네이버 제휴 언론 기사를 이용한 실시간 이슈분석시스템 설계 및 구현

이종화 (부경대학교 경영대학 경영학부 박사, 주저자 newjwcom@daum.net)
이현규 (부경대학교 경영대학 경영학부 교수, 교신저자 hyunqlee@pknu.ac.kr)

Design and Implementation a Real-time analysis system for issues on Naver's partnership online Newspaper

Jong-Hwa Lee (Candidate Ph. D., Graduate School of Business, Pukyong National University, First Author)
Hyun-Kyu Lee (Professor. College of Business Administration, Pukyong National University, Corresponding Author)

-원고매수: 16 페이지

[교신저자 연락처]
◎ 이현규
- 주소 : 부산 남구 용소로 45, 부경대학교 경영대학 경영학부 교수
- 전화번호 : 051-629-5744, 휴 대 폰 : 010-3222-1950
- E-mail주소: hyunqlee@pknu.ac.kr

◎ 이종화
- 휴 대 폰: 010-3526-6050
- E-mail주소: newjwcom@daum.net

네이버 제휴 언론 기사를 이용한 실시간 이슈분석시스템 설계 및 구현

Design and Implementation a Real-time analysis system for issues on Naver's partnership online Newspaper

● 목차 ●

··· Abstract ···

언론은 매체를 통해 발생한 사실을 사람들에게 알리는 역할을 하며 신문, 텔레비전, 라디오, 잡지 등의 매체로 구분한다. 요즘은 정보 통신 기술과 인터넷의 발달로 여러 매체의 다양한 뉴스를 보다 쉽게 접할 수 있으며 여론이 형성될 수 있는 정보를 제공하는 언론의 영향력은 아주 크다고 볼 수 있다.

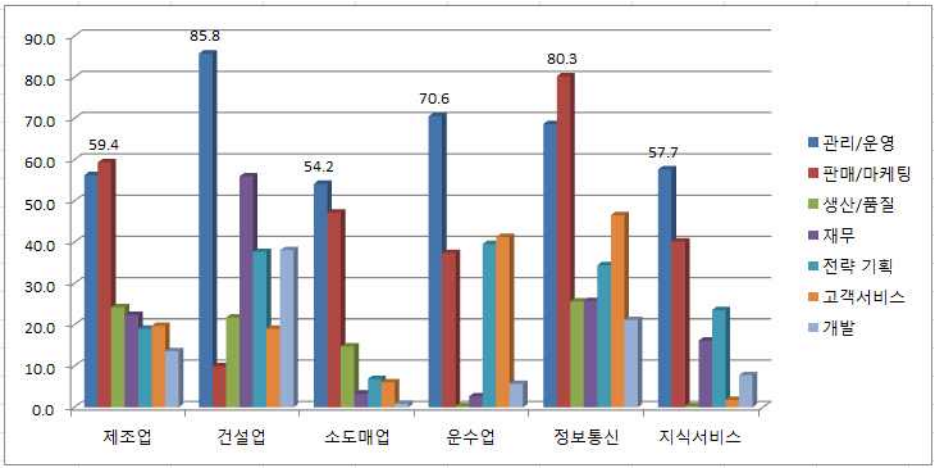
문화체육관광부 정기간행물 현황 등록알림표의 “인터넷 신문” 발간 통계자료에 따르면 2005년 286곳의 인터넷 신문을 시작으로 2010년에는 2,484곳에 이르며 2015년에는 6,605곳으로 집계되었다. 인터넷의 보급과 모바일 기기의 등장, 네트워크의 발달로 무려 10년 사이 23배 이상 증가한 수치이다. 사용자들의 사용 가치를 반영하듯 많은 연구자들의 연구 대상으로 활용되어 왔다. 하지만 대부분의 기사 분석 연구들은 최신 정보를 반영하지 못하기 때문에 시간이 지남에 따라 실시간으로 바뀌는 여론의 방향을 정확히 예측하기가 어려웠고, 그래서 최신 정보를 업데이트하여 실시간으로 생성되는 데이터를 빠르게 분석할 수 있는 방법에 대한 연구가 필요하였다.

본 연구는 네이버에서 제공되는 뉴스스탠드(newsstand) 서비스에 실시간 업데이트되는 기사들을 대상으로 실시간 분석 시스템을 구축하고자 한다. 온라인 포털 사이트 네이버 뉴스 서비스에서 제휴된 157개 언론사를 방송, 통신, 경제, 스포츠, 연예, 매거진, 영자지, IT, 지역, 언론사등의 범주로 분류하였으며 키워드 검색 분석, 일별 이슈 분석, 키워드 시계열 분석을 실시간으로 분석하여 여론의 방향을 보다 빠르게 확인할 수 있었고 무엇보다 개발 과정에 사용되는 Server/Client 관련 언어(Language)는 누구든지 쉽게 학습하고 수정하고 배포할 수 있는 오픈 소스 소프트웨어(Open Source Software)를 활용하였다.

Key Words : Text Mining, predictive analytics, R Program, Open Source Software

Ⅰ. 서 론

“논어”는 중국 최초의 어록(語錄)이며 춘추전국시대의 유교의 근본문헌으로 알고 있다. 공자와 그 제자와의 문답으로 전개되며 공자의 행적이나 발언 등 인생의 교훈이 되는 말들이 함축성있게 표현되어 있다. 현대인들은 옛 선인 공자처럼 모든 행적이나 말들을 제자가 아닌 ICT기술이 대신하여 모든 것을 더욱 상세하게 기록하고 있다. 물론 개개인 모두를 대상으로 실시간으로 쌓아 가고 있다. 비정형의 다양한 데이터, 문자 데이터(SMS, 검색어), 영상 데이터(CCTV, 동영상), 위치 데이터 등 민간 분야, 공공 분야의 모든 업종에서 데이터를 양산 중에 있다. 이러한 거대한 데이터는 기업, 정부 할 것 없이 모든 기관에서 수집 및 분석 하고 있다. <그림 1>은 국가통계포털인 e-나라지표 자료를 분석한 것으로 2015년 3,500여개의 기업을 대상으로 빅데이터 산업별 활용분야 조사에 따르면 제조업은 관리/운영분야와 판매/마케팅분야에 빅데이터를 활용하는 것으로 나타났다. 건설업은 관리/운영분야에 86% 가까이 빅데이터를 활용하고 있다고 조사되었다. 소도매업, 운수업, 지식서비스 분야도 관리/운영 업무에 데이터를 사용하고 있고 정보통신 분야는 판매/마케팅 분야에 높은 빅데이터 이용 현황을 확인 할 수 있었다(lee et al., 2016; www.index.go.kr).



<그림 1> 2015년 산업별 빅데이터 활용 분야

네트워크의 발달로 지역간 경계가 사라지고 글로벌 환경은 국내외 기업의 무한 경쟁시대로 발달하고 있다. 또한 IT기술의 융합으로 산업간 경쟁의 벽이 무너지면서 모든 산업이 한 울타리에서 무한한 경쟁을 벌이고 있는 것이다. 구글(Google)은 데이터

회사에서 이제는 자동차를 개발하는 새로운 시장에 도전하여 무인 자동차시장의 선두에 서 있다. 한국을 5대 자동차 생산국으로 만든 현대자동차 역시 IT 업체인 Google, Apple를 비롯한 IT와의 경쟁을 피할 순 없을 것이다. 전자신문에 따르면 구글은 자율주행차의 누적 주행 거리만 해도 무려 322만킬로미터로 도심 주행시 발생할 수 있는 상황들에 대한 데이터를 확보하고 있으며 그러한 빅데이터(BigData) 기반 무인 주행 자동차를 개발하여 2020년경 자율주행차를 상용화할 계획이라고 밝혔다. (http://www.etnews.com/ 20161006000003).

사용자 정보와 관계 정보 그리고 소비자 형태에 따른 고객 데이터 관계 분석, 페이스북, 트위터, 언론에서의 이슈 정보를 분석하는 SNS 비정형 데이터 분석, 이미지나 동영상의 의미 분석과 콘텐츠 소비 형태 또는 선호도 분석등의 대용량 멀티미디어 분석, 실시간 사물 센서 데이터 분석이나 RFID, 원격 헬스 모니터링과 같은 M2M 센서 정보 분석 등과 같이 데이터 분석 기술이 발전되고 있다(조성룡, 2012).

빅데이터 시대에는 단순히 관계형 데이터베이스에 잘 정리된 정형 데이터뿐 아니라 인터넷, 매일 3억 부 이상 발행되는 신문, 1400억 건 이상 발송되는 이메일, 4억 건 이상 발생되는 트윗 등의 비정형 빅데이터를 효과적으로 분석하는 것이 무엇보다 중요해졌다(http://www.worldometers.info/kr/). ETRI의 보고에 따르면 2012년 2.8제타바이트에 이른 빅데이터는 2020년에는 40제타바이트로 급격히 증가할 것이며, 그 중 20%는 정형 데이터, 나머지 80%는 비정형 데이터가 될 것으로 예상된다고 한다 (https://www.etri.re.kr/).

본 연구는 비정형 데이터 비중이 커지는 웹 환경에서 텍스트 중심 마이닝 연구를 하고자 한다(김용현・허의남, 2014). 웹 환경에서 데이터 수집, 가공 및 처리, 통계/분석 작업, 시각화 과정을 실시간으로 처리할 수 있는 프로세스를 설계하고자 한다. 서버프로그래밍, 클라이언드 프로그래밍 구현과 웹 환경에 사용자 인터페이스를 제공함으로써 입력과 출력이 가능한 시제품을 만들고자 한다. 본 연구를 통해 최신 데이터를 반영한 이슈 정보 분석을 해보고자 한다.

Ⅱ. 선행 연구

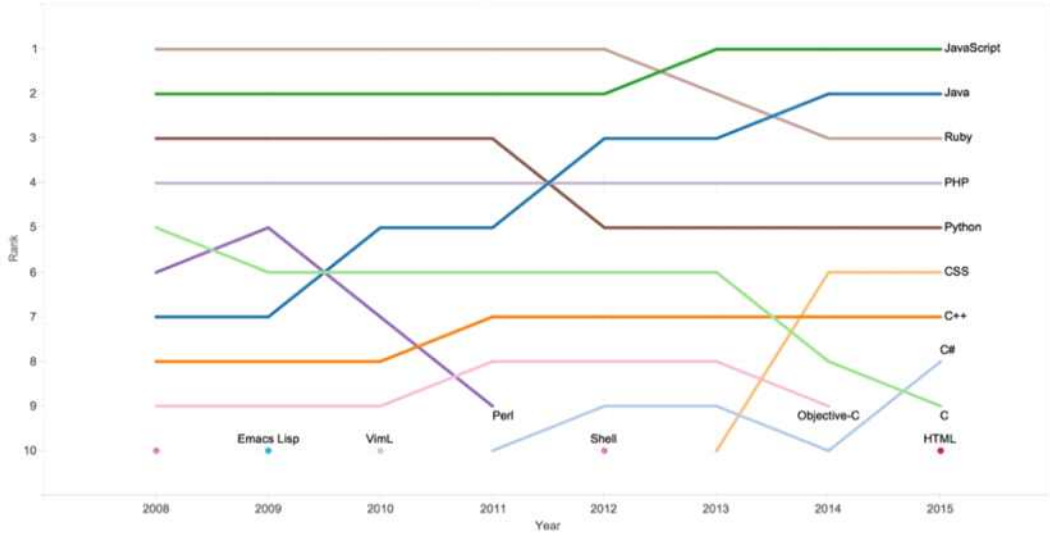
웹 환경의 정보들을 실시간으로 제어하기 위해서 웹 페이지의 개발 환경을 이해되어야 할 것이다. 인터넷에서 정보를 검색하는데 사용하는 소프트웨어인 웹 브라우저 (Web Browser) 즉, 웹 페이지를 보여주는 프로그램이 있다. Internet Explorer, Google Chrome, Apple Safari, Mozilla Firefox, Opera software Opera 등 다양한 브

라우저로 사용자들은 인터넷을 즐기고 있다. 이러한 웹페이지를 구현하는 도구들을 먼저 살펴본다.

1. 자바스크립트(Javascript)

웹페이지는 정보 제공자인 Server측과 정보를 요구하는 Client측으로 나누어진다. 웹 페이지의 내용과 모양을 제어하기 위하여 ASP(Active Server Page), JSP(Java Server Page) 등의 서버 페이지 제작 언어들이 존재한다. 현재 대부분의 웹 페이지에서 클라이언트 페이지 제작 언어로는 HTML(Hyper Text Markup Language), CSS(Cascading Style Sheets), JavaScript 등이 담당한다. 웹 페이지의 큰 틀을 제작할 때는 HTML언어를 사용하며, 페이지 내 글씨체나 색깔과 같은 디자인적 요소들을 표현할 때는 CSS가 담당한다. JavaScript는 웹 페이지의 동작을 담당하는데 '버튼을 선택했을 때 시간을 보여줘'라는 형태의 명령을 내릴 수 있는 풍부한 효과를 넣을 수 있다(정원기·문수묵, 2010; 최민아 외, 2013).

분산 버전 관리 툴인 'Git'을 사용하는 프로젝트를 지원하는 웹호스팅 서비스를 하고 있는 'GitHub'는 2013년이후 CSS와 함께 웹 페이지 구축에 많이 사용되고 있다고 한다. <그림 2>를 살펴보면 객체 지향 언어인 JavaScript, Java, Ruby 등이 개발자들 사이에 많이 사용되고 있다는 것을 알 수 있다. 웹페이지 구축을 위한 언어로는 단연 JavaScript가 많이 사용된다는 통계 자료를 볼 수 있다. 또한 PHP와 CSS 그리고 HTML을 볼 수 있다(김진국 외, 2015; http://github.com/).



<그림 2> GitHub 사용자의 소스 코드 언어 순위

2. 제이쿼리(JQuery)

디자인을 변경하거나 애니메이션 등의 효과를 자바스크립트로 작성할 경우 엄청난 소스 코드를 입력해야하는 기능들이 Jquery를 사용하면 이 모든 것을 쉽고 빠르게 작성할 수 있다. 웹 페이지를 개발하는 개발자의 생각하는 방법을 바꾼 언어로 자바 스크립트 라이브러리이다(장영현 외, 2011). Jquery 환경에서 사용하는 Ajax 기법은 대화식 웹 페이지 제작을 위한 웹 개발 기법이다. Ajax는 페이지 이동 없이 고속으로 화면 전환이 가능하며 수신하는 데이터 양을 줄일 수 있고, 클라이언트에게 처리를 위임할 수도 있다. 단점은 Ajax와 호환성에 문제가 있는 브라우저는 Ajax 기법을 사용한 페이지를 볼 수 없다는 것이다(https://wikipedia.org/).

<그림 3>웹페이지에서 JavaScript로 작성된 소스 코드와 JQuery로 작성된 소스 코드를 비교한 예문이다. wrapper라는 아이디를 가진 div태그 안에 ul태그가 있고 그 자식으로 li태그가 4개가 있고 li태그의 자식으로 span태그가 있는 html 구성에서 세 번째 'li'태그의 글자색을 빨간색으로 변경하려고 할 때 두 언어의 기법을 이용하여 표현한 것이다.

Language	Source Code

Java Script	window.onload = function(){
	//페이지가 로드되면
	var wrapper = document.getElementById('wrapper');
	//'warpper' id를 찾아서 wrapper 변수에 대입
	var ul = wrapper.getElementsByTagName('ul');
	//'ul' 태그를 찾아서 ul 변수에 대입
	var li = ul[0].getElementsbyTagName('li');
	//'li' 태그를 찾아서 li 변수에 대입
	
	for(var i=0; i<li.length; i++)
	{
	// li의 개수만큼 반복
	
	var l = li[i];
	// 각 li를 l 변수에 대입
	if(l.className=='three')
	{
	//l의 class가 tree인지 비교해서 맞으면
	var s = l.getElementsByTagName('span');
	//l의 자식인 'span'태그를 s 변수에 대입
	s[0].style.color='red';
	//'span'의 0번째요소 즉, 텍스트의 글자색 변경
	}
	}
	}

jQuery	$(document).ready(function(){
	//페이지 로드되면
	$('#wrapper > ul > .three > span').css('color','red');
	//wrapper id의 자식인 'ul' 태그의 자식인 three class의 자식인 span 태그의 글자색을 red로 변경
	});

<그림 3> Java Script와 jQuery 소스 코드 비교

3. 크롤링(Crawling)

웹 페이지를 그대로 가져와서 데이터를 추출해 내는 과정을 의미하며, 이번 연구의 가장 많은 노력과 시간이 소요된 과정이다. 크롤링 대상 페이지를 웹 클라이어트를 통해 접속하고 접속된 웹 클라이언트를 통해 HTML 파일을 파싱(Parsing)하여 연구

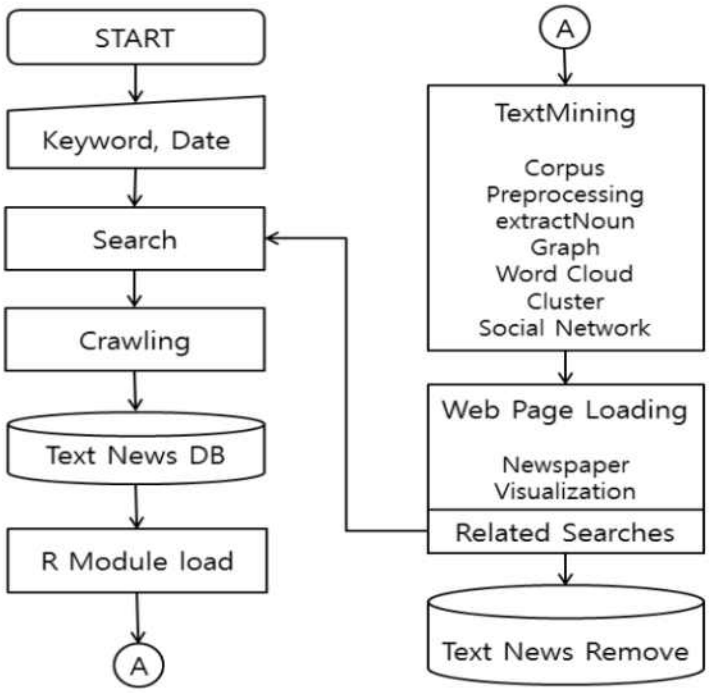
자가 원하는 내용을 가져오는 과정이다. 최민석(2015)은 페이스북의 크롤링 과정을 실증 연구를 통하여 연구하였다. 리눅스(Linux) 크론(Cron)을 이용하여 데이터 수집하며 MySql DB에 결과물을 저장하였다. HTML, CSS, JavaScript 언어를 활용하여 사용자 인터페이스를 구현하였다. 일반적인 웹서버와 DB를 이용하여 접근성과 호환성을 확보하였고 적은 비용으로도 시스템 구축이 가능하였다. 무엇보다 오픈 소스 API를 이용하여 간단한 웹 프로그래밍 작업으로도 자료 수집에 큰 문제가 없었다고 한다(구홍서, 2000; 손수아・박석천, 2015; Cachia et al., 2007; Black, 2008).

Ⅲ. 연구방법 및 모형

인터넷 신문은 2005년부터 시작하여 현재 2,500곳의 인터넷 신문을 운영하고 있다(http://www.index.go.kr). 인터넷의 보급과 모바일 기기의 등장, 네트워크의 발달로 무려 10년 사이 급성장한 상업이기도 하다. 중앙지를 비롯한 언론사들은 사용자들의 사용가치를 반영하듯 학계에서도 많은 연구자들이 이슈 분석, 선거 분석의 연구 대상으로 활용되어 왔다(Kam and Song, 2012). 하지만 대부분의 기사 분석 연구들은 최신 정보를 반영하지 못하기 때문에 시간이 지남에 따라 실시간으로 바뀌는 여론의 방향을 정확히 예측하기가 어려운 것이 한계점으로 나타났다(Lee and Lee, 2015).

본 연구는 최신 정보를 업데이트하여 실시간으로 생성되는 데이터를 빠르게 분석할 수 있는 웹 페이지를 구현하였다. 포털 사이트 검색 기능처럼 사용자가 궁금해 하는 이슈 단어와 해당 기간을 선택하면 관련된 단어가 있는 기사들을 크롤링하여 서버 메모리에 로딩되고 R 프로그래밍을 통하여 시각화하는 텍스트마이닝 처리가 되며 시각화 된 결과는 이미지 처리되어 물리적인 저장 파일로 처리된다. 이후 마이닝 결과를 웹 페이지에 Reload 하므로 사용자의 질문에 대한 이슈 분석 결과를 제공하게 된다.

<그림 4>는 본 연구의 실시간 이슈분석 시스템의 프로세스이며 웹페이지는 Google Chrome 웹 브라우저에 최적화되어 설계되었다. Linux CentOS 서버 환경에 별도의 실행 파일 없이도 JavaScripe 환경의 언어를 동적으로 실행이 가능하며 문서를 빠르고 쉽게 작성할 수 있는 장점의 PHP(Hypertext Preprocessor) 스크립터 언어를 사용하였다(구홍서, 2000). HTML, JavaScript, jQuery, Ajax 등의 스크립터 언어를 사용하여 사용자 인터페이스를 구축하였으며 신문 기사 텍스트 분석엔 R program을 사용하였다.



<그림 4> 본 연구의 프레임워크

웹 페이지 로딩 시간을 줄이기 위하여 각 페이지를 불러오는 모든 작업은 ajax를 이용하여 실시간으로 페이지 이동 없이 설계되었다(이상준・이동훈, 2015). 메인페이지는 검색어, 시작날짜, 끝날짜, 검색버튼, 연관검색어, 기사내용(제목, 날짜, 언론사, 내용), 텍스트마이닝 분석 결과 영역으로 구성된다.

메인모듈에서 검색어 입력 후 검색 버튼을 누르면 ajax를 이용하여 ajax_search_news 모듈로 검색어, 날짜를 전송한다.

ajax_search_news 모듈은 스누피 클래스를 이용하여 검색어와 날짜에 해당하는 기사 리스트를 불러온 후 다시 기사리스트에서 각각 기사 페이지로 들어가 기사제목, 날짜, 언론사, 기사내용을 크롤링한다(그림 5). 그리고 MySql DB에 저장한다. 스누피 클래스는 웹 브라우저를 시뮬레이션하는 PHP 클래스로 웹 페이지의 콘텐츠를 검색하고 양식을 게시하는 작업이 자동화처리되며 인증통과, 프록시 호스트 지원, 웹페이지 텍스트 가져오기, 쿠키 헤더 내용 가져오기등의 클래스들을 제공하고 있다(임상석, 2012).

```
$.ajax({
		url:'./ajax_search_news.php',
		type:'post',
		data:{
			sch_txt : $('#sch_txt').val(),
			startDate : $('#startDate').val(),
			endDate : $('#endDate').val()
		},
		success:function(data){

			$('#result_text').html(data);

			//rscript실행하기
			$.ajax({
				type:'GET',
				url:'../r_test.php',
				success:function(html) {
					var url = "./ajax_news_list.php";
					$('#ifr_search').attr('src', url);

					data_reload();
				}
			});
		}//success end
}); //ajax end
```

<그림 5> ajax_search_news, r_test 코딩 과정

스누피 과정이 성공적으로 실행이 되면 메인페이지로 돌아와서 결과를 메모리에 탑재하고 빅데이터 분석 도구인 R을 실행하기 위해 r_test 모듈을 로드한다(그림 4). r_test 모듈은 텍스트마이닝을 처리하기 위한 과정이 탑재되었다. 먼저 마이닝 처리에 필요한 라이브러리 함수를 로드하고 DB에 저장된 값을 Corpus 과정을 통하여 말뭉치로 묶는다. 불필요한 불용어 및 공백 제거, 크롤링 과정에서 빈번히 나타나는 웹 소스 등을 제거한다. 또한 뉴스 기사들에 흔히 등장하는 “기자”라는 단어와 이메일 주소 등을 함께 제거한다. 이렇게 준비된 말뭉치는 명사 추출 과정을 거치며 2글자 이상인 단어들의 리스트가 완성이 된다. R의 table 함수를 이용하여 단어별 빈도를 계산하면 시각화 전 단계의 모든 과정을 오류없이 처리한 것이다. 마이닝 결과를 한 장의 그림으로 표현하는 Word Cloud Analysis, 유사성을 기초하여 동질적인 집단으로 분류하는 Cluster Analysis, 그래프 이론으로 연결 구조와 연결 강도 등을 바탕으로 키워드 간의 사용 빈도 분석을 통해 서로의 영향력을 측정하는 Social Network Analysis 등을 진행하였다. 모든 시각화된 분석 결과는 png 파일 형식으로 저장하여 메인 페이지에 업로드할 준비를 한다. 또한 메인페이지에 제공할 연관 검색어 리스트도 함

께 R에서 추출하였다.

크롤링, 분석과정을 마무리하고 사용자 인터페이스로 넘어가는 메인 페이지 구성을 살펴보면 <그림 6>과 같다.

<그림 6> 연구 메인 페이지 구성

사용자의 검색어, 날짜 조건에 관련된 기사를 크롤링하여 기사 내용을 확인할 수 있다. 기사 내용은 기고된 날짜, 언론사, 기사내용 순으로 페이지에 로드하여 메인모듈에 iframe으로 기사를 출력한다. iframe은 웹 문서 중간에 다른 웹 문서나 텍스트를 원하는 크기로 로드 가능한 태그이다. 그리고 연관검색어 top10을 출력하기 위해 ajax_top10 모듈을 로드하고 결과를 지정된 위치에 출력하고, 연관검색어로 등록된 단어 클릭했을때는 검색어를 입력하고 검색버튼을 클릭하는 것과 동일하게 크롤링이 실행되며 반복적인 분석 작업이 진행된다. 분석 작업의 결과물인 시각화된 Word Cloud, Cluster, Social Network 결과 이미지가 반영될 수 있도록 이미지를 reload를 한다. 사용자에게 제공되는 모든 과정을 시스템 에러 없이 처리한 결과가 도출된다.

Ⅳ. 연구 알고리즘 실험 및 결과

본 연구는 네이버에서 제공되는 뉴스스탠드(newsstand) 서비스에 실시간 업데이트

되는 기사들을 대상으로 실시간 분석 시스템을 구축하고자 한다. 온라인 포털 사이트 네이버 뉴스 서비스에서 제휴된 157개 언론사를 방송, 통신, 경제, 스포츠, 연예, 매거진, 영자지, IT, 지역, 언론사등의 범주로 분류하였으며 키워드 검색 분석, 일별 이슈 분석, 키워드 시계열 분석을 실시간으로 분석하여 여론의 방향을 보다 빠르게 확인할 수 있는 실시간 시스템을 구현 하였다.

실시간 이슈 분석 시스템은 실험용 Server에서 자료 검색, 크롤링, 텍스트마이닝, 사용자 인터페이스까지 구축하는 과정을 구현하였다. 리눅스 환경에서 PHP, Java Script, jQuery, R progrma 등을 연동하여 사용자 개입 없이 시스템에서 전 과정을 분석, 처리하도록 구현하였다. 웹 프로그램 언어들의 연동은 구현이 상대적으로 자유로우나 PHP에서 R 프로그램 연동은 "Rscript" 명령을 통하여 미리 준비된 텍스트마이닝 소스 코드를 실행할 수 있었다.

[Linux Command] Rscript textmining.R

[PHP Command] $abc = `Rscript real.R 5`;
[PHP Command] print_r($abc);

리눅스 환경에서 "textmining.R"로 코딩된 R 프로그램을 CUI환경에서 실행하기 위하여 "Rscript textmining.R"로 커맨드 해야 한다. PHP 프로그램에서 배열의 키와 그에 해당하는 값을 출력하는 출력문은 "print_r" 명령어를 활용하여 "Rscript textmining.R" 출력 및 실행하는 과정을 삽입하였다. 기사 검색 및 크롤링 과정이 종료되면 텍스트마이닝 처리 모듈인 textmiining.R를 연동했다. 이러한 과정을 통하여 사회의 이슈 분석을 일자별 분석으로 실시하였다.

연구 당시 일어난 가장 큰 사건 사고 였던 "경부고속도로 울주 부근 관광버스 화재 10명 사망"이라는 비극적인 뉴스 기사를 발생일인 2016년 10월 13일부터 2016년 10월 17일까지 5일간을 실험의 대상으로 정하였다.

<표 1>는 연구 대상 키워드를 "고속도로"로 검색하여 시스템에 크롤링된 뉴스 기사 건수의 통계이다.

키워드	일자	뉴스 기사 통계	대기시간
고속도로	2016-10-13	385건	22초
	2016-10-14	1,116	35초
	2016-10-15	251	30초
	2016-10-16	361	27초
	2016-10-17	360	25초

<표 1> 실험 데이터 통계

<그림 7>이슈가 된 사건은 저녁 10시경 발생한 교통사고로 실험 이슈의 기사 건수와 일반적인 고속도로 관련 뉴스 건수가 많이 차지하다 보니 빈도 분석에 크게 못 미쳐 실험 이슈가 눈에 띄지 않고 있다.

<그림 7> 2016-10-13 분석 결과

<그림 8> 전날 사고 발생 다음 날로 모든 방송 매체가 Top 뉴스로 다루면서 “버스”, 화재“, ”경부”, “운전기사”, “울산”, “승객”, “타이어”, “조사”, “부부”, “울주군”, “분리”, “탈출” 등 사고 관련 뉴스에 등장하는 키워드들의 빈도가 높은 것을 알 수 있다. “졸음”, “결함”, “상황”, “신원”, “원인”, “바퀴” 등 사고 경위에 관련된 단어들이 등장하기 시작한 시기이기도 하다.

<그림 8> 2016-10-14 분석 결과

<그림 9> 2016-10-15 분석 결과

<그림 9> 3일째 되는 뉴스들은 버스, 화물차 등 대형차들의 고속도로 운전 실태 등 사고의 원인으로 재조명하는 단어들이 등장하고 있다. “운전자”, “운전기사”의 관련된 증명에 대하여 집중적으로 보도하고 있는 것을 알 수 있다.

<그림 10.> 4일째 사고 수습과 뒤 처리를 위한 기사들이 등장하고 있다. 사망자의 형태를 알아보지 못할 정도로 회손되어 “감정”, “DNA”, ”부검“ 등의 키워드가 등장하고 있다. ”고인“들의 안타까운 사연들을 소개하는 기사도 등장하고 있는 것을 알 수 있다.

<그림 10> 2016-10-16 분석 결과

<그림 11> 2016-10-17 분석 결과

<그림 11> 고속버스 관련 법인 “여객법” 개정에 대한 뉴스들이 등장하며 “감식” 결과, 관련 법 “강화”, 운전 기사 “이씨”의 “과실”여부 “확인”과 안정을 찾고 있는 “승객”들의 증언 관련 기사들이 나타났다.

Ⅴ. 결론 및 향후 과제

본 연구는 인터넷 신문 기사를 이용한 실시간 이슈분석 시스템을 설계하고 구현, 그리고 실험까지 진행하였다. 그 결과를 종합적으로 정리해보면 다음과 같다.

최신 정보를 업데이트하여 실시간으로 생성되는 데이터를 빠르게 분석할 수 있는 웹 페이지를 구현하였다. 포털 사이트 검색 기능처럼 사용자가 궁금해하는 이슈 단어와 해당 기간을 선택하면 관련된 단어가 있는 기사들을 크롤링하여 서버 메모리에 로딩되고 R 프로그래밍을 통하여 시각화하는 텍스트마이닝 처리가 되며 시각화 된 결과는 이미지 처리되어 물리적인 저장 파일로 처리된다. 이후 마이닝 결과를 웹 페이지에 Reload 하므로 사용자의 질문에 대한 이슈 분석결과를 제공하게 된다. 본 연구 과정에는 누구나 자유롭게 소프트웨어를 코딩하여 유용한 기술을 공유하며 사용자들이 서로의 기술을 함께 업데이트 가능한 오픈 소프트웨어를 활용하였으며 리눅스(Linux), MySql DB, HTML, CSS, JavaScript, jQuery, R program등이 사용되었다.

웹 페이지는 Google Chrome 웹 브라우저에 최적화되어 설계되었고, Linux CentOS 서버 환경에 Java 환경의 언어를 동적으로 실행하며 PHP(Hypertext Preprocessor) , HTML, JavaScript, jQuery, Ajax 등의 스크립터 언어를 사용하여 사용자 인터페이스를 만들었으며 신문 기사 텍스트 분석엔 R program을 사용하였다. 또한, 웹 페이지 로딩 시간을 줄이기 위하여 각 페이지를 불러오는 모든 작업은 ajax를 이용하여 실시간으로 페이지 이동 없이 설계되었다.

실험은 네이버에서 제공되는 뉴스스탠드(newsstand) 서비스에 실시간 업데이트되는 기사들을 대상으로 실시간 분석 시스템으로 크롤링하여 실험을 하였다. 연구 당시 일어난 가장 큰 사건 사고 였던 “경부고속도로 울주 부근 관광버스 화재 10명 사망”이라는 비극적인 뉴스 기사를 발생일인 2016년 10월 13일부터 2016년 10월 17일까지 5일간을 실험의 대상으로 정하였다. 연구 대상 키워드를 “고속도로”로 검색하여 3,000여 뉴스 기사를 일별 빈도 분석한 결과를 도출하였다. 시스템에 크롤링된 뉴스 기사 건수의 통계이다. 키워드 검색시 분석 결과가 클라이언트에 도달되는 시간은 약 25초 내외로 나타났다.

빅데이터 분석은 방대한 데이터 내에서 패턴을 찾아 예측을 하는 것이다. 가장 단

순한 빈도 분석을 실시간으로 처리할 수 있는 시스템을 구축하여 보다 빠른 이슈의 흐름을 파악하는데 혁신적인 검색 도구라 본다. 관련 뉴스 기사들 전체를 분석하는 전수조사 방식이 가능하기 때문에 무엇보다 의미가 있는 실시간 분석 방법이다. 많은 연구자들이 고객의 소리에 귀 기울이며 소비자의 패턴과 니즈 분석, 의미 기반 분석 등의 기능을 향후 추가하여 개발하고자 한다. 단순 빈도 분석을 넘어 연관 분석, 군집 분석, 분류 분석, 예측 분석까지 실시간 처리가 가능한 시스템 구현을 기대한다.

참 고 문 헌

구흥서, "WWW과 데이터베이스 연동기술의 조사분석," 정보과학회지, 제18권, 제4호, 2000, pp. 32-40.

김용현, 허의남, “효율적인 빅데이터 분석을 위한 로그 데이터기반의 로그분석 지원 시스템,” 한국정보과학회 학술발표논문집, 2014, pp. 936-938.

김진국, 최성수, 지수영, 류관희, “데이터베이스 연동을 통한 빅데이터 분석결과 가시화,” 한국빅데이터서비스학회 논문지2, 2015, pp. 1-10.

손수아, 박석천. "IoT 기반 실시간 시각화 알고리즘을 이용한 스마트가드닝 시스템 설계 및 구현," "인터넷정보학회논문지", 제16권, 제6호, 2015, pp. 31-37.

이상준, 이동훈, “빅데이터 로그를 이용한 실시간 예측분석시스템 설계 및 구현,” 정보보호학회논문지, 제25호 제6호, 2015, pp. 1399-1410.

임상석, "Open source기반 HTML5 mobile web application platform 구조 분석 및 성능 최적화 방법," 한국통신학회지(정보와통신), 제29권, 제9호, 2012, pp. 10-17.

장명현, 박대우, 이예원, “Live site 개념을 도입한 웹사이트 저작도구의 실시간 자동화 모델에 관한 연구,” 한국컴퓨터정보학회, 제19권, 제2호, 2011, pp. 175-177

정원기, 문수묵, “웹 페이지 자바스크립트 분석과 자바스크립트 엔진의 성능 평가,” 한국정보과학회, 제37권, 제2호, 2010, pp. 181-182.

조성룡, "데이터 환경의 변화와 분산 데이터베이스 시스템," 정보과학회지, 제30권, 제5호, 2012, pp. 21-28.

최민석, "홍보 효과 증진을 위한 페이스북 팬페이지 분석 시스템 개발," 디지털융복합연구, 제13권, 제12호, 2015, pp. 135-142.

최민아, 강동옥, 강지훈, 이광근, “자바스크립트의 성질과 핵심 언어 정의,”, 한국정보

과학회 학술발표논문집, 2013, pp. 1592-1594.

Black, E. W., “Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication?,” *Online Information Revie*, Vol. 32, No. 1, 2008, pp. 73~88.

Cachia, R., R. Compano, and O. D. Costa, “Grasping the potential of online social networks for foresight,” *Technological Forecasting and Social Change*, Vol. 74, No. 8, 2007, pp. 1179~1203.

Kam M. and Song M., “A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis”, *Journal of intelligence and information systems*, Vol. 8, No. 3, pp. 53-77, 2012.

Lee J, Le HS, Lee H. “Research on Methods for Processing Nonstandard Korean Words on Social Network Services,” *Journal of the Korea Industrial Information Systems Research*, Vol. 21, No. 3, 2016, pp. 35-46.

Lee, J. H. and Lee, H. K., “A Study on Unstructured Text Mining Algorithm through R Programming based on Data Dictionary,” *Journal of the Korea Society Industrial Information System*, Vol. 20, No. 2, 2015, pp. 113-124.

http://github.com/

http://www.etnews.com/20161006000003

http://www.index.go.kr

http://www.worldometers.info/kr/

https://wikipedia.org/

https://www.etri.re.kr/