신문 기사 분석을 통한 연관어 비주얼라이저

김현진, 문성영, 정용기, 이정준
한국산업기술대학교 컴퓨터공학과
e-mail: khjnkhjn@hanmail.net, o2i2o2i2@naver.com,
penter379@naver.com, jjlee@kpu.ac.kr

Visualizer of Associated Word by Analyzing News Articles

Hyun-Jin Kim, Sung-Young Moon, Yong-Gi Jeong, Jeong-Joon Lee
Dept. of Computer Engineering, Korea Polytechnic University

요 약

신문기사 분석을 통한 연관어 비주얼라이저는 신문 기사의 단어를 추출하여 단어 간 연관도를 분석하여 다양한 그래프로 표현하는 시스템이다. 인터넷 신문사의 뉴스 기사들을 수집하고 형태소 분석을 통해 기사별로 단어의 출현 횟수를 데이터베이스에 저장하고 단어와 단어 간의 연관성을 분석한다. 단어 간 연관성을 측정하기 위한 기준으로 두 단어 간 동일기사에 존재여부, 동일날짜에 존재여부를 이용한다. 이 값을 바탕으로 웹 페이지 상에서 다양한 그래프로 상위 연관성을 가진 단어들을 표현한다. 표현 되는 그래프는 다양한 형태의 그래프로 단어와 단어사이에 연관성을 보다 쉽게 파악 할 수 있다.

1. 서론

최근 몇 년 사이 빅 데이터가 빠른 시간 안에 처리 가능해 짐에 따라 다양한 주제의 빅 데이터를 이용하여 의미 있는 분석결과를 얻어 내는 것이 중요한 이슈로 자리 잡았다. 분석 시 각 데이터 자체의 의미를 넘어서 데이터와 데이터 간의 연관성을 나타내는 의미는 중요한 부분으로 인식 되고 있고 데이터 간에 연관성을 찾기 위한 많은 분석 방법이 연구되고 있다. 하지만 기존 연관 검색어나 일반적인 검색으로는 과련 정보의 연관성을 포괄적이고 깊게 알기 힘들다[1].

본 논문은 신문기사를 이용해 기사에 출현한 단어 간의 연관성을 분석하여 분석결과를 다양한 그래프로 표현하는 시스템의 설계 및 구현에 대해 설명하고자 한다.

신문기사 분석을 통한 연관어 비주얼라이저는 신문 기사를 이용함으로 신뢰도 있는 정보를 바탕으로 기사에 등장하는 기사의 출현횟수와 자주 함께 등장하는 단어들을 분석하여 여러 가지 그래프 형태로 분석결과를 출력한다. 이를 통해 트렌드에 대한 분석 속도를 향상시키고 최근 이슈와 사회 흐름을 보다 쉽게 파악 할 수 있다. 또 단어와 단어 사이에 연관도를 쉽게 파악하여 한 단어와 연관된 언론의 관심을 유추할 수 있도록 하였다.

2. 관련 연구

본 논문의 시스템 개발을 위한 단어 Filtering 방식과 분석 방법 인 출현 횟수 분석과 연관도 분석에 대해 간단히 설명하고자 한다.

2.1 필터링(Filtering)

형태소 분석을 통해 얻은 신문기사의 단어들 중 그 단어 자체만으로 의미를 찾기 어려운 단어들이 존재하는데 이러한 단어들은 분석결과에 영향을 미쳐 제대로 된 분석결과를 얻을 수 없게 된다. 이러한 문제를 해결하기 위해서 추출 단어 모음으로부터 데이터를 읽어 들일 때 필터를 이용하여 의미 없는 단어들을 걸러준다. 예로 들면, “것”, “말”, 등과 같이 의미 없는 명사이름이 있다.

2.2 출현 횟수 분석기(Wordcount Analyzer)

출현 횟수 분석은 한 달 분량의 신문기사에서(약 1300000개) 단어들의 누적 출현 횟수를 계산하고 누적 출현 횟수를 기준으로 내림차순으로 정렬하여 파일 형태로 결과물을 산출한다. 출현 횟수 분석을 통해 산출 된 결과물을 바탕으로 분석 기간 안에 해당 되는 이슈 단어와 단어의 추이를 파악하고 시각화 할 수 있다.

2.3 연관도 분석기(Association Analyzer)

연관도 분석은 한 달 분량의 신문기사에서(약 1300000개) 단어 간의 연관도를 분석하는 기능을 수행한다. 단어 간의 연관도를 측정하기 위해 적용한 기준으로는 같은 기사 여부와 기사 출현 날짜가 같은 가이다. 이 두 기준에 만족하는 단어 간에는 연관 가중치를 부여하여 단어들 간의 연관도를 측정하여 파일 형태로 결과물을 산출한다. 연관도 분석을 통해 산출 된 결과물을 바탕으로 단어들 간의 연관도를 쉽게 파악 할 수 있고 시각화 할 수 있다[2].

3. 시스템 설계 및 구현

3.1 개발 환경

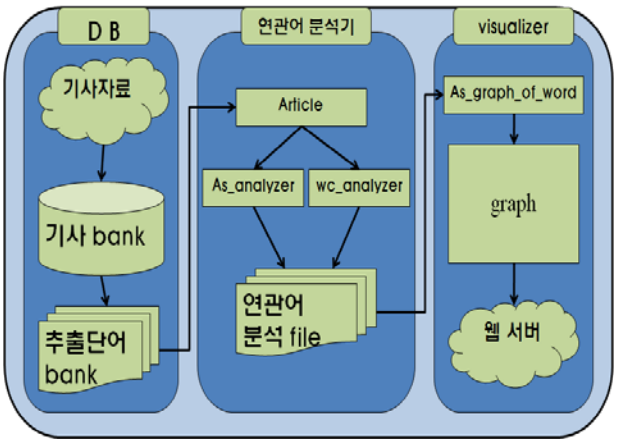
아래의 <표 1>은 본 논문 시스템의 개발 환경이다. 비주얼라이저 부분의 그래프 표현을 위해 flash와 html5를 이용하였다.

운영체제	Windows
툴	Eclipse
플랫폼	Flash, HTML15
언어	Java

<표 1> 개발 환경

3.2 시스템 동작 흐름도

본 논문의 전체 시스템 동작 흐름은 <그림 1>과 같다. 먼저 신문 기사 자료를 수집. 기사은행를 만들고 기사은행에서 형태소 분석을 통해 추출 단어 은행를 생성한다. 연관어 분석기 부분의 기사 클래스는 추출 단어 은행에서 한 달 분량의 데이터를 메모리로 읽어오고 출현 횟수 분석기와 연관도 분석기가 각각 출현 횟수 분석과 연관도 분석을 수행한다. 각 분석을 통해 결과물로 분석 파일이 생성 되고 이 분석 파일을 이용하여 비주얼라이저 부분에서 플래시와 구글 차트를 이용한 다양한 그래프를 웹에서 서비스 하게 된다.



<그림 1> 동작 흐름도

3.3 핵심 알고리즘

본 논문에서 설명하는 중요한 분석기능은 출현횟수를 분석하여 결과를 산출하는 출현 횟수 분석과 단어들 간의 연관도를 측정하여 결과를 산출하는 연관도 분석이 있다. 이 분석을 통해 산출된 결과물을 바탕으로 시각화한 부분을 웹에서 서비스하게 된다.

3.3.1 출현 횟수 분석기(Wordcount Analyzer)

출현 횟수 분석 방법은 <그림 2>와 같이 해시테이블을 이용하여 단어들의 출현 횟수를 누적 계산한다. 해시테이블은 중복을 막아주기 때문에 단어를 해시테이블에 입력하기 전에 그 단어가 테이블에 존재하는지를 파악하고 존재한다면 이미 넣어져 있는 값과 추가 출현 횟수를 더하여 누적 입력시킨다. 만약 테이블에 존재하지 않는 단어라면 그대로 입력 하게 된다. 이 과정을 거치면 단어들의 누적 출현 횟수가 해시테이블에 정의되고 이를 <그림 3>과 같이 해시테이블에 키가 아닌 값을 기준으로 정렬하여 결과를 파일에 출력하는 기능을 수행하게 된다.

```
for(int i =0;ac.get_word(i)!=null;i++){
  String key = new String(ac.get_word(i));
  if(ht.get(key)==null){
    ht.put(ac.get_word(i), ac.get_count(i));
  }
  else{
    ht.put(ac.get_word(i), (Integer)ht.get(key)+ac.get_count(i));
  }
}
```

<그림 2> 출현 횟수 분석 알고리즘

```
List<Map.Entry<String,Integer>>list
=new ArrayList<Map.Entry<String, Integer>>(sparse.entrySet());
Collections.sort(list,newComparator<Map.Entry<String, Integer>>() {
  public int compare(Map.Entry<String, Integer> e1,
                     Map.Entry<String, Integer> e2){
    if (e1.getValue() == e2.getValue())
      return e1.getKey().compareTo(e2.getKey());
    else
      return e2.getValue().compareTo(e1.getValue());
  }
});
```

<그림 3> 해시 테이블 정렬 알고리즘

3.3.2 연관도 분석기(Association Analyzer)

연관도 분석 방법은 <그림 4>와 같이 가장 먼저 모든 단어들을 해시테이블에 넣어준다. 해시테이블은 중복을 막아주기 때문에 이 해시테이블에는 모든 종류의 단어들이 들어있게 되고 이 단어들을 하나씩 꺼내어 기준단어로 정하고 분석을 하게 된다. 해시테이블에서 꺼내어 기준단어를 정하게 되면 그 단어와 같은 단어를 메모리에 올라와 있는 데이터에서 찾게 되고 찾은 단어와 같은 날짜의 단어들에게 연관 가중치를 부여하게 된다. 연관 가중치는 같은 기사에 소속된 단어가 같은 날짜를 가진 단어보다 연관성이 높다고 판단하여 차등 부여한다. 본 시스템에서는 동일 기사 소속 단어에는 5의 가중치를 같은 날짜를 가진 단어에게는 1의 가중치를 부여한다. 이러한 과정을 통해 기준 단어와 관련된 단어와의 연관도를 저장한다.

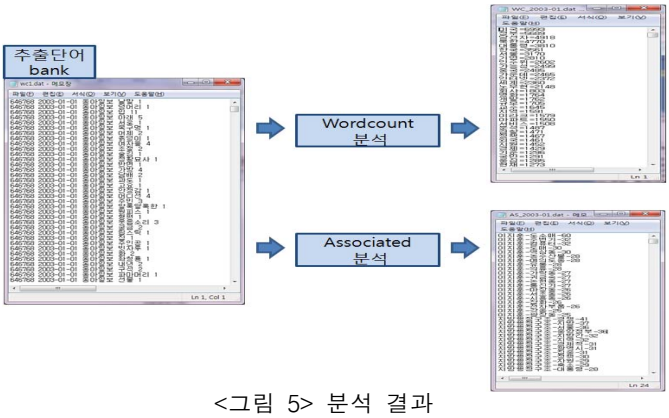
```
while(sstd.hasMoreElements()){
   Hashtable sparse=new Hashtable();
   System.out.println(s+"번째 단어 분석중");
   std = sstd.nextElement().toString();

   for(int i =0; ac.get_word(i)!=null; i++){
     if(std.equalsIgnoreCase(ac.get_word(i))){
        std_word= ac.aw[i];
        String key = new String();
        int t = Integer.parseInt(std_word.date.substring(8));
        stempdate = std_word.date;
        atempdate = ac.get_date(j);
                        if(std_word.id==ac.get_id(j)){
            if(sparse.get(key)==null){
               sparse.put(key,(weight_id+ac.get_count(j)));
            }
            else{
              sparse.put(key,
              (Integer)sparse.get(key)+(weight_id+ac.get_count(j)));
            }
        }
     }
}
```

<그림 4> 연관도 분석 알고리즘

4. 개발 내용

4.1 분석 결과

추출 단어 모음을 이용하여 출현 횟수 분석과 연관도 분석을 통해 얻은 결과파일은 <그림 5>와 같다. 추출 단어 모음에 기사번호, 기사날짜, 언론사이름, 단어, 출현횟수를 분석하게 되면 출현 횟수분석을 통해 가장 출현횟수가 높은 순서로 정렬되어 단어와 출현 횟수의 형식으로 결과물을 저장한다. 연관도 분석은 기준단어와 연관단어, 연관도의 형태로 결과물을 저장한다.



<그림 5> 분석 결과

4.2 웹 페이지 구성

4.2.1 주요 기능

웹 페이지는 크게 세 가지의 기능을 제공한다. 홈 화면에서는 트렌드 슬라이드를 제공하고 소셜 인사이트 화면에서는 노드그래프, 버블그래프, 트리그래프를 제공한다. 기타 그래프 화면에서는 구글 차트를 이용한 원그래프 트리맵 그래프 등을 제공한다.

4.2.2 각 메뉴별 화면

각 메뉴별 화면 구성은 <그림 6>와 같다. (a)는 홈 화면으로 출현횟수 분석을 통한 트렌드 슬라이드를 제공 한다. 최근 한 달간 가장 많이 출현한 단어들에 대한 정보와 사진을 제공하고 검색이나 선택을 통해 해당 키워 드와 관련된 다른 그래프를 볼수 있는 화면으로 이동한다.(b)는 어바웃 화면으로 개발자의 정보와 간단한 설명을 볼수 있다. (c)는 소셜 인사이트 화면으로 검색을 통해 노드그래프, 버블그래프, 트리그래프를 볼수 있는 기능을 제공한다. (d)는 기타 그래프 화면으로 구글차트를 이용하여 원그래프, 트리맵그래프, 꺽은선그래프, 테이블차트를 제공한다 [3][4].

(a) 홈 화면 (b) 어바웃 화면

(c) 소셜 인사이트 화면 (d) 기타그래프 화면

<그림 6> 화면 구성

4.2.3 주요 그래프

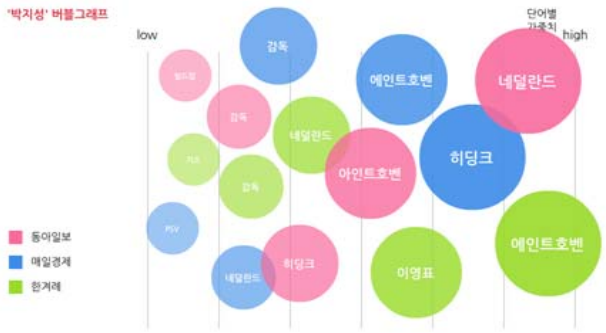
<그림 7>는 참고문헌[5, 6, 7]를 참고하여 구성한 노드그래프와 테이블이다. 노드 그래프로 중심키워드와 연관단어에 대해 연관도에 따라 거리를 차등 적용하고 연관도 테

이블을 이용해 단어 간 연관도를 보여준다[5, 6, 7].

No.	연관어	연관도
1.	네덜란드	892
2.	히딩크	851
3.	감독	806
4.	이영표	711
5.	거스	651
6.	월드컵	639
7.	에인트호벤	521
8.	계약	493
9.	아인트호벤	466
10.	PSV	460
11.	연봉	343
12.	진출	296
13.	프로축구	291
14.	구단	245
15.	송종국	201

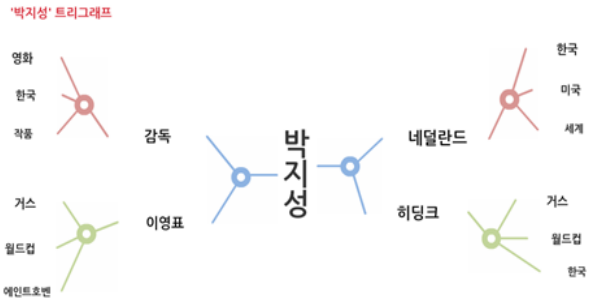
<그림 7> 노드 그래프와 연관도 테이블

<그림 8>는 참고문헌 [8, 9]을 참고하여 구성한 버블 그래프이다. 버블 그래프로 중심키워드와 연관단어에 대해 언론사 별로 연관도를 표현해준다[8, 9].



<그림 8> 버블 그래프

<그림 9>은 참고문헌[10]을 참고하여 구성한 트리그래프이다. 트리 그래프로 중심단어와 연관단어를 표현하고 연관단어의 연관단어까지 2-level의 연관단어들을 보여준다[10].



<그림 9> 트리 그래프

5. 결 론

본 논문에서는 신문 기사 분석을 통한 연관어 비주얼라이저의 설계 및 구현에 대해 설명하였다. 이를 위하여 다양한 분석 기법을 참고하여 단어들의 출현 횟수를 분석하는 출현 횟수 분석과 단어들 간에 연관도를 분석하는 연관도 분석을 구현하여 신문기사를 분석 하였고 분석 결과를 바탕으로 트렌드 슬라이드, 노드그래프, 버블그래프, 트리그래프, 구글차트를 이용한 그래프를 웹에서 검색을 통해 확인할 수 있도록 구현 하였다.

본 시스템을 활용하면 보다 쉽게 최근 이슈와 사회 흐름을 파악 할 수 있고 단어들 간에 연관성을 표현하는 그래프를 통해 단어와 연관된 대중들의 의견을 유추할 수 있을 것으로 사료된다. 또한 연관도 측정방법에 대한 검증과 연관도 측정을 위한 새로운 기준 도입에 대한 연구가 진행 될 것이다.

참고문헌

[1] 이규명의 경제학-빅 데이터(Bigdata)의 경제학, http://news.lec.co.kr/gisaView/detailView.html?gisaCode=L0010002007470006&tblName=tblNews

[2] 다음소프트, http://www.daumsoft.com

[3] 구글 차트, https://developers.google.com/chart/

[4] J.D.Gauchat, HTML5 CSS3자바스크립트의 정석 에어콘 출판 2012.07.23

[5] 국회 본회의 표결 시각화, http://politiz.org

[6] OEDC 행복 지수, http://oecdbetterlifeindex.org

[7] 소셜 메트릭스, http://insight.som.co.kr

[8] 유럽의 에너지 시각화, http://energy.publicdata.eu

[9] 세금 정보 시각화, http://wheredoesmymoneygo.org

[10] Nathan Yau, 비주얼 라이즈 디스, 에어콘 출판 2012.04.26