기업 마케팅 전략을 위한 SNS 및 Web 데이터 분석 시스템 설계

이 병 관*, 정 은 희**, 정 이 나***

A Design of SNS and Web Data Analysis System for Company Marketing Strategy

ByungKwan Lee*, EunHee Jeong**, YiNa Jung***

요 약

본 논문에서는 기업 이미지에 타격을 줄 수 있는 부정적인 SNS와 Web 데이터를 빠르게 분석하여 기업 마케팅 전략에 활용할 수 있는 SNS 및 Web 데이터 분석 시스템을 제안한다. 본 논문에서 제안하는 시스템은 SNS 및 Web Data를 수집하는 데이터 수집 모듈(Data Collection Module), 수집된 데이터를 저장하는 HBase 모듈(Hbase Module), 수집된 데이터의 의미 분석을 수행한 후 데이터의 의미를 평가 및 분류하는 데이터 분석 모듈(Data Analysis Module) 그리고 관리자에 의해 요청된 질의어에 따라 기업과 관련된 SNS와 Web데이터를 이용하여 최적화된 Map Reduce 과정을 수행하는 PSH 모듈(Priority Scheduling Hadoop Module)로 구성된다. 본 논문은 이런 모듈들을 통하여 SNS와 Web 데이터를 보다 효율적으로 관리하여 이 분석 결과를 기업 마케팅 전략에 활용할 수 있다.

ABSTRACT

This paper proposes an SNS and Web Data Analytics System which can utilize a business marketing strategy by analyzing negative SNS and Web Data that can do great damage to a business image. It consists of the Data Collection Module collecting SNS and Web Data, the Hbase Module storing the collected data, the Data Analysis Module estimating and classifying the meaning of data after an semantic analysis of the collected data, and the PHS Module accomplishing an optimized Map Reduce by using SNS and Web data involved a Businesse. This paper can utilize this analysis result for a business marketing strategy by efficiently managing SNS and Web data with these modules.

Keywords SNS, Web data analysis, Marketing, Hadoop, Opinion mining

Ⅰ. 서 론

현재 통신의 발달로 인해 현대인들은 SNS, Mobile, Web등 다양한 통신매체와 함께 생활하며 정보를 공유하고 있다. 특히, SNS(Social Network Service) 및 Web을 통한 정보 공유는 기업 및 기업의 제품 홍보와 같은 마케팅 전략으로 사용되고 있다. 하지만, 기업의 부정 이미지 또는 제품의 단

* 제1저자 관동대학교 컴퓨터학과 교수(bklee@kd.ac.kr)
** 교신저자 강원대학교 지역경제학과 교수(jeongeh@kangwon.ac.kr)
*** 제2저자 관동대학교 전자계산공학과 석박사통합과정(lupinus07@nate.com)
접수일자 : 2013년 10 월 04 일, 수정일자 : 2013년 11 월 05 일, 심사완료일자 : 2013년 12 월 05 일

점이 공유되면 순식간에 기업의 전체 이미지에 타격을 줄 수 있을 정도로 파급력이 강력하다.

이에 본 논문에서는 SNS 및 Web 마케팅을 보다 효율적으로 활용하며, 기업의 이미지에 타격을 줄 수 있는 데이터를 빠르게 파악하고 대응할 수 있는 SNS 및 Web 데이터 분석 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 관련연구를 살펴보고, 3장에서는 본 논문에서 제안하는 SNS 및 Web 데이터 분석 시스템을 설계한다. 그리고 4장에서 결론을 맺는다.

Ⅱ. 관련연구

2.1 MapReduce

대표적인 SNS인 Facebook은 하루에 100억장 이상의 사진을 저장하고 있다고 한다. 데이터 관리 업체들은 이렇게 수많은 데이터가 업로드 되고, 저장된 데이터를 사용하기 위하여 얼마나 빨리 읽어낼 수 있을까 하는 문제점이 대두되기 시작하였다. 지금까지 데이터를 저장할 수 있는 공간은 매우 늘어났지만 이 데이터를 읽는 속도는 늘어난 공간만큼 빨라지지 않았기 때문이다.

이런 문제점을 해결하기 위하여 많은 양의 데이터를 입력 받아서 처리하는 어플리케이션을 여러 컴퓨팅 노드들로 이루어진 컴퓨팅 클러스터를 이용해서 데이터를 빠르게 처리할 수 있도록 만들어 줄 수 있는 컴퓨팅 모델인 맵리듀스가 개발되었다[1, 2, 3].

맵리듀스는 맵(map)과 리듀스(reduce)의 2가지 함수를 조합하여 데이터를 처리한다. 즉, 사용자로부터 맵리듀스에서 처리하도록 받은 일을 잡(job)이라고 하고, 이를 분산환경에 맞게 나누어 맵함수와 리듀스함수를 이용하여 잡을 분산 병렬 처리함으로써 빠르게 처리한다[4].

2.2 Opinion Mining

오피니언 마이닝이란 대량의 정보에서 의견을 뽑아 평가 하는 것이다. SNS의 대중화로 스스로 말하는 의견을 수집하기 때문에 설문지와 같이 형식적인 데이터로 의견을 조사하는 것보다 정보의 질이 높다. 오피니언 마이닝은 유용한 정보의 특징들을 추출하고 추출된 특징이 어떤 의미를 나타내는가에 대한 어휘정보도 추출된다. 추출된 특징과 의견을 나타내는 어휘가 어떤 의미로 사용되었는지 판단하고, 의견에 대한 성향이 밝혀진 의견 정보들을 요약하는 단계로 실행된다[5, 6].

Ⅲ. SNS 및 Web 데이터 분석 시스템 설계

본 논문에서는 기업 마케팅 전략을 위한 SNS 및 Web 데이터 분석 시스템을 제안한다. 제안하는 시스템은 데이터 수집 모듈, HBase 모듈, 데이터 분석 모듈, PSH(Priority Scheduling Hadoop) 모듈로 구성되며, 이러한 모듈들에 의해 처리된 기업 평가 데이터 분석 결과를 기업에게 제공하는 기업 마케팅 전략을 위한 SNS 및 Web 데이터 분석 시스템을 제공하고자 한다.

3.1 SNS 및 Web 데이터 수집 모듈 설계

본 논문에서 제안하는 SNS 및 Web 데이터 수집 모듈은 특정한 관심이나 활동을 공유하는 사람들 사이의 관계망을 구축해 주는 온라인 서비스인 SNS, E-news, 그리고 블로그 같은 시스템을 통해 기업에 대한 이미지를 추출 하도록 설계한다.

즉, 데이터 수집 모듈은 FaceBook, Twitter와 같은 SNS에 실시간으로 업데이트 되는 데이터 중에서 그룹명을 의미하는 Keyword 혹은 그룹의 제품명을 의미하는 Keyword가 포함된 데이터를 추출하여 HBase에 전송하도록 설계한다.

데이터 수집 모듈이 수집하는 정보는 Keyword, 등록일 및 시간, 작성자, 내용, 첨부파일로 구성된다.

3.2 HBase 모듈 설계

HBase는 HDFS(Hadoop Distributed File System)에 구현한 분산 컬럼 기반 데이터베이스로 수백, 수억만 건의 행과 컬럼을 처리하는 대용량 테이블을 지원하기 위한 목적으로 개발된 시스템으로 HDFS와 MapReduce 등과 함께 사용되기에 최적화 되어있다.

본 논문에서는 3.1절의 데이터 수집 모듈에서 수집한 비정형 데이터를 그림 1과 같이 HBase에

저장하도록 설계한다.

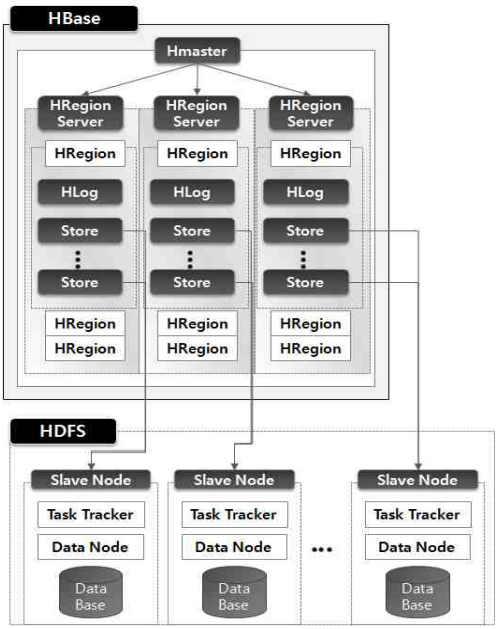


그림 1. HBase의 구성 및 흐름
Fig. 1 The component and flowchart of HBase

3.3 데이터 분석 모듈

데이터 분석 모듈은 SNS와 Web을 통해 수집된 기업의 평가 혹은 기업 제품의 평가에 대한 데이터를 분석한 후, 기업 자체 혹은 제품이나 서비스에 대한 이미지를 판단하고 분류한 내용을 Hbase에 다시 업데이트 하도록 설계한다.

3.3.1 평가 주제어 추출 및 형태소 분석기를 통한 품사 태깅 과정

평가 주제어 추출 및 품사 태깅 과정은 HBase 테이블에서 추출한 Column:"평가내용" 값 즉, Value값을 전달받아 이 데이터를 기반으로 평가 내용의 주제어를 추출하고 품사를 태깅하는 역할을 수행한다.

형태소 분석기를 통한 품사 태깅과정은 다음과 같으며, 그림 2는 주제어 탐색 및 태깅 과정에 대한 흐름도를 설명한 것이다.

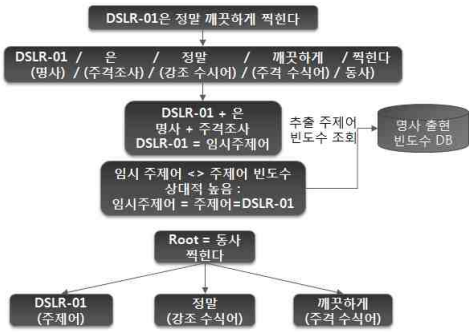


그림 2. 주제어 탐색 및 태킹 과정
Fig. 2 The search process of keyword and Tragging process

[1 단계] HDFS로부터 평가내용인 "DSLR- 01은 정말 깨끗하게 찍힌다."를 추출하여 Value 변수에 저장한다.

[2 단계] Value를 오픈 소스인 형태소분석기를 통해 품사를 태깅 한다.

[3 단계] 분석된 형태소에 주격조사 "은","는", "이","가'를 탐지하여 주격조사 앞에 명사를 임시 주제어로 선정한다(그림 2의 "DSLR -01 + 은).

[4 단계] 임시 주제어를 명사 출현 빈도수 DB를 조회한 결과 상대적으로 높은 빈도수로 출현한 경우 이를 주제어로 선정한다 (그림 4의 DSLR-01이 주제어로 선정).

[5 단계] 태깅된 형태소중 동사를 기준으로 트리를 구성 한다.

위의 단계를 거쳐 트리화된 데이터를 이용하여 긍정 및 부정에 대한 의미방향 추출과정을 수행한다.

3.3.2 긍정,부정 판단을 위한 의미 방향 추출 설계

본 논문에서 긍정, 부정 판단을 위한 의미방향 추출 과정은 평가 주제어 추출 및 형태소 분석기를 통한 품사 태깅 과정에서 형성된 형태소 트리를 기준으로 하위 계층의 서술어의 극성값을 이용하여 의미방향을 추론하도록 설계한다.

의미방향을 추론하는 과정은 다음과 같은 단계로 수행된다.

[1 단계] 트리의 하위 계층 가운데 주격 서술어인 "깨끗하다"를 추출하여 서술어 긍정, 부정 출현 빈도수 데이터베이스를 조회하여 "깨끗하다"라는

표현이 긍정적으로 출현한 횟수와 부정적으로 출현한 횟수와 긍정,부정 서술어의 총 출현 빈도수를 추출하여 극성값 연산을 수행한다. 극성값 연산 공식은 식 1과 같다.

$$\text{극성값} = \frac{\text{긍정표현빈도} - \text{부정표현빈도}}{\text{총긍정표현빈도수} + \text{총부정표현빈도수}} \quad \text{.....식(1)}$$

예를 들어, “깨끗하다”를 기준으로 긍정 표현 빈도수가 400회이고 부정표현 빈도수가 60회 그리고 총 긍정표현 부정표현의 빈도수를 합산한 결과 38,000이라는 값이 추출되었을 경우 “깨끗하다”의 극성값은 약 0.00895의 값을 가진다.

[2 단계] 앞 단계에서 연산된 극성값을 양수와 음수로 구분하여 양수인 경우 극성값을 +1로 음수인 경우 -1로 대체한다.

[3 단계] 트리의 하위 계층 가운데 주격 서술어 앞에 강조 서술어가 있는 경우 서술어의 극성값을 2배 증가시켜준다. 그림 2의 경우, 트리의 하위 계층 주격 서술어 “깨끗하다”앞에는 강조 서술어 “정말” 이라는 단어가 있기 때문에 깨끗하다의 극성값은 {+1 x 2}로 총 +2의 극성값을 가지게 된다.

3.4 PSH 모듈 설계

본 논문에서 제안하는 PSH 모듈은 Hadoop기반 Map, Sort, Reduce과정을 통해 Hbase에 저장된 데이터를 수집 및 분류하여 모니터링할 수 있게 해주는 역할을 수행한다.

이는 기업의 마케팅을 위한 모니터링 과정에서 질의가 들어오는 경우 질의사항을 Key값으로 하여 Web 혹은 SNS 데이터 가운데 Key값과 일치하는 데이터만을 수집하고 이를 다시 오피니언 마이닝을 통해 분석한 결과대로 정렬한 후, 이들을 병합하여 모니터링 시스템에 전달해주는 역할을 수행한다.

3.4.1 Map 모듈 설계

Map 모듈은 기업의 마케팅 목적으로 모니터링을 수행할 때 모니터링 관리자의 질의에 맞는 데이터를 추출하기 위한 목적으로 실행된다. Map 모듈은 그림 3과 같이 구성되며 각각의 Map 모듈은 의미 분석을 마치고 Sort Module에 전달하게 된다.

모니터링 관리자가 이렇게 저장된 데이터를 마케팅을 목적으로 특정 제품에 대한 평가 혹은 시간대별 기업의 평가 등을 조회할 때, 제품명을 Key값으로 검색하여 목적에 맞는 데이터만을 추출한다. 이러한 값은 다시 중복되는 Key값끼리 묶어주는 Sort Module로 전송된다.

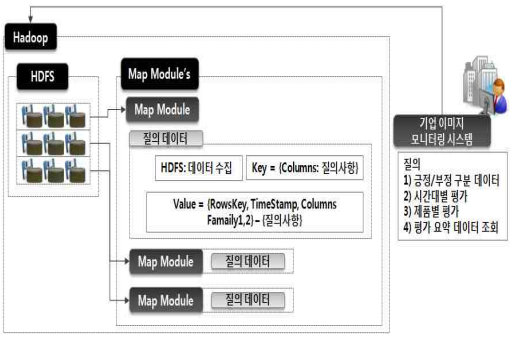


그림 3. Map 모듈 절차
Fig. 3 The process of Map module

3.4.2 Sort 모듈 설계

Sort 모듈은 Reduce 모듈에 전달되기 전에 Map 모듈에서 맵핑된 수많은 {Key, Value} 쌍을 분류하는 모듈이다. 각각의 Sort 모듈은자신이 담당하는 Map 모듈에서 맵핑된 데이터의 Key값에 기준을 두어 하나의 Key에 대하여 Buffer를 생성하고, 동일한 Key에 여러 개 존재하는 Value값을 통합하는 과정을 수행한다.

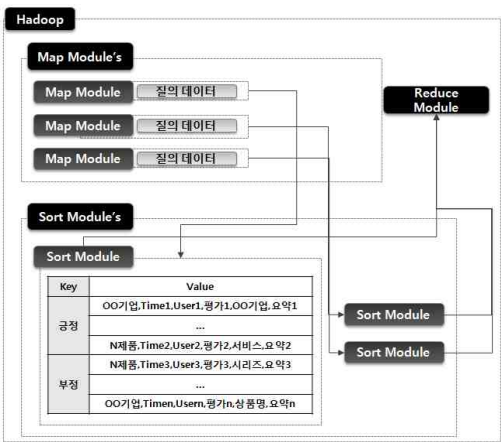


그림 4. Sort 모듈 절차
Fig. 4 The process of Sort module

그림 4는 Sort 모듈의 처리절차를 설명한 것이다. 그림 4와 같이 수집된 데이터를 모니터링 하기 위하여 Map 모듈에서 {Key,Value}쌍을 의미방향과 기타정보로 구성되었을 경우 각 Map 모듈과 연결된 Sort 모듈에서 이를 다시 수집하여 같은 Key 즉, 의미방향이 “긍정”인 데이터와 “부정”인 데이터를 분리하여 통합한다.

3.4.3 Reduce 모듈

Reduce 모듈은 각각의 Sort 모듈에서 Key값으로 정렬된 데이터를 모두 수집하여 데이터를 통합하고, 정렬하여 기업 이미지 모니터링 시스템에 전달하는 역할을 수행한다.

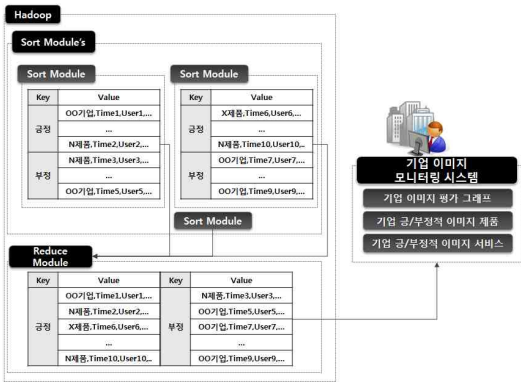


그림 5 Reduce 모듈 절차
Fig. 5 The process of Reduce module

Reduce 모듈은 그림 5와 같이 Sort 모듈 내의 모든 데이터를 수집하여 같은 Key값끼리 통합한다. 예를 들어 Key값이 “긍정”인 데이터만을 추출하여 통합하고, 다시 “부정”인 데이터 끼리 통합하는 과정을 수행하며 최종적으로 통합된 데이터를 기업에게 제공하도록 설계한다.

Ⅳ. 결 론

본 논문에서 기업 마케팅 전략을 위한 SNS 및 Web 데이터 분석 시스템을 제안하였다. 본 논문에서 제안한 SNS 및 Web 데이터 분석 시스템의 기대 효과는 다음과 같다.

첫째, SNS에 급속도로 퍼지는 기업 혹은 기업 제품에 대한 데이터를 수집한다.

둘째, 기업에 대한 수집 데이터를 자동으로 긍정 혹은 부정 데이터로 분류하여 저장하기 때문에 제품 혹은 기업에 대해 개선사항 혹은 발전해야할 사항을 빠르게 파악할 수 있다.

셋째, SNS 혹은 Web을 통해 빠르게 번지는 기업의 부정적 데이터에 대한 수습시간을 모니터링을 통해 단축시킬 수 있다.

향후, 본 논문에서 제안한 PSH 모듈은 차후 스케줄링과 필터링 기능 탑재를 추가할 예정이며 이를 통해 분석 및 모니터링 시간을 보다 단축시킬 수 있을 것으로 예상된다.

참 고 문 헌

[1] Jeffrey Dean, Sanjay Ghemawat, “MapReduce Simplified Data Processing On Large Clusters,” Google Inc., Communications of the ACM, Vol.51, No.1, pp.107-113, 2008.

[2] White Tome, Cutting Doug, “Hadoop: The Definitive Guide,” O'REILLY, 2009.

[3] 황인성, “맵리듀스 모델의 성능 향상을 위한 데이터 분배 및 작업 진행 스케줄링”, 인하대학교 석사학위논문, pp.11-28, 2011년 2월.

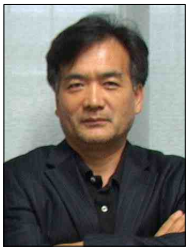
[4] 배혜찬, “맵리듀스 환경에서 블룸 필터를 사용한 탄력적 조인 연산 처리”, 서울대학교 석사학위논문, pp.12-30, 2013년 2월

[5] Jung-Yeon Yang, Jaeseok Myung, Sang-goo Lee, "A Sentiment Classification Method Using Context Information in Product Review Summarization," Journal of KIISE, vol.36, no.4, pp.254-262, 2009.

[6] 곽혜림, “오피니언 마이닝을 이용한 선호도 분석 및 예측기법”, 서울과학기술대학교 석사학위논문, pp.5-33, 2011년 8월

저자약력

이 병 관(ByungKwan Lee) **정회원**



1979년 부산대학교
기계설계학과 학사
1986년 중앙대학교
전자계산공학과 석사
1990년 중앙대학교
전자계산학과 박사
1988년 3월~현재 관동대학교
컴퓨터학과 교수

<관심분야> 네트워크 보안, 인터넷 보안, IoT
빅데이터

정 은 희(EunHee Jeong) **정회원**

1991년 강릉대학교
통계학과 학사
1998년 관동대학교
전자계산공학과 석사
2003년 관동대학교
전자계산공학과 박사
2003년 9월~현재 강원대학교
지역경제학과 부교수

<관심분야> 네트워크 보안, 전자상거래 보안,
빅 데이터

정 이 나(YiNa Jung) **정회원**

2011년 관동대학교
컴퓨터학과 학사
2011년 3월~ 현재 관동대학교
전자계산공학과
석박사 통합과정 중

<관심분야> 네트워크 보안, 빅데이터, IoT