점진적 학습 기술 기반 범용적인 분류기 구조설계 방법의 설계 및 구현

Design and Implementation of a Generic Classification System Based on Incremental Learning Technology

민 병 원, 오 용 선
목원대학교 정보통신융합공학부

Min Byung-Won, Oh Yong-Sun
Mokwon Univ.

요약

전통적인 마이닝 기법은 다양한 디지털 매체와 센서 등에서 생산되는 빅데이터를 처리하기 어려울 뿐 아니라 신규 데이터 누적 시 전체 데이터를 재분석 해야하는 비효율성과 대용량의 문서를 학습함에 있어 메모리부족 문제, 학습 소요시간 문제 등이 있다. 이러한 문제를 해결하기 위하여 본 논문에서는 자질축소 기법에 의존하지 않고 대량의 문서를 자유롭게 학습하고 부분적인 자질 추가 변경 시에 변경요소만을 추가 반영할 수 있는 범용적이고 일반적인 분류기의 구조설계 방법을 설계 및 구현하였다.
점진적 학습 모듈은 일반적인 학습 방법이 데이터의 추가 및 변동시마다 모든 데이터를 재학습하는 데 반해, 기존의 학습 결과에 증분된 데이터만 재처리 없이 추가적으로 학습한다. 재학습을 위해 사용자는 작업 수행 중 자원 관리를 통해 기존에 처리된 데이터를 자유롭게 가져와서 새로운 데이터와 병합이 가능하다. 이러한 점직적 학습 효율성은 빅데이터 기반 데이터 처리에 주요한 특성인 데이터 생산 속도를 극복하기 위한 좋은 대안이 될 수 있음을 확인하였다.

I. 서론

비즈니스 환경 변화에 따라 DBMS와 관련하여 실시간 데이터의 처리 및 정보 활용에 초점이 맞춰진 기술개발이 이루어지고 있으나 누적되는 스트리밍 데이터를 효율적으로 분류할 수 있는 기술이 부재하며[8], 임베디드 DBMS와 공간정보시스템(GIS)과 연계된 공간데이터 관리에 대한 시장의 관심도가 높아짐에 따라, 임베디드 DBMS는 다양한 디바이스와 연계되고 있으며 U시티 등 공공GIS의 확산에 따라 실시간 센서데이터 처리의 수요가 증가 하고 있다[1-4].

빅데이터 시대의 도래로 분석해야 할 데이터의 양적 증대와 더불어 시계열 분석을 위한 정형 및 비정형 데이터 분석의 요구가 증대 되고 있다[5-6]. 전통적인 마이닝 기법은 다양한 디지털 매체와 센서 등에서 생산되는 빅데이터를 처리하기 어려울 뿐 아니라 신규 데이터 누적 시 전체 데이터를 재분석 해야 하는 비효율성이 있다. 이와 같이 대용량의 데이터가 수시로 추가되는 환경에서의 데이터 마이닝을 위해, 데이터 전체를 반복 학습하지 않고 증분 데이터만을 학습하는 기술, 점진적인 학습 기술이 필수적이다[7-9].

본 논문에서 제안하는 점진적 학습 점진적 학습은 데이터가 동적으로 추가되는 환경에서 전체 데이터를 반복적으로 학습하지 않고 증분된 데이터만을 학습하여 기존의 학습 결과와 통합할 수 있는 학습 방법이다. 점진적 학습 모델에 기반한 빅데이터 학습기술은 대용량의 문서를 학습함에 있어 메모리부족 문제, 학습 소요시간 문제 등을 해결하기 위해 널리 사용하는 자질축소 기법에 의존하지 않고 대량의 문서를 자유롭게 학습하고 부분적인 자질추가 변경 시에 변경요소만을 추가 반영할 수 있는, 범용적이고 일반적인 분류기의 구조설계 방법에 관한 것이다.

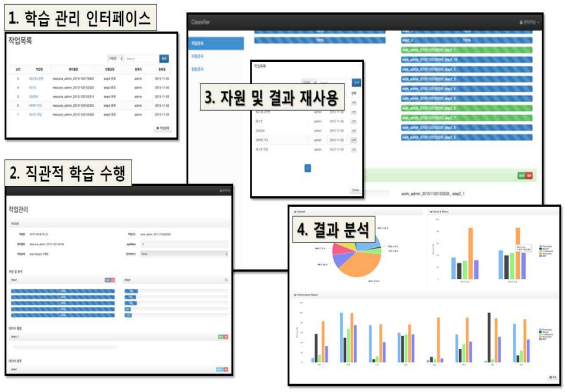
Ⅱ. 본론

자동 범주화 기술을 실제 서비스에 응용하고자 할 때, 경우에 따라서는 수백만 건 이상의 정보자원을 학습하고 해석해야 하는 경우가 있다. 일반적으로 효율적인 문서처리를 위해 자질 선정 기법을 사용하는데, 이는 정보량의 축소뿐만 아니라 성능의 향상을 위해서도 필요한 과정으로 알려져 있다. 그러나, 실시간 (또는 준실시간)으로 대용량의 데이터가 쏟아지는 빅데이터 환경에서 대용량의 문서학습을 하는 과업에서는, 자질 특성을 분석하고 이를 제거하는 과정에 소요되는 시간과 컴퓨팅 자원 문제에서 자유로울 수 없으며, 자질선택 및 축소기법의 적용 역시 한계가 존재하게 된다.

본 점진적 학습 기술에서 사용하는 기법은 작은 용량의 매트릭스를 다수 생성하여 정보 손실 없이 동적으로 결합하는 방법이다. 데이터베이스별로 여러 개의 작은 단위의 학습 결과(분류기)를 조합하는 예시이다. 여러 개의 분할된 복수의 분류기로 구성하여 동적으로 결합하여 최종 분류기를 생성할 수 있다.

본 논문에서 개발한 시스템은 웹 기반의 인터페이스를 통해 사용자 접근성을 높였고, 운영체제에 관계없이 윈도우, 리눅스 등 다양한 환경에서 구동이 가능하다. 또한 전문 지식이 없는 사용자도 쉽게 분류 작업을 가능하도록 직관적인 인터페이스를 구현하였다.

(그림 1) 학습 관리 인터페이스와 같이 웹 브라우저에서 점진적 학습 기능을 용이하게 수행하기 위해 사용자 권한에 따른 작업 및 자원 관리 기능을 개발하였다. 사용자는 자원 관리 기능을 통해 소유하고 있는 데이터를 손쉽게 데이터베이스로 적재할 수 있고, 업로드 된 자원을 기반으로 데이터 분류 작업을 수행하고 관리 할 수 있다.



▶▶ 그림 1. 웹기반 통합 인터페이스 개발 화면

Ⅲ. 결론

점진적 학습 모듈은 일반적인 학습 방법이 데이터의 추가 및 변동시마다 모든 데이터를 재학습하는 데 반해, 기존의 학습 결과에 증분된 데이터만 재처리 없이 추가적으로 학습한다. 재학습을 위해 사용자는 작업 수행 중 자원 관리를 통해 기존에 처리된 데이터를 자유롭게 가져와서 새로운 데이터와 병합이 가능하다. 이러한 점직적 학습 효율성은 빅데이터 기반 데이터 처리에 주요한 특성인 데이터 생산 속도를 극복하기 위한 좋은 대안이 될 수 있다.

폭증하는 빅데이터를 처리할 수 있는 상용 수준의 점진적 학습 모델로 비정형 데이터 뿐만 아니라 IoT 기술이 활용된 다양한 정형 센서 데이터의 분류가 가능해짐으로서 거의 모든 유형의 데이터 처리가 가능하여 산업 전반에 큰 파급효과가 클 것으로 기대되며, 점진적 학술기술을 통해 신속하고 정확한 빅데이터 분석을 수행할 수 있으며, 데이터 저장 효율을 증대시켜, 거대 매트릭스(Large-scale sparse matrix)를 PC급의 클러스터 환경에서도 계산 처리할 수 있어 초고성능 컴퓨팅 환경에서 더욱 큰 효율을 기대할 수 있다.

▌참 고 문 헌▐

[1] Gim, Jangwon, Hwang, Myung-Gwon, Song, Sa-Kwang, Kim, Jinhyung, Jeong Do-Heon, & Jung, Hanmin. Researcher history tracking service for prescriptive analytics based on researcher activities. Journal of KIISE: Computing Practices and Letters, 20(6), 359-363. 2014

[2] Jeong, Do-Heon. A study on automatic database selection technique using the maximal concept strength recognition method. Journal of the Korean Society for Information Management, 27(3), 265-281. https://doi.org/10.3743/kosim.2010.27.3.265, 2010

[3] Jeong, Do-Heon, Kim, Hwan-Min, Kim, Hye-Sun, & Shin, Ki-jeong. The relationship between the specificity of S&T terms and auto-classification accuracy. Proceedings of the 14th Conference of Korean Society for Information Management, 31-36, 2007

[4] Lee, Jae-Yun. Improving the performance of a fast text classifier with document-side feature selection. Journal of Information Management, 36(4), 51-69. https://doi.org/10.1633/jim.2005.36.4.051, 2005

[5] Lee, Jae-Yun. A novel clustering method for examining and analyzing the intellectual structure of a scholarly field. Journal of the Korean Society for Information Management,23(4), 215-231. https://doi.org/10.3743/kosim.2006.23.4.215, 2006

[6] Lee, Won-Goo, Shin, Sung-Ho, Kim, Kwang-Young, Jeong, Do-Heon, Yoon, Hwa-Mook, Sung, Won-Kyung, & Lee, Min-Ho. Semi-automatic management of classification scheme with interoperability. The Journal of the Korea Contents Association, 11(12), 466-474. https://doi.org/10.5392/jkca.2011.11.12.466, 2011

[7] R. Burke, "Hybrid Recommender Systems:Survey and Experiments," User Modeling and User-Adapted Interaction, Vol.12, No.4, pp.331-370,2002(11).

[8] Francesco Ricci, Recommender Systems Handbook, Springer, 2011.

[9] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.6, pp734-749,2005.