개체추출기법을 이용한 관계성 도출기법

김종희* · 이은석** · 김정수*** · 박종국**** · 김종배*****。

*,**,***,****,*****。숭실대학교 소프트웨어특성화대학원

A Study of Relationship Derivation Technique using object extraction Technique

Jong-hee Kim* · Eun-seok Lee** · Jeong-su Kim*** · Jong-kook Park**** · Jong-bae Kim*****。

*,**,***,****,*****。Graduate School of Software Soongsil Universityy

E-mail : skyimo@hanmail.net*, geoles95@ssu.ac.kr**, kjjjsss1@gmail.com***, eszero@gmail.com****

kjb123@ssu.ac.kr*****。

요 약

최근, 산재된 비정형 데이터 분석 등을 통한 빅데이터 활용에 대한 요구들이 증가하고 있으나, 아직까지 이에 대한 연구들이 부족한 실정이다. 따라서 본 연구에서는 수집된 웹 정보에서 개체들을 추출하여 이들 간의 관계를 집단지성 기술과 언어처리 기술을 통해 자동 분석해 냄으로써 문장단위의 의미기반 분석을 할 수 있는 기법을 제시한다. 이를 위해, 수집된 정보를 DBMS에 정형화된 형태로 저장한 후 형태소와 자질정보를 분석한다. 획득한 형태소 중 관심개체, 주변개체, 비관심 개체를 분류하고 개체간 속성인식기법을 이용하여 각 개체간의 관계를 정도, 범위, 성격 등으로 분석한다. 그 결과, 긍정·부정의 판단이 가능한 개체간의 관계성 도출기법을 제시함으로써, 특정 키워드를 대상으로 분석된 정보들의 연관도를 분석할 수 있었다. 이 연구를 통해, 최근 실시간 대용량 처리 시스템에 적합한 시스템을 설계하여 이를 부가가치가 높은 서비스에 적용할 수 있는 방법을 제시하였다.

ABSTRACT

Despite increasing demands for big data application based on the analysis of scattered unstructured data, few relevant studies have been reported. Accordingly, the present study suggests a technique enabling a sentence-based semantic analysis by extracting objects from collected web information and automatically analyzing the relationships between such objects with collective intelligence and language processing technology. To be specific, collected information is stored in DBMS in a structured form, and then morpheme and feature information is analyzed. Obtained morphemes are classified into objects of interest, marginal objects and objects of non-interest. Then, with an inter-object attribute recognition technique, the relationships between objects are analyzed in terms of the degree, scope and nature of such relationships. As a result, the analysis of relevance between the information was based on certain keywords and used an inter-object relationship extraction technique that can determine positivity and negativity. Also, the present study suggested a method to design a system fit for real-time large-capacity processing and applicable to high value-added services.

키워드

자동분석, 정보 분석기, 개체추출기, 속성인식기

Ⅰ. 서 론

개방과 공유를 표방하는 웹 2.0의 온라인 공간에는 매우 많은 지식과 정보들이 산재해 있으며, 수많은 웹 이용자들의 생각이나 감정 등이 텍스트, 이미지 등의 형태로 존재하게 된다. 따라서 주로 기업 등에서 마케팅의 목적으로 특정한 주제(상품)에 대한 텍스트 분석을 통해'텍스트 마이닝'에 대한 연구가 활발하게 진행되고 있다. 하지만, 기존의 '텍스트 마이닝' 기법은 한 가지 주제에 대해 단순하게 정량적으로 분석하고 있다. 따라서, 본 연구에서는 웹 환경에 산재된 정형 및 비정형 정보들을 정제된 형태로 주기적으로 수집하고 이들 수집된 정보에서 관심 개체들과 주변

개체들을 추출하여 개체간 “관계 성격”, “관계 정도” 을 집단지성 기술과 언어 처리 기술을 통해 자동 분석해 낼 수 있는 기법의 개발을 목표로 한다. 이를 통해 기존의 ‘텍스트 마이닝’ 에서 분석하는 범위를 넘어서 관심개체와 주변개체간의 관계정도, 관계범위, 관계성격을 분석해 낼 수 있다.

II. 관련 연구

데이터는 정형화된 정도에 따라 정형, 반정형, 비정형 데이터로 구분된다. 차이는 고정된 필드에 저장되느냐의 유무에 따라 구분되며, 반정형 데이터의 경우 비정형 데이터와 같이 고정 필드에 저장되어 있지는 않으나, 메타데이터나 스키마 등의 정보를 포함하는 데이터를 의미한다[2][4]. 본 논문에서는 인터넷에서 수집된 데이터의 분석을 위해 텍스트마이닝 기법을 사용한다[3]. 구조화된 텍스트에서 유용한 정보를 발견하는 것과 같은 기존 텍스트 마이닝의 문제점은 오직 구조화된 데이터베이스의 형태로만 사용할 수 있다는것이다[5]. 본 논문에서는 텍스트 마이닝 기법과 상호 응용되어 텍스트 마이닝을 통해 분류된 단어 및 문장들을 사용하여 본 연구의 핵심인 개체간에 관계 도출을 목적으로 오피니언 마이닝 분석을 수행한다.[1].

III. 개체간의 관계성 도출 기법

개체간의 관계성 도출을 위해서 인터넷상의 정보를 선택적으로 수집하여 데이터베이스에 정형화된 형태로 저장하고 수집된 정보는 정보분석을 거치며 형태소정보 및 자질정보로 나뉘어 분석한다. 획득한 형태소 중에서 개체추출기를 통해 관심개체와 주변개체 그리고 비관심개체로 분류하고 개체간 속성인식기를 이용하여 각 개체간의 관계를 정도, 범위, 성격 등으로 파악한다. [그림 1]은 데이터 분석의 절차를 나타낸다.

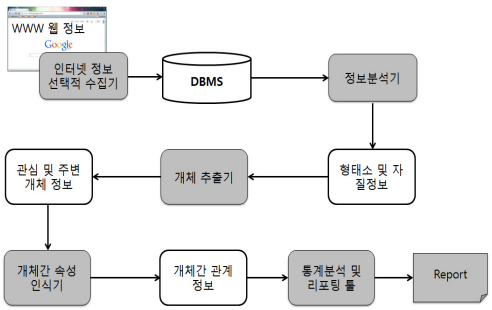

그림 1. 데이터 분석 기본모델

본 논문에서는 정보분석기, 개체추출기, 속성인식기에 대해서만 설명한다. 본 연구는 개체 추출에 목적이 있으므로 단순 품사 정보 외에도 각 형태소의 자질 정보를 분석할 수 있어야 한다. 이를 위해 “개체 추출기”를 통해 각 형태소의 세부 자질 정보를 추출한다. 이 연구에서의 “개체추출기”로는 정규표현식을 통한 주민번호, 전화번호, 날짜 정보와 같은 정형화된 패턴 탐색과 인명에 대한 확률과 통계를 활용하여 인명 개체를 추출할 수 있도록 하였다. 주민번호, 전화번호, 날짜, 이메일과 같은 정보들은 다음과 같은 정규표현식에 의해 표현이 가능하며 해당 항목의 문자열이 정규식 패턴과 일치 하는지 여부를 확인하고 모든 필요 개체들을 탐색하여 추출한다. 아래와 같은 정규 표현식을 사용하면 한 개의 단어 또는 구문만이 아닌 텍스트 패턴의 일치 여부를 판단할 수 있다.

표 1. 항목별 정규표현식

항목	정규표현식
이메일	[0-9a-y]([-_.]?[0-9aa])*@[0-9a-y]([-_.]?[0-9a-y])*.[a-y]{2,3}
주민번호	[0-9]{6}-[1-4][0-9]{6}
전화번호	([0-9]{2,4}-[0-9]{3,4}-[0-9]{3,4}[^0-9])|([^0-9][0-9]{3,4}-[0-9]{3,4})
날짜형식 : YYYY-MM-DD YYYY/MM/DD	[12][0-9]{3}(-|/)(0[1-9]|1[012])(-|/)(0[1-9]|[12][0-9]|3[0-1])
중간 생략	
날짜형식 : 2004/11/22 05:32:56	[0-9]{4}(-|/|\.)[0-9]{2}(-|/|\.)[0-9]{2}+[0-9]{1,2}:[0-9]{1,2}:[0-9]{1,2}

동시 발생 빈도에 의해 두 개체가 얼마나 관계성이 높은지를 알 수 있으며 이를 “관계의 정도”라고 한다. 개체 추출기에 의해 추출된 “개체”들 간의 동시발생빈도 정보들을 통해 “개체”간의 연관도를 분석한다. 이때 단어 발생 빈도가 단순하게 높다는 이유로 고빈도 단어는 무조건 연관도가 높게 측정되는 현상을 보정하기 위해 PMI(Point-wise Mutual Information)를 사용하여 연관도를 계산한다. PMI가 높을수록 연관도가 높은 것이며 반대인 경우는 연관도가 낮은 경우이다. PMI에 의해 알아낸 연관도 만으로도 관계성이 높은지 낮은지는 알 수 있지만 관계의 성격이 좋은지 나쁜지는 알 수 없다. 연관도에 성격을 부여하기 위해서는 각 개체 동시에 나타난 문장의 긍정·부정 판단을 수행하고 각 개체간의 긍정·부정관계를 매칭 하여 연관의 정도뿐만 아니라 연관의 성격을 알아 낼 수 있다.

IV. 연구결과 및 결론

"개체간 관계성 도출 기법"을 검증하기 위하여, 여러 SNS에서 수집한 정보들을 일정한 형태로 정리하여 MongoDB에 저장하였다. 주요 테이블을 구성하는 데이터의 수집기간은 약 한 달간으로 수집된 총 데이터의 량은 약 22만 건이다. 수집된 SNS 데이터를 정보 분석과정을 거쳐 얻은 약 22만개 키워드에서 개체 추출과정을 통해 얻은 상위 개체 30개를 선정하였다. 총 7개의 키워드로 수집된 SNS 정보로부터 파생된 새로운 개체들이 추출되었고 이들 전체 개체들의 상호간 관계성을 수치화시키면 다음과 같다.

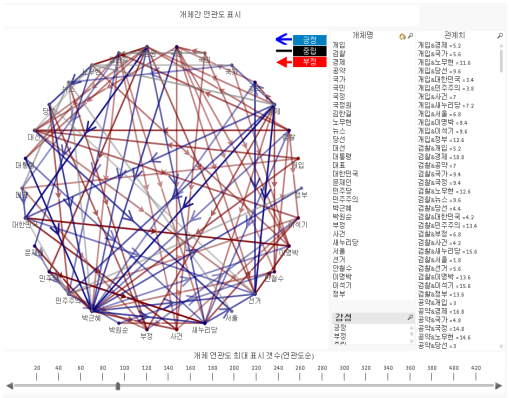

그림 2. 실험 결과에 따른 전체 개체의 상호연관도(긍정·부정 포함)

본 논문에서는 기존의 텍스트 마이닝을 뛰어넘는 긍정·부정의 판단이 가능한 개체간의 관계성 도출기법을 제시함으로써 특정 키워드를 대상으로 분석된 정보들의 연관도를 분석할 수 있었다. 또한 최근 실시간 대용량 처리 시스템에 적합한 시스템을 설계하여 이를 부가가치가 높은 서비스에 적용할 수 있도록 효율성과 실용성을 극대화 시켰다.

참고문헌

[1] 김진옥, 이선숙, 용환승, "한글 텍스트의 오피니언 분류 자동화 기법", 정보과학회논문지, 데이터베이스 38.6, 2011

[2] 양창준, "미래의 창, 빅 데이터", 한국정보통신기술협회, TTA Journal Vol.140pp. 6-23, 2012

[3] 컴퓨터연구정보센터(CSERIC), "텍스트마이닝(TextMining)", (http://www.cseric.or.kr/new_cseric/yungoostep/content.asp?idx=903&startpage_view=900&startpage=900&page=1)

[4] Gartner,'Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data', 2011.6

[5] Text Mining with Information Extraction University of Texas, Austin, (http://www.cs.utexas.edu/~ml/papers/discotex-melm-03.pdf)