**연구노트**

**(수학과)이 재 봉**  2017.05.17 (목)

|  |  |  |
| --- | --- | --- |
|  | | |
| **연 구**  **주 제** | **특허분석을 통한 미래기술 예측** | |
| **텍스트 마이닝 기술 조사** | |
| **구 분** | **연구내용** | **진행률** |
| **연 구**  **내 용** | ■ 텍스트 마이닝  텍스트 마이닝은 대용량 비정형 데이터로부터 패턴(즉, 유용한 정보와 지식)을 추출하는 반자동화된 과정이다. 데이터마이닝이 범주형, 순서형 또는 연속형 변수에 의해 정형화되어 레코드로 정리가 된 정형 데이터베이스에 저장된 데이터에서 타당하고 새로우며 잠재적으로 유용하고 궁극적으로는 이해할 수 있는 패턴들을 확인하는 과정이라면 텍스트마이닝은 목적과 사용하는 과정이 같다는 점에서 데이터마이닝과 동일하지만, 텍스트마이닝 과정에서 사용하는 입력물들은 PDF파일, 텍스트 요약, XML파일 등과 같은 비정형(또는 덜 정형화된) 데이터 파일들의 집합이다. 근본적으로 텍스트 마이닝은 텍스트 기반의 데이터 소스에 구조를 입히는 것으로부터 출발해서 데이터마이닝 기술과 도구들을 사용하여 이 정형화된 텍스트 기반 데이터로부터 적절한 정보와 지식을 추출하는 과정이라고 생각할 수 있다.  [기술분야]  1. 정보추출  텍스트마이닝은 미리 정해진 객체와 순차들을 패턴 매칭으로 찾아내어 텍스트 안의 중요 어구와 관계들을 확인한다. 아마도 가장 널리 사용되는 정보추출 형태는 객체추출일 것이다. 객체추출이란 이름이 있는(named) 객체의 이름 인식, 동일 지시어 찾기, 그리고 관계 추출 등을 뜻한다.  2. 주제 추적  사용자의 프로필과 사용자가 본 문서들을 바탕으로 그 사용자가 관심을 가질 만한 다른 문서들을 예측할 수 있다.  3. 요약  독자가 시간을 절약하도록 문서를 요약할 수 있다.  4. 범주화  문서의 주된 주제를 찾아내서 그 주제에 맞게 미리 정해진 범주로 문서를 이동한다.  5. 군집화  미리 정해진 범주가 없어도 유사한 문서들끼리 모아 줄 수 있다.  6. 개념의 연결  공통된 개념들을 찾아냄으로서 관련 문서들을 연결하고, 이로써 사용자로 하여금 기존의 탐색법으로는 찾을 수 없던 정보들을 찾게 도와줄 수 있다.  7. 질의응답  주어진 질의에 대해 지식 구동 패턴 매칭을 통해 가장 좋은 답변을 찾는다. |  |
| **차 주**  **계 획** | ■ 계획  - 텍스트 마이닝 자연언어 처리기술 | |
| **이 슈** | ■ **해당사항 없음** | |

|  |  |
| --- | --- |
|  |  |