

TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석

Analysis of Changes in the Housing Market Using TF-IDF Weight Model

박 종 영* · 서 충 원**

Park, Jong Young · Suh, Chung Won

目 次

- | | |
|--------------|------------|
| I. 서 론 | 3. 선행 연구 |
| 1. 연구배경과 목적 | III. 분석 결과 |
| 2. 연구범위와 방법 | IX. 결론 |
| II. 관련 이론 | 〈abstract〉 |
| 1. 데이터마이닝 이론 | 〈참고문헌〉 |
| 2. 분석 모형 | |

ABSTRACT

1. CONTENTS

(1) RESEARCH OBJECTIVES

The purpose of this study is to look for the changing characteristics of the housing market by analyze the news story by using the TF-IDF weight models.

(2) RESEARCH METHOD

Data were collected for 280 of articles using Naver and Google from 2012 to 2014. The collected data underwent a preprocessing to remove stopwords. It used Textom Solution to analyze Word frequency and N-gram.

(3) RESEARCH FINDINGS

The 2012 housing market's primary keywords was slump and acquisition tax, capital gains tax, reconstruction, small-sized apartments. The 2013 primary keywords

* 주 저 자 : 강남대학교 부동산학과 박사과정. pakka74@naver.com

** 교신저자 : 강남대학교 부동산학과 교수. 도시계획 공학박사. suhwon@kangnam.ac.kr

▷ 접수일(2015년 7월 17일), 수정일(1차 : 2015년 8월 13일, 2차 : 2015년 9월 12일), 게재확정일(2015년 12월 18일)

was Auction, remodeling, presidential election, reconstruction, unsold house. 2014 main keywords was rising, recovery, Rental deposit, Monthly rent.

2. RESULTS

The Presidential Election is providing a diverse policy. Acquisition tax cuts and abolition of capital gains tax on reviving the sagging housing market sentiment continued contributions, and the takeover and housing market is fizzing. Reconstruction and contributed in revitalizing the price to increase and ease. Rent is up and tenants took part in a market is leading. Therefore, real estate measures play an important role in the housing market.

3. KEY WORDS

- Housing market, Word frequency, N-gram, TF-IDF, Text Mining

국문초록

최근 전 세계적으로 빅데이터에 대한 관심과 활용에 대한 이야기들이 흥미로운 소재를 안고 소개되고 있다. 사회 현상에 대한 복잡하고 다양한 데이터를 통계적 사고와 마이닝적 관찰을 통해 얻어내는 패턴의 발견인 것이다. 빅데이터 분석 방법 중 텍스트마이닝은 텍스트 문서의 복잡한 특성을 전처리 과정을 통해 정제된 후 문서가 전달하고자 하는 메시지를 분석하여 의미를 찾는 기법이다. 본 연구는 부동산 뉴스 기사를 수집하여 주택시장과 관련된 단어를 빈도수, N-gram 순열, TF-IDF 가중치를 구하여 시기별로 다루어지는 핵심 단어를 발견하므로서 당시의 주택시장 상황을 파악하고 연구에 적용된 분석기법을 소개하고자 한다. 텍스트 문서 분석에 중요한 역할을 하고 있는 빈도수, N-gram, TF-IDF 가중치 모델을 적용해 보면서 부동산학에서의 활용방안을 제시해보고 이를 통해 텍스트마이닝 기법이 부동산 연구 분야에 적극적으로 활용될 수 있도록 그 관심이 점차 커지길 기대해본다.

핵심어 : 주택시장, 텍스트마이닝, 빈도분석, 엔-그램, 역문서빈도

I. 서 론

1. 연구배경 및 목적

디지털 혁명과 소셜미디어의 등장 그리고 데이터의 저장 기술이 발전하면서 빅데이터에 대한 사회적 관심이 커지고 있다. 블로그, SNS,

뉴스 매체 등을 통해서 방대한 정보를 생산해내고 그 정보는 공유자들로부터 또 다른 정보로 재생산되어 쏟아져 나온다.

토마스 멀론(Thomas Malone)과 같은 일부 학자들은 빅데이터를 통해 인간 행동을 미리 예측할 수 있는 세상이 열리고 있다고 주장하기도 하였다. 빅데이터는 수십억 테라바이트의 양

적(Volume) 개념과 데이터의 종류가 문자, 음성, 영상 등으로 다양성(Variety)을 갖추고 있으며 데이터의 수집·저장 속도(Velocity)가 매우 빠르다는 3V로 정의했으나 최근에는 빅데이터의 분석을 통해 가치(Value) 있는 정보를 추출하여 활용하는 시대로 4V로 표현한다.¹⁾

빅데이터에서의 가치(Value)는 ‘복잡하게 얽혀있는 현상속에서 의미있는 패턴이나 규칙을 찾아내는 것’으로 표현된다. 날마다 새롭게 만들어지는 엄청난 데이터 홍수 속에서 기업들은 앞다퉈 데이터 분석을 통한 의미있는 가치를 찾아내려고 노력하고 있다.

빅데이터 분석 방법 중의 하나인 텍스트마이닝은 미국질병통제예방센터에서 발표하는 독감 예측보다 구글에서 2주 먼저 발표하는 독감트렌드를 예측할 만큼 신속하고 정확한 분석사례로 회자되고 있다.

최근 우리나라 기업들도 소비자의 구매패턴이나 상품평 등의 소비자 구매심리를 파악하여 기업마케팅 전략에 뉴스기사 마이닝(Mining)이나 소셜네트워크(트위터, 페이스북, 블로그, 질문과 답변 등)분석을 활용하고 있다. 뉴스기사 마이닝이나 소셜네트워크 분석은 실시간으로 올라오는 텍스트를 기반으로 분석자가 의도하는 패턴을 찾아내는 텍스트마이닝 방법의 하나이다.

부동산 관련 기업이나 지자체, 정부출연기관들도 필요성을 인식하고 빅데이터 분석 소프트웨어나 솔루션 라이선스를 구매하는 비율이 늘어나고 있다. 이는 부동산 관련 통계를 새로운 시각에서 폭넓은 가치를 찾고자 하는 시대적 요구에 부응하는 시도라고 본다.

부동산은 다양한 분야에서 서로 연관성을 가지고 복잡하게 작용하고 있으며 이를 이해하고 해석하려는 학문적 연구들은 그동안 많은 성

과를 이루어냈다. 부동산 빅데이터를 필연적으로 받아들여야 하는 시대에서 우리가 이해하고 있는 부동산을 새로운 방법으로 의미 있는 패턴을 찾아내는 것은 중요한 일이다.

부동산 뉴스보도가 주택시장에 대한 사회적 이슈와 시장 상황에 대해 어떻게 표현하고 있는지를 살펴보고 패턴을 찾아보는 것은 의미 있는 일이다. 이를 위해서는 텍스트로 이루어져 있는 뉴스기사 메시지의 의미 구조와 규칙을 파악하는 분석이 필요하다.

본 연구는 부동산에 대한 뉴스기사 텍스트 문서를 점수화 하여 가중치를 부여함으로써 시기별 주택 시장전망에 대한 판단의 근거를 제시할 수 있다는 가정에서 출발한다. 이를 통해 구축된 경험 값은 주택가격 마이닝기법 중 나이브베이즈, 로지스틱 회귀분석에 활용될 수 있다는 점에서 연구의의를 찾을 수 있다.

1. 연구범위 및 방법

2012년부터 2014년까지 최근 3년간 네이버, 구글의 뉴스란 항목에서 ‘주택 시장 전망’ 키워드 검색을 통해 280건의 관련 기사를 수집하였다. 네이버를 주요 검색으로 이용하였고 구글은 부족한 자료를 추가로 찾기 위한 보조 검색으로 하여 중복 기사 수집을 사전에 차단하였다.

추출한 기사는 분기별로 분류한 후 코퍼스(Corpus)정규화²⁾ 하였고, 자료 집단 내에서 관련성이 떨어지는 내용이나 광고성 내용은 제거하였다.

정제된 자료는 단어빈도수 분석 솔루션인 Textom을 이용하여 순위에 따라 주요 단어를 추출·분류하였다. 조사, 접속사, 문장의 어미, 의미 없는 숫자, 특수 문자 등은 제거하여 데이터 전처리 과정을 거쳤다. 이렇게 분류된 단어는

1) 김정선·권은주·송태민, “소셜 빅데이터를 활용한 국내 빅데이터 수요공급 예측”, 한국경영정보학회지, 한국경영정보학회, 2014, 2014No 1, p.709.

2) 코퍼스(Corpus)- 문서를 모아놓은 것을 말한다. 문서는 단어 하나하나로 이루어진 토큰(Token)으로 구성되어 있으며, 데이터마이닝 기법에서 흔히 사용되는 개념이며 단어빈도, 유사도, 군집화, TFIDF, 단어주머니, n-그램 순열 등에서 자주 등장하는 용어이다. 본 논문에서의 코퍼스는 분기별로 신문기사를 모아놓은 문서이다.

N-gram 순열을 이용하여 단어-단어 구조로 연결하여 추출된 단어가 표현하고자 하는 의미를 명확하게 분류해주었다. 끝으로 문서에서 핵심적인 역할을 하는 단어의 중요도를 평가하는 TF-IDF 가중치를 계산하였다.

TF-IDF 가중치 모델은 부동산 뉴스기사 키워드의 출현 빈도수를 추출하여 시기별 시장 상황을 직관적으로 살펴보고 이를 기반으로 단어와 문서간의 가중치 값을 구하므로써 문서 집합의 주제 탐색과 핵심 단어를 정하는 모델이다.

II. 관련 이론

1. 데이터마이닝 이론

데이터마이닝(Data Mining)은 원시상태의 자료를 전처리 과정을 통해 통계뿐만 아니라 그와 관련된 패턴인식 기술들을 이용하여 대용량의 데이터로부터 의미있는 정보를 찾는 분석 방법이다.³⁾ 의미있는 패턴과 규칙을 발견하기 위하여 분석 도구를 이용하여 데이터를 분석하는데 대표적인 분석방법은 분류(Classification), 추정(estimation), 군집(Clustering), 연관규칙(Association Rule), 예측(Prediction), 프로파일링, 텍스트마이닝(Text Mining) 등을 들 수 있다.⁴⁾

분류 분석은 데이터마이닝에서 가장 많이 사용되는 분석방법으로서 과거 경험치가 포함된 데이터의 분류 특성을 파악하고 이를 토대로 새로운 데이터에 대해서 분류, 규칙, 특성 등을 찾아내어 몇 개의 그룹으로 나누어 예측하는 방법이다. 대표적인 분류기법에는 로지스틱 회귀분

석, 나이브베이지, 기계학습추론(Support Vector Machine), 인공신경망, 의사결정트리 기법이 있다. 트리기반에 의한 분류 분석은 특정 범주에 영향을 미치는 다른 범주들의 계층적인 구조를 파악할 수 있다.⁵⁾ 주택담보대출 차입자의 연체가능성에 대해서 연체자와 비연체자로 구분하여 로지스틱 회귀분석, 인공신경망분석, 의사결정트리분석을 통해 모형별 예측력을 비교하여 대출기간, 금리조정기간, 주택규모 등이 연체와 연관성을 가지고 있음을 분석한 사례가 있다.⁶⁾

군집 분석은 유사한 속성을 갖는 객체들끼리 묶어 그룹핑하는 과정이며 군집간의 유사도나 거리함수에 기반한다. 자연휴양림 이용객의 만족도 및 재방문의도 그리고 추천의도를 설문하여 미혼이면서 소득100만원 미만 그룹과 소득이 200~300만원인 기혼 그룹으로 군집화하여 고객 그룹에 따라 행동의도의 차이를 알아본 분석사례가 있다.⁷⁾

연관규칙은 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙으로서 장바구니 분석이라고도 한다. 월마트 사례가 대표적인 연관규칙분석인데 허리케인이 지나가면 어떤 일이 벌어질지에 대한 과거 경험을 통해 손전등 외에 배터리, 생수, 통조림 등 고립과 단전·단수에 대비한 비상물품을 구매하는데 연관성 있는 물품들간의 관계를 찾아내는 것이다.

끝으로, 텍스트마이닝은 방대한 텍스트 문서에서 의미 있는 정보를 추출하는 것을 말하며 분석 대상은 신문기사, SNS, 블로그, 질문게시판, 설문지 등이 있다. 텍스트 분석을 통해 과거뿐만 아니라 현재 시점에서 실시간으로 이슈가 되는 키워드를 찾아내어 의미를 분석하는 과정

3) 김완섭·이수원, “데이터마이닝을 이용한 설문조사의 심층분석”, 한국정보처리학회지, 한국정보처리학회, 2006, 제9권 제4호, p.73.

4) 김중우·김선태, 경영을 위한 데이터마이닝, 한경사, 2009, pp.21-25.

5) 네트워크 관계는 링크(link)와 노드(node)로 표현하는데 계층형 트리는 링크를 계층화하여 표현하는 것으로 노드 쌍이 부모와 자식 관계(parent-child relationship)형식일 때 사용하는 방식이다.(이수상, 네트워크 분석방법론, 논형, 2012, pp134-173.)

6) 유재술, 데이터마이닝기법을 활용한 주택담보대출 연체가능성 분석 연구, 한성대학교 대학원 박사학위논문, 2015, pp.70-84.

7) 신현규·신홍철, “군집분석을 통한 자연휴양림 이용객의 시장세분화”, 한국콘텐츠학회 논문집, 한국콘텐츠학회, 2010, 제10집 제3호, pp.364-372.

이다. 분석 대상 문서자료의 형태가 일정하지 않은 대표적인 비정형 텍스트 데이터로서 숫자, 형용사, 조사 등의 불용어를 제거하는 전처리 과정이 필요하다. 텍스트마이닝은 큰 분류에서의 데이터마이닝 분석방법이며 최근에 각 분야에서 활발하게 연구되고 있다. 소셜네트워크분석은 노드(Node)와 링크(Link)관계를 두고 연결 관계를 찾는 분석기법이다.⁸⁾ 학회지의 주저자와 공동저자, 교신저자간의 학술교류 관계에서 위상구조, 확산·진화과정을 찾거나 후원금을 주고 받은 정치인들의 연결 관계를 찾는 대표적인 사례가 있다.

2. 분석 모형

1) 자료의 수집

뉴스 기사는 전달하고자 하는 내용이 텍스트만으로 구성되어 있지 않다. 표 형식, 이미지, 광고 URL, 분양상품의 기획성 기사, 중복내용의 기사 등 불필요한 내용들이 다수 포함되어 있다. 무엇보다도 기사 작성자가 글의 검색을 위하여 달아 두는 태그(tag)에는 지나치게 많은 단어들이 포함되어 있어 찾고자 하는 키워드 내용과 관련성이 떨어지는 기사들이 검색되는 경우가 많다. 결국, 자료의 수집과정에서 대략적으로 문

서를 스킴오버(Skim over)로 분류하여 관련 없는 문서들은 제거해야 한다.

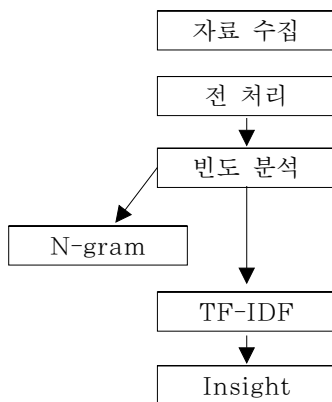
2) 전처리(Preprocessing)

전처리(Preprocessing)는 분석 대상의 데이터를 제대로 분석하기 위한 정제 과정이다. 수집과정에서 불필요한 기사들을 제거하지만 미처 발견하지 못하는 불필요한 텍스트들이 문서 내에 포함되어 있는 경우가 많다. 실질적으로 의미가 없는 조사, 어미, 접속사 등의 불용어도 제거해야 한다. 경우에 따라 접속사를 불용어로 취급할지는 따로 고민하는 경우도 있다. 문서 내에서 문맥상 비슷한 의미를 전달하는 단어들은 한 묶음으로 하여 대표성을 가지는 단어로 정제해줘야 한다. 동일한 의미를 가지는 단어가 포함되어 있을 경우 문서에서 핵심적인 단어를 분류하더라도 잘못된 방향으로 분석될 수 있기 때문이다.

〈표 1〉 비슷한 의미의 단어 묶음 처리

처리 전	처리 후
금리, 이자	금리
2주택자, 다주택자	다주택자
수요자, 매수자	수요자
방안, 대책	대책
전세값, 전세가격	전세값
전세시장, 전세	전세
상승, 올라가는	상승
하락, 내려가는, 떨어지는	하락
거품, 버블	거품
분양시장, 분양	분양
급증, 증가	증가

〈그림 1〉 분석 알고리즘



3) 빈도 분석(Frequency)

뉴스기사나 소셜네트워크 등과 같은 비정형 데이터인 코퍼스(corpus) 언어분석은 언어학의 한 분야로 언어통계학적으로 중요하게 다루어지고 있다. 대표적인 분석 방법이 빈도 분석인데

8) A.L. 바라바시. 링크, 21세기를 지배하는 네트워크과학, 동아사이, 2015, pp.23-48.

동일한 의미를 표현하더라도 빈도가 높은 단어는 빈도가 낮은 단어에 비해서 기억을 더 잘하는 심리적 연관성과 관련이 있으며 이를 빈도효과(frequency effect)라고 한다.⁹⁾ 예를 들어, '구매'라는 단어보다는 '매수'라는 단어를 더 잘 기억한다. '방안' 보다는 '대책', '양도소득세' 보다는 '양도세', '버블' 보다는 '거품'이라는 표현에 더 익숙하다. 따라서, 빈도수는 해당 문서에서 표현하고자 하는 메시지를 함축적으로 전달하므로 연구자가 적절한 단어를 선택하여야 한다.

언어자료는 빈도가 높은 단어와 낮은 단어 간의 차이가 커서 대표적인 멱함수 분포를 보이며 표본에 대한 평균과 표준편차는 적용 될 수 없다. 결국, 언어자료는 수집한 데이터의 전부를 전수 조사하는 것을 원칙으로 한다.

언어자료와 관련된 분포통계는 지프의 법칙(Zipf's law)을 적용하며 빈도와 빈도순 사이는 반비례 관계로 다음과 같이 표현한다.

$$f \times k = r, \quad (f \propto \frac{1}{r}) \quad (1)$$

f : 빈도, r : 빈도순, k : 상수

4) N-gram 순열

빈도분석을 통해 나온 단어들은 원래의 자기 자리에 있던 순서를 완전히 무시하고 처리되므로 단어 자체가 가지는 의미를 제대로 표현하지 못할 수 있다. 이를 보완하기 위해 문서에서 인접한 단어들을 쌍으로 묶어 표현하는 N-gram 순열 기법을 적용한다. 대표적으로 '상승'과 '증가'라는 단어는 '집값-상승' 또는 '금리-상승' 그리고 '거래량-증가' 또는 '세부담-증가'와 같이 긍정과 부정의 의미를 모두 포함하고 있기 때문에 N-gram 순열은 문서의 의미를 정확히

전달하는 중요한 기법이라고 할 수 있다. N-gram 순열에서 요약된 단어들이 해당 문서들의 내용을 함축적으로 표현해주기 때문에 이를 기반으로 TF-IDF의 빈도 단어 선정에 상당 부분 도움이 된다.

5) TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency: 역문서 빈도) 가중치 모델은 텍스트마이닝을 위해서 문서 내부의 단어가 얼마나 중요한지를 평가하기 위한 표현방식이다. TF-IDF 가중치 값이 큰 단어일수록 속해 있는 문서의 주제나 의미를 결정지를 가능성이 크며 이 측정치를 주요 키워드를 추출할 수 있는 척도로 활용할 수 있다.¹⁰⁾

이 모델은 SNS나 뉴스기사에서 자주 언급되는 이슈들을 분석하는데 유용하게 쓰여진다. 실시간으로 떠도는 이슈가 되는 내용에 대해 출처, 유용한 정보, 네트워크 관계 등을 발견할 수 있다.

빈도수가 높은 단어와 문서에 점수를 부여한 후 확률모형을 통해 주가 변동을 예측한 연구는¹¹⁾ 역문서 빈도가 출현 단어와 해당 문서의 중요도를 얼마나 핵심적으로 파악하고 있는지 보여주는 사례라고 볼 수 있다.

TF(Term frequency)는 한 문서 내에서 등장하는 단어의 빈도를 나타내는데 단어와 문서 간의 중요도를 나타내기 위한 것이다. 문서 내에서 많이 출현 할수록 상대적으로 더 중요하다는 의미이다. 하지만, 지나치게 빈도수가 높은 단어는 불용어 수준의 단어에 해당될 수 있다. 따라서, TF 분석 이후 연구자가 의도하는 주제에 따라 불용어 처리 여부를 결정해야 한다.

문서 d_j 에서 단어 t_i 의 중요도는 다음과 같다.

9) 김동성, 언어자료 분석을 위한 통계학, 한국외국어대학교 출판부, 2010, pp.33-37.

10) 이성직·김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출기법", 한국전자거래학회보, 한국전자거래학회, 2009, 제14집 제4호, p.61.

11) Fawcett·Provost, Data science for business, O'REILLY·Hanbit, 2014, pp.312-323.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

$n_{i,j}$: 문서 d_j 에서 단어 t_i 가 나오는 횟수

$\sum_k n_{k,j}$: 문서 d_j 에서 나오는 모든 단어횟수

DF(Document frequency)는 문서 빈도인데 자주 등장하는 단어가 몇 개의 문서에 등장하는지를 나타낸다.

IDF는 DF의 역수이며 로그를 취해 준다. 단어 간의 거리를 일정하게 유지하기 위해서 로그를 취해 주는데 자연로그나 상용로그 중 선택하면 된다. DF는 전체 문서 수(d_j)에 대한 단어(t_i)의 중요도를 나타낸다. 값이 작은 단어는 흔한 단어일 가능성이 크며 값이 큰 단어는 희귀한 단어일 가능성이 크다. 희귀한 단어는 IDF가 높게 나오다가 단어(t_i)가 나오는 문서가 늘어나면 IDF가 감소하면서 임계치인 0에 가깝게 내려간다. 대부분의 불용어는 대부분의 문서에 포함되므로 IDF는 0에 가깝게 나온다.¹²⁾

$$IDF_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (3)$$

$|\{d_j : t_i \in d_j\}|$: t_i 가 나오는 문서 수

$|D|$: 전체 문서 수

TF-IDF 값은 TF와 IDF를 곱한 것이며 TF-IDF 가중치 값에 따라 문서 내에서 단어(t_i)의 출현 빈도가 높고 전체 문서에서 단어(t_i)가 출현하는 문서들의 수가 적은 단어가 중요한 단어로 평가된다.

$$TDIDF_{i,j} = TF_{i,j} \times IDF_i \quad (4)$$

TF-IDF 가중치는 하나의 문서에서 TF가 크고 전체 문서에서 DF가 작을수록 커진다. 값이 클수록 상대적으로 문서내에서 핵심적인 단어이다. 가령, '상승'이라는 단어의 빈도는 매우 높아 불용어 수준이지만 포함하는 문서 수가 적으면 문서내에서 많은 빈도로 언급되고 있다는 의미로서 핵심적인 메시지를 담고 있는 문서로 볼 수 있다.

3. 선행 연구

TF-IDF 가중치에 대한 연구는 컴퓨터 정보, 언어, 경영학 등에서 다양하게 진행되었다. TF-IDF 가중치 값은 그 자체로서 활용가치와 해석적 의미를 가지지만 다른 마이닝 기법과 콜라보를 통해 더 의미있는 가치를 얻어낼 수 있다.

조태호·서정현(2000)¹³⁾은 TF-IDF에 근거하여 수집된 뉴스기사의 내용을 재료로 주요 키워드를 추출하고 문서내 빈도와 역문서빈도, 단어의 위치를 고려한 신경망 접근 모델로 키워드 판별 성능을 향상시킬 수 있음을 보여주었다. 역문서 빈도는 키워드 분류 및 선택에 중요한 역할을 해주는데 분류마이닝 기법의 하나인 신경망을 통해 검증하였다.

조은경(2015)¹⁴⁾은 수많은 텍스트가 어떤 내용에 대해서 어떤 의견이나 감정을 가지고 있는지를 영화감상평을 재료로 하여 '좋다-나쁘다'의 범주로 분류하는 감성분석을 하였다. TF-IDF를 활용하여 감성단어에 영향을 주는 함수적 어휘를 구하고 이를 통해 얻어진 값은 베이즈 확률을 구하여 단어의 감성분류를 도출하였다.

이성직·김한준(2009)¹⁵⁾은 분야별 뉴스기

12) Fawcett·Provost, 전게서, p.300.

13) 조태호·서정현, "문서의 키워드 추출에 대한 신경망 접근", 한국정보과학회보, 한국정보과학회, 2000, 27권 제2호, pp.317-319.

14) 조은경, "감성분석에서 함수적 어휘와 대하어성 추출", 한국어 의미학 학술지, 한국어 의미학회, 2015, 제47권, 제0호, pp.1-26.

15) 이성직·김한준, 전게논문, pp.59-73.

사에서 의미있는 키워드를 추출하기 위하여 기존의 TF-IDF 가중치 모델에 6가지 변형식을 제안하여 추출능력이 향상됨을 관찰하고 키워드의 사회적 관심도의 변화에 따라 그 순위가 변화됨을 관찰하였다.

뉴스기사는 기자의 주관적인 의지가 많이 포함되므로 오피니언 마이닝 기법으로 분류하기도 한다. 주택시장 전망에 대한 전문가의 견해와 시장지표를 재해석하여 기사를 작성하므로 문서에 대한 작성자의 감성이 스며들어 있다. TF-IDF 기법은 주관적이거나 혹은 객관적이라고 하더라도 그 문서가 담고 있는 감정을 핵심 단어를 통해 속성을 찾는 과정이다.

Ⅲ. 분석 결과

연도별 키워드 검색을 통해 단어 빈도수를 분석한 결과 2012년에 많은 빈도수로 언급된 용어는 '소형', '수익형' 단어였다. 저금리 기조가 계속되면서 시세차익보다는 안정적인 임대수익에 초점을 맞추면서 오피스텔이나 도시형생활주택

이 투자 상품으로 인기를 누렸던 시점이 기사에 반영된 것으로 보인다. '금융위기'의 여파와 가계 부채 '증가', 대출 '규제' 부담과 주택가격 '하락' 그리고 주택가격 '회복'에 대한 기대감, 주택 거래량 '감소'등의 부정적인 단어를 통해 당시의 주택시장의 상황을 엿볼 수 있다. '충선대선'이 있는 해로 부동산 관련 공약들이 '회복'에 긍정적인 신호를 보낼지 관심이 몰렸던 시기이기도 하다.

전국 미분양 주택이 서울과 '수도권'을 중심으로 줄어들었던 시기이기도 하며 충선 공약의 대부분이 개발보다는 '전월세' 안정화에 초점이 맞춰져 있었고 각종 규제완화 '대책'의 일환으로 '양도세' 중과 및 분양가상한제 '폐지'에 대한 관심이 증폭된 시기였음을 알 수 있다.

2013년은 대선이 끝나고 부동산 정책변화의 기대감으로 세금관련 '취득세', '양도세'와 규제완화 측면의 '재건축', '리모델링'에 대한 단어 빈도가 많았음을 짐작할 수 있다.

'전세'가격 '상승'과 '오피스텔'과 같은 수익형 부동산의 '공급' '증가', '월세' 비중의 상승, 주택가격 '하락'에 대한 빈도수가 높은 것을 확인할 수 있으며 전세 가격의 급등은 월세 비중의 증가로 이어지는 악순환이 계속 이어지고 있음

〈표 2〉 연도별 주요 단어 빈도수

빈도수		주요 단어	분기별 기사 수	전 처리 후 사용된 단어 수
				총 단어 출현 빈도수 분석 대상 기사 수
2012	1q	소형, 수익형, 충선대선, 재건축, 하락, 금융위기, 상승, 회복, 공급, 거래량, 뉴타운	38	179
	2q	대책, 집값, 상승, 투기지역, 강남, 중국, 양도세, 감소, 침체, 월세, 분양, 완화, 회복	19	5153
	3q	거래량, 양도세, 하락, 침체, 감소, 중국, 대책, 투자, 지속, 취득세, 폐지, 종과세, 인하	21	
	4q	수도권, 하락, 상승, 전세, 수요, 비율, 공급, 증가, 규제, 전세값, 침체, 취득세, 재건축	27	105
2013	1q	취득세, 대책, 양도세, 전세, 중소형, 분양, 침체, 연장, 폐지, 박근혜, 다주택, 대출	14	162
	2q	대책, 리모델링, 수직증축, 오피스텔, 공급, 기대, 증가, 상승, 분양, 침체, 양도세	20	5139
	3q	취득세, 재건축, 인하, 전세, 분양, 지방, 9억원, 미분양, 서울, 건설사, 청약, 지속	17	
	4q	상승, 전세, 집값, 하락, 월세, 공급, 수요, 취득세, 대출, 경매, 입주, 지방, 입주물량	34	85
2014	1q	재건축, 상승, 강남, 전세, 폐지, 대책, 거래, 다주택, 월세, 규제, 가세, 투자자, 회복세	13	173
	2q	상승, 재건축, 하락, 강남, 분양, 증가, 물량, 공급, 수요자, 과천, 가계대출, 과세, 규제	14	6278
	3q	재건축, 상승, 규제, 완화, 지방, 공급, 대출, 수요, 투자, 효과, 분양, 기대, 증가, 물량	17	
	4q	상승, 내년, 재건축, 상승세, 전세값, 증가, 금리, 월세, 대책, 수요, 하락, 공급, 분양	46	90

을 확인할 수 있다.

2012년과 비교해서 규제완화측면의 단어 빈도가 많았으며 점점 심화되고 있는 전세난에 대한 언급이 많았음을 엿볼 수 있다.

‘건설사’의 시공실적 악화와 ‘미분양’ 적체로 인한 부담이 최고조에 달했던 시기에 관련 단어들이 대거 등장함을 알 수 있다. 부동산에 대해 전반적으로 ‘하락’, ‘침체’, ‘감소’와 같은 부정적인 의견이 대부분이었음을 알 수 있었다.

2013년 4분기부터 2014년은 전반적으로 ‘상승’과 관련된 단어가 분기별로 상위권에서 빈번하게 언급되고 있다. 거듭되는 주택시장 침체에서 벗어나기 위한 정부의 거래활성화 대책이 효과를 보인 것으로 판단된다.

주택가격이 2013년 5월부터 저점을 통과한 후 현재까지 상승국면에 있는 것으로 파악되면서 정부의 세제완화, 주택거래지원, 대출지원 등이 거래활성화에 긍정적으로 작용하므로써 정책의 중요성을 다시 한번 확인할 수 있다.¹⁶⁾

〈표 3〉은 빈도수가 높은 주요단어가 어떤 의미로 다른 단어와 연결되어 있는지를 결합 빈도수로 보여준다. 2012년은 ‘소형-아파트’, ‘금융위기-이후’, ‘수익성-부동산’, ‘충선-대전’, ‘강남-재건축’, ‘부동산-대책’, ‘다주택-양도세’, ‘취득세-감면’, ‘정부-출범’, ‘양도세-감면’, ‘미분양-아파트’ 순으로 단어 간 조합 빈도를 보여준다. 대부분이 정부의 대책에 대한 단어 조합으로서 시장가격 하락과 침체라는 당면한 문제를 활성화 대책을 통해 벗어나려는 의도가 담겨 있다. 주택가격과 전세가격 변화는 과거 외환위기 기간에 비해 글로벌 금융위기 기간에 더 많은 경제변수와 관련이 있다.¹⁷⁾ 장기 불안심리와 침체가 지속되고 경제 지표에 민감하게 반응하면서 부동산 가격의 불안정 심리가 가라앉지 않고 언론에서 자주 언급되고 있는 것으로 보여진다.

2013년은 ‘취득세-감면’과 ‘수직증축-허용’, ‘수도권-아파트’, ‘금리-인하’, ‘전세-월세’, ‘부동산-대책’순으로 나타났으며 취득세 영구인하와 다주택자 양도세 중과 폐지 그리고 아파트

〈표 3〉 주요 단어 N-gram 순열

	1분기	2분기	3분기	4분기
2012	소형-아파트(20)	부동산-대책(11)	다주택-양도세(17)	취득세-감면(17)
	금융위기-이후(18)	서울-수도권(8)	최저치-기록(7)	정부-출범(12)
	수익성-부동산(15)	집값-상승(8)	일반세율-적용(6)	양도세-감면(12)
	충선-대전(13)	투기지역-해제(6)	인구구조-변화(6)	미분양-아파트(11)
	강남-재건축(12)	시장-침체(6)	글로벌-경제(5)	수요-전환(10)
2013	취득세-감면(16)	취득세-감면(14)	강남-재건축(17)	집값-하락(22)
	중소형-아파트(8)	수직증축-허용(12)	취득세-인하(14)	수도권-아파트(18)
	양도세-중과(6)	1기-신도시(9)	6억원-이하(11)	금리-인하(18)
	분양가-상한제(5)	기대-어렵다(6)	9억원-이하(9)	전세-월세(18)
	생애-최초주택(4)	주택-거래량(6)	취득세-영구인하(9)	부동산-대책(17)
2014	강남-재건축(20)	가격-상승(11)	규제-완화(23)	집값-상승(25)
	리모델링-수직(12)	전셋값-상승(9)	재건축-연한(13)	규제-완화(15)
	초과이익-환수(9)	토지-보상금(9)	가격-상승(12)	전세값-상승(15)
	규제-완화(9)	선진-방안(8)	회복세-보이기(7)	월세-전환(14)
	연속-상승(8)	강남-재건축(7)	미분양-아파트(7)	상승-전망(11)

16) 노정휘·성주환, “주택시장의 경기변동과 인과관계에 관한 연구”, 부동산학보, 한국부동산학회, 2015, 제61집, p.216.

17) 김리영·서원석, “시기별 거시경제 환경변화에 따른 주택시장의 가격변동성 연구”, 부동산학보, 한국부동산학회, 2013, 제53집, p.111.

리모델링 수직증축허용, 전세값 고공행진이 이슈가 되었던 해이다.

2014년은 ‘강남-재건축’, ‘리모델링-수직’, ‘가격-상승’, ‘규제-완화’, ‘집값-상승’, ‘전세값-상승’, ‘월세-전환’, ‘상승-전망’ 순으로서 대부분 상승과 관련된 단어들이 등장하므로써 주택가격이 상승하면서 회복하고 있음을 유추해 볼 수 있다. 하지만, 전세값 상승과 임대인의 월세 전환 비중이 높아지면서 새로운 주택 문제가 되기도 하는데 금융위기 이후에는 매매가격 변화율과 전세가격의 변화율이 동반 상승하는 관계에 있으며¹⁸⁾ 특히, 전세가격과 CD금리는 금융위기 이후 주택매매가격에 영향 관계에 있다고 한다.¹⁹⁾ 고공행진하는 전세수요가 매매수요로 전환되지 못하고 있는 이유 중의 하나로 정부의 저

금리 정책을 들기도 한다.²⁰⁾

〈표 4〉에서는 해당 단어가 포함된 코퍼스 내에서 상위 10개 단어들의 중요도를 TF-IDF 가중치 값을 통해 살펴보았다. 단어의 가중치 값이 높을수록 단어가 포함된 문서내에서 핵심적인 메시지를 담고 있다는 의미이며 값이 낮을수록 많은 문서에서 다루고 있다는 의미이다.

2012년은 ‘총선대선’, ‘재건축’, ‘양도세’, ‘침체’, ‘폐지’, ‘대책’, ‘지방’, ‘공급’, ‘소형’, ‘중국미국’, ‘투자’, ‘취득세’에 대해서 가중치 값이 높게 나타났다.

4월의 총선과 연말 대선이 부동산 시장에 중요한 변수가 되므로 재건축 완화나 양도세 증가 폐지 및 취득세 인하 조치에 대한 언급과 중국과 미국의 대외경제 상황, 지방 미분양 주택에

〈표 4〉 주요 단어 TF-IDF 가중치

구분	2012				2013				2014			
	1q	2q	3q	4q	1q	2q	3q	4q	1q	2q	3q	4q
단어 TF-IDF	총선대선	양도세	폐지	지방	중소형	취득세	재건축	경매	재건축	수요자	재건축	인하
	35.95	38.56	33.53	24.62	10.19	36.07	21.71	32.26	16.93	9.84	26.73	26.34
	재건축	침체	대책	공급	양도세	리모델링	취득세	취득세	집값	재건축	회복세	전세
	19.53	23.78	29.41	17.26	7.05	23.07	18.44	18.12	9.73	8.99	13.82	23.80
	소형	중국미국	투자	취득세	취득세	재건축	9억원	하락	초과이익	상승	효과	월세
	13.97	23.42	28.13	14.43	6.75	19.61	15.07	15.96	9.40	8.91	11.56	23.51
	수익성	감소	거래량	전세	대책	분양	미분양	대출	전세	강남	상승	상승
	13.21	20.46	26.53	12.27	6.72	9.95	10.68	15.05	6.86	8.02	9.53	15.52
	거래량	강남	지속	증가	폐지	생애최초	건설사	월세	발표	신도시	투자	공급
	12.76	19.50	25.98	12.25	6.22	8.89	10.55	14.98	6.45	7.60	8.29	13.38
	회복	집값	양도세	대출	박근혜	감면	분양	공급	회복	증가	공급	분양
	12.33	19.13	20.82	8.17	4.70	6.94	9.17	14.98	2.27	6.56	8.07	11.11
	공급	거래량	취득세	하락	전세	공급	인하	물량	규제	과세	대출	대책
	12.11	18.00	16.03	7.97	4.22	6.49	8.76	9.84	2.50	5.42	5.14	10.24
	금융위기	상승	침체	상승	연장	상승	지방	상승	상승	하락	기대	금리
	9.69	9.97	6.65	6.11	4.03	5.97	7.46	8.84	1.94	5.11	4.92	10.18
	상승	대책	하락	침체	분양	대책	전세	전세	완화	분양	수요	수요
	8.75	9.73	6.59	5.80	4.03	4.36	7.14	8.22	1.59	4.53	3.73	8.01
	하락	투기지역	상승	수요	침체	침체	세율	수요	바닥	규제	완화	하락
	6.10	9.47	4.46	5.68	4.03	3.93	6.55	7.22	0.00	2.48	0.00	7.79

18) 성주환·오준석, “잠재성장모형을 이용한 금융위기 이전과 이후 서울 아파트 전세가격과 매매가격의 관계에 대한 연구”, 부동산학보, 한국부동산학회, 2013, 제54집, p.202.

19) 전해정, “글로벌 금융위기 전후로 거시경제변수와 부동산시장 간의 관계에 대한 연구”, 부동산학보, 한국부동산학회, 2014, 제58집, p.41.

20) 정인호, “주택전세 자기자금이 주택매매가격에 미치는 영향에 관한 연구”, 부동산학보, 한국부동산학회, 2014, 제57집, p.218.

대한 우려, 소형 주택 수요에 대한 변화를 기사에서 중요하게 다루었다고 볼 수 있다.

한편, 가중치 값이 상대적으로 낮은 ‘하락’, ‘침체’, ‘감소’와 같은 단어는 대부분의 기사에서 다루어졌다고 볼 수 있는데 당시 주택시장 전망에 대한 부정적인 의견이나 당시 상황을 이해할 수 있는 단어라고 할 수 있다.

2013년은 ‘취득세’, ‘재건축’, ‘양도세’, ‘경매’, ‘중소형’의 가중치 값이 높게 나타났는데 총선과 대선이 끝나고 부동산 시장에 긍정적으로 작용할 것이라는 기대감으로 정부에 정책적인 변화 요구와 발표 이후 시장에서의 변화를 기사에서 비중 있게 다뤘다고 볼 수 있다.

취득세 감면 정책은 아파트 가격보다 거래량과 관계가 있어 거래활성화에 긍정적인 요인으로 작용하는데, 당시 취득세 감면정책에 힘입어 주택거래량이 2006년 이래 최고의 거래량인 9만건을 기록하였다. 거래량 증가가 가격 상승을 동반하지는 않았지만 세금 정책이 시장에 얼마나 탄력적으로 작용하는지 확인하였던 시기이기도 하다²¹⁾

전환기 부동산시장에서 반복적으로 나타나는 경매시장도 이 시기에 가중치 값이 높게 나타났다. 특히, 경매 주택의 경우는 중소형으로 실수요자의 입찰이 늘어나 낙찰가율이 높아지는 시기였기도 하다.

이 시기에 가중치 값이 상대적으로 낮게 나온 ‘전세’, ‘상승’과 같은 단어를 통해 많은 사람들이 자주 언급하고 중요하게 다루었지만 재건축 완화나 세제 완화가 더 중요한 이슈였음을 추정해볼 수 있는데 2013년은 세제완화를 통해 시장 분위기 반전에 나섰다면 2014년은 주택시장을 이끄는 재건축 시장에 초점을 두고 있었다.

2014년은 ‘상승’, ‘회복세’, ‘재건축’, ‘전세’, ‘월세’가 상대적으로 가중치 값이 높게 나타났다. ‘상승’과 ‘회복’은 대체적으로 시장에 대한 긍정의 메시지를 담고 있음을 알 수 있다. 전월세 문제에 대한 가중치 값도 높은 것으로 나타나면서 서민

주거안정이라는 측면에서 주택가격의 상승과 전세가격의 상승이라는 양면의 문제를 안고 있기도 하다.

‘바닥’이나 ‘완화’는 가중치 값이 0으로 나오므로 모든 기사에서 다루었다고 볼 수 있고 상승국면에 있는 시기에 부동산 경기 바닥론이나 규제완화에 대한 내용이 자주 등장했다고 볼 수 있다.

V. 결론 및 시사점

2012년부터 2014년까지 단어빈도수와 N-gram 순열 그리고 TF-IDF 가중치를 통해 핵심 단어를 분류하여 시기별 주택시장 분위기를 유추해 볼 수 있었다.

2012년은 총선과 대선이 부동산정책에 변화를 주므로써 시장 회복을 기대하는 여론이 많았음을 유추할 수 있었으며 금융위기 이후에 미국과 중국의 시장회복을 국내 시장 회복에 대한 기대감으로 드러내기도 하였다.

과거 매수자들이 자본차익을 기대했던 중대형 평형은 가치하락이 컸던 만큼 거래량뿐만 아니라 가격 정체에서 좀처럼 벗어나지 못한 반면 이를 학습한 베이비부머 이후 세대들의 실수요 주택거래가 늘어나면서 소형평형과 수익형 부동산이 시장을 주도했다고 볼 수 있다.

2013년 이후로는 취득세 감면과 양도세 중과 폐지가 침체된 주택시장을 회복하는데 기여하고 매수심리도 살아나 반등의 계기가 되었음을 확인 할 수 있었다. 주택시장의 선도적 역할을 하는 재건축 규제완화는 직간접적으로 거래활성화에 기여하고 있다고 판단해 볼 수 있었으며 거래량이 늘어나고 가격이 상승하면서 분양시장에도 훈풍이 불었고 치솟는 전세가격으로 인해 실수요자들이 분양시장에 참여하는 등 부동산 대책이 주택시장에서 차지하는 역할 비중이 크다

21) 임대봉, “주택거래량과 주택가격에 관한 연구”, 부동산학보, 한국부동산학회, 2014, 제58집, p.207.

는 것을 알 수 있었다. 그러나, 부동산 대책이 주택시장 거래 정상화에 효과를 보인 반면 서민 주거 안정 측면에서는 많은 과제를 안고 있어 효과적인 대책이 시급하다고 볼 수 있다.

‘바닥’과 ‘상승’ 단어를 대부분의 기사에서 비중있게 다루면서 주택시장의 침체에서 벗어나 상승 분위기로 전환되고 있음을 분석을 통해 직관적으로 확인할 수 있었고

‘재건축’, ‘회복’, ‘전세’ 등 상위권으로 나타

나는 가중치 값은 해당 시기에 이슈가 되는 내용을 파악하는데 본 연구가 실증적으로 도움이 된다고 판단해본다.

본 연구는 텍스트마이닝 기법으로 대량의 뉴스기사를 통해 주택시장의 분위기를 가늠할 수 있는 핵심 단어와 단어의 중요도를 계량화하여 객관적인 정보를 찾아냈다는 점에서 일차적 의의가 있다. 부동산 관련 소셜미디어, 뉴스 등

비정형 데이터의 수집과 패턴을 실시간으로 모니터링하여 정책 자료로 활용할 수 있음을 제안해본다.

또한, 기존의 통계적 사고를 바탕으로 패턴 발견을 위한 마이닝적 사고를 접목시키므로써 텍스트 정보를 과학화하여 연구의 지평을 확장하였다는데 의미를 부여하고 싶다.

그러나 매체성향을 고려했을 때 언론사별로 구분하여 분석하였으면 좀 더 명확한 정보 분류가 되었을 것이라는 점과 단어의 가중치 값에 한정하여 해석하므로써 다양한 논의로 이어지지 못하였다는 한계가 있다.

향후 연구에서는 데이터마이닝의 다른기법들을 접목하여 언론사별, 전문기관별로 세분화하여 좀 더 입체적으로 패턴을 파악하는 것이 필요할 것이다.

參考文獻

- 김동성, 언어자료 분석을 위한 통계학, 한국외국어대학교 출판부, 2010.
- 김리영 · 서원석, “시기별 거시경제 환경변화에 따른 주택시장의 가격변동성 연구”, 부동산학보, 한국부동산학회, 2013, 제53집,
- 김완섭 · 이수원, “데이터마이닝을 이용한 설문조사의 심층분석”, 2006, 한국정보처리학회지, 한국정보처리학회, 제9권 제4호.
- 김정선 · 권은주 · 송태민, “빅데이터를 활용한 국내 빅데이터 수요공급 예측”, 한국경영정보학회지, 한국경영정보학회, 2014, Vol.2014-No1.
- 김종우 · 김선태, 경영을 위한 데이터마이닝, 한경사, 2009,
- 노정희 · 성주한, “주택시장의 경기변동과 인과관계에 관한 연구”, 부동산학보, 한국부동산학회, 2015, 제61집.
- 성주한 · 오준석, “잠재성장모형을 이용한 금융위기 이전과 이후 서울 아파트 전세가격과 매매가격의 관계에 대한 연구”, 부동산학보, 한국부동산학회, 2013, 제54집.
- 신현규 · 신홍철, “군집분석을 통한 자연휴양림 이용객의 시장세분화”, 한국콘텐츠학회 논문집, 한국콘텐츠학회, 2010, 제10집 제3호.
- 이수상, 네트워크 분석방법론, 논형, 2012.
- 유재술, “데이터마이닝기법을 활용한 주택담보대출 연체가능성 분석에 관한 연구”, 한성대학교 일반대학원 박사학위 논문, 2012.
- 이성직 · 김한준, “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출기법”, 한국전자거래학보, 한국전자거래학회, 2009, 제14권 제4호.
- 임대봉, “주택거래량과 주택가격에 관한 연구”, 부동산학보, 한국부동산학회, 2014, 제58집.

- 전해정, “글로벌 금융위기 전후로 거시경제변수와 부동산시장 간의 관계에 대한 연구”, 부동산학보, 한국부동산학회. 2014, 제58집.
- 정인호, “주택전세 자기자금이 주택매매가격에 미치는 영향에 관한 연구”, 부동산학보, 한국부동산학회. 2014, 제57집.
- 조태호·서정현, “문서의 키워드 추출에 대한 신경망 접근”, 한국정보과학회지, 한국정보과학회. 2000, 제27권 제2호.
- 조은경, “감성분석에서 함수적 어휘와 대하여성 추출”, 한국어의미학술지널, 한국어의미학회. 2015, 제47권.
- A.L. 바라바시. 링크, 21세기를 지배하는 네트워크과학, 동아시아, 2015.
- Fawcett·Provost, *Data science for business*. O'REILLY·Hanbit, 2014.