학교폭력과 자살사고를 예방하기 위한 감성분석 시스템의 설계

김영택†

요 약

현 청소년들의 학교내 생활환경에서 문제점으로 대두되는 폭력 및 자살사고 발생률 증가에 대한 예방 차원의 빅 데이터 처리 분석 시스템을 목표로 연구하였고 설계의 경제성과 용이성, 적용의 신속성 등을 고려해서 많은 이용률을 가지고 있는 오픈 소스인, 하둡 시스템(Hadoop system)의 맵리듀스(MapReduce) 알고리즘과 분산 병렬 환경을 위한 HDFS(Hadoop Distibuted File System) 구성을 사용하여 실험하였다. 연구에서 사용된 분석기법은 기존의 통계적인 분석기법들이 가지는 난이도를 피하기 위해 상업적인 사회 망의 비정형 대화 자료를 이용해서 폭력성 어휘에 대한 단어 수(word count) 분석을 적용하여 폭행, 자살사고를 사전에 감지하여 예방하는 감성분석(sentiment analysis) 시스템을 텍스트 마이닝 관점에서 제안하여 실험하였다.

주제어 : 감성분석, 빅 데이터, 맵 리듀스, 하둡 시스템, 키-값 매핑, 워드 카운트, 텍스트 마이닝, 폭력성 어휘, 자살사고

Design of a Sentiment Analysis System to Prevent School Violence and Student's Suicide

YoungTaek Kim†

ABSTRACT

One of the problems with current youth generations is increasing rate of violence and suicide in their school lives, and this study aims at the design of a sentiment analysis system to prevent suicide by uising big data process. The main issues of the design are economical implementation, easy and fast processing for the users, so, the open source Hadoop system with MapReduce algorithm is used on the HDFS(Hadoop Distributed File System) for the experimentation. This study uses word count method to do the sentiment analysis with informal data on some sns communications concerning a kinds of violent words, in terms of text mining to avoid some expensive and complex statistical analysis methods.

Keywords : Sentiment Analysis, Big Data, Map Reduce, Hadoop System, Word Count, Key Value Mappinpg, Text Mining, Violent Word, Suicide Accident

† 정 회 원: 경성대학교 컴퓨터공학부 교수
논문접수: 2014년 10월 7일, 심사완료: 2014년 11월 05일, 게재확정: 2014년 11월 11일

1. 서론

1.1 연구목적 및 필요성

본 연구는 현 국내 교육정책에서의 가장 난제로 분류되어지고 있는 청소년 학생들의 폭력이나 자살관련 문제에 대한 예방책을 마련하기 위해 빅 데이터 개념을 적용한 근실시간 처리 (near real time processing) 분석기법을 제안하고, 폭력성향을 사전에 미리 파악 할 수 있는 감성분석 시스템 설계를 목표로 연구하였다.

그 목표를 위해 빅 데이터 관련 오픈소스를 사용한 가상 및 실 자료를 이용한 실험을 통해서 현 교육현장의 환경에 적응 할 수 있는 방법을 설계하여 그 결과의 사회적인 가용성에 대해 고찰하였다. 이 목적을 수행하기 위해 본 연구에서는 하둡(Hadoop) 시스템의 분산 병렬 실행기능을 사용하였고 맵리듀스(MapReduce) 알고리즘을 적용한 키-값 매핑(key-value mapping) 기법을 적용한 단어 수 분석(word count analysis)을 주된 도구로 활용하였다. 여기서 하둡 시스템의 분산 병렬 구조인 HDFS (Hadoop Distributed File System)가 이용되는 것은 먼저 오픈소스의 경제성과 더불어서 본 제안 의 적용 환경이 전국적인 광역성 뿐 만 아니라, 각 지역별, 학교단위별, 학교 내의 학년별, 학급별 등의 각기 다른 단위의 분산적인 교육 환경에서 해당 교사들의 운영에 의해서 처리가 되며, 병렬성을 유지하면서 구현될 수 있다는 장점이 있기 때문이다.

본 연구에서 계획했던 실험의 문제점은 먼저, 자료의 수집에 있었고 학생들의 최우선선호 사회망인 페이스 북(Face Book) 관련 시스템 상에서의 개인 대화 정보의 추출 및 실험적 사용은 불법이기 때문에 본 연구에서는 필요한 자료를 공개된 사회 망 자료의 일부를 가상적으로 설정 할 수밖에 없었다. 그렇다면. 실험 결과의 교육적 목적 사용은 이 기법의 정당성과 효율성이 인정이 되고나서 교육당국에 의한 교사 학생 간의, 혹은, 각급 교육기관 간의 합의에 의한 사용 허가를 전제로 적용하고자 한다. 그런 이유로 인해서 본 연구에서는 이미 사회적으로 공개성과 인기를 유지하고 있는 '일간베스트' 사회망의 일부 자료를 예시로 활용하였다. 이 연구의 결과는 어디까지나 교육용 목적의 사용에 대한 제안에 목표를 두고 있다.

현 세대의 ICT 분야에서 빠르게 진화하고 있는 분석기법 중에서도 특히, 감성분석 혹은, 견해분석(opinion analysis) 등이 상업적으로나 사회적으로 활발하게 개발되어지고 있다. 대부분 이 분야 연구개발의 실적이 주로 사회 경제적인 효율성, 마케팅, 혹은 일부의 정치적 응용과 관련이 되어 있는 실정이다. 하지만 본 연구에서는 같은 맥락의 기법이라도 교육적인 분야에 적용 할 수 있는 여지를 고찰하였고 실험적으로 분석을 적용하여 그 가용성을 제안 하였다.

물론, 종래의 각종 수학적인, 통계적인 정량적 처리기법의 분석들이, 정형적인 자료에 대한 타당성이 있는 결과와 해석을 보장 할 수 있지만, 빅 데이터 처리가 가지는 새로운 응용환경에 대한 비정형 자료에 대한 적응성은 통계적인 기법들이 가질 수 없는 여러 가지 현실적인 장점을 보장받기 때문에 훨씬 더 효율성이 있는 것으로 알려져 있다[1].

이런 장점의 이유는 먼저, 빅 데이터의 주된 대상 자료가 일반적인 정형성 자료 뿐 만아니라, 소셜 미디어 중심의 비정형성 자료들을 주로 취급하기 때문에 인위적으로 만들어진 인터뷰나 설문조사와 같이, 소비자나 일반인의 정해진 규격의 환경에서 생산된 것이 아닌, 대상자들이 자발적으로 표현한 '날 것 그대로'의 데이터라는 점에서 그 관련자들을 가장 잘 파악할 수 있는 각종 마케팅과 혹은, 정치적인 지표로 각광을 받고 있다. 이러한 소셜 데이터에서 소비자가 특정 대상에 대해 느끼는 좋고 싫음, 그리고 나아가 그 이유를 분석해 주는 감성 분석의 시도는 소셜 마케팅 기업들에게 필수적인 전략도구로서의 기능을 발휘하고 있는 실정이다.

본 연구에서는 이런 사회적인 변화추세가 현실적으로 청소년들의 생활 및 교육적인 변화에 적용이 가능한가를 실험하는데 초점을 두고 있고 또한 현 세대에서 진행되고 있는 스마트 교육열풍이 가지고 있는 여러 가지 역기능 및 순기능에 대한 복합적인 현상 중에서도 역기능에 속하는 sns의 대화 문제점을 주된 관점으로 고찰하였으

며 그에 대한 해결책에 도움을 주는 기능을 빅 데이터 처리기법으로 수행하고 제안 하였다.

또한, 연구의 필요성은 일반적인 청소년들의 현실문제에 감성분석을 응용하기 위한 것이었고, 이것을 위해서 하인리히(Heinrich)법칙을 이론적인 기반으로 하였다. 그 법칙의 내용은, 각종 사고와 관련된 사회적인 보험 통계 분석에서 300:29:1 비율로 소형:중형:대형 재해사고의 전조현상이 발생하고 있다는 것이 하인리히의 분석이었다, 그래서, 본 연구에서는 학교 현장의 청소년의 폭력성 관련사고의 연쇄적인 관계를 사전에 감지하는데 이 법칙이 도움이 될 수 있다는 점에 착안하고 있다.

만약, 연쇄적 사건, 사고의 연결고리를 어느 정도 빅 데이터 분석에서 예측 가능하다면, 재해사고의 도미노이론이 안전관리 이론으로 발전하게 할 수가 있다, 그래서, 학생들의 교내 및 가정에서의 인성지도와 폭력관련 사고방지를 위한 안전지킴이로서의 역할을 하는 분석시스템 제안이 본 연구의 방향이 된다. 여기서, 이 법칙의 300:29:1 사고발생 비율 자체는 별 의미가 없다. 마찬가지로, 적용되는 학생들의 폭력 행동의 유형을 어느 특정한 유형의 사고나 사건으로 규정 할 필요는 없다. 다만, 모든 폭력성 관련 특징에 대한 포괄적인 유형을 의미하고 있다[2][3][4].

본 연구에서 주된 폭력성 관련 행동은 대부분이 언어적인 형태로부터 먼저 나타나고 있다는 점을 착안하고 있다. 다시 말해서, 300개정도의 언어폭력 발생이 29개 정도의 폭력사고를 유발하고 있고 종래에는 그것이 1건의 자살 사고까지 연계될 수 있다는 것이다.

또한, 이런 폭력적인 언어 형태가 스마트사회와 접목되어서 표출되고 있다는 점은 이미 널리 알려져 있다. 소위 말하는 사회 망을 통한 전달이 폭발적으로 증가하고 있는 현상이며, 이 점을 상업적으로나 정치적으로 누군가에 의해서 이용당하고 있다는 점 때문에 빅 데이터 처리를 통한 비정형성 자료의 분석으로 폭력사고를 예방해야 한다는 의미가 있다.

여기서, 가장 크고 심각한 결과로 나타나는 사고형태가 자살이고, 어떤 자살사고의 발생 이전에 다른 유형의 사고가 그 당사자에게 이미 발생하고 있다는 것을 미리 분석한다면 자살을 막을 수 있는 상황이 될 것이다.

일반적으로, 자살사고 이전에 먼저 관련 학생들 간에 욕설이 난무하였을 것이고 또 특정 학생에게 집중 되면서 왕따(집단 따돌림) 현상과 폭력이 동반되었을 것이 분명하다. 그렇다면, 이 과정에서의 각 단계별 발생 건수는 하인리히 법칙에서의 300:29:1 의 비율과 거의 유사한 특징을 가지고 있다는 것을 암시하고 있다.

만일, 욕설과 협박 행동이 표출되어 전달되는 과정이 스마트 화 되어버린 사회 망 의존 세대의 의식과도 관계가 있기 때문에 그에 대한 교육적인 대비책이 하루빨리 만들어져야하는 현실에 당면하고 있다. 본 연구에서 사회 망 중심의 대화 자료 중에서도 특히, 폭력성 언어를 내포한 비정형성 자료에 초점을 두고 빅 데이터 분석을 시도하는 것이 바로 이런 현실적 문제점의 해결을 위한 길이기 때문이다.

1.2 연구 목표

본 연구의 초점은 청소년들의 폭력성 대화를 통해 그들의 감성을 분석하여 폭력이나 자살 사고를 예방하기 위해서는, 먼저 학교 교사들에 의한 분석 시스템의 관리 및 사용을 염두에 두어야 할 것이다. 그렇기 때문에 실제 사용의 용이성 및 그 효과에 대한 긍정적인 결과를 도출하는 것을 목표로 하였다. 마찬가지로, 그것을 위해서는 선택적으로 넓은 빅 데이터 처리 기법들 중에서도 어떤 기법이 지역적으로 분산 운영되고 있는 학교 현장에 적합 할 것인지를 고려하기 위한 연구이기도 하다.

여기서, 제일 먼저 생각해야 할 관점은 수요자 측면에서의 접근 용이성이다.

다시 말해, 교사들의 손쉬운 자료 획득과 처리의 간편함이 우선적으로 생각 되어야 한다는 뜻이 된다. 이 목표를 위해서 본 연구에서는 비정형성 자료로 부터 폭력성 어휘의 발생 빈도를 측정하여 폭력에 대한 의지를 사전 분석하는 기법의 취급 용이성을 실험하였고, 더불어서 상대적으로 다른 여러 가지 빅 데이터 분석기법이 가지는 복잡도 문제의 수학적 구성 원리를 외형적으로

비교하여서 본 연구에서 선택한 분석기법에 대한 결론에 적용하였다.

두 번째로 고려해야 할 사항은, 시스템의 구현 경비이기 때문에 오픈 소스 환경을 선택해야 만 하였다. 특히, 여기서 가장 분산처리 능력과 병렬성에서 우수한 기능을 보여주는 Hadoop 시스템을 적용하는 실험을 하였다.

2. 본론

2.1 빅 데이터 분석 및 관련 비교연구

2.1.1 Mining, 계량분석

본 연구에서 적용하고자 하는 감성분석 기법 이외에도 일반적인 분석기법으로서 응용분야에 따른 여러 가지의 분류가 가능하다. 예를 들어서, 기존의 사회 망 대상의 감성분석 목적으로 어휘에 대한 분석을 위해 많이 이용되는 SVM (Support Vector Machine) 뿐만 아니라, 데이터 마이닝에서의 대규모 상품이나, 서비스, 이메일 등의 각종 정형 및 비정형 자료들의 거래기록에서 의미 있는 관계자료 추출을 위한 분산 환경 프레임워크인 하둡과 시스템 R의 통계분석 기법 등이 활발히 연구되어지고 있다. 또한, 계량정보 분석기법으로 분류되는 문헌정보의 서지분석이나, 과학적 지식분석 등은 과학기술 분야 문헌의 생산과 유통을 중심으로 계량적 분석을 하고 있다. 아울러서, 웹 구조의 이용분석을 통한 각종 지식 개체들 간의 연결 관계까지 분석을 시도하고 있는 실정이고 이것을 위한 Web 3.0 개발로 인한 시멘틱 웹의 발전이 펼쳐지고 있는 현실이다[5].

2.1.2 네트워크 분석, 복잡계 분석

네트워크 분석기법으로는 분석대상을 네트워크의 집단으로 구성하고 각 노드의 중심성 역할을 여러 가지 방향에 따라 구분하여 처리하고 있다.

여기서는 중심성 분석, 하위집단 분석 등으로 분류되어서 연구되어지고 있다.

중심성 분석에는 기본 중심성, 확장 중심성 등의 기법이 있고 하위집단 분석에는 컴포넌트, 파당, 위성분석 등의 기법들이 Netminer, UCINET/NetDraw와 같은 여러 가지 툴을 활용하여 연구가 진행되고 있다. 또 다른 기법으로, 무질서한 구조적 시스템에서부터 어떤 일정한 질서를 찾아내는 기법이 되는 복잡계 분석도 활발하게 연구되어진다. 여기서는 일반적으로 과학적, 기술적인 해법이 적용되기 힘든 분야, 예를 들어 카오스현상, 요동현상, 나비효과 등 특정한 패턴의 현상에 대한 분석도 인공 지능적 측면에서 많은 연구가 진행되어지고 있다[5[6[7]]].

위에서 제시한 여러 가지 기법이 공통적으로 가지는 시스템의 복잡도 및 추상화 된 구현, 혹은 사용의 어려움으로 인하여 어느 방법에서도 본 연구의 목적으로는 적용이 힘들다는 사실이 문제점이 되기 때문에 위의 기법과는 다르게 간편한 설계 및 교사들의 손쉬운 사용을 위하여 가장 기본적인 방법의 단어 수에 의한 가중치측정 방법을 제안하였다.

그 이유로서는, 적용난이도가 상대적으로 높은 방법들은 먼저 수학적 상관관계를 비정형데이터에 적용해야 하는 어려움을 내포하고 있고 또 분석 결과의 일관성은 유지 될지 모르지만 복잡계 분석에서의 경우와 같은 특수성에 대한 적용에는 감성분석에서는 문제가 있을 수 있기 때문이다.

하지만, 본 연구에서 사용되는 텍스트 마이닝 기법에서 일반적으로 사용하는 단어수의 분석 기법은 사용하는 특정단어의 출현빈도나 동시 출현빈도 측정을 기반으로, 해당 관련사항의 중요도(weight), 혹은 의미(opinion)를 주관적으로 추정하는 기초적인 방법에서도 많은 성공사례가 전해지고 있기 때문에 실험에 적용하였다.[5][6][7].

3. 연구 실험

3.1 하둡 시스템 실험환경

본 연구에서 사용하는 빅 데이터 처리 환경으로서 오픈 소스 기반의 대용량 병렬 분산 처리를 목표로 하는 프레임워크 지원이 가능한 HDFS와 맵 리듀스 기법으로 구성 되어 키-값 매핑으로 엔티티를 처리하는 방법을 선택하였다. 기존의 또 다른 프레임워크로서 존재하는 스플렁크(Splunk) 시스템은 상용 시스템이기 때문에 구현의 경제성

문제로 인하여 연구대상에서 고려되지 못하였다. 또한, 실험에서 사용된 비정형자료는 공개 자료인 '일간베스트' 소셜 미디어의 일부분 자료에 실험적으로 적용하여 분석하였다. 여기서 개인정보의 실험적 사용은 실제적으로 법적인 제한을 받을 수 있었기 때문에, 공개적인 실험에 사용하기에는 가장 적합한 미디어로서 이 회사의 자료를 선택하였다.

특히, 폭력성 어휘의 출현빈도가 폭력 사고 및 자살사고와 연관성 있다는 가정과 더불어 학교 교육환경에서 스마트장비 사용의 급속한 확대현상과 더불어서, 교육 담당자의 학생들에 대한 상담용 사회 망 통신에 대한 확대를 의무적으로 할 수 있다는 가능성 조건을 미리 가정하고 실험하였다. 예를 들어서, 장래에는 학생과 교사 사이에 교육적인 ,혹은 개인적인 소셜 미디어를 통한 대화가 각 기업에서 실행 하는 것처럼 학생지도에서도 적극 활용 될 수 있음을 가정하기 때문에, 우선 가상적 실험을 위해서 실제적으로 필요한 자료를 위해 기존의 소셜 미디어로 부터 자료를 넛치(Nutch) 크롤러 시스템 오픈 소스를 사용해 수집하였다.

3.2 하둡시스템, 맵리듀스와 단어수의 분석기법

하둡 시스템의 구성은 크게 stand alone, pseudo distributed, fully distributed 등 세 가지 환경의 구축이 가능하다. 본 연구에서는 하나의 서버에 가상적으로 분산 시스템 구축이 가능한 pseudo distribution 방법을 사용하였고, 윈도스 환경에서 자바로 구현 하였다. pseudo distribution의 가상 병렬구조 플랫폼을 10개 노드의 컴퓨터 구성으로 실험하였다.

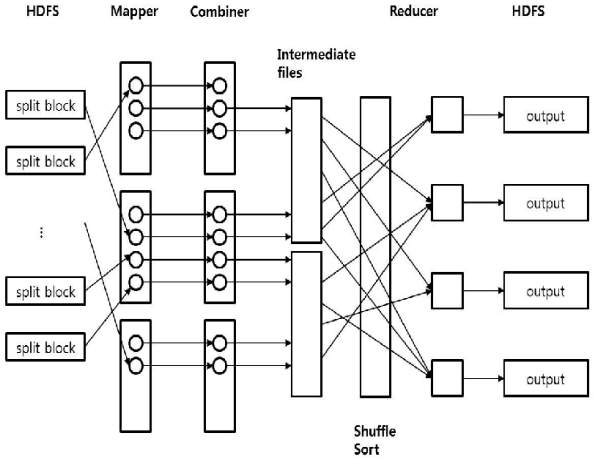
일반적으로 키-값 매핑 방법을 사용하는 맵리듀스 방식의 특징은 다음 절에 명시된< 그림 1> 에서 보인 것과 같다. 그 장점으로서는, 데이터 웨어하우스(DW)의 수평적인 확장 접근 방식으로 구성되는 MPP(Massibly Parallel Processing) 구조의 병렬 DBMS나, 혹은, 보통의 SAN(Storage Area Network), NAS(Network Attached Storage)와는 다르게, 분산 파일 시스템 전용 구조인 HDFS를 보통의 내장 하드디스크를 가진 일반 컴퓨터로 구현할 수 있다는 점이다, 또 각 노드 컴퓨터 간에는 약한 상관관계를 유지하기 때문에 그 연결의 수를 얼마든지 확장 할 수 있으며, 그중 어느 컴퓨터의 시스템의 장애 문제도 충분히 복구하기 위해서 정해진 스플릿 단위(64MB 디폴트 블럭)의 모든 파일에 대한 삼중 복사를 기본으로 구성해 처리하고 있다.

그래서, 만약 이 시스템을 각 급 학교 환경에 적용한다면 교사 한 명당 기존의 사용 중인 하나의 pc를 각 슬레이브 노드로 병렬분산 시킬 수 있고, 또 혹은, 하나의 중간 마스터 노드로도 사용이 가능하다. 또 다른 중요한 특징으로서는, 병렬 프로그래밍에 익숙하지 못한 프로그래머라도 쉽게 데이터의 병렬처리에 접할 수 있도록 하고 있다 하지만, 여기서 구성되는 병렬구조에서 나오는 생산성은 이미 다른 연구에서 아주 우수하게 평가되어 지고 있다.

3.3 HDFS와 맵리듀스 과정

GFS(Google File System)을 기반으로 하는 하둡 분산처리는 크게 HDFS와 맵리듀스 과정으로 구성된다, HDFS는 마스터 노드와 다수의 슬레이브 노드들로 이루어지고, 마스터 노드의 네임노드 및 잡 트래커의 제어에 의해서 각 슬레이브 노드의 데이터 노드 및 테스크 트래커들의 스토리지와 단위작업에 대한 스케쥴링을 수행하여 대규모 작업의 분산 병렬처리의 효율성을 구현하고 있다.

여기서 큰 파일을 여러 개의 블록으로 분리 저장된 상태에서 아래와 같은 맵리듀스 과정을 구성하고 있다.



<그림 1> HDFS의 맵리듀스 과정

<그림 1>과 같이 블록의 스플릿 과정을 거친 입력 데이터 소스를 분산된 여러 매퍼에 의한 키-값 매핑을 처리하고 나서 그 중간 결과는 소팅과 셔플 과정을 거쳐서 리듀서 에게 전달되며 리듀서 태스크가 모든 키 값을 통합하여 결과를 출력한다.

3.4 분석기법의 비교

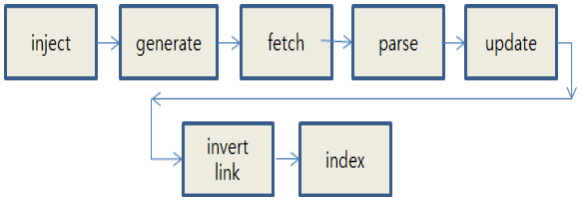
빅 데이터의 텍스트 마이닝 관련 분석업무에서 가장 기본적으로 많이 이용되는 전문용어 인식 기법은, 일반적으로, 키 값의 후보 용어추출(candidate term extraction)과 용어의 상대적인 가중치 할당 (termhood weight assignment) 의 두 단계로 구분되어 처리된다. 이 기법은 입력된 문장 내의 여러 가지 품사들의 패턴을 몇 가지 종류의 필터를 사용하여 먼저 추출한다. 또한, 여기서 추출된 후보용어는 출현빈도에 의해 가중치가 주어지며, 출현 빈도를 위한 통계적 연구로서는 다이스 계수(dice coefficient)를 사용하여 용어들의 응집도를 계산한 F. Smadia et.al (1996)의 연구와 더불어서 다어절 용어에 내포된 부분적인 문자열의 전문성을 측정하는 C-value 기법을 제안한 K.Frantzi et.al(2000)의 방법도 상당한 수준의 인식률을 보여준다. 그 외에도 후보용어의 출현빈도로 가중치를 측정하는 기법으로서는, 구글 정규거리 (Normalized Google Distance) 등이 사용되어진다[8][9] [10].

그렇지만, 본 연구에서의 목표가 일반 교사 사용자의 간편성과 학생들의 비정형자료에 대한 활용성에 초점을 두고 있기 때문에 위에서 언급된 고차원적인 난이도를 포함한 통계적 기법과는 거리를 두는 방법을 사용 하였고 그 점이 바로 단순하게 특정 단어, 특히, 폭력성 어휘를 사용하는 횟수를 단어 수 분석 하는 키-값 매핑으로 방향을 잡은 것이다. 이 기법의 구현을 위해서는 하둡 시스템의 접근성을 이용하였다.

다음 절에서 실험의 결과를 보였다.

4. 실험 결과

본 실험에서의 첫 번째 단계에서 먼저 크롤러 오픈 소스 넛치 시스템을 이용해 <그림 2>와 같은 수집단계를 실행하였다. 일반적으로, 맵리듀스 실행에 적합한 넛치 시스템을 이용하였지만 다른 크롤러를 사용해도 무방하다. 그 단계는 <그림 2>와 같이 보여준다.

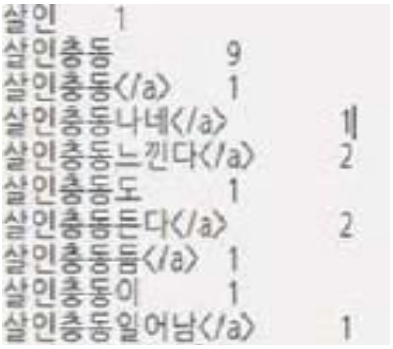


<그림 2> 크롤러 넛치 시스템의 자료 수집

또한, 본 실험에서 적용한 사회 망의 예에서 공개와 자료수집이 가능한 대중사이트인 '일간베스트' 대화내용을 대상으로 단어 수 분석을 수행한 결과의 일부가 아래와 같이 보여 진다. 하지만, 본 연구의 실제 대상이 되는 교사에 의한 학생들 간에 벌어지는 대화의 경우와는 다른 양상이 되지만 시스템의 구성에 있어서는 같은 맥락의 실험이 되는 것이다.

여기서 보여 지는 키 용어로서 사용되어진, '죽**' 관련 어휘의 출현횟수가 값으로 산출되어진다.

<그림 3> 단어 수 분석의 예(1)
후보용어 '죽**' 대상



<그림 4>단어 수 분석의 예(2)
후보용어 '살인**' 대상

이 실험에서 보여 지는 가중치 밸류 값이 가상적인 일부 자료이기 때문에 작은 수치가 보여 진 것이고 그 포맷은 어느 특정한 집단의 대화내용에서 일어나는 위험성을 분석하는데 별 어려움 없이 적용 되고 있다.

여기서, 키 단어로서 청소년들의 감성적인 지도와 분석에 필요한 각기 다른 여러 가지 종류의 어휘를 적용 할 수 있는 교육적인 연구와 더불어서 학교 현장에서 이 과정을 쉽게 교사들에게 사용 할 수 있게 하는 소프트웨어를 오픈 소스로 공급을 할 수 있는 투자가 필요할 것이다.

5. 결론

본 연구에서는 빅 데이터 분석 응용의 확산 추세에서 대부분의 적용이 학술적이거나 기업, 연구기관, 정부 부처 등의 비 생활적인 분야에 국한되어 있었다는 점이 분명하기 때문에, 연구 방향을 생활관련, 특히 교육적인 분야에 대한 응용으로 잡았고, 그 중에서도 학교 내에서의 사회 망 대화에 나타나는 폭력성 대화의 흔적을 빅 데이터 분석을 통해서 실시간적으로 단어 수를 측정 분석하여 폭력성 대화의 사전 탐색을 교사들에 의해서 가능하게 하고 미리 조치하여 폭행과 더불어서 학생들의 자살을 방지하는 방법을 제안 하였다.

윗 절에서 명시한 각종 다른 통계적 분석기법의 복잡성을 피 할 수 있었기 때문에, 본 연구에서의 단어 수 분석 기법의 이용이 상대적인 단순성과 더불어서 오픈 소스인 하둡 시스템의 경제성 있는 구현이 연구의 가장 큰 목표였고, 실험적으로 충분히 구현 및 실현 가능성을 보여 준 제안 이었다고 사료되어진다.

향후 좀 더 연구해야 할 관점으로서는, 교육계에서의 이런 분석에 대한 관심과 더불어서 어떻게 하면 실사회 학생들 사이에서 무섭게 확산되고 있는 사회 망의 이용률과 교사들이 이 망을 통한 학생들과의 대화를 얼마나 자연스럽게 유도할 수 있을까 하는 점이고, 본 연구의 제안인 폭력성 언어의 사용빈도 측정이 자살사고의 미연방지와 밀접하게 연관된다는 점을 학교 현장에서 직접 적용하는 방법을 연구 할 필요성이 있다.

참 고 문 헌

[1] 김민경, (2011), **학습동기, 자기 효능감 및 흥미가 수학에서의 비정형 문제 해결력에 미치는 영향**, 중앙대학교 대학원 논문집

[2] 김민주, (2014), **하인리히법칙**, 미래의창

[3] 이영직, (2009), **세상을 움직이는 100가지 법칙(하인리히에서 깨진 유리창 까지)**, 스마트비지니스

[4] 김민주, (2008), **하인리히 법칙**, 토네이도

[5] 박종만, 엄태원, 김하진 (2012), **빅 데이터 분석기술동향과 활성화 과제**, 한국통신학회 논문지 제29권 제11호

[6] 김상락, 강만모, (2014), **빅 데이터 분석 기술의 오늘과 미래**, 한국정보과학회지 32(1) 8-17 1015-9908

[7] 박대민, (2013), **뉴스기사의 빅 데이터분석 방법으로서 뉴스 정보원 연결망 분석**, 한국 언론학보 제57권 6호

[8] F. Smadia, K. R. McKeownnn. V. Hatzivassiloglou,(1996),**Translating collections for bilingual lexicons: A statistical approach**, Computational Linguistics

[9] K. Frantzi, S. Ananiadou, H. Mima, (2000) **Automatic recognition of multi-words terms: the C-value /NC-value method,** International Journal on Digital Libraries.

[10] S. K. Song, et.al., (2011), **Multi -words Terminology Recognition Using Web Search,** Communications in Computer and Information.

[11] Wang, Y., Wang, S., (2010) **Research and Implementation on Spatial Data Storage and Operation Based on Hadoop system,** Proc. of Int. Conf. on Geoscience and Remote Sensing

[12] **MapReduce Tutorial,** (2008), The Apache Software Foundation,

[13] Borthakur, D., (2007) **The Hadoop Distributed File System Architecture and Design,** The Apache Software Foundation

[14] **Splunk 5.0.2 Tutorial,** (2013), Splunk System

[15] Cilibrasi, R., Vitanyi, P., (2007), **The Google Similarity Distance,** IEEE. Trans. Knowledge and Data Engineering Vol.19, No.3,

[16] Frantzi, K., Anariadou, S., (2000), **Automatic recognition of multi -word terms : the C-value method** International Journal on Digital Library

김 영 택

1982 Fairleigh Dickinson Univ. Computer Sci.(학사)
1983 Polytech Univ. Computer Sci.(석사)
1986 Polytech Univ.(박사수료)
1987 수성대학교 전자계산학과
1988~현재, 경성대학교 컴퓨터공학부
관심분야: 인공지능
E-Mail: ytkim@ks.ac.kr