들이 존재하며, 그중 간단하면서 가장 강력한 방법으로는 TF-IDF(Term Frequency - Inverse Document Frequency) 방식을 많이 사용하고 있다. Spark(1972)는 TF-IDF가 여러 문서에 동시에 출현하는 단어는 범용적인 확률이 높다는 전제아래 역문헌빈도수(IDF : Invert Document Frequency)를 제시하였다. Salton(1976)에 의해 한 문서내에 자주 출현하는 단어는 그 문서를 대표할 수 있다는 명제를 통해 문헌내의 단어빈도수(TF : Term Frequency)를 계산하는 방식이 제시되었다. Wu & Salton(1981)은 이러한 두가지 방식의 가중치 중요도(term weight)를 발표하였다. 즉 TF-IDF에 대해 좀더 자세히 살펴보면, TF-IDF정보 검색과 텍스트 마이닝에서 이용하는 키워드의 가중치를 구하는 방법으로 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.

이와 같이 TF-IDF[1]는 빈도를 이용하는 방법으로 오랜 기간 동안 검증되어 왔으나, 복잡한 계산방식과 데이터 추출하는 방법과 범위에 따라 오차율이 크게 발생하는 등 많은 보완이 필요하다. 이에 따라, 현재의 트랜드나 현황을 분석하는데 어려움이 따르며, 이러한 현재의 트랜드를 분석하기 위하여 기존의 중요도를 보완하여 새로운 알고리즘이 필요하게 되었다.

1) 추가정보는 "Salton G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill"를 참조 바람

$$TF-IDF = TF \times \frac{1}{DF}$$

TF: 문서내 특정단어의 빈도수

DF: 여러문서내의 특정단어 빈도수

IDF: DF의 역수

2. 연구범위 및 방법

2.1 연구개요 및 프로세스

본 연구에서는 인터넷의 정보와 텍스트마이닝 방법을 활용하여 건설분야의 미래 주요 키워드를 분석하였다.

연구범위는 '제3회 과학기술예측조사'의 건설분야를 대상으로 분석하였다.

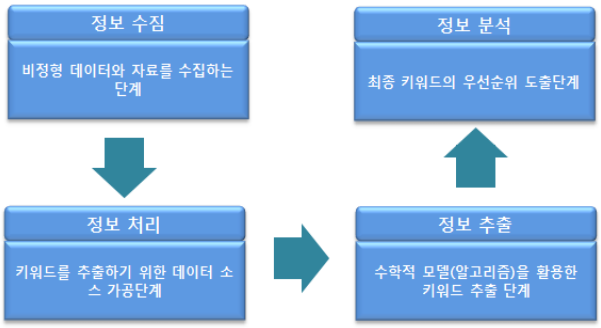

그림 1. 텍스트마이닝 프로세스

2.2 연구방법

(1) 인터넷을 활용한 연구방법

인터넷이 급속히 증가하면서 인터넷에는 수많은 데이터들이 넘쳐나고 있으며, 1999년에는 약 3천만개의 웹사이트가 있는 것으로 조사되었으며, 2007년도에는 약 1억개, 현재에도 기하급수적으로 증가하고 있다. 이렇게 기하급수적으로 증가한 인터넷의 실시간 데이터나 중요한 정보를 찾기 위하여 우리는 많은 시간과 노력을 필요로 하게 되었다. 이로 인해 인터넷 데이터를 활용하여 우리가 필요로 하는 정보를 어떻게 도출할 것인가에 대한 많은 연구가 진행되고 있으며, 텍스트 마이닝, 검색엔진, 로봇 등의 다양한 기법들이 도입되고 있으며 현재 많은 연구가 진행되고 있다.

네트워크분야에서도 IT(information technology)기술과 인터넷을 활용한 새로운 방법들이 도입되기 시작하였으며, 일부 연구에서는 인터넷의 방대한 데이터를 활용하여 데이터 마이닝, 텍스트 마이닝, 검색엔진을 이용한 방법 등이 연구되고 있다. 예를 들면, 2008년 구글에서는 미국의 109대 상원선거에서 기존의 출구조사를 한 결과와 구글의 검색결과로 분석한 결과가 유사한 것으로 알려지므로 인터넷의 유용성에 대한 연구가 활발하게 진행되고 있다.

본 연구에서도 인터넷의 다양한 정보(웹페이지, 논문검색, 특허검색, 블로그 등)를 검색결과값을 빈도수로 산정하였다. 특히, 여기에 사용되어진 구글 검색엔진은 페이지랭크(page rank)라는 알고리즘(algorithm)을 통해 구현되었으며, 웹페이지에 연결된 하이퍼링크를 수에 따른 가중치를 분석하여 검색 결과값으로 표현한 일종의 네트워크 알고리즘으로 만들어진 검색엔진이다.[2] 이렇게 구현된 검색엔진은 기존의 검색엔진과 달리 임의로 검색결과를 조작하기 어려워 검색결과에 대한 신뢰성이 높은 것이 특징이다. 최근 구글에서는 Open API(application programming interface)를 제공하여 논문검색사이트, 특허검색사이트, 구글 사전 등을 제공하여 다양한 목적을 위해 사용하게끔 공개되어 있다. 일반적으로 API는 운영체계나 언어가 어떤 기능을 제어할 수 있도록 제공되는 인터페이스였으나, 웹 2.0에서는 웹의 특정한 서비스를 이용하도록 제공되는 인터페이스로 개념이 확장되고 있는 개념이다.

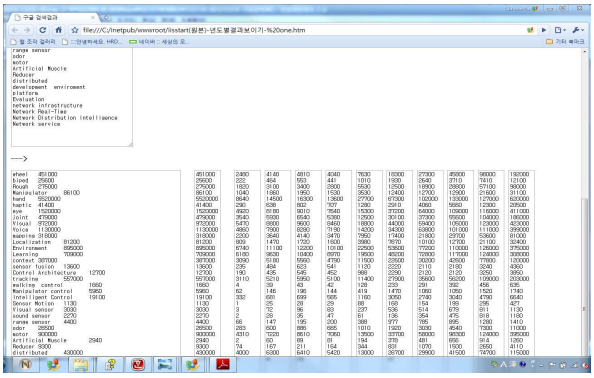

그림 2. 구글검색 결과값

(2) 텍스트마이닝의 수학적 알고리즘

본 연구의 텍스트마이닝에서 정보추출단계에서 사용되어지는 수학적 알고리즘은 TF-IDF[3]는 빈도를 이용하는 방법으로 오랜 기간 동안 검증되어 왔으나, 복잡한 계산방식과 데이터 추출하는 방법과 범위에 따라 오차율이 크게 발생하는 등 많은 보완이 필요하다. 또한, 시간의 변화에 따른 트랜드를 분석할 수 없는 한계점을 가지고 있다.

본 연구의 TF-DI(Term Frequency - Data Index)는 미래의 트랜드를 분석하기 위한 텍스트 마이닝의 알고리즘으로, 특정 키워드가 연도별로 얼마나 중요한지를 나타내는 가중치를 분석하여 TF-IDF를 변형하여 단점들을 보강하고 특정목적(트랜드 분석)을 위하여 개발되었다.

TF-DI의 가장 중요한 원리는 문서내의 중요한 키워드들을 도출하여 인터넷의 정보량에 따라 단어의 빈도수를 분석하는 방식을 사용한다. 또한, 문서군내의 특정단어의 문서 간 빈도수를 사용하는 것이 아니라 연도별 가중치를 사용함으로써 트랜드 분석이 가능하도록 설계되었다. 즉 TF-IDF의 빈도수가 중요하다는 원리를 이용하지만, 트랜드를 분석하기 위하여 문서의 중요도는 인터넷을 활용한 연도별 가중치 분석으로 중요도를 분석하고 있다. 이러한 분석은 현재의 인터넷의 정보를 가중치로 이용하는 방법으로 현재의 데이터나 자료가 잘 반영되어 있다는 장점이 있다.

TF-IDF와 TF-DI의 가장 큰 차이점은 첫째, TF-IDF 빈도수분석은 문서군들에 속해진 문서들의 빈도수를 이용하는 것이고, TF-DI은 중요하다고 생각되는 보고서, 논문 등의 중요문서들의 키워드들을 인터넷의 최근 검색 결과값을 들을 빈도수로 이용하는 것이다. 이 빈도수는 특정한 문서군에 국한되지 않으며, 최근의 트랜드를 반영하는 지표가 된다. 둘째, TF-IDF는 문서간 특정단어의 속한 문서들의 수에 따라 중요도를 판별하였으나, TF-DI에서는 시간의 개념을 변수로 사용하여 중요도를 산정하였다. 이는 시간개념을 도입함으로써, 최근까지의 트랜드를 분석할 수 있는 유용한 방법으르 제시하고 있다. 셋째, TF-IDF의 중요도는 어떤 문서군을 선택하는냐에 따라 편차가 심하게 발생하나, TF-DI는 인터넷을 이용한 결과값들을 사용하기 때문에 오차도 적을뿐더러 다양한 조건하에 값들을 중요도를 추출하고 트랜드를 분석할 수 있다

2) http://www.wikipedia.org 참조바람

3) 추가정보는 “Salton G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill”를 참조 바람

이러한 접근방식의 차이에 의해 TF-IDF와 TF-DI는 중요도를 판별하는 유사한 스타일로 볼수 있으나, 시간의 개념이 포함되면서 완전히 다른 중요도를 도출하는 방법으로 인식된다.

이러한 TF-DI의 세부적인 알고리즘을 살펴보면 2가지의 명제를 중심으로 개발되었다.

<명제>

1. 문서나 프로세스에 의해 도출된 키워드는 빈도수가 높은 것이 중요하다.

표 1. TF-IDF와 TF-DI의 비교분석

분석내용	TF-IDF	TF-DF	비교분석
특성	특정 문서내의 단어를 추출하여 문서와의 관계에 따라, 단어의 우선순위를 도출이 가능함.	다양한 문서에서 단어를 추출하고 인터넷을 활용하기 때문에, 미래예측을 위한 트랜드 분서과 우선순위 도출이 가능함	TF-IDF는 검색엔진의 알고리즘으로 사용가능하나, TF-DF는 검색엔진의 결과를 활용함
사용용도	특정키워드 도출, 특정문서의 우선순위 도출	특정키워드 도출, 미래 트랜드 분석을 위한 키워드 우선순위 도출	트랜드분석의 용이성 TF-IDF < TF-DF
자료의 수집 및 접근 용이성	문서내의 단어 빈도수는 문서의 종류와 문서군이 중요하며, 적절한 문서와 문서군이 아닌 경우 전혀 다른 결과를 초래하므로 적절한 너무 많은 시간과 경비를 소요함	문서를 기준으로 하지 않기 때문에 키워드를 추출하기 위해서 문서의 종류나 문서군에 제약을 받지 않고, 특정한 상황이나 조건을 가정하여 다양한 종류의 문서에서 필요한 키워드를 추출할 수 있음	자료수집의 용이성 TF-IDF < TF-DF
문서의 종류에 대한 효용성	문서내의 단어 빈도수는 논문, 특허 등 문서의 종류에 따라 많은 영향을 받음	인터넷을 통한 단어빈도수를 도출하기 때문에 문서의 종류나 영향을 많이 받지 않으므로, 보고서, 논문, 컨퍼런스 자료등 특정목적에 맞는 다양한 단어를 선택할 수 있음	문서의 제한성 TF-IDF > TF-DF
최신 트랜드 분석 가능 여부	미래예측시 최신 연구경향을 반영하기 위해 논문을 대상으로 하나, 최신 논문의 범위와 논문의 수를 판단하거나 분류하기가 어려움이 따르며, 또한 해당 전문가의 도움이 반드시 필요함	인터넷을 활용하기 때문에 최신 트랜드나 추세를 반영하기 쉬우며, 데이터 소스에서도 논문, 특허, 블러그 등 다양한 소스를 통해 똑 같은 조건내에서 추세를 파악할 수 있음.	최신 트랜드 반영 TF-IDF < TF-DF
연도별 분석 용이성	TF-IDF 알고리즘상 연도별 분석이 거의 불가능하면, 만약 분석하기 위해서는 해당문서의 시간별로 분류해야하기 때문에 거의 불가능함	TF-DF는 연도별 빈도수가 발생하고 연도별 가중치를 염두하고 설계하였기 때문에 트랜드 분석 용이함	시계열분석 용이성 TF-IDF < TF-DF
분석의 오차 범위	문서의 내의 단어의 갯수가 많아질수록 해당 문서는 더 낮은 similarity 값을 가지게 되므로, 문서의 크기가 클수록 저평가가 될 확률이 높음	특정 키워드의 단어수가 아무리 많아도 별로 영향을 받지 않고, 단어가 많을수록 단어별로 분석하기가 용이함.	분석의 오차 TF-IDF > TF-DF
대상문서의 범위의 한계	워낙에 광범위한 주제를 다루는 문서를 대상으로 하는 경우에는 문서에 포함된 단어의 양도 많아지고, 비슷한 주제를 다루는 문서라고 하더라도 전혀 다른 단어 구성을 나타낼 확률이 높음	각 단어별로 연도 및 개별 분석이 가능하고, 특정 주제를 묶어서 분석이 가능하기 때문에,	다양한 키워드 분석 TF-IDF < TF-DF
실제 활용도와 및 시스템 개발의 용이성	문서가 많아질수록 계산하는 과정이 복잡하고 느려지고, 시스템개발이 많은 데이터베이스와 인덱스를 사용하므로 시스템 개발시 시간과 능	구조가 간단하고 심플하여 계산과정이 속도가 빠르며, 구글의 검색 Open API를 활용하기 때문에 유연한 시스템개발이 가능함.	활용도 및 개발용이성 TF-IDF < TF-DF

2. 연도별 키워드 빈도가 높다는 것은 중요한 키워드이다.

첫 번째 명제는 특정 키워드의 빈도수가 높다는 것은 특정 키워드가 중요하다는 것을 의미하므로, 인터넷의 노출 빈도가 높은 키워드를 빈도수로 선정하였다. 즉, 인터넷 검색엔진에 의한 검색결과 값을 빈도수로 선정하였다.

두 번째 명제는 특정키워들의 빈도수가 최근으로 올수록 중요한 요소라고 예상하여, DF(data frequency)에서는 키워드의 연도별 빈도수와 연도별 가중치르 부여하여 산정하였다(최근년도로 갈수록 가중치가 높음)

이러한 두가지 명제를 중심으로 TF-DI에서는 키워드의 상대적 빈도수와 연도별 가중치의 곱으로 표현하여 아래와 같은 수식을 도출하였다.

$$(TF\text{-}DI)_i = \sum_{j=1}^{n} TF_j \times DI_j$$

$$TF_j = \frac{i - frequency}{(total\,frequency)_j}$$

$$DI_j = \frac{j}{n}$$

$i =$ 키워드 구분

$j =$ 시작되는 연도의 첫 횟수$(j = 1)$

$n =$ 분석기간의 연도별 횟수$(j = 1, 2, 3 \ldots)$

3. 연구내용

3.1 정보수집단계

미래과련 키워드를 수집하기 위하여, 2010.2월에 발표된 「과학기술 미래비전」의 내용중 건설분야와 관련된 기술들을 추출하였다.

■ 과학기술 미래비전 4대 미래모습
- ○자연과 함께하는 세상
- ○풍요로운 세상
- ○건강한 세상
- ○편리한 세상(6대 트랜드 중 건설관련 2개 분야)

- 복합공간과 생태도시 개발 기술은 육상 공간의 과밀화와 생활 패턴 변화 등에 대응하기 위해서 점차 그 중요성이 높아질 것이다.
- 새로운 물류·운송 수단의 등장으로 교통수단의 효율성이 증대되어 이동 시간이 단축되고 생활권이 확대될 것이다.

2개의 트랜드의 내용에 언급된 건설 및 교통분야의 29개 기술들을 도출하였다.

3.2 정보처리단계

건설분야 13개의 기술과 교통분야 16개의 기술들로 분류하였으며, 분류된 기술들은 영문 키워드로 재가공되었다.

정보처리 과정에서는 도출된 대상 자료와 기술리스트를 중심으로 각 기술에 해당되는 정보를 추출할 수 있는 데이터 소스를 가공하는 단계로 수집된 자료(문서)를 기본으로 관련 키워드의 추출과 추출된 키워드의 수정 및 검토를 통한 영문화 하였다.

표 2. 한글 및 영문 주요키워드

No	기술명(한글)	기술명(영문)
1	대규모 구조물	large structure
2	지하공간	deep underground
3	인공섬	Artificial island
4	해저터널	undersea tunnel
5	해양도시	ocean city
6	첨단 폐기물 시설	Waste Manifest System
7	초고층 빌딩	Skyscraper
8	도시방재	urban disaster prevention
9	시설물 안정성	facility safety
10	국토공간정보 활용	spatial information
11	무인 시공현장	automated construction
12	건축물 이동 기술	building motion
13	우주기지 건설	space station construction
14	무인 자동 운전 시스템	unmanned vehicles
15	대심도 지하도로	deep road
16	대심도 지하철도	deep railroad
17	자기부상열차	magnetic levitation train
18	수소자동차	hydrogen car
19	연료전지 교통기관	fuel cell transportation
20	첨단 ITS	intelligent transportation system

21	첨단 자동차 안전기술	Transportation safety
22	파일럿 정보시스템	Head Up Display(HUD)
23	실시간 차량정보 시스템	Vehicle Information Communication System(VICS)
24	개인선택형 대중교통	Personal Rapid Transit(PRT)
25	지능형 화물 운송 및 관리	intelligent logistics
26	유비쿼터스 통합물류 정보기술	ubiquitous logistics
27	진공터널 등의 고속 물류운송 튜브망	tube transportation
28	대량 물류 수송수단	large transportation
29	우주 수송기	space transportation

3.3 정보처리단계

정보처리 단계에서 도출된 영문 키워드를 중심으로 TF-DI(Term Frequency - Data Index)를 활용하여 키워드의 가중치 도출하였다. 가중치는 구글 검색엔진을 활용하여 연도별 검색 결과 커리값을 가중치로 도출하였다. 검색기간은 2000.1.1. ~ 2009.12.31.(10년간)로 설정하여 TF-DI에 의해 주요 키워드를 선정하였다. 특히, 범위가 크거나 대표성이 있는 기술은 제외하고 같은 기술의 여러 가지 세부기술들은 우선순위가 높은 기술로 선정하였다(예, 초고층 빌딩 설비 및 환경, 초고층 빌딩 설계, 초고층 빌딩 계획 등은 중요도 지수가 높은 것을 선정). 또한, 다른 기술에 비해 TF-DF 너무 낮은 것은 제외하였다(예, ubiquitous 물류, 공간정보 기반 인프라 기술, 친환경 주거, U-Transportation 기술, 중소형항공기)

표 3. TF-DI에 따른 기술의 우선순위

No	분류	기술명(한글)	TF-DI
7	건설	초고층 빌딩	1.8556025
5	건설	해양도시	1.0999907
10	건설	국토공간정보 활용	0.6536529
2	건설	지하공간	0.5906837
21	교통	첨단 자동차 안전기술	0.439549
22	교통	파일럿 정보시스템	0.169426
29	교통	우주 수송기	0.1127485
1	건설	대규모 구조물	0.1015096
20	교통	첨단 ITS	0.0991385
14	교통	무인 자동 운전 시스템	0.0594938
3	건설	인공섬	0.0588668
18	교통	수소자동차	0.0544719
9	건설	시설물 안정성	0.0479931
28	교통	대량 물류 수송수단	0.0371752
11	건설	무인시공현장	0.0271896
24	교통	개인선택형 대중교통	0.026965
4	건설	해저터널	0.0117184
17	교통	자기부상열차	0.0109483
12	건설	건축물 이동 기술	0.0101684
25	교통	지능형 화물 운송 및 관리	0.007569
27	교통	진공터널 등의 고속 물류운송 튜브망	0.0050366
13	건설	우주기지 건설	0.0047533
15	교통	대심도 지하도로	0.004304
26	교통	유비쿼터스 통합물류 정보기술	0.004052
8	건설	도시방재	0.0026285
19	교통	연료전지 교통기관	0.0013928
6	건설	첨단 폐기물 시설	0.0013735
23	교통	실시간 차량정보 시스템	0.0010955
16	교통	대심도 지하철도	0.0005029

3.4 정보분석단계

정보분석 단계에서는 최종적으로 29개의 기술들을 도출하였다. 특히, 1~10위까지의 기술을 분석할 결과 표4와 같다.

건설분야에서는 초고층빌딩과 대규모 구조물의 초고층화, 지하공간, 해양도시를 개발하는 기술들은 현재의 공간을 좀 더 효율적으로 활용하기 위한 중요한 미래기술로 도출되었다. 또한, IT기술을 활용한 국토공간정보의 활용한 기술들도 IT와 결합되어 중요한 기술로 도출되었다. 특히, 구글, 마이크로소트프 등의 회사에서는 원천 공간정보를 확보하는데 주력하고 있으며, 대부분의 상용화된 지도서비스는 가장 큰 포털에서 중요한 서비스로 자리잡고 있다.

교통분야에서는 첨단 IT를 활용하여 교통의 안정성과 관련기술의 고도화를 주력하는데 필요한 기술이 미래의 주요 기술들로 도출되었다. 특히, 우주 수송기와 같이 먼 우주를 여행할 수 있는 관련기술과 이에 따른 운전의 안정성을 극대화하는 기술들이 미래의 기술들로 도출되었다.

표 4. 미래 주요 기술들

우선순위	분류	기술명(한글)
1	건설	초고층 빌딩
2		해양도시
3		국토공간정보 활용
4		지하공간
8		대규모 구조물
5	교통	첨단 자동차 안전기술
6		파일럿 정보시스템
7		우주 수송기
9		첨단 ITS
10		무인 자동 운전 시스템

5. 결 론

본 연구에서는 트랜드를 분석하기 위하여 새로운 텍스트마이닝기법과 수학적 알고리즘을 연구하였다. 또한, 인터넷 검색엔진을 이용하여 검색결과값을 주요 키워드의 빈도로 사용하였다. 이러한 원리로 주요 키워드의 미래 트랜드를 분석할 수 있었다. 특히, 본 연구에서 개발된 TF-DI는 인터넷을 활용하여 시간별 추이를 분석하는데 용이 할 것으로 예상된다. 이러한 연구결과로 다음과 같은 결론은 낼 수 있었다.

첫째, 인터넷을 활용한 트랜드 분석은 향후 아주 중요한 리소스를 제공하는 역할을 할 것으로 예상되며, 본 연구는 인터넷 데이터를 어떻게 활용하는 것에 대한 하나의 방법을 제시하고 있다.

둘째, 텍스트마이닝을 이용한 트랜드 분석은 미래 키워드 분석하는 하나의 방법으로 향 후 중용한 연구 테마가 될 것으로 예상된다.

셋째, 본 정량적인 연구방법은 전문가를 활용한 정성적인 미래예측을 보조하거나 예측의 일부를 대처하는 하나의 방법으로 활용될 것으로 예상된다.

참고문헌

1. Byeongwon Park, 2007. Development of method and framework for Korean Technology Foresight Program, KISTEP Press

2. Byungnam Kahng, 2009. Complex Networks Science, Jipmoon Press, Seoul

3. C. Tsallis and M.P. de Albuquerquer, 2000 C. Tsallis and M.P. de Albuquerquer, Eur. Phys. J. B13(2000)

4. Cornelia Daheim, 2007. Regional Foresight in Europe 2 Examples: Duesseldorf and Linz, WFS Conference Minneapolis

5. D.J. Watts and S.H. Strogatz, 1998. Collective dynamics of 'small-world' networks, Nature 393

DongWon Sohn, 2002. Social network analysis, Kyungmoon Press

6. G. Sabidussi, 1966G. The centrality index of a graph. Psychometrika, 31(4) 581-603

7. Jae-Yun Lee, 2006. A Study on the Network Generation Methods for Examining the Intellectual Structure of Knowledge Domains, Korean Society for Library and Information Science 40 333-355

8. James A. Dator, 2002. Advancing Futures: Futures Studies in Higher Education, Westport Press

9. Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant, 2009. Detecting influenza epidemics using search engine query data, Nature 457, 1012-1014

10. M. Faloutsos, P. Faloutsos, and C. Faloutsos, 1999. On power-law relationships of the internet topology,

11. Martin Hilbert, Ian Miles, and Julia Othmer, 2009. Foresight tools for participative policy-making in inter-governmental processes in developing countries: Lessons learned from the eLAC Policy Priorities Delphi, Technological Forecasting and Social Change, Volume 76, Issue 7 880-896

12. Mats Lindgren,Hans Bandhold, 2002. Scenario planning: The link between future and strategy, Macmillan Press

13. Michael Marien, 2002. Future Studies in the 21st Century: A Reality-Based. View, Futures, Vol. 34 261-281

14. Mikko Syrjnen, Yuko Ito, Eija Ahola, 2009. Foresight for Our Future Society: Cooperative project between NISTEP(Japan) and Tekes(Finland), Tekes & NISTEP Press

15. OECD, 2001. Governance in the 21st Century: FUTURE STUDIES, OECD Press, Paris

논문접수일 (2012. 5. 16)

심사완료일 (1차 : 2012. 5. 30, 2차 : 해당 없음)

게재확정일 (2012. 6. 4)