텍스트 마이닝을 이용한 상황 정보 분석 및 예측 프로세스에 관한 연구

정세훈* · 강주희* · 김종찬** · 심춘보*

*순천대학교 멀티미디어공학과, **순천대학교 컴퓨터공학과

A novel on Context Information Analysis and Prediction Process using Text Mining

Se-hoon Jung* · Joo-hee Kang* · Jong-chan Kim** · Chun-bo Sim*

*Dept. of Multimedia Engineering, Sunchon National University

**Dept. of Computer Engineering, Sunchon National University

[E-mail : iam1710@hanmail.net, cbsim@sunchon.ac.kr

요 약

최근 IoT 및 인공지능 기술을 활용한 상황 정보 예측 서비스가 각광을 받고 있다. 본 논문에서는 특정 메타 데이터(Meta Data)로부터 입력되는 정보를 기반으로 상황 정보 분석 및 예측하는 프로세스를 제안한다. 주성분 분석 및 데이터의 집단화(Corpus), 문서 매트릭스(Document Matrix), 단어 빈도수(Frequency)에 따른 데이터 전처리 과정을 통해 상황정보 데이터를 확보한다. 또한 연관 규칙 분석을 통해 분류된 데이터의 연관성을 분석하여 예측 데이터의 연관성을 확보한다. 제안하는 상황 정보 분석 및 예측 모델은 R을 적용하여 설계한다.

키워드

Text Mining, Association Rules, R, Prediction Process

Ⅰ. 서 론

과거에는 데이터로부터 유용 가능한 정보를 얻기 위해서 프로그래머가 직접 수작업 분석하였지만 최근엔 데이터의 분류 및 분석 기법을 이용하여 데이터의 의미를 분석하는 연구[1]가 늘어나고 있다. 신경망(Neural Network) 알고리즘, 유전자(Genetic) 알고리즘, 군집분석(Cluster Analysis), 의사결정나무(Decision Trees), 서포트 벡터 기계(Support Vector Machines) 등을 기반으로 대용량 자료를 데이터 마이닝(Data Mining) 기법[2]으로 분석이 어느 정도 가능해 지고 있다. 빅데이터(BigData)를 분석할 수 있는 데이터 마이닝 기법은 메타 정보의 요약과 미래에 대한 예측을 목표로 자료에 포함된 관계, 패턴, 규칙 등을 탐색하고 이를 통계적으로 모형화함으로써 유용한 지식을 추출할 수 있다. 이러한 데이터 마이닝 기법 중 텍스트 마이닝 분석은 분석 요구사항에 맞는 추천 서비스와 분석 군집화 처리 비용 기술을 포함하고 있으며, 빅데이터 분석 기술과 함께 중요성이 높아지고 있다. 이에 본 논문에서는 SNS상의 데이터를 기반으로 상황 정보 데이터간의 연관성과 규칙을 텍스트 마이닝 기법을 적용한 데이터 분석 및 예측 프로세스 설계를 R[3]로 제안한다. 제안하는 상황 정보 예측 데이터 처리 프로세스는 다양한 예측 프로세스의 의사결정을 위하여 텍스트 마이닝기반의 예측 데이터 정확성 및 신뢰성을 향상시키는 것을 주목적으로 한다.

Ⅱ. 관련 연구

텍스트 마이닝[2]은 데이터 마이닝의 한 분야로서 자연어처리 기술과 문서처리 기술을 적용하여 문서요약, 특성추출 등의 연구에 활용되고 있다. R[3]은 S 언어에서 확장된 통계 및 분석학적 프로그램 언어로 데이터마이닝, 통계 등을 포함한 분석 프로그램을 수행한다.

Ⅲ. 상황 정보 분석 및 예측 프로세스

그림 1은 상황정보 분석 및 예측 프로세스를 위한 텍스트 마이닝기반의 분류 및 연관분석 처리도이며, 제안하는 프로세스는 메타 데이터의 분

류 및 분석을 위하여 표본값을 분석 클래스 범위를 규정한다. 분석 클래스 집단을 분석하기 위한 전처리 과정으로 비정형 데이터인 문서의 명사 필터링 과정을 적용하는 메타 데이터 샘플링 단계에 적용한다. 구분된 단어를 기준으로 빈도수를 측정하여 빈도수가 높은 단어간의 연관 규칙 설정 후 분류 데이터의 연관성을 분석한다.

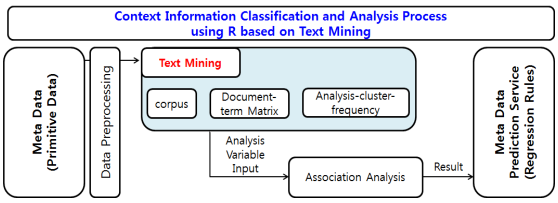

그림 1. 분석 및 예측 프로세스 구성도

3.1 분석 대상 및 분석 방법

본 연구에서는 요일에 따른 사용자의 상황 정보 분석 및 예측을 위하여 특정 SNS 사용자의 데이터를 활용하였다. 또한 분석을 위하여 R 프로그래밍 기반의 tm, arules, apriori 패키지를 활용하였다.

3.2 전처리 및 텍스트 마이닝 프로세스

상황 정보 예측을 위하여 분류 프로세스는 메타 데이터로부터 클래스를 추출한다. 데이터 예측을 위한 데이터의 전처리 과정은 데이터의 손실을 막고 통합적인 메타데이터 분류를 위해 임시 데이터베이스에 임시로 저장하게 된다. 그림 2는 메타 데이터의 변수 범위를 줄이기 위하여 전처리 과정에서 적용된 주성분 분석의 결과화면이다.

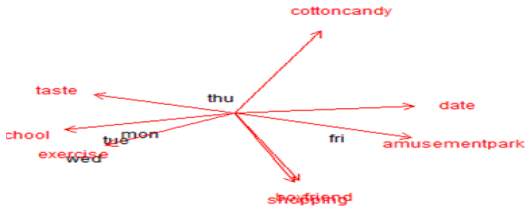

그림 2. 주성분 분석을 통한 변수 분석

그림 3은 메타 데이터의 분류를 위한 텍스트 마이닝 전처리 코드로 명사에 해당하는 단어를 필터링 처리하기 위한 코드이다.

```
data <-function(x){ nchar(x) <= 4 && nchar(x) >= 2} filter2 <- function(x){Filter(data1, x)} data2 <- sapply(data1, filter2)
```

그림 3. 텍스트 마이닝을 위한 전처리 코드

그림 4. 키워드 추출 결과(frequency>25)

그림 4는 텍스트 마이닝을 통해 분류된 추출 키워드의 Wordcloud 결과이다. Frequency는 25로 이하로 설정하였다.

3.3 연관 규칙 분석

키워드 추출을 통해 목요일날 생성되는 상황 데이터는 shooping, date, amusementpark, boy friend, exercise가 추출되었으며, 각 단어별 연관 규칙을 분석한 결과는 그림 5와 같다. 발생빈도와 동일 일자에 발생된 상황정보 단어의 연관성 분석도이다.

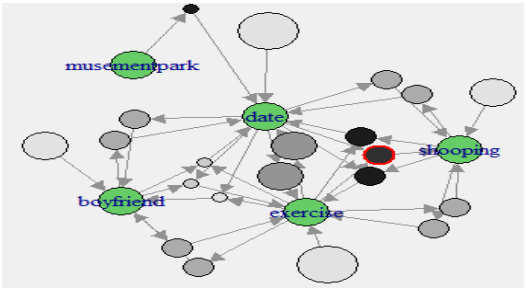

그림 4. 연관분석을 통한 상황 정보 예측

Ⅳ. 결 론

본 논문에서는 SNS상의 일상적으로 작성하는 데이터를 기반으로 사용자의 상황 정보 분석 및 예측 프로세스를 제안하였다. 데이터 분류과정은 데이터의 집단화와 문서간의 매트릭스, 단어의 빈도수를 통해 연관성이 높은 데이터셋을 확보하였다. 확보된 데이터셋을 기반으로 단어간의 연관성을 분석하여 특정 단어 작성 시 사용자의 상황정보 또는 예측 서비스를 제공할 수 있다.

감사의 글

본 논문은 한국정보통신산업진흥원에서 지원하는 2014~2015년 지역 SW 융합 지원 사업(S0413-15-1233)의 사업수행으로 인한 결과물임을 밝힙니다.

참고문헌

[1] S. H. Jung et al., “A novel on classification and Analysis process based on R for Reliability improvement of data prediction system”, ICFICE2015, vol.7, no.1, pp.355-358, 2015.

[2] K. A. Jang et al., "Project Failure Main Factors Analysis using Text Mining in Audit Evaluation," Journal of KIISE, vol.42, no.4, pp. 468-474, 2015.

[3] R, "http://www.r-project.org/“