Journal Of The Korean Society Of Rural Planning
Vol. 23, No. 3, 2017 (37-51)
Issn 1225-8857(Print) / Issn 2288-9493(Online)
Http://Dx.Doi.Org/10.7851/Ksrp.2017.23.3.037

한국농촌계획 온톨로지 구축을 위한 상호정보 기반 단어연결망 분석

이제명

교토대학교 지역환경과학전공

Word Network Analysis based on Mutual Information for Ontology of Korean Rural Planning

Lee, Jemyung

Postdoctoal associate, Division of Environmental Science and Technology, Kyoto University, Japan

ABSTRACT : There has been a growing concern on ontology especially in recent knowledge-based industry and defining a field-customized semantic word network is essential for building it. In this paper, a word network for ontology is established with 785 publications of Korean Society of Rural Planning(KSRP), from 1995 to 2017. Semantic relationships between words in the publications were quantitatively measured with the 'normalized pointwise mutual information' based on the information theory. Appearance and co-appearance frequencies of nouns and adjectives in phrases are analyzed based on the assumption that a 'noun phrase' represents a single 'concept'. The word network of KSRP was compared with that of WordNet™, a world-wide thesaurus network, for the verification. It is proved that the KSRP's word network, established in this paper, provides words' semantic relationships based on the common concepts of Korean rural planning research field. With the results, it is expecting that the established word network can present more opportunity for preparation of the fourth industrial revolution to the field of the Korean rural planning.

Key words : The Fourth Industrial Revolution, Information Theory(IT), Pointwise Mutual Information(PMI), Phrase analysis, Semantic word network

I. 서 론

지식, 정보 기반의 '4차 산업혁명(The Fourth Industrial Revolution)'이 부각되면서 다양한 분야에서 이에 대응하기 위한 노력이 이루어지고 있다. 농촌 지역의 개발에도 이러한 흐름이 반영되어 농촌을 전통적인 1차 산업을 위한 공간에서 1, 2, 3차 산업을 연계한 '6차 산업'을 위한 공간으로 개발하기 위한 노력에 이어(Yang et al., 2014), 이제는 지식과 정보를 활용하는 맞춤형 농산업 및 관련 산업을 활성화할 수 있는 공간으로 개발하기 위한 방안을 모색하고 있다(Choi and Choi, 2016). 기존에도 센서를 통해 실시간으로 정보를 수집하고 활용하는 정밀농업(Park and Chang, 2001; Nam et al., 2011)이나, 농촌지역의 자원정보에 관한 데이터베이스 구축(Lee and Lee, 2006; Park et al., 2014b)과 같이 농업, 농촌 활성화를 위한 다양한 정보기술의 활용방안이 활발히 연구되어 왔다. '4차 산업혁명' 이슈는 여기에서 더 나아가 농업과 농촌의 개발에 '빅데이터(Bigdata)', '온톨로지(Ontology)'와 같이 다양한 분야에서 축적된 지식과 정보를 활용하는 정보기술과 '사물인터넷(Internet on Things, IoT)'과 같이 실시간으로 정보를 주고받는 통신기술, 그리고 '시멘틱 웹(Semantic Web)', '인공지능(Artificial Intelligence, AI)'과 같이 축적된 정보를 인간의 사고방식과 닮은 논리적인 추론을 통해 분석하는 정보처리기술을 융합하여 적용함으로써 농업의 산업역량과 농촌의 지역역량을 극

Corresponding author : Lee, Jemyung
Tel : +81-75-753-6159
E-mail : lgm00@snu.ac.kr

대화할 수 있는 방안을 마련할 수 있는 기회를 열어주고 있다. 이러한 이점에 힘입어 각 지자체, 연구소 및 중앙정부를 포함한 다양한 기관에서 '4차 산업혁명'에 대응한 산업 및 지역개발을 준비하고 있으며(MSIP, 2016; Choi, 2017; Hyun and Ham, 2017), 농업·농촌의 개발에도 기존 데이터베이스 활용 및 빅데이터 구축 등을 통해 '4차 산업혁명'에 대응하기 위한 움직임을 보이고 있다(MAFRA, 2017a, 2017b).

'4차 산업혁명'의 핵심은 기존 지식체계 간의 연결을 통해 정보의 활용성을 극대화하는 데에 있으며 이를 구현하기 위한 도구로서 '시멘틱 웹(semantic web)'의 개념을 채용하고 있다(Grangel-González et al., 2016). '시멘틱 웹'은 '빅데이터' 이슈 등을 통해 구축된 방대한 지식체계를 이용, 자료 사이의 연관관계를 정의하고 이를 바탕으로 논리적인 추론을 통해 정보의 의미를 분석하고 사용자에게 필요한 정보를 제공할 수 있는 지능형 정보전달 방식으로 이해할 수 있다. 예를 들어, '배'라는 항목의 자료를 제공함에 있어 사용목적이 농업과 관련되어 있을 때는 과일인 '배'에 관한 자료 내에서 정보를 분석하는 등의 서비스가 가능한 정보처리 방식이다. 이러한 기능을 구현하기 위해서는 사전에 '배-식품-농업' 과 같은 지식체계가 구축되어 있어야 하는데 이러한 역할을 하는 것이 '온톨로지(ontology)'이다. 따라서 '4차 산업혁명'에 대응하기 위해서는 해당 분야의 '온톨로지' 구축이 필수적이며, '4차 산업'에 대응하는 농촌개발 계획을 수립하기 위해서는 '농촌계획' 분야에 대한 맞춤형 지식체계, 즉 '농촌계획 온톨로지' 구축이 필수적이다.

도서정보(Weinstein, 1998), 하천(Yoo and Yoon, 2000), 생명정보(Yang et al., 2004), 지반정보(Lee, 2010)뿐만 아니라 작황정보(Lee et al., 2008), 농작물정보(Lee, 2009) 등 다양한 분야에서 고유의 온톨로지를 구축하기 위한 연구가 진행되었다. '농촌계획' 분야에서도 '농촌정보'의 활용성을 높이기 위한 온톨로지 활용에 관한 연구(Lee et al., 2006; Lee and Lee, 2006)가 진행되기도 하였으며, 이를 위한 데이터베이스 설계에 관한 연구(Lee et al., 2005)가 진행되는 등 온톨로지 데이터베이스의 설계와 활용에 관한 연구가 활발히 진행된 바 있다. 그러나 '농촌계획' 분야 '온톨로지'의 내용(contents)을 어떻게 구축할 것인가에 대한 연구는 부족한 실정이다.

'4차 산업혁명'에 효과적으로 대응하기 위해서는 기존에 구축된 지식자원의 재이용과 활용성 증진을 도모하여야하며 이를 위해서는 지식 간 연계가 핵심적인 요소이다. '온톨로지'의 구축은 해당 분야에서 통용되는 지식, 용어 혹은 개념 간의 연결망(network)을 바탕으로 이루어지는데(Thomas, 1993), 이 연결망이 정보의 활용성을 극대화시킴으로써 컴퓨터 프로그램의 논리적 추론 과정을 더욱 사람의 사고방식에 근접하도록 도와준다(Berneers-Lee, 2001). 따라서 '농촌계획 온톨로지' 구축을 위해서는 '농촌계획' 분야의 단어 간, 용어 간, 혹인 개념 간 네트워크 분석이 선행되어야 한다. 본 연구에서는 '4차 산업혁명'에 대응하기 위한 한국 '농촌계획' 분야의 단어 네트워크를 구축하기 위해 '개념'을 공유하는 단어 사이의 연관성을 분석하고자 하였다. 또한 한국 '농촌계획' 분야에 대한 방대한 지식자원이 축적되어 있는 한국농촌계획학회의 논문자료를 분석에 활용하고자 하였다.

'농촌계획' 분야의 '온톨로지'를 구축함으로써 향후 사회 전반에 걸쳐 막대한 영향력을 미치게 될 지식체계에 '농촌계획' 분야에서 통용되는 개념을 반영할 수 있는 토대를 마련할 수 있다. 예를 들어, 일반적인 연관 개념으로 구축한 온톨로지에서는 '농촌'과 관련한 개념으로 '낙후', '고령화' 등이 도출되나 '농촌계획 온톨로지'에서는 '녹색관광', '농촌어메니티' 등으로 개념을 연결시킬 수 있게 된다. 이러한 온톨로지 구축을 위해서는 해당 분야의 지식체계를 반영한 개념 간 연결망 분석이 필요하다. 본 연구는 '농촌계획' 분야의 논문자료를 통해 의미 기반 단어 네트워크를 구성함으로써 '농촌계획 온톨로지' 구축에 활용할 수 있는 기초자료를 마련하는 데에 그 의의가 있다.

본 논문은 '농촌계획 온톨로지' 구축을 위한 기초연구로서 '농촌계획' 분야에서 통용되는 단어(용어) 간 연결망을 분석하는 것을 연구목적으로 삼고, 연결망 분석에 있어 '한국농촌계획' 분야 고유의 온톨로지를 구축할 수 있도록 동의어나 유의어와 같이 기존에 구축된 연결망이 아닌 '농촌계획' 분야에서 통용되는 '개념'을 바탕으로 단어 연결망(network)을 분석하는 것을 목표로 설정하였다. 이를 위해, 한국농촌계획학회(KSRP)에 1995년부터 2017년 3월 까지 게재된 논문 785편의 텍스트(text)를 분석하여 단어 간 연관성을 계측하고, 이를 통해 단어 연결망을 구축하였다. '개념'에 기반한 단어 사이의 연관성을 분석하기 위해 구(phrase) 단위에서의 단어 사용 패턴을 분석하였으며, 정보이론(Information Theory)에서 등장한 '정규화된 점별 상호정보(Normalized Pointwise Mutual Information, NPMI)'를 활용하여 단어 사이의 연관성을 정량화하였다.

II. 기본 이론

1. 4차 산업혁명과 농촌계획

‘4차 산업혁명’은 전자, 정보기술을 통한 생산의 자동화를 이루었던 ‘3차 산업혁명’의 기반 위에 정보통신기술과 정보처리기술, 그리고 광범위한 영역의 데이터베이스의 융합을 통해 이루어지고 있으며, 전산화(digital)된 산업, 사회, 학계의 각 분야를 통합할 뿐만 아니라 정보(information)를 다루는 기술을 통해 물리적, 생물학적 경계를 허물고 있다는 점에서 기존 산업혁명과 구분된다(Schwab, 2015). 이와 같은 ‘4차 산업혁명’의 핵심은 정보, 지식체계의 통합 및 그 활용에 있다.

정보를 문서에 기록하던 방식에서 전자매체에 기록하는 방식으로 변경된 전산화가 이루어진 이후 정보저장 및 처리기술의 발달에 따라 정형화된 정보를 처리하던 수준에서 최근에는 비정형 데이터로 불리는 방대한 양의 일반 언어를 직접 분석하고 이로부터 정보를 추출하는 이른바 ‘빅데이터’(Mashey, 1998)로 불리는 기술을 통해 다양한 형태의 지식과 정보를 저장하고 활용할 수 있게 되었다. 이렇게 축적된 지식체계에서 인간의 논리적인 추론방식을 적용하여 자료를 분석하고 사용자에게 맞는 정보를 제공하기 위한 도구로서 ‘시멘틱 웹’(Berneers-Lee, 2001)의 개념이 등장하였으며, 이를 위해 필요한 지식체계를 제공할 수 있도록 논리적인 추론에 활용할 수 있는 형태로 자료간의 연계를 표현하고 저장한 ‘온톨로지’(Thomas, 1993)가 여러 분야에서 구축되어 활용되고 있다. 이러한 지식체계는 이른바 ‘4차 산업혁명’을 통해 기존 산업 및 지식체계와 결합하면서 사회 전반에 영향을 미치고 있다. ‘4차 산업혁명’은 기존에 구축된 지식과 정보를 인간의 사고방식과 닮은 형태로 유기적으로 연결하고, 사용자가 필요한 혹은 필요할 것으로 예상되는 정보를 논리적인 추론을 통해 제공함으로써 기 구축된 정보의 효용성을 최대화 한다는 데에 그 의의가 있다.

농업과 농촌개발 분야에서는 이미 ‘6차 산업’에 관한 특례법 지정 등 기존 1, 2, 3차 산업 기반 간의 유기적인 연계를 통해 산업간 시너지 효과를 극대화하고자 하는 노력을 기울여 왔다(Park et al., 2014a). 이러한 산업 인프라 간의 연계를 위한 노력이 지식체계의 유기적인 연계와 통합을 도모하는 ‘4차 산업혁명’의 흐름과 결합한다면 그 시너지 효과는 더욱 극대화될 수 있을 것으로 기대되고 있다. ‘4차 산업혁명’에 대응하기 위해서는 해당 분야의 지식체계, 즉 온톨로지가 구축되어 있어야 한다. ‘농촌계획’ 분야에 관한 정보 및 지식 체계는 본 학회인 농촌계획학회의 논문집 그리고 농촌진흥청에서 진행한 농촌어메니티 자원도 조사사업 데이터베이스(Park et al., 2014b) 등 활용가능한 정보자원이 풍부하게 축적되어 있는 것으로 판단된다. 이러한 정보자원을 ‘4차 산업혁명’ 이슈에 맞추어 활용할 수 있는 방안을 마련하기 위한 기초연구로서 ‘농촌계획’ 분야의 온톨로지 구축에 관한 연구가 필요하다. 이를 위해 본 연구에서는 본 ‘한국농촌계획학회’에 축적된 논문 자원을 활용하여 ‘농촌계획’ 분야의 온톨로지 구축에 활용할 수 있는 단어 연결망을 분석하고자 하였다.

2. 온톨로지 구축

사람과 더 자연스러운 의사소통이 가능한 지능형 시스템을 만들기 위해서는 사람처럼 사물이나 대상을 인식하고 연관되는 다른 사물이나 대상, 또는 개념을 떠올리는 데에 사용할 수 있는 지식체계가 필요하다. 이러한 필요성이 의해 등장한 온톨로지(ontology)는 특정 분야에 대해서 공유할 수 있는 개념들의 연결망을 정형화된 형태로 기술한 것을 의미하며(Thomas, 1993), 컴퓨터가 사람의 직관적인 사고방식을 이해할 수 있는 형태로 정보를 처리하고 표현할 수 있는 시멘틱 웹의 논리적 기반이 된다(Kim et al., 2005). 온톨로지는 대상이나 개념을 표현하기 위한 단어(용어)와 이들 간의 관계를 구조화하고 표준화하기 위한 목적으로 만들어지며, 해당 집단이 공유하고 있는 합의된 개념과 관계를 나타낼 수 있어야 한다(Uschold and Gruninger, 1996; Benjamins et al., 1998; Staab et al., 2001; Mun and Woo, 2006)

초기의 온톨로지는 관련분야의 전문가가 직접 구축하였으나 이는 시간과 비용이 과다하게 소요되는 단점이 있어 다양한 분야에 적용하기에는 어려움이 있었다(Gu et al., 2006). 전자사전데이터나 WordNet™(Miller, 1995)과 같이 기존에 단어 사이의 관계가 정의된 시소러스(thesaurus)를 활용하는 방안이 활용되기도 하였으나(Kang, 2004; Song, 2005), 이러한 방법은 일반적인(general) 온톨로지 구축에는 용이하나 ‘농촌계획’과 같이 특정 분야의 온톨로지를 구축하는 데에는 한계가 있다. 이를 해결하기 위해 특정 분야와 관련된 문헌 내에 사용된 단어의 통계적인 특성으로부터 ‘개념’간의 관계를 도출하는 방안이 온톨로지 구축에 활용되었다(Maedche and Staab, 2000). 통계적인 특성으로부터 온톨로지를 구성하는 경우에는 Yarowsky(1995)의 연구와 같이 연관어 중심으로 분석이 이루어졌다. 이 때, 분석대상 자료가 특정 분야에 대해 일정한 형식에 맞추어 개관적으로 작성되어 있을 경우 좋은 결과를 나타내는 것으로 알려져 있다(Kim et al., 2005). 논문 자료는 이러한 분석에 활용하기에 적합한 조건을 갖추고 있는 것으로 여겨져, 특정 분야에 대한 논문 분석을 통해 해당 분야의 온톨로지를 구축하기도 하였다(Lee et al., 2014). 본 연구에서는 ‘한국농촌계획학회’에 등재된 논문 785편 내에 사용된 단어의

통계적 특성을 분석함으로써 '농촌계획' 분야의 온톨로지 구축에 활용할 수 있는 연관어 중심의 단어 네크워크를 구축하고자 하였다.

3. 단어네크워크 구축

온톨로지를 구축하기 위한 단어 네트워크를 구성은 단어 쌍의 연관성 분석을 통해 이루어진다. 특정 단어가 동시에 출현하는 빈도가 높은 것은 두 단어가 나타내는 개념이 서로 연관되어 있는 것으로 해석되어(Ding et al., 2001), 온톨로지 구축에 활용할 단어 네크워크의 구성은 주로 단어의 동시사용 빈도분석을 통해 이루어졌다(Shim and Choi, 2013; Lee, 2016). 단어 분석 범위에 있어 논문에 사용한 단어 전체를 분석하는 것보다 논문의 내용이 간결하고 명확하게 제시되어 있는 제목, 초록, 키워드를 분석하는 것아 보다 명확한 결과를 도출하는 것으로 연구되었다(Hyvönen et al., 2003; Mun and Woo, 2006; Lee et al., 2014). 단어 쌍의 동시사용 판별기준에 있어서는 문서의 '주제'를 바탕으로 단어 네트워크를 구성하는 경우에는 동일 문서 내 동시 사용된 빈도를 분석하여 활용하였으며(Cambrosio et al., 1993), '개념'에 기반한 단어 네트워크를 구축하는 경우에는 동일 '어구(phrase)' 내 동시 사용된 빈도를 활용하였다(Sanderson and Croft, 1999; Lawrie et al., 2001). 이는 하나의 문서는 하나의 '주제'를 담고 있으며, 하나의 '어구'는 하나의 '개념'을 나타낸다는 가정에 바탕을 두고 있다.

단어 네트워크 구성을 위한 단어 간 연관성 분석에 있어 동시사용 빈도만을 활용하는 경우, 단어마다 사용한 빈도가 다르다는 점을 고려할 수 없어 단어 쌍의 연관성 크기를 상대적으로 비교하는 데에 어려움이 발생한다. 이러한 문제점을 보완하고자 '상호정보(Mutual Information)'가 단어 사이의 연관성 분석에 도입되었다(Kang, 2004). Fano(1961)가 제안한 '상호정보'는 정보 집합이 공유하고 있는 정보의 양을 정량화하기 위해 등장한 개념이며, Shannon(1948)에 의해 정립된 '정보이론(Information Theory, IT)'에 기반을 두고 있다. '상호정보'는 정보통신기술(Information and Communication Technology, ICT) 분야에서 주로 활용되어오다, Church and Hanks(1990)의 연구를 통해 언어로 이루어진 텍스트 분석에 적용하기 시작하였다. '상호정보'를 활용할 경우 논문과 같은 문서 집단 내에서 개체 사이의 연관성을 정량화할 수 있는 이점이 있으나 이는 문서와 같은 단어 군집 사이의 비교에 적용할 수 있는 개념이며 이를 단어 사이의 연관성 분석에 활용하기 위해서는 '점별 상호정보(Pointwise Mutual Information, PMI)'를 적용하여야 한다. 그러나 PMI에서도 개체의 사용빈도 격차에 따라 값이 편향되는 문제가 발생할 수 있어(Role and Nadif, 2011), 단어 네트워크 분석과 같이 빈도의 편차가 큰 대량의 데이터 분석에 활용하기 위해서는 PMI를 보정한 '정규화된 점별 상호정보(Normalized Pointwise Mutual Information, NPMI)'를 적용하여야 한다(Bouma, 2009; Damani, 2013). 본 연구에서는 '농촌계획' 분야의 '개념' 기반 단어 네트워크를 구축하고자 '한국농촌계획학회'에 등재된 논문의 제목, 초록, 키워드로부터 '어구' 단위의 단어 동시사용 빈도를 분석하고 이로부터 NPMI를 계량함으로써 단어 사이의 연관성을 정량화하였다.

III. 연구대상 및 연구방법

1. 연구대상

본 연구에서는 '한국농촌계획학회(KSRP)'에 1995년부터 2017년 3월 까지 등재된 논문의 영문 제목(title), 초록(abstract), 및 키워드(keyword)를 분석에 활용하였다. 분석 대상 텍스트로부터 추출한 '어구(phrase)'를 중심으로 단어의 사용빈도를 분석하였다. '어구'의 분석에 있어 '명사구(noun phrase)'를 구성하는 단어를 연구대상을 설정함으로써 '개념'을 중심으로 단어 사이의 관계를 분석하고자 하였다. 단어는 명사와 형용사를 중심으로 분석하였으며, 총 785편의 논문을 대상으로 분석에 활용한 '명사구'의 수와 평균 구성 단어 수, 그리고 분석 단어의 총 수와 종류를 정리하여 Table. 1에 나타내었다. 또한, 분석 대상 단어 중 사용빈도가 높은 단어를 명사와 형용사로 구분하여 Table. 2에 정리하였다.

Table 1. Statistics of nouns, adjectives and noun phrases in the KSRP publications

Noun	Total	65,859
	Sort	5,210
Adjective	Total	16,938
	Sort	2,239
Noun-Phrase	Total	24,681
	Avg.[1)]	2.45

1) Average words, nouns and adjectives, in a noun phrase

Table 2. Nouns and adjectives frequently used in the KSRP publications

Noun	Frequency	Adjective	Frequency
study	1,673	rural	1,930
village	1,604	agricultural	426
area	1,269	local	411
landscape	1,078	regional	264
development	936	spatial	221
analysis	773	urban	210
result	706	natural	200
project	629	important	179
resource	611	new	160
system	542	various	158
community	517	environmental	145
resident	452	economic	143
facility	447	public	138
planning	434	high	136
model	400	traditional	135
tourism	399	second	128
land	392	social	123
management	375	basic	121
method	374	necessary	120
environment	372	total	119

2. 연구방법

본 연구에서는 '농촌계획' 분야에서 통용되는 '개념'을 바탕으로 한 온톨로지 구축에 활용할 수 있는 단어 네트워크를 구성하기 위해 '어구' 단위에서의 단어 동시사용 빈도를 통해 NPMI를 계량하고 이로부터 단어 네크워트를 구성하는 연구방법을 적용하였다. 분석 텍스트에서 '어구', 그 중에서도 '명사구'를 추출하고 이들 '명사구'를 구성하는 명사와 형용사를 분석하기 위해 자연어처리기(Natural Language Process; NLP)인 OpenNLP™ (Baldridge, 2005)를 활용하였다. 이를 통해 분석대상 텍스트에서 '명사구'를 추출하고, 이들 '어구'를 구성하는 명사와 형용사의 사용빈도 및 동일 '어구'에서의 동시사용 빈도를 분석한다. 이들 빈도분석결과를 활용하여 단어 사이의 NPMI를 계량함으로써 단어 쌍의 연관성을 정량적으로 분석하게 된다. NPMI는 -1.0~1.0의 범위를 가지며 단어 쌍의 연관성이 높을 수록 1.0에 가까운 값을 가지게 되며, 이를 통해 단어 사이의 연관성의 많고 적음을 정량적인 수치를 통해 상대적으로 비교할 수 있게 된다. 분석된 단어 쌍의 NPMI를 바탕으로 연관성이 높은 단어 쌍을 활용하여 단어 네트워크를 구성하게 되는데, 본 연구에서는 사용빈도가 높은 상위 50개의 단어를 중심으로 단어네트워크를 구성하였다. 단어 네트워크 구성에 있어 NPMI가 높은 상위 단어 10개의 연관어를 네트워크 구성에 활용하였으며, 연관 단어의 경우 전체 텍스트 자료에서 사용빈도 5 이상, 전체 단어-단어 NPMI, 즉 연관도가 상위 5% 이내의 조건을 만족하는 단어만을 사용해 단어 네트워크를 구성한다. 이러한 연구과정은 Figure 1에 정리하였으며 빈도분석, NPMI 계산 등 연구과정에 필요한 도구는 JAVA™ 프로그래밍언어를 통해 프로그램을 작성하여 분석에 활용하였다.

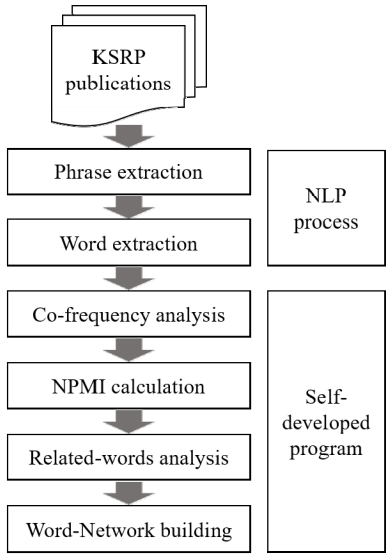

Figure 1. Process of KSRP ontology building

3. Normalized Pointwise Mutual Information (NPMI)

Fano(1961)가 제안한 '점별상호정보(PMI)'는 두 정보개체(entity) 사이의 관계를 계량하는 정보단위로서 '정보이론'(Shannon, 1948)에 기반을 두고 있다. 본 연구에서는 표기의 구분에 있어, 일반적으로 '점별상호정보'를 언급하는 경우에는 'PMI'로 표기하고, 개체 'x'와 'y'사이의 '정별상호정보' 값을 언급할 경우에는 'pmi(x,y)'로 구분하여 표기하기로 한다. pmi(x,y)는 두 개체의 관계를 정량화한다는 점에서 두 정보집합의 관계를 정량화하는 '상호정보(MI)'와 구분된다. 일반적으로 개체 x와 개체 y의 발생빈도로부터 각각의 발생 확률 p(x), p(y)와 동시발생확률 p(x,y)로 구성된 식(1)을 통해 산정한다.

$$pmi(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

pmi(x,y)는 식(1)에 의해 개체 x와 개체 y의 발생확률

에 따라, 또한 동시발생확률 p(x,y)에 따라 상대적인 값을 가지며, 각 개체의 발생확률 p(x), p(y)가 낮은 상태에서 동시발생확률 p(x,y)이 높을수록 높은 값을 가지게 되며, 값의 범위는 식(2)와 같다.

$$-\infty \le pmi(x,y) \le \min[-\log p(x), -\log p(y)] \quad (2)$$

동시발생 빈도가 '0' 이어서 p(x,y)이 '0'일 경우 pmi(x,y)는 $-\infty$ 로 산정되며, 개체 x와 개체 y가 독립인 즉 p(x,y)=p(x)·p(y)일 경우 pmi(x,y)는 '0'으로 산정된다. 개체 x혹은 개체 y가 상대 개체와 항상 같이 사용될 경우는 $-\log p(x)$ 와 $-\log p(y)$ 중에서 작은 값으로 산정된다.

PMI는 최대값이 두 개체의 빈도, 즉 발생확률에 영향을 받게 되어 일반적으로 발생빈도가 적은 개체들 사이에서 높은 값이 산정되는 특성이 있다. 따라서 본 연구에서와 같이 개체의 빈도차이가 큰 개체들 사이의 연관성을 분석하는 경우에는 이러한 값의 왜곡을 완화하기 위해 PMI를 정규화한 '정규화된 점별상호정보(NPMI)'를 적용하게 된다. NPMI는 식(3)과 같이 정의되며, 식(4)와 같이 일정한 범위의 값을 가지므로, 단어 쌍의 NPMI 수치 간에 상대적인 비교가 가능하게 된다.

$$npmi(x,y) = \frac{pmi(x,y)}{-\log p(x,y)} \quad (3)$$

$$-1.0 \le npmi(x,y) \le 1.0 \quad (4)$$

npmi(x,y)는 개체 x와 개체 y의 동시발생 빈도가 '0'일 경우 -1.0, 개체 x와 개체 y가 독립인 p(x,y)=p(x)·p(y)인 경우 '0', 그리고 개체 x혹은 개체 y가 상대 개체와 항상 같이 사용될 경우는 1.0으로 산정된다. 이러한 'NPMI'의 특성을 단어의 통계적인 특성분석에 활용하여 각 단어의 어구 내 사용빈도와 단어 쌍의 동시사용빈도를 식(3) 적용한다면 단어와 단어 사이의 연결성을 비교 가능한 정량적인 값으로 수치화 할 수 있다. 본 연구에서는 식(3)의 'NPMI'를 통해 단어 간 연관성을 정량화하였다.

IV. 연구결과

1. 단어 간 NPMI

본 연구에서는 단어 쌍의 NPMI를 계량함으로써 단어 사이의 연관성을 정량화하였다. 단어 사이의 NPMI 수치가 높을수록 연관성이 높은 것으로 해석하였으며, NPMI가 상위 5% 이내에 드는 단어 관계를 연관성을 가지는 단어 쌍으로 해석하였으며 최대 연관어는 10개로 한정하였다. '한국농촌계획학회' 논문 785편에서 사용빈도가 높은 것으로 분석되어 Table. 2에 정리하였던 명사와 형용사에 대하여 연관성이 높은 것으로 분석된 단어 분석결과를 각각 Table. 3과 Table. 4에 나타내었다.

명사의 분석결과를 나타낸 Table. 3의 결과를 살펴보면, '연구(study)'의 경우 일반적으로 논문과 연관되어 사용하는 '사례(case)', '분석(analyze)', '실증(empirical)' 등과 연관성이 높은 것을 나타났다. '분석(analysis)'와 연관성이 높은 단어는 '회귀(regression)', '비교(comparative)', '빈도(frequency)', '군집(cluster)' 등으로 분석되어 한국농촌계획학회의 논문에서 적용 비중이 높은 '분석'방법을 나타내는 것으로 판단하였다. 이와 비슷하게 '방법(method)'과 연관성이 높은 단어는 't-test', '가치평가(valuation)', '점수산정(scoring)', '설문조사(surveying)' 등으로 나타나 본 학회에서 많이 사용하고 있는 분석'방법'이 연관어로 도출되었다. 연관성 분석결과가 한국농촌계획학회 고유의 개념 간 연결고리를 나타내기도 하였는데 '마을(village)'의 경우 '인구감소(depopulation)'이 연관성이 높은 것으로 분석되어 본 학회에서는 마을의 인구문제에 관해 비중 있게 다루어지고 있는 것을 확인할 수 있었다. 또한 '개발(development)'는 '균형잡힌(balanced)', '지속가능한(sustainable)'과 연관성이 높은 것으로 분석되어 개발에 관한 본 학회의 논문에서 '개발'의 개념과 연관된 단어를 확인할 수 있다. 뿐만 아니라 '자원(resource)'는 '어메니티(amenity)'와 높은 연관성을 나타내 본 학회에서 '농촌어메니티자원'관련 논문을 통해 주요하게 다루어지는 개념 간의 연결을 보여주고 있다. 이밖에도 '지역(area)'는 '도시화(urbanized)'와 연관성이 높게 나타나는 등 NPMI를 활용한 연관성 분석을 통해 '한국농촌계획학회' 논문에서 각 개념들이 어떻게 연결되어 있는지를 나타내는 자료를 도출할 수 있었다.

형용사의 분석결과를 나타낸 Table. 4의 결과를 살펴보면, '경제의(economic)'과 연관성이 높은 단어는 '성장(growth)', '가치(value)', '실행가능성(feasibility)' 등과 같이 일반적으로 '경제'와 연관성이 있는 단어가 도출되었다. '농업의(agricultural)'의 경우 '생산(product)'와 같이 일반적인 연관어가 분석되기도 하였으나 '가뭄(drought)', '특산품(specialty)'와 같이 본 학회의 관심사를 나타내는 단어가 연관성이 높은 것으로 분석되기도 하였다. 본 학회와 연관성이 높은 '농촌(rural)' 단어의 경우 '마을(village)', '어메니티(amenity)', '관광(tourism)', '개발(development)'과 같은 단어와 높은 연관성을 보여 본 학

회의 농촌에 관한 관심사항을 나타내주고 있는 것으로 판단하였다. 또한 '도시(urban)'의 경우 지역 '확장'과 연관된 단어인 'sprawl', 'fringe', 'expansion' 등이 연관성이 높은 것으로 나타나 '한국농촌계획' 분야에서 '도시'와 관련한 개념의 특성을 확인할 수 있었다. '지역적(regional)'과는 '불평등(inequality)', '안전(safety)', '활성화(vitalization)'의 단어가, '공간적(spatial)'과는 '계량경제학의(econometric)', '구조(structure)', '교류(interaction)'의 단어가 연관성이 높은 것으로 분석되는 등, 형용사의 경우에도 NPMI를 활용한 연관성 분석을 통해 '한국농촌계획' 분야에서 통용되는 개념을 바탕으로한 연관어 분석을 수행할 수 있었다.

2. 단어네트워크 비교: KSRP vs. WordNet

Table 3. Related words with nouns in KSRP

Noun	Related words (NPMI)
study	case(0.67), present(0.47), previous(0.45), depopulation(0.41), show(0.39), selected(0.36), analyze(0.36), future(0.35), further(0.35), empirical(0.34)
village	fishing(0.47), grove(0.39), waterfront(0.39), mountain(0.39), seaside(0.37), eco(0.37), marketplace(0.36), over-depopulated(0.36), rural(0.35), appraisal(0.30)
area	urbanized(0.41), less-favored(0.40), favored(0.38), reserve(0.37), inland(0.37), biosphere(0.36), mountainous(0.35), rural(0.35), greenbelt(0.34), metropolitan(0.33)
landscape	prototype(0.46), prototypal(0.45), scenic(0.38), ordinance(0.36), draft(0.35), architecture(0.35), composition(0.34), conservation(0.33), continuity(0.33), superior(0.33)
development	endogenous(0.49), permission(0.43), balanced(0.43), project(0.43), comprehensive(0.42), sustainable(0.41), bottom-up(0.37), hot-spring(0.35), eco(0.30)
analysis	regression(0.53), comparative(0.47), cluster(0.47), frequency(0.44), shift-share(0.44), importance-performance(0.43), descriptive(0.43), multiple(0.42), principal(0.42), interpretation(0.41)
result	burden(0.43), following(0.42), above(0.38), analysis(0.37), estimation(0.34), test(0.32), assessment(0.31), empirical(0.31), main(0.31), best(0.30)
project	community-building(0.45), Saemaul(0.44), comprehensive(0.43), development(0.43), art(0.40), bottom-up(0.38), modernization(0.38), reclamation(0.36), empowerment(0.36)
resource	amenity(0.60), societal(0.40), cultural(0.39), human(0.38), indigenous(0.37), intangible(0.37), superior(0.37), natural(0.34), ecocultural(0.32), analyzing(0.32)
system	incentive(0.46), agricultural-product(0.44), propulsion(0.43), web(0.43), forecasting(0.43), web-based(0.43), tentative(0.42), week(0.41), voucher(0.41), information(0.39)
community	center(0.46), mongolica(0.46), youth(0.46), finder(0.44), revival(0.42), eupmyun(0.42), viable(0.41), spirit(0.38), close(0.35), pension(0.35)
resident	local(0.52), regard(0.48), consciousness(0.44), original(0.44), respect(0.41), participatory(0.35), leader(0.35), urbanites(0.34), visitor(0.32), corresponding(0.32)
facility	sanitation(0.48), disposal(0.46), convenience(0.44), legislation(0.43), protected(0.43), hardware(0.42), waste(0.42), medical(0.38), unused(0.36), public(0.36)
planning	collaborative(0.56), process(0.41), strategic(0.39), rational(0.37), recovery(0.37), storytelling(0.36), longterm(0.32), use(0.32), ordinance(0.32)
model	logit(0.60), econometric(0.56), input-output(0.54), gravity(0.54), mediation(0.52), multi-level(0.52), probit(0.51), mushroom(0.48), equation(0.45), additive(0.44)
tourism	green(0.52), fair(0.50), intermediary(0.46), healing(0.44), motivation(0.41), experiential(0.40), eco-cultural(0.37), storytelling(0.34), rural(0.32), wildflower(0.29)
land	use(0.68), cover(0.57), suitability(0.51), fencing(0.51), consolidation(0.49), cultivated(0.49), price(0.48), tidal(0.45), Saemangeum(0.44)
management	initiative(0.47), conflict(0.46), police(0.42), fringe(0.41), diversification(0.41), ability(0.39), pesticide(0.38), efficient(0.35), intensive(0.34), official(0.32)
method	valuation(0.56), contingent(0.55), decomposition(0.46), multi-dimensional(0.46), scoring(0.42), surveying(0.40), improved(0.39), move(0.39), t-test(0.39)
environment	natural(0.52), living(0.49), diagnostic(0.42), clean(0.41), ecology(0.40), friendly(0.39), pleasant(0.39), indoor(0.39), context(0.38), countryside(0.37)

본 연구에서 NPMI를 통해 '한국농촌계획학회' 논문을 분석하여 구축한 단어 네트워크(이하 KSRP 네트워크)와 사전데이터의 시소러스(thesaurus)를 통해 구성한 단어 네트워크의 차이를 살펴보기 위해 Figure 2와 Figure 3에 각각 본 연구에서 구축한 단어 네트워크와 대표적인 시소러스인 WordNet™(Miller, 1995)을 통해 구축한 단어 네트워크(이하 WordNet 네트워크)를 나타내고 이를 비교하였다. 비교를 위한 네트워크 구축에 있어 연결(link) 차수는 2차로 한정하였으며, '농촌(rural)', '도시(urban)', '계획(planning)', '자원(resource)', '개발(development)'의 5가지 단어에 대한 소규모 네트워크를 비교하였다.

KSRP 네트워크에서는 '농촌(rural)'이 1차 연결에서는 '마을(village)', '어메니티(amenity)', '관광(tourism)' 등과 연결되어 2차 연결에서는 '보존(reserve)', '녹색(green)', '지속가능한(sustainable)', '숲(grove)', '자원(resource)' 등과 연결된 반면 WordNet 네트워크에서는 1차 연결에서

Table 4. Related words with adjectives in KSRP

Noun	Related words (NPMI)
rural	village(0.35), area(0.35), amenity(0.34), tourism(0.32), minor(0.30), water-front(0.30), development(0.30), marketplace(0.30), green-village(0.28)
agricultural	product(0.59), machinery(0.54), heritage(0.48), agrimanufacturing(0.42), alpine(0.42), legislation(0.40), lease-holders(0.40), highschool(0.39), drought(0.39), specialty(0.37)
local	government(0.59), resident(0.52), strengthening(0.51), economy(0.50), innovation(0.50), archive(0.47), autonomous(0.47), autonomy(0.46), capability(0.42), authority(0.42)
regional	inequality(0.51), disparity(0.48), revitalization(0.41), preparedness(0.41), safety(0.40), inputoutput(0.40), dependency(0.39), vitalization(0.38), commercialization(0.38), endogenous(0.38)
spatial	econometric(0.57), extent(0.55), autocorrelation(0.51), structure(0.50), interaction(0.46), distribution(0.45), diffusion(0.44), allocation(0.43), depth(0.43), dependency(0.42)
urban	sprawl(0.58), fringe(0.57), expansion(0.52), rank(0.47), urbanites(0.44), initiative(0.44), apartment(0.41), architectural(0.37), farming(0.36), people(0.35)
natural	environment(0.52), hazard(0.47), beauty(0.46), purification(0.45), societal(0.40), recreation(0.39), adjustment(0.38), manmade(0.38), monument(0.38), artificial(0.37)
important	clue(0.54), criterion(0.53), factor(0.46), role(0.44), component(0.42), preliminary(0.42), issue(0.41), traits(0.40), question(0.36), necessity(0.34)
new	act(0.54), methodology(0.46), expressway(0.42), paradigm(0.40), formulation(0.39), attempt(0.39), kind(0.39), finding(0.38), protected(0.36), manufacturing(0.36)
various	fact(0.51), kind(0.46), desire(0.44), therapy(0.37), flooding(0.37), background(0.35), intangible(0.35), format(0.35), tangible(0.35), menu(0.33)
environmental	pollution(0.53), sanitation(0.52), color(0.52), formative(0.44), issue(0.42), friendliness(0.41), protection(0.41), domain(0.38), consciousness(0.38), loading(0.38)
economic	huge(0.52), growth(0.46), vitalization(0.45), validity(0.45), burden(0.45), impact(0.43), feasibility(0.42), value(0.41), satisfied(0.40), social(0.38)
public	partnership(0.54), private(0.51), commercial(0.50), transportation(0.50), utility(0.48), official(0.43), space(0.41), interest(0.40), benefit(0.40), worker(0.40)
high	affinity(0.48), resolution(0.47), red(0.46), pepper(0.42), school(0.42), medium(0.41), emotional(0.39), locality(0.37), connectivity(0.37), salt(0.37)
traditional	culture(0.51), fence(0.46), food(0.45), dwelling(0.42), stone(0.39), scenery(0.38), floor(0.37), art(0.37), epistemology(0.37), era(0.35)
second	round(0.65), quadrant(0.54), friendliness(0.52), home(0.50), step(0.50), phase(0.42), first(0.37), cause(0.37), global(0.37), group(0.36)
social	learning(0.52), capital(0.51), economical(0.47), interaction(0.43), network(0.42), potentiality(0.42), competence(0.42), cultural(0.40), welfare(0.39), domain(0.39)
basic	principle(0.45), data(0.44), programming(0.42), functionality(0.39), material(0.39), information(0.38), authority(0.38), concept(0.36), topographic(0.32), word(0.31)
necessary	improving(0.64), part(0.53), unit(0.44), facility(0.35), period(0.35), component(0.32), role(0.31), maintenance(0.31), condition(0.28), program(0.27)
total	nitrogen(0.64), coliform(0.60), sum(0.54), phosphorus(0.53), length(0.51), biotop(0.48), number(0.48), variance(0.44), meter(0.43)

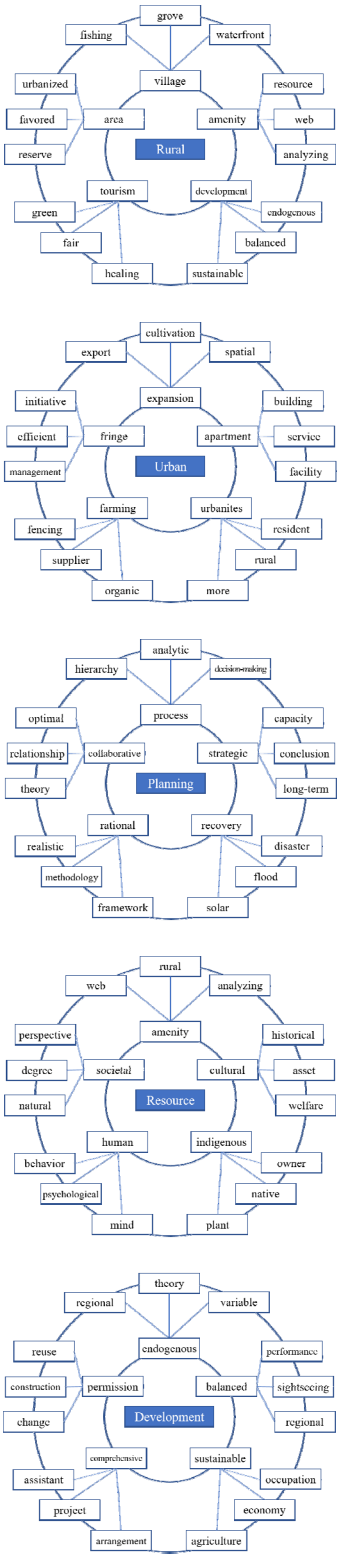

Figure 2. Related words in KSRP publications, with 'rural', 'urban', 'planning', 'resource' and 'development'

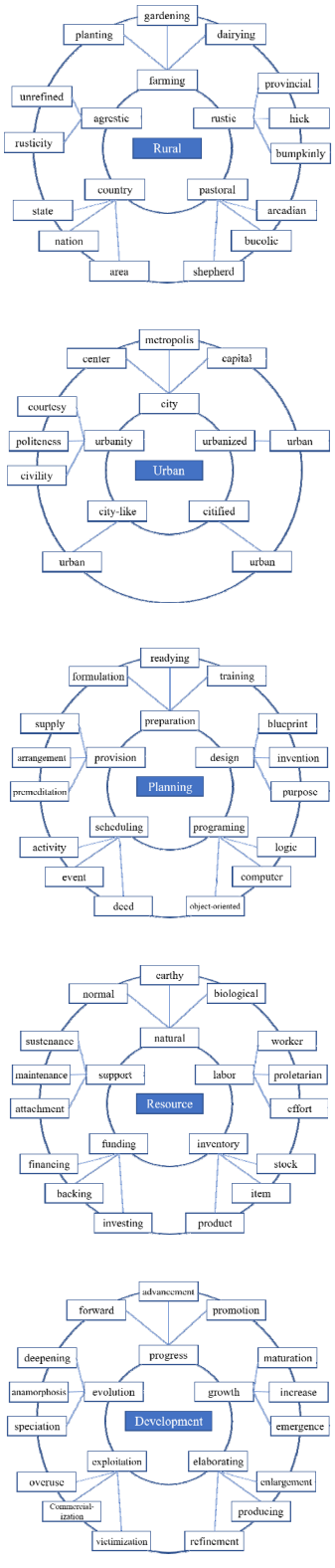

Figure 3. Related words in WordNet™, with 'rural', 'urban', 'planning', 'resource' and 'development'

‘영농(farming)’, ‘시골(rustic)’, ‘촌스러움(agrstic)’ 등과 연결되고 2차 연결에서는 ‘축산(dairying)’, ‘지방의(provincial)’, ‘세련되지 않은(unrefined)’, ‘전원의(bucolic)’ 등과 연결되었다. 후자의 결과를 통해, 일반적인 관점에서는 농촌에 대한 개념이 농업과 연결되거나 전원과 같이 한적한 곳, 혹은 촌스럽고 세련되지 못한 것과 연관되어 있음을 확인할 수 있었다. 이와 달리 ‘농촌계획’ 분야의 관점에서는 농촌에 대한 개념이 보존해야할 공간이나 어메니티와 같은 자원이 있는 곳, 숲이나 수변이 있는 마을, 혹은 균형잡히고 지속가능한 내생적 발전 등과 연결되어 있는 것으로 나타났다. ‘도시(urban)’ 네트워크에서도 이러한 차이는 발견되었는데, WordNet 네트워크에서는 ‘도시(city)’, ‘중심지(center)’, ‘대도시(metropolis)’, ‘수도(capital)’, ‘세련(urbanity)’ 등이 ‘도시(urban)’과 연관되어 있었으나 ‘농촌계획’ 분야 네트워크에서는 ‘확장(expansion)’, ‘영농(farming)’, ‘관리(management)’, ‘시설

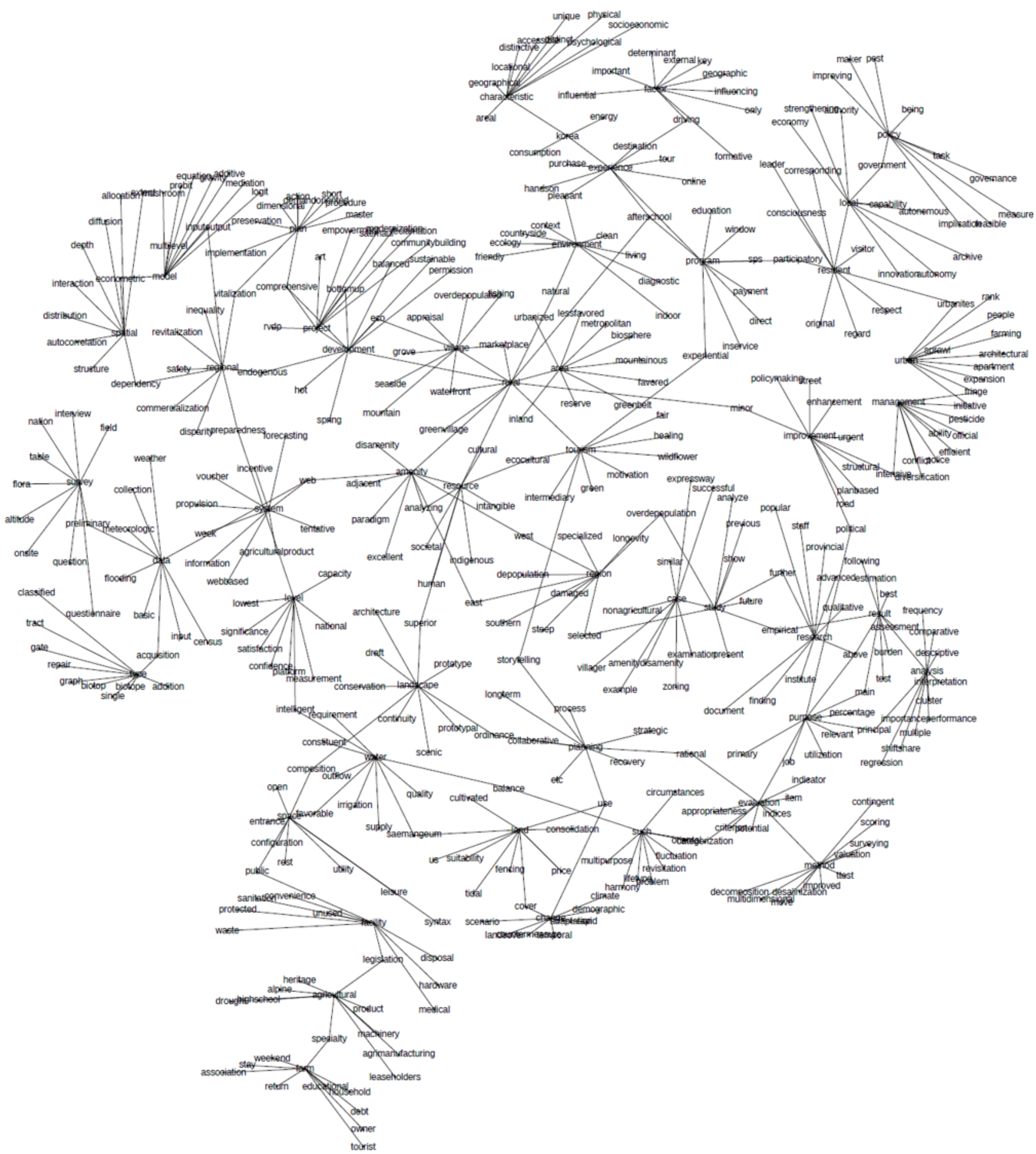

Figure 4. Word network in KSRP publication, built by NPMI

(facility)', '유기농(organic)' 등이 '도시(urban)'과 관련되어 있는 것으로 나타났다. 이와 같이, '한국농촌계획학회'의 논문을 분석하여 구축한 단어 네트워크는 일반적인 사전데이터의 시소러스로 구성한 단어 네트워크와 구분되는 '농촌계획' 분야 고유의 개념에 기반을 둔 단어 연결 관계를 나타내고 있는 것으로 판단하였다.

이외에도 '계획(planning)'이 WordNet 네트워크에서는 '준비(preparation)'-'공식화(formulation)', '일정관리(scheduling)'-'활동(activity)', '설계(design)'-'청사진(blueprint)' 등과 연결되는 반면 KSRP 네트워크에서는 '의사결정(decision-making)'-'과정(process)', '장기(long-term)'-'전략(strategic)', '공동의(collaborative)'-'관계(relationship)' 등으로 연결되어 '농촌계획' 분야에서 보다 주요하게 다루어지는 개념 간 연결을 나타내고 있는 것으로 판단하였다. '자원(resource)'의 경우도 WordNet 네트워크에서는 '자연(natural)'-'생물학적(biological)', '노동(labor)'-'노동자(worker)', '자금(funding)'-'투자(investing)' 등의 일반적인 개념 간 연결을 나타내었으나, KSRP 네트워크에서는 '농촌(rural)'-'어메니티(amenity)', '문화적(cultural)'-'역사적(historical)', '사람(human)'-'행동(behavior)' 등과 연결이 되어 '농촌계획' 분야에서 '자원'의 개념과 연관된 개념을 나타내었다. '개발(development)'에서도 비슷한 결과가 도출되어 WordNet 네트워크에서는 '진전(progress)'-'발전(advancement)', '성장(growth)'-'증가(increase)', '개발·착취(exploitation)'-'상업화(commercialization)'과 같은 연결이 도출되었으나, KSRP 네트워크에서는 '내생적(endogenous)'-'지역적(regional)', '균형잡힌(balanced)'-'지역적(regional)', '종합적인(comprehensive)'-'프로젝트(project)' 등의 '농촌계획' 분야에서 '개발'을 다루는 관점이 반영단 단어들로 네트워크가 구성된 것을 발견할 수 있었다.

이상의 비교결과를 통해 '한국농촌계획학회'의 논문을 분석하여 구축한 단어 네트워크를 통해 일반적인 연관 단어 분석을 통해서는 구축할 수 없는 '한국농촌계획' 분야 고유의 개념 간 연결을 보여주는 단어 네트워크를 구성하였음을 확인할 수 있었다.

3. KSRP 단어네트워크

본 연구에서 분석한 '한국농촌계획학회' 논문 내 단어 간 '정규화된 점별 상호정보(NPMI)'를 통해 구축한 단어 네트워크를 Figure. 4에 나타내었다. 네트워크 구축 프로그램으로는 Gephi^{TM}(Bastian, et al. 2009)를 사용했으며, 강하게 연결된 대상은 가까이 표시하고 약하게 연결된 대상은 상대적으로 멀게 표시하는 중력공식을 이용한 'ForceAtlas2'(Jacomy et al., 2014) 알고리즘을 이용하여 네트워크를 도식화하였다. 사용빈도 상위 50위의 단어와 연결된 단어로 구성한 네트워크를 구축하였으며, 단어 네트워크 구성에 있어 노드(node)는 전체 논문 자료에서 사용빈도가 5 이상인 단어로만 구성하고 연결(edge)은 NPMI 상위 5% 이내에 드는 연결만을 활용하였다. 각 단어별 연관 단어 구성에 있어 최대 연결 단어 수는 10개로 한정하였다.

V. 결 론

'3차 산업혁명'의 정보화시대를 거쳐 인공지능, 빅데이터, 사물인터넷, 시멘틱웹 등으로 대표되는 '4차 산업혁명'에 대한 대비로 분주해진 시대를 맞고 있다. 정보와 관련된 여러 저장 및 처리기술의 통합과 이의 광범위한 영역에 대한 활용이 '4차 산업혁명'을 가능하게 하고 있으나 그 기저에는 온톨로지를 통한 여러 정보체계의 통합과 기존에 구축된 정보자원의 재활용이 뒷받침이 되어야 한다. 이는 '4차 산업혁명'을 준비하기 위해서는 여러 분야에 독립적으로 축적되어 오던 지식이 온톨로지를 통해 하나의 유기적인 정보체계로 연결되어야 함을 의미한다. 따라서 '4차 산업혁명'에 대응하는 과정에서 각 분야 고유의 온톨로지를 구축하고 이를 전체 온톨로지에 통합할 수 있는 표준언어로 구현하는 작업이 무엇보다 중요하다. 이는 본 한국농촌계획학회에도 동일하게 작용되며, 농촌계획 분야와 같이 기존에 축적된 유용한 정보자원이 많은 분야에서는 고유 온톨로지의 구축이 더욱 필요하다고 할 수 있다. 본 연구에서는 이러한 '한국농촌계획' 분야 고유의 온톨로지 구축에 활용할 수 있는 단어 네트워크를 구축하고자 하였다. '한국농촌계획' 분야에서 통용되는 개념을 바탕으로 단어 네트워크를 구축하고자 '한국농촌계획학회'에 등재된 논문 자료를 바탕으로 단일 개념을 담고 있는 '어구' 단위의 단어사용 통계적 특성을 분석하고, 이를 정보이론에 등장하는 '상호정보'에 대입하여 단어 간 연관성을 정량화하고, 산정한 연관성을 바탕으로 단어 네트워크를 구축하였다. 본 연구의 과정과 그 결과를 요약하면 다음과 같다.

'한국농촌계획학회'에 등재된 논문 785편으로부터 영문 제목, 초록, 키워드를 추출하고 이 텍스트 자료로부터 자연어처리기(NLP)를 활용하여 명사와 형용사를 추출하였다. 이 과정에서 '개념'을 바탕으로 단어와 단어 사이의 연관성을 분석하고자 명사구에 대한 정보도 함께 추출하였다. 각 단어에 대하여 어구 단위 단어사용 빈도와 단어 간 동시사용 빈도를 분석하고 이 자료를 정보이론

의 '정규화된 점별 상호정보(NPMI)'에 적용함으로써 단어와 단어 사이의 연관성을 정량화 하였다. 산정된 단어-단어 간 연관성 수치를 적용하여 단어 네트워크를 구성하였다. 이러한 과정을 통해 본 한국농촌계획학회 논문 내에서 사용된 단어들의 어구 단위 사용빈도와 단어 쌍의 동시사용빈도를 분석할 수 있었으며, 이들 수치를 적용하여 산정한 NPMI를 통해 단어 사이의 연관성을 정량화함으로써 연관어 분석을 수행할 수 있었다. 분석한 연관어를 바탕으로 한국농촌계획학회의 단어 네트워크를 구성할 수 있었으며, 이를 사전데이터의 연관어로 구성된 WordNet™을 통해 구성한 단어 네트워크와 비교함으로써 본 연구에서 구축한 단어 네트워크가 일반적인 연관어 기반의 단어 네트워크와 구분되는 고유한 개념 간 연결관계를 보여줌을 확인할 수 있었다. 본 연구에서 구축한 단어네트워크에서는 '농촌(rural)'이 '어메니티(amenity)'나 '관광(tourism)'과 연결되고, '개발(development)'이 '종합적(comprehensive)', '지속가능한(sustainable)'과 연결되는 등 농촌계획 분야에서 '농촌'이나 '개발'이 어떠한 개념과 연결되어 있는지가 본 연구의 결과를 통해 드러났으며, 이를 통해 농촌계획 분야 고유의 온톨로지 구축에 활용할 수 있는 단어 연관관계가 구축되었음을 확인할 수 있었다.

특정 분야의 온톨로지를 구축하기 위해서는 있어 해당 분야에서 공유하는 개념을 바탕으로 서로 연관된 단어 쌍을 분석하고 이들이 연결된 네트워크를 구축하는 작업이 선행되어야 한다. 본 연구를 통해 한국농촌계획학회 785편의 논문이 공유하고 있는 단어와 단어 사이의 잠재적인 의미관계를 분석하고 이 중 연결성이 강한 단어 쌍을 통해 단어 네트워크를 구축할 수 있었다는 점, 이를 통해 농촌계획 분야 온톨로지 구축을 위한 기초자료를 마련할 수 있었다는 점에서 본 연구의 의의가 있을 것으로 판단하였다. 또한 기존에 구축된 정보자원을 활용함으로써 해당 분야의 개념 간 연결에 관한 지식체계를 구축할 수 있었다는 점에서도 본 연구의 의의가 있는 것으로 판단하였다. 이러한 단어 네트워크는 향후 온톨로지 구축에 활용할 수 있으며, 기 구축된 온톨로지와 결합함으로써 농촌계획 분야 고유의 지식체계를 타 분야 지식체계에 반영할 수 있다는 점에서 '4차 산업혁명'에 대비하는 기초자료로 활용할 수 있을 것으로 기대된다. 또한 '농촌계획' 단어 네트워크 및 온톨로지와 같은 고유의 지식체계는 맞춤형 농촌정보 검색과 같은 기초적인 활용에서부터 인공지능을 활용한 농촌계획 관련 정보의 수집 및 가공 등 다양한 정보기술 분야에 활용할 수 있을 것으로 기대한다.

다만 본 연구의 결과를 통해 구성할 수 있는 온톨로지는 기본적인 개념 간 연결을 나타내는 기초적인 단계에 머무르고 있어 보다 완전한 온톨로지 구성을 위해서는 세부적인 관계설정에 대한 분석이 필요하다. 본 연구의 목적은 기초적인 온톨로지 구축에 활용할 수 있는 개념 기반 단어 연관어 분석에 있으므로 구체적인 단어 연결의 속성을 분석하지는 않았으나 이러한 부분은 추후 연구를 통해 보완될 수 있을 것으로 기대한다.

References

1. Baldridge, J., 2005, The opennlp project. URL:http://opennlp.apache.org/index.html.
2. Bastian, M., Heymann, S. and Jacomy, M., 2009, Gephi: an open source software for exploring and manipulating networks, International AAAI Conference on Weblogs and Social Media.
3. Benjamins, P.V., Fensel, D. and Gómez-Pérez, A., 1998, Knowledge Management through Ontologies, International Conference on Practical Aspects of Knowledge Management (PAKM-98), pp.29-30.
4. Berneers-Lee, T., 2001, The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, 284(5), pp.34-43.
5. Bouma, G., 2009, Normalized (Pointwise) Mutual Information in Collocation Extraction, Proceedings of the Biennial GSCL Conference, pp.31-40.
6. Cambrosio, A., Limoges, C., Courtial, J.P. and Lavile, F. 1993, Historical scientometrics? Mapping over 70 years of biological safety research with coword analysis, Scientometrics, 27(2), pp.119-143.
7. Choi, Y., 2017, Legal Issues of Regional Informatization in the Fourth Industrial Revolution, The Journal of Public Policy & Governance, 10(4), pp.35-57.
8. Choi, Y.J., Choi, S.J., 2016, Analysis of Buyer's Behavioral Difference by Types of Online Shopping - Utilizing online sales data of Yangpyeong Sumi Village -, Journal of Kyonggi Tourism Research, 26, pp.49-66.
9. Church K.W. and Hanks P., 1990, Word association norms, mutual information, and lexicography, Computational Linguistics, 16(1), pp.22-29.
10. Damani, O.P., 2013, Improving Pointwise Mutual

Information (PMI) by Incorporating Significant Co-occurrence, Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp.20–28.
11. Ding, Y., Chowdhury, G.G. and Foo, S., 2001, Bibliometric cartography of information retrieval research by using co-word analysis, Information Processing & Management, 37(6), pp.817-842.
12. Fano, R.M., 1961, Transmission of Information: A Statistical Theory of Communication, The MIT Press. pp.21-61.
13. Grangel-González, I., Halilaj, L., Coskun, G., Auer, S., Collarana, D. and Hoffmeister, M., 2016, Towards a Semantic Administrative Shell for Industry 4.0 Components, 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp.230-237.
14. Gu M.S., Hwang J.H., Ryu, K.H. and Hong J.E., 2006, Semi-Automatic Ontology Generation about XML Documents using Data Mining Method, The KIPS Transactions : Part D, 13(3), pp.299–308.
15. Hyun, S.H. and Ham, Y.S., 2017, Analysis on the Effects to Local Tax Income according to Regional Industrial Structure's Productivity: Focused on Local governments in GyeongGi Province, The Korean Journal of Local Government Studies, 20(4), pp.25-45.
16. Hyvönen, E., Saarela, S., and Kim V., 2003, Ontogator: combining view- and ontology-based search with semantic browsing, Proceedings of XML Finland 2003, Open Standards, XML, and the Public Sec-tor, Kuopio.
17. Jacomy, M., Venturini, T., Heymann, S. and Bastian, M., 2014, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS ONE, 9(6): e98679, https://doi.org/10.1371/journal.pone.0098679
18. Kang, S.J., 2004, Ontology Construction and Its Application to Disambiguate Word Senses, The KIPS transactions. Part B, 11(4), pp.491-500.
19. Kim, H.S., Choi, I. and Kim, M., 2005, A Statistical Approach for Extracting and Naming Relation between Concepts, The KIPS(Korea Information Processing Society) Transactions : Part B, 12(4), pp.479-486.
20. Lawrie D., Croft B., and Rosenberg, A., 2001, Finding topic words for hierarchical summarization, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM, New York, NY, USA, pp.349-357.
21. Lee Y., 2016, Dynamic ontology construction algorithm from Wikipedia and its application toward real-time nation image analysis, Journal of the Korean Data & Information Science Society, 27(4),pp.979-991.
22. Lee, H.J., Lee, J.M., Park, M.J., Kim, H.J., Lee, J.J., 2006, Development of Rural Amenity resrouces Information System Using Ontology and Web-GIS, Journal of the Korean Society of Rural Planning, 12(4), pp.13-22.
23. Lee, H.R., 2009, Implementation of Information Retrieval and Management System Based on Ontology Using Object Oriented Design Pattern, Journal of the Korean Association of Geographic Information Studies, 12(4), pp.146-157.
24. Lee, H.R., Baek, J.H., Baek, J.H., 2008, Prototype of Crops Information System based on Ontology and WebGIS, Journal of the Korean Association of Geographic Information Studies, 11(3), pp.43-51.
25. Lee, J., Kim, Y., Shin, H. and Song, K., 2014, A Study on Ontology Based Knowledge Representation Method with the Alzheimer Disease Related Articles, Journal of Internet Computing and Services, 15(3), pp.125-135.
26. Lee, J.M., Lee, J.J., 2006, Integration with External Information Using Ontology for Rural Amenity Resources Information Service, Journal of the Korean Society of Rural Planning, 12(4), pp.53-61.
27. Lee, J.M., Suh, K., Kim, H.J., Lee, J.J., 2005, Design of Integrated Database Schema for Imptoving Usability of Rural Information, Journal of the Korean Society of Rural Planning, 11(2), pp.43-49.
28. Lee, S.H., 2010, Development of Subsurface Spatial Information Model with Cluster Analysis and Ontology Model, Journal of the Korean Association of Geographic Information Studies, 13(4), pp.170-180.
29. Maedche, A. and Staab, S., 2000, Semi-Automatic Engineering of Ontologies from Text , In Proceedings of the 12th Internal Conference on Software and Knowledge Engineering, pp.231-239.
30. Mashey, J.R., 1998, Big Data ... and the Next Wave

of InfraStress, Usenix, Slides from invited talk.
31. Miller, G.A., 1995, WordNet: a lexical database for English, Communications of the ACM, 38(11), pp.39-41.
32. Ministry of Agriculture, Food and Rural Affairs (MAFRA, 농식품부), 2017a, High-quality Bigdata-map construction for the fourth industrial revolution of agriculture and food sector(in Korean: 농식품분야 4차혁명 위해 고품질 빅데이터 지도 구축), A press releas, MAFRA.
33. Ministry of Agriculture, Food and Rural Affairs (MAFRA, 농식품부), 2017b, Application of public databasse of agriculture and food, lead to new foundation in the era of the fourth industrial revolution (in Korean: 농식품 공공데이터 활용, 4차 산업 혁명시대 새로운 창업 선도), A press releas, MAFRA.
34. Ministry of Science, ICT and Future PLanning(MSIP), 2016, A comprehensive mid- and long-term plan of inteligence and information society for the fourth industrial revolution (in Korean: 제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책), MISP and ministries concernced.
35. Mun, H.J. and Woo, Y.T., 2006, Concept Extraction Technique from Documents Using Domain Ontology, The KIPS Transactions : Part D, 13(3), pp.309-316.
36. Nam, W.H., Kim, T., Choi, J.Y., Kim, J.T., La, M.C., 2011, Wireless Sensor Network Development using RFID for Agricultural Water Management, Journal of the Korean Society of Agricultural Engineers, 53(5), pp.43-51.
37. Park, J.G. and Chang, Y.C., 2001, A Contruction of Spatial Database for Precision Farming, Kon-Kuk Journal of Natural Science and Technology, 12, pp.61-72.
38. Park, J.H., Hwang, J.H., Lee, S.W., 2014a, The effect of the 6th industrialization in agriculture on farm and off-farm income, Journal of the Korean Society of Rural Planning, 20(4), pp.193-208.
39. Park, M., Kim, S.B., Kim, E.J., Rhee, S., Song, Y., Lim, C.S., Choi, J.A. and Chin, H.S., 2014b, The Current State of the Korean Rural Amenity Resource Database, Journal of the Korean Society of Rural Planning, 20(4), pp.263-276.
40. Role F. and Nadif, M., 2011, Handling the Impact of Low Frequency Events on Co-occurrence based Measures of Word Similarity - A Case Study of Pointwise Mutual Information, Proceedings of KDIR 2011 : KDIR- International Conference on Knowledge Discovery and Information Retrieval, pp.226-231
41. Sanderson, M. and Croft, B., 1999, Deriving concept hierarchies from text, In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99), ACM, New York, NY, USA, pp.206-213.
42. Schwab, K., 2015, The Fourth Industrial : what it means, how to respond, Foreign Affairs, Accessed July 6. 2017. from https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution.
43. Shannon, C.E., 1948, A Mathematical Theory of Communication, The Bell System Technical Journal, 27, pp.379-423.
44. Shim, J.H. and Choi, M.G., 2013, An Analysis of the Intellectual Structure of Venture-Creation Studies to build an Entrepreneurship Ontology, Knowledge Management Research, 14(4), pp.75-86.
45. Song, D.G., 2005, A Study of Methodology for Automatic Construction of OWL Ontologies from Sejong Electronic Dictionary, Language and Information, 9(1), pp.19–34.
46. Staab S., Studer, R., Schnurr, H.P. and Y. Sure, 2001, Knowledge processes and ontologies, IEEE Intelligent Systems, 16(1), pp. 26-34.
47. Thomas R.G., 1993, A Translation Approach to Portable Ontology Specifications, Stanford Knowledge System Laboratory Technique Report KSL-92-71, pp.1-2.
48. Uschold, M. and Gruninger, M., 1996, Ontologies: Principles, methods and applications, Knowledge Engineering Review, 11(2), pp.93-136.
49. Weinstein, P.C., 1998, Ontology-based metadata: transforming the MARC legacy, In Proceedings of the third ACM conference on Digital libraries (DL '98), pp.254-263.
50. Yang, J.I., Lee, J.H., Hwang, D.Y., 2014, Empirical Study on the Activation Plan for 6th Industrialization of Rural Agricultural Resources: Focus on the Field Experts and the Complementary Demand, Journal of the Korean Society of Rural Planning,20(3),

pp.111-120.

51. Yang, K.A., Yang, H.J., Yang, J.D., 2004, Bio-Ontology Generation Using Object-Oriented Ontology Manager, The KIPS transactions. Part B, 11(4), pp.437-448.
52. Yarowsky D., 1995, Unsupervised word sense disambiguation rivaling supervised methods, In Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL '95), Association for Computational Linguistics, Stroudsburg, PA, USA, pp.189-196.
53. Yoo, S.B. and Yoon, H.K., 2000, Searching River Information using Ontology, Journal of the Korea open GIS association, 2(2), pp.117-126.

- Received 12 July 2017
- First Revised 4 August 2017
- Finally Revised 4 August 2017
- Accepted 4 August 2017