빅 데이터의 분석을 통한 정보 자동 요약 시스템

윤다영°*, 이현화*, 송재오**, 이상문*
*한국교통대학교 컴퓨터정보공학과
**(주)디엘커뮤니케이션즈
e-mail:{yunda0616°*, 123574hh*}@naver.com, jeo@web-d.co.kr**, smlee@ut.ac.kr*

Automatic Information Summary System using by Big Data Analysis

Da Young Yun°*, Hyun Hwa Lee*, Jeo Song**, Sang Moon Lee*
*Dept. of Computer Sci. & Info. Engineering, Korea Nat'l Univ. of Transportation
**Research Center, DLCOMS Co.,Ltd.

● 요 약 ●

오늘날 인터넷상에서는 무수히 많은 디지털 데이터가 생성되고 있으며, 그 디지털 데이터는 기존의 소프트웨어로는 처리할 수 없을 정도로 그 양이 방대해지고 있다. 이러한 데이터들을 사용자의 검색의도에 따라 문장 분석, 키워드 추출, 요약문 생성 등의 방법을 통하여, 사용자에게 개인화된 정보를 제공하기 위한 빅 데이터의 분석을 이용한 정보 자동 요약 시스템을 제안한다.

키워드: 데이터 마이닝(Data Mining), 클러스터링(Clustering), 빅 데이터(Big Data)

I. 서 론

세계적인 시장조사기관인 IDC에 의하면 2011년도 전 세계 디지털 정보의 양은 약 1.8 제타바이트(zettabyte)였으며, 현재의 데이터 증가 추세를 반영하면 2020년에는 35 제타바이트가 넘을 것으로 전망되고 있다. 우리 사람들은 이 데이터를 생성시킬 뿐만 아니라 이 데이터를 개개인의 필요성에 따라 각종 검색엔진에서 검색을 통하여 정보로 활용을 한다. 하지만 한 번의 검색만으로도 수도 없이 많은 양의 정보가 검색이 되는데, 이것이 바로 빅 데이터라고 불리는 것으로, 이렇게 데이터의 양이 많고 실시간으로 갱신이 요구되는 처리에는 빅 데이터 처리 기법을 사용하는 것이 적절하다. 따라서 본 논문에서는 이 빅 데이터 속에서 사람들이 가장 많이 공유하고 선호하며, 가장 정답이 될 수 있는 데이터들을 정보로 자동 요약하여 보여주는 정보 자동 요약 시스템을 제안한다.

II. 관련 연구

데이터마이닝(Data Mining)[1]은 대용량 데이터베이스에 존재하는 데이터 간의 관계, 패턴, 규칙 등을 찾아내고 모델화하여 의사결정을 돕는 유용한 정보로 변환하는 일련의 과정이다. 다음으로 클러스터링(Clustering)[2]은 비슷한 데이터끼리 모아 몇 개의 그룹으로 분류하는 것으로 추천 엔진에서는 사전에 취미나 기호가 비슷한 사용자를 클러스터링 해두고, 해당 그룹을 추천하는 효율적인 방법을 사용한다. 또한, 최근 이슈화되고 있는 빅 데이터(Big Data)[3]의 활용을 위해 주안점을 두어야 하는 부분에 대해 Doug[4]는 빅 데이터의 특징 3가지를 3V로 표현하였으며, 하둡 HDFS 등의 분산 파일 시스템 환경이 주목을 받기 시작했고, 단순한 분산 프로그래밍 프레임워크를 넘어 기존 RDBMS의 변형 형태인 NoSQL이 이슈화되기 시작했다.

III. 시스템 설계 및 구현

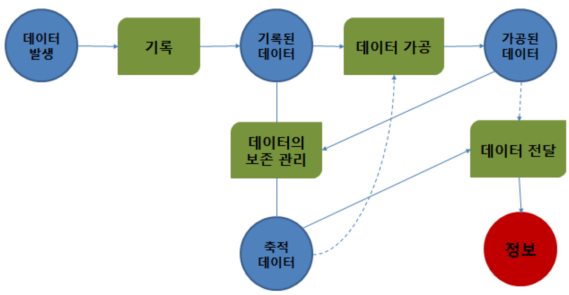

그림 1. 자료처리 과정
Fig. 1. Data Processing Flow

그림 1은 정보 자동 요약 시스템을 도식화 한 것이다. 제안하는 시스템은 우선, 발생된 데이터를 기록한 후에 기록된 데이터를 가공하고, 가공된 데이터를 전달하여 정보로 DB에 저장시킨다. 기록되거나 가공된 데이터는 데이터의 필요 여부에 따라 데이터를 보존하고 관리하며, 사용자의 요구에 따라 다시 데이터로 전달될 수 있다. 특히 과거의 분석 결과를 단순히 보여주고 끝내는 것이 아니라 지속적으로 관리, 저장하여 과거의 분석 결과와 계속 비교하여 데이터의 변화의 추이를 전체적으로 분석하도록 하였다.

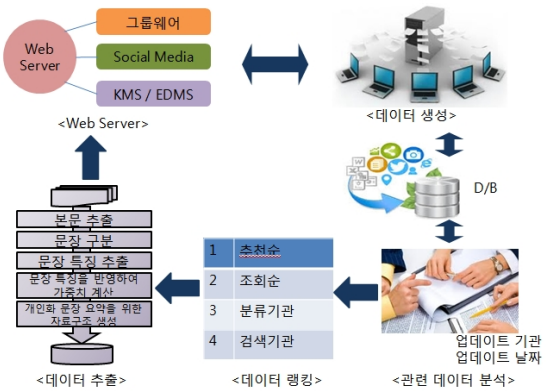

그림 2. 시스템 구성도
Fig. 2. System diagram

그림 2는 본 논문에서 구현한 시스템의 전체적인 구성을 보여주고 있다. 제안하는 시스템은 관련 데이터 분석의 업데이트 기관, 업데이트 날짜를 이용하여 믿을 수 있는 최근의 데이터를 추출한다. 아울러 형태소 분석기에서는 웹상의 문서를 읽어 올 경우 태그와 같은 불필요한 기호를 불용어 사전을 이용하여 제거한다. 이와 같은 형태소 분석 과정을 거쳐 추출된 데이터는 그 빈도를 파악하여 순위별로 점수를 부여한다. 이때 만약 "요약하면", "결론은", "다시 말하면" 등과 같은 상용구가 포함된 문장은 무조건 추출한다. 본 시스템은 크게 데이터 분리를 위한 문장 분석, 키워드 추출, 요약문 생성 기법에 접근하여 주어진 데이터 랭킹 정보에서 요약 문장을 데이터로 추출하는 데이터 추출 기능을 나누어 구현되었다.

Ⅳ. 결 론

본 논문에서는 빅 데이터를 분석하는 기술을 토대로 웹상의 수많은 데이터를 이용하여 각 자료 간의 관계, 패턴, 규칙 등을 찾아내어 좀 더 분석적인 정보를 요약해주는 정보 자동 요약 시스템을 제안하였다. 제안 시스템의 적용은 웹상에서 발생한 데이터들을 빅 데이터의 분석 기법을 활용하여 웹상의 사용자들에게 좀 더 확실한 정보를 제공할 수 있다. 그러나 현재 시스템은 그 정확도의 신뢰성에 대해 신뢰도가 높지 않으므로, 향후에는 예측 신뢰도를 높이기 위해 각종 자료의 축적 및 분석 알고리즘을 최적화할 필요가 있다.

본 논문은 중소기업청 산학연 첫 걸음 사업을 통해 작성된 논문입니다.

참고문헌

[1] Woohyun Hwang, Ja-Hee Kim, Wan-Sung Jang, Jung-Sik Hong, Deuk-Su Han, "Fault Pattern Analysis and Restoration Prediction Model Construction of Pole Transformer Using Data Mining Techniques", 2008.

[2] Byoung - Hee Kim, "Agglomerative Clustering Methods based on Information Theory", 2003.

[3] J.M Seo, Jeo Song, C.R Lee, Sangmoon Lee, "Functional Cosmetics Trend Analysis System using Big Data From The Girls High School Of SNS", KSCI Proc. Vol. 21, No. 1, 2013.1.

[4] Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", META Group Inc., 2001.

[5] Dongwook Kim, at al., "A Document Summary System based on Personalized Web Search Systems", 2010.