

국가이미지 분석을 위한 위키피디아 실시간 동적 온톨로지 구축 알고리즘 및 적용

이영환¹

¹ 건국대학교 기술경영학과

접수 2016년 5월 10일, 수정 2016년 7월 2일, 게재확정 2016년 7월 22일

요약

이 연구에서는 국가이미지를 실시간 측정하기 위하여 위키피디아의 키워드를 추출하여 반자동 동적 온톨로지를 구축하는 알고리즘을 개발하였다. 이 DCDKC (double-crossing double keyword collection)라 명명된 알고리즘에 의해 추출된 국가이미지 온톨로지인 위키온토 (WikiOnto)는 동적으로 변화하는 국가이미지를 실시간으로 자동 측정하여 제공할 수 있도록 하였다. 따라서 국가이미지 관리에 효율성을 기하고 뜻밖에 있을 지도 모르는 국가이미지의 급박한 변화에 선제적으로 대응할 수 있는 도구를 개발하고자 하였다. DCDKC 알고리즘은 과거 오프라인 설문조사에 상당부분 의존하던 각종 국가브랜드지수 모델과는 달리 온라인 실시간 동적 국가이미지 모니터링 시스템에 잘 적용될 수 있음을 보여주었다. DCDKC 알고리즘을 적용한 시스템의 유용성을 검증하기 위하여 아시아의 3대 수출국인 한국, 일본, 중국의 국가이미지를 측정하기 위한 동적 온톨로지 위키온토를 구축하고 국가 이미지의 동적 변화를 실시간으로 추적가능함을 확인하였다. 결론적으로 DCDKC를 이용한 위키온토의 구축을 통한 국가이미지 실시간 관리체계에 대한 가능성을 확인하였을 뿐더러 저렴하고 탁월한 솔루션이 될 수 있음을 증명하였다.

주요용어: 국가이미지, 온톨로지, 키워드추출, 위키피디아, 국가브랜드

1. 서론

전 세계 국가들의 무한 경쟁 속에서 자국의 브랜드 이미지와 글로벌 경쟁력을 강화시키는 것은 매우 중요하다. 국가의 경쟁력이 곧 기업의 경쟁력과 직결되기 때문이다. 이에 이명박 정부의 국가브랜드 위원회 발족으로부터 현 박근혜 정부의 창조경제에 이르기까지 우리나라는 국가 브랜드 및 이미지 역량강화를 목표로 하는 정책을 지속적으로 수행해 왔다. 그러나 한국은 2014-2015 WEF 보고서에 의하면 국가경쟁력이 26위로 싱가포르, 일본, 홍콩, 타이완, 말레이시아 보다 뒤떨어져 있고 2007년 11위를 기록한 이래 지속적으로 하락하여 역대 최하위 순위를 기록하고 있다 (<http://reports.weforum.org/global-competitiveness-report-2014-2015/rankings>).

이러한 필요를 인정한 연구자들은 빅데이터와 인터넷의 발달을 이용하여 최근 인터넷의 발달로 소셜 네트워크서비스 (SNS)의 활용이 증가되면서 소비자 간의 교류가 인터넷 상에서 직접적으로 이루어지는 것을 이용하여 국가이미지를 자동으로 측정하고 변화를 분석하는 시스템을 연구하고 있다. (Ahn과 Whang, 2012) 특히 온라인상에서 각 언어권의 사용자들이 관심이 있는 국가와 관련된 단어를 언급하고, 클릭하고, 리트윗하고, '좋아요' 버튼을 누르고, 공유하는 등의 행위 데이터를 수집하고 분석하여 특

¹ (05029) 서울시 광진구 능동로 120, 건국대학교 경영대학 기술경영학과, 조교수.
E-mail: nicklee@konkuk.ac.kr

정 국가에 대한 인식과 변화의 과정을 파악할 수 있는 모델을 개발하고 시범·적용하여 유용성을 확인하는 연구가 진행되고 있다 (Lee 등, 2015) (Chun, 2015) (Song 등, 2015).

이런 국가이미지를 연구하는 연구팀들은 다음과 같은 문제에 부딪히게 된다.

1) 국가이미지 측정의 객관성 확보: 인터넷에서 교환되는 단어 혹은 개념 중에서 어떤 단어가 특정 국가의 이미지에 영향을 미치는 것인지 결정하는 것 자체가 지극히 주관적이어서 결과물인 국가이미지 측정의 객관성을 담보하기가 어렵다 (특정인의 이름이나 특정 사물 등을 국가이미지와 연결시켜서 측정해야 할 것인가 하는 점은 어렵고 주관적일 수 있다. 예컨대, 한 국가의 현직 대통령의 이름과 관련된 단어를 국가이미지에 반영하는 것은 두말할 나위 없이 당연하다고 말할 수 있다. 그러나 현직대통령의 친척은 어떻게 해야 할지 하는 점은 매우 주관적인 판단이 필요하다). 그럼에도 불구하고 국가이미지 측정의 객관성을 확보하는 것이 필요하다.

2) 국가이미지의 동적 변화 대응: 시시때때로 변화하는 국가이미지는 측정 단어나 개념의 동적인 변화를 요구한다 (한국의 국가이미지에 영향을 주는 단어가 시간에 따라 변화한다는 것은 두말할 나위가 없다. 예컨대, 세월호 참사 이전에는 어느 누구도 “세월호” 혹은 “울돌목”이 한국의 국가이미지와 관련된 단어라고 생각하지 않을 터이나 세월호 참사 이후에는 확실히 “세월호”는 국가이미지와 관련되게 되었다). 따라서 국가이미지의 변화에 따라서 단어와 개념의 변화를 반영하여 측정하는 것이 필요하다.

3) 국가이미지의 실시간 측정: 국가이미지의 실시간 측정은 여러 가지 예측하지 못한 사고 등에 대해 등에 실시간으로 대응할 수 있는 능력이 기반이 된다는 점에서 매우 중요하다.

본 연구에서는 이 논문에서는 위의 세 가지 요건을 충족하면서 동시에 국가이미지 측정의 객관성을 확보하고 동적으로 자동 추출할 수 있는 온톨로지 (ontology)를 구축하는 알고리즘을 개발하고자 한다.

국가 이미지 온톨로지를 구축함에 있어 본 연구는 특정 국가와 관련된 단어를 최대한 수집하면서 관련 국가이미지 연관성을 고려하고 객관적으로 적절할 수 있는 도출할 수 있는 방안을 제시하고자 한다. 특히 여기서 추출된 국가이미지는 타 국가의 국가이미지와 비교 가능할 수 있도록 추출해 내는 것이 필요하다.

이를 위하여 본 연구는 국가이미지의 동적 변화에 대응할 수 있고 실시간 측정이 가능한 알고리즘인 DCDKC (Double-Cross Double-Keyword Collection)를 고안하였다. 이의 유용성을 검증하기 위하여 한중일 삼국의 동적 국가이미지 온톨로지인 위니온토 (WiniOnto)를 실시간 (near real-time)으로 추출하였다. 위키온토는 8641개라는 방대한 양의 키워드로 이루어진 한중일 국가이미지 측정을 위한 온톨로지이다. 마지막으로 본 연구에서는 이를 이용하여 위키피디아에 나타난 한중일 삼국의 국가이미지를 실시간 측정하고 비교하였다.

2. 국가이미지 측정 모델

2.1. 국가 이미지

ICT의 발전으로 시공간 제약의 극복과 자본의 유동성 증가로 기업, 자본, 그리고 관광객을 유치하기 위해 국가 간의 경쟁이 심화되면서 국가이미지의 중요성이 지속적으로 증가하고 있다 (Bradley 등, 2002; Choi, 2008). 따라서 국가에 있어 이미지는 매우 중요한 자산으로 인식되고 있고 국가이미지를 측정하려는 노력이 증가하면서 다양한 방법론이 적용되고 있다. 그러나 국가이미지는 각 개인이 가지고 있는 주관적인 경험이나 정보가 종합되어 심상이 형성된다는 점에서 이미지를 측정하는 것은 어렵다. 국가이미지는 개인의 관점으로부터 시작됨에 따라 방법론에 차이가 있지만 측정방법은 일반적으로 설문 조사 혹은 전문가 인터뷰를 통해 연구 되고 있다.

그런데 최근 빅데이터 기술이 발전되고 SNS 데이터와 같은 비정형 데이터 처리 기술이 발전됨에 따라 국가이미지 연구에 적용하려는 노력이 증가하고 있지만 (Cha 등, 2013), 데이터의 편의성에 따른 한계점이 있다.

2.2. 국가이미지 측정 방법 및 선행연구

국가이미지는 주관적인 심상의 형성이 개인마다 다르고 복잡하기 때문에 형성 경로 및 요인은 중요하다. 그러므로 형성 경로 및 요인에 따른 구체적인 국가이미지의 세부 요소를 정의할 필요성이 있다.

심상의 이미지를 형성하는 요인은 관계적 경로, 텍스트 경로, 그리고 경험적 경로 크게 세 요인으로 볼 수 있다 (Moffitt, 1994). 현재까지 국가이미지 측정방법은 텍스트 경로와 경험적 경로를 기반으로 설문조사에 치중되어 있다. 그러나 SNS의 데이터는 두 요인뿐만 아니라 관계적 경로까지 포함할 수 있다는 점에서 다 방면의 국가이미지를 측정할 수 있다.

형성 경로뿐만 아니라 형성 요인에 있어 다수의 연구가 존재하며 이를 종합하면, 국가이미지를 이루는 형성 요인은 연구자의 주관에 따라 민주화 혹은 안정과 같은 정치환경 그리고 공공질서 및 복지과 같은 사회적 환경, 물가와 성장 측면의 경제적 측면 그리고 예술, 여가와 같은 문화적 요소 그리고 제품이나 혁신과 같은 마케팅적 요소이다 (Wang, 1978; Roth와 Romeo, 1992; Martin과 Eroglu, 1993; Shin, 1993). 결국, 국가이미지 형성요인의 경우 주관적 견해에 영향을 받으나 공통적인 요인을 추출해 통합하면 정치, 사회, 경제, 문화 네 요인으로 볼 수 있다.

국가이미지의 측정방법은 연구자들의 주관적인 형성요인을 기반으로 일반적으로 설문조사방법을 통해 분석했다. 국가이미지 연구 중 해외 연구는 대표적으로 Anholt의 NBI (Nation Brand Index), Anholt의 CBI (City Brands Index), Futurebrand의 CBI(Country Brand Index), 그리고 Anholt의 GCI (Good Country Index)가 있다. 또한, 국내연구로는 SERI (삼성경제연구소)와 PCNB(국가브랜드위원회)의 NBDO(Nation Brand Dual Octagon)과 산업정책연구원의 국가 및 산업 브랜드 맵이 있다.

Simon Anholt는 리서치 전문업체인 GfK Roper와 협업해 국가 그리고 세계 유명 도시들의 평판과 이미지를 측정하기 위한 방법을 2005년, 2006년 각각 개발했다. 그것이 Anholt-GfK Roper NBI와 CBI이다. 이 모형들은 20개국의 2만 명 패널을 대상으로 50개국에 대한 국가이미지와 도시이미지를 온라인 설문조사를 통해 연구되고 있다. 측정 요인은 NBI의 경우 해당 국가의 관광, 수출, 정부, 투자와 이민, 문화유산, 국민과 같이 6개의 영역이며 CBI의 경우 존재감, 잠재력, 시민, 생활기반 시설, 도시의 활기, 장소적 측면 총 6요인으로 나누어진다 (Anholt, 2005; 2006). CBI는 각각 6요인으로 구성되어 있어 Hexagon 모델이라 한다.

FutureBrand는 글로벌 브랜드 컨설팅 기업이며 2010년부터 매년 지속적으로 국가의 브랜드 가치를 측정하기 위해 CBI를 개발했다. 조사대상은 2700명 이상의 국제 관광객으로서 CBI는 관광에 대한 국가경쟁력을 연구하고 있다. FutureBrand의 CBI는 최종적으로 국가의 순위를 구하기 위해 통계자료와 설문조사의 정량적 기법과 전문가 조사와 같은 정성적 기법의 3단계의 과정을 거친다. 측정요인은 인지도, 친숙도, 연상, 선호도, 여행 고려, 여행 선택 및 방문, 옹호와 같은 7개 요인을 전 과정의 기반으로 한다 (Futurebrand, 2010).

2014년 Simon Anholt는 기존 NBI 그리고 CBI와 데이터 및 개념의 측면에서 차별성이 있는 GCI를 개발했다 (<http://www.goodcountry.org/overall>). GCI는 UN과 기타 국제기구의 데이터를 활용하며 국가가 인류 공익측면에서 기여도를 측정해 지수화 시켜 순위 정하는 것을 목표로 한다. 연구결과는 보고서 혹은 논문의 형태로 존재하지 않으며 인터넷 사이트를 통해 공개되어 있다. 대다수의 측정은 데이터를 통해 정량적 기법을 사용하고 있으며 세부요인은 크게 과학 및 기술, 문화, 국제적 평화 및 보안, 세계 질서, 지구 및 기후, 번영 및 평등, 건강 및 웰빙 총 7가지이다. GCI 모델의 특징은 기존의 델파

이 기법과 같은 정성적인 방법 혹은 설문조사에 의존한 정량적 방법이 아닌 국가의 객관적 통계 데이터를 기반으로 한 정량적 분석이라 할 수 있다.

국내의 국가이미지 분석연구로는 대표적으로 삼성경제연구소와 국가브랜드위원회에서 공동으로 개발한 국가브랜드지수 모델 (SERI-PCNB NBDO)이 있다 (SERI, 2012). NBDO는 국가의 실제 현황 및 위치와 국가이미지 두 요소를 고려한 모델이다. 국가의 객관적인 요인은 통계 데이터를 기반으로 하고 국가이미지는 설문조사를 이용하여 측정한다. 국가 브랜드적인 측면과 국가의 구성요소를 결합하여 SERI 자체적인 국가 브랜드 요인을 도출하였는데 이는 국부 측면의 경제와 기업, 과학 및 기술 요인과 정부 측면의 정책 및 제도, 인프라적 요소이다. 또한 문화적 측면의 문화 및 역사, 지리적 요소 그리고 국민 측면의 유명인사 및 국민으로 나눈다. 이 모델의 특징은 기존의 국가이미지의 한계점을 분석하여 이를 극복할 수 있는 종합적 지수체계를 만드는데 초점이 맞춰 있으며 설문조사와 통계 데이터를 근거로 한 정량적 조사를 통합적으로 실행했다는 특징이 있다. 그러나 50개국을 대상으로 설문조사를 시행한 점과 한국을 중심으로 국가이미지를 분석했다는 점에서 신뢰성 부분의 약점이 존재한다 (SERI, 2012).

산업정책연구원과 대한무역투자진흥공사가 주축이 되어 개발한 국가 및 산업 브랜드 맵은 국가의 이미지적인 측면보다는 국가와 그 안에 속한 산업들의 브랜드에 초점을 맞춰 제품 및 서비스를 식별해 브랜드적인 구성요소를 개발해 분석한 방법이라 할 수 있다 (산업정책연구원, 2010). 이 연구는 한국을 기준으로 국가의 브랜드 및 이미지 그리고 산업의 현황을 파악하는데 중점을 두었다. 그래서 산업마다 비교국가 상이한 특징이 있다. 따라서 본 연구에서는 Figure 2.1과 같이 각 연구자들이 제시한 분류표를 정리하고 비교하여 국가이미지를 위한 4대 분야를 도출하고 이를 바탕으로 세부분야까지 도출하고자 한다.

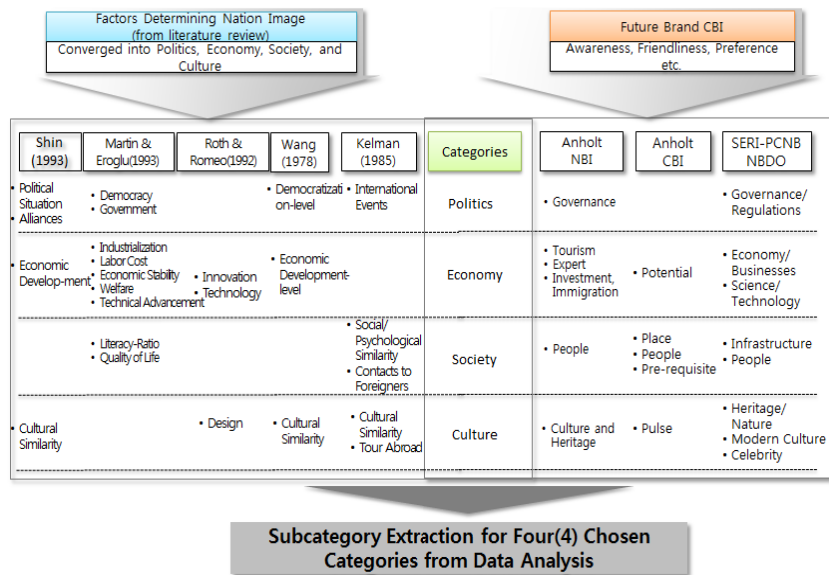


Figure 2.1 Four(4) Categories and subcategories of nation image ontology

알고리즘을 이용하여 어떤 단어의 이미지를 자동 추출하는 방법은 최근 인공지능의 자연어 처리 분야에서 딥러닝을 이용한 연구로 활발히 진행되고 있다. 특히 구글이 word2vec를 발표하고 그 프로그램을 오픈소스로 공개하면서 단어의 의미를 추출하고 이를 자연어 처리에 이용하는 연구가 다각도로 진행되는 중이다 (Tomas Mikolov 등, 2013; Xin 2014; Garten 등; Goldberg와 Levy, 2014)

본 연구와 word2vec의 연구는 다음과 같은 유사한 점이 있다.

- 단어에 대한 이미지 (혹은 의미) 추출을 자동화하려고 시도하였다.
- 단어에 대한 이미지를 이용하여 다른 단어와 측정하고 비교하는 것을 자동화하려고 하였다.

본 연구와 word2vec 커뮤니티의 연구는 다음과 같은 차별성이 있다.

- word2vec은 빅데이터 상에서 단어 간의 동시발생 (co-occurrence) 빈도에 의해 문맥상 단어의 이미지를 추출하려고 한 반면 본 연구는 단어의 의미론적 유사성을 추출하는 알고리즘을 구현 적용하여 동적인 온톨로지를 구축한 후 클릭수를 측정하여 단어의 이미지 변화를 측정하려고 시도하였다.
- word2vec은 어떤 단어사이의 이미지의 유사도를 비지도학습 등에 의해 장기간에 걸쳐 단어의 의미를 학습하려고 시도한 반면 본 연구는 단어의 시간에 따른 이미지 변화를 측정하고 모니터링하려고 시도하였다.
- 본 연구의 대상인 국가이미지는 대상이 되는 개개인의 주관적인 심상의 교집합을 측정하고 그 변화를 분석하는 반면 word2vec은 기본적으로 개인이 갖고 있는 심상의 공통부분을 추출하여 자연어로 사용하였다.

2.3. 온톨로지 풀(Pool)로서 위키피디아

국가이미지의 대다수의 분석방법은 설문조사방법이었으며 몇몇 연구들은 국가의 통계데이터를 이용하였다. 이와 같은 분석방법은 객관적인 결론을 도출할 수 있지만 국가이미지의 형성경로로서 경험적 경로를 통한 분석방법이라 할 수 있다. 객관적이며 기존보다 개량된 국가이미지를 측정하기 위해 Mofitt (1994)이 정의한 경험적 경로 외 관계적 경로 및 텍스트적 경로로부터 데이터 추출이 필요하다.

최근 빅데이터의 분석기술이 발달함에 따라 비정형 데이터를 이용한 연구들이 증가하고 있다. 데이터의 특성상 위의 경험적, 관계적, 텍스트적 경로 모두를 포함한다. 그러므로 국가이미지를 측정하는데 있어 빅데이터를 활용하는 것에 의의가 있다. 그 중 ‘트위터’나 ‘페이스북’과 같은 SNS 빅데이터를 일반적으로 사용하고 있으나 개방된 API로부터 추출된 데이터의 경우 편의 (bias)된 데이터일 가능성이 높으며 양질의 데이터를 수집하기 위해 비용이 드는 단점이 있다. 뿐만 아니라 대부분의 소셜미디어를 통해서 발생하는 데이터는 사회관계적 경로를 통해 발생하므로 의견의 쏠림현상을 보일 수 있어 객관성에 의문이 제기 될 수 있다. 반면 위키피디아는 사회관계적 경로로 유입되는 경우가 적고 객관적 관점을 유지하기 위하여 많은 노력을 들이고 있어 국가이미지 온톨로지의 추출을 위한 기본 풀 (pool)로서 활용될 수 있다.

3. 국가이미지 온톨로지 키워드 추출 알고리즘

본 연구를 통하여 국가 이미지 키워드를 추출하고 동적 온톨로지인 WiNiOnto 구축하기 위한 “양방향 이중 키워드 수집 (Double Crossing Double Keyword Collection; DCDKC)”이라고 명명한 알고리즘을 개발하였다.

3.1. DCDKC (Double-Crossing Double Keyword Collection) 알고리즘

DCDKC 알고리즘은 Figure 3.1에 도시된 바와 같이 국가이미지에 관련된 키워드를 추출하는데 그 국가와 관련된 단어를 최대한 수집하면서 관련 키워드의 객관적 적절성을 유지하기 위해 다음과 같은 과정을 거치도록 고안되었다. 즉 어떤 키워드 A와 연관이 있는 모든 키워드를 찾으려면 위키피디아 영문판 460만 페이지를 다 뒤져보는 방법이 가능할 것이나 이는 현실적으로 계산복잡성 (Computational

Complexity)이라는 극복할 수 없는 문제를 가지고 있다. 이를 극복하기 위한 현실적 대안은 키워드 A를 중심으로 양방향 키워드를 단순히 수집하면 된다.

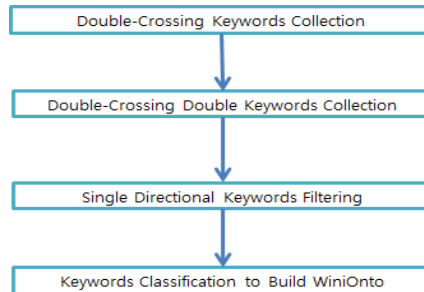


Figure 3.1 DCDKC Algorithm in Abstract

1) 양방향 키워드 수집: 위키피디아에서 어떤 키워드 A가 다른 키워드 B와 연관이 있는지 알려면 A의 위키 페이지에서 B를 언급했는지를 확인하면 된다. A는 B와 연관이 있다고 할 수 있다. 또한 B의 위키 페이지에서 A를 언급한다면 A와 B는 서로 연관이 있다고 확실히 말할 수 있다. 그러나 일방적으로 한쪽에서 다른 쪽을 언급한다면 연관의 확신이 없다. 따라서 본 연구에서는 국가이미지와 관련하여 키워드가 서로를 언급하여 확실한 연관이 있는 단어들을 추출한다 (Figure 3.2).



Figure 3.2 Double-crossing keyword collection example

2) 이중 키워드 수집: 양방향 키워드만을 수집할 경우 특정 국가와 확실하게 연관된 단어는 추출할 수 있었으나 각 카테고리별 구체적 의미를 갖는 단어의 추출은 불가능하다는 단점이 있다. 즉 한국이 K-Pop과 관계가 있지만 K-Pop에 속한 원더걸스 등은 구체적 단어는 추출되지 않는다. 이를 보완하기 위하여 본 연구에서는 이중 키워드 수집을 적용한다. 양방향 관계성을 갖는 A와 B를 추출할 때 한 번 더 B와 양방향 관계성을 갖는 C를 추출한다면 A와 관계있는 단어를 충분히 추출할 수 있다고 가정할 수 있다 (Figure 3.3).

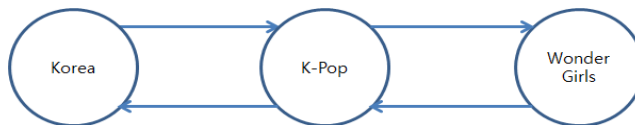


Figure 3.3 Double-crossing double keyword collection example

3) 단방향 필터 적용: 이중 키워드 수집은 특정 국가와 확실하게 연관된 단어는 추출할 수 있으나 이행성 규칙 (transitivity rule)이 성립되지 않는다는 단점이 있다. 즉 이중 키워드 방식으로 수집된 많은 키워드는 대부분 국가의 이미지와 관계있는 키워드이지만 경우에 따라서는 전혀 연관성 없는 키워드가 추출될 수 있다. 이를 방지하기 위하여 추출된 키워드 C가 A를 언급하는지 확인하는 작업을 거쳐서 마지막으로 온톨로지 하부 분류 키워드의 집합이 완성이 된다.

4) 키워드 카테고리 적용: 이렇게 키워드 집합은 다시 대분류와 중분류의 분류작업을 거쳐 국가 이미지 측정을 위한 위니온토 (WiNiOnto) 온톨로지가 완성이 된다.

3.2. 위니온토 구축 과정

본 연구에서는 DCDKC 알고리즘의 유용성을 입증하기 위하여 본 연구에서는 한중일 삼국의 온톨로지를 구축하고 이를 이용하여 삼국의 국가이미지를 추출하고 비교하는 작업을 다음과 같이 수행한다.

- 1) 각 국가별 이름 (Korea, Japan, China)으로 된 위키피디아 페이지를 기준으로 각 페이지에 수록된 링크가 수록된 1차 키워드를 추출한다 (Table 3.1의 “1st Collection”에 해당하는 키워드가 추출된다.)
- 2) 1차 추출된 키워드들의 위키피디아 페이지에 들어가서 페이지 내용을 파싱한 다음에 해당 페이지 내에 각 국가이름이 들어있는지 확인한다 (Table 3.1의 “1st Filtering”에 해당하는 키워드가 추출된다.)
- 3) 1차 추출된 키워드 페이지에서 다시 링크가 수록된 2차 키워드를 추출한다 (Table 3.1의 “2nd Collection”에 해당하는 키워드가 추출된다.)
- 4) 2차 추출된 키워드에서 1차로 추출한 키워드가 들어있는지 확인한다.
- 5) 2차 추출되고 1차 추출된 키워드가 들어있는 위키피디아 페이지에서 원래 국가의 이름도 추출이 되는지 확인한다 (Table 3.1의 “2nd Filtering”에 해당하는 키워드가 추출된다.)
- 6) 추출된 키워드를 분류한 결과 2차 필터링까지 수행한 결과 약 1만개에서 2만1천개 가량의 키워드를 추출하였다.
- 7) 같은 뜻의 단어이거나 유사한 단어를 통합하는 수작업을 하고 추출된 단어 중 관계없다고 판단되는 키워드를 제외한다 (Table 3.1의 “Keywords After Manual Inspection”에 해당하는 키워드가 추출된다.)
- 8) 마지막으로 선택된 키워드를 위의 Figure 2에서 도출된 대분류와 세부 키워드로 분류하여 국가이미지 동적 온톨로지 WiniOnto를 완성한다.

Table 3.1은 각 단계마다 도출된 키워드의 숫자를 보여 주고 있다. 최종 도출된 키워드의 총 개수는 한국이 중국과 일본에 비해서 압도적으로 많았고 특별히 재검토를 하였으나 아쉽게도 정확한 이유를 밝힐 수가 없었다. 다만 한국은 한류의 특성이 두드러진 현상이 있었는데 그로 인하여 사람이름이나 드라마 이름 등 고유명사가 특별히 많은 이유로 추정된다.

Table 3.1 The Number of Keywords Collected in Each Step for Korea, Japan, and China

Country	1 st Collection	1 st Filtering	2 nd Collection	2 nd Filtering	Keywords After Manual Inspection
Korea	472	362	59,556	12,841	7,733
Japan	548	458	71,354	21,974	4,534
China	591	477	67,328	16,322	2,706

Figure 2.1에서 제시되었던 정치, 경제, 사회, 문화 각 4개의 대분류와 20개의 세부 카테고리로 하여 알고리즘을 통해 도출된 키워드에 적용해 카테고리를 수작업으로 분류하였다. 그러나 총 키워드 14,973개 중 20개의 세부 카테고리에 분류할 수 없는 경우가 다수 있었다. 이 경우는 새로운 카테고리를 생성해 적용하였고 그 기준은 다음과 같다.

첫째, 정치영역에서는 한국의 경우 정부 관련보다 정당 및 정치인에 대한 키워드가 많아 정당 정치인에 대한 카테고리를 추가하였다.

둘째, 경제영역에서는 과학과 기술을 분류했고 한국과 일본의 경우 제품관련 키워드가 다수 있어 생성했다. 여기에서 제품 카테고리에서는 콘텐츠에 대한 내용을 포함한다.

셋째, 사회영역에서는 재난이나 특별사건 등이 일부 있어 사건 카테고리를 생성했고 ‘아줌마’, ‘386세대’, ‘재수생’ 등 세대와 관련된 키워드가 있어 생성했고 법이나 제도 그리고 시스템관련 키워드가 있어 제도 및 시스템 카테고리를 생성했다.

넷째, 언론 및 미디어의 경우 역시 키워드가 많아 카테고리를 생성했다.

다섯째, 문화의 경우 한류열풍에 따라 TV프로그램이 다수 있어 영화와 TV프로그램을 분류했다. 그리고 문화재와 명절과 같은 축제 키워드가 다수 있어 새로 카테고리를 생성했다.

4. 국가이미지 온톨로지 위니온토 적용

4.1. 뷰카운트 추출

이렇게 해서 DCDKC 알고리즘에 의해 완성된 최종 온톨로지인 WiniOnto의 유용성을 확인하기 위해 각 키워드마다 실제 페이지뷰 (page view) 카운트를 추출하였다. 위키피디아는 뷰카운트의 Dump파일을 일일단위로 제공하고 있고 Dump파일을 가공해서 뷰카운트 (view count)를 보여주는 웹사이트가 다수 존재하기 때문에 그 중 유명한 웹사이트를 임의로 ‘Wikipedia article traffic statistics (Wikipedia article traffic statistics; <http://stats.grok.se/>)’ 사이트를 선택하였다. 해당 웹사이트 내 뷰카운트 추출 첫단계로 위키피디아 언어와 뷰카운트를 보고 싶은 달을 선택하고 카테고리명을 입력하면 뷰카운트가 표시되고 임의로 기간을 선택해 뷰카운트를 볼 수 있다. 이 웹사이트의 경우 임의 기간을 선택할 수 있을 뿐만 아니라 달별로 뷰카운트를 제시한다는 특성을 갖고 있어 2008년 1월부터 2014년 8월까지 매달의 뷰카운트를 추출했다.

Table 4.1 Category page views for Korea, Japan, and China (2008.7.~2014.6.)

	Politics	Society	Culture	Economy	Total
Korea	14,629.2 (11.0)	23,533.3 (17.6)	76,429.9 (57.2)	19,010.9 (14.2)	133,603 (100.0)
China	7,545.59 (4.7)	76,513.5 (47.3)	27,369.7 (16.9)	50,393.8 (31.1)	161,823 (100.0)
Japan	12,907.9 (6.6)	62,881 (32.1)	64,184.4 (32.8)	55,655.4 (28.5)	195,629 (100.0)
Total	35,082.7 (7.14)	162,928 (33.18)	167,984 (34.21)	125,060 (25.47)	491,054 (100.0)

(Unit: 10,000, % in parenthesis, $\chi^2=67976.3$, $df=6$, $p < 0.0001$)

추출된 뷰카운트는 Table 4.1에서 보여주는 바와 같다. 온톨로지 키워드로 분류된 페이지 뷰에 한중일 국가간의 정치, 사회, 문화, 경제의 이미지 차이가 있는지 카이제곱 통계량으로 통계적 분석을 하고 대응분석을 통해 이미지 시각화를 시도하였다.

4.2. 한중일 국가이미지 분석 및 시각화 결과

위키온토에서 추출된 한중일 국가간 심층적 비교분석 결과는 다른 논문에서 자세히 논의된 바와 같이 대응분석을 실시하고 매우 흥미로운 결과를 다수 도출하였다 (대응분석에 대한 기술적 이론적 논의는 다른 논문에서 하였으므로 여기서는 생략한다). 특히 국가이미지를 분석하는 중에 ‘한식세계화’와 ‘한류우드’ 등의 정책 효과성까지도 검증될 수 있다는 것을 발견하는 의외의 결과까지도 얻을 수 있었다 (Lee 등, 2015). 본 연구는 또한 대응분석의 결과를 시각화를 통해 누구나 쉽게 이해할 수 있도록 구글 모션차트를 이용하여 시각화된 모션 그래픽 차트로 도시하였다. Figure 4.1에서 보듯이 한·중·일 3국 간의 정치, 사회, 문화, 경제 인지도 즉 대분류 간의 상대적인 관련성을 시각적으로 표현하였다. 한국은 문화와, 중국은 사회와 독보적으로 매우 밀접한 관련이 있다. Table 4.1을 자세히 관찰하면 일본은 한국과 중국에 비해 상대적으로 사회와 문화의 페이지 뷰가 적지 않음에도 경제의 클릭뷰가 Figure 4.1에는 상당히 높게 나타나는 것을 볼 수 있다. Figure 4.1은 3국간의 국가 간 이미지에 있어서 상당한 차이가 존

재하고 있음을 보여준다 (실시간으로 도시된 그래프는 그림 아랫부분의 화살표 단추를 누르면 그래프가 시간축을 따라 움직이므로 대분류와 국가이미지 간의 시간적 변화를 알 수 있다).

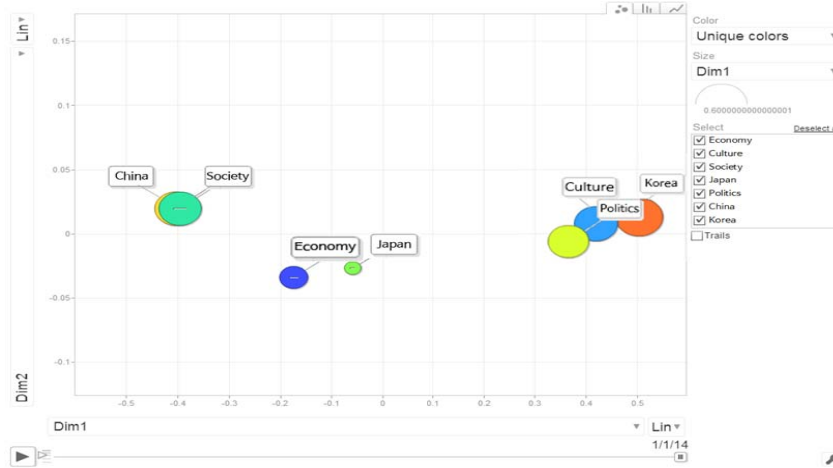


Figure 4.1 Nation image comparison of the three countries, Korea, China, and Japan through WikiOnto categories shown in Wikipedia click views

국가이미지의 연도별 변화를 좀 더 자세히 볼 수 있도록 다시 도시하면 Figure 4.2와 같다. 흥미롭게도 대분류 연도별 변화추이를 보면 주황색선으로 표시된 처음 그래프가 시작하는 시점인 2008년 중반 이후 한국이 문화영역에 근접하여 위치함으로써 한국이 문화영역에서 중국과 일본에 비하여 압도적인 강세를 보이고 있다는 것을 보여주고 있다. 하지만 초기에는 한국과 문화 사이에 다소의 갭이 보이고 있어 일본과 중국에 대하여 문화적으로 관심이 있는 인터넷 사용자들도 있었다는 것을 관찰할 수 있다. 그러나 2012년 중반을 기점으로 문화가 한국에 더욱더 가까워지는 것을 볼 수 있는데 이를 세부 키워드로 분석해보면 영화가 한류의 새로운 축으로 편입하기 시작한 시점인 것을 볼 수 있다.

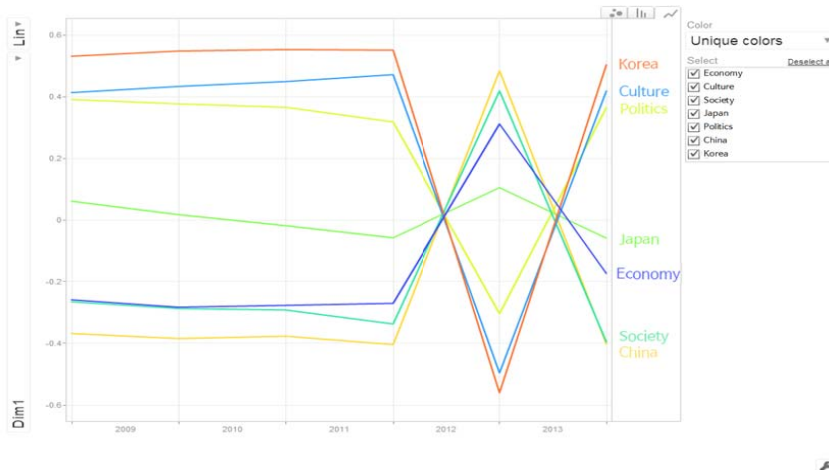


Figure 4.2 Dynamic nation image changes of the three countries, Korea, China, and Japan through WikiOnto categories shown in Wikipedia click views

2012년 중반에는 한류에 있어서 많은 변화가 나타난다. 그것은 영화에서의 변화뿐만이 아니고 한류 음악에서 싸이의 등장이다. Figure 4.3을 보면, 한국이 문화에서 음악의 비중이 중국과 일본에 비해 압도적으로 높은 데 그 이유는 빅뱅을 비롯하여 소녀시대가 지속적인 한류음악을 유지하는 힘이었던 싸이는 압도적인 분량으로 문화영역의 우위를 지키는데 일조한 것으로 보인다. 흥미롭게도 이와 비슷한 현상이 비슷한 시점에 영화에서도 나타난다. 그것은 아마도 2012년 사이의 등장이 한국에 대한 관심을 폭발적으로 이끌게 된 부수적 효과라고 할 수 있다.

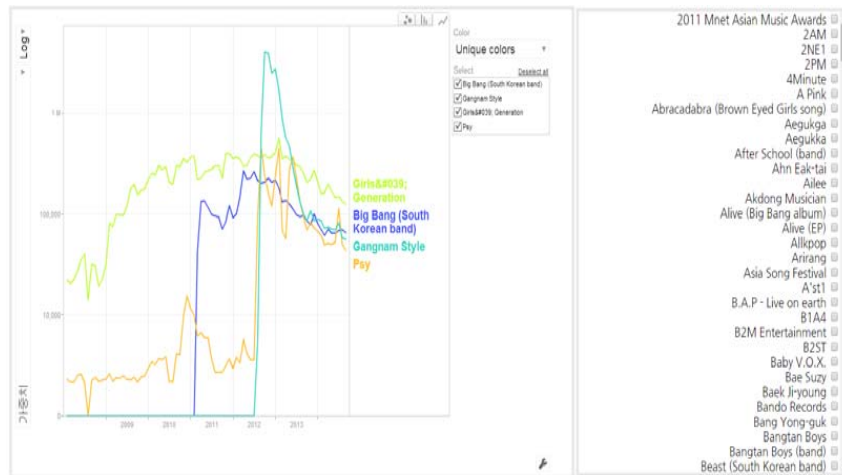


Figure 4.3 Click view changes of Girls Generation, Big Bang, Kangnam Style and Psy in Korean music keywords

5. 결론

본 연구에서 연구자들은 국가이미지를 실시간 측정하기 위하여 위키피디아의 키워드를 이용하여 온톨로지를 구축하는 DCDKC 알고리즘이 개발하였다. 이 알고리즘에 의해 추출된 국가이미지 온톨로지인 위키온토는 변화하는 국가이미지를 동적으로 실시간으로 자동 측정하여 제공할 수 있도록 하였다. 따라서 국가이미지 관리에 효율성을 기하고 뜻밖에도 있을 지도 모르는 국가이미지의 급박한 변화에 선제적으로 대응할 수 있는 도구를 개발하고자 하였다.

이는 과거 설문조사에 의존하던 각종 국가브랜드지수 모델과는 달리 온라인 지표와 그에 맞는 도구를 개발하고 이를 위키피디아에서 네티즌이 키워드를 클릭하는 숫자를 분석하여 위키온토의 유용성을 검토하였다.

본 연구의 결과는 다음과 같이 요약될 수 있다.

첫째, 국가이미지 온톨로지 구축 알고리즘인 DCDKC의 유용성이다. 위키피디아에서 추출된 결과에서 보듯이 DCDKC를 사용하여 구축된 온톨로지로 시험적으로 한중일 삼국의 국가이미지를 추출하고 많은 시사점을 발견하였다. 특히 연구초기에는 생각지도 못한 국가정책 수행평가 등에도 사용될 수도 있다는 시사점을 포함하여 DCDKC 알고리즘의 유용성은 충분히 확인되었다 (Lee 등, 2015).

둘째, 국가이미지 실시간 관리체계에 대한 가능성이다. 국가이미지를 실시간 관리하는 것은 무엇보다도 중요하다. 예컨대 2012년부터 한류의 한 축으로 자리 잡을 수 있는 가능성을 보여주는 영화산업의 경우 국가이미지 관리차원에서 초기 발견을 통해 적절한 대응을 한다면 국익에 크게 도움이 될 수 있으리라고 본다.

셋째, 위키피디아를 이용한 한중일 국가 이미지 분석 및 이용가능성 확인이다. DCDKC에서 추출한 온톨로지 위키온토의 적용성은 단지 위키피디아에서 그치지 않는다. 본 연구에서는 단지 알고리즘의 유용성을 위하여 위키피디아를 선택하여 한중일 삼국의 국가이미지를 분석하고 비교한 바 국가이미지 분석에 저렴하지만 탁월한 역할을 할 수 있음을 증명하여 확인하였다.

본 연구가 국가이미지를 과거와는 다르게 실시간 관리할 수 있는 방법론을 제시하였다는 점에서 새로운 기여를 하였다고 할 수 있으나 다음과 같은 한계점도 다수 발견되었다.

첫째, DCDKC의 완전자동화의 어려움이다. 위키피디아의 키워드를 추출하면서 위키피디아의 많은 부분이 완벽하게 되어있지 않아서 적지 않은 양의 수작업이 필요했다는 점은 아쉬운 점이다. 특히 키워드를 추출했을 때 클리닝을 해야하는 점도 있었고 알고리즘의 하자는 발견하지 못했으나 아무런 이유도 없이 따라 들어오는 키워드들도 다수 있었으므로 다수의 키워드들을 매뉴얼로 지워야 했다. 다행히도 한 번 수작업한 키워드는 초기에 제외할 수 있다는 점에서 반복으로 작업하면 다수의 수작업의 양은 상당량 줄어들 것이다.

둘째, 키워드 분류의 어려움이다. 본 작업을 통해 키워드 분류를 수작업에 의존했다. 일단 온톨로지가 작동하면 수작업의 양은 절대량이 줄어들 것으로 보지만 향후 완벽한 자동화를 위해서는 자동클러스터링 알고리즘을 추가하는 것도 한 방법일 것으로 본다.

References

- Anholt, S. (2006). The Anholt-GMI city brands index: How the world sees the world's cities. *Place Branding*, **2**, 18-31.
- Anholt, S. (2005). Anholt nation brands index: How does the world see America? *Journal of Advertising Research*, **45**, 296-304.
- Ahn, Changwon, Seungkoo Whang (2012). Big data technology and primary issues. *Journal of the Korean Data & Information Science Society*, **30**, 10-17.
- Ahn J. S. (2005). Studies on the effect of multi-dimensional attributes of nation image on the product and brand evaluation: Focusing on Chinese consumers. *International Business Journal*, **6**, 63-90.
- Bradley, A., Hall, T. and M. Harrison, (2002). Selling cities: Promoting new images for meeting tourism. *Cities*, **19**, 61-70.
- Choi, C. W. (2008). A case study on the nation brand strategy of the United Kingdom to China: Focus on 'Think UK' campaign. *The East-West Channel*, **23**, 81-114.
- Chun, H. (2015). The comparison of coauthor networks of two statistical Journals of the Korean Statistical Society using social network analysis. *Journal of the Korean Data & Information Science Society*, **26**, 335-346.
- Futurebrand (2010). *2010 Country Brand Index*, Futurebrand, <http://bit.ly/1NobMQN>.
- Hall, C. J. (1986). *National image: A conceptual assessment*, Paper presented to the International Communication Association, Boston.
- Kim, C. H., Son, Y. S. and W. B. Na. (1997). Measuring nation's image in cross cultural context. *Proceedings of the 6th Symposium on the Cross Cultural Consumer and Business Study*, 148-149.
- Kim, B. C. (2007), *A Study on the Effect of Nation Image to Product Image*, Master Thesis, Hanyang University, Seoul.
- Lee, J., Lee C., and Yang S. (2012). A exploratory empirical study on the factors impacting nation image in Korea: Variables focused on government, enterprise, and civil activities. *Korean Journal of Business Administration*, **22**, 1583-1601.
- Lee, Youngwhan, Chun, H., and Sawng, Y. (2015). A study on the nation image of the big three exporting countries in East Asia shown in Wikipedia English-edition. *Journal of the Korean Data & Information Science Society*, **26**, 1071-1085.
- Li, Z. G., Murray, L. W. and D. Scott. (2000). Global sourcing, multiple country-of-origin facets, and consumer reactions. *Journal of Business Research*, **47**, 121-133.
- Martin, I. M. and S. Eroglu. (1993). Measuring a multi-dimensional construct: country image. *Journal of Business Research*, **28**, 191-210.

- Moffitt, M. A. (1994). A cultural studies perspective toward understanding corporate image: A case study of State Farm insurance. *Journal of Public Relations Research*, **6**, 41-66.
- Nagashima, A. (1970). A comparison of Japanese and U.S. attitude Toward foreign products. *Journal of Marketing*, **34**, 68-74.
- Papadopoulos, N. G. (1993). What product and country images are and are not. In *Country-of-origin images : Impact and role in international marketing*, International Business Press, New York.
- Roth, M. S. and J. B. Romeo. (1992). Matching product category and country image perceptions: A framework for managing country-of-origin effects. *Journal of International Business Studies*, **23**, 477-497.
- Samsung Economic Research Institute. (2013). *Research report on 2012 nation brand index*, SERI Issue Papers, Seoul, Korea.
- Shin, J. (1993). *Antecedents and moderators of consumer ethnocentrism, and the effect of source country characteristics on consumer's attitudes toward purchasing imported product categories*, Ph. D. Thesis, University of South Carolina, South Carolina.
- Song, T. M., Song, J. and Cheon, M. K. (2015). Predicting tobacco risk factors by using social big data. *Journal of the Korean Data & Information Science Society*, **26**, 1047-1059.
- The institute for industrial policy studies (2007). *2007 The assessment report on nation and enterprise brand value*, The Institute of Industrial Policy Studies, Seoul.
- Wang, C. K. (1978). *The effect of foreign economic, political and cultural environments on consumers' willingness to buy foreign products*, Ph.D. Thesis, Texas A&M University, Texas.

Dynamic ontology construction algorithm from Wikipedia and its application toward real-time nation image analysis

Youngwhan Lee¹

¹William Miller School of Management of Technology, Business College, Konkuk University

Received 10 May 2016, revised 2 July 2016, accepted 22 July 2016

Abstract

Measuring nation images was a challenging task when employing offline surveys was the only option. It was not only prohibitively expensive, but too much time-consuming and therefore unfitted to this rapidly changing world. Although demands for monitoring real-time nation images were ever-increasing, an affordable and reliable solution to measure nation images has not been available up to this date. The researcher in this study developed a semi-automatic ontology construction algorithm, named “double-crossing double keyword collection (or DCDKC)” to measure nation images from Wikipedia in real-time. The ontology, WikiOnto, can be used to reflect dynamic image changes. In this study, an instance of WikiOnto was constructed by applying the algorithm to the big-three exporting countries in East Asia, Korea, Japan, and China. Then, the numbers of page views for words in the instance of WikiOnto were counted. A collection of the counting for each country was compared to each other to inspect the possibility to use for dynamic nation images. As for the conclusion, the result shows how the images of the three countries have changed for the period the study was performed. It confirms that DCDKC can very well be used for a real-time nation-image monitoring system.

Keywords: Automated keyword extraction, dynamic ontology, nation image, Wikipedia, brand management

¹ Assistant professor, William Miller School of Management of Technology, Business College, Konkuk University, Seoul 05029, Korea. E-mail: nicklee@konkuk.ac.kr