Journal of The Korea Society of Computer and Information
Vol. 19, No. 4, April 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.4.043

국가별 관심도 측정을 위한 온톨로지 기반 위키피디아 사용 데이터 분석

김 현 희*, 조 진 남*, 김 동 건*

An Ontology-based Analysis of Wikipedia Usage Data for Measuring degree-of-interest in Country

Hyon Hee Kim*, Jinnam Jo*, Donggeon Kim*

요 약

본 논문에서는 위키피디아 사용 데이터를 분석하여 국가별 관심도를 측정하는 기법을 제시하였다. 먼저 해당 국가에 대한 분야별 관심도를 측정하기 위해서 위키피디아 카테고리로부터 개념 계층 구조를 추출하여 관심도 온톨로지를 구축하였다. 관심도 온톨로지는 국가에 대한 관심 분야를 정치, 경제, 사회, 그리고 문화로 분류하고 각 대분류에 대해 다시 세부 분야으로 분류하였다. 다음으로, 특정 국가의 페이지에서 자주 편집된 기사들의 제목을 관심도 온톨로지에 매핑하여 분야별 페이지 뷰를 분석하였다. 마지막으로 한국, 중국, 그리고 일본에 대한 국가별 관심도를 측정하고 국가별로 위키피디아 사용자들의 관심 분야가 다른지 판별하기 위해서 카이 제곱 독립성 검정을 실시하였다. 실험 결과는 위키피디아 사용자들의 관심 분야가 각 국가와 연관성이 있음을 보여준다. 본 연구는 기존의 설문조사 방식으로 국가 이미지를 측정하는 경우보다 적시에 그리고 유연하게 분야별 관심도를 측정할 수 있는 방안을 제시하며, 위키피디아 사용 데이터 분석 결과를 국가 이미지 개선을 위해 분야별로 재고할 방향을 제시한다.

▸ Keywords : 위키피디아 사용 데이터 분석, 온톨로지 기반 분석, 관심도 온톨로지, 분야별 관심도 측정

Abstract

In this paper, we propose an ontology-based approach to measuring degree-of-interest in country by analyzing wikipedia usage data. First, we developed the degree-of-interest ontology called DOI ontology by extracting concept hierarchies from wikipedia categories. Second, we map the title of frequently edited articles into DOI ontology, and we measure degree-of-interest based on DOI ontology by analyzing wikipedia page views. Finally, we perform chi-square test of independence to figure out if interesting fields are independent or not by country. This approach shows interesting fields are closely related to each country, and provides degree of interests by

•제1저자 : 김현희 •교신저자 : 김현희
•투고일 : 2014. 3. 7, 심사일 : 2014. 3. 20, 게재확정일 : 2014. 3. 27.
* 동덕여자대학교 정보통계학과(Dept. of Information and Statistics, Dongduk Women's University)

country timely and flexibly as compared with conventional questionnaire survey analysis.

▸ Keywords : Analysis of Wikipedia Usage Data, Ontology-based Analysis, Degree-of-Interest Ontology, Measuring degree-of-interest by fields

I. 서 론

소셜 네트워킹 서비스(SNS) 사용자들의 태그, 평가글, 그리고 페이지 뷰 등의 웹 사용 데이터를 분석하여 제품 마케팅이나 개인화된 추천 서비스에 적용하고자 하는 노력이 활발해지고 있다. 특히 제품을 생산하는 국가의 이미지가 제품 구매 및 충성도에 영향을 미치는 것으로 알려짐에 따라[1], SNS 사용 데이터로부터 국가 이미지를 측정하여 국가 이미지를 향상시키고 마케팅에 활용할 필요성이 요구된다.

현재 가장 많이 사용되고 있는 안홀트 국가 브랜드 지수[2]는 문화 및 유산, 국민, 수출, 관광, 정부, 그리고 투자 및 이민의 6가지 분야로 분류하고 설문 조사를 통해 1년에 한번씩 국가 브랜드 지수를 발표하고 있다. 한 국가의 브랜드 지수를 계산하기 위해서 10개국 10,000명의 소비자에게 설문조사를 실시하고 이중 다시 1,000 명의 대표 샘플을 추출하여 브랜드 지수를 계산한다.

이와 같이 설문 조사를 이용하여 국가 브랜드 지수를 산출하면 비용이 비싸고 미리 정의된 6개의 분야에 대해 측정하게 되어 각 분야별 구체적인 정보를 얻을 수 없다는 단점이 있다. SNS 사용자 데이터를 분석하여 국가 이미지를 측정하면 시간 및 공간적 제약이 없고 비용 면에서 효율적이다. 또한 현재 안홀트 국가 브랜드 지수에서 얻을 수 없는 각 분야별 구체적 관심 분야를 파악할 수 있다.

제품 이미지 분석에는 트위터 데이터가 많이 활용되고 있다[3,4]. 트위터는 140자 이내의 단문을 전송하는 특성 때문에 대부분 사용자의 단편적 감정이나 단순 정보를 포함한다. 또한 그 전파 속도가 상당히 빠르므로 신제품에 대한 고객의 반응을 파악하는데 성공적으로 사용될 수 있다. 그러나 잘 알려진 바와 같이 트윗의 80% 이상이 개인적인 잡담을 포함하므로 국가 이미지와 같은 객관전 판단을 필요로 하는 정보 분석에는 적절하지 못하다.

본 연구에서는 국가에 대한 사용자들의 관심 정도를 분석하기 위해 위키피디아를 선정하였다. 위키피디아는 사용자들이 협업하여 페이지를 생성하고 편집할 수 있는 온라인 백과사전[5]으로서 대부분 객관적인 정보를 포함하기 때문이다. 위키피디아 사용 데이터 중에서 편집 횟수와 클릭 횟수를 분석하였는데, 상위 편집된 기사의 제목은 해당 국가에 대한 전문적 지식을 가진 사용자들의 관심 정도를 나타내고 클릭 횟수는 일반적인 사용자들의 관심 정도를 나타낸다고 볼 수 있다.

국가에 대한 관심도 혹은 국가 이미지에 대한 측정은 다양한 방식으로 이루어지고 있다. Pappu[1]는 Aaker가 정의한 브랜드 자산의 요소인 인지도, 품질, 그리고 충성도[6]를 국가 이미지에 적용하여 국가 이미지와 제품 이미지와의 상관관계를 분석하였으며 이밖에 기본 정의 요소를 변형하여 국가 이미지를 측정하고 있다. 본 연구에서는 웹 사용 데이터가 편집 횟수와 클릭 횟수임을 고려하여 국가에 대한 관심도로 한정하여 측정하였다.

본 연구에서는 위키피디아 사용 데이터로부터 해당 국가에 대한 관심 분야와 분야별 관심 정도를 측정하기 위해서 관심도 온톨로지를 개발하였다. 먼저 관심도 온톨로지의 상위 클래스를 정치, 경제, 사회, 그리고 문화의 대분류로 나누고 각 대분류에 대한 소분류는 위키피디아 카테고리 구조로부터 추출하였다. 다음으로 상위에 편집된 기사들의 제목을 관심도 온톨로지에 매핑한 후, 선정된 기사들에 대한 클릭 횟수를 관심 정도로 계산하였다.

관심 정도는 관심도 지수로 나타나는데 해당 국가의 기사에서 상위 편집된 20개의 기사를 선정한 다음, 다시 선정된 20개의 기사 내용 중에서 상위 편집된 20개의 기사를 선정한다. 이렇게 선정된 총 400개의 기사의 제목에 대한 클릭수를 관심도 지수로 환산하였다. 국가별로 관심 분야에 차이가 있는지 판별하기 위해서 한국, 중국, 그리고 일본의 3개국을 선정하고 위키피디아 영어판에서 2013년 5월 1일부터 2013년 5월 31일까지 한 달간 자료를 수집하였다.

분야별 관심도 지수에 대한 카이 제곱 독립성 검정을 실시한 결과 국가와 대분류 사이에는 연관성이 있음을 알 수 있

었다. 좀 더 구체적으로 한국의 경우는 정치와 문화 분야에 기댓값보다 많은 관심을 보이고 있고, 중국의 경우는 경제와 정치 분야에 기댓값보다 많은 관심을 보이고 있음을 알 수 있다. 일본의 경우는 사회와 문화 분야에 기댓값보다 많은 관심을 보이고 있고 정치 및 경제 분야는 기댓값보다 적은 관심이 나타났다.

정치 분야의 소분류는 통치자, 군대, 정부, 행정, 그리고 대외관계로 분류하였으며 한국은 대외 관계에 많은 관심이 나타났고 중국은 통치자와 정부에 일본은 군대와 통치자에 많은 관심이 나타났다. 경제 분야의 소분류는 재정, 산업, 인프라, 복지 그리고 과학기술로 분류하였다. 경제 분야에서 한국은 복지와 재정 분야에, 중국은 산업과 과학기술 분야에 그리고 일본은 인프라 분야에 기대치보다 높은 관심도를 보이고 있다.

사회 분야의 소분류는 지리, 교육, 역사, 언어, 종교, 그리고 민족으로 분류하였으며 한국의 경우 지리, 중국은 민족, 그리고 일본은 역사와 밀접한 관계가 있음을 알 수 있었다. 문화 분야은 예술, 음악, 엔터테인먼트, 필름, 그리고 스포츠로 분류하였으며 한국은 음악과 중국은 필름 및 음식, 그리고 일본은 스포츠와 엔터테인먼트와 밀접한 관계를 보이고 있었다.

국가의 이미지가 제품 이미지에 많은 영향을 미침에 따라 국가마다 국가 브랜드 관리를 위한 노력이 활발히 이루어지고 있다. 설문 조사를 통한 국가 브랜드 지수 계산은 많은 시간과 비용이 소요되고 필요시마다 브랜드 지수를 추출할 수 없다는 제약점을 갖는다. 본 연구에서 제안한 방법과 같이 위키피디아 사용 정보를 활용하면 국가의 다양한 분야에 대한 브랜드 지수를 보다 구체적이고 효율적으로 측정하고 분석할 수 있다.

본 연구의 공헌은 온톨로지를 기반으로 하여 위키피디아 사용 데이터로부터 국가의 관심 분야 및 관심도 지수를 측정한 것이다. 카이 제곱 검정을 통해 한국, 중국, 그리고 일본의 3개국이 국가별로 관심 분야가 다르며 대분류를 다시 소분류로 구체화하여 구체적으로 관심 지수를 나타내는 분야들을 확인할 수 있었다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 관련 연구를 살펴보고, 3장에서 본 연구에서 제안하는 관심도 온톨로지를 자세히 설명한다. 4장에서는 실험 방법 및 분석 결과를 서술하고 마지막으로 5장에서 결론 및 향후 연구를 제시한다.

II. 관련 연구

온톨로지는 개념에 대해 컴퓨터가 해석하고 처리할 수 있도록 정형화된 언어로 정의한 것[7]으로 전통적인 방식은 전문가 지식을 수작업으로 구축하는 것이다[8]. 이 방법은 전문가 지식이 정확하게 온톨로지화 될 수 있다는 장점은 있으나 시간 및 인적 제약 때문에 실용적 온톨로지 구축이 어렵다. 따라서 웹 문서와 같은 정보 자원으로부터 지식을 추출하여 온톨로지 구축을 자동화하는 연구가 활발해지고 있다[9].

특히 온라인 백과사전인 위키피디아는 텍스트 정보에 카테고리나 인포박스와 같은 구조적 지식을 포함하고 있으므로 웹 문서보다 지식 추출이 용이하다. YAGO[10] 온톨로지는 위키피디아 카테고리 구조와 인포박스로부터 Is-A 관계와 다른 의미 관계를 추출하여 구축되었다. DBpedia[11] 역시 위키피디아로부터 구조적 정보를 추출하여 시맨틱 웹으로 표현한 온톨로지이다. 이밖에도 단어 사이의 관련성을 위키피디아에서 링크된 거리로 계산하는 등[12] 온톨로지 구축에 다양하게 활용되고 있다.

위키피디아 사용 데이터 또한 그 분석 결과를 의사 결정에 활용하고자 하는 시도가 이루어지고 있다. Moat[13]와 그의 동료들은 위키피디아에서 편집 횟수와 클릭 횟수를 분석하여 주식 시장의 동향을 예측하였다. 이 연구에서 주식 시장의 변화가 있기 전에 사용자들은 관련 기사들을 편집하거나 클릭하는 동향을 보였다.

국가 이미지 측정은 대부분 5점 척도의 설문조사를 통해 이루어지고 있다[14]. 설문조사는 시간과 인력이 많이 필요하다는 비용적 문제 외에도 다음과 같은 문제점들을 안고 있다. 설문 조사를 실시할 경우 문제수가 많아질수록 정확도가 떨어질 수 있고, 응답자들이 제공한 점수가 실제 응답자들이 평가한 점수와 차이점을 보일 가능성도 있다.

본 연구에서 시도한 것과 같은 위키피디아 사용 정보는 사용자들의 웹 사용 패턴이라는 암시적 정보로부터 국가에 대한 관심도를 추출한 것이므로 응답자가 인위적으로 점수를 제공할 가능성을 방지할 수 있다.

III. 본 론

본 장에서는 먼저 제 1절에서 관심도 온톨로지 개발 프로세스를 살펴보고, 제 2절에서 관심도 온톨로지의 기본 구조 및 정의 규칙에 대해서 자세히 살펴본다. 마지막으로 제 3절에서 관심도 온톨로지의 작동 방식을 구체적으로 설명한다.

1. 관심도 온톨로지 개발 프로세스

그림 1은 관심도 온톨로지 개발 프로세스를 나타낸다. 관심도 온톨로지는 크게 클래스 정의와 인스턴스 등록 단계로

나눈다. 클래스 정의는 다음과 같은 방식으로 수행되었다.

먼저 관심도를 측정하기 전에 관심 분야에 대한 대분류를 정의하였다. 대분류는 가장 일반적으로 사용되는 정치, 경제, 사회, 그리고 문화의 4대 분야로 분류하였고 그림 1의 오른쪽 그림에서 보이는 바와 같이 관심도 온톨로지에서 상위 클래스로 정의된다. 각 대분류 분야에 속하는 소분류 분야는 위키피디아 카테고리로부터 추출하여 분류하였다. 대분류와 소분류 간의 계층 구조를 정의하고 소분류를 하위 클래스로 생성하여 Is-A 관계를 형성하였다.

본 연구에서 사용된 인스턴스는 자주 편집된 기사들의 제목에 해당되며 이를 키워드라고 정의하였다. 먼저, 한국, 중국, 그리고 일본에 대한 위키피디아 기사 중에서 가장 많이 편집이 일어난 20개의 키워드들을 선정하였다. 해당 키워드들을 클릭하였을 때 하이퍼링크로 연결된 페이지들은 해당 키워드를 기사의 제목으로 사용한다. 이 기사들에서 다시 자주 편집된 20개의 키워드를 선정하여 국가별로 약 400개의 키워드를 선정하였으며 총 약 1,200개의 키워드를 인스턴스로 처리하였다.

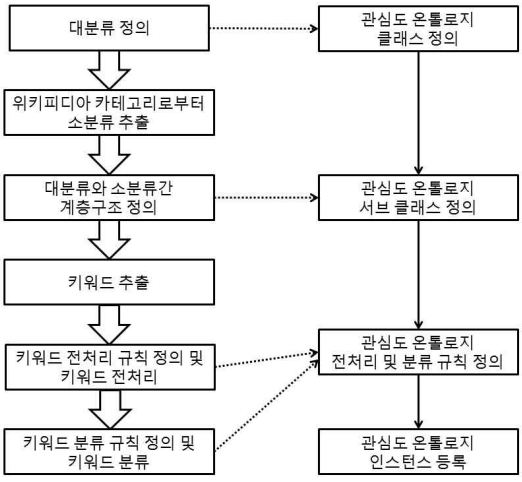

그림 1. 관심도 온톨로지 개발 프로세스
Fig. 1.Development Process of DOI Ontology

추출된 모든 키워드를 인스턴스로 활용할 수 있는 것은 아니다. 키워드들 중에는 중복된 경우도 있고, 같은 의미를 갖는 경우도 있으며 간혹 적절치 않은 키워드가 선정되는 경우도 있었다. 따라서 이러한 키워드들에 대한 전처리를 위해서 관심도 온톨로지에 전처리 규칙을 정의하였다.

마지막으로 전처리한 키워드들은 적절한 클래스에 인스턴스로 등록된다. 이때 분류가 모호한 인스턴스의 경우 그 분류 기준을 온톨로지에 정의한다. 정의된 규칙 외에 유사한 패턴이 발견될 경우 다시 온톨로지에 추가하여 점차적으로 분류 규칙을 확장하고 자동화가 가능하도록 하였다.

2. 관심도 온톨로지 구조

관심도 온톨로지는 위키피디아 사용자들의 국가에 대한 관심 정도를 분야별로 알아보기 위한 도메인 온톨로지이다. 그림 2는 관심도 온톨로지의 기본 구조를 나타낸다. 먼저 최상위 클래스로서 *DOI:IFields* 관심 분야 클래스를 정의하였다. 관심 분야 클래스는 D*OI:Politics*, *DOI: Economy*, *DOI: Society*, 그리고 *DOI:Culture*의 정치, 경제, 사회, 문화의 4개의 대분류 관심 분야 클래스로 구성된다. 대분류 관심 분야 클래스는 일반적인 사전적 정의라기보다는 국가에 대한 관심 정도를 추출하기 위해 도메인에 맞게 재정의하였다.

그림 2는 관심도 온톨로지의 계층 구조를 보여준다. 정치 클래스는 "국가의 운영 또는 운영에 영향을 미치는 활동"을 정의한다. 경제 클래스는 "개인 혹은 기업 및 국가의 전반적인 경제활동"을 정의한다. 사회 클래스는 "국가를 이루는 조직의 구조적 시스템"을 정의한다. 문화 클래스는 "국가의 가치관 및 행동 양식 등을 나타낼 수 있는 요소"들을 정의한다.

대분류 관심 분야 클래스의 하위 클래스인 세부 관심 분야 클래스를 정의하기 위해서 기존의 국가 브랜드 자산 평가 모형인 안홀트의 국가 브랜드 지수와 위키피디아 카테고리 구조를 고려하였다. 위키피디아 카테고리 구조는 국가별로 약간의 차이점은 있으나 공통적인 카테고리를 먼저 추출한 다음, 선정된 키워드들을 포함할 수 있는 카테고리들을 하위 클래스들로 정의하였다.

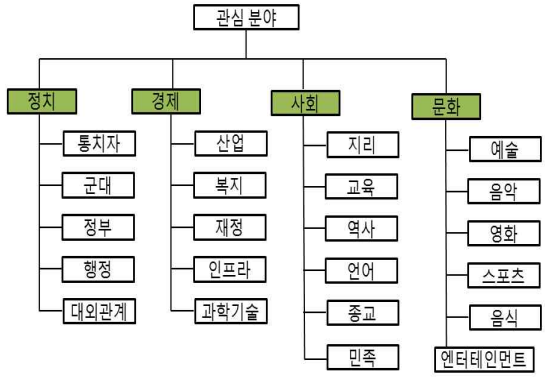

그림 2. 관심도 온톨로지 계층 구조
Fig. 2.Class Hierarchy of DOI Ontology

먼저 정치 분야에 대한 관심도를 측정하는 정치 클래스를 살펴보면 통치자, 군사, 정부, 행정, 그리고 대외관계를 하위

클래스로 갖는다. 통치자 클래스에 해당하는 키워드들은 각 나라의 대통령에 대한 키워드, 대통령직에 대한 서술 등을 포함한다. 정부 클래스는 국가의 정부 조직에 대한 키워드를 포함하며, 행정 클래스는 행정적 관리 관련 키워드를, 그리고 대외 관계는 외국과 해당 국가에 대한 정보를 포함한 키워드들이 속한다. 마지막으로 군사 클래스는 국가의 군사 관련 키워드를 포함한다.

경제 클래스는 국가의 경제적 측면에 대한 관심도를 측정하기 위한 클래스이다. 국가의 산업 전반에 걸친 정보를 나타내는 산업 클래스, 복지에 대해 나타내는 복지 클래스, 재정에 대한 키워드를 표현하는 금융 클래스, 산업을 유지하기 위한 토대를 설명하는 키워드들로 구성된 인프라 클래스, 그리고 과학 기술에 대한 정보를 나타내는 과학기술 클래스를 하위 클래스로 갖는다.

사회 클래스는 사회 전반에 걸친 인지도를 측정하는 클래스이다. 국가의 지리 정보를 나타내는 지리 클래스, 교육 정보를 나타내는 교육 클래스, 역사 정보를 나타내는 역사 클래스, 한 국가에서 사용하는 언어에 대해 나타내는 언어 클래스, 국가의 종교에 대한 정보인 종교 클래스, 마지막으로 국가를 구성하는 민족에 대한 정보인 민족 클래스로 구성된다.

마지막으로 문화 클래스는 국가의 문화적 측면에 대한 인지도를 측정하는 클래스이다. 예술, 음악, 영화, 스포츠, 음식, 그리고 전통 문화 및 연예 관련 정보를 나타내는 엔터테인먼트의 하위 클래스로 구성된다.

최상위 클래스인 관심 분야 클래스는 *DOI:KID*, *DOI:Keyword*, 그리고, *DOI:Value*의 세 개의 속성을 갖는다. *DOI:KID* 속성은 각 키워드의 유일한 식별자로서 자동 생성된다. *DOI:Keyword*는 선정된 키워드들을 뜻하고, *DOI:Value*는 해당 키워드의 클릭 횟수를 의미한다. 모든 선정된 키워드들은 단지 한 개의 세부 관심 분야 클래스에만 속할 수 있으며, 만일 해당 키워드의 세부 관심 분야 클래스가 명확하지 않을 경우 온톨로지 관리자에 의해 보다 관련 깊다고 판단되는 클래스의 인스턴스로 등록된다.

3 키워드 매핑 및 관심도 지수 측정 기법

선정된 키워드들의 관심도 온톨로지로의 매핑은 다음과 같은 방식으로 이루어졌다. 먼저, 각 단어들은 관심도 온톨로지의 세부 관심 분야 클래스의 인스턴스로서 등록된다. 이 때, 포괄적인 개념을 나타내는 키워드들은 수동적으로 해당 클래스의 인스턴스로서 등록하고 포괄적인 개념의 구체적인 내용, 즉 고유 명사 등은 미리 동의어로 정의하여 동일한 클래스의 인스턴스로서 분류하도록 하였다.

예를 들어, “president of south korea” 라는 키워드가 선정되면, 이는 온톨로지 관리자가 정치 클래스의 세부 분야 클래스인 통치자 클래스의 인스턴스로 등록한다. 이때, 한국의 대통령에 대한 고유 명사들 즉, “Choi Kyu-hah”, “Park geun-hye” 등은 “president of south korea”의 실질적인 고유 명사이므로 해당하는 고유 명사들은 미리 규칙으로 정의한다. 만일, 이 중 선정된 키워드에 “Choi Kyu-hah”가 등장하였을 경우, 정의된 규칙에 따라 통치자 클래스의 인스턴스로 분류된다.

이러한 규칙 부류에 속하는 키워드들은 다음과 같다. Vietnam, USA와 같은 국가명은 정치 문서의 대외관계에서 등장하므로 세부 관심 분야인 대외 관계 클래스에 매핑하도록 미리 정의하였다. 각 나라의 지역명은 사회 문서의 지리 영역에서 등장하므로 세부 관심 분야인 지리 클래스에 매핑하도록 하였다. 다수의 스포츠인이나 음악가의 경우 미리 각 세부 분야 클래스에 매핑되도록 정의하였다. 키워드에 따라서 다중의 클래스에 속할 경우도 있었으나, 원칙적으로 대표적인 한 개의 클래스에만 속할 수 있도록 정의하였다.

관심도 지수는 크게 대분류 관심도 지수와 세부 관심도 지수로 나뉜다. 대분류 관심도 지수는 정치, 경제, 사회, 문화의 대분류 관심 분야의 관심도를 측정하며, 세부 관심도 지수는 세부 관심 분야에 대한 관심도를 측정한다. 먼저, 각 세부 분야 클래스의 *DOI:Value*들의 합으로서 세부 분야의 관심도 지수를 측정한다. 대분류 관심도 지수는 대분류 관심 분야 클래스의 하위 클래스인 세부 관심 분야 클래스들의 관심도의 합으로 측정된다.

IV. 온톨로지 기반 관심도 분석 결과

본 장에서는 각 국가와 분야별 관심도가 서로 연관성이 있는지 알아보기 위해서 카이 제곱 독립성 검정을 실시한 결과를 자세히 살펴본다. 제 1절에서는 국가별 관심도 분석을 위한 실험 방법을 구체적으로 서술하고, 제 2절에서는 국가별 대분류 관심 분야 분석 결과를 보여준다. 마지막으로 제 3절에서는 국가별 세부 관심 분야의 분석 결과를 제시한다.

1. 실험 방법

개발된 관심도 온톨로지를 기반으로 국가별 관심도 지수를 측정하기 위해서 영문 위키피디아로부터 한국, 중국, 그리고 일본의 분야별 관심도 지수를 측정하였다. 먼저 위키피디아에서 키워드를 추출하기 위해 편집 횟수가 상위에 랭크된 20개

의 키워드를 선정하고 다시 각 20개의 키워드를 제목으로 하는 기사내용 중에서 편집 횟수가 상위에 랭크된 20개의 키워드를 수집하였다. 자주 편집된 키워드를 추출하기 위해서 Condor 툴[15]을 사용하였으며, 국가별로 편집 횟수가 상위에 랭크된 400개의 키워드들을 DOI 온톨로지의 해당 클래스로 매핑하였다. 또한, 각 인스턴스의 속성으로 각 키워드들에 대한 페이지뷰를 수집하여 3개국 약 1,200여개의 키워드에 대해 분석을 실시하였다.

자료 수집은 영문 위키피디아에서 2013년 5월 1일부터 2013년 5월 31일까지 한 달간 편집된 키워드들과 선정된 키워드들에 대한 페이지뷰를 수집하였다. 상위 편집 횟수를 기반으로 키워드를 추출할 경우 키워드 페이지가 양방향으로 연결된 페이지들만 추출하여 상호간의 연관성이 높은 키워드들만 선정하였다. 관심도 지수는 관심 분야별로 속하는 키워드에 대한 페이지 뷰의 합으로 계산하였다.

2. 국가별 대분류 관심 분야 분석 결과

표 1은 한국, 중국, 그리고 일본의 대분류 관심 분야별 페이지 뷰 분석결과를 나타낸 것이다. 각 표에서 첫 번째 행은 페이지 뷰, 두 번째 행은 각 범주별 열기준 비율(column percentage), 그리고 세 번째 행은 카이제곱 독립성모형의 피어슨 잔차(Pearson residual)를 나타낸다. 피어슨 잔차는 추정된 기댓값과 실제 관측값의 차이를 기댓값의 표준오차로 나눈 것으로 기댓값보다 페이지뷰가 많으면 양의 값, 기댓값보다 페이지뷰가 적으면 음의 값을 가진다.

한국의 경우 위키피디아 사용자들은 정치 분야(48.3%)에 가장 많은 관심도를 나타냈으며, 다음으로 사회(31.9%), 문화(18.5%), 그리고 경제(1.3%)의 순으로 관심도를 나타냈다. 중국의 경우는 사회(52.6%), 정치(37.5%), 문화(6.8%), 그리고 경제(2.7%)의 순으로 나타났다. 일본은 한국 및 중국과 다른 경향을 보이는데 사회(76.4%) 분야에 가장 큰 관심을 보이고, 다음으로 문화(14.6%), 정치(8%), 그리고 경제(1%)의 순으로 나타났다.

국가별로 관심 분야의 페이지 뷰가 독립성에서 벗어난 정도를 구체적으로 알아보기 위한 잔차 그래프가 그림 3이다. 그림 3에서 순서대로 1은 정치, 2는 경제, 3은 사회, 그리고 4는 문화를 나타낸다. 세로축 좌표는 잔차값을 나타내며 양의 값은 기대보다 관심도가 높음을 나타내고, 음의 값은 기대보다 관심도가 낮음을 나타낸다.

표 1. 관심 분야별 관심도 지수, 열 비율, 피어슨 잔차
Table 1. DOI index, column percentage, and Pearson residual by interesting fields

	한국	중국	일본
정치	9,224,655	10,038,803	1,252,664
	(0.483)	(0.375)	(0.080)
	1675.16889	616.20242	-2475.63249
경제	243,379	727,183	158,361
	(0.013)	(0.027)	(0.010)
	-218.81619	454.11330	-283.94425
사회	6,088,724	14,156,884	12,038,181
	(0.319)	(0.529)	(0.764)
	-2161.31797	70.91621	2210.35258
문화	3,533,699	1,823,128	2,306,112
	(0.185)	(0.068)	(0.146)
	967.13061	-1171.80753	306.07334
계	19,157,103 (1.000)	37,487,384 (1.000)	9,841,771 (1.000)

X-squared = 9253949, DF=6, p-value < 2.2e-16

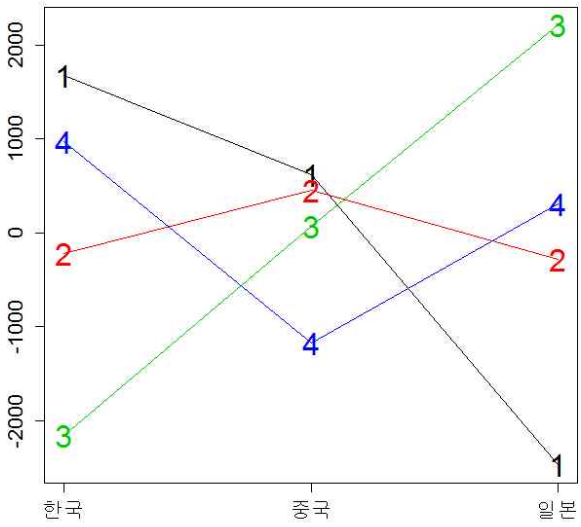

그림 3. 관심 분야별 잔차 그래프
(1=정치, 2=경제, 3=사회, 4=문화)
Fig. 3 Residual Graph by Interesting Fields of DOI
(1=politics, 2=economy, 3=society, 4=culture)

국가별로는 한국의 경우 정치와 문화 분야가 기대보다 많은 관심도를 보이고 있으며 사회 분야에 대한 관심도가 현저히 적게 나타났다. 중국의 경우 정치, 경제, 그리고 사회가 기대보다 많은 관심도를 보이고 있으며 문화 분야가 기대보다 적은 관심도를 보인다. 일본의 경우 사회와 문화가 기대보다

많은 관심도를 보이고 정치와 경제가 기대보다 적은 관심도를 보이고 있다.특히 정치 분야에 대한 관심도가 현저히 낮음을 알 수 있다.

3. 국가별 세부 관심 분야 분석 결과

본 절에서는 정치, 경제, 사회, 그리고 문화의 각 세부 관심 분야가 국가와 연관성이 있는지 알아보기 위해서 카이 제곱 검정을 실시한 결과를 나타낸다.

3.1 정치 분야 분석 결과

정치 분야의 세부 관심 분야는 행정, 대외관계, 정부, 군사, 그리고 통치자로 나뉜다. 표2는 각 국의 정치 분야의 세부 분야에 대한 페이지 뷰와 열기준 비율, 그리고 피어슨 잔차를 나타내며, 그림 4는 잔차 그래프를 나타낸다. 그림 4에서 순서대로 1은 행정, 2는 대외관계, 3은 정부, 4는 군사, 그리고 5는 통치자를 나타낸다.

한국의 세부 관심 분야는 대외관계(81.6%), 통치자(17.9%), 행정(0.3%), 군사(0.1%), 정부(0.1%)의 순이다. 중국은 대외관계(54%),통치자(31.3%), 군사(8.1%), 행정(5.1%),정부(1.4%)순이고, 일본은 대외관계(58.4%), 통치자(26%), 군사(15.3%), 정부(0.3%),행정(0%)순이다. 세 국가 모두 대외관계, 통치자에 많은 관심을 보이고 있다.

표 2. 정치 분야 관심도 지수, 열 비율, 피어슨 잔차
Table 2. DOI index, column percentage, and Pearson residual by politics

	한국	중국	일본
행정	27577	411128	0
	(0.003)	(0.051)	(0.000)
	-582.645	678.964	-180.423
대외관계	7524832	4355824	731283
	(0.816)	(0.540)	(0.584)
	1245.660	-1135.116	-239.390
정부	10901	116489	4263
	(0.001)	(0.014)	(0.003)
	-301.983	330.397	-51.029
군사	12851	655735	191974
	(0.001)	(0.081)	(0.153)
	-917.361	627.493	588.070
통치자	1648494	2526257	325144
	(0.179)	(0.313)	(0.260)
	-639.202	621.536	45.688
계	9224635	8066433	1252664
	(1.000)	(1.000)	(1.000)

X-squared = 2304228, DF=8, p-value < 2.2e-16

좀 더 구체적인 카이제곱 검정 결과는 그림 4에서 보인다. 한국은 대외관계에 상대적으로 많은 관심을 보이는 반면, 중국은 대외관계에 대한 관심도가 상대적으로 낮고, 나머지 분야의 관심도가 높은 것으로 나타났다. 일본은 군사에 대한 관심도가 상대적으로 높은 것으로 나타났다.

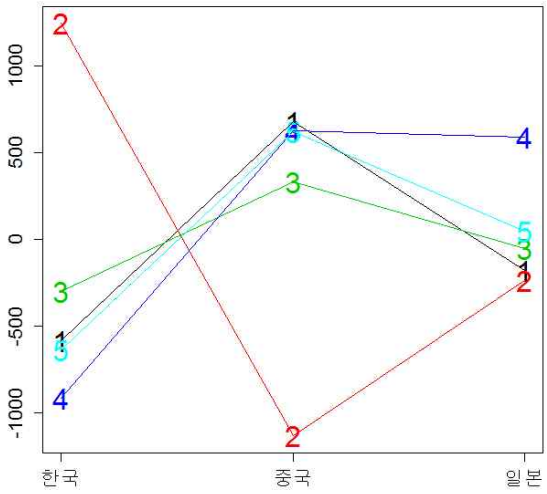

그림 4. 정치 분야의 세부 분야별 관심도 잔차 그래프
(1=행정, 2=대외관계, 3=정부, 4=군사, 5=통치자)
Fig. 4. Residual Graph by Detailed Fields of Politics
(1=administration, 2=foreign relation, 3=government, 4=military, 5=ruler)

3.2 경제 분야 분석 결과

경제 분야는 세부 관심 분야로 재정, 산업, 인프라, 과학기술, 그리고 복지로 나뉜다. 표 3은 삼국의 페이지 뷰, 국가별 각 분야가 경제에서 차지하는 비율, 그리고 잔차를 나타낸다. 그림 5는 경제 분야의 세부 분야별 관심도 잔차 그래프이다. 순서대로 1은 재정, 2는 산업, 3은 인프라, 4는 과학기술, 그리고 5는 복지를 나타낸다.

한국의 세부 관심 분야는 재정(86.5%), 복지(13.2%), 산업(0.2%), 인프라(0.1%), 과학기술(0.0%)의 순이다. 중국은 재정(53%), 산업(27.2%), 인프라(11.1%), 과학기술(8.8%), 복지(0.0%)순이고, 일본은 인프라(89.2%), 재정(10.8%), 그리고 산업 (0.0%), 복지(0.0%), 과학기술(0.0%) 순이다.

경제 분야의 경우 기본적으로 대분류 중 가장 관심도 지수가 적은 영역이라 분석에 있어 상위에 편집된 키워드가 세부 분야에 한 번도 나타나지 않은 경우가 있었다. 한국의 경우 과학기술 관련 키워드가 등장하지 않았고, 중국의 경우 복지와 관련된 키워드가 등장하지 않았으며, 일본의 경우, 산업, 복지,

그리고 과학기술 분야와 관련된 키워드가 한 번도 등장하지 않았다. 이는 삼 국 모두를 고려하였을 때 과학기술, 복지 분야에 대한 관심도가 매우 적음을 나타내며 시사하는 바가 크다.

표 3. 경제 분야의 관심도 지수,열 비율, 피어슨 잔차
Table 3. DOI index, column percentage, and Pearson residual by economy

	한국	중국	일본
재정	154482	230749	17162
	(0.865)	(0.530)	(0.108)
	332.055	17.871	-368.685
산업	363	118301	0
	(0.002)	(0.272)	(0.000)
	-202.609	327.064	-190.171
인프라	139	48251	141199
	(0.001)	(0.111)	(0.892)
	-274.002	-312.609	670.096
과학기술	0	38104	0
	(0.000)	(0.088)	(0.000)
	-109.782	176.107	-101.679
복지	23556	0	0
	(0.132)	(0.000)	(0.000)
	284.261	-177.204	-79.166
계	178540	435405	158361
	(1.000)	(1.000)	(1.000)

X-squared = 635830.2, DF=8, p-value 〈2.2e-16

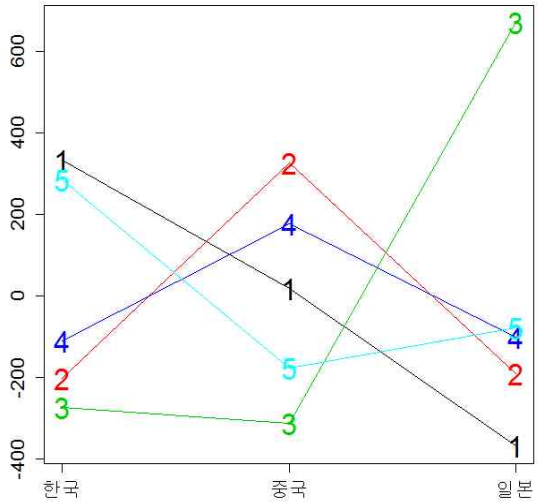

그림 5. 경제 분야의 세부 분야별 관심도 잔차 그래프
(1=재정, 2=산업, 3=인프라, 4=과학기술, 5=복지)
Fig. 5. Residual Graph by Detailed Fields of Economy
(1=finance, 2=industry, 3=infrastructure, 4=science&technology, 5=welfare)

보다 구체적인 국가와 경제 분야 세부 분야와의 연관성은 그림 5에서 보인다. 한국의 경우 복지와 재정 분야에 상대적으로 많은 관심을 보인 반면, 중국은 산업과 과학기술 방면에 관심도가 많고, 재정과 인프라 분야는 낮은 것으로 보인다. 일본은 인프라 분야에 상당히 높은 관심도를 보이고 재정 분야에는 상대적으로 관심이 적은 것으로 나타났다.

3.3 사회 분야 분석 결과

사회 분야는 세부 분야로 인구, 교육, 지리, 역사, 언어, 민족, 그리고 종교로 나뉜다. 표 4는 삼국의 페이지뷰, 사회 분야의 세부 관심 분야에서 차지하는 비율, 그리고 잔차를 나타낸다.

한국의 경우, 인구와 종교를 제외하고 역사(67.6%), 지리(17.3%), 언어(9.6%), 교육(3.8%), 민족(1.7%) 순으로 관심도를 갖는다. 중국의 경우는 역사(67.7%), 언어(18.3%), 지리(8.6%), 민족(5%), 종교(0.3%) 그리고 교육(0.1%)의 순으로 관심도를 보인다. 마지막으로 일본의 경우는 역사(84.5%), 지리(9.9%), 언어(2.8%), 민족(1.7%), 인구(0.6%), 그리고 교육(0.5%)의 순으로 관심도를 보인다.

그림 6은 사회 분야의 세부 분야별 관심도 잔차 그래프이다. 한국의 경우 교육과 지리 분야가 많고 역사, 언어, 그리고 민족 분야에 대한 관심도가 적음을 알 수 있다. 중국의 경우 지리와 역사 분야에 대한 관심도가 적고 언어와 민족에 대한 관심도가 많음을 알 수 있다. 마지막으로 일본의 경우는 지리, 언어, 민족에 대한 관심도가 적은 반면 역사에 대한 관심도가 상대적으로 많음을 알 수 있다.

표 4. 사회 분야의 DOI index, 열 비율, 피어슨 잔차
Table 4. DOI index, column percentage, and Pearson residual by Society

	한국	중국	일본
인구	0	0	74931
	(0.000)	(0.000)	(0.006)
	-132.077	-241.996	355.015
교육	230024	10644	63964
	(0.038)	(0.001)	(0.005)
	803.416	-450.637	-187.365
지리	1051471	1213359	1186280
	(0.173)	(0.086)	(0.099)
	584.048	-342.866	-120.540

역사	4105503	9564405	10175234
	(0.676)	(0.677)	(0.845)
	-399.134	-713.883	1054.976
언어	584650	2589827	331659
	(0.096)	(0.183)	(0.028)
	-110.227	1201.426	-1143.224
민족	105057	708698	206113
	(0.017)	(0.050)	(0.017)
	-224.288	531.112	-363.435
종교	0	37806	0
	(0.000)	(0.003)	(0.000)
	-93.762	220.324	-150.184
계	6076705	14124739	12038181
	(1.000)	(1.000)	(1.000)

X-squared = 3166677, DF=12, p-value <2.2e-16

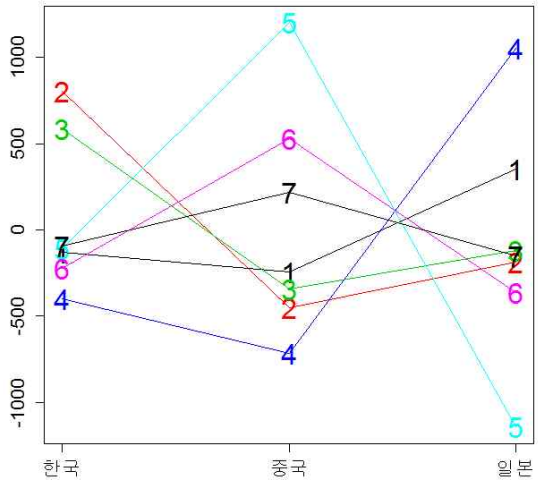

그림 6. 사회 분야의 세부 분야별 관심도 잔차 그래프 (1=인구, 2=교육, 3=지리, 4=역사, 5=언어, 6=민족, 7=종교)
Fig. 6. Residual Graph by Detailed Fields of Culture (1=demography, 2=education, 3=geography, 4=history, 5=language, 6=people, 7=religion)

3.4 문화 분야 분석 결과

문화 분야에 대한 세부 분야는 예술, 음식, 오락, 필름, 음악, 그리고 스포츠로 분류된다.

표 5. 문화 분야의 DOI index, 열 비율, 피어슨 잔차
Table 5. DOI index, column percentage, and Pearson residual by Culture

	한국	중국	일본
예술	32966	27365	0
	(0.009)	(0.016)	(0.000)
	36.431	130.369	-160.106
음식	42667	427723	0
	(0.012)	(0.249)	(0.000)
	-542.163	1140.990	-460.038
오락	293757	0	470835
	(0.084)	(0.000)	(0.214)
	-163.593	-505.750	645.449
필름	117673	513801	153807
	(0.033)	(0.299)	(0.070)
	-605.904	941.665	-206.457
음악	2653131	10046	203304
	(0.755)	(0.006)	(0.092)
	1959.545	-1165.408	-1066.849
스포츠	372920	736783	1375692
	(0.106)	(0.429)	(0.624)
	-1248.734	300.855	1087.481
계	3513114	1715718	2203638
	(1.000)	(1.000)	(1.000)

X-squared = 6041213, DF=10, p-value <2.2e-16

표 5는 삼국의 문화 분야에 대한 세부 관심도 지수를 나타낸다. 한국은 음악(75.5%), 스포츠(10.6%), 오락(8.4%), 필름(3.3%), 음식(1.2%), 예술(0.9%)순으로 나타났고, 중국은 스포츠(42.9%), 필름(29.9%), 음식(24.9%), 예술(1.6%), 음악(0.6%), 오락(0%) 순으로 나타났다. 일본은 스포츠(62.4%), 오락(21.4%), 음악(9.2%), 필름(7%), 음식(0%), 예술(0%)로 나타났다.

그림 7은 문화 분야의 세부 분야별 관심도 잔차 그래프이다. 한국의 경우 음악 분야에 대한 관심도가 많고 나머지 분야는 기대보다 적은 관심도를 나타낸다. 중국의 경우는 오락과 음악 분야에 대한 관심도가 적고 나머지 부분은 관심도가 많음을 나타냈다. 일본의 경우는 예술, 음식, 필름, 그리고 음악에 대한 관심도가 기대보다 적고 오락과 스포츠에 대한 관심도가 많음을 알 수 있다.

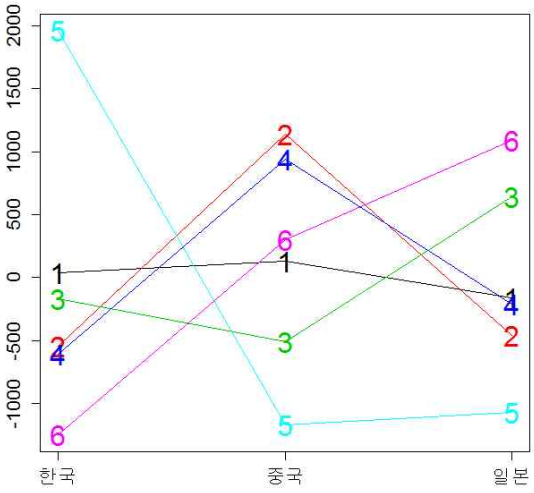

그림 7. 문화 분야의 세부 분야별 관심도 잔차 그래프 (1=예술, 2=음식, 3=오락, 4=필름, 5=음악, 6=스포츠)
Fig. 7. Residual Graph by Detailed Fields of Culture (1=art, 2=cuisine, 3=entertainment, 4=film, 5=music, 6=sports)

V. 결론 및 향후 연구

본 논문에서는 위키피디아 사용 데이터를 수집하여 한국, 중국, 그리고 일본의 분야별 관심도를 측정 및 분석하였다. 먼저, 국가의 분야별 관심도를 파악하기 위해서 관심도 온톨로지(DOI ontology)를 개발하였다. 관심도 온톨로지는 정치, 경제, 사회, 그리고 문화의 대분류 관심 분야 클래스와 각 대분류 관심 분야 클래스에 해당하는 세부 관심 분야 클래스를 갖는다. 다음으로 국가별로 자주 편집된 키워드 약 400개를 해당 클래스에 매핑하여 인스턴스로 등록하였다. 각 인스턴스에 대해 사용자들의 페이지 뷰를 수집하여 페이지 뷰의 합으로서 세부 분야별 및 대분야별 관심도 지수를 산출하였다.

먼저, 각 국가별로 관심 분야와 연관성이 있는지 알아보기 위해서 카이제곱 검정을 실시한 결과, 국가는 관심 분야와 밀접한 연관이 있는 것으로 나타났다. 한국과 중국의 경우, 정치 분야에 대한 관심도가 높은데 비해, 일본의 경우는 사회 분야에 대한 관심도가 다른 두 나라의 두 배 가까이 높은 점수를 보이고 있다. 또한 일본의 경우는 문화 분야보다도 적은 10% 안팎의 정치 분야에 대한 관심도를 가지고 있다.

한국, 중국, 그리고 일본 중에서 문화 분야의 관심도에서 가장 높은 점수를 받은 나라는 한국이며 특히 세부 항목 중에서 음악이 대부분을 차지하는 것으로 보아 최근 K-팝의 영향이 관심도 중 문화 분야에 영향을 미쳤을 것으로 예측된다.

과학 기술 분야는 대분류로 처리하고자 하였으나 관심도가 거의 나타나지 않아 사회 분야의 세부 분야로 처리하였다. 이는 독일이나 스위스, 영국 등의 유럽 국가와 차이를 보이는 것으로 과학 기술 분야의 키워드가 거의 등장하지 않았다는 것은 주목할 만하다. 따라서 이 분야에 있어서 계속적인 관심과 국가 차원의 위키피디아 관리가 필요할 것으로 보인다.

향후 다양한 국가들에 대한 관심도 측정을 할 수 있도록 일반적인 프레임워크를 개발할 예정이다. 개발된 온톨로지를 바탕으로 온톨로지 학습 모듈을 추가하여 다양한 국가를 고려할 수 있도록 할 예정이다.

참고문헌

[1] R. Pappu, P. Quester, and G. R. W. Cooksey, "Country Image and Consumer-Based Brand Equity: Relationships and Implications for International Marketing." Journal of International Business Studies, Vol. 38, No. 5, pp. 726-745, June, 2007.

[2] S. Anholt, "Anholt Nation Brands Index: How Does the World See America?" Journal of Advertising Research, Vol. 45, No. 3, pp. 296-304, sept. 2005.

[3] M. Ghiassi, J. Skinner and D. Zimbra, "Twitter Brand Sentiment Analysis: A Hybrid System Using n-gram Analysis and Dynamic Artificial Neural Network", Expert Systems with Applications, Vol. 40, pp. 6266-2686, Nov. 2013.

[4] M. M. Mostafa, "More than Words: Social Networks' Text Mining for Consumer Brand Sentiments", Expert Systems with Applications, Vol. 40, pp. 4241-4251, Aug. 2013.

[5] Wikipedia, http://en.wikipedia.org/

[6] D. A. Aacker, "Managing Brand Equity", The Free Press: New York, Sept. 1991.

[7] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, Vol. 5, No. 2, pp. 199-220, June, 1993.

[8] H. H. Kim, "A Tag-based Music Recommendation Using UniTag Ontology", Journal of the Korea Society of Computer and Information, Vol. 17,

No. 11, pp. 133-140, Nov. 2012.
[9] W. Wong, W. Liu and M. Bennamoun, "Ontology Learning from Text: A Look Back and into the Future", ACM Computing Surveys, Vol. 44, pp.1-35, Aug. 2012.
[10] F. M. Suchanek, G. Kasneci and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia", In Proc. of World Wide Web Conference, pp. 697-706, Alberta, Canada, May, 2007.
[11] M. Morsey et al., "DBpedia and the Live Extraction of Structured Data from Wikipedia", Electronic Library and Information Systems, Vol. 46, No. 2, pp. 157-181, June, 2012.
[12] S. P. Ponzetto and M. Strube, "Knowledge Derived From Wikipedia For Computing Semantic Relatedness", Journal of Artificial Intelligence Research, Vol. 30, pp. 181-212, Oct. 2007.
[13] H. S. Moat et al., "Quantifying Wikipedia Usage Patterns Before Stock Market Moves", Scientific Reports, Vol. 3, No. 1801, pp. 1-5, May, 2013.
[14] P. Kotler and D. Getner, "Country as brand, product, and beyond: A Place Marketing and Brand Management Perspective", Journal of Brand Management, Vol. 9, No. 4, pp. 249-261, Jan. 2002.
[15] Condor, http://www.ickn.org/condor.html

저 자 소 개

김 현 희

1996: 이화여자대학교 전자계산학과 공학사.
1998: 이화여자대학교 컴퓨터공학과 공학석사.
2005: 이화여자대학교 컴퓨터공학과 공학박사.
현 재: 동덕여자대학교 정보통계학과 조교수
관심분야: 추천 시스템, 온톨로지 빅데이터 분석
Email : heekim@dongduk.ac.kr

조 진 남

1980: 연세대학교 응용통계학과 경제학사.
1982: 연세대학교 응용통계학과 경제학 석사.
1992: Virginia Tech 통계학 박사
현 재: 동덕여자대학교 정보통계학과 교수
관심분야: 실험계획법, 표본조사
Email : jinnam@dongduk.ac.kr

김 동 건

1986: 연세대학교 경영학과 학사.
1990: Virginia Tech 통계학 석사.
1995: Virginia Tech 통계학 박사.
현 재: 동덕여자대학교 정보통계학과 교수
관심분야: 데이터마이닝, 통계계산
Email : dongg@dongduk.ac.kr