Journal of Korea Multimedia Society Vol. 17, No. 6, June 2014(pp. 671-680)
http://dx.doi.org/10.9717/kmms.2014.17.6.671

SNS상의 비정형 빅데이터로부터 감성정보 추출 기법

백봉현†, 하일규††, 안병철†††

An Extraction Method of Sentiment Infromation from Unstructed Big Data on SNS

Bong-Hyun Back†, Ilkyu Ha††, ByoungChul Ahn†††

ABSTRACT

Recently, with the remarkable increase of social network services, it is necessary to extract interesting information from lots of data about various individual opinions and preferences on SNS(Social Network Service). The sentiment information can be applied to various fields of society such as politics, public opinions, economics, personal services and entertainments. To extract sentiment information, it is necessary to use processing techniques that store a large amount of SNS data, extract meaningful data from them, and search the sentiment information. This paper proposes an efficient method to extract sentiment information from various unstructured big data on social networks using HDFS(Hadoop Distributed File System) platform and MapReduce functions. In experiments, the proposed method collects and stacks data steadily as the number of data is increased. When the proposed functions are applied to sentiment analysis, the system keeps load balancing and the analysis results are very close to the results of manual work.

Key words: Big Data, Sentiment Analysis, SNS, Unstructed data analysis

1. 서 론

최근 들어 소셜 네트워크 활성화로 SNS(Social Networking Service)에서 발생하는 대량의 데이터로부터 정보를 추출하여 이를 정치, 경제, 개인 서비스, 연애 등 다양한 분야에 활용하고자 하는 노력이 계속되고 있다. SNS 상의 데이터를 빠르게 분석하여 의미있는 정보를 추출하고, 이를 통해 대중들이 요구하는 의견과 생각들을 실시간으로 파악하여, 제품을 생산하고 서비스를 제공하는 다양한 분야에서 활용할 수 있도록 하는 기술이 필요하다. 또한 이러한 정제된 유효하고 다양한 정보들을 빅데이타 처리 분석 기술을 통해 보다 효율적으로 관리하고 시각화하는 기술도 필요하다. 따라서, 본 연구에서는 소셜네트워크에서 발생되는 다양한 데이터, 특히 비정형 데이터를 효율적으로 처리할 수 있는 빅데이터 처리플랫폼을 제안한다.

SNS 서비스는 정보 전달 대상 간의 상호 연결 방법이 용이하며, 데이터 작성 형식이 비교적 자유롭기 때문에 발생하는 데이터는 대부분 비정형 데이터(unstructured data)이다. 비정형 데이터는 숫자 데이터와 달리 그림이나 영상, 문서처럼 형태와 구조가 복잡해 정형화되지 않은 데이터로 정의할 수 있다[1]. SNS 상에서 발생하는 수많은 비정형 데이터로

※ Corresponding Author : ByoungChul Ahn, Address: (712-749) 280 Daehak-Ro, Gyeongsan, Gyeongbuk, Republic of Korea, TEL : +82-53-810-2556, FAX : +82-53-810-4630, E-mail : b.ahn@yu.ac.kr
Receipt date : Jan. 28, 2014, Revision date : Apr. 17, 2014
Approval date : May. 12, 2014

† Dept. of Computer Engineering, Yeungnam University (E-mail : wefbbh@naver.com)
†† Dept. of Computer Engineering, Yeungnam University (E-mail : ilkyuha@ynu.ac.kr)
††† Dept. of Computer Engineering, Yeungnam University

부터 의미있는 정보를 추출하기 위해서는 우선 비정형 데이터에 대한 처리가 필요하다. 비정형 데이터 분석은 형태소 분석을 기반으로 다양한 분석 방법들이 연구되고 있다[2-4]. 그러나 다양한 방송 매체와 젊은 계층들로부터 새로운 유행어와 협의되지 않은 단어는 컴퓨터를 통한 언어 분석과 감성 분석이 어려워지고 있고 이에 대한 유효성 검정이 더욱 어려워지고 있다.

현재, 빅데이타 처리를 위한 다양한 오픈소스 프로젝트들을 하둡에코시스템(Hadoop ECO system)[5]을 사용한다. 빅데이터 처리에 사용되는 데이터베이스는 전통적인 관계형 데이터베이스보다 덜 제한적인 일관성 모델을 이용하는 데이터 저장 및 검색을 위해 NoSQL(Not-Only SQL)[6]을 이용한다. 현재 업계 및 학계에서 NoSQL데이터베이스에 관한 많은 연구가 진행되고 있으며, 구글의 BigTable[7], 아마존의 Amazon DynamoDB[8], 오픈소스 프로젝트의 Apache HBase[9], Cassandra[5], MongoDB[10] 등이 대표적이다.

특히 본 연구에서 사용되고 있는 MongoDB는 CAP(Consistency, Availability, Partition tolerance) 이론에 따라 데이터베이스를 분류하였을 때 Consistency와 Partition tolerance를 만족하는 CP형 데이터베이스로서, 현재 오픈소스 프로젝트로 진행되고 있으며, key-value의 방식으로 JSON(JavaScript object notation)형태의 문서데이터를 저장한다. 이는 스키마가 없으며, 정규표현 검색 및 배열데이터의 특정값 포함여부 등의 검색조건 등에 유연하게 대응할 수 있다. 전통적인 RDBMS에 비하여 대량의 데이터를 병렬로 처리할 수 있으며 MapReduce 기법 등을 사용하여 데이터 클러스터링 연산, 통계, 데이터 추출 및 필터링이 가능하다.

감성 분석(sentiment analysis)은 자연언어처리와 전산언어학 그리고 텍스트 분석론을 활용하여 원자료에서 주관적인 정보를 발견하고 추출하는 과정이다[11]. 빅데이터로부터 사용자의 감성을 분석하기 위한 연구가 진행되고 있다[12-15]. 감성의 종류를 분석하고 분류하는 작업은 크게 세 가지의 단계로 나눌 수 있다. 첫 번째 단계는 감성 정보가 들어 있는 주관적인 생각이나 느낌을 표현하는 문장을 추출하고, 다음 단계에서 문서 또는 문장의 극성(긍정, 부정)을 나눈다. 마지막 단계는 문서 또는 문장이 어느 정도의 주관성을 갖는지 그 강도를 구하는 강도 분류이다[16,17].

2. 비정형 SNS 감성 데이터 분석 방법 제안

2.1 시스템 구성

본 연구에서는 다양한 대용량 SNS 데이터로부터 데이터를 안정적으로 수집하고 저장하기 위한 하둡에코시스템을 기반으로한 병렬적 HDFS(Hadoop distributed file system)을 사용하고, 대량의 비정형 데이터를 분석하여 사용자의 감성을 효과적으로 분석할 수 있는 MapReduce[18]기반의 감성분석 알고리즘 및 사전을 제안한다. 전체적인 시스템의 구성은 그림 2와 같다. 제안된 시스템은 Hadoop EcoSystem

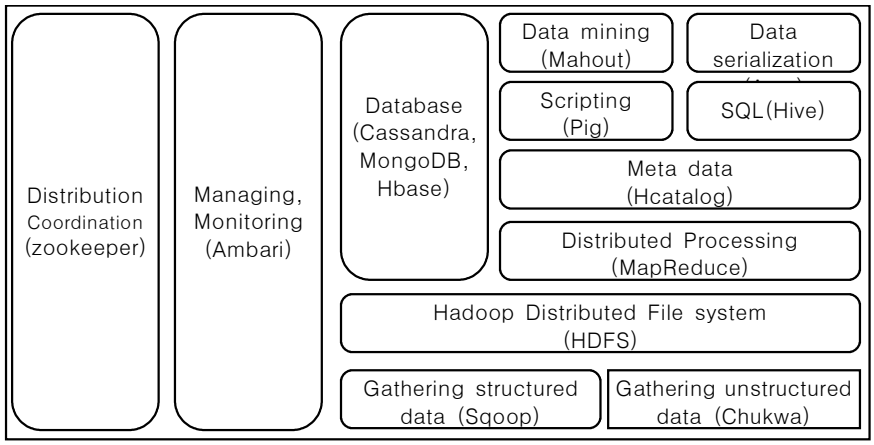

Fig. 1. Hadoop ECO System.

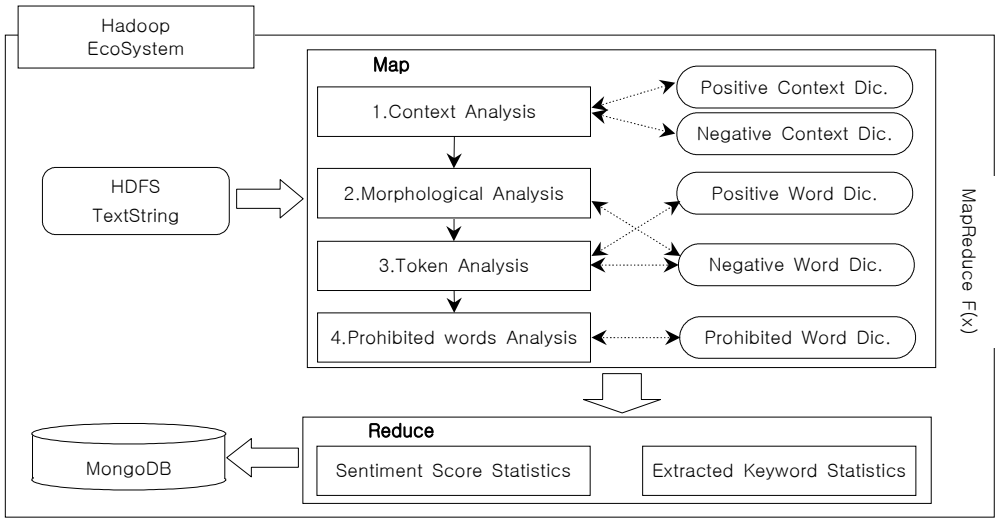

Fig. 2. The Proposed System.

을 기반으로 하고 있으며, HDFS는 SNS 사이트의 API를 통해 전달받은 데이터를 효율적으로 분산처리하여 적재하는 역할을 한다. 적재된 텍스트 형태의 자료는 MapReduce기반의 제안된 함수에 의해 4단계에 걸쳐 감성분석이 이루어진다. 감성분석 시에는 제안된 5종류의 감성분석사전을 참조하여 보다 정확한 감성분석이 이루어질 수 있도록 한다. 감성분석 결과는 다양한 통계를 위해 사용될 수 있으며, MongoDB에 저장된다.

2.2 HDFS의 구성

HDFS는 분산처리구조의 파일처리 시스템이다. HDFS는 입력받은 대량의 데이터를 적절하게 분산하여 적재하는 역할을 한다. 본 연구에서 제안하는 HDFS는 그림 3과 같이 구성하였다. 이 시스템은 리눅스 기반의 4대의 서버로 병렬로 연결되며, 각각 데이터를 저장하기 위한 Node들의 chunk는 64MB로 구성되며, 장애 복구를 위해 NFS를 이용한 네임서버를 이중화한다. 구성된 서버의 기능은 표 1과 같다.

2.3 MapReduce 함수 구성

MapReduce는 분산 컴퓨팅을 지원하기위해 구글에서 개발한 소프트웨어 프레임워크로 맵(Map)과 리듀스(Reduce)라는 함수의 개념을 이용하여 병렬 프로그래밍을 가능하게 한다. 본 연구에서는 맵 함수는 4개로 분류되며, 감성분석을 위한 각 단계에 적용

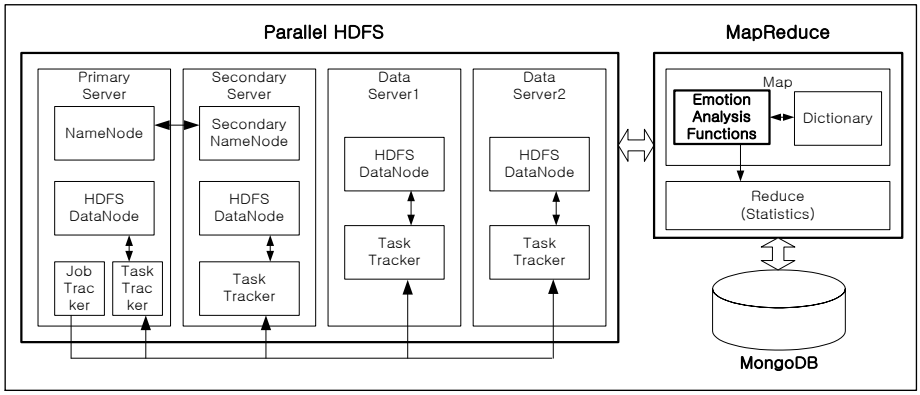

Fig. 3. The proposed HDFS.

Table 1. Servers of the proposed HDFS

Server	Components	Role
Primary Server (Master Node)	NameNode, DataNode MapReduce, Crawler	Main server for parallel distribution process Name node(controlling other servers) Data node, Data loading
Secondary Server (Slave Node 1)	Secondary NameNode DataNode	Backup server of main server Data node, Data loading
Data Server1 (Slave Node 2)	DataNode	Data node, Data loading
Data Server2 (Slave Node 3)	DataNode	Data node, Data loading

된다. 제안된 맵함수는 긍정/부정 문맥 분석, 형태소 분석, 토근 분석, 금칙어 분석 함수이다. 각 함수의 상세한 역할은 아래와 같다.

첫째, 긍정/부정 문맥 분석 함수이다. 이 함수는 먼저 정확도를 높이기 위해 한 문장 단위로 문맥을 검사하고, 긍정문맥 사전과 부정문맥 사전을 이용하여 패턴(정규식)매칭을 실시한 후, 원본 자료(트위트 text)를 긍정과 부정으로 카운팅하며, 긍정과 부정의 카운트가 동일하면 긍정으로 처리하고 판단 불가일 경우 형태소 분석으로 이관한다. 문맥분석을 위한 알고리즘은 그림 4와 같다.

둘째, 형태소 분석 함수이다. 이 함수에서는 한나눔의 한글형태소 분석기를 이용하여 링크, 특수기호 등 분석에 불필요한 요소를 제거한 후, 긍정어절과 부정어절 사전을 비교하여 각각의 카운터를 계산한다. 또한 긍정 또는 부정 카운터의 수치가 동일하다면 긍정으로 처리, 판단 불가한 상태라면 토큰분석으로 이관한다.

```
//Context Analysis
public double calculateSentenceScore(String keyword, String source)
{
    keyword = keyword.toLowerCase();
    source = source.toLowerCase();
    // processing of the keyword
    source = source.replaceAll(keyword, "#KW#");

    int pCount = 0;
    int nCount = 0;
    // context analysis by minimum sentence unit
    List<String> lst = getTwitterSentences(source);
    for (String s : lst) {
        pCount += getPostiveCount(keyword , s);
        nCount += getNegativeCount(keyword , s);
    }
    if (pCount == 0 && nCount == 0) {  return 0.0; }
    int result = pCount - nCount ;
    // if the result is zero then assign 1 to the result(positive)
    if (result == 0) { result = 1;  }
    return result;
}
```

Fig. 4. Context Analysis Algorithm

셋째, 토튼 분석 함수이다. 이 함수는 원본자료(트위트 text)의 토큰을 공백으로 분리하고, 한글 형태소 분석기를 사용하여 형태소를 분석한 결과와 긍정어절과 부정어절 사전을 비교하여 각각의 카운터를 계산하며, 긍정 또는 부정카운터의 수치가 동일하다면 긍정으로 처리, 판단 불가한 상태라면 금칙어 분석으로 이관한다.

넷째, 금칙어 분석 함수이다. 이 함수에서는 최종 분석 단계로 상위의 과정에서 분석이 이루어지지 못했을 경우 금칙어 사전을 기반으로 금칙어스코어 계산 한다. 형태소 분석, 토큰 분석 및 금칙어 분석을 위한 알고리즘은 그림 5와 같다.

2.4 감성분석 사전

제안한 MapReduce 함수에서 사용하는 사전은 모두 5가지 종류이다. 즉, 긍정어(em_positive), 부정어(em_negative), 긍정문맥(p_context_pattern), 부정문맥(n_context_pattern), 금칙어(abuses) 사전이다. 제안된 감성분석 사전은 감성분석을 위한 각 맵 함수에서 사용된다. 각 사전의 역할은 표 2와 같다.

3. 실험 결과 및 고찰

3.1 비정형 SNS 데이터 수집

제안한 시스템의 성능분석을 위한 데이터 수집은 Topsy와 트위터(Twitter)를 통해 이루어졌다. Topsy는 트위터나 구글플러스 등 SNS 서비스에서 사용자의 활동을 분석해 통계 기법으로 정리해주며, 하루 5억 건의 방대한 데이터를 분석하여 제공한다. Topsy에서 제공하는 API key는 상업용 키(commercial

```
//Morphological Analysis
//if the result is zero in previous stage, this stage is processed
public double calculateSentenceScore(String source) {
    double sc = 0.0;
    List<MorphemeTag>verifyMorphemeList = cleansingWord(source);
    double posIdx = getMorphemeScore(verifyMorphemeList,
                    MorphemeCalculator.s_posDic);
    double negIdx =  getMorphemeScore(verifyMorphemeList,
                     MorphemeCalculator.s_negDic);
    if (posIdx == 0 && negIdx == 0){
      sc = 0.0;
    }
    else{
      sc = posIdx - negIdx;
      // is the sc is zero, then assign the posIdx value to the sc
      //value as a positive value
      if (sc == 0) { sc = posIdx;  }
    }

// Token Analysis
    if (sc == 0.0)
    {
       List<String> tokenList = Arrays.asList(source.split(" "));
       List<MorphemeTag>verifyMorphemeList= leansingWord(source);
            .....

       if (posIdx == 0 && negIdx == 0){
           sc = 0.0;
       }
       else{
          sc = posIdx - negIdx;
          if (sc == 0) {  sc = posIdx;   }
       }
    }

//Prohibited word Analysis
    if (sc == 0){
       List<MorphemeTag>verifyMorphemeList = cleansingWord(source);
       for (MorphemeTag tag : verifyMorphemeList)
       {                    .....              }
    }
    return sc;
}
```

Fig. 5. Morphological, Token and Prohibited word Algorithms.

key)가 아닌 일반용 키는 하루에 최대 7,000 쿼리를 요청할 수 있으며, 과거 데이터의 수집이 완료된 후 지속적인 증분 데이타의 경우 Twitter4j를 이용하여 트위터에서 제공하는 데이터를 수집한다. 트위터 제공 데이터는 현시점으로부터 최대 1주일 과거 데이터만을 수집 가능하며, 트위트에서 제공되는 키(토큰)는 15분 동안 450개의 쿼리를 사용할 수 있다. 본 연구에서는 크론을 통해 매 4시간마다 데이터 수집 모듈을 실행하도록 하였다. 그림 6은 Twitter4j를 이용하여 트위터로부터 데이터를 수집하는 과정을 보여주고 있다.

3.2 실험 환경

제안 시스템의 성능분석을 위한 환경은 표 3과 같다. 실험환경은 4대의 서버를 하둡기반의 병렬시스템으로 구성하였으며, 사용 운영체제는 CentOS 6.3×64를 사용하였다.

3.3 실험 분석 및 평가

제안 시스템의 성능분석을 위해 아래와 같은 4 가지의 테스트를 진행하였다.

첫째, 데이터 량에 따른 시스템성능 실험이다. Topsy API를 통해 수집된 표 4와 같은 7개 셋트의 실제 트위터 데이터에 대해 수집시간 및 처리 시 시스템의 부하를 테스트하였다.

그림 7은 데이터 셋트별 크롤링 시간과 HDFS 적재시간을 비교한 것이다. 그림 8과 그림 9는 데이터 셋트별 크롤링과 HDFS 적재시 각 노드별 메모리 부하와 CPU부하를 나타낸 것이다. 각 데이터 셋트의 데이터 건수에 대하여 크롤링 시간과 HDFS 적재시간이 비례하여 안정적으로 증가하는 것을 볼 수 있다. 따라서 제안된 시스템에서 데이터를 수집하여 적재하는데 네트워크 부하나 시스템부하는 미비하며, 수초~수분의 시간 내에 안정적인 데이터 수집과 적재가 가능함을 알 수 있다.

Table 2. Dictionaries for Emotion Analysis

Dictionary	Role	application
Positive Context Dictionary	set of positive context patterns, compute the number of positive context in source sentence	Context Analysis
Negative Context Dictionary	set of negative context patterns, compute the number of negative context in source sentence	〃
Positive Word Dictionary	set of positive word patterns, compute the number of positive word in source sentence	Morphological/ Token Analysis
Negative Word Dictionary	set of negative word patterns, compute the number of negative word in source sentence	〃
Prohibited Word Dictionary	set of prohibited words	Prohibited Word Analysis

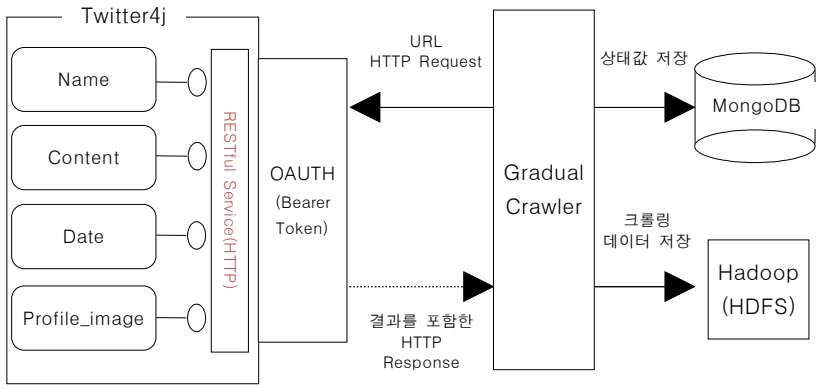

Fig. 6. Data Crawling using Twitter4j.

Table 3. Experimental Environment

Components	Roles
OS, RE	Use of Hadoop for distributed storage, Supporting Java environment for processing some business logic
Crawler, HDFS Layer	Crawler: Gathering the source data from various SNSs HDFS: Distribution File system, Data storage
MapReduce Layer	Sentence Analysis, Text Mining, Emotion Analysis
MongoDB	Storing analyzed results by MapReduce in MongoDB
WAS, Web Server	Supporting Web applications using analyzed results

Table 4. Data Sets for experiment and analysis

Data Set number	Number of data	extraction period (day)	API
1	2,106	1	Topsy API
2	11,672	6	〃
3	20,000	10	〃
4	40,788	20	〃
5	79,080	36	〃
6	90,014	44	〃
7	100,497	52	〃

그림 8과 같이 슬레이브 노드 1(SN1)에서 3(SN3)까지의 노드들은 메모리 사용량이 최소 0.03%에서 최대 3.93% 사용한 것으로 나타났고, 마스트 노드(M)의 경우 최소 0.6%에서 최대 7.31%를 사용한 것으로 나타났다. 슬레이브 노드의 경우 데이터를 분산 적재함으로써 메모리부하가 낮게 나왔으며, 마스터 노드의 경우 전체 슬레이브 노드의 분산 데이터처리를 위해 슬레이브 노드보다 약 2배가량 더 많은 메모리 자원을 사용하는 것으로 나타났다.

그림 9과 같이 슬레이브 노드 1(SN1)과 2(SN2)의 경우 최소 0.0%에서 최대 2.8%의 CPU 사용량을 나타냈으나, 슬레이브 노드 3(SN3)의 경우 최소 0.0%에서 최대 11.4%의 CPU사용량을 나타냈다. 이는 병렬분산 처리 시 HDFS시스템의 자동병렬처리 과정에서 슬레이브 노드 3을 주로 사용하기 때문인 것으로 파악된다. 마스트 노드의 경우 최소 5.0%에서 최대 7.9%의 CPU사용량을 나타냈다. 따라서 CPU 사용량에 있어서도 제안된 시스템은 데이터의 수집과 적재 시 안정적인 환경을 제공하는 것으로 나타났다.

둘째, 데이터량에 따른 감성분석 시간과 시스템

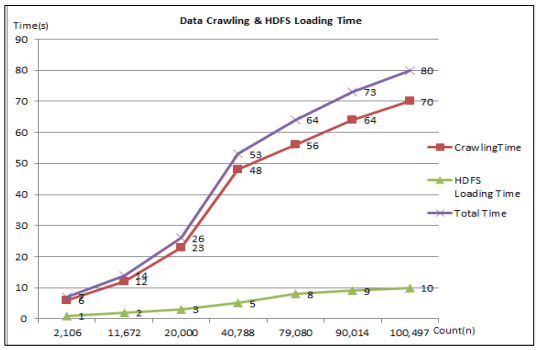

Fig. 7. Crawling Time and HDFS Loading Time.

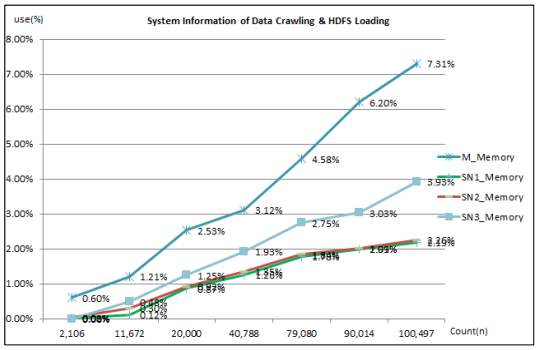

Fig. 8. Memory Consumption of Data Crawling and HDFS Loading.

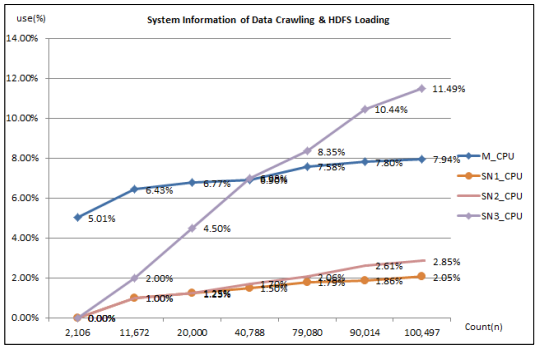

Fig. 9. CPU Consumption of Data Crawling and HDFS Loading.

부하 테스트이다. 첫 번째 실험에서 사용한 데이터 셋을 가지고 감성 분석 시 소요되는 분석시간과 시스템 부하정도를 실험하였다. 그림 10은 데이터 셋트별 감성분석 소요시간을 비교한 것이고 그림 11과 그림 12는 감성분석 시 각 노드별 CPU부하와 메모리부하를 각각 비교한 것이다. 그림 10과 같이 7개의 데이터 셋트들에 대해 감성분석 시간은 35초에서 68초의 시

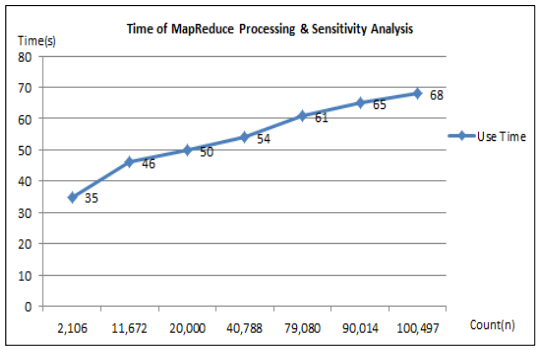

Fig. 10. Time of MapReduce Processing for Sentiment Analysis.

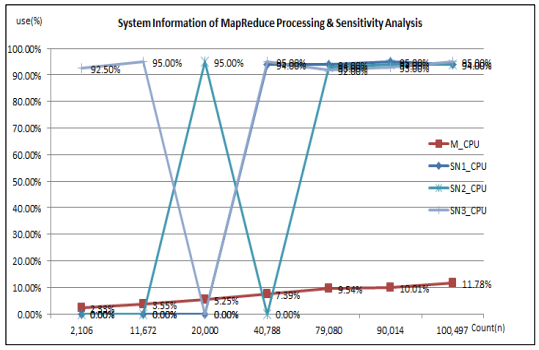

Fig. 11. CPU Consumption of MapReduce Processing for Sentiment Analysis.

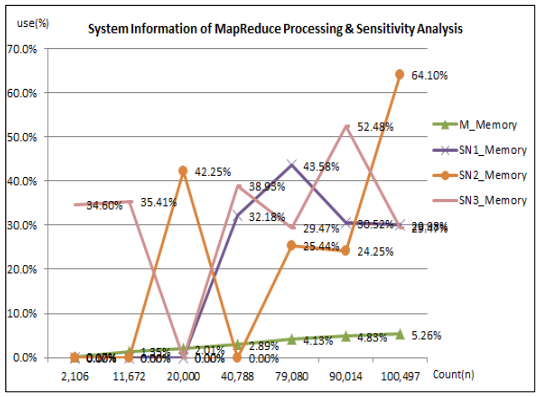

Fig. 12. Memory Consumption of MapReduce Processing for Sentiment Analysis.

간이 소요되어 데이터건수와 비례하여 감성분석 시간이 안정적으로 증가하는 것으로 나타났다.

그림 11에서 마스트 노드(M)는 실제 분석처리를 하지 않고 하위 슬레이브 노드를 관리하므로 CPU의 사용량이 낮은 반면, 슬레이브 노드는 분석을 하므로

CPU자원을 사용한다. 데이터 건수 40,788인 데이터 셋트까지는 각 슬레이브 노드가 데이터를 상호 병렬처리하는 것으로 나타났으나, 데이터 건수 79,080인 데이터 셋트 이후는 분산된 데이터량이 많아짐에 따라 모든 슬레이브 노드들이 CPU를 최대로 사용한다. 따라서 제안된 시스템은 일정한 수준까지 안정적인 상호병렬처리가 이루어지는 것으로 볼 수 있다. 그림 12도 그림11과 같이 슬레이브 노드가 분석을 병렬처리를 하므로 많은 메모리를 사용한다.

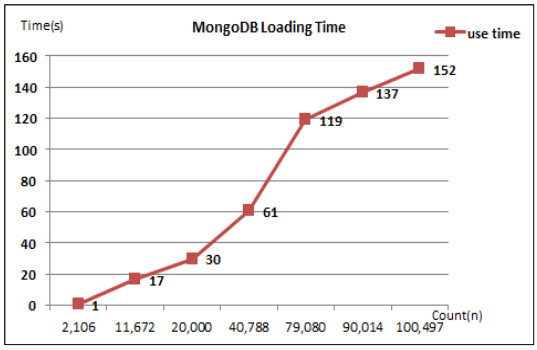

Fig. 13. Time of Data Loading to MongoDB.

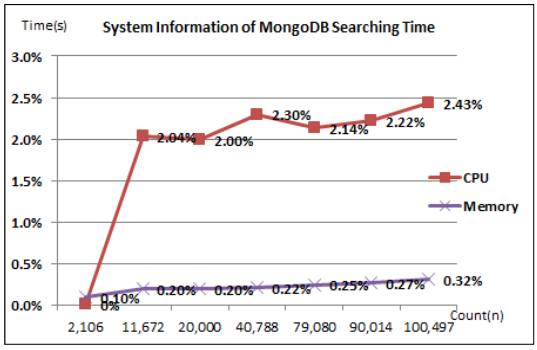

Fig. 14. Load Valance in MongoDB searching.

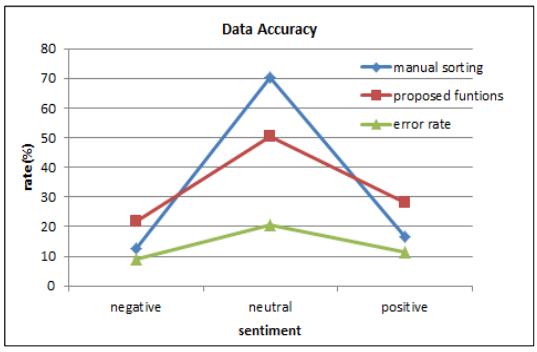

Fig. 15. Data Accuracy.

따라서, 제안된 시스템과 알고리즘은 자원할당 측면에서 단일 노드에만 시스템부하가 집중되지 않고 상호 병렬처리됨으로써 안정적인 병렬 분석 환경을 제공하는 것으로 나타났다.

셋째, 데이터 조회시 처리시간과 인덱싱 사용유무에 따른 속도비교 테스트이다. 첫 번째 실험에서 사용한 데이터 셋에 대하여 MongoDB에 저장된 데이터를 적재 및 조회할 때 소요되는 시간과 시스템의 부하 테스트를 진행하였다. 그림 13은 데이터 셋트별 MongoDB에 데이터가 적재되는 시간을 비교한 것이고, 그림 15는 MongoDB조회시 시스템의 부하를 비교한 것이다.

그림 13과 같이 데이터 셋트의 데이터 건수에 비례하여 MongoDB에 적재되는 시간이 증가하였다. 그림 14에서는 MongoDB에 적재된 데이터 조회시 메모리 사용량은 데이터 건수에 비해 큰 차이가 없는 것으로 나타났고, CPU사용량의 경우 2%정도, 메모리 사용량 0.2% 정도로 데이터 조회시 시스템 I/O에 의한 성능저하는 없는 것으로 판단할 수 있다.

마지막으로 감성분석 결과의 정확도 및 오차 비교에 관한 테스트이다. “애플”을 키워드로 하여 추출한 2,106건의 데이터 셋에 대해 긍정 또는 부정의 대한 정확도를 제안된 시스템 분석 결과와 인간이 직접 지각으로 느끼는 정도를 비교 분석하였다. 그림 15는 제안 시스템의 감성분석(형태소 분석) 결과와 인간의 직시적인 감성분석 결과를 비교한 것이다. 그림 15와 같이 중립 감성의 경우 상대적으로 오차가 높았다. 부정과 긍정의 감성인 경우 비교적 오차가 적었으나, 제안된 시스템에 의한 감성분석결과가 인간의 직시적인 감성 분석결과에 상당히 근접하고 있음을 알 수 있다.

4. 결 론

본 연구에서는 SNS로부터 발생하는 대량의 비정형 데이터로부터 사용자의 감성을 분석할 수 있는 빅데이터처리 시스템과 알고리즘을 제안하였다. 제안된 시스템은 하둡 에코시스템기반 병렬 HDFS시스템을 구성하며, MapReduce로 4개의 주요기능을 가진 함수를 구성하였다. 또한 감성분석을 위한 5가지 종류의 데이터 사전을 사용하였다. 제안한 시스템은 실험을 통해 다음과 같은 결론을 얻었다. 첫째,

데이터량에 따른 시스템성능은 데이터를 수집하여 적재하는데 네트워크 부하나 시스템부하는 미비하며, 수초~수분의 시간 내에 안정적인 데이터 수집과 적재가 가능함을 알 수 있었다. 둘째, 데이터량에 따른 감성분석 시간과 시스템 부하 실험에서 데이터 건수에 따라단일 노드에만 시스템부하가 집중되지 않고 상호 병렬처리됨으로써 안정적인 병렬 분석 환경을 제공하는 것으로 나타났다. 셋째, 데이터 조회 시 처리시간과 인덱싱 사용유무에 따른 속도비교 실험에서 데이터 건수가 증가함에 따라 인덱싱의 사용이 유용함을 알 수 있었다. DB에 적재된 데이터 조회 시 메모리 사용량은 데이터 건수에 비해 큰 차이가 없는 것으로 나타났고, 데이터 조회시 시스템 I/O에 의한 성능저하도 없는 것으로 판단할 수 있었다. 마지막으로 감성분석 결과의 정확도 및 오차 비교에 관한 실험에서 제안된 시스템에 의한 감성분석결과가 인간의 직시적인 감성 분석결과에 상당히 근접하고 있음을 알 수 있었다.

REFERENCE

[1] Big Data: The Next Frontier for Innovation, Competition, and Productivity(2011), http://www. mckinsey.com/ (accessed Jan., 27, 2014)

[2] C.S. Lee and M.H. Wang, "Automated Ontology Construction for Unstructured Text Documents," *Data & Knowledge Engineering*, Vol. 60, Issue 3, pp. 547-566, 2007.

[3] B. Lee, J. Lim, and J. Yoo, "Utilization of Social Media Analysis using Big Data," *Journal of the Korea Contents Association*, Vol. 13, No. 2, pp. 211-219, 2013.

[4] M. Song and S. Kim, "A Study of Improving on Prediction Model by Analyzing Method Big data," *The Journal of Digital Policy & Management*, Vol. 11, No. 6, pp. 103-112, 2013.

[5] Hadoop ECO system, http://www.revelytix.com/?q=content/hadoop-ecosystem (accessed Jan., 27, 2014)

[6] J. Han and K. Du, "Survey on NoSQL Database," *Proceeding of 6th International Conference on Pervasive Computing and Applications*, pp. 363-366, 2011.

[7] F. Chang and R.E. Gruber, "Bigtable: A Distributed Storage System for Structured Data," *ACM Transactions on Computer System*, Vol. 26, Issue 2, pp.1-26, 2008.

[8] S. Sivasubramanian, "Amazon dynamoDB: A Seamlessly Scalable Non-relational Database Service," *Proceeding of the 2012 ACM SIGMOD'12*, pp. 729-730, 2012.

[9] L. George, *HBase: The Definitive Guide*, O'REILLY, Sebastopol, Calif., 2011.

[10] K. Chodorow, *MongoDB: The Definitive Guide 2nd Edition*, O'REILLY, 2013.

[11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1-135, 2008.

[12] S. Mukherjee and P. Bhattacharyya, "Sentiment Analysis in Twitter with Lightweight Discourse Analysis," *Proceeding of COLING 2012*, pp. 1847-1864, 2012.

[13] N. Godbole and S. Skiena, "Large-Scale Sentiment Analysis for News and Blogs," *Proceeding of the ICWSM'2007*, pp.1-4, 2007.

[14] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Proceeding of the LREC'2010*, pp.1320-1326, 2010.

[15] S. Kim and B. Hwang, "Propensity Analysis of Political Attitude of Twitter Users by Extracting Sentiment from Timeline," *Journal of Korea Multimedia Society*, Vol. 17, No. 1, pp. 43-51, 2014.

[16] H. Tang, S. Tan, and X. Cheng, "A Survey on Sentiment Detection of Reviews," *Expert Systems with Applications*, Vol. 36, pp. 10760-10773, 2009.

[17] S. Gilbert and N. Lynch, "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services," *ACM SIGACT*, Vol. 33, No. 2, pp. 51-59, 2002.

[18] J. Dean and S. Ghemawat, "MapReduce;

Simplified Data Processing on Large Clusters," *Communications of the ACM*, Vol. 51, No. 1, pp. 107-113, 2008.

백 봉 현

1999년 동국대학교 전자계산학과 이학사
2002년 영남대학교 컴퓨터공학과 공학석사
2014년 영남대학교 컴퓨터공학과 공학박사

2005~2009 일본 SecuAvail 시스템엔지니어
2010~현재 ㈜아르고스 대표이사
관심분야 : 빅데이터처리, 센서네트워크, 데이터마이닝, 개인정보 보호

하 일 규

1992년 영남대학교 전산공학과 학사
2001년 영남대학교 정보처리교육 전공 석사
2003년 영남대학교 컴퓨터공학과 박사

1992년~1995년 증권감독원 전산업무실
2002년~현재 영남대학교 컴퓨터공학과 강사, 객원교수
관심분야 : 센서네트워크, 소셜네트워크분석, 빅데이터처리

안 병 철

1976년 영남대학교 전자공학과 학사
1986년 오레곤주립대 전기 및 컴퓨터공학 석사
1989년 오레곤주립대 전기 및 컴퓨터공학 박사

1976년~1984년 국방과학연구소 연구원
1989년~1992년 삼성전자 수석연구원
관심분야 : 센서네트워크, 임베디드시스템, 빅데이터처리, 멀티미디어처리