http://dx.doi.org/10.5392/JKCA.2013.13.04.427

소설 등장인물의 텍스트 거리를 이용한 사회 구성망 분석

Analysis of Social Network According to The Distance of Characters Statements

박경미*, 김성환**, 조환규**
부산대학교 U-Port 정보기술산학공동 사업부*, 부산대학교 정보컴퓨터공학부**

Gyeong-Mi Park(miya11@pusan.ac.kr)*, Sung-Hwan Kim(sunghwan@pusan.ac.kr)**,
Hwan-Gue Cho(hgcho@pusan.ac.kr)**

요약

복잡계 과학의 발달에 따라 많은 사회 네트워크들이 분석되고 있다. 사회 네트워크는 현재 인문, 경제, 웹 사이언스 등 다양한 분야에 응용되고 있다. 최근, 소설의 등장인물을 이용한 네트워크와 실제 사회 네트워크의 특성을 비교하는 다양한 연구가 진행되고 있다. 그러나 기존의 등장인물 네트워크는 대부분 미리 정리된 인명사전을 이용하므로 주요한 몇몇 인물들 사이의 연관성은 밝힐 수 있으나, 한번 이상 등장한 모든 인물의 전체적인 사회적 구조는 설명하지 못하고 있다. 본 연구에서는 소설로부터 등장인물을 직접 추출하고, 등장인물 사이의 거리를 사용하여 상관관계를 설정하여 네트워크를 구축한다. 제안방법은 소설 텍스트로부터 등장인물의 출현빈도와 등장인물들 사이의 연관성의 발생 빈도를 이용하여 연관성 가중치를 구할 수 있으며, 이 연관성 가중치를 사용하여 노드의 수를 조절하여 K-critical 네트워크를 구성한다. 제시한 K-critical 네트워크는 분석대상 소설에 등장하는 인물들끼리 얼마나 긴밀하게 연관되어 있는지를 정량적으로 파악하는 매우 중요한 정보를 줄 수 있음을 실험을 통하여 제시할 수 있었다.

■ 중심어 : | 사회 네트워크 | 그래프 이론 | 데이타 마이닝 |

Abstract

With the fast development of complex science, lots of social networks are studied. We know that the social network is widely applied in analyzing issues in human culture, economics and web sciences. Recently we witness that some researchers began to compare the social network constructed from fiction literatures(literature social network) and the real social network obtained from practice. But we point that previous approaches for literature social network have some drawbacks since they completely depend on the biographical dictionary constructed for a designated literature. So since the previous approach focus on the few important characters and peoples around them, we can not understand the global structure of all characters appeared in the literature at least once. We propose one method to extract all characters appeared in the literature and how to make the social network from that information. Also we newly propose K-critical network by applying frequency of the named characters and the strength of relationship among all textual characters. Our experiment shows that the K-critical measure could be one crucial quantitative measure to compute the relationship strength among characters appeared in the object literature.

■ keyword : | Social Network | Graph Theory | Information Mining |

* 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음
접수번호 : #130212-007
접수일자 : 2013년 02월 12일
심사완료일 : 2013년 03월 27일
교신저자 : 조환규, e-mail : hgcho@pusan.ac.kr

I. 서 론

최근 복잡계 연구의 일환으로 네트워크에 관한 연구가 다양한 분야에서 활발하게 진행되고 있다[1-3]. 네트워크란 노드(node)라 불리는 구성요소와 그들 사이를 연결하는 링크(link)로 이루어진 하나의 기하학적 구조로 네트워크를 기반으로 노드와 노드 사이의 상관관계를 분석하여 복잡한 현상을 쉽게 이해할 수 있다[4]. 이러한 네트워크는 공학, 생명과학, 정보 분야 등에 널리 응용되고 있다. 특히 사회과학 분야에는 국가, 경제, 소설, 문화 등 다양한 분야의 핵심 구성원 또는 구성성분을 노드로 하고 이들 사이의 상관관계를 연결하는 링크로 구성된 복잡하고 다양한 네트워크에 대한 연구가 지속되고 있다[4][23]. 다양한 네트워크 연구들 중에서, 실제 사회가 아닌 소설 속에 등장하는 인물들을 노드로 하고 인물사이의 관계를 링크로 연결하여 구축한 네트워크를 분석하여 작가의 작품구상, 등장인물의 이해, 소설 전개에 따른 네트워크의 변화를 통해 동적 특성을 분석하는 연구들이 진행되고 있다.

문학 소설의 등장인물들의 관계를 바탕으로 구축한 네트워크에 대한 연구들은 크게 두 가지로 네트워크의 정적인 구조적 특성을 분석하는 것[5][6]과 시간의 흐름에 따라 네트워크의 동적 변화를 분석하는 것[7][8]으로 분류할 수 있다. 먼저, 네트워크의 정적인 구조적 특성을 분석하는 것은 최단거리, 연결선수의 분포, 견고성, 네트워크 다양성 등을 분석하는 것이다[9-11]. 이것은 주로 네트워크의 구조적 특성을 분석하고 소설의 등장인물들의 연결망 역시 멱함수 분포를 나타내며 척도 없는 네트워크의 특성을 제시함으로서 사회 네트워크와 유사하다는 것을 보여주었다. 두 번째, 소설의 시간적 흐름에 따라 주요 등장인물과 연관성의 변화에 의해 네트워크를 구축하고 분석하는 연구가 소설 삼국지와 대하소설 토지를 이용하여 이루어졌다[7][8]. 이들은 등장인물 네트워크를 구축하고 구조적인 분석, 동적변화 및 중심인물과 소설 속에 전개되는 이야기와 연계하여 네트워크의 동적 변화를 기반으로 분석하였다.

소개된 소설 등장인물 네트워크에 대한 연구의 대부분은 소설의 등장인물들과 그들 사이의 연관관계를 파악할 때 소설을 수동으로 여러 번 읽어서 관계를 설정하거나 미리 정리한 인명사전을 기초로 이루어졌다. 전자의 경우는 직접은 소설을 읽어서 등장인물과 관계를 파악하므로 많은 시간과 노력이 필요하다. 또한 후자의 경우는 모든 소설의 등장인물에 대한 정리된 인명사전이 존재하는 것은 아니다. 인명사진이 존재하는 경우에도 소설의 등장인물의 출현빈도나 등장인물들 사이의 관계의 발생빈도를 측정할 수 없다.

본 연구는 등장인물 이름 텍스트 사이의 거리를 이용하여 사회망(Social Network) 구축 및 분석 방법을 제안한다. 제안방법은 소설 문서로부터 등장인물 인명 텍스트(Character Name Text)를 추출하고, 추출된 등장인물들의 인명 텍스트의 위치 정보에 따라 그들 사이의 거리를 계산하여 연관관계를 설정한다. 그리고 각 등장인물들 사이에 설정된 연관관계의 빈도를 계산하여 연관성 가중치를 구하고, 이 연관성 가중치에 따라 수준별로 노드 수를 K개로 제한한 K-critical 네트워크를 구축하여 소설의 주요 등장인물들 사이의 연관성과 구조적 특성을 분석한다.

본 연구의 주된 목적은 다양한 언어로 작성된 문학 소설의 등장인물을 이용하여 자동으로 사회망을 구축하고, 구조적 분석을 수행하고, 이 분석된 결과를 이용하여 문학 소설의 구조를 분석하여 작가의 작품의 특성을 분석하는데 도움을 제공할 수 있다. 또한 등장인물들 사이의 연관관계의 빈도를 측정하여 이 값을 사회망 연결 가중치로 사용하여 네트워크의 노드 수를 동적으로 조절할 수 있어, 소설의 중심 등장인물과 그들의 주된 연관성을 자동으로 요약할 수 있다.

본 논문의 구성은 다음과 같이 구성된다. 관련연구에서는 본 논문의 연구와 관련된 개체명 인식과 등장인물을 이용한 사회 네트워크에 관련된 여러 기술을 소개한다. 그리고 등장인물 이름 텍스트 추출하고, 문장 사이 거리를 계산하는 방법을 소개한 후, 본 연구에서 제안하는 거리기반 네트워크 구성을 위한 연관성 계산 및 사회 네트워크를 구축하는 방법을 소개한다. 그리고 제안 방법의 타당성을 검증하기 위한 실험 및 분석을 수행하고 결과를 평가하고, 마지막 결론을 맺는다.

II. 관련 연구

2.1 개체명 추출

소설의 등장인물 네트워크 구축은 먼저 소설 텍스트에서 등장인물을 지칭하는 텍스트를 추출하여야 한다. 문서로부터 개체명(인명, 지명, 조직명 등)을 추출하는 방법은 규칙 기반 방법과 통계 기반 방법으로 나눌 수 있다. 규칙 기반 방법은 정교한 규칙을 수동으로 작성하거나 수동 작성된 규칙을 학습 코퍼스를 이용하여 수정하는 방법을 사용한다. 이때, 고유명사 사전이나 접사 사전, 결합명사 사전과 같은 다양한 사전을 이용한다. 규칙 기반의 방법은 적용할 분야의 특성을 활용한 방법[13], 문장 내에 자주 발생하는 문맥 기반 방법[14], 접사 사전과 결합 규칙을 활용한 방법[12], 규칙과 문맥을 다단계로 적용한 방법으로 나눌 수 있다.

통계 기반의 방법은 학습 코퍼스의 유무에 따라 교사 학습을 이용한 연구와 비교사 학습에 기반한 방법[17]으로 나눌 수 있으며, 교사 학습 방법은 학습방법에 따라 HMM에 기반한 방법[16]과 Trigram과 이를 확장한 방법[18], 결정 트리 기반[19], 최대 엔트로피 모델에 기반한 방법[15] 등이 있다. 소개한 방법들 중에 비교사 학습에 기반한 방법은 높은 정확성을 얻은 것으로 평가되었다. 그러나 한국어에서 비교사 학습을 위한 구문 분석 결과를 대량으로 얻기는 매우 어렵다. 또한 얻어진 구문 태그 결과에 대한 신뢰성도 약하기 때문에 직접 이러한 방법을 적용하기는 어렵다.

한국어 문서에서 개체명 인식을 위해서는 고유명사의 출현 패턴을 이용하거나 성씨 정보를 이용해 인명일 가능성을 있는 경우를 추출하여 뒤에 호칭이나 지위를 나타내는 말이 나올 경우 색인어로 채택하는 연구가 있었다[20]. 그리고 문서에 나타나는 미등록어 추정을 위해 형태소 분석에 실행한 어절을 입력어로 해서 어절 내의 조사와 어미, 접미사, 명사를 고려하여 미등록어를 추출하고, 고유명사인 경우 범주를 추정하는 연구도 있었다[21]. 본 연구에서는 성씨 정보를 이용하여 추출된 인명과 형태소 분석 결과 미등록어를 추출하고, 이들의 어미를 분석하여 인명 텍스트를 추출한다.

2.2 소설 등장인물 사회망 구성

문학 소설 등장인물들의 관계를 바탕으로 네트워크의 구축 및 분석에 관한 다양한 연구가 있어왔다[5]. 이것은 소설의 등장인물들에 대한 인간관계 네트워크를 구축하고 이 연결망의 특성을 일반적인 사회 연결망의 특성과 비교 분석하는 것이다.

등장인물 네트워크는 대부분 단어 수준의 텍스트를 기반하여 수행된다[22]. 이 작업은 작가의 단어 사용의 스타일과 어휘 패턴에 집중되고 있으며, 소설이 현실세계를 정확하게 표현하기 위한 양식으로 문학적 이론을 검증하거나 평가하는데 기여한다. 이 이론에 따르면, 소설의 형식(줄거리 구상 작업, 캐릭터, 기본적인 범주)과 실제 사회 공간의 변화 사이의 관계에 대한 이론을 추정하는 것이다. 등장인물을 이용한 사회망은 다양한 문학작품을 사용하여 네트워크 구축 및 분석에 대한 연구가 이루어졌다. 다음의 [표 1]은 소설 등장인물을 이용한 사회망 분석에 대한 주된 연구 주제를 요약한 것이다.

표 1. 주된 등장인물 사회망 구축 및 분석에 대한 연구 요약

내용	참고	언어
19세기 영국 소설 등장인물 네트워크	[5]	영어
연속드라마 등장인물 분석에 의한 사회 그룹화	[6]	영어
세익스피어 작품 등장인물 네트워크[9]	[9]	영어
그리스 신화에 의한 복잡계 분석 [10]	[10]	한글
소설 토지에 의한 복잡계 분석[11]	[11]	한글
소설 토지의 네트워크 동적 변화분석[8]	[8]	한글
삼국지의 등장인물 동적변화 분석[7]	[7]	한글

등장인물 사회망 분석은 크게 네트워크의 정적인 구조적 특성을 분석하는 것과 소설의 내용상 시간의 흐름에 따라 네트워크의 구조적 동적 변화를 분석하는 것으로 나눌 수 있다. 네트워크의 정적인 구조를 분석한 대표적인 연구들은 리어왕과 헴릿 등과 같은 섹스피어 작품에 등장하는 인물들의 네트워크[9], 그리스-로마 신화 네트워크[10], 박경리의 대하소설 토지에 나오는 인물들에 대한 네트워크[11] 등이 있다. 이것은 네트워크의 정적인 특징인 최단거리, 연결선수의 분포, 견고성, 네트워크 다양성과 같은 구조적인 면에 한정되어 진행되었다. 그러나 소설로부터 이러한 네트워크의 대부분

은 시간과 공간에 따라 끊임없이 변화되는 동적인 네트워크이기 때문에 소설 삼국지 등장인물 네트워크 동적 변화 분석[7]과 대하소설 토지 등장인물 네트워크 동적 변화 분석[8]과 같이 시간에 따라 동적으로 변화하는 네트워크 분석에 대한 연구가 이루어졌다. 여기서 소개한 대부분 소설 등장인물 네트워크는 대부분 미리 정리된 인명사전을 이용하거나 수작업에 의해 인물들 사이의 연관성을 설정하여 네트워크를 구축한다. 그러나 모든 소설에 대해 미리 정리된 인명사전이 존재하지 않으며, 또한 수작업에 의한 연관성을 설정하기는 매우 번거롭다.

소설의 텍스트에서 대화 인용을 검색하여 연관성을 설정하여 네트워크를 구성하는 연구가 다수 19세기 소설 및 연재물을 이용하여 이루어졌다[5]. 이 방법은 소설 텍스트에서 인용부호(“ ”)를 탐색하여 대화에 참여하는 등장인물을 찾아서 연관성을 설정하였다. 그러나 소설에서 직접적인 대화뿐만 아니라 상황 설명 등으로도 연관성이 이루어질 수 있다. 그러므로 본 연구에서는 소설 텍스트에서 인명 텍스트를 검출하고 그들 사이의 거리를 사용하여 연관관계를 설정하여 사회망을 구축한다.

III. 등장인물 사회망 구축 전처리 과정

3.1 등장인물 텍스트 추출

문학 소설의 등장인물을 이용한 사회망 구축은 등장인물들 사이의 연관성을 이용하여 네트워크를 연결한다. 이 작업은 등장인물을 지칭하는 텍스트에 대한 추출 작업과 그들 사이의 연관성을 설정하는 작업이 필요하다.

본 연구는 소설 텍스트로부터 등장인물을 지칭하는 텍스트를 추출하여 등장인물 리스트와 등장인물 마크 파일을 작성하는 전처리 과정과 등장인물들 사이의 거리를 이용하여 연관성을 설정하여 사회망을 구축하는 과정으로 구성한다. 다음의 [그림 1]은 본 논문에서 제안한 사회망을 구축하는 과정을 나타낸 것이다.

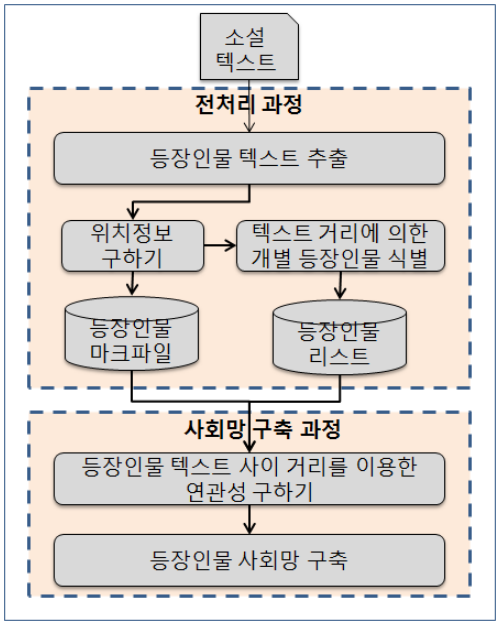


그림 1. 제안된 등장인물 사회망 구축 과정

제안 시스템의 첫 번째 단계는 사회망 구축을 위한 전처리 과정으로 (1) 소설의 텍스트로부터 인명 텍스트를 추출하고, (2) 추출된 인명 텍스트의 위치 정보를 포함한 등장인물 마크 파일(Character Mark File)과 (3) 텍스트 사이 거리를 사용하여 여러 단어로 이루어진 한 등장인물 단어들을 묶어서 개별 등장인물을 식별하여 등장인물 리스트(Character List)를 작성하는 것이다.

먼저 소설로부터 인명 텍스트를 추출하기 위하여 객체명(인명, 지명, 조직명) 인식이 필요하며, 객체명 인식 방법은 언어의 특성에 따라 달라진다. 본 연구에서는 한글 소설과 영문 소설을 데이터로 사용한다. 그러므로 등장인물 텍스트 추출 과정은 소설의 언어에 따라 [그림 2]와 같이 영문과 한글 부분을 분류하였다. 영문의 경우 모든 문장 구성요소들이 단어로 구성되어 있고, 고유명사 표기에 대문자를 사용하기 때문에 대문자로 시작하는 단어를 객체명 후보로 간단하게 추출할 수 있다. 그리고 추출된 개체명 후보 단어에 대하여 품사의 정보, 문자의 특성 등과 같은 객체명에 대한 사전 정보와 객체명 후보의 주변 단어(Mr. Dr. Co, Bank, University, City 등)를 분석하여 개체명을 인식할 수

있다. 영문 텍스트에서 객체명 인식을 위한 응용 프로그램들이 구현되어 있다. 스텐포드 개체이름 인식기(SNER : Stanford Name Entity Recognizer)는 영어 문서에서 객체명 인식을 위한 응용 소프트웨어로 스텐포드 자연어처리 연구팀에 의해 개발 되었고, 개체명 인식에 대한 우수한 성능이 알려져 있다. 그러므로 본 연구에서도 영어 소설에서 객체명을 추출하기 위해 SNER을 사용하였다.

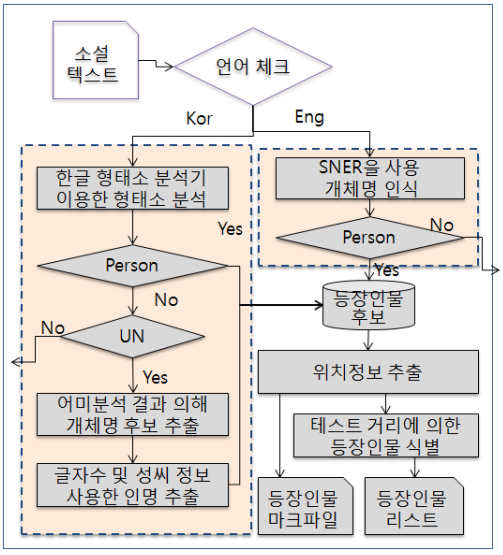


그림 2. 소설 텍스트로부터 등장인물 식별을 과정

한글은 영어 문서와 다르게 대문자 구분이 없기 때문에 고유명사 후보를 쉽게 추출할 수 없다. 또한 문장을 구성하는 어절은 하나의 품사가 아닌 "명사+ 조사"로 구성되거나 또는 "명사 + 명사" 구성된 복합명사 등과 같이 여러 형태소의 복합 요소로 구성되어 있다. 그러므로 한글의 경우 개체명 인식을 위해 먼저 어절의 복합 요소에 대한 형태소 분석을 수행하고 그 결과를 바탕으로 인명후보를 선택한다. 본 실험은 형태소 분석을 위해 KAIST 시맨틱웹 연구센터에서 개발한 한나눔 형태소 분석기를 사용하였다.

본 논문은 소설에서 등장인물 텍스트를 추출하므로 형태소 분석된 결과가 인명(Person)이면 바로 등장인물 후보(Person Name) DB에 추가한다. 형태소 분석에서 인명은 고유명사로 미리 만들어진 형태소 사전을 참고하여 구현되며, 이러한 형태소 사전은 모든 인명이 포함되어 있지 않다. 또한 소설의 경우 성+ 이름이 아닌 별칭(소설 토지의 예 : "윤씨부인", "간난이네" 등)으로 표기되는 경우도 있다. 이러한 이유로 소설의 등장인물에 대한 명칭 텍스트의 많은 부분이 형태소 분석 결과 미등록 단어(UN : undefined noun)로 분류되므로 이에 대한 처리가 필요하다. 본 연구에서 형태소 분석결과에서 미등록어(UN)로 태깅된 어절의 어미의 조사(인명에 많이 붙는 조사로 ", "가", "은", "는", "의", "에게", "부인", 등)와 문자열 길이 $l(2 < l \leq 3)$을 검사하여 인명 후보를 선택하여 등장인물 후보 DB에 저장하였다. 등장인물을 인명 DB에 저장하였다. 등장인물 인명 DB에 중복되어 저장된 모든 인명의 위치정보를 구하고 이것을 포함하여 등장인물 마크 파일을 작성한다. 그리고 한 등장인물은 여러 단어로 된 이름을 사용하거나 같은 단어가 여러 사람을 지칭하는 경우도 있다. 그러므로 텍스트 사이의 거리를 사용하여 한 등장인물을 나타내는 텍스트를 식별하여 등장인물 리스트를 작성한다. 이 과정은 다음절에서 자세히 다룰 것이다.

3.2 등장인물 텍스트의 거리 계산

본 연구는 앞 절에서 소개한 방법으로 추출한 등장인물 텍스트는 개별 등장인물을 식별하고 등장인물 사이의 연관성을 계산하기 위해 등장인물 단어(word)의 위치와 텍스트 사이의 거리 계산이 필요하다.

텍스트의 위치와 텍스트 사이의 거리를 계산하기 위해 다음과 같이 정의한다. 하나의 문서를 구성하는 텍스트를 T로 정의하고, 문서의 텍스트 T를 구성하는 각각의 문장은 S_i라고 하고, 각 문장을 구성요소인 단어 $W_{i,j}$로 정의한다. 단어 $W_{i,j}$에서 i는 T에서 문장 S_i의 순서번호를 나타내고 j는 문장 i에서 단어의 순서번호를 나타낸다. 각 문장 S_i에서 첫 번째 단어는 $W_{i,1}$로 표기한다. 그리고 문장 S_i의 단어의 개수는 $|S_i|$로 나타낸다. 이와 유사하게 $|T|$ 는 문서의 텍스트 T에서 문장의 개수를 나타낸다. 그리고 텍스트 T에서 단어 $W_{i,j}$의 순서를 $wordrank(W_{i,j})$라 하고 다음의 식(1)

과 같이 정의한다.

$$wordrank(W_{i.j}) = \sum_{k=1}^{i-1} |S_k| + j, \ldots\ldots (1)$$

식(1)에서 $|S_k|$는 문장 S_k의 내의 단어의 개수를 나타내고, j는 S_k에서 $W_{i,j}$의 순차적 위치를 나타낸다.

앞 절에서 소설 텍스트 T에서 모든 등장인물을 나타내는 텍스트를 $\{C_i\}$라고 표기한다. 즉 C_i는 소설에서 나타나는 모든 인명을 지칭하는 단어로 아래 [그림 3]에서 가장자리에 테두리로 표시한 모든 단어들을 말한다. 즉 [그림 3] 예에서 C_1는 첫 번째 이름을 나타내는 단어 "Harry"이고, C_2는 두 번째 이름을 나타내는 단어 "Potter"로 각각 $C_1 = Harry$, $C_2 = Potter$이다. 마찬가지고 C_{10}과 C_{11}은 10번째와 11번째 이름 단어인 5번째 줄의 "James"와 "Potter"로 각각$C_{10} = James$, $C_{11} = Potter$이다. 이것을 이용하여 C_3의 위치정보를 식(1)을 사용하여 구하면, C_3 = $W_{2,1}$ 두 번째 줄 첫 번째 단어인 "Potter"이므로 식(1)을 사용하여 단어의 위치를 구하면 다음과 같다.

$$wordrank(W_{2,1}) = \sum_{k=1}^{2-1} |S_k| + 1 = |S_1| + 1 = 7$$

같은 방법으로 마찬가지로 C_5과 C_{10} 위치를 구하면 $workrank(C_5) = |S_1| + |S_2| + 1 = 6 + 9 + 1 = 16$이 고 $workrank(C_{10}) = 34$이다.

Harry Potter walked across Hedwig's cage. #
Potter and Dursley had been absent for two nights. #
Dursley worried about Hedwig. #
Ron Weasley had pretended for ten years. #
Harry is the son of Lily and James Potter. #

그림 3. 등장인물 추출 및 거리 계산 샘플 데이터, 문장은 "#" 기호 사용하여 구분, 모든 등장인물 단어 C_i는 가장자리에 테두리를 표시

[그림 3]의 예제에서 한사람의 등장인물을 지칭하는 단어는 1개 이상의 단어로 이루어져 있는 것을 발견할 수 있다. 예를 들면 "Harry Potter"라는 등장인물을 지칭하는 단어는 "Harry", "Potter", "Harry Potter" 등으로 나타난다. 또한 "Potter"라는 단어는 "Harry Potter" 또는 "James Potter" 두 명의 등장인물이 같은 이름으로 지칭하는 경우도 있다. 그러므로 이 두 가지 경우 어떤 이름을 나타내는 단어가 특징인물을 지칭하도록 선택하여야 한다. 본 연구에서는 이 문제를 해결하기 위해 단어 사이의 거리를 이용하였다. 먼저 2개 이상의 이름 단어 C_i가 서로 인접하여 나타나면 한사람의 등장인물을 지칭하는 것으로 선택한다. 또한 한 단어가 두 사람을 지칭하는 경우에는 그 단어와 나머지 다른 단어 사이 거리를 이용한다. [표 3]의 예에서 두 번째 줄 C_3 =Potter의 경우 "Harry Potter와 James Potter 둘 중에서 선택하기 위하여 C_3과 C_1 사이의 거리와 C_3와 C_{10} 사이의 거리를 비교하여 더 짧은 거리인 것을 선택한다. 단어 사이의 거리 계산은 두 단어 $W_{i,j}$와 $W_{k,l}$ 라고 할 때 다음과 같이 계산하였다.

$$dist_{word} = |wordrank(W_{i,j}) - wordrank(W_{k,l})|, (2)$$

위의 식을 사용하여 모든 이름을 나타내는 단어의 $wordrank()$를 구하면 [표 2]와 같다.

표 2. 등장인물 단어 거리 계산을 위한 기초자료

C_i	W_{ij}	$wordrank(W_{i,j})$
C_1	$W_{1,1}$	1
C_2	$W_{1,2}$	2
C_3	$W_{2,1}$	7
C_4	$W_{2,3}$	9
C_5	$W_{3,1}$	16
C_6	$W_{4,1}$	20
C_7	$W_{4,2}$	21
C_8	$W_{5,1}$	27
C_9	$W_{5,6}$	32
C_{10}	$W_{5,8}$	34
C_{11}	$W_{5,9}$	35

[표 2]의 기초 자료를 사용하여 두 단어 사이의 거리가 1인 것을 이용하여 동일 인물을 나타내는 단어는 다음의 [표 3]과 같이 추출하였다.

표 3. 동일 인물을 나타내는 단어 구하기

C_i, C_j	$dist_{word}(C_i, C_j)$	words
C_1, C_2	|2-1|=1,	Harry Potter
C_6, C_7	|21-20|=1	Ron Weasley
C_{10}, C_{11}	|35-34|=1	James Potter

한글 소설의 경우 “성+이름”, “이름”을 한 인물을 나타내도록 다음의 예와 같이 이름 리스트를 구성하였다.

표 4. 한글 등장인물 리스트

Character List
1 최서희 서희 2 김길상 길상 3 김상현 상현

그러나 등장인물을 나타내는 경우 “성+이름” 또는 “이름” 이외의 다른 방법으로 나타내거나 이름이 바뀌는 경우도 있다. 예를 들면 토지에서 “봉순이”가 “기화”로 이름이 바뀌는 경우와 “소설 삼국지”에서 “유비”와 “유현덕”이 같은 사람을 나타내는 경우이다. 그러나 이것은 소설 텍스트의 내용분석이 필요하다. 그러므로 본 논문에서는 주요 등장인물을 중심으로 알려진 위키 사전 및 인명사전을 참고하여 수동으로 두 개 이상으로 이름으로 나타나는 등장인물을 하나의 등장인물로 묶어 주었다.

또한 같은 단어가 여러 사람을 나타내는 경우 $dist_{word}()$가 가까운 것을 선택하였다. 예를 들어 C_3과 C_1의 거리를 계산하면 $dist_{word}(C_3, C_1) = 6$이고, C_3와 C_{10} 사이의 거리를 계산한 결과는 $dist_{word}(C_3, C_{10}) = 27$이다. 그러므로 C_3의 Potter의 “Harry Potter”를 지칭하는 것으로 선택하였다.

본 연구에서 등장인물들 사이의 연관성을 설정하기 위하여 등장인물 단어가 포함된 문장들 사이의 거리를 사용하였다. 연관성 계산을 위한 두 등장인물 각각 C_i, C_j가 각각 다른 등장인물일 때, 이것을 각각 X, Y라고 두고 각각 문장 S_i와 S_j의 단어이면 두 단어 사이의 문장거리는 $dist_{stete}(X, Y) = |j - i|$로 계산한다. 문장 거리를 계산한 결과가 $dist_{stete}(X, Y) = 0$이면 두 등장인물이 같은 문장에 나타나는 것으로 서로 연관성이 매우 높은 것이다.

[표 2]의 예제에서 첫 번째 줄에 있는 Harry와 4 번째 줄에 있는 Ron 사이의 거리를 계산할 때 문사 사이 간격을 이용하여 계산하면 $dist_{state}(s_1, s_4) = |4 - 1| = 3$가 된다. 본 연구에서 등장인물들 사이의 연관성 계산을 문장사이의 거리를 이용하였고, 문장의 거리 범위는 0~5까지 사용하였다.

VI. 등장인물 거리를 이용한 사회망 구성

이 장에서는 소설 텍스트로부터 사회네트워크를 구축하는 과정을 설명한다. 분석된 소설 텍스트를 T라고 하고 T로부터 구축된 연관성 그래프 $G_T(V, E)$를 구축한다. 모든 등장인물 단어는 그래프(G: Graph)에서 노드(V: Vertex)로 매핑 시키고, 등장인물들 사이의 관계를 엣지(E : Edg)로 연결하여 네트워크를 구축한다. 여기서 v_i, v_j는 두 등장인물을 나타내는 서로 다른 노드이다. 두 노드 v_i, v_j 사이의 엣지 연결은 두 노드 사이의 연관성 가중치가 상수 임계값 이상이면 연결한다. 본 연구에서 두 노드 사이의 연관성 가중치는 등장인물 텍스트의 단어나 문장을 기반으로 측정하기 때문에 문장 사이의 거리 $dist_{state}()$를 적용하고, 다시 단어 사이의 거리 $dist_{word}()$에 의해 비교하였다. 소설 텍스트 T에서 대부분의 등장인물들은 1번 이상 출현하며, 두 등장인물 X, Y도 인접한 위치에 1번 이상 중복되어 나타난다. 그러므로 두 등장인물 X, Y의 연관성 가중치는 다음 식과 같이 정의할 수 있다.

$$weight(X, Y) = \sum_{i=1}^{n} \alpha^{k_i}, \quad \text{.....(3)}$$

, 식(3)에서 $k = dist_{state}(w_x, w_y)$이며, α는 제어 변수로 $0 < \alpha < 1$이다. 본 연구의 실험에서 $\alpha = 0.7$을

사용하였다. 그래서 두 캐릭터가 하나의 문장에 있다면 $0.7^0 = 1$이며, 두 등장인물의 문장사이 거리 차가 1이면 0.7 이고, 문장사이 거리 차가 5이면 $0.7^5 = 0.16807$이 된다. 만약 제어 변수 $\alpha = 0.9$ 이면, $\alpha^5 = 0.59049$이다. 제어 변수 $\alpha = 1$이면, 문장사이의 거리를 무시하고 모든 가중치가 1이 된다. T에서 두 등장인물이 한 문장에 있을 때 연관성이 매우 높으며, 문장사이 거리가 멀어지면 연관성이 낮아진다. 본 연구에서는 $\alpha = 0.7$을 사용하여 두 주인공의 텍스트의 문장사이 거리가 멀어질수록 α^k의 값이 작아지도록 하였다. 다음의 [표 3]은 [표 2]의 문장 샘플을 식(3)를 사용하여 연관성 가중치를 구한 것이다.

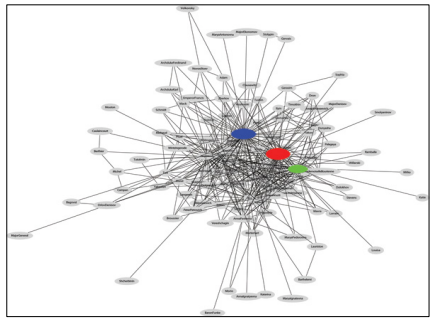
표 5. 연관성 가중치 계산 예

관계(X->B)	거리	weight(X,Y)
Harry->Ron	1, 2, 3	1.533
Harry->Jemes	0, 3, 4	1.583
Jemes->Lily	0	1

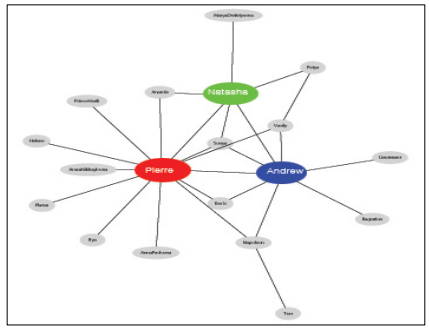
[표 5]에서 "Harry->Ron"의 연관성 가중치는 첫 번째 줄 "Harry"에서 거리 3인 것(0.343)과, 두 번째 줄에 "Potter에서 거리 2인 것(0.49), 5번째 줄에서 거리 1인 것(0.7)을 모두 더하여 0.343+0.49+0.7=1.533이다. 제안 방법은 등장인물들 사이의 거리를 이용하여 연관성 가중치를 구하고, 이 값을 사용하여 그래프 G에서 노드 v_i와 노드 v_j 사이의 연결을 제어하여 가변적으로 그래프를 그릴 수 있다. 즉, $weight(X, Y) \geq t$ 이면 두 등장인물 X와 Y사이를 연결한다. 여기서 t는 연관성 가중치 수준을 결정하는 임계 값(threshold) 이다.

일반적으로 소설에서 주인공을 비롯한 주요 등장인물들은 출현 빈도가 높으며, 서로 밀접하게 연결되어 있다. 연관성 가중치 수준(t)을 적절히 조절하여 네트워크를 구성하면 소설의 주요등장인물과 그들 사이의 연관성을 요약할 수 있다. t의 값은 연관성 가중치 최대값과 최소 갑에 의해 간격을 선택한다.

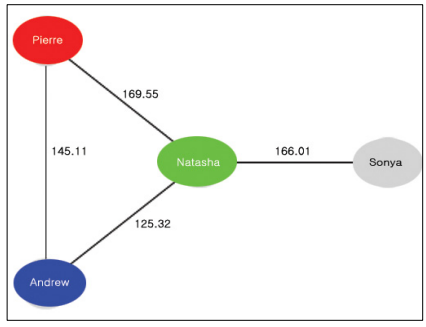
다음의 [그림 4]은 영문 소설 "War and Peace" 주인공들 사이의 연관성 가중치의 임계값 t 변화에 따른 사회 네트워크를 나타낸 것이다.



(a) $weight(X, Y) \geq 1$



(b) $weihgt(X, Y) \geq 30$



(c)$weiht(X, Y) \geq 110$

그림 4. 영문 소설 "War and Peace "의 등장인물 연관성 가중치 임계값 t 변화에 따른 사회네트워크. (a)는 $t = 1$일 때 이고, (b)는 $t = 30$, (c)는 $t = 100$을 사용하여 네트워크를 구축한 예

[그림 4]에서 (a)는 $t = 1$일 때 이고, (b)는 $t = 30$, (c)는 $t = 110$을 사용하여 네트워크를 구축한 예이다. $weight(X, Y)$를 계산에 사용한 거리 k는 $0 \leq k \leq 3$ 범위를 사용하였다. [그림 4]에서 3 주인공의 노드의 색을 "Pierre" 빨간색(Red), "Andrew" 파란색(Blue), "Natasha"

녹색(Green)으로 지정하였다.

위의 제안방법에서 연관성 가중치를 계산할 때 두 주인공 텍스트의 문장 사이의 거리 k는 중요한 역할을 한다. 즉 $k=0$일 경우 보다 $0 \le k \le 5$의 범위를 사용하였을 때 등장인물들 사이에 연관성이 높아진다. 소설 전쟁과 평화에서 $k=0$일 때 Pierre와 Andrew 사이의 연관성 가중치는 $weight(X, Y)=53$이고, $0 \le k \le 5$일 때는 $weight(X, Y)=154.14$이다. 제안방법은 연관성 가중치를 사용하여 노드 사이 연결을 결정하기 때문에 k의 범위가 네트워크 연결에 영향을 미친다.

다음의 [표 6]는 k의 범위에 따라 소설 전쟁과 평화에서 세 주인공인 “Pierre”, “Andrew”, “Natasha” 노드에 연결된 링크의 수를 나타낸 것이다. 연관성 가중치는 $weight(X, Y) \ge 1$이다.

표 6. k의 범위 변화에 따른 링크의 수

k의 범위	Pierre	Andrew	Natasha
$k=0$	43	50	23
$k \le 1$	56	59	32
$k \le 2$	62	64	35
$k \le 3$	63	68	37
$k \le 4$	65	69	39
$k \le 5$	67	71	40

V. 실험 및 평가

5.1 실험데이터

이 장에서는 본 논문에서 제안한 소설의 등장인물 이름 텍스트의 거리를 기반으로 구축한 네트워크를 실제 사회 네트워크의 특성과 비교 분석하고, 제한방법의 효율성을 평가한다. 먼저 등장인물 이름의 텍스트 사이의 거리를 이용하여 구축된 네트워크가 사회 네트워크의 특성인 멱함수(Power) 법칙을 따르는 척도 없는 네트워크인지 분석한다. 그리고 네트워크를 연결하기 위해 텍스트 거리를 이용한 연관성 설정의 유효성을 분석한다. 실험을 위해 한글과 영문으로 작성된 소설을 각각 3개 씩 총 6개의 소설을 [표 7]와 같이 데이터로 사용하였다.

표 7. 실험데이터: 소설 텍스트

Title	단어	문장	등장인물
삼국지(한글)	642,222	121,779	829
토지(한글)	1,446,187	176,378	467
해리포터(한글)	907,306	85,005	273
Harry Potter(영문)	1,279,375	126,112	299
War and Peace(영문)	584,618	30,912	240
gone with the wind(영문)	428,478	23,797	92

[표 7]에서 등장인물의 수에 대한 결정은 제안 방법에 의해 검출된 이름을 위키 사전 및 이미 알려진 인명사전을 참고하여 선택한 결과이다.

표 8. 등장인물 이름 겸출 결과

Title	초기검출	오검출	선택등장인물
삼국지(한글)	915	86	829
토지(한글)	509	42	467
해리포터(한글)	296	23	273
Harry Potter(영문)	322	23	299
War and Peace(영문)	237	17	120
gone with the wind(영문)	106	14	92

초기 검출은 제안 방법에 의해 이름 텍스트로 검출된 결과이고, 오검출은 주요 등장인물 중 한 등장인물이 두 개 이상의 이름으로 나누어진 경우이고, 선택 등장인물은 위키 사전 및 알려진 인명사전을 참고하여 나누어진 이름을 하나의 이름으로 묶어서 최종 등장인물로 수동으로 선택한 것이다. 여기서 선택된 등장인물은 출현빈도가 3이상인 것을 사용하였다. 본 논문에서 한 등장인물이 두 개 이상의 이름으로 나타나는 오검출을 수동으로 제거하였으나 소설의 텍스트 내용 분석들을 통하여 자동으로 오검출을 제거 방법 연구가 필요하다.

5.2 등장인물 출현빈도 및 가중치 검사

일반적으로 사회 네트워크는 척도 없는 네트워크로 멱함수 법칙을 따르고 있다. 제안 방법으로 구축된 네

트워크도 멱함수 법칙을 따르는지 분석하기 위해 먼저 등장인물의 출현 빈도를 살펴보았다.

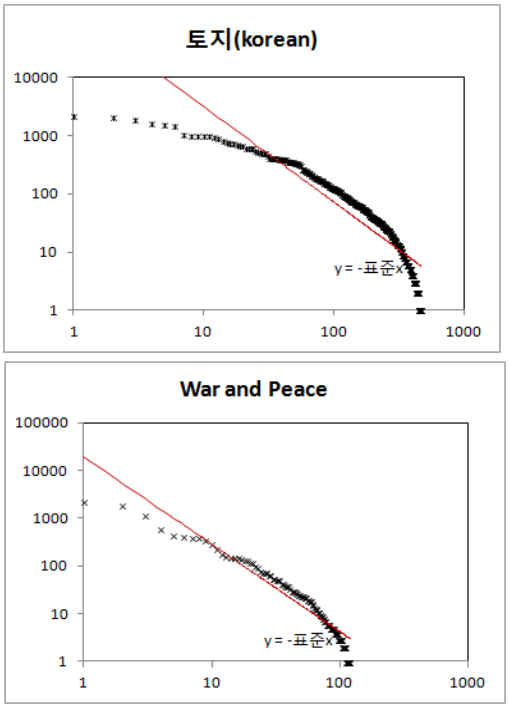


그림 5. 등장인물 출현빈도. X축은 등장인물 번호, y축은 출현빈도를 나타내며, X, Y축 모두 log 눈금 간격을 사용.

[그림 5]는 한글 소설 "토지"와 영문 소설 "Harry Potter"에 대한 등장인물 출현빈도를 나타낸 것이다. [그림 5]에서 한글 소설 "토지"와 영문 소설 "War and Peace" 모두 특정 등장인물의 출현 빈도가 매우 높게 나타나므로 전체 등장인물의 출현 빈도는 멱함수 법칙을 따르고 있다. 등장인물의 출현빈도가 높은 것은 여러 등장인물과 인접해 있을 확률이 높기 때문에 여러 다른 등장인물들과 연결될 확률이 높다.

다음 [그림 6]는 등장인물의 출현빈도와 연결된 노드와 상관관계를 한글과 영문으로 작성된 소설 "Harry Potter"를 이용하여 실험한 결과를 그래프로 나타낸 것이다.

[그림 6]에서 대부분의 등장인물은 출현빈도가 높으면 연결된 노드의 수도 많다는 것을 확인 할 수 있다. 제안된 방법은 등장인물 출현빈도가 멱함수의 법칙을 가지므로 등장인물 네트워크도 사회망의 특징인 멱함수 법칙을 가진다는 것을 확인하였다.

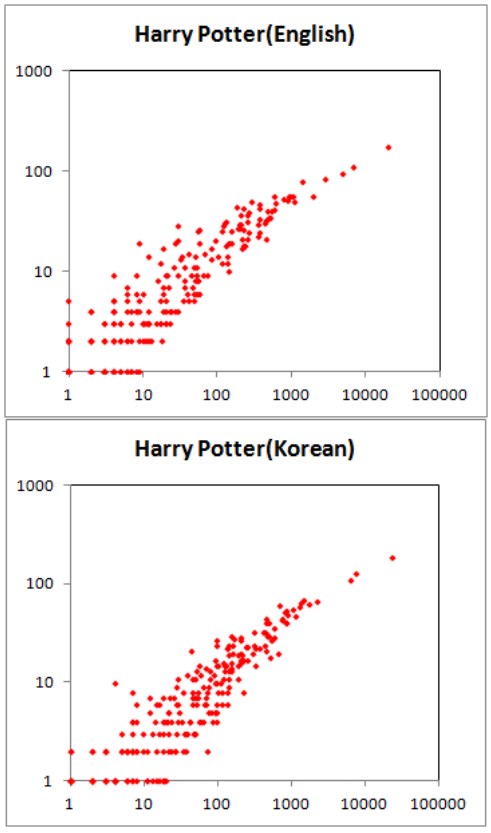


그림 6. 등장인물 출현빈도와 연결된 노드와의 상관관계. x축 등장인물 출현빈도, y축 등장인물에 연결된 노드의 수.

5.3 문장거리에 변화에 따른 노드의 링크 수 변화

제안 방법은 등장인물들 사이의 연관성 가중치를 구할 때 문장 사이의 거리를 여러 단계로 구별하여 실험하였다. 두 등장인물 사이의 연관성을 구할 때, 문장 사이 거리의 범위를 좁게 두면 네트워크에 연결되지 못하는 노드의 개수가 많아지고, 범위가 넓어지면 연관성이 없는 노드들도 연결되는 결과를 가져올 수 있다. 그러므로 본 연구에서 연관성 계산을 위한 적절한 문장거리의 범위를 찾기 위한 실험을 수행하였다, [표 9]는 등장인물 네트워크 전체 노드들의 연결의 합과 출현빈도가

가장 높은 주인공 노드에 연결된 노드의 수를 나타낸 것이다. 문장사이 거리 범위가 넓어질수록 전체 네트워크에 연결된 노드의 수와 주인공에 연결된 노드 수가 증가하는 것을 확인할 수 있다.

표 9. 문장거리 변화에 따른 전체 네트워크에 연결된 노드 수와 주인공에 연결된 노드 수.

Title	$k=0$	$k\leq 1$	$k\leq 2$	$k\leq 3$	$k\leq 4$	$k\leq 5$
삼국지 (한글)	9,566 (320)	17,230 (452)	21,939 (504)	25,415 (554)	28,213 (567)	30,686 (582)
토지 (한글)	4,291 (93)	7,102 (129)	8,829 (164)	1,0217 (181)	11,295 (187)	12,274(196)
해리포터 (한글)	3,201 /184	5,412 (221)	6,593 (237)	7,416 (245)	8,034 (251)	8,587 (257)
Harry Potter (영문)	3,165 /(176)	4,806 (216)	5,662 (231)	6,270 (241)	6,764 (252)	7,186 (260)
War and Peace (영문)	686 /(43)	1,001 (56)	1,158 (62)	1,279 (63)	1,371 (65)	1455 (67)
gone with the wind (영문)	624 (42)	889 (46)	1,018 (49)	1,112 (52)	1,182 (56)	1,246 (59)

[표 9]는 하나의 셀에 전체 네트워크에 연결된 노드 수는 위쪽에 표시하고, 주인공에 연결된 노드의 수는 아래쪽 괄호()에 표시하였다. [표 9]에서 문장 사이 거리 범위가 넓어지면 네트워크에 연결되는 노드의 수는 증가하지만 문장 거리 범위가 너무 좁아지면 실제 연관성이 있는 노드들이 연결이 되지 않는 경우가 발생한다. 그러므로 본 실험에서는 문장거리 범위 변화에 따른 연결 실패(노드가 네트워크에 연결되지 못한 경우)를 검사 하였다.

표 10. 문장거리 변화에 따른 연결 실패율

Title	$k=0$	$k\leq 1$	$k\leq 2$	$k\leq 3$	$k\leq 4$	$k\leq 5$
삼국지 (한글)	0.072	0.063	0.011	0.003	0.002	0.001
토지 (한글)	0.143	0.071	0.049	0.045	0.038	0.030
해리포터 (한글)	0.201	0.075	0.031	0.007	0.000	0.000
Harry Potter (영문)	0.221	0.060	0.020	0.010	0.003	0.003
War and Peace (영문)	0.181	0.091	0.057	0.050	0.033	0.025
gone with the wind (영문)	0.122	0.082	0.061	0.020	0.000	0.000

[표 10]은 등장인물이 문장거리 변화에 따라 네트워크에 연결되지 못한 경우인 연결 실패를 나타낸 것이다. [표 10]에서 문장의 거리 범위가 넓어질수록 연결 실패율이 줄어드는 것을 확인할 수 있다. 즉, 문장사이 거리 범위가 넓어지면 등장인물사이의 연결성이 강해지므로 연결 실패율은 감소하는 것을 확인할 수 있다.

5.4 등장인물 가중치에 의한 노드 수 변화

제안 방법은 네트워크를 구축할 때 등장인물들 사이에 연관성 가중치를 구하였다. 이 연관성 가중치는 임의의 임계값을 사용하여 네트워크를 수준별로 분리할 수 있다. 다음의 표는 영문 소설 “전쟁과 평화(War and Peace)”를 사용하여 연관성 가중치의 수준에 따른 네트워크에 연결된 노드 수의 변화를 보여준다. [표 11]에서 문장사이 거리 범위는 $k=5$을 사용하였다.

표 11. 연관성 가중치 레벨 변화에 따른 네트워크 연결 노드 수 변화(War and Peace 사용)

level	Nodes	Edges	cut value(t)
1	4	7	140
2	5	11	70
3	18	54	30
4	52	127	10
5	108	684	1

본 연구에서 제안한 텍스트 거리 기반 연관성 가중치를 사용하여 구축한 네트워크를 임의의 임계치에 의해 분리하면, 소설의 주요등장인물을 쉽게 찾을 수 있다. 이것을 소설의 주요 내용을 요약하는데 유용하게 활용할 수 있다. 소설의 주요 내용은 주요 등장인물과 관련된 내용이므로, 주요 등장인물 텍스트가 포함되어 있는 문장을 추출하여 소설의 중요한 내용을 요약할 수 있다.

IV. 결론 및 향후 과제

본 논문은 텍스트 사이의 거리를 사용하여 연관성을 자동으로 계산하여 네트워크를 구축하는 방법을 제안하였다. 제안 방법은 다양한 문학 소설에서 등장인물

텍스트를 추출하고, 텍스트 사이의 거리를 사용하여 연관성을 설정하고, 거리와 연관성의 빈도에 따라 가중치를 구하여 여러 단계의 네트워크를 구축하여 분석하였다. 본 실험은 결과는 다음과 같이 요약할 수 있다.

제안된 네트워크가 실제 사회 네트워크의 특성과 비교하기 위해 등장인물 출현빈도와 노드의 링크를 분석한 결과 멱함수(Power)의 법칙을 가지고 있음 확인하였다.

제안 방법에 의해 구축한 네트워크는 출현빈도가 높은 노드(등장인물)가 많은 노드와 연결되어 있는 것을 확인했다. 이것은 실제 소설의 주요 등장인물이 소설에서 자주 등장하고 많은 인물과 연관성이 있음을 나타낸다.

등장인물 사이의 연관성 가중치를 사용하여 수준별 네트워크 구축하여 핵심 노드(등장인물)를 쉽게 요약할 수 있음을 확인하였다. 이것은 실제 소설에서 중심 등장인물들은 서로 간에 높은 연관성을 가지고 있기 때문에 소설의 중심인물에 의한 소설 중심내용분석이 용이하다.

제안 방법의 네트워크의 연결성은 문장의 거리 범위에 비례한다는 것을 확인하였다. 연관성 계산에 문장 사이의 거리 범위가 넓으면 네트워크의 연결성은 좋아지나 잘못 연결된 오류가 높아지는 것을 확인하였다.

제안 방법은 다양한 언어로 작성된 소설로부터 문장이나 의미적 분석 없어 간단한 단어 사이 거리에 의해 쉽게 연관성을 계산할 수 있어 여러 소설로부터 쉽게 네트워크를 구축할 수 있음을 보여주었다.

향후 연구 과제로 소설의 등장인물은 이름 이외의 여러 가지 별칭이나 대명사를 사용하여 소설에 등장하고, 또한 소설은 여러 가지 고유명사가 많이 등장하므로 여기에 대한 더 많은 연구가 필요하다.

참 고 문 헌

[1] R. Solomonoff and A. Rapoport, "Connectivity of random nets," Bull. Math. Biophys, Vol.13, No.1, pp.107-117, 1951.

[2] P. Erdös and A. Rényi, "On random graphs," Publicat. Math., Vol.6, No.2 pp.290-297, 1959.

[3] Y. Y. Liu, J. J. Slotine, and A. L. Barabasi, "Controllability of complex network," Nature, Vol.473, No.7346, pp.167-173, 2011.

[4] Linton Freeman, "The Deveolpment of Social Network Analysis," Empirical Press, 2006.

[5] D. K. Elson, N. Dames, and K. R. McKeown "Extracting Social Networks from Literary Fiction," Proc. ACL, pp.138-147, 2010.

[6] P. Matthews and L. Barrett, "Small-screen social groups: soap operas and social networks," J. Cul. Evol. Psychol, Vol.3, No.1, pp75-86, 2005.

[7] 김윤경, 신현일, 구자을, 김학용, "소설 삼국지 등장인물 네트워크의 동적 변화 분석," 한국콘텐츠학회논문지, 제9권, 제4호, pp.364 - 371, 2009.

[8] 김학용, "대하소설 토지 등장인물 네트워크의 동적 변화 분석", 한국콘텐츠학회논문지, 제12권, 제11호, pp.519-526, 2012.

[9] D. N. Stiller and R. Dunbar, "The small world of Shakespeare's play", Proc. Nat'l. Acad. Sci, Vol.14, No.4, pp.397 - 408, 2003.

[10] 최연무, "복잡계 네트워크로서의 그리스 신화", 물리, 제49권, 제3호, pp.298-302, 2004.

[11] 김상락, "문학 작품에서의 복잡계 연결망 분석: 소설 토지를 중심으로", 물리, 제50권, 제4호, pp.267-271, 2004.

[12] J. Fukumoto, F. Masui, M. Shimohata, and M. Sasaki, "Oki Electric Industry: Description of the Oki System as Used for MET-2," MUC-7, Columbia, MD, 1998.

[13] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward Information Extraction: Identifying protein names from biological papers," PSB'98, 1998.

[14] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유 명사의 추출과 분류", 한국정보과학회 논문집, 제27권, 제2호, pp.170-172, 2000.

[15] K. Uchimoto, O. Ma, M. Murata, H. Ozaku, and

H. Isahara, "Named entity extraction based on a maximum entropy model and transformation rules," Proc. ACL pp.152-160, 1998.

[16] G. D. Zhouand and J. Su, "Named Entity Recognition using an HMM-based Chunk Tagger," Proc. ACL, pp.473-480-201, 2002.

[17] M. Collins and Y. Singer, "Unsupervised Models for Named Entity Classification," EMNLP/VLC-99, 1999.

[18] M. Sassano and T. Utsuro, "Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition," Proc. ACL, pp.705-711, 2000.

[19] H. Isozaki, "Japanese named entity recognition based on a simple rule generator and decision tree learning," ACL, pp.314-321, 2001.

[20] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 제12회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.292-299, 2000.

[21] 양장모, 김민정, 권혁철, "언어정보를 이용한 한국어 미등록어 추정", 한국정보과학회, pp.957-960, 1996.

[22] A. A. j. Ahmed Hassan and D. Radev. "Extracting Signed Social Networks From Text," Proc. ACL, pp.4-12, 2012.

[23] 정진수, 김학용, "고구려, 백제, 신라 왕조실록 인명 네크워크 분석", 한국콘텐츠학회논문지, 제11권, 제5호, pp.474-480, 2011.

저 자 소 개

박 경 미(Gyeong-Mi Park) 정회원

- 2000년 : 한국방송통신대하교 컴퓨터과학과(공학사)
- 2003년 : 부경대학교 정보공학과(공학석사)
- 2010년 : 부경대학교 전자계산학과(이학박사)
- 2011년 8월 ~ 현재 : 부산대학교 U-Port 정보기술산학공동사업단 박사후연구원

<관심분야> : 컴퓨터 비젼, 영상처리, HCI

김 성 환(Sung-Hwan Kim) 정회원

- 2011년 : 부산대학교 정보컴퓨터공학부(공학사)
- 2013년 : 부산대학교 컴퓨터공학과 석사
- 2013년 ~ 현재 : 부산대학교 컴퓨터공학과 박사과정 중

<관심분야> : 한글언어처리, 정보검색, HCI

조 환 규(Hwan-Gue Cho) 정회원

- 1984년 : 서울대학교 계산통계학과(이학사)
- 1986년 : KAIST 대학원 전산학과(공학석사)
- 1990년 : KAIST 대학원 전산학과(공학박사)
- 1990년 3월 ~ 현재 : 부산대학교 컴퓨터공학과 교수

<관심분야> : 계산이론, 생물정보학