비정형 문서 데이터를 활용한 클러스터 기반 질의 시스템에 관한 연구

이은정, 최형선, 김양중*

한국공학대학교 소프트웨어융합공학과

{neperz, winter765p, zeroplus*}@tukorea.ac.kr

A Study on Cluster based Query System for Data Usability of Unstructured Document

Eunjeong Lee, Hyeongseon Choi, Yangjung Kim*

Tech University of Korea

요 약

AI와 빅데이터 기술의 발전에 힘입어 다양한 서비스들로 이용자의 편의를 제공하고 있는 상황이다. 어느 때 보다 더 비정형 데이터 처리는 이러한 서비스를 제공하기 위한 정보의 원천임으로, 광범위한 데이터 정보를 수집과 가공 그리고 정제를 통해서 정보의 신뢰도를 높여야 할 것이다. 비정형 문서의 텍스트와 이미지의 가공을 통해서 중요한 데이터의 가치를 높이고 활용도를 높이는 시스템은 데이터가 중심이 되는 서비스의 핵심이 될 것이다. 따라서 본 논문에서는 클러스터 기반의 비정형 문서 정보를 검색하기 위한 질의시스템을 요구사항 및 구조에 관해 연구한 결과를 제안하고자 한다.

Ⅰ. 서 론

AI와 빅데이터 기술이 서비스 융합의 중심에 자리매김하고 있는 상황에서 더욱 데이터 정보의 가치가 중요한 상황이다. 정형, 반정형 외에도 비정형 데이터의 중요도가 높아, 이에 따른 데이터 마이닝 및 가공처리의 시스템과 프레임워크도 개발되어 활용되고 있지만, 비정형 문서에서의 데이터 추출, 가공, 정제를 통해 융합 서비스의 데이터 정보의 폭을 넓힐 필요가 있다. 이에 비정형 문서 데이터 정보를 클러스터 단위로 보관 및 관리를 하면서 이를 하나의 거대 마스터 클러스터 관리를 통해 분산된 정보를 찾을 수 있도록 하는 클러스터 시스템이 요구되며, 이를 통해 분산 클러스터의 데이터 관리 방법과 클러스터 마스터에서의 비정형 데이터를 질의하는 전체구조가 정의되어야 할 것이다.

본 논문에서는 비정형 문서를 통한 방대한 데이터 사용성을 높이는 클러스터 기반의 질의 시스템의 전체 구조 및 기능부의 요구사항을 연구하며, 이에 분산 클러스터 데이터 가공처리 및 저장하는 기능부를 정의하며, 마스터 클러스터에서 이용자의 요청을 받아 분산된 슬레이브 클러스터의 비정형 문서 정보를 질의하는 전체적인 시스템 구조를 연구하고자 한다.

Ⅱ. 본 론

데이터 마이닝으로 시작되어 다양한 사회 이슈 및 트랜드를 분석하고자 하는 정보의 시대가 도래한 이후, 웹과 각종 사회관계망 서비스에서의 트랜드를 분석 혹은 이슈 및 재난과 같은 특정 키워드 중심의 정보를 빠르게 파악하는데 비정형 데이터가 유용하게 사용되면서, 비정형 데이터를 기반한 다양한 산업 파급효과가 증대됨에 더 많은 서비스들이 창출되고 있다. 텍스트, 이미지, 오디오 및 비디오에서 이러한 데이터 정보를 수집하고 정제해 시각화 함으로써, 정보력의 새로운 기술응용은 AI를 위한 정보의 활용의 극대화를 이끌어 내고 있다. 이와 더불어, 비정형 문서에서의 이러한 정형 데이터를 추출하려는 서비스들 또한 비정형 데이터 만큼의 중요한 자료를 정형 데이터로 데이터베이스를 구축해 정보의 선택폭을 넓히려는 노력이 기울여 왔다. 이러한 비정형 문서로부터 데이터 정보를 추출하고 이를 분산 슬레이브 클러스터들에 보관하고 이를 마스터 클러스터 관리부를 통해 이용자의 질의를 분석해 결과를 제공하는 시스템의 구조 및 내부 기능별 특성을 분석하고자 한다. 방대한 비정형 데이터를 저장하는데는 한계가 있고 중앙집중된 데이터 저장소를 별도로 통합하는데에 어려움이 있는 바, 제안되는 시스템 구조는 이를 해결하고 효율성을 높일 것으로 기대된다.

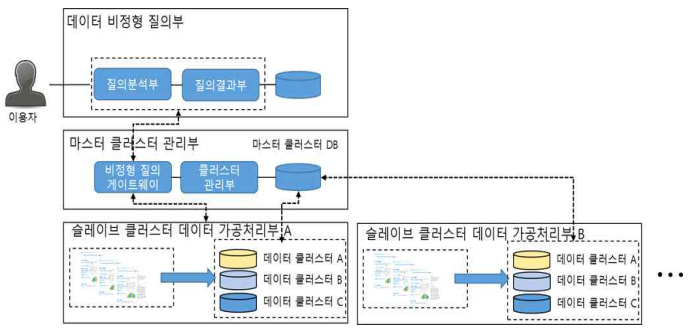

그림 1. 비정형 문서 데이터를 위한 클러스터 기반 질의 시스템 구조도

그림 1에서 볼 수 있듯이, 비정형 문서에서 데이터를 추출해 분산 슬레이브 클러스터에 저장하기 위해서는 세 개의 기능부가 구조적으로 정의된다.

- 데이터 비정형 질의부:
  질의분석부와 질의결과부의 내부기능을 갖고 있으며, 이용자의

질의를 먼저 분석하고 마스터 클러스터에 관리부에 전달해 원하는 질의결과를 가져오도록 함

- 마스터 클러스터 관리부:
  분산된 슬레이브 클러스터 데이터 가공처리부들의 정보를 갖고 있으며, 이용자의 질의에 따른 결과를 산출해 최적의 결과를 제공하도록 하는 기능부
- 슬레이브 클러스터 데이터 가공처리부:
  비정형 문서[1]로부터 추출된 텍스트 및 이미지 데이터의 전처리 정보[2]를 로컬 저장하고 있으며, 데이터 분류에 따른 빠른 처리를 위한 인덱스 기반의 데이터베이스를 갖으며 복수의 슬레이브 클러스터 데이터 가공처리부가 존재할 수 있으며 해당 정보는 마스터 클러스터 관리부에 등록되어 메타검색이 될 수 있도록 함

그림 2. 슬레이브 클러스터 데이터 가공처리부 기능구조도

그림 2에서 볼 수 있듯이, 분산화 되어 있는 슬레이브 클러스터는 비정형 문서에서 텍스트 데이터 및 이미지 데이터를 추출한다. 텍스트 데이터는 비정형 데이터의 시멘틱 추출까지 단어 또는 문장으로 수집하는 것과 동일하나, 문서구조의 특성을 고려한 기능을 수행한다. 마찬가지로 이미지 데이터 추출에 있어서, 이미지 픽셀 단위 추출을 통해 데이터화 될 수 있으며, 이런 데이터를 딥러닝을 통해 의미있는 특성을 추출하게 된다. 이리 추출된 문서 데이터는 데이터를 실제 클러스터 서버에 저장하게 되며 보안모듈을 통해 접근제어를 하게 된다. 마스터 클러스터에 시멘틱 비정형 질의는 바로 비정형 문서 데이터 질의처리부에서 받게 되어 시멘틱 질의를 분석하거나 관련 이미지 질의를 처리할 수 있도록 프로세스를 갖는다.

본 논문에서 제안된 마스터 클러스터에서 이용자의 요청을 받아 분산된 슬레이브 클러스터의 비정형 문서 정보를 질의하는 전체적인 시스템 구조를 통해서, 이용자의 비정형 시멘틱 질의를 효율적인 분산 데이터 저장소를 마스터 클러스터를 통해 메타 질의를 하는 시스템의 구조 및 각각 기능부의 요구사항을 제안하며 이를 통해 방대한 양의 비정형 문서로부터의 데이터를 효율적으로 관리하고 데이터 질의의 결과를 풍부하고 신뢰적으로 응답받도록 한다.

Ⅲ. 결론

본 논문에서는 방대한 비정형 문서로부터 데이터를 추출하고 클러스터별로 저장되어 데이터 저장소의 집중되는 부담을 줄여 분산 클러스터에 저장되어 있는 비정형 데이터 정보를 마스터 클러스터 관리부를 통해 시멘틱 질의를 처리하는 시스템의 구조 및 기능부 정의를 하였으며 이에 요구사항을 확인했다. 향후, 이러한 시스템의 구현 및 변화하는 지식정보 체계에 대응하도록 하는 추가적인 모듈을 정의해 운영 모범사례로 확장할 계획이다.

ACKNOWLEDGMENT

본 연구는 고용노동부 및 한국산업인력공단의 '2024년 고숙련 마이스터 사업'과 ㈜에스에이티정보(www.satu.co.kr)의 지원을 받음.

참 고 문 헌

[1]https://learn.microsoft.com/ko-kr/azure/ai-services/document-intelligence/concept-layout?view=doc-intel-4.0.0

[2]김주엽, 김민영, 정유철(2021) 문서 객체의 랜덤 배치를 통한 문서 레이아웃 분석기