한국컴퓨터정보학회지
제19권 제2호, 2011. 12.

Big Data 관리를 위한 SNS포탈의 기술동향

• 김시우(숭의여자대학 인터넷정보과)

I. 서론

빅데이타에 대한 관심은 스마트폰과 SNS의 발전과 더불어 더욱 커지고 있다. 데이타에는 정형화된 데이타와 비정형화된 데이타가 있다. 최근 논의가 되고 있는 Big Data는 그것이 정형화된 것이든, 아니든 상관없이 너무 덩치가 커서 어떻게 할 수 없는 데이타를 말한다. 특히 기업에서 기존에 모아둔 분석 데이타들과 최근에 다양한 고객들의 활동 정보 - SNS에 기반한 - 들은 대부분 비정형화 Data이거나 정형화하기 전에 먼저 Data로 추출되어 쌓인 데이타들이다. 해외의 한 조사에 따르면, 기업에 쌓이는 개인정보의 양은 1.2년마다 두배씩 증가하는 것으로 나타났다. 웹 2.0, IPTV, 모바일 컴퓨팅 등 IT 기술의 발전으로 폭발하는 대용량 데이터를 안정적으로 저장하고 신속히 분석하는 것이 기업의 핵심 경쟁력으로 부상하고 있다. 미국의 온라인 쇼핑몰 이베이에는 고객 데이터를 분석하고 가공하는 일을 맡은 직원만 6000명에 달한다.

대용량 정보가 늘면서 정보를 세는 단위도 기가나 테라를 넘어 이제는 페타, 엑사, 제타까지 등장하고 있다. EMC가 지난 6월 말에 발표한 자료에 따르면, 올해 만들어지는 디지털 정보량은 무려 1.8ZE(제타바이트, 10^{21}). 이는 대한민국 국민 모두가 17만 847년 동안 쉬지 않고 매 분마다 트위터에 3개의 글을 게시할 경우 생성되는 데이터량이다.

부산에서 아내를 살해한 혐의로 구속된 대학교수 강모씨는 한 포털 사이트에서 '사체 없는 살인'이란 검색어를 사용한 흔적이 발견돼 용의선상에 올랐다. 경찰은 주요 포털 업체에 공문을 보내 강씨 명의로 가입된 아이디가 있는지 확인한 뒤 접속기록과 검색어 등을 뒤져 이를 확인했다. 강씨가 '흔적'을 남긴 곳은 이곳뿐이 아니었다. 그는 스마트폰 문자메시지 애플리케이션인 '카카오톡'을 통해서도 내연녀에게 "맘 단단히 먹어라"는 내용의 문자 메시지를 보낸 것으로 나타났다. 디지털 공간 곳곳에 자신의 관심사와 개인정보를 '흘리고' 다닌 것이다. 지난해 가수 MC몽 역시 네이버 지식인에 어금니를 어떻게 뽑는지 남긴 질문 때문에 '군 복무를 회피하려 했다'는 의심을 받기 시작했다.

개인들이 인터넷에 남긴 정보는 경찰에만 유용한 것이 아니다. 기업들이 이를 활용하고 있다. 기업들은 웹사이트 방문 기록, 온라인 검색통계, 소셜미디어 소통 기록 등을 그러모아 경영에 활용하고 있다. 한 사람의 개별적인 정보가 수만~수억건씩 모이면 이를 통해 전혀 새로운 패턴을 찾아낼 수 있다. 기업들은 이를 통해 소비자들의 취향이나 행태의 변화를 예측할 수 있다. 미국 구글은 웹사이트(google.org)에서 독감 유행 정보를 예보하고 있다. 독감 증상이 있는 사람이 늘면 '감기' 관련 주제를 검색하는 빈도가 함께 증가하는 것에 착안되어, 시간·지역별 독감 유행 정보를 제공하는 것이다. 구글의 예보는 보건당국보다 앞서 독감 유행 징후를 감지하는 것으로 알려져 있다.

스마트폰 보급 이후에는 위치정보와 결합된 '맞춤 정보'를 제공하는 광고 · 마케팅 기법도 비즈니스의 새로운 패러다임으로 부상했다. 모바일 광고의 가장 첨병에 선 것은 포털. 야후코리아는 2010년 12월 SNS를 통한 사용자 참여형 온라인 광고 '소셜 애드'를 출시했다. '소셜애드'는 SNS 사용자들이 자주 업데이트하는 콘텐츠인 소셜커머스 상품 소식이나 영화 예고편 등과 같은 흥미 및 정보성 광고를 클릭하면 페이스북

이나 트위터와 같은 SNS를 통해 지인들에게 알리고 공유할 수 있는 사용자 참여형 온라인 광고다. 야후코리아는 호주, 대만, 인도 등에서 '소셜애드'를 미리 써본 결과 일반 배너 광고 대비 광고 도달률이 최대 10배 증가한 것으로 나타났다.

또한, FourSquare는 미국에서 가장 빨리 성장하는 위치 기반 서비스 포탈이다. 한국에서도 서비스를 확대하고 있고, 실시간으로 파악되야 하는 위치 정보 시스템은 빅데이타 처리의 어려움을 보여주는 예로 떠오르고 있다.

2011년 구글의 성장은 놀라운 수준이다. 올해 6월 프라이빗 베타 서비스를 시작하고 3개월 뒤 모든 사용자에게 서비스를 공개했고, 첫 번째 API도 릴리즈했다. 더불어 구글 버즈를 폐쇄하고 Ripples 기능 추가했으며 구글 리더, 유투브, 구글뮤직과도 통합했다. 즉 구글 플러스로 모든 기능을 할 수 있게 한 것이다. 이에 트위터는 Facebook에게 SNS부분을 잠식 당하고 구글 플러스에 인포메이션 네트워크시장을 빼앗길 위기이다. 즉 모바일 저널리즘 시장에서 갖고 있는 트위터의 위상이 위협받고 있다. 트위터는 자신의 위치와 협력업체들의 관계도를 명쾌히 볼수 있는 트위터버스를 통해 생존 전략을 표현하고 있다.

현재 구글 플러스 사용자는 5천만명을 돌파했으며 이러한 추세는 역대 어떠한 소셜네트워크 서비스보다도 급속한 성장이다. 구글플러스의 서비스 런칭 및 개편 과정에서 눈여겨 봐야 할 부분들은, 구글이 구글플러스를 위해서 기존 서비스들을 통폐합하며 선택과 집중을 하고 있다는 사실이다.

그림1. 트위터버스

II. 관련 연구

1. 관련연구

1.1 해외 동향

빅데이타에 대한 관심은 아마존북에서 제일 먼저 이루어졌다. 아마존은 검색자의 검색어분석을 통해 특정 사용자의 관심사를 파악하고 이중 실제 구매자에 대해서 더 관심을 기울여서 기존 구매자의 흥미를 유발할 수 있는 새 책을 추천한다.

이를 위해 아마존은 Dynamo를 개발하였다.

구글은 빅데이타 관리를 위해 초기에 MapReduce 라고 하는 프로그래밍 모델과 대용량 데이터 분산처리프레임워크 과 대용량 데이터를 효과적으로 저장하고 확장할 수 있는 GFS (구글파일시스템) 기술을 확보하고 이를 적극적으로 활용하고 있었고, 이를 바탕으로 구글만의 검색기술과 검색서비스를 가능하게 하였으며, 이를 빅 테이블이라고 명명하였다.

구글이 가진 기술을 참고해서 등장한 다양한 맵리듀스프레임워크중에서 가장 주목을 받고 그 기반으로 커다란 에코시스템을 갖추게 된 것이 바로 자바 기반의 아파치 하둡 (Apache Hadoop) 이다. 구글이 발표한 분산 프레임워크 논문을 바탕으로 야후가 오픈소스로 개발한 하둡은 예전 리눅스의 등장으로 OS 시장에 있어서 틀을 크게 바꾸었듯이 빅데이터(대용량데이터)분석 시장에 있어서 커다란 대안으로 등장을 하고 있다. 야후 내부에서 사용하던 이 기술이 오픈소스로 발표되면서 크게 주목을 받으면서 사실상 현재 페이스북, 트위터, 링크드인, 이베이, 아마존 등 많은 글로벌 인터넷, 커머스 업체들은 빅데이터 처리를 위해서 하둡의 사용은 보편적이 되었으며, 이를 기반으로 한 다양한 기술과 도구들이 나타나고 있다.

국내의 대표 포털 네이버, 다음 등 국내 대표 인터넷 기업들 뿐 아니라 S클라우드를 준비하고 있는 삼성전자와 같은 제조사 역시 스마트폰, 스마트 디바이스를 위한 컨텐츠 서비스와 이를 통해서 발생하는 엄청한 로그 데이터 처리에 관심을 기울이고 있다. 최근 인터넷 기업뿐 아니라 글로벌 대기업이나, 금융회사들이 자신들의 트랜잭션 분석이나 사용로그 분석을 위해서 하둡에 대해서 크게 관심을 가지고 있고 오라클, IBM, EMC, SAS 등의 DW 시장의 강자들이 자신들의 솔루션에 하둡을 결합해서 제품과 솔루션을 내놓는 것을 봐도 하둡을 기반으로 하는 대용량데이터분석시장의 큰 변화를

느낄 수 있다.

하둡은 크게 두개의 요소로 나뉘어져 있다.

맵리듀스프레임워크(MapReduceFramework)와 하둡분산파일시스템(HDFS)이다. 하둡은 다양한 분산파일시스템과 연동할 수 있도록 구현되어 있고 대표적으로 아마존의 클라우드 서비스를 이용해서 하둡 어플리케이션을 개발하는 이들은 아마존의 분산파일시스템인 S3 을 이용하고 있다.

하둡으로 데이터 분석 로직을 손쉽게 구현할 수 있는 프로세싱언어인 pig와 SQL과 같은 언어를 제공하는 hive이 등장하였다. 최근엔 오픈소스 통계툴로 유명한 R이 하둡과 연동되면서 이 세가지가 데이터 분석시장을 지배할 것으로 전망된다.

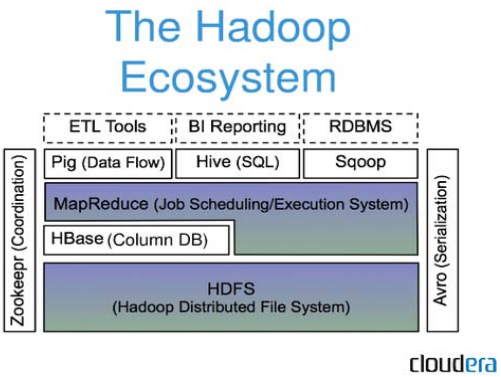

그림 2. 하둡 구조도

아파치 마하웃(Apache Mahout) 프로젝트는 다양하고 중요한 마이닝 알고리즘들을 하둡 프레이워크상에서 구현해서 오픈소스로 공유하자는 차원에서 만들어졌고 현재 0.5 버전이 릴리즈된 상태이다. 이미 많은 사람들이 마하웃의 알고리즘을 직접 이용하거나 최적화해서 자신들의 각 분야에서 활용하고 있다. 향후 아파치 마하웃 프로젝트는 꾸준히 성장해서 시간이 지나면 하둡 기반의 대용량 마이닝 알고리즘을 제공하는 주요 소스가 될 것으로 예상된다.

이와 더불어 하둡파일시스템(HDFS)을 기반으로 하는 대용량 데이터베이스인 HBase 역시 주목을 받고 있다. 이 역시 구글의 BigTable의 아키텍처를 참조해서 만든 오픈소스 대용량 데이터 스토어 기술이다. 최근 NoSQL 데이터베이스라해서 오라클 DBMS, MSSQL, MySQL 과 같은 관계형데이터베이스의 한계 또는 확장성 등의 단점을 해결할 수 있는 대안으로 보다 단순한 아키텍처을 가졌지만 분산컴퓨팅 환경에 적합한 데이터 스토어 기술들이 등장하고 있다. 그 대표적인 것으로 바로 이 HBase을 들 수 있다. 이밖에 하둡파일시스템을 기반으로 하지 않지만 BigTable과 유사한 형태의 Cassandra와 같은 기술들이 함께 주목을 받고 있다.

Cloudera와 같은 경우는 안정적인 하둡배포판을 만들고 컨설팅 및 교육을 하고 있고, HortonWorks 라는 회사가 하둡의 차세대 버전의 아키텍처와 버전 업그레이드를 진행하고 있다. HortonWorks 는 Cloudera 와 달리 하둡 코어아키텍처 차후 버전 개발을 향후 주요 사업으로 발표하였다.

그림 3. 빅데이타 관련 툴들

1.2 국내 동향

국내에서는 대표적으로 넥스알(NexR)이 하둡 및 클라우드 기술을 기반으로 다양한 컨설팅 및 사업을 추진했었고 작년말 KT 에 자회사로 인수되면 크게 주목을 받았다. 넥스알은 꾸준히 국내의 하둡 오픈소스 커뮤니티의 활동을 적극 지원하고 있고, 최근에는 RHive 라고 하는 R 와 Hive 을 결합한 시스템을 오픈소스로 공개하고 해외시장을 노리고 있다. 넥스알의 '빅데이터 어날리틱스 플랫폼'은 정형 및 비 정형데이터에 대한 통합관리가 가능한 하둡기반의 빅데이터 관리 솔루션으로 확장성을 주요 장점으로 내세우고 있다.

그루터는 클라우드 기반 소셜 데이터 분석 등으로 유명한 국내 클라우드 벤처 기업 그루터(www.gruter.co.kr)는 클라우드 컴퓨팅 환경 구축에 필요한 오픈소스 기반 미들웨어를 관리, 모니터링할 수 있는 '클라우몬(Cloumon)'을 최근 출시했다. '클라우몬'은 이러한 어려움을 덜어주기 위해 개발된

것으로, 클라우드 시스템에서 세계적으로 가장 많이 사용되고 있는 오픈소스 기반의 미들웨어인 주키퍼, 카산드라, 하둡, H베이스에서의 취약한 관리 기능을 보완해 준다.

사실 구글, 야후, 트위터, 페이스북, 링크드인 등 웬만한 인터넷서비스 기업들은 자체 팀을 꾸려서 이러한 대용량 데이터 분석 기술과 자신들만의 프레임워크를 개발하고 플랫폼화하고 있다.

네이버는 페이스 북과 트위터가 유입되기 전까지는 국내 웹 환경에서 독보적인 위치를 차지하던 사이트였다. 하지만 스마트폰이 활성화되고 SNS가 차지하는 영향력이 커지면서 입지가 점점 줄어들게 되었다. 그래서 네이버는 미 투데이를 인수했다. 미 투데이도 트위터 창업 시기와 비슷한 시기에 만들어져서 운영을 해오던 SNS인데 이걸 NHN이 인수하게 되면서 사람들에게 많이 알려지기 시작했다. 하지만 네이버는 평소에 하던 대로 스타 마케팅을 이용했고 10대 이용자들을 많이 끌어들이게 되었다. 하지만 이 전략은 10대 이용자를 잡는 데는 성공했지만 SNS 자체를 너무 가볍게 만들어버리는 크나큰 실수였다. 네이버는 소셜 허브로의 도약을 시도해야 한다.

비교해서, 야후 홈페이지에 가면 다른 웹사이트들과 자유자제로 연동이 되게 구조를 만들어 놓았다. 네이버가 검색시장은 구글에, SNS시장은 페이스북과 트위터에 빼앗기고 있는 현실에서 빅데이타 관리에 좀 더 신경을 써야 한다.

인터넷 사용자들이 포털을 방문하는 가장 큰 이유는 검색과 함께 자신들의 개인화서비스를 쓰기 위해서다. 이 때문에 네이버는 지난해부터 '소셜과 개인화'에 초점을 맞춘 서비스를 준비해왔고, 지난해 12월 네이버 미 오픈베타 서비스를 시작했다. 네이버 소셜홈 '네이버 미(me)'는 개인화웹서비스(PWE, Personal Web Environment)와 소셜네트워크서비스(SNS)가 결합된 형태의 새로운 홈이다.

블로그나 카페, 미 투데이 등에 업데이트된 내 소식과 친구들의 새 글을 한눈에 확인할 수 있으며 뉴스, 스포츠, 영화 같은 다양한 콘텐츠를 구독하고 공유할 수 있는 것이 큰 특징이다. 즉, 네이버미의 최종 목표는 '페이스북+구글'이라고 봐도 무방할 것이다. 또한 네이버 미에서 커뮤니케이터의 역할을 할 네이버 톡도 관심을 끌고 있다. 네이버 톡'은 웹과 모바일 환경의 제약을 뛰어넘어 누구나 자유롭게 대화하고 정보를 공유할 수 있는 새로운 형태의 메시징 서비스이다. 네이버 로그인 만 으로도 추가 기능 설치 없이 네이버 톡을 이용할 수도 있고 데스크톱 애플리케이션을 통해 웹페이지 접속 없이 이용할 수도 있다. 네이버 미가 개인화서비스였다면, 네이버 톡은 지인들과의 커뮤니케이션을 돕는 보완제의 역할을 하게 될 것으로 보인다.

네이버는 모바일과 스마트TV등 새로운 디바이스 환경에 맞춘 검색 플랫폼 확대 및 서비스 고도화 전략을 펼칠 예정이다.

다음은 PC 시대에서는 네이버에게 항상 밀리는 2인자였다. 하지만 모바일 세계가 열리고 나서는 상황이 역전되고 있다. 다음 view를 통해 사람들과 소통하고, 적극적인 모바일 웹과 검색 투자, 그리고 다음 앱을 통해 스마트폰 사용자들에게 접근하고 있다. 소셜 검색을 통해 네이버에게 밀렸던 시장을 확보하려고 노력하고 있다. 다음은 지난해에 이어 올해에도 '라이프 온 다음(Life On Daum)'이란 슬로건을 중심으로 사용자가 생활 속에서 다음의 서비스를 이용할 수 있는 플랫폼을 만들 계획이다. '라이프 온 다음'이라는 슬로건을 효과적으로 구현하기 위해 PC웹, 모바일, 디지털사이니지, IPTV를 연결하는 멀티스크린 전략을 펼치고 있다.

특히 다음은 빠르게 확산되고 있는 다양한 디바이스를 망라하는 이용자 접점 확대에 주력하는 동시에 '라이브'와 '개인화' 패러다임을 적극 수용함으로써 앞으로도 오픈소셜플랫폼으로 진화하기 위한 다양한 서비스들을 선보일 계획이다. 또한 최근 트위터와의 제휴를 통해 '트위터로 소셜을 밀고, 요즘(yozm)으로 내부 서비스를 강화'하는 전략도 추진될 예정이다.

다음은 지난해 '라이브 온 다음(Live on Daum)'을 모토로 개방과 연결을 전제로 한 오픈소셜플랫폼으로 진화하기 위해 실시간 검색과 소셜웹 검색, My소셜 검색 등을 선보이는 동시에 소셜네트워크서비스(SNS) '요즘'을 오픈하며 이용자를 중심으로 한 차별화된 소셜 서비스들을 제공하고 있다. 다음은 올해에도 SNS에서 생산되는 콘텐츠를 다음에서 공유할 수 있도록 하는 한편 다음의 다양한 콘텐츠들도 외부 SNS 등과의 연결기능 및 범위를 더욱 확대해 진정한 개방과 공유를 실현함으로써 이용자 편의성을 더욱 높여나간다는 계획이다.

다음과 트위터와의 제휴에 따른 소셜전략도 주목된다. 이번 제휴로 실시간 트위트 정보가 다음 첫 화면의 '라이브 스토리'에 노출되며 트위터와 다음의 사회관계망 서비스인 '요즘'을 연동해 한쪽에 글을 남기면 양쪽에 동시에 등록되는 기능 등의 연동 서비스를 제공할 예정이다.

트위터와 다음 카페, 블로그 연동을 시작으로 소셜 플랫폼을 깔기 시작하고, 자사의 SNS인 요즘을 추가로 도입해 시너지를 낸다는 계획이다.

특히 모바일 서비스에 주력해 1000만 스마트폰 사용자를 눈앞에 둔 국내 모바일 시장을 공략한다는 계획이다. 최근 다음은 모바일 메신저인 '마이피플'에 무료통화(m-VoIP)기능을 추가로 탑재하면서 카카오톡의 아성에 도전하고 있다. '마이피플'은 자주 연락하는 지인들과 편리하게 대화하고 연락할 수 있는 유무선 인스턴트 메시지 서비스로, 아이폰, 안드로이드폰을 비롯해 모바일웹, PC웹 등 디바이스에 구애 받지 않고 지인들과 무료로 문자 메시지와 음성 쪽지 등을 이용할 수 있는 서비스이다. 이번 무료통화 기능 탑재로 마이피플 사용자가 카카오톡을 넘어선다면 다음은 가장 큰 모바일 플랫폼을 가지게 된다.

다음은 모바일 광고가 새로운 광고 채널로 급부상하는 가운데 국내 최초로 모바일웹 배너 광고를 선보이는데 이어 키워드 광고 상품인 '프리미엄링크' 광고를 모바일웹을 통해 동시에 노출해 광고 효과를 극대화하는 '모바일 키워드' 광고를 제공하고 있다. 이와 함께 모바일 어플리케이션 탑재형 광고인 '인앱애드(in-app ad)'와 모바일웹을 포괄하는 모바일 광고 플랫폼 'AD@m'을 오픈해 애플리케이션 개발자, 모바일 사이트 운영자 등 다양한 플랫폼의 운영자들이 손쉽게 등록해 광고를 노출하고 수익을 낼 수 있도록 했다.

키워드를 입력하기 어려운 모바일의 단점을 극복해줄 음성검색, 코드검색, 사물검색, 초성검색 등과 더불어 모바일에서 유용한 실시간 검색, 장소검색, 음악검색 등 다양한 검색 기능을 선보이며 이용자들의 편의성 및 만족도 향상에 기여했다.

다음은 개인 클라우드 컴퓨팅 시장에도 진입할 예정이다. 가상화 기술과 분산 컴퓨팅기술을 바탕으로 서버의 자원을 효율적으로 이용하는 동시에 개인PC나 스마트폰 같은 단말기에 구애 받지 않고, 언제나 일치된 이용자 환경을 구현할 수 있는 장점을 살려 개인 저장 공간 서비스와 같은 관련 서비스를 제공할 예정이다.

싸이월드는 2000년대 중반부터 대학생들에게 유명세를 타면서 우리나라에서 유저 층을 크게 벌려나갔다. 흔히 싸이월드의 실패 요인은 네이버보다 더 심각한 폐쇄성에 있다. 그동안 그 어떤 서비스의 API도 개방하지 않았다. 이로 인해서 다른 서비스들은 싸이월드의 데이터를 활용할 수 없어,

드파티 개발자들의 수도 적어질 수밖에 없었다.

지난 2009년부터 페이스 북, 트위터 사용자의 급증으로 인해 한때 '싸이월드'의 위기설이 닥치면서 SK컴즈에서는 '넥스트 싸이월드'를 구상하기 시작했다. 당시 싸이월드와 페이스 북·트위터의 가장 큰 차이점은 '타임라인'의 부재였다.

싸이월드는 개인화의 성격이 강해 지인들의 정보를 습득하는데 있어 페이스 북·트위터보다 부족한 모습을 보였던 것이 사실이다. 이에 SK컴즈는 싸이월드의 사용자를 그대로 흡수하면서 해외서비스의 장점인 타임라인을 도입한 'C로그'를 지난해 출시하고 올해 본격적으로 서비스할 것이라고 밝혔다. C로그는 단순히 '타임라인을 적용한 싸이월드'를 넘어 구독, 스크랩 등의 기능도 탑재돼 있다. 최근의 SNS 사용 트렌드를 반영해 SK컴즈 내부 서비스는 물론이고 외부사이트의 콘텐츠를 즉시 구독할 수 있다.

SK컴즈가 구상하고 있는 브랜드 C로그는 소셜커머스 시스템과 C로그를 결합한 서비스로 기획되고 있다. 이는 지난 17일 구글이 선보인 소셜검색의 강화판과도 맥락이 같다. 구글은 검색 시 지인들의 활동과 검색어 매치를 통한 결과물을 최상단에 배치하고 있다. 검색서비스의 질은 사용자가 원하는 검색결과를 노출시켜주는 것에 달려있다.

네이버와 다음은 오랫동안 지식인, 카페와 같은 서비스를 통해 DB를 구축해뒀기 때문에 덩달아 검색결과물이 많다. 이에 SK컴즈는 네이트 검색결과에 일촌 공개 게시물까지 보여줄 계획이다. 현재 싸이월드 미니홈피 게시물의 90% 이상이 일촌 공개이기 때문에 이 게시물들을 검색결과에 반영하면 네이버 지식인, 다음 카페에 못지않은 검색 DB를 확보할 수 있다는 생각이다.

요즘의 네이트는 베플 이라는 사용자 참여를 이끌어내는 시스템을 적용해서 3개 포털 뉴스 중 가장 많은 유저, 특히 젊은 층의 참여를 크게 이끌어냈다. 그와 더불어 네이트온의 강화, 네이트 앱스토어를 통해 새로운 포털로 도약하는 중이다. 거기다가 네이트 판을 통해 블로그와 SNS영역에서 가장 발전하고 있다.

야후는 '소셜서비스'에서 돌파구를 찾고 있다. '소셜펄스'라는 서비스를 통해 페이스 북과 트위터 등 소셜네트워크서비스(SNS)를 한곳에서 관리할 수 있도록 한 것이다. 즉 데이터의 관리만 하고 API로 SNS와 연결만 한 것이다. 야후가 제일 잘하였던 그야말로 포탈의 기능에서 자기의 위치를 찾은 것이다.

III. 본론

국내에서의 대용량데이터분석시장은 SNS 포탈들이 가장 필요성을 느끼고 있다. 데이터를 분석하기위해서는 세단계의 과정이 필요하다. 먼저 데이터의 수집이다. 이를 위해 포탈들은 API를 이용하여 자료를 수집한다. 또한 Web2.0 : XML, RSS, Trackback 같은 기술도 이용하게 된다. 이후에는 유효한 데이터를 추출하는 과정이 필요하다. 의사 결정 시스템 DSS(Decision Support Systems)이라 불리우며 이에는 CRM, CIODS(Customer Information Operational Data Store), DT(Decision Tree), COM(Co-Occurrence Matrix), K-Means Clustering기법을 이용하게 된다. 마지막으로 데이터에 대한 분석기술이다. 이에 OLAP (On-Line Analytical Processing)이 멀티미디어 데이터에 주로 쓰였으나 Ontology 기법이 주 분석 방법으로 각광 받고 있다.

가장 주목을 받고 있고 다양한 시도가 이루어지고 있는 것은 바로 실시간 대용량 데이터 분석 기술이다. 이는 소셜네트워크데이타의 특징이다. 이 데이터에 대한 분석은 세단계로 나누어진다.

첫 번째, 소셜네트워크의 위상학적 구조(Network Topology Structure)를 분석을 해서 네트워크의 전반적 특성을 파악하게 된다.

두 번째, 네트워크 구조의 시간에 따른 진화를 분석한다. 세 번째, 소셜네트워크 상에서 각 노드(사용자)가 생산, 확산시키는 콘텐츠(포스트, 댓글, 리트위, 동영상, 링크 등) 흐름을 분석한다.

이를 종합하여, 각 개인 또는 그룹의 소셜네트워크 내 영향력(Influence), 관심사, 성향 및 행동 패턴을 분석 추출하게 된다.

먼저 네트워크 구조 분석이다. SNS에서 사용자는 동일한 관심사를 갖는 다른 사람과 관계 맺기(친구 맺기)를 통해 인맥을 형성한다. 즉, 노드가 다른 노드와 링크를 형성하면서 소셜네트워크가 형성 된다. 이렇게 관계 맺기를 통해 형성된 소셜네트워크의 구조를 분석할 때, 글로벌 구조 분석, 커뮤니티 분석, 노드 역할 분석으로 세분된다.

글로벌 구조 분석은 네트워크에서 정의되는 각종 지표(네트워크 직경, 친구끼리 친밀도, 친구 수 분포 등)를 분석하여 네트워크의 전반적인 모습을 파악한다. 커뮤니티 분석은 네트워크 내 조밀한 연결 구역을 명명하는 작업이라고 할 수 있다.

이 커뮤니티 내에 존재하는 노드들은 외부보다 이들 커뮤니티 멤버들끼리 조밀한 연결 관계를 맺고 있다. 이러한 관계를 맺게 된 원인은 그들이 갖고 있는 콘텐츠 또는 프로필의 유사성에 기인하기 때문에 타깃 마케팅을 할 때, 매우 중요한 데이터로 사용될 수 있다. 마지막으로, 커뮤니티 내 각 노드들의 역할 분석을 하게 된다. 이를 통하여 특정 그룹 내 각 노드들의 특성이 파악되어 최종 타깃팅을 위한 판단 근거를 제공할 수 있게 된다.

두 번째, 네트워크 진화 분석, SNS에서는 시시각각 유저가 가입 및 탈퇴, 그리고 새로운 관계 맺기 형성 및 끊기가 빈번히 이뤄진다. 이러한 개별적인 행동들이 모여서 다음과 같은 현상이 발생한다. 즉, 서로 다른 두 커뮤니티가 합쳐지거나 특정 커뮤니티가 여러 개로 갈라지는 현상이 발생한다. 물론, 지속적으로 유지되면서 팽창하는 것이 있는 반면 점점 그 세가 작아지고 결국은 사라져가는 커뮤니티도 있을 수 있다.

이러한 현상의 원인으로 기존 상품 또는 정보에 대한 관심사가 같은 별개 커뮤니티들의 이합집산이 있을 수 있다.

또 새로운 토픽의 출현, 팽창, 소멸 현상도 커뮤니티 생성 소멸의 원인이 될 수 있다. 따라서 거시적 관점에서 새로운 토픽이나 트랜드 변화를 추적하기 위해선 네트워크 구조에 대한 진화 분석은 필수적이다.

세 번째, 네트워크 정보흐름 분석, SNS 유저는 자신의 일상생활 및 관심사에 대한 콘텐츠를 SNS에 포스팅하여 자신의 친구와 공유(sharing)한다. 또한 친구 포스트를 또 다른 친구들과 공유하기도 한다. 이러한 개인들의 움직임이 소셜네트워크에 의미 있는 거시적 정보흐름 현상으로 나타나게 된다. 이들 정보 흐름들을 관심 토픽별, 또는 키워드별로 추출하여 네트워크를 재구성하게 되면 특정 정보 흐름 네트워크를 구축할 수 있게 된다.

이렇게 형성된 정보흐름 네트워크에 대해 커뮤니티 분석을 하고 소속 커뮤니티 내 노드들의 역할을 분석하게 되면 커뮤니티별로 영향력이 높은 오피니언 리더(Opinion Leader)를 추출할 수 있다. 이렇게 선택된 오피니언 리더들과 기업들과의 교류는 기업이나 상품 이미지에 지대한 영향을 끼치게 될 것이다. 그런데 이들과 교류 시 주의할 점이 있다. 기존의 홍보나 광고 매체를 통한 일방적인 정보 캐스팅 방식은 오히려 반감을 불러일으키고 말 것이다. 신뢰감 확보를 통해, 오피니언 리더들이 스스로 맘에 드는 기업들에 대한 홍보 및 마케팅

을 해줄 것이기 때문이다. 그리고 이들 뒤에 있는 많은 사람들은 자신들이 신뢰하고 있는 오피니언 리더들이 제공하는 정보를 믿음을 갖고 받아들이게 될 것이다.

이러한 실시간 데이터 분석을 위해서 주목 받는 기술 중에 하는 Complex Event Processing (CEP) 라고 하는 기술이다. 다시 말하면 실시간으로 발생하는 복수의 이벤트로부터 특정 패턴을 찾아내서 원하는 데이터 처리나 알림 서비스가 가능하게 하는 기술이라고 할 수 있다.

TIBCO, Oracle, IBM과 같은 솔루션업체들은 이미 CEP 솔루션을 제공하고 있고 이밖에도 EsperTech 라는 회사는 Esper 라고 하는 자바와 닷넷에서 사용할 수 있는 CEP 엔진을 오픈소스로 공개하고 있다. 그림 2는 CEP관련 이벤트의 흐름도이다.

IBM은 기존 텍스트나 정형화된 이벤트 스트림 뿐아니라 실시간으로 센서로 부터 쏟아져 들어오는 대용량 데이터 스트림에서부터 이미지, 동영상, 음향 데이터 등에도 적용이 가능한, InfoSphere Stream 이라는 스트림 프로세싱 엔진을 상용화해서 내놓고 있다.

페이스북의 경우에는 하둡과 HBase 을 기반으로 페이스북의 실시간 메신저 서비스를 구현 하였다.

마지막으로 대용량 데이터 분석 분야에서 주목해야 할 부분은 대용량 데이터 비주얼라이제이션 부분이다. 대용량 데이터의 형태, 용량, 내용등에 대한 표현을 그래픽으로 나타내는 것에 대한 연구이다.

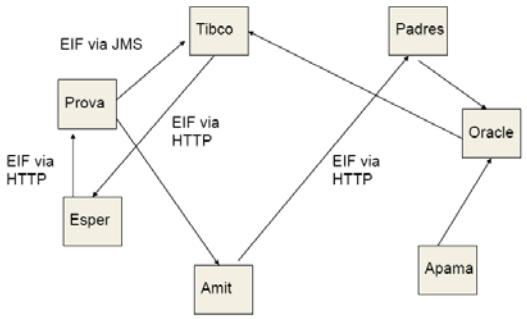

그림 4. CEP 관련 흐름도

IV. 결론

위에서 여러 각도로 기존과 향후 연구 동향들을 분석하였다.

SNS의 확산에 따라 대용량 데이터의 분석 기술과 이의 활용 방법에 대한 연구가 활발히 이루어지고 있다. 데이터의 저장 데이터의 추출 데이의 분석 세단계가 SNS와 연계되어 보여지는 특성들이 앞으로의 향후 연구 과제들이 될 것이다. 하둡과 같은 저장 시스템 R과 같은 분석시스템 그리고 소셜네트워크의 진화에 의한 새로운 포탈의 형성등 이분야의 연구 과제는 무궁무진하다.

앞으로 데이터의 비쥬얼라이제이션에 대한 연구가 향후 실행 할 계획이다.

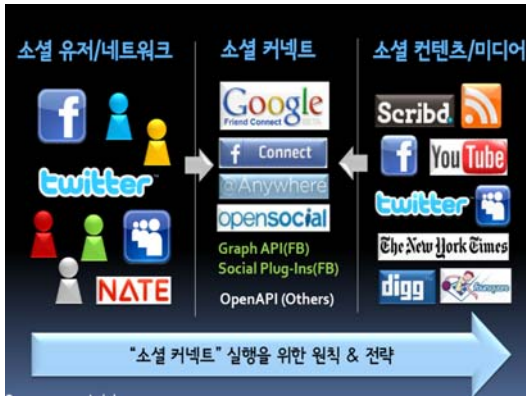

그림 5. 소셜네트워크 융합전략

참고문헌

[1] Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., and Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization.. Psychological Science,

[2] Bonneau, Joseph, Anderson, Jonathan, Anderson, Ross, and Stajano, Frank. (2009). Eight Friends are Enough: Social Graph Approximation via Public Listings. In proceedings of the Second ACM Workshop on Social Network Systems.

[3] 데이터분석기술
http://kims.wordpress.com
[4] 메소스 프로포절
http://wiki.apache.org/incubator/MesosProposal
[5] NextR
http://blog.daum.net/commman/675579
[6] 정교한 SNS마케팅
http://www.fnnews.com/view?ra=Sent0901m_View&corp=fnnews&arcid=0922180465&cDateYear=2011&cDateMonth=01&cDateDay=02
[7] 트위터 인포메이션 네트워크
http://www.zdnet.co.kr/column/column_view.asp?artice_id=20111205101936
[8] CEP Model
http://www.slideshare.net/isvana/ruleml2011-cep-standards-reference-model/download
[9] 국내 SNS
http://itopen.tistory.com/64
[10] 네이버 전략
http://news.nate.com/view/20110223n11394
[11] SK컴즈전략
http://ddaily.co.kr/news/news_view.php?uid=74669
[12] 다음 전략
http://news.nate.com/view/20110224n09592
[13] 소셜네트워크 분석기술
http://www.ciobiz.co.kr/news/articleView.html?idxno=5181

저자소개

김 시 우
1990: Univ. of Michigan
전자계산학과 학사.
1991: Univ. of Michigan
산업공학과 공학석사.
2007: KAIST
정보및 통신 공학박사
현 재: 숭의여자대학
인터넷정보과 교수
관심분야: 모바일S/W,
온톨로지