Lead Article

빅데이터를 이용한 경기판단지표 개발: 네이버 검색 경기지수 작성과 유용성 검토

이긍희* · 황상필**

빅데이터 시대가 개막되면서 검색로그, 소셜네트워크의 대화글, 블로그의 게시글 등 비정형 데이터를 이용하여 경제를 분석하기 시작했다. 이중 검색로그를 바탕으로 만들어지는 검색통계는 키워드 중심으로 생성되므로 소셜네트워크 또는 블로그의 텍스트 분석을 통해 얻어지는 데이터에 비해 오류가 적고 의도성도 낮다.

본 연구에서는 우리나라에서 점유율이 가장 높은 포털인 네이버에서 찾은 호황, 불황 관련 검색 데이터를 바탕으로 BSI 방식으로 네이버 검색 경기지수를 작성하였다. 또한 네이버 검색 경기지수의 유용성을 확인하기 위해서 네이버 검색 경기지수에 대해 교차상관분석, 전환점 분석, 예측력 분석을 실시하였다. 교차상관분석 결과 네이버 검색 경기지수는 경제심리지수와 매우 밀접하고 경기에 2개월 선행하는 것으로 나타났다. 또한 2008년 글로벌 금융위기 기간을 포함한 기간에서 네이버 검색 경기지수를 포함한 예측모형이 임의보행모형과 AR(1)모형에 비해 대체로 우수하게 나타났다. 네이버 검색 경기지수는 주별로 작성되면서 속보성이 있으므로 밀접한 관련성이 있는 경제심리지수를 보완하여 경기판단에 도움을 줄 것으로 판단된다.

JEL Classification Number: C1, C8, E3

핵심 주제어: 검색데이터, 경기지표, 경기실사지수, 경제심리지수, 예측력

* 교신저자, 한국방송통신대학교 정보통계학과 교수(E-mail : geunghee@knou.ac.kr, 02-3668-4695)
** 한국은행 조사국 계량모형부 팀장(E-mail : hwangsp7@bok.or.kr, 02-759-4200)

이 논문은 한국은행의 재정지원을 받아 작성된 것임. 논문 작성에 유익한 의견을 주신 심사위원, 경제분석 편집위원과 조사국 계량모형부 김준한 부장, 김용복 박사께 감사드립니다.
논문 투고일: 2014.9.30, 논문 수정일: 2014.11.22, 게재 확정일: 2014.12.16.

Ⅰ. 연구배경

인터넷의 확산, 스마트 기기 확산, 데이터 관리·분석 기술의 발전 등으로 다양한 형태의 데이터가 광범위하게 수집, 축적되면서 데이터로부터 인사이트를 창출할 수 있는 빅데이터 시대가 개막되었다. 빅데이터는 기존의 방식으로는 저장, 관리, 분석이 어려울 정도로 규모가 크고, 순환 속도가 빠르며, 숫자, 텍스트, 이미지, 음성 및 영상 등 형식이 다양한 데이터로 기존의 방식으로는 관리와 분석이 매우 어려운 데이터를 의미한다 (이긍희 외 2014).[1] 빅데이터의 예로는 네이버, 구글 등의 포털 검색통계, 트위터, 페이스북, 카카오톡 등의 소셜미디어 데이터, 유전자 정보, 위치정보, 행정데이터, 카드거래정보와 통화정보 등이 있다.

빅데이터 시대에서 경제주체들은 다양한 온라인 활동을 하고 있으며 활동의 흔적은 포털의 로그, 소셜네트워크의 대화글 또는 블로그의 게시글 등으로 저장된다. 이렇게 저장된 온라인 활동의 흔적을 바탕으로 축적된 빅데이터를 바탕으로 다양한 통계 분석이 실시되고 있다. 일반적으로 다수의 사람들이 생각하는 것이 현실로 실현될 가능성이 높다. 이러한 사람들의 생각 또는 관심은 소셜네트워크 상에 쓰여 져서 유통되는 글의 키워드와 포털로부터 검색되는 다수의 키워드를 통해 짐작할 수 있다. 따라서 이를 집계하면 그 결과로부터 키워드 관련 사건들이 현실에 실현될 가능성을 추측할 수 있다.

빅데이터는 통신사, 카드사 등 기업은 물론 정부 및 공공기관을 통해 활발히 연구되어 실용적인 결과를 창출하고 있다. 정부 및 공공기관은 자체 보유 데이터를 공개하여 빅데이터 산업을 육성하고 있고, 카드사는 카드 거래 데이터를 이용하여 카드 이용자들의 소비행태를 분석하여 신상품 개발 및 시장분석에 활용하거나 특이한 결제패턴을 검출하여 카드의 부정사용을 예방하고 있다. 통신사는 축적된 통신 내역 데이터를 바탕으로 고객들의 라이프 스타일 유형에 따른 상권 잠재력과 유망업종 분석하고 있다. 2013년 서울시는 심야 통신데이터와 택시의 심야 승하차 데이터를 바탕으로 심야버스노선을 선정하였다. 구글은 검색

1) 협의적 의미의 빅데이터는 규모(Volume), 다양성(Variety), 속도(Velocity)의 3V로 정의된다. 광의적 의미의 빅데이터는 협의적 정의의 빅데이터에 이를 관리·분석하기 위해 필요한 인력과 조직 및 관련 기술까지 더하여 정의된다. 본 연구에서는 언급되는 빅데이터는 협의적 의미의 정의이다.

을 이용하여 독감 예측하였고, 각종 선거에서 소셜네트워크 데이터를 이용하여 민심을 파악하고 있다.

빅데이터를 이용한 경제분석은 아직 초기 단계이다. 빅데이터 경제분석은 분석 대상 데이터와 분석방법이 기존 분석과 다른 점이 있다. 분석 대상 데이터로는 표본조사 등으로 작성된 경제통계보다는 구조화되지 않은 구글 검색 데이터, 트위터 데이터, 거래 데이터 등 경제주체의 온라인 활동 결과로 축적된 데이터가 이용된다. 분석방법은 기존의 인과관계 중심의 계량모형보다는 상관관계 분석, 기계학습 분석 방법이 이용된다(Einav and Levin, 2013).

빅데이터를 이용한 경제분석의 예로는 구글 검색통계를 이용한 경제분석, 트위터 게시글을 이용한 경제분석 등이 있다. Choi and Varian (2012)은 구글 검색통계를 이용한 모형을 바탕으로 미국 소매판매, 자동차 판매 등을 예측하였고, Chamberlin (2010)은 구글 검색통계와 소매판매 데이터 간 관계를 바탕으로 구글 검색통계가 영국 소비 패턴과 밀접하게 움직이는 것을 밝혔다. Zimmermann (2009)은 구글 검색통계가 독일 실업 데이터와 밀접하게 움직이고 있음을 밝힌 후 구글 검색통계를 이용하여 독일 실업 데이터를 예측하였고, D'Amuri (2009)는 구글 검색통계를 이용하여 이탈리아 실업률 예측하였다. Suhoy (2009)는 구글 검색통계를 이용하여 이스라엘 2008년 경기하강을 예측하였고, McLaren (2011)은 구글 검색통계를 이용하여 영국 노동 및 부동산 시장을 분석하고 구글 검색통계의 유용성을 평가하였다. 김지은 (2013)은 구글 검색통계와 우리나라 실물지표 간 대응성 분석과 예측력 분석을 통해 구글 검색통계의 유용성을 점검하였다. 소셜네트워크 데이터를 이용한 분석으로는 트위터(twitter) 데이터가 주로 이용되고 있다. UN 글로벌 펄스는 트위터 게시글의 감성분석 결과와 실업률 변화를 비교하였고, 트위터 게시글을 분석하여 글로벌 경제 위기 이후 미국과 인도네시아 사람들의 행태를 분석하였다(김정미, 2012). 한편 Chetty et al. (2011), Finkelstein et al. (2012)과 Einav et al. (2013)는 공공 데이터 및 거래 데이터를 이용하여 경제분석을 실시하였다.

우리나라 경제분석 및 경제예측에서 빅데이터 활용은 2014년 현재 미미한 수준이나 향후 데이터의 축적과 공개가 확산되면서 그 활용도는 크게 높아질 가능성이 높다. 특히 빅데이터 중 검색통계 및 소셜네트워크 텍스트 분석 데이터는 속보성이 높아 이들을 통해 경기를 빠르게 판단할 수 있다.

본 논문에서는 속보성이 높은 검색통계를 이용하여 경기판단에 도움을 줄 수 있는 지수를 작성하고 그 유용성을 점검하고자 한다. 본 논문의 구성은 다음과 같다. 제Ⅱ장에서 검색과 관련한 인터넷 이용현황과 포털 점유율 현황을 살펴보고, 제Ⅲ장에서는 네이버 트렌드를 바탕으로 경기 관련 검색 통계를 발굴하고 검토한다. 제Ⅳ장에서는 네이버 검색 경기지수를 작성하는 방안을 마련하고, 제Ⅴ장에서 네이버 검색 경기지수의 유용성을 점검한다. 마지막으로 제Ⅵ장에서 연구내용을 요약하고 향후 과제를 정리한다.

Ⅱ. 인터넷 사용 현황과 검색 데이터

인터넷 검색 데이터를 체계적으로 분석하기 위해서는 인터넷 사용 현황과 포털 점유율 관련 통계 등을 파악할 필요가 있다.

1. 우리나라 인터넷 사용 현황

2013년 인터넷이용실태조사에 따르면 2013년 7월 기준으로 보면 만 3세 이상의 인터넷 이용자수는 4,008만 명으로 전체의 82.1%를 차지하고 있다(한국인터넷진흥원, 2013).[2] 한편 스마트폰의 보급이 빠르게 확대되면서 2013년 스마트폰 보유 가구 비율이 79.7%로 빠르게 증가하여 장소에 구분 없이 인터넷을 사용하는 비율도 2012년 58.3%에서 91.0%로 대폭 높아졌다. 인터넷 이용의 주목적은 자료 및 정보 획득으로 그 비중이 91.3%이다. 이와 같이 경제활동과 밀접한 세대들의 인터넷 사용률이 높고, 인터넷 관련 주요 활동이 검색이므로 검색 결과를 정리·요약한 데이터는 경제활동과 밀접하게 움직인다.

검색은 주요 포털의 검색 엔진을 통해 이루어지므로 포털의 점유비율을 우선 살펴보아야 한다. <Table 1>은 2010년 이후 우리나라 포털의 점유비율을 정리한 표인데 이를 보면 네이버의 점유율이 71.5%, 다음의 점유율은 19.3%, 구글의 점유율은 4.7%로 나타났다. Return on Now (2014)의 31개국 2013년 국가별 검색 포털 점유비율을 살펴보면 우리나라, 중국, 일본, 러시아를 제외하면 구글의 검

2) 연령별로 보면 10대, 20대 30대는 100%에 가깝게 인터넷을 사용하고 있으며, 40대는 96.8%, 50대는 80.3% 사용하고 있다. 따라서 향후 인터넷 이용률은 빠르게 높아질 것으로 보인다.

〈Table 1〉 Search Engine Market Share in Korea

Search Engine	Average in period	Maximum	Minimum
Naver	71.7%	81.6%	58.5%
Daum	19.2%	27.7%	13.8%
Google	4.6%	6.8%	1.7%

Notes: Based on data from 2010.1.1. to 2014.6.30. Average in period is on a daily basis.
Source: http://trend.logger.co.kr/trendForward.tsp

색시장 점유비율이 가장 높게 나타났다.[3)]

2. 검색 통계의 특성

온라인 활동에 얻을 수 있는 데이터로는 크게 소셜네트워크 텍스트와 검색 로그로 구분된다. 소셜네트워크 텍스트로는 대표적으로 트위터, 페이스북, 블로그, 카카오톡, 라인 등을 통해 사용자간 주고받거나 게시하는 텍스트가 있다. 이러한 텍스트는 그대로 분석되기 보다는 텍스트마이닝 기법을 적용하여 주요 단어를 추출·집계하게 된다. 개별 연구자가 소셜네트워크 텍스트로부터 주요 키워드를 수집하기 위해서는 서버 구축, 수집 기술 확보 등의 비용이 소요된다. 일부 업체에서는 소셜네트워크 상의 여러 종류 텍스트를 수집하여 데이터베이스로 구축하고 유료로 주요 키워드 시계열을 제공하기도 한다.

검색 통계는 사용자가 검색 포털에 검색어를 입력하여 주요 정보를 찾을 때 검색 질의어별 검색수를 요약하여 정리한 것이다.[4)] 검색 통계로 분석 기간 중 검색어의 변화를 상대적으로 파악할 수 있고 검색 통계의 이용에 비용이 소요되지 않는다. 검색통계의 바탕인 검색어는 키워드 중심이므로 비정형인 소셜네트워크 데이터를 바탕으로 이루어진 텍스트 분석보다 상대적으로 오류가 적고 의도성이 낮아 검색통계는 소셜네트워크 데이터보다 객관적이라고 평가받고 있다.

3) 다른 나라의 경우 구글의 검색 점유율이 매우 높으므로 검색데이터 관련 경제분석에서 구글 검색데이터를 이용하고 있다. 참고로 구글의 점유율이 31개국 중 22개국이 90%이상이었다.
4) 포털에서 검색 엔진을 통해 검색이 이루어지는데 검색 엔진은 웹 크롤러 등을 이용하여 문서를 수집하고, 색인 검색의 과정을 거쳐 검색어와 가장 잘 맞는 문서를 보여준다.

3. 검색통계 작성방법

검색포털에서는 검색 통계를 트렌드라는 사이트를 통해 제공하고 있다. 2014년 9월 현재 제공되고 있는 검색 통계 서비스로는 네이버 트렌드, 구글 트렌드가 있다. 본 연구에서는 국내 포털 점유율이 가장 높은 네이버 검색 통계를 이용하였다.[5)]

가. 네이버 트렌드

네이버는 통합검색에서 발생하는 검색어 통계를 볼 수 있는 서비스로 beta 버전을 제공하고 있다(http://trend.naver.com). 이 사이트에서는 검색기기를 PC와 모바일로 구분하여 검색통계를 제공하고 있다. PC 검색통계는 2007년 1월부터 현재까지, 모바일 검색통계는 2010년 7월부터 현재까지 주간으로 작성되고 있다.

네이버 트렌드의 검색통계는 사람들이 검색한 검색 수를 단순히 합하여 발표하지 않고 상댓값으로 표현되고 있다. 네이버 트렌드의 검색통계 관련사항은 네이버 트렌드 도움말에 나타나 있는데 이를 정리하면 다음과 같다. 첫째, 그래프로 나타난 검색통계는 총 검색횟수를 0~100 숫자로 환산하여 표시되고 있다. 특정 검색 키워드가 검색에서 가장 많이 검색된 주간 평균값을 100으로 두고 나머지 기간의 평균 검색횟수를 상댓값(정수)으로 환산하여 그래프로 표현하고 있다. 또한 검색통계값은 최댓값을 기준으로 작성되기 때문에 별도로 검색된 검색통계 간 값들을 서로 비교할 수 없다. 검색통계값 0은 검색이 없었다기보다는 최대 검색량에 비해 검색량이 충분하지 않음을 주로 의미한다. 검색 키워드는 최대 5개까지 비교할 수 있으며 검색 키워드 내 띄어쓰기한 공백 제거와 대문자·소문자 변환은 자동으로 이뤄진다. 그러나 검색 키워드에 맞춤법 오류가 있거나 다르게 표기한 경우 다른 검색어로 처리되고 있다.

나. 구글 트렌드

구글 트렌드(www.google.com/trends)는 검색통계를 국가, 도시, 언어에 따라 구분하여 제공하는 사이트이다. 구글 트렌드 검색통계는 2004년부터 구할 수 있

5) 다음(daum)은 2013년 12월 검색 통계의 공표를 중단하였고, 구글은 국내 포털 점유율이 낮아서 구글의 검색 통계를 이용하는 데에는 제약이 있다.

으나 국가분류에서 한국을 지정할 수 있도록 한 시점은 2013년 12월부터이다. 검색통계는 주 단위로 작성되는데 최대 검색량을 100으로 하여 상대적 값으로 표준화되어 작성된다. 발표되는 검색통계는 키워드의 모든 검색 결과를 이용하지 않고 그 중 일부를 표본추출하여 작성된다. 키워드 검색 결과는 일별로 수집되는데 한 사람이 당일에 반복되어 검색하는 경우 하나의 검색으로 처리된다.

구글 트렌드 중 가장 유명한 사이트가 구글 독감 트렌드(www.google.org/flutrends)인데 여기에서는 독감 관련 구글 검색 통계를 사용하여 전 세계 독감 유행 수준을 실시간으로 예측하고 있다.[6)]

Ⅲ. 경기 관련 검색 데이터의 검토

경제주체는 경제활동과 관련한 다양한 검색을 하고 있으며 경제활동과 밀접한 것으로 나타났다. 경기판단과 관련해서는 호황, 불황 관련 중심으로 검색어 목록을 만들고 호황, 불황과 관련성 높은 검색어를 찾았다.

1. 경기 관련 검색통계의 선정

경기 관련 용어를 불황 관련 단어, 호황 관련 단어, 경제단어로 나누고 이들 검색어가 호황과 불황의 상관관계를 살펴보았다(<Table 2>).[7)]

네이버 트렌드를 이용한 주요 호황과 불황 관련 검색통계의 추이는 <Figure 1>과 같다. 불황 관련 검색어는 글로벌 금융위기 등을 반영하여 경제통계와 밀접하게 움직이는 것으로 나타났지만 호황 관련 검색통계는 경기와 관련성이 크지 않게 나타났다. <Figure 2>는 구글 트렌드를 이용한 불황 및 호황 관련 검색통계 추이인데 이를 보면 구글 트렌드에는 불황 및 호황 관련 검색통계가 충분하지 않은 것으로 나타났다.

6) Ginsberg et. al. (2009)은 독감 관련 주제를 검색하는 사람의 수와 독감 증상이 있는 사람 수 간에 관계가 밀접함을 밝혔다.

7) 경제 단어로는 이외도 여러 단어가 있으나 지나치게 단기적 변동이 큰 단어는 제외하였다. 특히 임금동결, 실업, 구직과 고용 관련 단어도 추가적으로 포함했으나 계절성 등이 크거나 실제 불황과 밀접하지 않은 것으로 나타나 비교에서는 제외하였다.

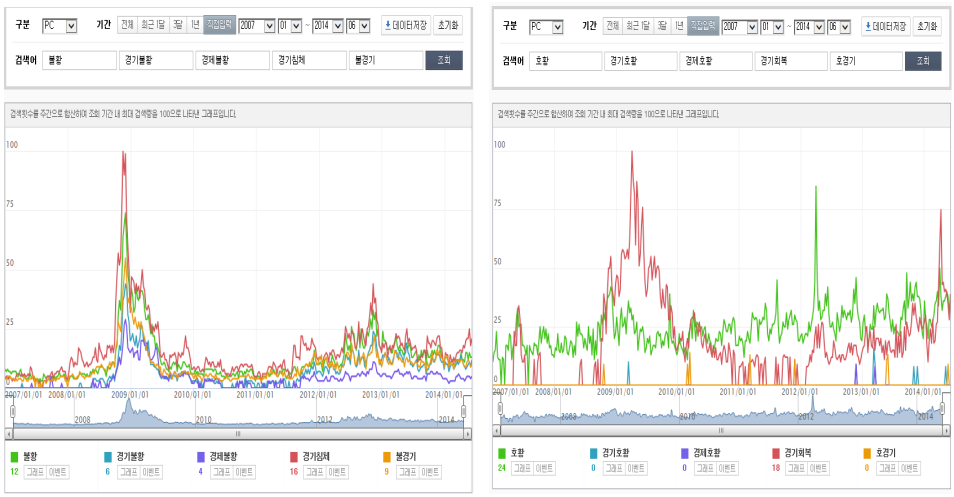
〈Table 2〉 Search Terms Related to Business Cycles

Recession terms	Boom terms	Economic terms
금융위기	호황	경기
부도	흑자	경제
불황	채용	금리
신용불량	폭등	돈
외환위기	상한가	물가
정리해고	거품	환율
경기침체	경기호황	코스피
파산	경제호황	주가
폭락	경기회복	기름값
하한가	호경기	TV
해고	인플레이션	일자리
불경기		주식시장
고유가		통화
디플레이션		재정
체납		GDP
집값하락		경상수지
저성장		경제성장
경기불황		고금리
경제불황		저금리
		양적완화

Note: Search terms are written in Korean.

〈Figure 1〉 Internet Search Data Using Naver Trends

(a) Recession terms (b) Boom terms



Note: Naver search data are available on trend.naver.com.

〈Figure 2〉 Internet Search Data Using Google Trends

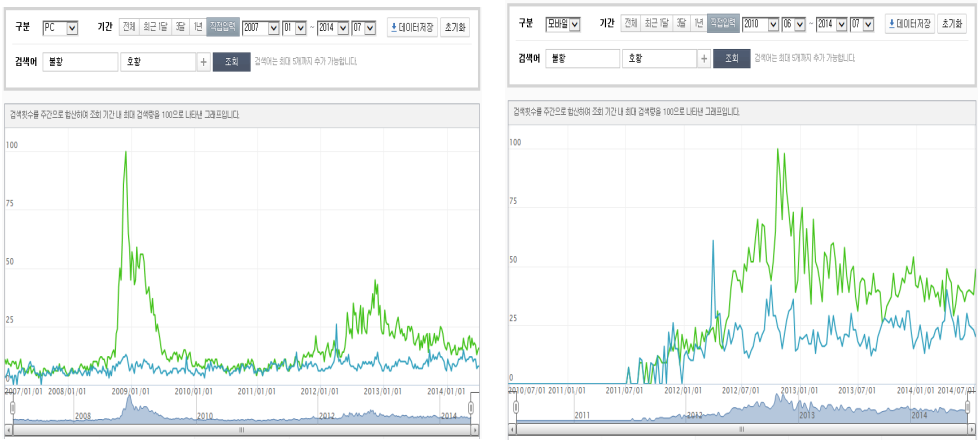
(a) Recession terms (b) Boom terms

Notes: Google search data are available on www.google.com/trends. Data for the term 'boom' cannot be obtained due to insufficient search volumes.

네이버 트렌드의 경우 검색기기를 PC와 모바일로 구분해서 검색통계를 제공하고 있다. <Figure 3>을 보면 PC의 검색통계와 달리 모바일의 검색통계를 보면 모바일 기기 확산에 따라 2011년 이후부터 검색어의 종류와 관련 없이 검색통계가 증가하는 모습을 보이고 있다.

〈Figure 3〉 Internet Search Data Using Naver Trends by Devices

(a) PC (b) Mobile



Note: The Naver Trends shows different search data for each devices.

3. 경기 관련 검색통계

‘불황’과 ‘호황’ 검색어와 다양한 경기 관련 검색어들 간 상관관계를 PC, 모바일로 나누어 구했다. <Table 3>은 경기 관련 네이버 PC 검색통계와 불황과 호황 검색통계 간 상관계수를 구한 결과이다. 이를 보면 ‘경기침체’, ‘불경기’, ‘경기불황’, ‘경제불황’의 검색통계가 ‘불황’ 검색통계와 매우 밀접하게 움직이는 것으로 나타났다. ‘호황’ 검색통계와 밀접하게 움직이는 검색통계는 찾기 어렵다.[8)]

<Table 4>는 경기 관련 모바일 검색통계와 ‘불황’과 ‘호황’ 검색어간 상관계수를 구한 결과이다. 분석 기간을 전 기간으로 확대해 보면 모바일의 확산이라는 요인으로 인해 검색통계 간 실제로는 유의한 관계가 없는데 유의한 관계가 있는 것으로 나타난다. 따라서 모바일 검색통계의 경우 상관계수 산출 기간을 2012년 이후로 한정하였다. 모바일 검색통계의 경우 ‘불황’과 상관관계가 높은 단어로는 ‘경기침체’, ‘불경기’, ‘경기불황’, ‘경제불황’ 등이 있으며, ‘호황’과 상관관계가 높은 검색어는 없는 것으로 나타났다.

경기 판단과 관련된 검색통계 중 불황 관련 검색 단어가 실제 경기와 밀접하게 움직이는 것으로 판단된다. ‘불황’ 검색통계와 관련성이 높은 검색어 군을 불황 검색어군으로 정했다. <Table 2, 3>의 상관계수 산출 결과를 바탕으로 불황 관련 검색 단어군으로 ‘불황’, ‘불경기’, ‘경기불황’, ‘경제불황’, ‘경기침체’로 정했다. <Table 2, 3>의 상관계수 계산 결과를 보면 ‘호황’과 의미 있게 높은 상관관계를 가지는 검색단어군을 찾기 어렵다. 따라서 호황 검색어군을 불황 검색어군의 반대어인 ‘호황’, ‘호경기’, ‘경기호황’, ‘경제호황’과 ‘경기회복’으로 지정했다.[9)] 불황 관련 검색어군과 호황 관련 검색어군을 결합할 때 단순 합계를 통해 불황 검색어군과 호황 검색어군의 대푯값을 PC와 모바일로 나누어 각각 구했다.

---

8) 경기순환과 관련 있는 데이터는 호황, 불황과 관련 없이 일정수준의 상관관계를 보인다. 상관관계가 의미를 가지려면 상대적으로 매우 높아야 한다. 실제로는 상관계수 행렬을 구해서 다양한 조합을 구해보았으나 지면의 한계로 불황과 호황 중심으로 상관계수를 정리했다.

9) ‘경기회복’의 경우 경기 불황기에 검색되는 경우 상대적으로 많아서 호황 검색어라고 말하기 어렵다. 그러나 ‘경기회복’의 검색은 소비자동향조사 및 기업경기조사에서 전월에 비해 경제상황이 나아지는지 여부를 묻는 경우를 고려할 때 경기가 좋아지는 신호로 볼 수 있다. 한편 제4절에서 기업경기실사지수, 소비자동향지수 및 경제심리지수와의 상관계수를 구해보면 경기회복을 호황 검색어로 포함하는 경우가 포함하지 않는 경우에 비해 상관계수가 0.01～0.08 높게 나타나 ‘경기회복’ 검색어를 호황 검색어로 포함하였다.

〈Table 3〉 Correlation Coefficients of Naver Search Terms (PC)

Search terms	Recession	Boom	Search Terms	Recession	Boom
금융위기	0.51	0.27	물가	-0.14	0.15
인플레이션	0.11	0.02	환율	0.60	0.43
부도	0.34	0.03	코스피	-0.07	0.13
불황	1.00	0.43	주가	-0.20	-0.28
신용불량	0.04	-0.37	기름값	0.04	-0.07
외환위기	0.23	0.15	TV	-0.17	-0.54
정리해고	0.54	0.32	일자리	0.00	-0.47
경기침체	0.94	0.44	주식시장	0.23	-0.10
파산	0.15	-0.20	통화	0.15	0.58
폭락	0.04	0.12	재정	0.07	0.37
하한가	0.12	0.29	GDP	0.09	0.43
해고	0.32	0.16	경상수지	0.17	0.17
불경기	0.95	0.44	경제성장	0.08	0.37
고유가	-0.07	-0.08	저금리	0.37	0.36
디플레이션	0.22	0.14	고금리	0.19	-0.23
호황	0.43	1.00	저성장	0.25	0.43
흑자	0.25	0.45	양적완화	0.09	0.39
채용	0.06	-0.20	체납	0.06	0.42
폭등	0.11	0.22	집값하락	0.12	0.13
상한가	0.06	0.10	경기불황	0.94	0.54
거품	-0.12	-0.19	경제불황	0.91	0.41
경기	0.15	-0.14	경기호황	0.13	0.17
경제	0.27	-0.10	경제호황	0.10	0.10
금리	0.05	-0.04	경기회복	0.56	0.32
돈	-0.06	-0.31	호경기	0.01	0.06

Note: Numbers are correlation coefficients between data for each search term and data for the term ‘recession’ or the term ‘boom’ in Korean.

〈Table 4〉 Correlation Coefficients of Naver Search Terms (Mobile)

Search terms	Recession	Boom	Search terms	Recession	Boom
금융위기	0.07	0.04	물가	0.07	0.16
인플레이션	0.29	0.57	환율	0.08	0.11
부도	0.07	0.35	코스피	0.05	0.15
불황	1.00	0.16	주가	0.10	0.10
신용불량	0.13	0.09	기름값	-0.25	-0.12
외환위기	0.20	0.27	TV	0.04	0.01
정리해고	0.09	0.05	일자리	0.19	0.25
경기침체	0.76	0.33	주식시장	0.00	-0.08
파산	-0.01	0.03	통화	0.08	0.29
폭락	0.12	0.17	재정	0.13	0.48
하한가	0.22	0.25	GDP	0.12	0.37
해고	0.13	0.15	경상수지	-0.01	0.16
불경기	0.66	0.22	경제성장	0.27	0.54
고유가	-0.36	0.17	저금리	0.42	0.32
디플레이션	0.36	0.19	고금리	0.28	0.20
호황	0.16	1.00	저성장	0.25	0.25
흑자	0.09	0.15	양적완화	0.03	0.09
채용	0.04	0.34	체납	0.15	0.26
폭등	0.07	0.33	집값하락	0.31	0.16
상한가	0.16	0.20	경기불황	0.74	0.21
거품	0.05	0.15	경제불황	0.69	0.22
경기	0.01	0.04	경기호황	0.00	0.00
경제	0.53	0.29	경제호황	0.02	-0.02
금리	0.21	0.13	경기회복	0.28	0.22
돈	-0.02	-0.12	호경기	0.13	0.25

Note: Numbers are correlation coefficients between data for each search term and data for the term ‘recession’ or the term ‘boom’ in Korean.

4. 불황 검색통계와 호황 검색통계의 작성

모바일 검색의 활용성은 높아지고 있으나 시계열의 길이가 짧고 모바일 기기 확산과 검색의 확산이 혼재되어 데이터의 활용에 제약이 있다. 그러나 2013년 인터넷 이용실태조사를 보면 스마트기기 확산으로 모바일을 통한 인터넷 서비스 이용이 많아지는 점을 감안하여 관련 정보를 혼합하여 분석하였다.[10] <Figure 4>는 네이버 트렌드를 이용한 불황 검색어군 대푯값과 호황 검색어군 대푯값에 대한 PC와 모바일 검색 결과이다. 이를 보면 2012년 이후에는 유사한 패턴을 보임을 알 수 있다.

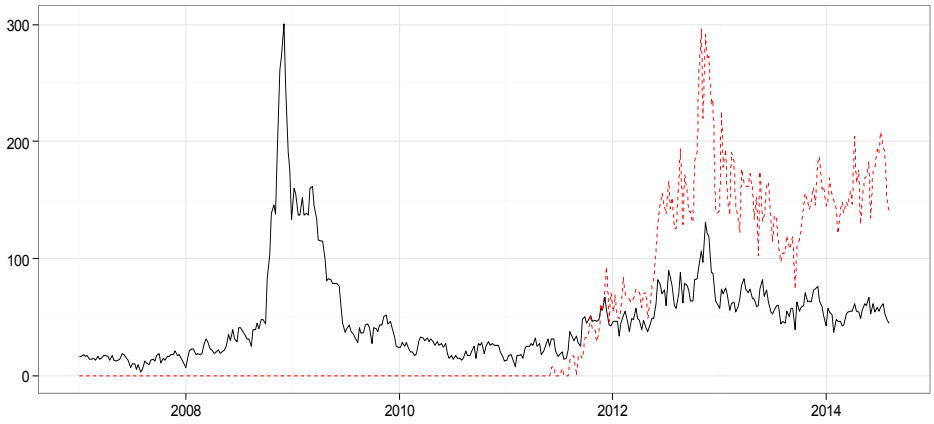
기기를 달리한 두 검색어군 통계를 결합하여 네이버 불황 검색통계와 호황 검색통계를 작성하는 과정을 정리하면 다음과 같다. 첫째, 불황 검색어군 5개를 네이버 트렌드에서 추출한 후 단순 합계하여 불황 검색통계를 작성했다. 둘째, 호황 검색어군 5개를 네이버 트렌드에서 추출한 후 합하여 호황 검색통계를 작성했다. 호황 검색어군 5개를 불황 검색어와 비교할 수 있도록 값들은 '불황' 검색어를 바탕으로 조정했다. 셋째, PC와 모바일에서 구해지는 호황과 불황 검색어군을 결합했다. 2012년 이후 PC와 모바일 검색통계의 상관관계를 보면 불황 검색통계의 경우 0.76, 호황 검색통계의 경우 0.60으로 나타났다. 구체적으로 살펴보면 PC 검색통계의 최댓값을 기준으로 모바일 검색통계를 조정한 후 모바일의 호황 검색통계와 불황 검색통계를 산출했다. 조정된 모바일 경기 관련 검색통계와 PC 경기 관련 검색통계를 같이 그려보면 <Figure 5>와 같은데 이를 보면 기기를 달리하더라도 동일한 검색어의 검색통계가 유사하게 움직이는 것으로 나타났다. 넷째, 2012년 이후 모바일 검색통계와 PC 검색통계를 평균하여 호황 및 불황 검색통계를 수정하였다.[11] 2012년 이전 기간의 경우 모바일 검색통계의 불안정성을 감안하여 PC 검색통계만으로 네이버 검색통계를 작성하였다. 호황과 불황과 관련된 조정된 모바일 검색통계와 PC 검색통계의 추이를 보면 <Figure 6>과 같다. 이를 보면 호황 관련 검색통계는 불황 관련 검색통계보다 적게 검색

10) 2013년 인터넷 이용실태조사를 보면 2013년 모바일 중심의 인스턴트 메신저 사용률은 82.7%로 2012년(60.1%)보다 크게 늘었고 2013년 인터넷뱅킹 이용자 중 모바일뱅킹 이용비율과 모바일쇼핑 이용비율도 각각 65.4%, 43.2%로 나타나 2012년(29.2%, 23.8%)보다 크게 늘었다.

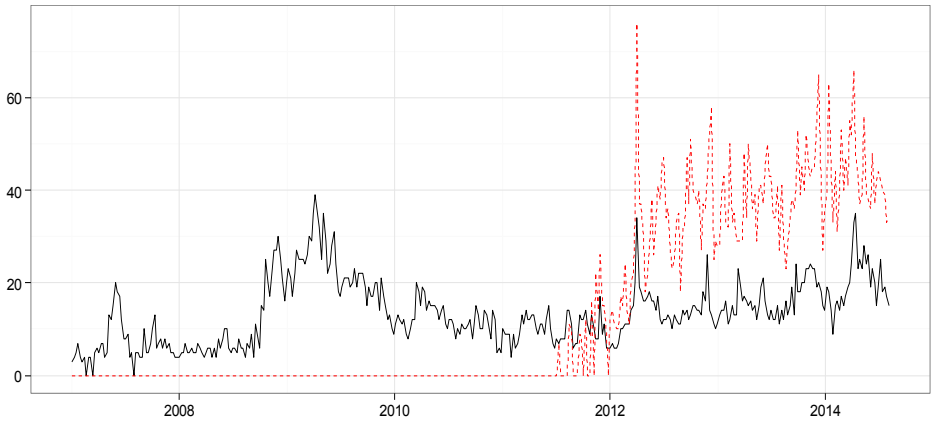
11) PC와 모바일 기기별로 검색량이 다르므로 PC와 모바일의 검색통계를 가중평균해서 결합해야 하지만 네이버 트렌드에서는 기기별 검색량을 알 수 없으므로 가중평균할 수 없다. 따라서 PC와 모바일의 검색통계를 패턴이 유지되도록 최댓값 기준으로 수준을 맞춘 후, PC와 모바일의 검색통계를 단순평균했다.

〈Figure 4〉 Naver Search Data by Devices

(a) Recession



(b) Boom

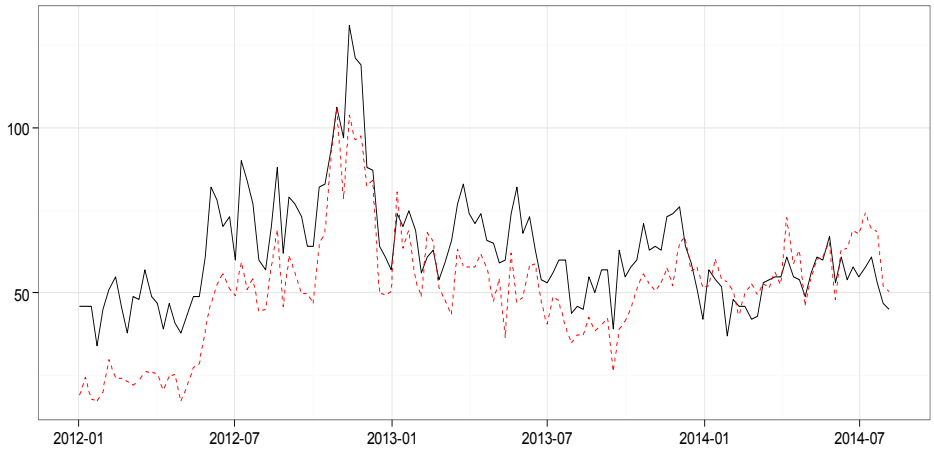


Notes : The solid line is the weekly Naver search data based on the PC. The dotted line is the weekly Naver search data based on mobile devices. Search terms are written in Korean.

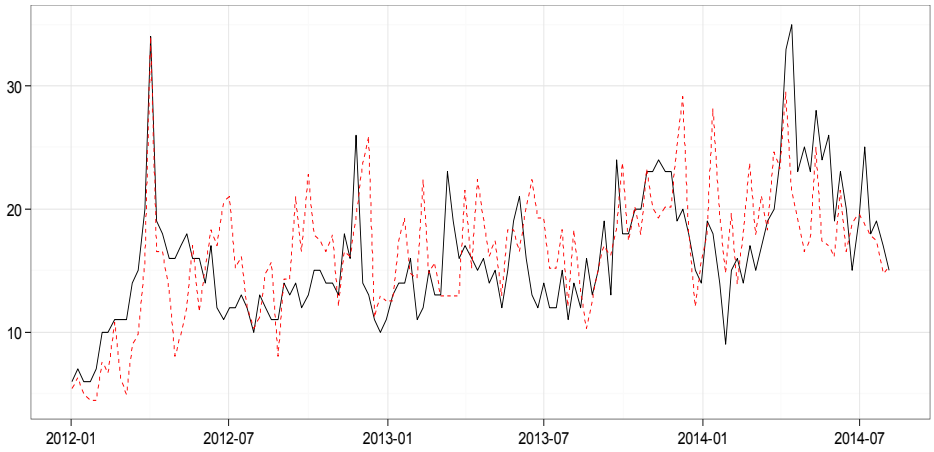
되었다. 2012년 이후 모바일과 PC 검색통계를 결합한 호황 및 불황 관련 검색통계는 PC 검색통계와 차이가 크지 않게 나타났다. <Figure 7>은 최종 불황 검색통계와 최종 호황 검색통계를 같이 그린 그래프이다. 이를 보면 불황 검색통계는 호황 검색통계보다 검색 빈도가 높고 변동이 큰 것을 알 수 있다.

〈Figure 5〉 Adjusted Naver Search Data by Devices

(a) Recession



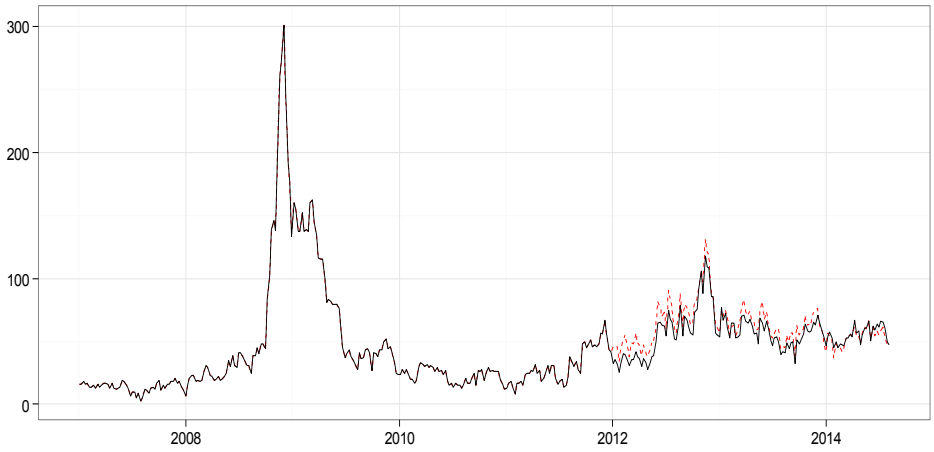
(b) Boom



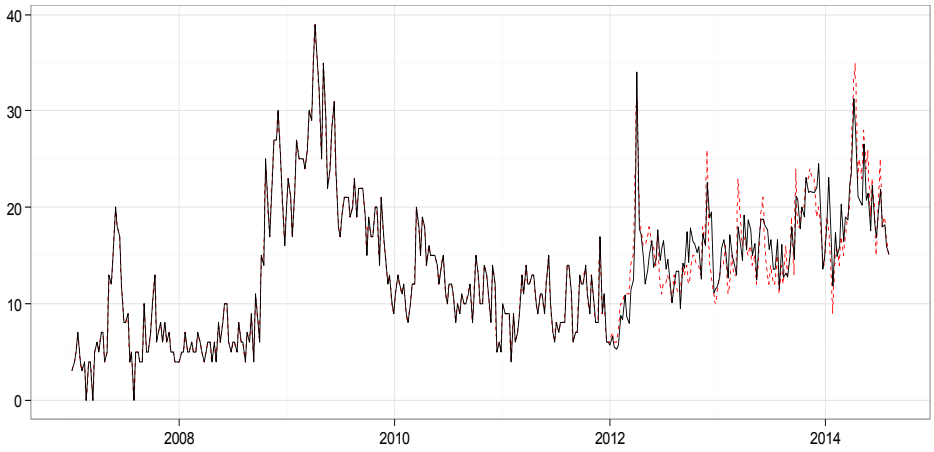
Notes: The solid line is the weekly Naver search data from PC. The dotted line is the weekly adjusted Naver search data from mobile devices. Search terms are written in Korean.

〈Figure 6〉 Composite Naver Search Data

(a) Recession

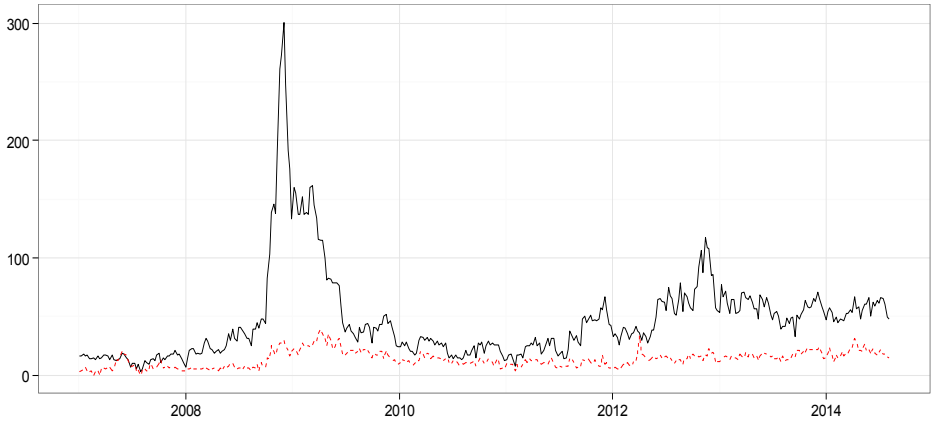


(b) Boom



Notes: The solid line is the weekly composite Naver search data made by combining the PC's Naver search data and the adjusted mobile's Naver search data. The dotted line is the PC's weekly Naver search data. Search terms are written in Korean.

〈Figure 7〉 Comparison of Composite Naver Search Data



Notes: The solid line is the weekly composite Naver search data for terms related to ‘recession’ in Korean. The dotted line is the weekly composite Naver search data for terms related to ‘boom’ in Korean.

Ⅳ. 네이버 검색 경기지수의 작성

1. 검색통계와 심리조사

검색은 일종의 자발적인 조사에 대한 응답이다. 어떤 사람이 ‘불황’을 검색한다면 그 사람 또는 주변 사람이 운영하거나 재직하고 있는 사업체가 업황이 나쁘거나 나빠질 가능성이 있다는 것을 의미한다. 즉 ‘불황’의 검색은 ‘운영하거나 재직하고 있는 사업체가 업황이 나쁘거나 나빠질 것으로 생각됩니까?’라고 질문에 검색하는 사람이 ‘예’라고 답한 것이다. 만약 수많은 사람이 ‘불황’에 대해 검색하고 이를 집계한다면 이는 일종의 불황에 대한 조사라고 할 수 있다. 그러나 검색은 표본조사와 같이 모집단을 대표하도록 설계되어 있지 않아서 모집단을 대표해서 구체적으로 파악하기 어렵지만 조사 응답률 하락 등 조사환경이 열악해지는 현 상황에서 표본조사를 보완할 수 있는 한 대안이 될 수 있다.

검색과 유사한 방식으로 경제주체의 판단을 물어보는 조사가 있다. 대표적인 조사로는 한국은행에서 조사하고 있는 기업경기조사와 소비자동향조사가 있다. 기업경기조사의 경우 전국 2,862개 법인기업(2014년 7월 기준)을 대상으로 전기

에 비해 업황, 재고, 인력 등의 사정이 나아졌는지 나빠졌는지 파악하는 조사이다. 조사된 결과를 기업경기실사지수(Business Survey Index : BSI)로 요약·정리하고 있는데 이 지수는 긍정 응답 업체 수의 비중에서 부정 응답 업체수의 비중을 차감하여 비율로 표현한 후 100을 더해서 작성된다. 소비자동향조사는 전국 도시 2,200 가구(2014년 7월 기준)를 대상으로 전기에 비해 생활형편, 가계수입 전망 등이 나아졌는지 파악하는 조사이다. 조사된 결과를 소비자동향지수(Consumer Survey Index : CSI)로 요약·정리하고 있다. 한편 경제심리지수(ESI : Economic Sentiment Index)는 기업과 소비자 모두를 포함한 민간의 경제상황에 대한 심리를 종합적으로 파악하기 위하여 BSI 및 CSI 지수(각각 40개 및 24개) 중 경기 대응성이 높은 7개 항목을 선정하여 이들의 표준화지수를 가중평균한 지수로 월별로 작성된다.

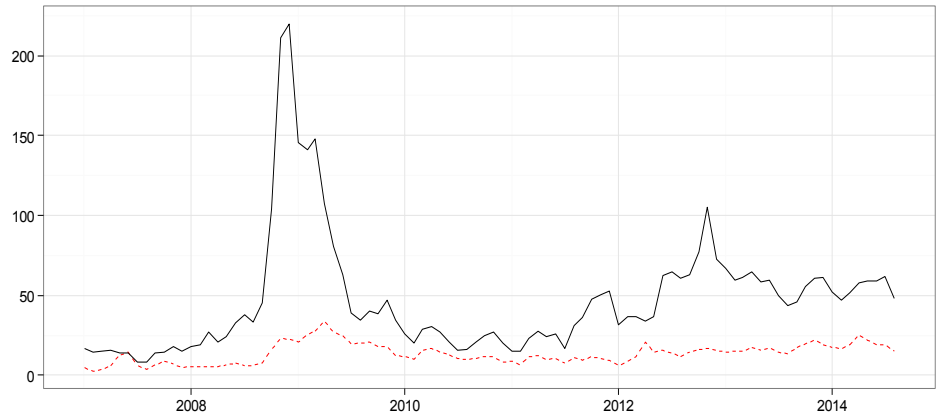
기업경기조사의 BSI와 마찬가지의 방식으로 네이버 검색 불황지수와 네이버 호황지수를 결합하여 BSI 방식의 네이버 검색 경기지수를 작성하였다. 네이버 검색 경기지수의 유용성을 파악하기 위해서 주별 네이버 검색 불황지수와 네이버 호황지수를 월별 지수로 전환하여 BSI, CSI, ESI와 비교할 필요가 있다.

2. 월별 네이버 검색 경기지수의 작성

경기 관련 네이버 호황, 불황 검색통계는 주별 데이터여서 발표되고 있는 경기관련 월별 경제통계와 직접 비교할 수 없다. 따라서 네이버 호황, 불황 검색통계를 월별 데이터로 전환하였다. 월별 네이버 호황, 불황 검색통계는 주별 네이버 호황, 불황 검색통계를 같은 값의 일별 데이터로 전환한 후 이를 월 단위로 재정리하여 생성하였다.[12] <Figure 8>은 월별 네이버 불황 검색통계와 호황 검색통계의 추이를 보여주고 있다.

12) 주별 데이터를 같은 값의 7일 데이터로 전환한 후 이를 해당하는 월 단위로 재정리한 후 평균하여 구했다. 월단위 결과와 주단위 결과가 같은 수준을 유지토록 하여 서로 비교가능하게 하였다.

〈Figure 8〉 Monthly Naver Search Data



Notes: The solid line is the monthly composite Naver search data for terms related to ‘recession’ in Korean. The dotted line is the monthly composite Naver search data for terms related to ‘boom’ in Korean.

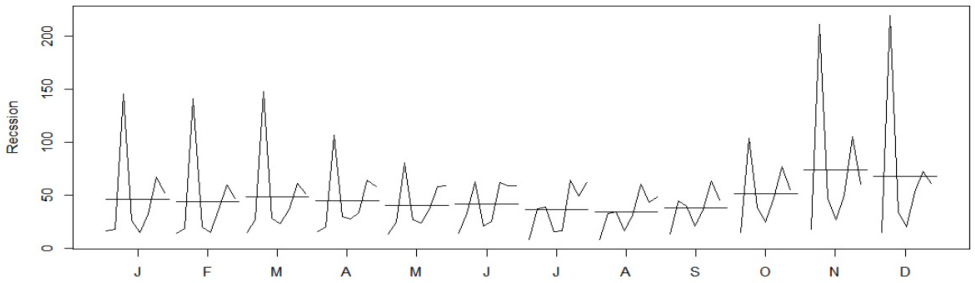
2. 월별 검색 데이터의 계절조정

월별 네이버 호황, 불황 검색통계는 BSI와 CSI와 마찬가지로 계절성을 가진다. <Figure 9>는 월별 네이버 호황, 불황 검색통계를 월별로 정리하여 다시 그린 그래프이다. 이를 보면 불황에 대한 검색은 평균적으로 11월, 12월에 많아지는 반면 호황에 대한 검색은 평균적으로 4월~6월에 많아지는 것으로 나타났다.

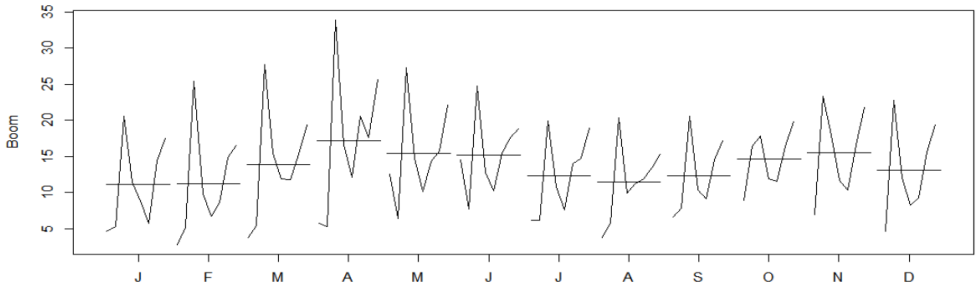
월별 네이버 경기 검색데이터의 계절성을 F검정, Kruskal-Wallis 검정 등으로 검정해보면 <Table 5>와 같은데 이를 보면 F검정, Kruskal-Wallis 검정 결과가 1% 유의수준에서 유의하게 나타나 불황, 호황 검색 데이터 모두 안정적 계절성을 가지는 것으로 나타났다. 이동계절성에 대한 F검정 결과를 살펴보면 호황 검색 데이터의 경우 이동 계절성이 유의하지 않으나 불황 검색 데이터는 5% 유의수준에서 이동 계절성이 유의하게 나타났다. 그러나 불황 검색 데이터의 이동계절성이 안정적 계절성보다 크지 않아서 호황 및 불황 검색 데이터 모두 식별 가능한 계절성이 있는 것으로 나타났다.

〈Figure 9〉 Seasonality of Monthly Naver Search Data

(a) Recession



(b) Boom



Note : Based on data from January 2007 to August 2014. X-axis denotes each month. Horizontal line denotes the average of each month data.

〈Table 5〉 Seasonality Test of Monthly Naver Search Indices of Business Cycles

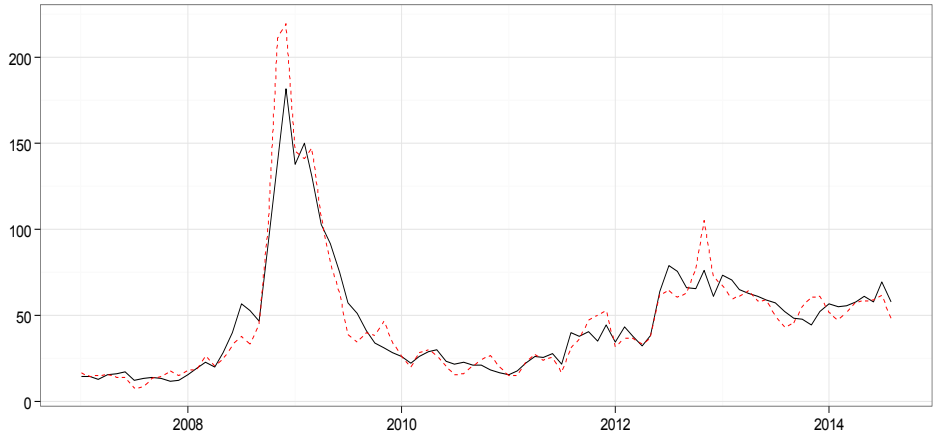
	Stationary seasonality		Moving seasonality	Identification of seasonality
	F test	Kruskal-Wallis test	F test	
Boom	8.326**	52.227**	0.461	Identified
Recession	14.405**	55.675**	3.495*	Identified

Notes : * and ** denote that statistics are significant at the 5% and 1% level, respectively.

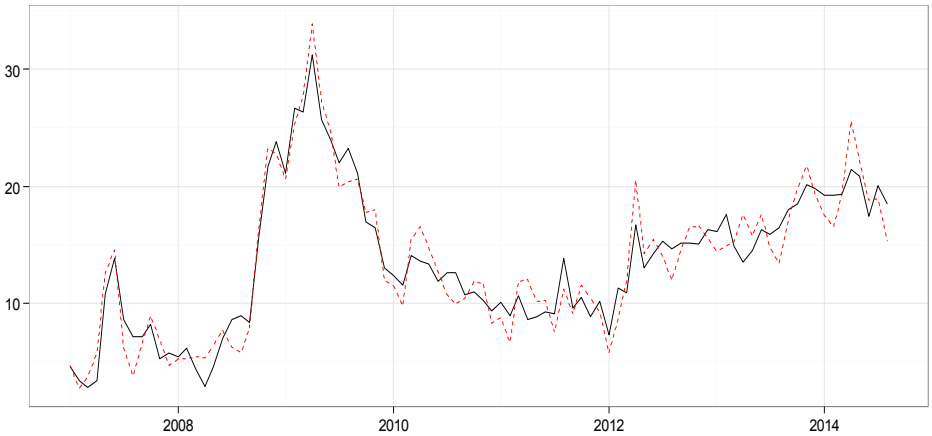
월별 네이버 호황 및 불황 검색데이터를 X-13ARIMA-SEATS의 X11필터를 이용하여 계절조정하면 <Figure 10>과 같다.

〈Figure 10〉 Seasonal Adjustment of the Monthly Naver Search Data

(a) Recession



(b) Boom



Notes: The solid line is the seasonally adjusted Naver search data. The dotted line is the original series.

3. 네이버 검색 경기지수의 작성

네이버 검색 경기지수(Naver search business index : NSI)는 월별 네이버 호황 및 불황 검색통계를 바탕으로 경기실사지수(BSI)와 같은 형태로 호황 검색통계의 평균에서 불황 검색통계의 평균을 차감한 후 100을 더한 지수를 작성하였다.[13)]

13) 검색어 5개를 단순평균하기 보다는 실물 지표와의 관련성과 주성분분석 등을 통해 가중평균할 수 있지만 본 논문에서는 초기 연구인 점을 감안해서 단순평균했다.

$$\text{네이버 검색 경기지수}(NSI) = \frac{(\text{호황검색데이터} - \text{불황검색데이터})}{5} + 100$$

네이버 검색 경기지수와 BSI, CSI, ESI간 상관계수를 구해보면 <Table 6>과 같다. 이를 보면 네이버 검색 (NSI)는 CSI보다 BSI간 상관관계가 높고, CSI와 BSI와 혼합된 ESI와 상관관계가 가장 높게 나타났다.

네이버 검색 경기지수의 계절조정은 호황 검색통계와 불황 검색통계를 각각 나누어서 계절조정한 후 결합하는 간접법으로 계절조정하였다.[14] <Figure 11>의 (a)는 월별 네이버 검색 경기지수(계절조정)를 ESI, BSI와 같이 보여주고 있는데 월별 네이버 검색 경기지수(계절조정)는 경제심리지수보다 평균적으로 작고 경기실사지수보다 크게 나타났다. 월별 네이버 검색 경기지수(계절조정)를 ESI의 표준편차와 최댓값을 감안하여 조정된 월별 네이버 검색 경기지수(계절조정)를 다음과 같이 구할 수 있는데 <Figure 11>의 (b)에서 보듯이 동 지수는 경제심리지수(ESI)와 매우 유사하게 움직인다.

〈Table 6〉 Correlation between Monthly Naver Search Business Index and Business Indicators

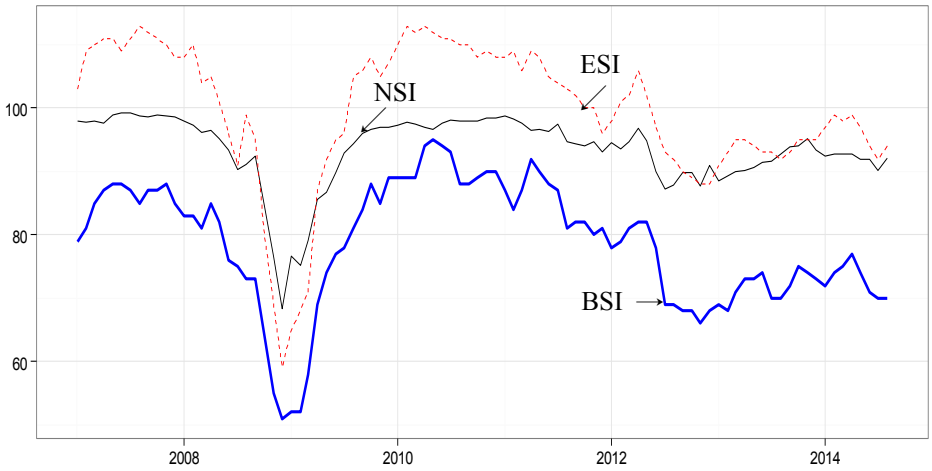
Business indicators	Naver search business index		Index average
	Seasonally adjusted series	Original series	
Business survey index (BSI, business condition)	0.907	0.853	78.7
Consumer survey index (CSI, business judgement)	0.615	0.597	75.8
Economic sentiment index (ESI)	0.960	0.925	99.5
Cyclical component of ESI	0.919	0.839	99.5

Notes: Based on data from January 2007 to August 2014. The Naver search business index is seasonally adjusted using the indirect method.

14) 직접법은 네이버 검색 경기지수를 산출한 후 동 지수를 직접 계절조정하는 방법이다. 간접법에 의해 계절조정된 네이버 검색 경기지수가 직접법에 의해 계절조정된 네이버 검색 경기지수에 비해 경기관련 지표와 밀접하게 나타나서 본 논문에서는 간접법에 의한 계절조정 네이버 검색 경기지수를 이용했다.

〈Figure 11〉 Comparison of Monthly Naver Search Business Index, ESI and BSI

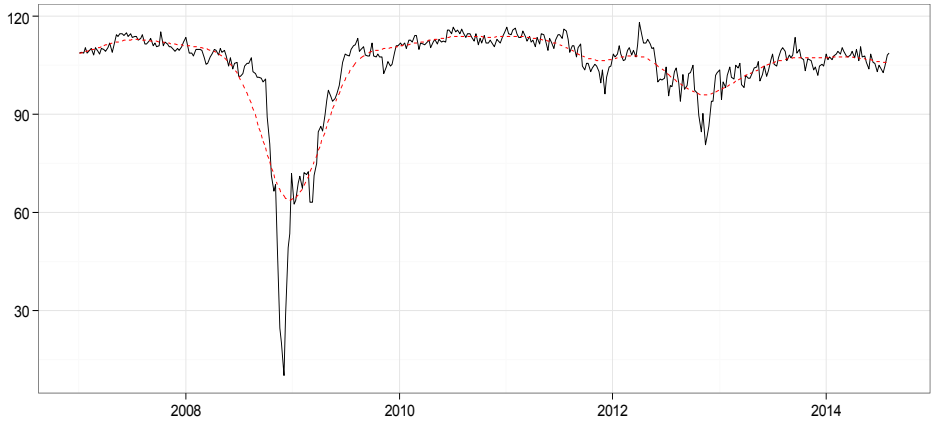
(a) Monthly Naver search business index



(b) Monthly adjusted Naver search business index

Notes: (a) The solid line is the monthly Naver search business index. The dotted line is the monthly ESI. The bold line is the monthly BSI.
(b) The solid line is the monthly adjusted Naver search business index. The dotted line is the monthly ESI.

〈Figure 12〉 Weekly Naver Search Business Index



Notes: The solid line is the weekly adjusted Naver search business index. The dotted line is the HP (Hodrick-Prescott) filter trend.

$$\text{조정된 } NSI = NSI \times \frac{\text{표준편차}(ESI)}{\text{표준편차}(NSI)} + (\text{최대값}(ESI) - \text{최대값}(NSI))$$

월별 네이버 검색 경기지수의 분석결과를 바탕으로 같은 방식으로 주별 네이버 검색지수를 작성했다. 주별 네이버 검색 경기지수는 주별 데이터이기 때문에 신호가 불분명하기 때문에 HP 필터를 이용하여 움직임의 추세를 살펴보았다. 또한 주별 네이버 검색 경기지수도 조정된 월별 지수와 같은 방식으로 ESI를 고려한 조정을 실시하였다. <Figure 12>는 조정된 월별 네이버 검색 경기지수와 이 지수에 HP필터 적용한 결과를 보여주고 있다. 네이버 검색 경기지수는 속보성이 있고 주별로 작성될 수 있어서 경제심리상태를 조기에 파악할 수 있다.

Ⅴ. 네이버 검색 경기지수의 유용성 점검

네이버 검색 경기지수의 유용성을 파악하기 위하여 교차상관분석, 전환점 분석을 실시하고 경제성장률과 민간소비증감률의 예측 가능성을 점검하였다. 교차상관분석에서는 교차상관계수의 최댓값을 바탕으로 시차구조를 점검하였고, 전

환점 분석은 경기 관련 계열의 정점과 저점을 구한 후 이를 비교하여 시차구조를 살펴보았다. 예측력 분석은 경제성장률과 민간소비증감률에 대해 네이버 검색 경기지수를 포함한 예측모형과 임의보행모형, AR(1)모형에 의한 예측 결과를 비교하였다.

1. 교차상관분석

네이버 검색 경기지수의 유용성을 점검하기 위해서 경제심리지수(ESI), 경기실사지수(BSI, 업황), 소비자동향지수(CSI, 경기판단), 경기동행지수 순환변동치(CCI Cycle), 경기선행지수 순환변동치(LCI Cycle), 실업률 간 교차 상관계수를 구했는데, 그 결과는 <Figure 13>과 같다. <Table 7>은 <Figure 13>에서 교차상관계수 최댓값(절댓값) 기준으로 정리한 결과이다. 이를 보면 계절조정 네이버 검색 경기지수는 경제심리지수와 동행하면서 상관계수도 0.96으로 매우 높아 동지표를 대체할 수 있을 것으로 보인다. 한편 계절조정 네이버 검색 경기지수는 경기동행지수 순환변동치와도 밀접하면서도 교차상관계수 최댓값 기준으로 2개월 선행하는 것으로 나타났다. 반면 계절조정 네이버 검색 경기지수는 경기선행지수 순환변동치에 3개월 후행하는 것으로 나타났다. 계절조정 네이버 검색 경기지수는 실업률에 13개월 선행하나 상관관계는 높지 않게 나타났다.

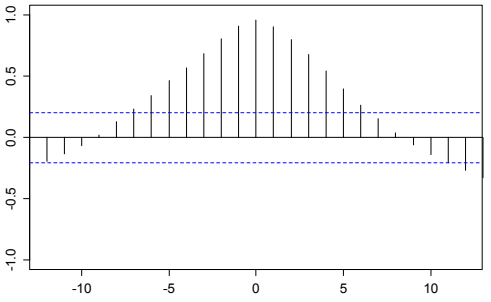
〈Table 7〉 Cross-correlation between Monthly Naver Search Business Index and Business Indicators

Business indicators	Cross-correlation with Naver search business index
Economic sentiment index (ESI)	0.960(0)
Business survey index (BSI, business condition)	0.907(0)
Consumer survey index (CSI, business judgement)	0.706(+5)
Cyclical component of coincident composite index	0.858(-2)
Cyclical component of leading composite index	0.749(+3)
Unemployment rate (seasonally adjusted)	-0.467(-13)

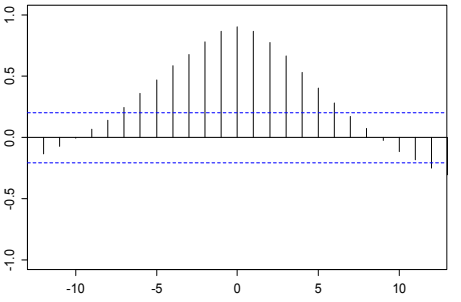
Notes: 1) Numbers in parentheses are time difference at the maximum of absolute coefficients.
2) + and - denote leading and lagging of Naver search business index, respectively.

〈Figure 13〉 Cross-correlation between Naver Search Business Index and Business Indicators

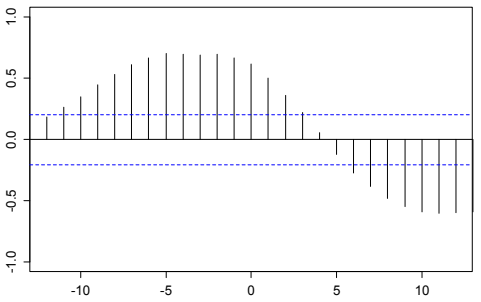
(a) ESI



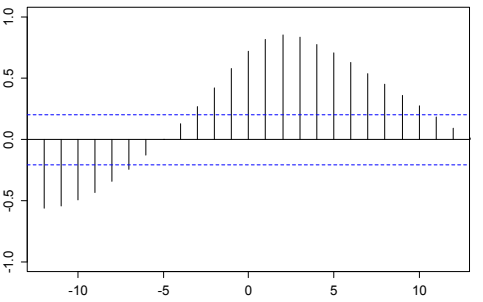
(b) BSI



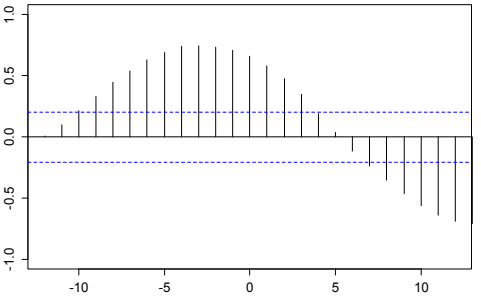
(c) CSI



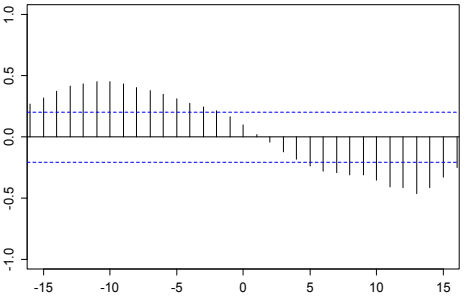
(d) CCI cycle



(e) LCI cycle

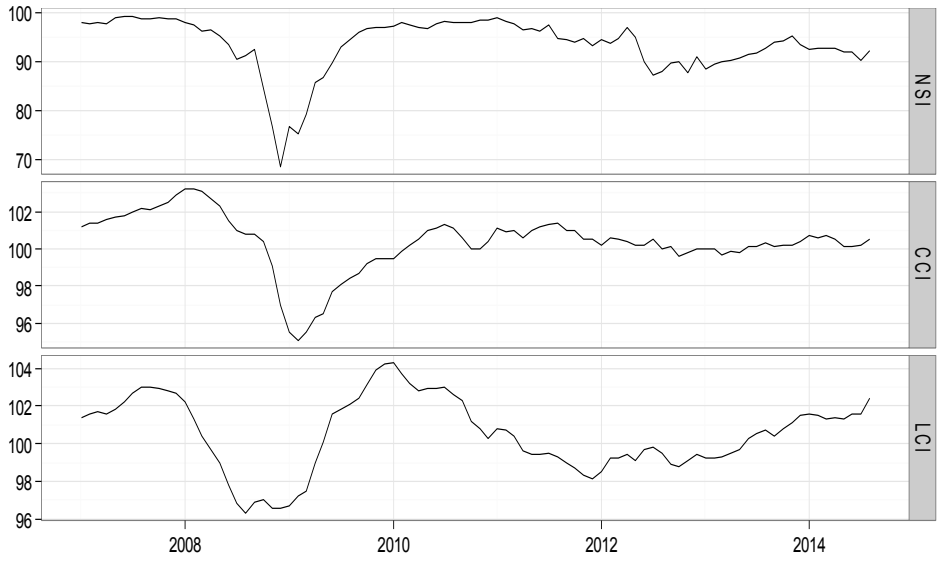


(f) Unemployment Rate



Note: Values on each figure denote cross-correlation coefficients.

〈Figure 14〉 Naver Search Business Index and Business Indicators



Notes: NSI is the monthly Naver search business index. CCI and LCI are cyclical components of the coincident composite index and the leading composite index, respectively.

<Figure 14>은 네이버 검색 경기지수를 경기선행지수 순환변동치 및 동행지수 순환변동치와 비교한 그래프인데, 이를 보면 계절조정 네이버 검색 경기지수가 유용한 경기지수가 될 가능성이 큼을 알 수 있다.

2. 전환점 분석

월별 계절조정 네이버 검색 경기지수의 유용성을 파악하기 위하여 Harding and Pagan (2002)의 방법에 따라 전환점 분석을 실시하였다. 월별 계절조정 네이버 검색 경기지수의 전환점을 동행지수 순환변동치, 선행지수 순환변동치와 경제심리지수 순환변동치의 전환점과 비교하였다. <Table 8>에 각 지수별 전환점이 정리되어 있는데 이를 보면 네이버 검색 경기지수는 동행지수 순환변동치 기준의 경기 정점에 6～7개월, 저점에 2～3개월 선행하여 다른 선행지표보다 시차구조가 안정적이다. 그러나 분석대상 시계열의 길이가 짧아 분석의 의미는 제한적이다.

〈Table 8〉 Turing Point Analysis of Monthly Naver Search Business Index and Other Business Indicators

		Reference date of business cycle		
		Trough	Peak	Trough
The 9th cycle	coincident index	2005. 4	2008. 1	2009. 2
	leading index	2003. 5(-11)	2007. 9 (-4)	2008. 8 (-6)
	ESI	-	2007.10 (-3)	2008.12 (-2)
	NSI	-	2007. 7** (-6)	2008.12 (-2)
The 10th cycle	coincident index	2009. 2	2011. 8	2012.10*
	leading index	2008. 8 (-6)	2010. 1 (-19)	2011.12 (-10)
	ESI	2008.12 (-2)	2010. 3 (-17)	2012.11 (-1)
	NSI	2008.12 (-2)	2011. 1 (-7)	2012. 7 (-3)

Notes: Numbers in parentheses are time difference. + and - denote leading and lagging of Naver searching business index, respectively.
* denotes that the date is not an official reference date and is obtained based on the cyclical component of coincident composite index.
** denotes that the date is estimated from the data due to shortness of series.

3. 예측력 비교

GDP와 네이버 검색 경기지수(NSI)의 발표 시점을 정리하면 <Table 9>와 같다. GDP 분기 속보치는 매분기 종료 후 1개월 이내, GDP 분기 잠정치는 매분기 종료 후 3개월 이내로 발표되고 있다. 네이버 검색 경기지수는 당월 데이터가 당월에 집계되므로 네이버 검색 경기지수는 잠정 GDP보다 2개월 빠르게, GDP 속보치보다 1개월 빠르게 발표된다. <Figure 15>는 네이버 검색 경기지수를 경제성장률과 민간소비증감률과 같이 그린 그래프인데 이를 보면 네이버 검색 경기지수가 경제성장률과 민간소비증감률과 어느 정도 밀접한 것으로 나타났다. 따라서 네이버 검색 경기지수는 GDP 또는 민간소비의 당기 예측에 유용하게 이용될 수 있다.

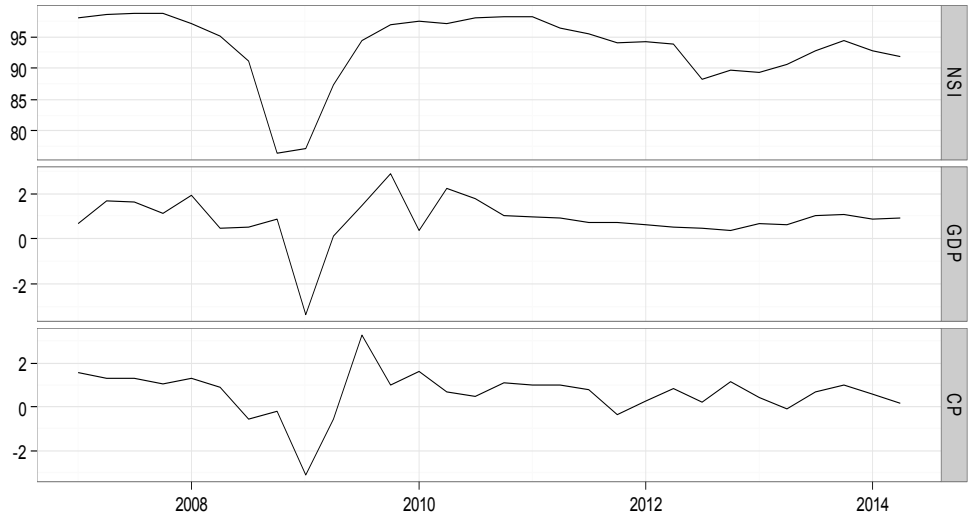
〈Table 9〉 Schedule of GDP Estimate Release and Naver Search Business Index Compilation

Month	1		2		3		4		5		6	
	1～15	16～31	1～15	16～28	1～15	16～31	1～15	16～30	1～15	16～31	1～15	16～30
GDP		Q4 A			Q4 P			Q1 A			Q1 P	
NSI		M1		M2		M3		M4		M5		M6
Month	7		8		9		10		11		12	
	1～15	16～31	1～15	16～31	1～15	16～30	1～15	16～31	1～15	16～30	1～15	16～31
GDP		Q2 A			Q2 P			Q3 A			Q3 P	
NSI		M7		M8		M9		M10		M11		M12

Notes : Q1, Q2, Q3 and Q4 denote each quarter. M1, ⋯, M12 denote each month. A and P denote release date of advance and preliminary GDP estimates by the Bank of Korea, respectively.

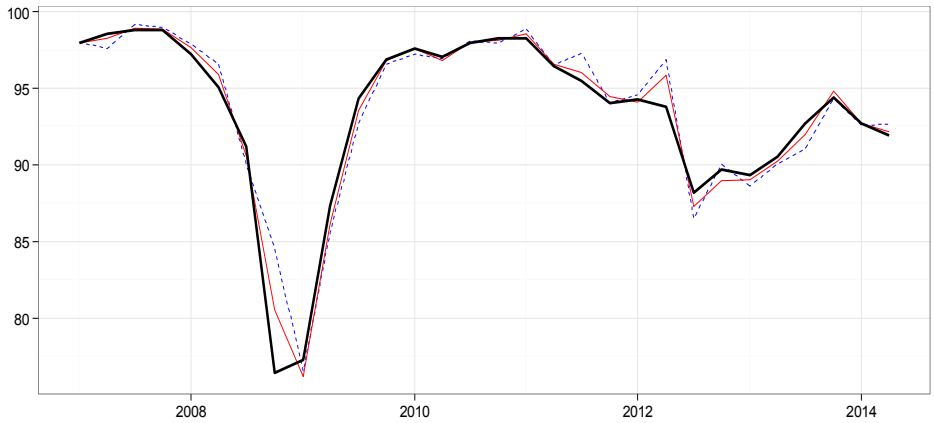
이 절에서는 네이버 검색 경기지수를 이용하여 분기 경제성장률과 민간소비증감률의 예측 가능성을 점검하였다. 예측은 네이버 검색 경기지수를 이용한 예측

〈Figure 15〉 Growth Rates of GDP and Private Consumption and Naver Search Business Index



Notes: NSI is the Naver search business index. GDP and CP are growth rates of GDP and private consumption, respectively.

〈Figure 16〉 Quarterly Naver Search Business Index Based on Available Monthly Data



Notes: The bold line is the quarterly Naver search business index based on 3 month data. The solid line is the quarterly Naver search business index based on 2 month data. The dotted line is the quarterly Naver search business index based on 1 month data

모형과 임의보행모형, AR(1)모형에 의한 예측 결과와 비교하였다[15]. 분기 네이버 검색 경기지수는 공표 시점에 따라 1개월, 2개월, 3개월간 월별 네이버 검색 경기지수를 평균하여 각월별로 작성하고 이를 활용하여 분기 경제성장률과 민간소비증감률을 예측하였다.[16] 1개월, 2개월, 3개월의 데이터를 바탕으로 작성된 월별 네이버 검색 경기지수는 <Figure 16>과 같은데 이를 보면 3개월 데이터를 평균한 분기 네이버 검색 경기지수가 1개월, 2개월 데이터를 평균하여 구한 분기 네이버 검색 경기지수 와 밀접하게 움직이는 것으로 나타났다. 3개월 데이터를 평균한 분기 네이버 검색 경기지수와 1개월, 2개월 데이터를 평균한 네이버 검색 경기지수 간 상관계수가 각각 0.95, 0.99이다.

경제성장률 (전기 대비)과 민간소비 증감률 (전기 대비)의 예측력을 비교하기 위해 다음 3개의 모형을 이용하였다.

15) AR모형은 모형선택기준인 AIC를 최소화는 모형인 AR(1)모형을 선택하였다.

16) 예를 들면 2014년 8월말의 경우 네이버 검색 경기지수는 7월과 8월의 2개월 데이터를 얻을 수 있으므로 이를 평균하여 3/4분기 네이버 검색 경기지수를 작성하고, 2014년 9월말의 경우 7월, 8월, 9월의 3개월 데이터를 얻을 수 있으므로 이를 평균하여 3/4분기 네이버 검색 경기지수를 작성한다.

① 모형 1 : $y_t = y_{t-1} + \epsilon_t$

② 모형 2 : $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$

· 추정기간 : 전 기간, 최근 7분기

③ 모형 3 : $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 NSI_t + \epsilon_t$

· 추정기간 : 최근 7분기

· 분기 네이버경기검색지수 : 1개월, 2개월, 3개월 평균해서 분기로 이용

여기서 y_t는 경제성장률 (전기 대비) 또는 민간소비 증감률 (전기 대비)이며 ϵ_t는 오차이다. β_i는 추정할 모수이며 NSI_t는 분기 네이버 검색 경기지수이다. 모형 1은 임의보행모형이며, 모형 2는 AR(1)모형이다. 모형 2의 경우 전 기간을 모형화하는 경우와 최근 7분기를 이용하는 경우로 나누어 살펴보았다. 모형 3의 경우 네이버 검색 경기지수와 GDP 및 민간소비와 관계의 변동성을 고려하여 최근 7분기를 이용하였고 설명변수인 네이버 검색 경기지수로 1개월, 2개월, 3개월 데이터를 바탕으로 작성된 통계를 이용하였다. 표본대상기간은 2007년 1/4분기부터 2014년 2/4분기이다.

경제성장률과 민간소비증감률 모형의 예측력은 다음의 MAE(mean absolute error)를 바탕으로 평가하였다.

〈Table 10〉 Comparison of Forecasting Performance for GDP Growth Rate

	Whole period (2008.4/4-2014.2/4)	Crisis period (2008.4/4-2009.4/4)	Period excluding crisis (2010.1/4-2014.2/4)
Model 1	0.782	2.580	0.282
Model 2			
(whole period)	0.739 (0.945)	2.383 (0.924)	0.283 (1.003)
(Recent 7 quarters)	0.868 (1.082)	2.374 (0.920)	0.450 (1.596)
Model 3			
(3 months)	0.647 (0.827)	1.184 (0.459)	0.498 (1.766)
(1 month)	0.717 (0.917)	1.628 (0.631)	0.463 (1.642)
(2 months)	0.706 (0.902)	1.460 (0.566)	0.496 (1.759)

Note: Numbers in parentheses are relative MAE(mean absolute error) compared with those from model 1.

〈Table 11〉 Comparison of Forecasting Performance for Private Consumption Growth Rate

	Whole period (2008.4/4-2014.2/4)	Crisis period (2008.4/4-2009.4/4)	Period excluding crisis (2010.1/4-2014 2/4)
Model 1	0.947	2.249	0.536
Model 2			
(whole period)	0.761 (0.804)	1.630 (0.725)	0.519 (0.968)
(Recent 7 quarters)	1.051 (1.110)	2.656 (1.181)	0.605 (1.129)
Model 3			
(3 months)	0.867 (0.916)	1.765 (0.785)	0.618 (1.153)
(1 month)	0.896 (0.946)	1.781 (0.792)	0.651 (1.215)
(2 months)	0.856 (0.904)	1.620 (0.720)	0.644 (1.201)

Note: Numbers in parentheses are relative MAE(mean absolute error) compared with those from model 1.

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |\text{예측치}_t - \text{실적치}_t|$$

예측력 평가는 실제 상황을 감안해서 2008년 4/4분기부터 2014년 2/4분기까지 매분기별로 당기예측을 실시하고 당기 예측치와 실적치를 비교했다. 시계열의 길이가 짧음에도 불구하고 2008년 4/4분기부터 예측을 시작한 이유는 글로벌 금융위기를 포함한 기간에 대해 예측모형들 간 당기 예측력을 파악하기 위해서이다. 예측력 평가 결과는 <Table 10, 11>에 정리되어 있다. <Table 10>의 경제성장률의 예측력 비교 결과를 보면 전체기간에 대해 분기 네이버 경기 검색지수를 이용한 모형 3이 우수하게 나타났다. 비교기간을 금융위기 구간과 그 외 구간으로 나누어 보면 네이버 경기 검색지수를 이용한 모형 3이 금융위기 기간에 예측력이 특히 우수하게 나타났다. 그러나 금융위기 이후 성장률이 낮고 변동성이 낮은 임의보행적 행태를 보이면서 금융위기 이후 기간에서는 임의보행모형과 AR(1)모형이 우수하게 나타났다. 1개월 정보로 분기 네이버 경기 검색지수를 추정한 후 이를 이용하여 예측모형을 작성하더라도 금융위기시 동모형은 임의보행모형과 AR(1)모형에 비해 예측력이 우수한 것으로 나타났다.[17] 한편 <Table 11>의 민간소비증감률 예측력 비교 결과인데, 이를 보면 전체기간에 대해 AR(1)

17) 금융위기 기간에는 언론 및 정책당국에서 불황에 대한 우려가 커짐에 따라서 경제주체의 불황에 대한 검색이 많아졌던 것으로 보인다.

모형과 분기 네이버 경기검색지수를 이용한 모형이 임의보행모형에 비해 예측력이 우수하게 나타났다. 예측모형의 예측력 평가는 데이터가 보다 더 축적된다면 보다 체계적으로 점검될 수 있을 것으로 판단된다.

Ⅵ. 요약 및 향후 과제

본 연구는 빅데이터를 이용하여 경기판단을 할 수 있는 방안을 모색하고 그 유용성을 점검하고자 하고자 진행되었다. 빅데이터로는 검색데이터, 소셜네트워크 데이터, 블로그 데이터가 있다. 검색데이터가 키워드 중심이므로 비정형인 소셜네트워크 데이터 또는 블로그 데이터의 텍스트 분석보다 오류가 적고 의도성 등이 낮아 객관적인 점을 감안하여 본 연구에서는 검색 데이터를 바탕으로 경기를 판단할 수 있는 지수를 작성하였다.

검색 데이터는 네이버, 다음, 구글 등 주요 포털에서 트렌드라는 서비스를 통해 요약해서 제공하고 있다. 구글의 경우 검색 국가분류에 한국이 2013년 12월부터 포함되어서 경기 관련 한글 키워드 검색 데이터가 상대적으로 부족하고, 다음의 경우 2013년 12월부터 서비스 중단되었다. 우리나라의 경우 포털의 점유비율이 네이버가 80% 내외로 높은 점을 감안하여 네이버 트렌드의 검색 데이터를 이용하였다.

네이버 트렌드를 통해 추출하여 네이버 검색 경기지수를 하였는데 작성과정을 정리하면 다음과 같다. 첫째, 경기판단 관련 경제 키워드를 호황, 불황 중심으로 추출하고 호황, 불황과의 관련성 및 대응성을 고려하여 각각 5개씩 추출하여 결합하였다. 둘째, 호황, 불황 검색통계를 월별 데이터로 전환하고 계절조정한 후 BSI 형식으로 네이버 검색 경기지수를 작성하였다. 월별 네이버 검색 경기지수 작성과 동일한 방식으로 주별 네이버 검색 경기지수를 작성하였다.

네이버 검색 경기지수의 유용성을 확인하기 위해서 월별 네이버 검색 경기지수에 대해 교차상관분석, 전환점 분석, 예측력 분석 등을 실시하였다. 교차상관분석 결과를 보면 네이버 검색 경기지수는 경제심리지수와 매우 밀접하게 움직이며 경기동행지수 순환변동치와도 밀접하면서도 교차상관계수 최댓값 기준으로 2개월 선행하는 것으로 나타났다. 전환점 분석을 위해 네이버 검색 경기지수는 경기 정·저점에 선행하는 것으로 나타났다. 마지막으로 경제성장률과 민간

소비증감률의 예측력을 비교하였는데 네이버 검색 경기지수를 포함한 모형의 예측력이 금융위기 기간 중 상대적으로 우수한 것으로 나타났다.

네이버 검색 경기지수는 주별로 작성되고 빠르게 데이터를 구할 수 있고 경제심리지수와 매우 밀접하게 움직이는 것으로 나타나 경제심리지수를 보완할 수 있는 지표로 활용될 수 있을 것으로 판단된다. 네이버 검색 경기지수의 유용성을 확인하기 위해서는 지속적으로 작성하면서 그 유용성을 점검할 필요가 있다. 네이버 검색 경기지수는 지속적으로 공표할 수 있도록 작성 방법을 가능한 단순화하였다.

네이버 검색 경기지수의 작성 연구결과를 보완하고 확장하기 위해서는 향후 다음의 내용을 검토할 필요가 있다. 첫째, 현재 이용한 호황과 불황 각 5개 검색어로 경기 상황 전체를 파악하기 어렵다. 따라서 경기 관련 검색어를 보다 발굴하여 검토할 필요가 있다. 둘째, 고용, 물가 등 경제 분야에 대해서도 검색어군을 확보하여 이와 관련된 검색지수를 작성할 필요가 있다. 이를 통해 고용상황, 물가상황을 조사 전에 파악할 수 있다. 셋째, 네이버 검색 경기지수를 작성할 때 검색어군을 단순 평균하였는데 검색어군을 결합할 때 주성분분석 등의 방법을 통해 가중평균을 이용하고 그 유용성을 검토할 필요가 있다. 넷째, 검색 경기지수가 주별로 작성되므로 혼합주기예측모형을 구성하여 네이버 검색 경기지수를 이용한 예측모형의 활용성을 높일 필요가 있다.

검색통계가 완전한 것은 아니다. Bulter (2013)는 2012년 말 구글 독감 트렌드의 예측 오류에 대해서 지적했다. 구글 독감 트렌드에서 예측오류가 발생했던 주요 이유는 사람들이 새로운 독감 바이러스에 대한 관심으로 독감 증상 없이도 독감 관련 키워드를 검색했기 때문이다. 불황, 호황 검색도 마찬가지의 예측 오류가 발생될 수 있다. 따라서 검색통계 하나만을 의존하기보다는 기존 조사통계와 비교하면서 검색 통계를 활용할 필요가 있다.

빅데이터는 기존에 불가능했던 미시적 경제분석, 속보성 높은 경제분석을 가능하게 하므로 검색통계, 소셜네트워크 데이터, 거래 데이터, 통신 데이터 등으로부터 유용한 빅데이터를 발굴하여 경제분석에 활용할 필요가 있다. 또한 빅데이터 분석에 적합한 분석방법인 데이터마이닝 방법 등을 경제분석에 적용하여 경제분석의 내용을 보다 풍성히 할 필요가 있다.

〈참고문헌〉

구글 트렌드, http://www.google.co.kr/trends/.

구글 독감 트렌드, http://www.google.org/flutrends/intl/ko/.

김민희 · 김나경 (2013), "검색데이터를 보면 소비트렌드가 보인다." *LG Business Insight*, 11/13, pp. 30-38.

김정미 (2012), "빅데이터로 알아가는 세상(UN Global Pulse의 빅데이터 분석사례)", 한국정보화진흥원.

김지은 (2013), "경기지표로서 인터넷 검색지표의 유용성 분석", 한국은행.

네이버 트렌드, http://trend.naver.com/.

네이버 트렌드 도움말, https://help.naver.com/support/service/main.nhn?serviceNo=606&categoryNo=9351.

이긍희 · 함유근 · 김용대 · 이준환 · 원중호 (2014), "빅데이터의 이해", 한국방송통신대 출판문화원.

인터넷 트렌드, http://trend.logger.co.kr/trendForward.tsp.

한국인터넷진흥원 (2013), "2013년 인터넷이용실태조사", 미래창조과학부 · 한국인터넷진흥원.

Bulter, D. (2013), "When Google got flu wrong," Nature News in Focus, http://www.nature.com/news/when-google-got-flu-wrong-1.12413#/fever.

Chamberlain, G. (2010), "Googling the present," *Economic and Labour Market Review*, Office for National Statistics, Vol. 4, No. 12.

Chetty, R., J. Friedman, and J. Rockoff (2011), "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," NBER Working Paper No. 17699.

Choi, H. and H. Varian (2012), "Predicting the Present with Google Trends," *Economic Record* 88, pp. 2–9.

Carriere-Swallow, Y. and F. Labbe (2011), "Nowcasting with Google Trends in an Emerging Market," *Journal of Forecasting* 32(4), pp. 289–98.

D'Amuri, F. (2009), "Predicting Unemployment in Short Samples with Internet Job Search Query Data," Bank of Italy Research Department. available at http://mpra.ub.uni-muenchen.de/18403/1/MPRA_paper_18403.pdf.

Einav, L., M. Jenkins, and J. Levin (2012), "Contract Pricing in Consumer Credit Markets," *Econometrica* 80(4), pp. 1387-1432.

Einav, L., C. Farronato, J. Levin, and N. Sundaresan (2013), "What Happened to Online Auctions?" Mimeo, Stanford University.

Einav, L., A. Finkelstein, S. Ryan, P. Schrimpf, and M. Cullen (2013), "Selection on

Moral Hazard in Health Insurance," *American Economic Review* 103(1), pp. 178-219.

Einav, L., D. Knoepfle, J. Levin, and N. Sundaresan (2013a), "Sales Taxes and Internet Commerce," NBER Working Paper No. 18018.

Einav, L. and J. D. Levin (2013), "The Data Revolution and Economic Analysis," Technical report, NBER Innovation Policy and the Economy Conference. NBER Working Paper 19035.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and the Oregon Health Study Group (2012), "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics* 127(3), pp. 1057-1106.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2009), "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* 457, pp. 1012-1014.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010), "Predicting Consumer Behavior with Web Search," *PNAS* 107(41), pp. 17486−17490.

Harding, D. and A. Pagan (2002), "Dissecting the Cycle: A Methodological Investigation," *Journal of Monetary Economics* 49(2), pp. 365−381.

Hellerstein, R. and M. Middeldorp (2012), "Forecasting with Internet Search Data," Liberty Street Economics Blog of the Federal Reserve Bank of New York, January 4.

Market Share Statistics for Internet Technologies, http://www.netmarketshare.com.

McLaren, N. and R. Shanbhoge (2011), "Using Internet Search Data as Economic Indicators," *Bank of England Quarterly Bulletin* 51(2), pp. 134-40.

Return On Now (2014), "2013 Search Engine Market Share By Country," http://returnonnow.com/internet-marketing-resources/2013-search-engine-market-share-by-country/.

Saiz, A. and U. Simonsohn (2008), "Downloading Wisdom from Online Crowds," Discussion Paper No. 3809, The Institute for the Study of Labor, Germany.

Suhoy, T. (2009), "Query Indices and a 2008 Downturn: Israeli Data," Bank of Israel Discussion Paper.

United Nations, http://www.unglobalpulse.org.

Vosen, S. and T. Schmidt (2011), "Forecasting Private Consumption : Survey-Based Indictor vs. Google Trends," *Journal of Forecasting*, 30(6), pp.565-578.

Wu, L. and E. Brynjolfsson (2009), "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales," MIT Sloan School of Management, http://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-3b3_paper.pdf.

Zimmermann K. F. (2009), "Google Econometrics and Unemployment Forecasting," *Applied Economic Quarterly* 55(2), pp. 107-120.

Business Cycle Indicator Using Big Data: Compilation of the Naver Search Business Index

Geung-Hee Lee*, SangPil Hwang**

With the advent of big data, there are a lot of economic analyses going on using amorphous data. Such are internet search queries, dialogues on social networks and blog posts. In this paper, we introduce a search business index based on the internet search data provided by Naver Trends considering market shares in Korea. The index is compiled based on the difference between search query data related to the business boom and recession. To check the usefulness of the newly compiled index, various analyses have been carried out. The analyses show that the index is highly correlated with the economic sentiment index and leads the business cycle by 2 months. Moreover, the forecasting performance of models with the index can be compared with benchmark models such as the random walk model. The results show that models with the index outperform the random walk model and AR(1) model during the global financial crisis period. The Naver search business index would be helpful in evaluating business cycles and complementing the economic sentiment index.

JEL Classification Number : C1, C8, E3

Key words : Internet search data, Business cycle indicator, Business survey index, Economic sentiment index, Forecasting.

* Corresponding Author, Professor, Department of Information Statistics, Korean National Open University, 86 Daehak-ro, Jongno-Gu, Seoul, 110-191, Korea. (E-mail: geunghee@knou.ac.kr)

** Head, Model-based Analysis Team, Macroeconomic Modelling Division, Research Department, The Bank of Korea (E-mail: hwangsp7@bok.or.kr)

This work was supported by the Bank of Korea.

Received 30 Septempber 2014; Received in revised form 22 November 2014; Accepted 16 December 2014