JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 1, pp. 163-170, January 2021
https://doi.org/10.9708/jksci.2021.26.01.163

A Development Method of Framework for Collecting, Extracting, and Classifying Social Contents

Eun-Sook Cho*

*Professor, Dept. of Software Engineering, Seoil University, Seoul, Korea

[Abstract]

As a big data is being used in various industries, big data market is expanding from hardware to infrastructure software to service software. Especially it is expanding into a huge platform market that provides applications for holistic and intuitive visualizations such as big data meaning interpretation understandability, and analysis results. Demand for big data extraction and analysis using social media such as SNS is very active not only for companies but also for individuals. However despite such high demand for the collection and analysis of social media data for user trend analysis and marketing, there is a lack of research to address the difficulty of dynamic interlocking and the complexity of building and operating software platforms due to the heterogeneity of various social media service interfaces. In this paper, we propose a method for developing a framework to operate the process from collection to extraction and classification of social media data. The proposed framework solves the problem of heterogeneous social media data collection channels through adapter patterns, and improves the accuracy of social topic extraction and classification through semantic association-based extraction techniques and topic association-based classification techniques.

▸ **Key words**: Social Media Data's Collection and Analysis, Framework, Semantic Association-based Extraction, Topic Association-based Classification

[요 약]

빅데이터가 여러 분야에서 다양하게 접목됨에 따라 빅데이터 시장이 하드웨어로부터 시작해서 서비스 소프트웨어 부문으로 확장되고 있다. 특히 빅데이터 의미 파악 및 이해 능력, 분석 결과 등 총체적이고 직관적인 시각화를 위하여 애플리케이션을 제공하는 거대 플랫폼 시장으로 확대되고 있다. 그 중에서 SNS(Social Network Service) 등과 같은 소셜 미디어를 활용한 빅데이터 추출 및 분석에 대한 수요가 기업 뿐만 아니라 개인에 이르기까지 매우 활발히 진행되고 있다. 그러나 이처럼 사용자 트렌드 분석과 마케팅을 위한 소셜 미디어 데이터의 수집 및 분석에 대한 많은 수요에도 불구하고, 다양한 소셜 미디어 서비스 인터페이스의 이질성으로 인한 동적 연동의 어려움과 소프트웨어 플랫폼 구축 및 운영의 복잡성을 해결하기 위한 연구가 미흡한 상태이다. 따라서 본 논문에서는 소셜 미디어 데이터의 수집에서 추출 및 분류에 이르는 과정을 하나로 통합하여 운영할 수 있는 프레임워크를 개발하는 방법에 대해 제시한다. 제시된 프레임워크는 이질적인 소셜 미디어 데이터 수집 채널의 문제를 어댑터 패턴을 통해 해결하고, 의미 연관성 기반 추출 기법과 주제 연관성 기반 분류 기법을 통해 소셜 토픽 추출과 분류의 정확성을 높였다.

▸ **주제어**: 소셜 미디어 데이터 수집 및 분석, 프레임워크, 의미 연관성 기반 추출, 주제 연관성 기반 분류

• First Author: Eun-Sook Cho, Corresponding Author: Eun-Sook Cho
*Eun-Sook Cho (escho@seoil.ac.kr), Dept. of Software Engineering, Seoil University
• Received: 2020. 12. 23, Revised: 2021. 01. 14, Accepted: 2021. 01. 14.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업혁명 시대에 들어서면서 빅데이터가 여러 산업 분야에서 다양하게 활용되고 있는 가운데 특히 SNS(Social Network Service) 등과 같은 소셜 미디어를 활용한 빅데이터 분석이 매우 활발히 진행되고 있다.

빅데이터 분야 중 소셜 미디어는 최근 가장 활발한 연구 분야 중 하나로서, 기존 연구들이 빅데이터를 통한 고객 및 사용자들에 대한 분석에 중점을 두고 있다면, 최근의 연구들은 소셜 미디어에서의 게시글, 사진, 반응 등과 같은 비정형 데이터들을 이용하여 선거결과를 예측[1]하거나 마케팅 캠페인의 수행에 활용[2]하는 등 다양한 분야에서 활용되고 있다.

한국IDC의 '국내 빅데이터 및 분석 2019~2023 시장 전망' 연구 보고서에 따르면 2019년 국내 빅데이터시장 규모가 전년 대비 10.9% 증가해서 1조 7644억원을 기록했으며, 2023년까지 연평균 성장률 11.2%를 기록하며 2조 5692억원 규모에 달할 것으로 전망하고 있다[3]. 특히 빅데이터 시장은 빅데이터에 대한 의미를 파악하고, 분석 결과를 시각화로 제공하는 어플리케이션 개발 솔루션으로 확대되고 있다[4].

현재 많이 사용되는 소셜 데이터 분석 도구들은 지정한 단어를 필터링한 후 분석하고 시각화 하는 네트워크 통계 분석 및 시각화 기능들을 제시한다[5,6]. 기업들은 기업의 마케팅 전략 성과를 측정하기 위한 방법으로 네트워크 통계 분석과 같은 도구들을 많이 사용하는데, 이런 경우는 분석 비용을 많이 필요로 한다. 특히 여러 사회 다양한 분야의 데이터들을 수집, 분석하여 향후 마케팅까지 연계하기에는 높은 사양의 컴퓨팅 자원 및 네트워크 자원을 필요로 할 뿐만 아니라 복잡한 플랫폼을 사용하기 때문에 구축비용이 많이 소요 될 뿐만 아니라 운영 노하우의 부족으로 소셜 데이터를 활용하는 데 많은 어려움이 발생한다. 따라서 본 논문에서는 이러한 문제를 해결하기 위하여 소규모 형태 플랫폼에서도 소셜 미디어 데이터를 수집, 분석, 분류할 수 있는 프레임워크를 개발하기 위한 방법을 제시한다.

본 논문의 2장에서는 관련연구로 토픽 모델링, 빅데이터 분석 관련 연구에 대해 설명하며, 3장에서는 본 논문에서 제안하는 소셜 컨텐츠 수집, 토픽 추출 및 분류·피드백을 위한 프레임워크 개발 과정과 주요 기법들을 설명한다. 4장에서는 3장에서 제안한 기법을 실제 실험한 결과와 프레임워크의 성능 평가 결과를 제시하고, 마지막으로 5장에서 결론과 향후 연구 과제를 제시한다.

II. Preliminaries

1. Related works

1.1 Topic Modeling

토픽 모델링은 텍스트 마이닝 분야에서 유용하게 사용되는 도구이다[7]. 토픽 모델링의 주된 목적은 방대한 다큐먼트 집합으로부터 유의미한 패턴을 찾아내는데 있다. 워드들로 구성된 이 벡터 형태의 패턴을 토픽이라고 한다. 즉 토픽은 문서들에 포함된 단어들의 확률분포로서, 연관관계가 높은 단어들로 이루어진다. 토픽 모델링은 방대한 문서들로부터 이러한 토픽들의 통계적인 집합을 추출하는 방법이다. 토픽 모델링에서 가장 많이 알려진 기법이 LDA(Latent Dirichlet Allocation)[8,9,10]으로서, 각 문서들은 토픽들의 특정 집합으로 여겨진다. 이 때 관측할 수 있는 유일한 변수들은 문서 속 특정 단어들이다. LDA는 토픽 분포와 토픽을 구성하는 단어들의 분포에 잠재된 변수들을 주출하기 위해 디리클레 다항분포에 기반 한 추론방법을 사용한다[4]. 그러나 LDA는 시간에 따라 변화하는 토픽 트렌드를 제공하지 못하고 있는 한계점을 지니고 있다.

1.2 Big Data Analysis Techniques

과거 빅데이터 분석 기법은 대부분 정형화 데이터를 기반으로 한 데이터 마이닝, OLAP, 통계 분석 등을 사용해 왔다[11]. 하지만 최근에는 빅 데이터를 이용한 새로운 분석 기법들이 등장하고 있는데, 특히, 비정형 데이터를 분석하는 방법으로 텍스트 분석을 시작으로 평판·감성 분석, 노출 추이 분석, 네트워크 분석 등과 같은 기법들이 사용되고 있다. 먼저 텍스트 분석이란 대표적 비정형 데이터인 텍스트 문서에 대한 분석이며, 대용량의 데이터에서 사용자가 관심을 갖는 정보를 키워드 수준이 아니라 문맥(context) 수준의 의미를 찾아내는 프로세스이다[12]. 이런 기법에서는 형태소 분석, 구문 분석, 개체명 인식, 이벤트 추출, 관계 분석 등과 같은 자연어 처리 기술이 핵심이다[13]. 텍스트 분석을 수행할 경우, 텍스트 문서로부터 주제 및 이슈를 추출하고 이들의 연관 관계를 분석하거나 시계열 분석이 가능하다. 이런 경우, 문서 내 키워드의 자동 분류와 군집 및 요약 등이 수행되며 자연어 처리 뿐만 아니라 기계학습과 딥 러닝(Deep Learning) 기술도 적용된다[14,15]. 그러나 이처럼 사용자 트렌드 분석을 통한 마케팅 전략을 수립하기 위한 소셜 미디어 데이터의 수집 및 분석에 대한 수요가 지속적으로 급증함에에도 불구하고, 다양한 소셜 미디어 채널들의 인터페이스 이질성으로 인한 동적 연동의 어려움과 복잡한 소프트웨어 플랫폼 구축 및 확장성에 대한 어려움을 해결하기 위한 연구가 미흡한 상태이다.

2. Process of Social Topic's Extraction and Classification

딥러닝 기술을 이용한 소셜 토픽 추출 및 분류 절차는 [Fig. 1]과 같은 순서로 진행되며, 이 때 미디어 컨텍스트는 사용자의 관심 주제에 대한 '관심 컨텍스트'와 관심 시점에 따른 변화에 대한 '시점 컨텍스트'로 구성된다.

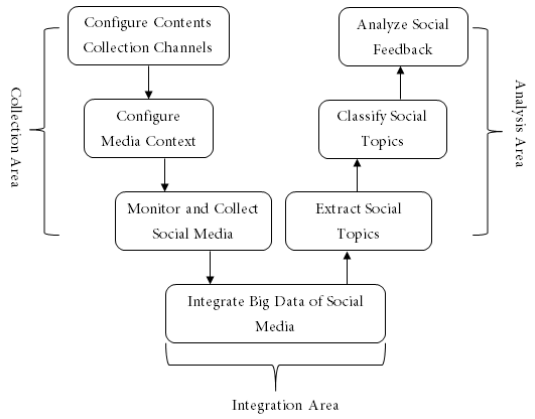

Fig. 1. Process of Extracting and Classifying Social Topic

다양한 소셜 미디어 수집 채널로부터 소셜 토픽을 추출하는 절차는 수집영역, 통합영역, 분석영역으로 구성되며, 각 영역에서 해결해야 하는 기술적 문제요소는 다음과 같다.

• 수집영역

첫째, 소셜 미디어 채널들이 다양하기 때문에 이러한 이기종 채널 관리의 어려움을 해결해야 한다. 둘째, 특정 관심주제 혹은 관심 시점을 기반으로 수집할 미디어 컨텍스트 설정의 어려움이 있다. 셋째, 다양한 미디어 컨텍스트 수집기 간의 이질성 문제를 해결해야 한다.

• 통합영역

첫째, 다양한 소셜 미디어 채널들로 인한 수집기 연동의 동적 연결을 위한 어댑터를 설계해야 한다. 둘째, 소셜 미디어 컨텐츠 통합 관리를 위한 통합 빅데이터 저장소를 설계해야 한다. 셋째, 비정형 소셜 미디어 컨텐츠의 변환, 전송, 수신, 해석을 위한 범용 컴포넌트를 설계해야 한다.

• 분석영역

첫째, 단어 연관성의 지속적 학습을 위한 딥러닝 기법을 설계해야 한다. 둘째, 소셜 미디어 빅데이터의 형태적 특성정보 기반 분석 및 소셜 토픽 추출 기법을 설계해야 한다. 셋째, 추출된 소셜 토픽을 주제별 또는 시점별 분류하는 기법을 설계해야 한다. 넷째, 연관된 다중 토픽 분석을 통한 소셜 피드백을 추출 기법을 설계해야 한다.

본 논문에서는 소셜 컨텍스트 수집, 추출, 분류에 이르는 과정들에서 발생한 이러한 기술적 이슈들을 해결하기 위해 다양한 채널로부터 동적으로 소셜 컨텐츠를 수집하여 의미 연관성을 기반으로 토픽들을 추출하고 주제 연관성을 기반으로 분류하는 프레임워크 개발 기법을 제시한다. 프레임워크로 개발하는 이유는 향후에 새로운 채널들이 발생하더라도 연동에 어려움이 없고, 또 필요한 컴포넌트의 추가나 기존 컴포넌트의 변경이 용이하다는 장점을 갖고 있기 때문이다.

III. The Proposed Scheme

이 장에서는 본 논문에서 제안하는 소셜 컨텐츠 수집에서부터 토픽 추출 및 분류에 이르는 기능들을 독립적인 컴포넌트들 구성한 프레임워크의 아키텍처와 각각의 컴포넌트들에 대해 설명한다.

1. Architecture Design of Framework

본 논문에서 개발할 프레임워크는 [11] 연구를 통해 설계된 메타모델을 기반으로 하고 있으며, 개발의 범위는 [Fig. 2]와 같이 소셜 컨텍스트들을 수집, 추출, 분류하는 과정들을 컴포넌트화 해서 구성한다. 각각의 컴포넌트는 현존하는 소셜 미디어 컨텐츠들을 수집하는 수집기, 의미 연관성 기반의 소셜 토픽 추출기와 주제 연관성 기반의 소셜 컨텐츠·피드백 분류기로 구성된다.

• 소셜 컨텍스트 수집기

다양한 소셜 미디어 채널로부터 소셜 컨텍스트를 수집하는 외부 소셜 미디어 컨텐츠 수집기와 동적으로 연동할 수 있는 어댑터를 개발하여 추가 확장 가능하도록 한다.

• 의미 연관성 기반 소셜 토픽 추출기

수집된 소셜 컨텍스트를 구성하는 단어들 간의 의미 연관성을 분석하여 해당 소셜 컨텍스트의 시맨틱 정보를 표현하는 토픽을 추출한다. 의미 연관성은 내재된 시맨틱 정보의 연결 관계를 나타내며 온톨로지로 표현된다. 본 연구에서는 딥러닝을 통해 의미 연관성 온톨로지의 자가 확장 구조를 개발한다.

• 주제 연관성 기반 소셜 컨텐츠/피드백 분류기

의미 연관성에 따라 추출된 소셜 토픽에 대해 동일한 주제의 소셜 컨텐츠와 피드백을 분류한다. 소셜 컨텐츠와 피드백을 분석하여 주제 키워드를 추출하고, 소셜 토픽의 의미 연관성을 기반으로 관련도를 측정하여 색인한다.

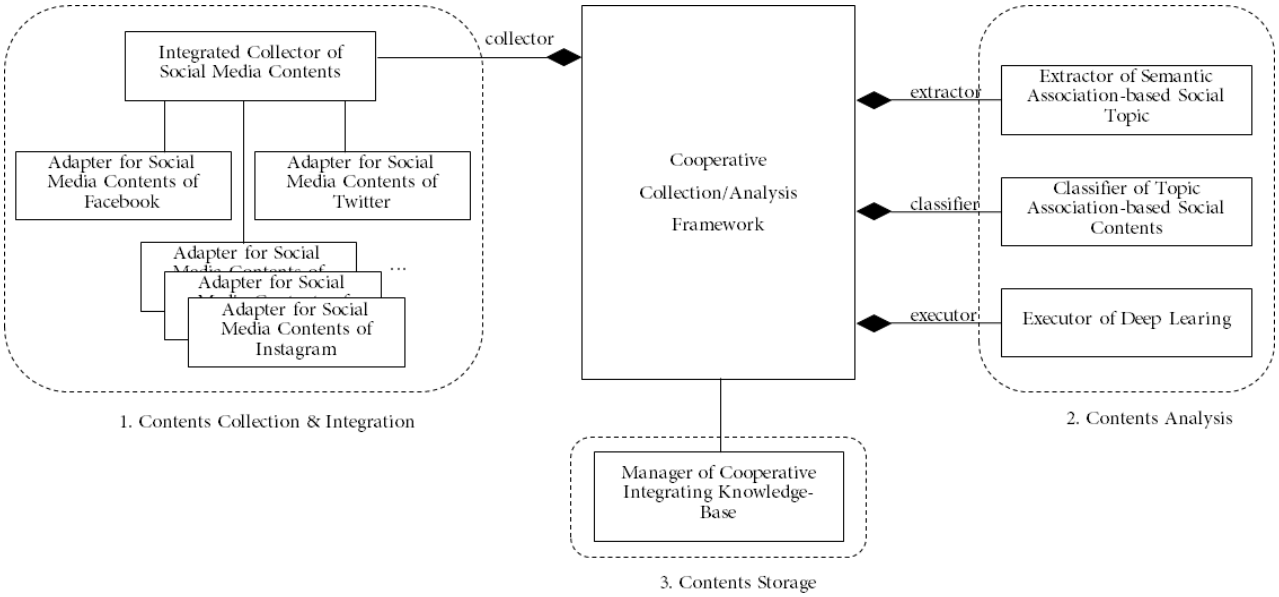

Fig. 2. Architecture of Framework

2. Collecting Social Context

본 논문에서 제시하는 프레임워크는 다양한 소셜 미디어 채널들을 통합하기 위한 통합 API를 개발하여, 트위터(Twitter), 페이스북(Facebook), 인스타그램(Instagram) 등 계속해서 늘어나는 다양한 소셜 미디어의 오픈API를 활용한 커넥터(Connector)를 지속적으로 개발하여 [Fig. 3]과 같이 확장 가능하도록 설계 한다.

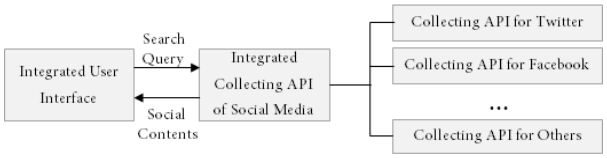

Fig. 3. Extensible Structure of Connector for Collecting Social Media

확장 가능한 통한 API 커넥터의 동작은 먼저 통합 사용자 인터페이스에서 검색 조건(검색어, 검색기간, 최대 개수 등)을 설정하고 검색을 실행한다. 소셜 미디어 커넥터 통합 API에 검색 쿼리가 전달된다. 소셜 미디어 커넥터에 등록된 다양한 확장 컴포넌트에 검색 쿼리를 브로드캐스팅한다. 확장 컴포넌트에서 리턴 된 수집 결과를 소셜 미디어 커넥터 통합 API에 전달한다. 통합 사용자 인터페이스로 결과를 리턴하여 화면에 표시한다.

본 논문에서는 다양한 소셜 미디어 컨텐츠 채널들로부터 제공되는 소셜 컨텍스트들을 수집하기 위해 각 채널들이 갖는 API의 이질성 문제를 효과적으로 해결하기 위해서 [Fig. 4]와 같이 어댑터(Adapter) 패턴을 적용한다.

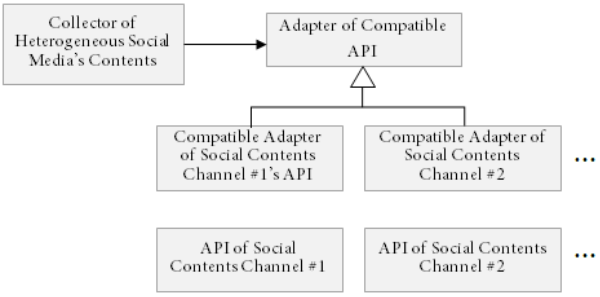

Fig. 4. Adapter Pattern based API Design

이 설계 기법을 적용하면, 각각의 이질적인 소셜 컨텐츠 채널 API에 의존하지 않고, API 호환 어댑터를 통해 소셜 컨텍스트를 동적으로 쉽게 연동하여 수집할 수 있다는 장점이 있다. 그리고, 새로운 채널이 추가되거나 다른 채널을 대체할 경우, 이기종 소셜 미디어 컨텐츠 수집기를 직접 수정하지 않고, 새로운 채널의 호환 어댑터만 구현하여 사용하면 객체 대체성(Substitutability)을 통해 교체가 가능하기 때문에 유지보수에 들어가는 비용이 크게 감소된다.

3. Extracting Social Topic based on Semantic Correlation

소셜 미디어 토픽 추출 및 분류 프레임워크는 소셜 컨텍스트를 관리하기 위한 온톨로지를 통해 소셜 컨텍스트의 연관관계를 정의하고, 이를 분석하여 소셜 토픽을 추출한다. 이 과정이 반복적으로 수행되면 소셜 컨텍스트의 집합과 소셜 토픽 간의 매핑 관계에 대한 표현이 가능하다.

[Fig. 5]는 상황에 적합한 소셜 토픽 추출 결과 및 사용자 피드백을 해석하여, 향후 소셜 토픽 추출 프로세스에

반영함으로써, 소셜 토픽 추출의 정확도를 계속 높여주는 기법을 나타낸다. 이 방식을 적용하면, 소셜 토픽 추출에 사용된 초기 온톨로지나 단어 간 의미 연관성 매핑에 사용하는 기초 데이터 등이 완전하지 않아도, 소셜 토픽 추출 프로세스의 실행을 반복하면서 점진적으로 그 정확성이 높아지게 된다.

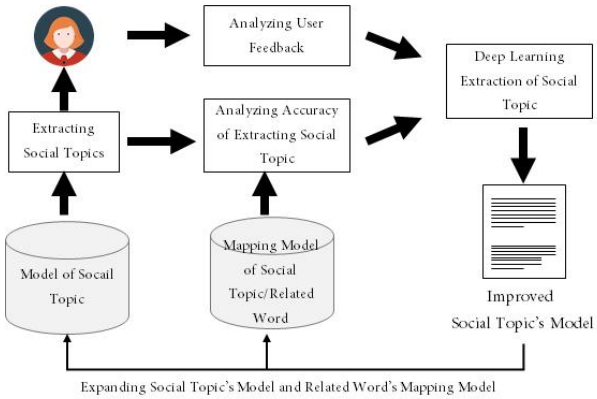

Fig. 5. Deep Learning for Improving Accuracy of Extracting Social Topic

3.1 Context-based Semantic Association's Measurement and Topic Extraction

소셜 미디어 컨텐츠는 비정형 데이터 형태로 여러 개의 단어들을 포함하고 있기 때문에, 의미정보를 담고 있는 소셜 컨텍스트를 기반으로 한 소셜 토픽 추출이 진행되어야 한다. 해당 소셜 컨텐츠를 구성하는 단어들 간의 연관 관계를 토대로 소셜 컨텍스트의 의미 연관성을 학습할 때, 소셜 토픽 추출의 정확도가 향상될 수 있다.

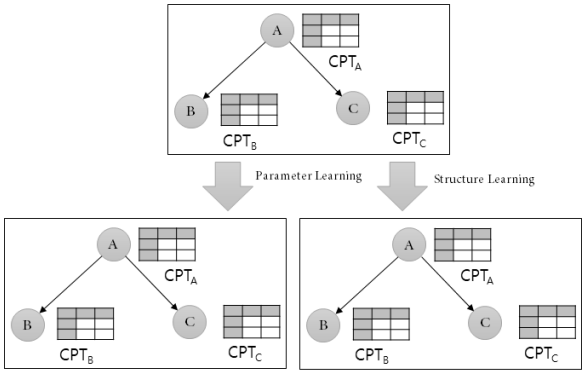

Fig. 6. Sematic Correlation's Learning of Social Contexts based on Bayesian Network

정확한 소셜 토픽 추출로 판단될 경우, 해당 추출 절차에 사용된 학습 데이터의 가중치를 높이고, [Fig. 6]과 같이 베이지안 네트워크(Bayesian Network) 모델을 이용해서 매개변수 학습과 구조 학습 모두를 수행한다. 매개변수 학습은 모델의 구조는 변하지 않고 모델의 각 노드의 CPT(Conditional Probability Table)을 개선시킴으로써 모델을 학습한다.

3.2 Extension of Semantic Association's Ontology and Improving Accuracy

소셜 컨텍스트는 해당 컨텍스트의 소유주와 작성 시간, 컨텐츠의 기본 구조를 갖으며, 소셜 컨텍스트 수집 API에 따라 보다 세분화된다. 따라서, [Fig. 7]과 같이 소셜 컨텍스트의 시그니처를 통해 시맨틱 정보의 기본 구조를 추출할 수 있다.

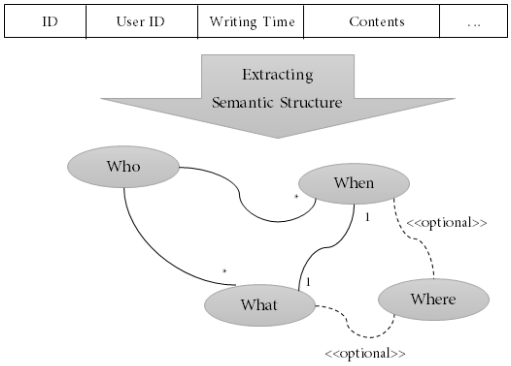

Fig. 7. Process of Extracting Semantic Structure form Social Context

다양한 소셜 컨텍스트로부터 시맨틱 구조를 추출하는 과정을 반복하여 시맨틱 구조는 점진적으로 확장되며, 이를 기반으로 향후 소셜 토픽 추출을 위한 온톨로지를 구축할 수 있다.

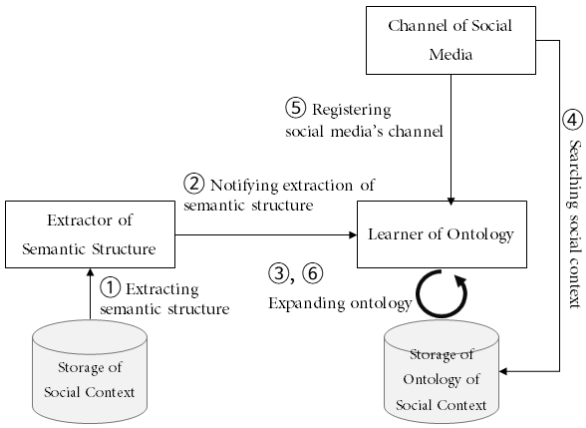

Fig. 8. Autonomous Extension Process of Social Context's Ontology

온톨로지는 [Fig. 8]과 같이 소셜 컨텍스트로부터 시맨틱 기본 구조를 추출하는 과정과 상호연동하며, 다양한 사용자의 소셜 컨텍스트 사용 이력을 학습하여 시맨틱 기본 구조를 자율적으로 확장한다.

4. Classifying Social Contents·Feedback based on Subject Correlation

4.1 Classification of Data(Contents/Feedback)

분류 대상 컨텐츠 및 피드백의 유형은 주로 순차형 데이터와 관계형 데이터로 구분된다. 이 두 종류의 집합은 서로 분석법이 다르며, 복잡한 상황 분석에서는 이 두 가지 유형의 집합을 함께 분석해야 하는 경우가 자주 발생한다. 그러므로 소셜 미디어 컨텐츠 및 피드백 분류기는 [Fig. 9]와 같은 절차를 가지는 두 단계 분석 (Double-Phase Analytics) 기법을 제공한다.

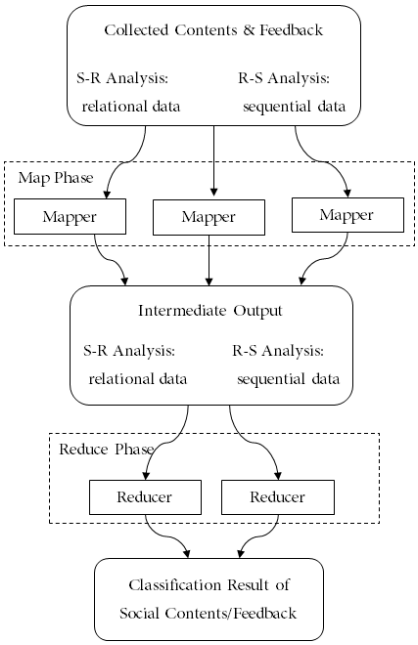

Fig. 9. 2 Phase Classification & Feedback of Social Contents

두 단계 분석 기법은 하둡(Hadoop)의 Map 단계와 Reduce 단계에서 순차형 데이터 집합과 관계형 데이터 집합을 따로 처리하는 기법이다. 두 단계 분석 기법은 Sequential-Relational(S-R) 분석과 Relational-Sequential (R-S) 분석 방법이다. S-R 분석은 Map 단계에서 순차형 데이터 집합을 먼저 분석하고 분석 결과로 나온 관계형 데이터를 Reduce 단계에서 분석하는 방법이고, R-S 분석은 Map 단계에서 관계형 데이터 집합을 먼저 분석하고 분석 결과로 나온 패턴 등의 순차형 데이터를 Reduce 단계에서 분석하는 방법이다.

4.2 Extracting Topic of Contents and Indexing Contents

소셜 미디어 데이터와 같은 비정형 데이터는 마이닝 기법을 사용하여 우선 분석 가능한 형태로 형식화하게 된다. 텍스트 마이닝은 [Fig. 10]과 같이 자연어처리 기술을 토대로 유의미한 정보를 추출하여 가공하는 것을 목적으로 한다. 텍스트 마이닝 기술을 통해 방대한 소셜 미디어 컨텐츠에서 의미 있는 정보를 추출해내고, 다른 정보와의 연계성을 파악하며, 소셜 미디어 컨텐츠가 가진 주제를 찾아낼 수 있다. 식별된 주제는 각 소셜 미디어 컨텐츠에 주제 연관 단어를 이용하여 색인되어 검색에 활용된다. 주제의 연관성 및 의미는 고정된 것이 아닌 시간 슬롯의 변화에 따라 함께 변경될 수 있기 때문에, 주제 색인 결과에 대한 지속적인 학습을 통해 변화에 대한 유연한 추적이 가능해야 한다.

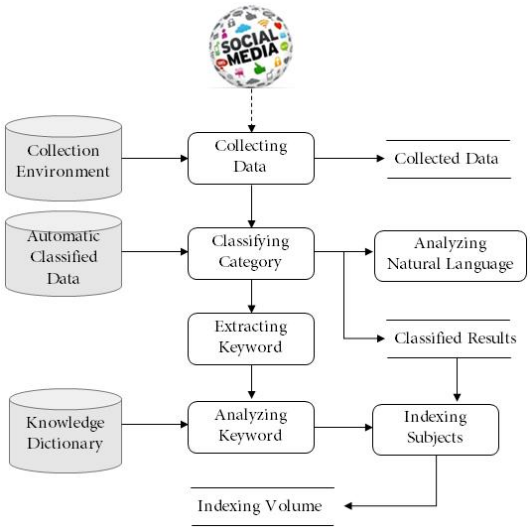

Fig. 10. Extracting Subjects and Indexing Process of Social Contents based on Text Mining

IV. Experiments and Evaluation

이 장에서는 본 논문에서 제시한 의미 연관성 기반 소셜 토픽 추출과 주제 연관성 기반 소셜 컨텐츠 분류 기법에 대한 실험과 평가 결과를 제시한다.

1. Experiments

제안한 소셜 토픽 추출 기법과 컨텐츠 분류 기법을 기반으로 협력형 수집·분석 프레임워크에 소셜 컨텐츠 수집기, 소셜 토픽 추출기, 그리고 컨텐츠 분류기로 개발해서 통합하였다.

소셜 토픽 추출을 위해서 네이버나 다음과 같은 주요 포털 사이트들을 대상으로 해당 사이트들의 뉴스 수집 및 분석을 통해 소셜 토픽 추출 실험을 [Table 1]과 같이 수행하였다.

Table 1. Lists of News

Year	Economics	Politics	Science	Society	Total
2013	35,865	42,914	19,933	46,606	145,318
2014	38,177	41,553	21,985	49,540	151,255
2015	86,874	97,295	38,536	112,158	334,863
2016	123,765	187,624	49,348	130,845	491,582
2017	124,455	211,147	44,242	141,190	521,034
2018	130,344	216,376	50,375	148,267	545,362
2019	137,236	230,193	53,912	152,723	574,064
2020	143,222	292,145	68,126	201,231	704,724

사물인터넷 관련 사회현상 분석 중 고용, 교육, 역기능에 대한 트렌드 변화 분석 결과는 [Fig. 11]과 같다.

2. Evaluation

제안한 기법을 가지고 소셜 토픽 추출의 정확도, 소셜 컨텐츠·피드백 분류의 정확도, 분석 시간, 시간당 처리량 등을 측정한 결과 [Table 2]와 같은 결과를 얻게 되었다.

V. Conclusions

기존 소셜 토픽 추출 기법의 문제점은 다수의 키워드들의 복합적 포함관계에 대해 확률적으로 계산해서 소셜 토픽 간의 유사도를 분석했지만, 일정 시점에서 구축된 통계적 모델이 추가적인 의미 또는 주제 키워드가 반영되지 못한다는 한계점을 갖고 있었다. 또한 비정형 데이터 형태인 소셜 컨텐츠들을 단순 키워드 포함 여부로만 연관성을 판별하는 것이 소셜 토픽 추출의 정확도를 저하시키는 문제점이 있다. 따라서, 본 논문에서 제시한 의미 연관성 기반 소셜 토픽 추출 기법은 이러한 문제점을 극복하기 위해 주기적으로 딥러닝을 통한 선행 학습을 통해 의미와 주제 연관성 기반으로 소셜 토픽을 추출하기 때문에 소셜 토픽 추출의 정확도와 컨텐츠 피드백 분류 정확도가 향상된다. 개발된 프레임워크 내에는 이질적인 다양한 소프트웨어 컨텐츠 채널들로부터 데이터를 수집하기 위한 통합 소셜 미디어 컨텐츠 수집기와 각 채널별로 연동할 수 있는 어댑터, 그리고 수집된 데이터를 가지고 딥 러닝을 통해 의미 기반의 소셜 토픽을 분석하고 추출하기 위한 추출기와 추출된 컨텐츠들의 주제를 기반으로 분류하는 분류기가 포함되었다. 향후 연구에서는 제시된 기법과 개발된 프레임워크를 보다 다양한 SNS 채널들에 적용하여 정확도와 성능에 대한 신뢰성을 확보하고, 프레임워크를 다양한 분야의 시스템에 적용할 것이다.

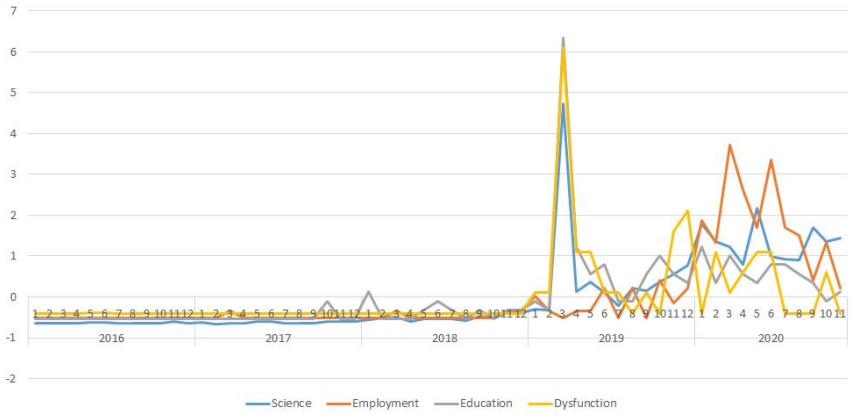

Fig. 11. Analysis Results of Trend Change of IoT-related Social Status

Table 2. Main Performance Indexes

Index	Unit	Expected Goal	Achievement Goal	Weight(%)
1. Accuracy of Extracting Social Contents	%	80 ↑	99.4%	20
2. Accuracy of Classification and Feedback Social Contents	%	80 ↑	100%	20
3. Num. of Media Data for Store	type	5 ↑	5	15
4. Num. of Collector for Social Media	num	5 ↑	5	15
5. Num. of Dictionary Items for Word Correlation	num	15,000 ↑	44,703	10
6. Analysis Time	sec	3 ↓	0.076	10
7. Throughput per sec	TPS	100 ↑	25,638/sec	10

ACKNOWLEDGEMENT

The present research has been conducted by the Research Grant of Seoil University in 2020.

REFERENCES

[1] Kushin, M.J. and M. Yamamoto, “Did Social Media Really Matter? College Students’ Use of Online Media and Political Decision Making in the 2008 Election”, Mass Communication and Society, Vol.13, No.5, pp.608-630, November, 2010. DOI:10.1080/15205436.2010.516863

[2] Michaelidou, N., N.T. Siamagka, and G. Christodoulides, “Usage, Barriers and Measurement of Social Media Marketing : An Exploratory Investigation of Small and Medium B2b Brands”, Industrial Marketing Management, Vol.40, No.7, pp.1153-1159. October 2011, DOI: 10.1016/j.indmarman.2011.09.009

[3] IDC, “Korea Big Data and Analytics Forecast, 2019-2023”, Wikibon, Jan 2020, https://www.idc.com/kr

[4] Byoung-Yup Lee, Jong-Tae Lim, and Jaesoo Yoo, “Utilization of Social Media Analysis using Big Data”, Journal of The Korea Contents Association, Vol.13, No.2, pp.211-219, February 2013, DOI: 10.5392/JKCA.2013.13.02.211

[5] Man-Mo Kang, Sang-Rak Kim, Sang-Moo Park, “Analysis and Utilization of Big Data”, Communications of the Korean Institute of Information Scientists and Engineers, Vol. 30, No. 6, 2012.6, pp. 25-32, June, 2012.

[6] Keun-Tae Kim, “Environment Challenge in Company for Big Data Analysis”, Korea Information Processing Society Review, Vol.19, No.2, March, 2012.

[7] Blei, D.M. and J.D. Lafferty, “Topic Models”, Text mining : Classification, Clustering, and Applications, 2009, Vol.10, No.71, pp.34, June, 2009, ISBN: 9781420059403

[8] Blei, D.M., A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation”, The Journal of Machine Learning Research, Vol.3, pp.993-1022, March 2003, DOI: 10.1162/jmlr.2003.3.4.-5.993

[9] Roo-daa Lee, Jin Man Kim, Joa Sang Lim, “Analysis of twitter topic using LDA”, Proceedings of Symposium of the Korean Institute of Communications and Information Sciences, , pp.1010-1011, UCI(KEPA): I410-ECN-0101-2016-567-002458809

[10] Tae Min Cho, Jee Hyung Lee, “Latent Keyphrase Extraction Using LDA Model”, Vol. 25, No. 2, pp.180-185, April, 2015, UCI(KEPA): I410-ECN-0101-2016-028-001346733

[11] Sangun Park, “Analysis of Social Media Contents about Broadcast Media through Topic Modeling”, Journal of Information Technology Services, Vol. 15, No. 2, pp.81-92, June, 2016, DOI: 1975-4256(pISSN)

[12] Kie-jin Park, “A Design on Informal Big Data Topic Extraction System Based on Spark Framework”, KIPS Transaction of Software and Data Engineering, Vol.5, No.11, pp.521~526, October, 2016, DOI: 2287-5905(pISSN)

[13] Jin-myeong Chung, Young-ho Park, Woo-ju Kim, “Social Media Analysis Based on Keyword Related to Educational Policy Using Topic Modeling”, Journal of Korean Society for Internet Information, No.19, Vol. 4, pp.53-63, Aug. 2017.

[14] Eun-sook Cho, et. al., “Development of Extracting System for Meaning·Subject Related Social Topic using Deep Learning”, Journal of the Korea Society of Digital Industry and Information Management, Vol.14, No.4, pp.35-45, December, 2018, 1738-6667(pISSN)

[15] Dong Wook Kim, Soo Won Lee, “News Topic Extraction based on Word Similarity”, Journal of the KIISE, Vol. 44, No.11, pp.1138~1148, DOI:10.5626/JOK.2017.44.11.1138

Authors

Eun-Sook Cho received the B.S. degree in Computer Science from DongEui University, Korea in 1993. She received the M.S and Ph.D degree in Computer Science from SoongSil University, Korea, in 1996 and 2000, respectively. Dr. Cho joined the faculty of the Department of Software Engineering at Seoil University, Seoul, Korea, in 2005. She is currently a Professor in the Department of Software Engineering, Seoil University. She is interested in framework modeling and development, Big Data, Service-Oriented Computing, and IoT Applications.