Journal of The Korea Society of Computer and Information
Vol. 19, No. 12, December 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.12.101

웰니스를 위한 빅데이터 분석과 의료 질 관리

조 영 복*, 우 성 희**, 이 상 호*

The Big Data Analysis and Medical Quality Management for Wellness

Young-Bok Cho*, Sung-Hee Woo **, Sang-Ho Lee *

요 약

의학기술의 발전과 소득수준의 증가로 “건강하게 오래살기”에 관심이 높아지면서 적극적으로 건강을 증진하고 유지하는 웰니스가 확대되고 있다. 또한 맞춤형 의료서비스에 대한 수요가 증가하고 방대한 의료 빅 데이터를 이용한 질병 예방의 움직임도 나타나고 있다. 이 논문에서는, 의료 시장에서 주요 관심분야로 부각되고 있는 웰니스를 지원하기위해 빅 데이터 기반의 의료 질 향상을 통한 환자중심의 의료서비스를 목적으로 한다. 환자를 약물에 의존적으로 치료만 하는 것이 아니라 식생활 개선을 기반으로 질병예방과 치료를 위해 빅데이터를 분석한다. 개인 트윗터를 분석해서 일상생활정보를 획득하고 웰니스 사전을 기반으로 질병예방과 치료를 목적으로 한다. 효율적인 빅데이터 분석을 위해 하둡노드를 증가하면서 데이터 처리시간을 실험하였다. 실험결과 저장시간의 경우 63%, 데이터 통합의 경우 18%, 전체 테스트 시간을 기준으로 26%로 하나의 노드로 처리하는 경우보다 세 개의 노드로 처리하는 것이 효율적임을 실험을 통해 확인하였다.

▸ Keywords : 클라우드컴퓨팅, IaaS, 개인의료정보레코드, 병원정보시스템, 웰니스

Abstract

Medical technology development and increase the income level of a "Long and healthy Life=Wellness," with the growing interest in actively promoting and maintaining health and wellness has become enlarged. In addition, the demand for personalized health care services is growing and extensive medical moves of big data, disease prevention, too. In this paper, the main interest in the market, highlighting wellness in order to support big data-driven healthcare quality through patient-centered medical services purposes. Patients with drug dependence treatment is not to diet but to improve disease prevention and treatment

•제1저자 : 조영복 •교신저자 : 우성희
•투고일 : 2014. 7. 22, 심사일 : 2014. 9. 5, 게재확정일 : 2014. 10. 2.
* 충북대학교 소프트웨어학과(Dept. of Computer Science, Chungbuk National University)
** 한국교통대학교 의료정보공학과(Dept. of Medical Informatics&Engineering, Korea National University of Transportation)

based on analysis of big data. Analysing your Tweets-daily information and wellness disease prevention and treatment, based on the purpose of the dictionary. Efficient big data analysis for node while increasing processing time experiment. Test result case of total access time efficient 26% of one node to three nodes and case of data storage is 63%, case of data aggregate is 18% efficient of one node to three nodes..

▸ Keywords : Cloud Computing, Infrastructure as a Service, Personal Healthcare Record, Hospital Information System, Wellness

I. 서 론

최근 모바일 인터넷과 소셜 미디어의 확산으로 소셜 네트워크에 대한 관심도가 증가하고 기하급수적으로 증가하는 데이터양은 다양한 소비체계에 변화를 주고 있다. 특히 정보통신기술(Information Communication Technology:ICT)이 다른 산업들과 융·복합되면서 빅 데이터의 활용은 매우 중요한 과제로 떠오르고 있다[1,2]. 이런 가운데 많은 사람들이 SNS(Social Networking Service)를 이용하면서 트위터(Tweeter)는 동영상 및 음성, 텍스트 등의 비정형 데이터를 생성하고, SNS에서 생성되는 정보들은 서비스의 특성상 사용자의 주관적인 의견이나 개인 생활에 대한 내용을 단어를 이용해 표현된다. 또한 전 세계적으로 의학 기술의 발전과 함께 질병의 예방, 진단과 치료 등 관심이 증가하고 우리나라 국민의 평균 기대수명이 증가하면서 함께 증가한 만성질환은 장기간 지속되어 의료산업 발달과 의료서비스 이용의 급격한 증가에 결정적 영향을 미치고 있는 것이 현실이다[3]. 보건의료 서비스는 질병으로 인한 치료의 의미뿐만 아니라 만성적인 질병에 대한 예방의 개념을 포함한다. 예방적 보건의료서비스는 질병이 발생하는 것을 예방하는 모든 서비스로 질병 발생 이전에 주로 일반적인 건강상태의 향상을 위해 제공되는 건강증진 서비스를 비롯하여 특정 발생을 예방할 수 있는 모든 서비스를 의미한다[4]. 헬스케어 분야에서는 빅 데이터를 활용해 환자의 의무기록은 물론 식생활습관, 직업이력, 등 광범위한 데이터 분석을 통해 치료방법을 개선하고 의료비용을 줄이려는 노력이 진행되고[4] 있으며 이것은 전 세계적으로 생활수준의 향상과 높은 삶의 질을 성취하고자 건강관리 패러다임이 변화되면서 예방 중심의 웰니스(Wellness)에 대한 관심이 증가한다[4,5,6]. 이 논문에서는 개인 트위터 등 빅 데이터 분석을 기반으로 일상에서 발생되는 실시간 빅 데이터 정보를 분석하여 보건의료 서비스의 질을 향상시킨다. 실시간으로 발생되는 개인 트윗터를 수집한 후 일상생활정보를 획득하고 수집된 정보는 하둡 분산 파일 시스템(HDFS)과 맵-리듀스(Map-Reduce)에서 구축된 웰니스 사전을 기반으로 특정 단어나 패턴을 추출하여 핵심 키워드를 분류한다. 이렇게 분류된 키워드를 기반으로 환자의 질병예방과 치료를 위해 저장/관리되고 저장된 정보들을 기반으로 보다 정확한 진단과 처방이 가능하도록 개인의료정보 데이터를 기반으로 환자중심의 의료서비스를 제공한다. 제안 논문에서는 효율적인 빅 데이터 분석을 위해 하둡노드를 증가하면서 빅데이터에서 발생된 데이터 처리시간을 실험하였다. 실험결과 데이터 저장시간의 경우 63%, 데이터 통합의 경우 18%, 전체 테스트 시간을 기준으로 26%로 하나의 노드로 처리하는 경우보다 세 개의 노드로 처리하는 것이 효율적임을 실험을 통해 확인하였다.

이 논문의 구성은 2장에서는 관련연구로 빅 데이터와 웰니스에 대한 연구동향과 근거중심보건의료와 의료의 질에 대해 기술한다. 3장에서는 빅 데이터를 기반으로 의료정보서비스 질 향상 방안을 제안한다. 마지막으로 4장에서는 결론으로 구성한다.

II. 관련 연구

1. 빅데이터와 웰니스

빅 데이터(Big Data)는 소셜 미디어의 성장과 스마트폰 등 다양한 휴대용 모바일 장치에서 생성되고 사용되는 지식의 확산 등으로 주목받기 시작했다. 그러나 이렇게 광범위한 범주 내에서 정확하고 신속하게 원하는 데이터를 얻는 것이 점점 어려워지고 있다는 문제점을 갖는다. 따라서 현재 제공되

는 기술보다 더 효과적인 데이터 저장, 검색, 분류, 처리, 분석 방법이 요구되고 있는 것이다. 최근 들어 빅 데이터의 활용이 사회적 이슈가 되면서 중요성이 점점 더 증가하고 있다. 빅 데이터의 효과적인 활용은 새로운 지식 생산이 가능하고 사회경제적 가치를 창출할 수 있기 때문이다. 현재 해외 기업들의 다양한 빅 데이터 활용이 이루어지고 있는데 민간 분야뿐 아니라 정부를 포함한 공공 부문에서도 빅 데이터를 활용하기 위해 노력하고 있다[1,2]. 또한 맞춤형 의료서비스에 대한 수요가 증가하면서 방대한 의료 빅 데이터를 이용한 질병 예방 움직임도 나타나고 있으며 치료에서 예방으로 의료공급자 중심에서 의료 소비자 중심으로 패러다임이 변화하고, 국내·외 의료계에서 빅 데이터가 활용되는 사례들도 점점 늘어나고 있다. 영국의 국가건강서비스(National Health Service)에서는 전국의 약국과 병원의 처방 데이터를 수집하여 국민 건강에 대한 예측을 수행하고 CPRD(Clinical Practice Research Data link)라는 사이트를 통해 다양한 데이터를 연구자들에게 제공하고 있다[2]. 현재 이용 가능한 데이터는 1차 및 2차 의료 기관으로부터 수집된 질병 등록 자료를 모두 포괄하고 있으며, 인구학적, 사회경제적 변수가 포함되어 있다. 이와 같이 영국 의료계는 빅 데이터 활용이 자리를 잡아가고 있다. 연구자들은 현재 유행하고 있는 질병의 발생 장소 및 전염 속도, 주요 질병의 분포, 연도별 증가 등에 대한 통계치를 확보하여 최종적으로 효율적이고 신속한 질병 관리가 가능해지게 해준다. 질병 예측과 예보의 측면에서도 다양한 빅 데이터 활용 사례가 존재한다. 미국의 존스 홉킨스 대학에서는 소셜 미디어인 트위터를 이용하여 질병 예보 시스템을 개발했으며, 이는 인플루엔자부터 알레르기까지 다양한 종류의 질병 추적이 가능한 기술을 구현하게 되었다. Seton Health Care Family와 IBM 공동개발 솔루션은 연간 200만 명 환자의 진료 정보를 분석, 추적하여 환자가 미래에 겪을 수 있는 질환, 증상을 예측하였다. 더불어 IBM은 심혈관 질환 예측을 통해 심근경색 발병 위험을 줄이는 솔루션도 개발할 수 있었다. 국내에서도 해외 사례들을 바탕으로 빅 데이터를 의료계에서 활용하고자 기술적인 도입이 이루어지고 있다.

웰니스(Wellness)란 건강한 상태를 유지하고 웰빙(well-bing)을 위한 잠재력을 극대화하기 위한 체계적인 노력(process)이다[4]. 웰니스 산업에 대한 관심은 점차 증가하는 추세인데 행복 추구 및 삶의 질을 목적으로 하는 비의료 영역에 IT 융합기술을 적용하는 건강 서비스가 선보이는 것도 이런 추세와 무관하지 않다. 세계 웰니스 시장은 미용 및 노화방지, 피트니스, 영양 및 체중 감량 등의 분야를 중심으로 많은 관심을 받고 있으며 특히 일상에서 웰니스 영역인 건강관리 예방, 진단, 사후관리 시장규모는 [표 1]과 같이 급성장할 것으로 예상하고 있다[5].

표 4. 영역별(예방·진단·치료·관리)산업규모 전망
Table 1. Specific Industrial scale view

(단위 : 억달러)

구분	예방	진단	치료	사후관리	합계
2010년	2,140	5,700	24,240	3,560	35,640
2015년	2,980	9,190	31,420	5,100	48,690
2020년	6,860	14,400	39,110	8,230	68,600

최근에는 시장의 니즈가 다양화, 다각화되고 개인주의가 심화됨에 따라 시장이 지향하는 웰니스의 가치가 더욱 중요해지고 있다. 미국의 HP2020(Healthy People 2020)은 미국 국민이 건강한 삶을 능동적으로 유지, 증진 할 수 있는 환경조성, 미래의 질병위협으로부터 자유롭기 위한 건강한 성장과 생활습관 확대 등을 총괄 목표로 제시하면서 정책방향이 의료서비스의 보편화 중심에서 웰니스 문화와 환경 도입 중심으로 전환되고 있는 것이다. 또한 마이크로소프트사는 미국내 거주자에게 무료로 제공되는 개인용EHR 서비스 HealthValut 제공을 통해 사용자의 혈액검사, 백신 기록 및 병력 등의 의료정보를 온라인에 저장하고 자신의 건강관리 및 응급 시 주요 의료 이력작성, 이들 정보의 의료기관이나 보험회사 등으로의 제공이 가능하게 되었다. 또한 파트너 기관으로부터 진료기록, 처방기록이 CCR또는 CCD형식으로 제공되고 이런 정보는 HealthValut에 기록되며 의료기기와 연결하여 데이터를 입력하는 형태의 서비스를 제공하고 있다[7].

유럽의 경우 2008년부터 고령층에게 IT기기를 활용한 건강관리와 긴급의료서비스를 제공 받을 수 있는 AAL(Ambient Assisted Living) 프로젝트를 추진하고 있다. AAL 프로젝트는 IT기술을 통해 고령자의 삶을 증진시키고 독립적인 생활을 할 있도록 도움을 주는 것을 목적으로 고령화의 가속화와 함께 급속도로 커지고 있는 블루오션인 웰니스 시장을 선점하고 재정 부담이 가중되는 것을 막겠다는 의지로 지속시키고 있다.

일본에서는 '전국 어디서나 과거의 진료 정보에 근거한 의료를 받을 수 있는 동시에 개인이 건강관리에 대응할 수 있는 환경을 실현하기 위해서 국민이 자기의 의료·건강 정보를 전자적으로 관리·활용하기 위한 전국 수준의 정보 제공 서비스를 창출한다'는 목표로 특정 건강진단, 특정 보건지도를 의무화하는 제도가 시작되면서 건강에 대한 관심이 증가하기 시작하였다. 2008년 4월부터 40세 이상 74세까지 가입자를 대상

으로 잘못된 생활습관으로 인한 질병예방을 위한 특정건강검진, 특정보건지도를 실시하도록 의무화가 시작되면서 건강관리에 대한 수요가 증가하고 Medical과 Wellness 서비스가 결합된 의료기록 관리서비스(PHR:Personal Health Record) 서비스를 추진하고 있다[6,7,11].

국내에서는 2000년 이후 u-Health 분야를 중심으로 서비스를 구현하였으나 국내의 건강관리 서비스 관련 법·제도가 일관성 있게 체계화 되고 관련 법률 간 연계조정이 되지 않아 적극적인 웰니스 서비스 시장 도입이 지연되고 있는 상황이다. 국내 기업들은 웰니스 서비스에 대한 대규 트자와 사업 참여를 모색하고 있으나 "의료행위" 위반 여부에 대한 불명확성으로 인하여 본격적인 투자가 소극적인 상황이다.

2. 근거중심의학과 의료의 질 관리

근거중심의학(Evidence Based Medicine :EBM)이란 최상의 연구근거(the best research evidence)를 의사의 숙련도(clinical expertise)와 환자의 가치 (patient's unique values and circumstances)에 접목시킨 것이다[3,8,13,14]. 최상의 연구 근거는 진단검사의 평가 연구, 예후인자의 예측력 평가 연구 및 치료, 재활, 예방서비스 효능을 평가하는 연구 등을 모두 포함하는 다양한 임상관련 연구에서 얻어지는 결과를 의미한다. 의사의 숙련도란 환자의 건강상태를 살펴 진단을 내리고, 환자가 받을 치료로 인한 편익과 위험을 예측하며, 치료에 대한 환자의 선호와 기대치를 파악하고, 여기에 임상경험과 기술을 적절하게 조화시킬 수 있는 능력의 정도를 의미한다. 마지막으로 환자의 가치란 치료과정에서 환자가 갖게 되는 처치에 대한 선호도 및 관심, 치유에 대한 기대치를 의미한다[3]. 우리나라에서는 1997년 세계가정의학회 아시아태평양학술대회에서 Smith Richard에 의해 '임상진료지침 및 근거중심의학'이라는 강연이 소개되면서 각종 논문이나 서적 등을 통해 EBM에 대한 연구가 진행되었으며, 의학 분야 이외에도 '근거중심'의 'Evidence based'라는 용어를 다양한 학문분야에 적용하여 근거중심보건의료(Evidence based Health care:EBH), 근거중심간호(Evidence based nurse:EBN), 근거중심 심혈관의학(Evidence based Cardiovascular Medicine:ECM)등으로 확대되었다. 2008년 보건복지부와 대한의학회가 서로 협력하여 임상진료지침 정보센터를 설립하는 등 임상진료지침의 개발에 박차를 가하면서 EBM의 중요성이 한층 더 대두되기 시작하였다. 특히 의료기관의 평가, 건강보험 적정성평가 제도 등 의료의 질 평가 정책들이 도입되면서 EBM에 대한 인식이 점차 확산되었으며 임상 의사들의 새로운 관심의 대상으로 떠오르고 있다. 의료분야에서 환자의 권리 신장 및 제한된 보건의료 자원의 적절한 활용 필요성과 더불어 EBM이 현재 보건의료가 안고 있는 문제를 해결 할 수 있는 하나의 방법론으로 각광을 받게 하는 요인들로 작용하고 있다"고 하여, EBM은 이제 선택이 아닌 필수요소로, 어떻게 시행하느냐가 문제로 다가오고 있음을 강조하였다. 한편 근거중심의학과 질 개선(quality improvement) 모두 근거와 실무의 간극을 줄이려고 한다는 점에서 공통적 특성이 있지만 기본 원리와 활동방식에 차이가 존재하고 있다.

이처럼 중재효과에 대한 정보를 주는 연구근거(research evidence)를 "global knowledge"혹은"global evidence"로 부르기도 한다. 반면에 질 개선 영역에서는 임상적 진료과정 혹은 보건의료체계의 변화를 추구하고, 반복적으로 나타는 체계적 오류를 발견하고 개선하는 것이 요구되며, 여기에는 진료과정이나 보건의료체계에 대한 정보의 과정적 지식(process knowledge)도 필요하다. Glasziou는 이와 같은 관계를 다음과 같이 제시하였다. 근거중심보건의료가 보건의료계의 주요한 흐름으로 정착되면서 임상진료지침도 근거중심적 방법을 통해 개발되기 시작했다. 앞서 언급한 체계적 고찰, 메타분석, 비용효과성 평가 등은 질병의 예방, 치료, 진단 및 재활에 필요한 중재를 권고하는데 있어 주요한 근거가 된다.

III. 빅 데이터를 기반으로 한 의료서비스의 질 관리

오늘날 의료산업은 첨단 IT 기반의 언제 · 어디서나 사용 및 접근이 가능한 스마트 기기를 이용해 다양한 미디어 콘텐츠를 통해 과거 치료중심의료 서비스에서 질병의 예방 및 관리에 초점을 맞춘 인간중심의료 서비스로 패러다임이 변화하고 있다[1,2,9]. 스마트 기기를 통한 원격진료는 의료서비스에 있어 과거 문제시 되었던 소외지역의 접근성 저하, 병원중심의 서비스 제공 등의 수동적 서비스에서 발전하여 지능화된

표 2. 의료서비스 발전에 영향을 미치는 IT기술
Table 2. IT affects the development of medical services and technology

ICT 기술	영향
가격-성능 개선	의료장비의 저렴화
이동통신	접근성, 즉시성의 향상
인터넷	네트워크 헬스케어의 활성화
소형화	센서, 모니터 등의 소형화
영상진단기계	MRI, CT, OCT
의료데이터 관리	EMR, PHR&Healthcare Big Data
융합기술	Pharmaceutical &devices

서비스 기술력이 급속도로 발전하고 있다. 〈표2〉는 의료서비스 발전에 영향을 미친 IT 기술을 정리한 것이다.

빅 데이터 분석 활용의 효과는 우선 경쟁 환경의 이해가 선행되어야 한다. 예를 들어 정보가 발생하는 소셜 네트워크의 구조와 정보전달 패턴의 파악이 매우 중요하다. 또한 트위터들의 소셜 네트워크 구조를 파악하여 소셜 미디어에 나타나는 정보의 경로를 분석한 후 다양한 커뮤니티 구조를 파악하는 것이 중요하다. 빅 데이터를 분석하기 위해서는 오픈소스 기반의 분산 데이터 저장 기술인 하둡(Hadoop)이 기존 데이터베이스로 관리하기 힘든 규모와 성격의 데이터를 처리하기 위한 기술로 주목받고 있다[9,10,11]. 이 논문에서는 개인의 소셜 미디어를 의료 질 향상 도구로 활용하기 위해 트윗 간의 연관 규칙을 기반으로 웰니스 사전을 구축하고 질병예방과 치료를 목적으로 생활패턴을 분석하여 웰니스를 위한 의료질 향상을 제안한다. 연관규칙을 이용한 소셜 미디어의 빅 데이터의 분석은 트위터들의 검색엔진을 통해 수집된 데이터를 축적하고, 이를 이용해 환자군의 일상생활 패턴을 분석한 후 웰니스를 위한 라이프 트래킹 측면의 건강이나, 라이프 사이클에 필요에 다양한 연관성이 있는 부분들을 개인의 성향에 맞게 분석하여 개인의료정보 레코드에 활용한다.

3.1 하둡 기반의 SNS 분석

하둡은 대용량의 데이터를 분산 처리하게 해주는 아파치 오픈소스 프로젝트로 맵-리듀스(map-reduce) 프레임워크를 기반으로 맵과 리듀스라는 두 개의 메서드로 구성된다. 맵 메서드는 키(key) 값을 읽어 필터링하거나 다른 값으로 변환하는 작업을 수행하고 리듀스는 맵 함수를 통해 출력된 결과 값을 새로운 키를 기준으로 그룹화 하여 집계 연산을 수행한 구성목록을 생성한다. [그림 1]은 제안 시스템에서 구성한 하둡 데이터 처리를 위한 구성으로, 대용량 데이터의 분산 저장 및 신속한 처리를 위해 다수의 컴퓨터를 네트워크로 연결하여 하나의 시스템과 같이 사용할 수 있도록 구성한다. 트위터가 작성한 트윗을 기반으로 특정 단어나 패턴을 추출하고 핵심 키워드를 분류한다.

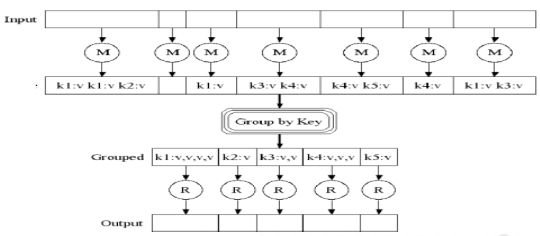

그림 1. 하둡 데이터병렬처리 구성
Fig. 1. Hadoop data parallelism configuration

[그림 1]의 입력 값으로 트윗에서 추출된 단어를 연관검색을 수행하기 위한 키(key)와 값(value)으로 매핑하여 그룹화하고 새롭게 생성된 그룹이 웰니스 사전 중 어느 영역에 포함되는지 관계를 검색하여 최종 출력으로 전달한다. 최종 출력단계에서 생성 결과는 웰니스 사전의 항목으로 추가된다.

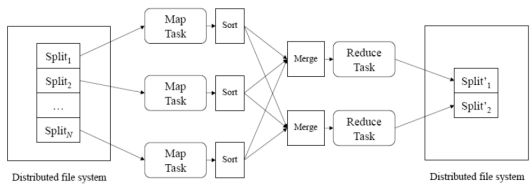

그림 2 맵 리듀스의 실행
Fig 2. Run Map-Reduce

[그림 2]는 제안 논문에서의 맵 리듀스 모듈로 워드 카운트(word count) 모듈을 통해 로그파일의 텍스트 문장을 키와 값으로 구성한다. 이것은 웰니스 정보를 추출하기 위한 것으로 웰니스 사전을 기반으로 트윗에서 크롤링된 데이터를 [그림 1]과 같이 하둡에서 그룹핑을 [그림2]의 분산화일 시스템에서 각각의 키와 값으로 우선순위를 정렬하고 중복데이터 제거를 수행하는 맵 리듀스 실행 과정을 나타낸 것이다.

SNS 사이트로부터 검출된 자료는 SOAP 메시지 파서에서 수집된 SOAP Body부분의 엘리먼트 이름과 값을 해쉬 테이블의 키와 값으로 저장한다. 이 엘리먼트는 사용자의 이름(name)과 몸무게(weight), 성별(sex), 지역(area)등의 정보를 가지고 HDFS에 저장된다. 워드카운트는 맵 리듀스 분석을 통해 웹 로그 및 쿼리 로그에서 단어의 빈도를 확인 할 수 있다. R 오브젝트를 기반으로 각 트윗의 단어 빈도를 계산하고, rhocollect 함수를 통해 개별 단어와 빈도수를 하둡 시스템으로 전송한다. [그림2]에서 1차적으로 계산된 리듀스 타스크(reduce task)를 단어(key)와 빈도(value)의 누적합으로 계산한 후 최종 결과 값을 HDFS에 분산 저장 한다. [그림 3]은 맵 리듀스에 HDFS에 저장된 데이터를 병렬로 처리하고 특징 값을 추출하기 위한 알고리즘이다. 맵 리듀스를 관리 동작시키고 Job의 메서드를 통해 매퍼/리듀서 클래스, InputFormat, OutputFormat 클래스, 리듀스 출력 포맷 클래스를 지정한다.

```
public class ACCDriver {
   public static void main(String[] args) throws Exception {

      Configuration conf = new Configuration();
      if (args.length != 2) {
         System.err.println("Usage: ACCDriver <input> <output>");
         System.exit(2);
      }

      conf.set("<tag>", "<AnnotatedECG");
      conf.set("</tag>", "</AnnotatedECG>");

      Job job = new Job(conf, "ACCDriver");

      job.setJarByClass(ACCDriver.class);
      job.setMapperClass(ACCMapper.class);
      job.setReducerClass(ACCReducer.class);

      job.setInputFormatClass(XmlInputFormat.class);
      job.setOutputFormatClass(TextOutputFormat.class);

      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(Text.class);

      job.setPartitionerClass(UserPartitioner.class);
      job.setCombinerClass(ACCReducer.class);


      FileInputFormat.addInputPath(job, new Path(args[0]));
      FileOutputFormat.setOutputPath(job, new Path(args[1]));

      job.waitForCompletion(true);
   }
}
```

그림 3. ACCDriver 소스코드
Fig. 3. ACCDriver Source Code

3.2 웰니스 데이터 수집

웰니스 데이터 분석을 위해 다수의 SNS 계정에 트윗 메세지를 수집한다. 수집된 데이터는 크롤링을 통해 SNS의 팔로윙된 데이터를 수집하고 이를 다시 서버로 전송한다. 〔그림 4〕는 크롤링된 메시지를 기반으로 웰니스 데이터 수집과 유효 단어의 판단 단계를 나타낸 것이다.

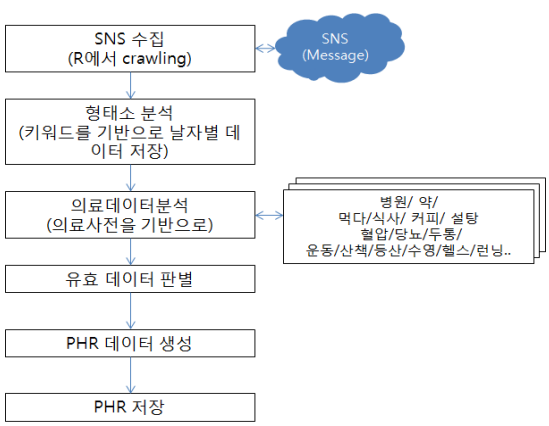

그림 4. SNS 데이터 크롤링 수집/판별/저장단계
Fig. 4. SNS Data crawling correction/distintion/save

〔그림4〕와 같이 크로롤링한 트윗의 형태소를 분석하고 웰니스 정보로 추출된 단어의 유효성 판단을 거쳐 PHR 데이터에 추가하여 wPHR을 생성한다. 〔그림 5〕는 wPHR을 하둡의 HDFS의 분산순차 저장 시스템 구성도이다. 클라이언트는 SNS에서 수집된 데이터를 네임노드의 하둡 파일 시스템(wphr)으로 생성하고 생성 데이터 목록을 반환한다. HDFS는 파일의 복제 개수만큼 데이터 노드를 설정하고 네임노드로 SNS 데이터를 전송한다. 네임노드는 데이터 노드에서 전송받은 메시지를 로컬 디스크에 저장한 후 순차적으로 HDFS의 데이터에 저장한다. 웰니스 데이터 분석은 기존 자연어처리(NLP) 기법을 이용하여 인간의 언어로 쓰인 텍스트 문장을 분석하고 문장에서 주어진 웰니스 정보(의료/건강 관련 용어)를 찾아내어 빈도를 계산한다.

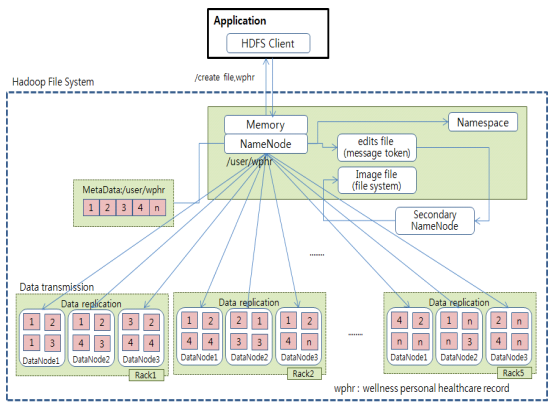

그림 5. HDFS 분산 저장
Fig. 5. HDFS distributed storage

IV. 제안모델 실험 및 평가

SNS의 트윗을 데이터 크롤링으로 수집된 하둡 시스템에 저장된 메시지를 문장단위로 분리하고 각 문장을 형태소 분석 후 토큰 배열로 변환한다. 변환된 토큰을 웰니스 사전에 웰니스 데이터로 수집하여 저장한다.

1. 실험 환경

이 논문의 실험은 〔표 3〕과 같이 실험환경을 구축하였다. cloudPHR 서버는 Inter server와 MS windows 7 그리고 VMware 9.0.2 버전을 사용해 SNS 데이터 쿼리를 조회하였다〔14〕. cloudPHR 서버는 하둡 파일 시스템과 데이터를 전송하여 레코드 생성을 위한 미들웨어 역할을 수행한다.

표 3. 실험환경
Table 3. Test Environment

항목	내용
HW	Intel(R) Core 16 2.4GHz, 16GB
OS	Linux kernel 3.8.2 Ubuntu 12.04.3 / Windows 7 32bit
DB	Mysql 5.5
tool	R i386 3.1

하둡파 일 시스템에서 분산처리 저장된 생체정보의 맵 리듀스 프로그래밍에서 맵퍼 코드를 이용해서 클래스는 키 값을 날짜로 지정하고 값은 웰니스 데이터 분석을 수행한 후 빈도를 계산하여 지정한다.

2. 평가

실험을 위해 두 사람의 트위터 데이터를 2014년 5월1일 ~ 8월 10일(102일간) 분석한 결과 〔표 4〕와 같다. 사용자1은 만성위궤양으로 병원에서 처방받은 약을 복용하는 40대 초반 여성이고, 사용자2는 갑상선 질환으로 치료를 받고 있는 40대 후반 여성으로 트위터를 분석한 결과이다.

표 4. SNS 트위터 문장 분석
Table 4. SNS Twitter sentence analysis

웰니스 사전	사용자1	사용자2
커피	264	86
초콜릿	63	18
외식/맛집	78	43
케익/빵	58	19
두통	37	47
눈충혈	10	19
소화/배부름/더부룩함	64	3
체중증가/감소	23	36
등산/산책	7	29
피곤함/졸림	31	63
퇴근/출근	28	41
건강	27	48

실험을 위해 웰니스 사전은 〔표 4〕와 같이 구성하였다. 사용자의 트위터를 통해 유병자들의 생활패턴과 치료 패턴 및 약품 복용에 대한 정보를 획득할 수 있다. 트위터에 남긴 문장을 분석해서 치료받는 환자들의 생활 패턴 정보를 확인할 수 있고, 또한 보균자의 경우 질병 발생률을 최소화 할 수 있도록 추적이 가능하기 때문에 예방의학 및 근거중심의학에 매우 효율적이라고 할 수 있다. 사용자1의 경우 추가적인 위경련으로 병원을 찾게 되면 담당의사는 환자의 생활패턴을 확인하고 "너무 잦은 카페인 섭취나 외식은 건강에 도움이 안 된다"는 것을 한 번 더 지적하고 사용자1의 생활패턴 변화를 유도해나감으로 의약품에 의존하기보다는 생활의 변화를 통한 환자중심의 치료가 가능하게 될 것이다.

표 5. 노드수에 따른 분석 소요시간(ms)
Table 5. analysis time of the number of nodes

Node	Data Storage	Data Aggregate	Total
N1	359,160	1,664,356	2,023,516
N2	334,650	1,328,975	1,663,625
N3	132,688	1,356,981	1,489,669

〔표 5〕는 클러스터 사이즈, 즉 하둡 노드수가 증가함에 따라 데이터를 분산 저장하는 과정(data storage)과 분산된 노드에서 분석하고 최종적으로 결과를 취합하는 과정에서 소요되는 시간을 측정한 결과이다. 〔그림 6〕은 〔표 5〕의 분석결과를 도식화하여 나타낸 것이다. 3개의 노드를 연결한 경우 data storage 과정은 63%와 data aggregate 과정은 18% 가량의 성능이 향상되었다. 전체 분석 시간은 26%이상 향상되는 결과를 보였다. 따라서 수집된 데이터의 양이 증가할수록 노드수를 증가하여 분산 처리를 수행하는 것이 보다 효율적임을 실험을 통해 확인할 수 있다.

그림 6. 노드수당 실험 결과
Fig. 6. Test result of the number of nodes

5. 결 론

현의학은 순수한 의학 지식 생산에서 질병의 치료, 진단, 예방에 이르기까지 기초와 응용이 융합된 학문으로 이제는 의학만의 힘으로는 결코 해결할 수 없는 문제들이 번번이 발생하게 된 것이다. 학문의 세분화와 전문화는 지식생산의 효율을 높일 수 있으나 개별 학문 각각의 고립을 초래하여 새로운 방향의 지식 생산에 방해가 되는 면도 있다. 그 해법 중 의 하나가 빅 데이터일 것이다. 빅 데이터를 이용한 분석 기법은 기존의 데이터 마이닝을 통한 다양한 분석기법, 기계학습, 인공지능, 연관규칙, 회귀분석 등으로 정형화된 데이터를 토대로 분석기법이 발전되어 왔고, 현 시점에서 의사결정을 가장 과학적으로 검증하는 도구로 활용되어 오고 있다. 하지만 서론 부분에서 논한 것처럼 소셜 미디어의 발달, 그리고 스마트 IT기기의 발달에서 실시간으로 대량의 데이터를 유발 하는 IT 환경의 패러다임에서 조금 더 기존의 데이터 마이닝을 통한 분석 기법들을 정형화 되지 않은 데이터들을 비즈니스적인 활용하고 분석하는데 초점이 맞추어져 있다. 빅 데이터의 분석은 무궁무진한 데이터 분석의 패러다임을 바꾸어줄 획기적인 대용량 데이터 처리기술 임에는 틀림이 없다.

이 논문에서는 이와 같이 분석된 빅 데이터의 결과를 기반으로 의료 서비스의 질 향상에 툴로 사용하였다. 의료 질 향상을 위해 개인생활에서 발생되는 패턴의 빅데이터 분석을 통해 질병관리를 위한 다양한 측면을 위한 U-헬스케어 서비스를 위해 유병자의 경우 추가적인 합병증이나 질병의 악화를 방지하고 보균자의 경우 질병 발생 원인을 의약품에만 의존하는 것이 아니라 의약품과 병행된 생활 밀착형 의료서비스를 제공함으로 의학에서 말하는 근거중심의료서비스를 통한 진정한 질병치료 및 질병예방이 가능하게 되었다. 향후 이 논문에서 수집된 개인 정보보보를 위한 지속적인 연구가 필요하다.

참고문헌

〔1〕 Sukja Ko, "Health Risk Prediction Using Big Health Data", The Journal of The Korea Institute for Health and Social Affairs, Vol.13, No.11, pp.43-52, 2012.

〔2〕 Taemin Song, " South Korea health and welfare big data trends and proposals ", Science and Technology Policy Institute, Vol.23, No.3, pp.56-73, 2013.

〔3〕 Straus, Sharon E., etal., Evidence-Based Medicine-How to Practiceand Teach EBM, 3thed. Elsevier, London, 2005.

〔4〕 Yongju Park, Gyeongeun Kong, "Wellness centered on big data with our overseas medical industry facts", www.digieco.co.kr, 2013.

〔5〕 Yeonghui Noh etl, "Wellness industry's business model analysis of a study on industrial development-policy report", The National IT Promotion Agency, 2012.

〔6〕 Seongsu Kim, "Building an effective business model of the industry of medical ICT convergence measures for them", www.digieco.co.kr, 2013.

〔7〕 Seongryeol Yoon, "A study on health information communication and security system for PHR service", Gachon University Graduate schools, department of computer engineering, a doctor's thesis, 2013.

〔8〕 Sunhyeong Jung, Jongryeol Park, " Study on Telemedicine system in Medical Law", Journal of The Korea Society of Computer and Information, Vol.17, No.12, pp.241-249, 2012.

〔9〕 Foto N. Afrati, Jeffrey D. Ullman, "Optimizing Multiway Joins in a Map-Reduce Environment," IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.9, pp.1282-1298, 2011.

〔10〕 Pramod Bhatotia, Alexander Wieder, Rodrigo Rodrigues, Umut A. Acar, Rafael Pasquini, "Incoop: MapReduce for Incremental Computations," In Proceedings of SOCC'11, 2011.

〔11〕 Sungsoo Kim, " Study on Big Data Utilization Plans of Medical Institutions", Journal of Digital convergence, Vol.12, No.2, pp.397-407, 2014

〔12〕 Daeseog Heo, "Evidence-based Healthcare in Korea", Korean Medical Association, pp.934-935, 2009.

〔13〕 Yongbin Kim, Problems of Personal Information Protection in Big Data Utilization and an Improvement Method Using PIMS, Kangwon National University industry graduate

school,computer information and telecommunication engineering, a master's thesis, 2013.

〔14〕 Youngbok Cho, Sunghee Woo, Sangho Lee, "The cloudHIS System for Personal Healthcare Information Integration Scheme of Cloud Computing Environment", Journal of the Korea society of computer and information , Vol.19, No.5 , pp.27-35, 2014.

저 자 소 개

조 영 복

2005: 충북대학교
전자계산학과 공학석사.
2012: 충북대학교
전자계산학과 공학박사
현 재: 충북대학교
의학과 박사과정
충북대학교 초빙교수
관심분야: 인증, 정보보안,
의료정보보호
Email : bogicho@cbnu.ac.kr

우 성 희

1993: 충북대학교
전자계산학과 이학석사.
1999: 충북대학교
전자계산학과 이학박사
현 재: 한국교통대학교
의료정보공학과 교수
관심분야: 침입차단 및 방지,
의료정보보호, 정보보안,
컴퓨터네트워크
Email : shwoo@ut.ac.kr

이 상 호

1989: 숭실대학교
전자계산학과 공학박사.
현 재: 충북대학교
소프트웨어학과 교수
관심분야: 컴퓨터네트워크, 정보보호,
데이터통신
Email : shlee@cbnu.ac.kr