설화 스토리 내 인물정보 추출 방법에 관한 연구

(A research for character information extraction method on Narrative Stories)

고병규*, 김정인*, 이은지*, 김판구**

(Byeong Kyu Ko, Jeong In Kim, Eun Ji Lee, Pan Koo Kim)

요 약

스토리텔링 기법을 사용하여 교육, 마케팅, 창작 등 다양한 분야에서 활용하고 있다. 특히 창작분야에서는 기존의 구전설화와 같이 짧고 이해하기 힘든 이야기를 현재의 상황과 배경에 맞춰 새롭게 창작하는 이야기들이 많아지고 있지만, 많은 이야기들이 나옴으로써 이야기의 소재 및 아이디어가 한계에 다다르고 있고 기존의 이야기를 바탕으로 하여 새롭게 각색하는 방법이 매우 힘든 실정이다. 따라서, 본 논문에서는 스토리텔링 저작 지원 소프트웨어 내 사용할 수 있는 스토리 개발을 위한 사전 단계인 텍스트 분석 및 인물정보 추출 방법을 통해 사용자로 하여금 기존의 이야기를 바탕으로 새로운 창작의 기회를 제공할 수 있는 시스템을 개발하고, 이에 대한 기술적인 내용을 기술한다.

■ 중심어 : 설화 콘텐츠 ; 객체 추출 ; 스토리텔링 저작 지원 시스템 ;

Abstract

Storytelling techniques have been using for utilizing in various fields such as education, marketing and so on. In creative content sector, especially, the amount of creating the a newly story content from existing oral tales which are short and quite hard to understand the context has been increased. However, there are some limitations for creating the ideas of the story according to the expansion of the stories while the approach for new adaptation of existing stories is quite challenging. In this paper, therefore, we described the preliminary steps of text analysis and object extraction method for story development which can be applied to the storytelling authoring supported software.

■ keywords : Narrative Contents ; Object Extraction ; Storytelling Writing Support System ;

Ⅰ. 서 론

스토리텔링은 이야기 하고자하는 정보를 상대방에게 쉽게 이해시키며, 기억하게하고 몰입과 공감을 이끌어내는 특성을 지닌 매우 효과적인 커뮤니케이션 형태이다[1]. 스토리텔링 기법을 사용하여 교육, 마케팅, 창작 등의 분야에서 상대의 이해도 및 응용능력 활성화를 위해 활용되고 있다[2,3,4,5,6]. 특히 창작분야에서는 기존의 구전설화와 같이 짧고 이해하기 힘든 이야기를 현재의 상황과 배경에 맞춰 새롭게 창작하는 내용의 이야기들이 많아지고 있다.

하지만, 많은 이야기들이 나옴으로써 이야기의 소재 및 아이디어가 한계에 다다르고 있고 기존의 이야기를 바탕으로 하여 새롭게 각색하는 방법이 매우 힘든 실정이다. 예를 들어, 하나의 이야기를 창작하는데 소요되는 인력은 최소 2~3명이며, 비용 또한 인력 및 기타 자료에 비례하여 많이 소요된다. 이를 해소하기 위해 국내외적으로 이야기 저작을 지원하는 소프트웨어가 개발되어 사용되고 있다. 현재 가장 많이 사용하는 소프트웨어는 국외에서 개발한 드라마티카 프로(Dramatica Pro), 파이널 드래프트(Final Draft)와 국내에서 개발한 스토리헬퍼(Story Helper)가 있다[7,8]. 이들 소프트웨어의 특징은 [표 1]과 같다.

표 1. 스토리텔링 저작 도구

저작도구명	특징
드라마티카 프로	-작가의 집필의도 파악을 위한 질문시스템 탑재 -도움말 및 예제 데이터베이스 제공
파이널 드래프트	-다양한 포맷별 특성화된 템플릿 제공 -캐릭터 이름 데이터베이스, 맞춤법 수정, 자동 Scene 넘버링 기능 제공
스토리헬퍼	-3만여개의 요소 데이터베이스 구축 -205개의 모티브, 36개의 에피소드를 정리하여 자유로운 저작을 할 수 있는 기반마련

출처:문화기술(CT) 심층리포트-스토리텔링 저작도구 연구동향과 사례분석

* 학생회원, 조선대학교 컴퓨터공학 ** 정회원, 조선대학교 컴퓨터공학과

이 논문(저서)은 2014년 교육부와 한국연구재단의 지역혁신창의인력양성사업의 지원을 받아 수행된 연구임 (NRF-2014H1C1A1073115)..

접수일자 : 2016년 06월 10일 게재확정일 : 2016년 06월 29일

수정일자 : 2016년 06월 27일

교신저자 : 김판구 e-mail : pkkim@chosun.ac.kr

[표 1]과 같이 다양한 기능의 스토리텔링 저작 도구가 상용화되고 있으며, 기능 또한 스토리 제작에 있어 많은 도움을 주고 있다. 하지만, 이야기에 저작을 위한 어플리케이션 레벨에서의 지원은 한계점이 있다. 그 이유는, 기존 데이터의 재활용성에 대한 정확한 수집 및 분석이 필요하지만, 모두 작가를 도와주는 보조작가 및 기타 지원자들에 의해 시간과 비용을 투자하고 있기 때문이다. 따라서, 텍스트 처리 및 분석을 통한 개체 추출 및 의미적 분석이 매우 필요한 것이다. 하지만, 영문 분석과 달리 한국어는 매우 복잡한 구조로 만들어져있다[9]. 즉, 동음이의어가 많고 조사나 어미와 같이 문법적 관계를 표현하는 말이 발달되어 정확한 문자구조 분석이 어렵기 때문에 기존에 한국과학기술원에서 개발한 한나눔 형태소 분석기를 사용하여 전처리 과정을 거친다.

한나눔 형태소 분석기는 한국어 처리 목적에 따른 플러그인 컴포넌트 아키텍처 구조를 지니고 있으며 필요에 따라 다양한 플러그인을 지원한다[10,11]. 또한, UI(User Interface)기능을 통해 이용자의 필요에 따라 플러그인 형태를 추가하는 방식으로 사용이 가능하며 다양한 필터를 추가함으로써 확장이 가능하도록 한다. 따라서 유연하게 한국어 데이터를 처리하기 위한 워크플로우를 구성할 수 있으며 다양한 목적에 맞게 활용할 수 있다[12,13,14].

본 논문에서는 앞서 설명한 스토리텔링 저작 지원 소프트웨어 내 사용할 수 있는 스토리 개발을 위한 중심 전단계인 텍스트 분석 및 개체 추출 방법에 대해 기술하고, 이를 토대로 사용자에게 제공할 수 있는 간단한 응용프로그램 개발하였다. 2장에서는 본 논문에서 사용할 데이터 전처리 방법과 전체 내용을 설명하고, 3장에서는 결론을 도출한다.

Ⅱ. 본 론

1. 설화 스토리 내 개체 추출 방법

본 논문의 전체 구성도는 (그림 1)과 같다.

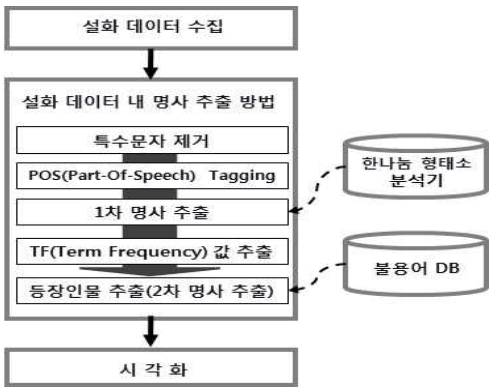

그림 1. 인물정보 추출을 위한 시스템 구성도

가. 설화 스토리 내 1차 명사 추출 방법

설화 스토리 데이터 수집을 위해 온라인상에 존재하는 국내 구전 설화 데이터를 수집하였다. 수집된 데이터는 불필요한 데이터인 특수문자를 제거한다. 이 부분에서 고려해할 사항은 문장의 끝에 오는 마침표를 제거하지 않는 것을 원칙으로 하지만, 대화체를 표현하는 큰따옴표는 삭제한다. 그 이유는 대화 문장을 통해 개체 간의 관계를 표현할 수 있으나, 본 연구에서는 이 부분에 대해 분류 하지 않고 전체 문장이 시나리오라고 가정하기 때문이다. [표 2]의 과정을 거쳐 전처리를 하게 된다.

표 2. 설화 스토리 원문에 대한 POS 태깅 예제

설화 원문	임금님은 이를 허락하고, 세 공주를 구하면 그 중 막내 공주와 결혼시키겠다고 하였다.	
특수문자 제거	임금님은 이를 허락하고 세 공주를 구하면 그 중 막내 공주와 결혼시키겠다고 하였다.	
POS Tagging	임금님/ncn+은/jxc 이/npd+를/jco 허락/ncpa 하고/jct+,/sp 세/nnc 공주/ncr+를/jco 구하/pvg+면/ecs	그/mmd 중/nbn 막내/ncn 공주/ncr+와/jct 결혼/ncpa+시키/xsva+겠/ep+다/ef+고/jcr 하/pvg+였/ep+다/ef+./sf

[표 2]와 같이 기본적인 특수문자 제거 후 POS Tagging을 하게된다. 이 과정에서 추출해야 할 태깅 데이터는 [표 3]과 같다[15].

표 3. 설화 스토리 원문 내에서 1차 명사 추출 태그 정의

정의 태그	정의
ncn	비서술성 명사
ncr	비서술성 직위 명사
ncpa	동작명사
ncps	상태명사
nbn	비단위성 의존명사
nbu	단위성 의존명사 외 5개

[표 3]과 같이 정의된 태그를 바탕으로 명사를 추출한다. 이는 한나눔 형태소 분석기 내에서 사용되는 Tag_Set 데이터를 바탕으로 구성하였으며, 명사와 관련된 태그정보를 포함하고 있다. 이를 통해 얻은 1차 명사 추출 정보는 총 380여개를 추출하였으며, 결과는 [표 4]와 같다.

표 4. 설화 스토리 원문 내에서 1차 명사 추출 태그 정의

무신, 도적, 옆구리, 비늘, 칼, 도적, 머리, 천장, 목, 공주들, 재, 목, 뿌리, 무신, 공주, 자신, 하인들, 바위, 자, 하인들, 공주, 임금님, 임금님, 하인들, 칭찬, 무신, 바위, 죽음, 구멍, 방법, …

1차 명사 데이터 추출 시 다양한 명사들이 추출되었으나, 중복 데이터가 많은 관계로 이를 최소화 시킬 필요성이 있으며 동시에 자연어 처리 분야에서 일반적으로 사용되는 TF(Term Frequency) 정보를 통해 불필요한 단어의 태그를 파악함으로써 2차 명사 추출(개체 추출)에 기반 데이터로 사용한다.

나. TF값 추출 및 2차 명사 추출 방법

TF 값 추출은 해당 단어의 출현빈도에 따라 단어의 중요도를 파악하기 위해 사용되는 일반적인 알고리즘이다. 본 논문에서 TF값에 중점을 두는 이유는 타 명사보다 스토리 내에서 주체가 되는 인물, 사건, 배경의 정보가 타 데이터 보다 높게 나올 가능성이 크기 때문에 TF값을 추출하게 된다. TF값 추출 결과는 [표 5]와 같다.

표 5. 명사별 TF값 계산 결과

명사	TF값	명사	TF값
공주	17	아귀	7
무신	15	그	6
도적	10	하인들	6
귀신	9	산	5
말	9	세상	5
사람	9	수	5
임금님	9	일	5
공주들	8	저	5

TF값 추출 후 총 명사의 개수는 기존 중복단어 제거 전보다 1/3가량이 감소하였으며, 115개의 명사로 축소되었다. 또한, 앞서 설명했던 인물, 사건, 배경 정보가 대부분 상위에 존재하는 것을 확인 할 수 있었다. 하지만, TF값 추출 후 1차 명사 추출 결과에서 불필요한 태그를 가진 명사가 파악되었다. 예를 들어, “수”, “일”, “저” 등의 경우 의존명사, 인칭대명사, 지시관형사 등으로 분류 되면서 전반적인 명사 추출 결과로 보았을 때 본 논문에서 추출하고자 하는 명사 데이터에 맞지 않은 것으로 파악되어 2차 명사 추출을 하게 된다. 또한, 한 단어의 품사가 한 개가 아닌 2개 이상인 단어가 존재한다. 이는 명확한 단어를 추출하는데 방해가 되는 단어들로 간주하여 불용어 명사로 정의하고 불용어 명사 DB를 구축한 후 2차 명사 추출 시 불용어로 제거를 하게 된다. [표 6]은 불용어 명사 목록을 나타내며 [표 7]은 명사 중 불용어 명사의 태그를 정의한 것이다.

표 6. 불용어 명사 정의

명사	명사	명사	명사
아구	아니	아니요	아무것
아나	아니나다를까	아닐세	아무때
아냐	아니야	아듀	아무려나
아녜요	아니에요	아름	아무렴
아뇨	어마	아무	아무리

표 7. 불용어 명사의 태그

태그명	설명	태그명	설명
xsnu	명사 파생 접미사	xsms	형용사 파생 접미사
xsnca	명사 파생 접미사	xsmn	형용사 파생 접미사
xsncc	명사 파생 접미사	xsam	부사 파생 접미사
xsna	명사 파생 접미사	xsas	부사 파생 접미사
xsns	명사 파생 접미사	ii	감탄사
xsnp	명사 파생 접미사	nbn	비단위성 의존명사
xsnx	명사 파생 접미사	nbu	단위성 의존명사
xsvv	동사 파생 접미사	nbs	비단위성 의존명사
xsva	동사 파생 접미사	npp	인칭대명사
xsvn	동사 파생 접미사	npd	지시대명사

[표 6], [표 7]의 불용어 태그 데이터를 기반으로 1차 과정에서 추출된 명사 데이터를 재분류하는 과정을 거친다.

```
public static void main(String[] args) {
 try {
         outputStream = new FileWriter("C:/my.txt");
         BufferedReader br = new BufferedReader(
         new InputStreamReader( new FileInputStream(dir1), "utf-8"));
       while((var_text = br.readLine()) != null){

if(var_text.contains("xsnu") ||
    var_text.contains("xsnca") ||
    var_text.contains("xsncc") ||
    var_text.contains("xsna") ||
    var_text.contains("xsns") ||
    var_text.contains("xsnp") ||
    var_text.contains("xsnx") ||
    var_text.contains("xsvv") ||
    var_text.contains("xsva") ||
    var_text.contains("xsvn") ||
    var_text.contains("xsms") ||
    var_text.contains("xsmn") ||
    var_text.contains("xsam") ||
    var_text.contains("xsas") ||
    var_text.contains("ii")||
    var_text.contains("nbn") ||
    var_text.contains("nbu") ||
    var_text.contains("nbs") ||
    var_text.contains("npp") ||
    var_text.contains("npd")) {

      outputStream.write(var_text);     }
        br.close();
   } catch(Exception e2){
     JOptionPane.showMessageDialog(null, e2);
   }
  }
}
```

그림 2 불용어 태그 정보를 활용한 명사 재추출 알고리즘

(그림 2)는 태그 정보를 바탕으로 2차 명사 추출을 하기 위한 알고리즘을 나타낸다. 이는 1차 명사 추출 시 해당 단어에 대한

품사 정보를 저장함으로써 시스템의 자원 소모를 최소화할 수 있다는 장점이 있다. 따라서, 본 논문에서 제안한 최종 인물정보는 [표 8]과 같다.

표 8. '지하국대적퇴치' 설화 내 인물, 사건, 배경 추출 결과

명사	TF	명사	TF	명사	TF
공주	17	집	4	대감님	2
무신	15	광주리	3	동정	2
도적	10	다음	3	땅	2
귀신	9	머리	3	마음	2
임금님	9	몸	3	목	2
공주들	8	소굴	3	물동이	2
아귀	7	수박	3	비늘	2
하인들	6	칭찬	3	소원	2
산	5	하인	3	아가씨	2
세상	5	결혼	2	엿구리	2
구멍	4	귀	2	의심	2
노인	4	나무	2	잠	2
막내	4	나뭇잎	2	치맛자	2
바위	4	납치	2	락	2

[표 8]의 결과를 살펴보면, "지하국대적퇴치"라는 설화 내 등장하는 인물정보는 "공주", "무신", "도적"이 주체가 되어 이야기를 이끌어가게 되는데 본 과정을 통해서 최종적으로 추출된 인물정보와 일치한다. 따라서, 설화 내 인물정보 추출은 매우 정확하게 추출된 것을 확인 할 수 있으며 최종 추출된 결과를 사용자가 쉽게 파악할 수 있도록 시스템화 하였다. 개발된 시스템은 (그림 3)과 같다.

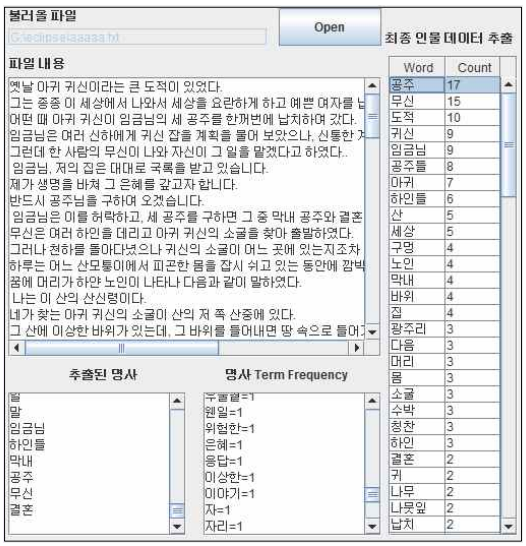

그림 3. 설화 콘텐츠 내 인물정보 추출 시스템

(그림 3)의 시스템의 구성은 설화 원천 데이터를 읽어들여 텍스트 전처리를 시행한 후, 본 논문에서 제안한 1차 명사 추출, 2차 명사 추출을 수행하고, 해당 명사 추출 결과를 사용자가 직접 파악함으로써, 스토리 창작에 도움이 될 수 있도록 리스트 형태로 표현하였다. 또한, 최종 인물 데이터 추출 결과를 바탕으로 각 인물 간 관련 단어를 추출하여 시각화하였다.

라. 시각화(Visualization)

인물 간 관련 단어 추출은 반자동형태로 추출되었으며, 추출 시 규칙은 첫째, 같은 문장내에 등장하는 명사를 추출한다. 예를 들면, "아귀 귀신이 임금님의 세 공주를 한꺼번에 납치하여 갔다."과 같은 문장이 있을 때, POS Tagging정보를 통해 명사를 추출하면 "아귀 귀신", "임금님", "세공주", "납치"와 같은 4개의 명사가 추출된다. 하지만, 4개의 명사 중 "납치"는 행위성명사로 분류되어 "아귀 귀신"과 "임금님", "세공주"간의 관계가 형성이 된다. 이를 바탕으로 문장 내 행위성 명사가 존재했을 시 기타 명사와 관련이 있는 것으로 정의하였다. 둘째는 앞서 추출한 TF 값을 바탕으로 2개 이상의 명사가 문장 내에 존재할 때 해당 명사들간에 관계가 있는 것으로 정의하여 관련된 단어로 판별 후 데이터를 표현하게 된다. [표 9]는 본 시스템에서 추출한 최종적으로 추출된 명사 간 관계단어를 나타낸다.

표 9. 인물 정보를 기반으로 한 관련 단어 리스트

인물명	관계단어	인물명	관계단어
아귀	산중	임금님	공주
아귀	바위	임금님	허락
아귀	땅속	임금님	막내공주
아귀	구멍	임금님	결혼
공주	수박	산신령	노인
공주	치맛자락	산신령	하얀머리
공주	문	산신령	말
공주	독한술	산신령	사라짐
무신	말	하인	새끼
무신	비늘	하인	광주리
무신	계책	하인	동정
무신	국록	하인	줄

[표 9]에서 정의한 인물 간 관계명사를 바탕으로 관련단어를 시각화 할 수 있는 모듈을 개발하였으며, 최종 추출 명사에 표현된 단어를 선택했을 시 표현한다. (그림 4)는 "지하국대적퇴치" 내 주체가 되는 각 인물과 관련된 단어를 시각화 한 모듈을 나타낸다.

그림 4. 주요 인물과 관련된 단어에 대한 시각화 예시

III. 결 론

본 논문에서는 다양한 스토리텔링 저작 지원 시스템에서 필요로 하는 한글 데이터 처리 부분 중 기존 스토리 내에 주체가 되는 등장인물, 사건, 배경을 자동으로 추출하여 등장인물을 중심으로 3가지 요소들과의 상관관계를 표현해주는 전반적인 방법을 제안하고 이를 기반으로 스토리텔링 저작 지원을 위한 텍스트 분석방법을 기술하였다.

그 결과, 기존 스토리 내에서 이야기를 이끌어가는 주체인 인물정보 및 이야기의 반전을 이끌어낼 뿐 아니라 인물과 인물 간 갈등 및 화해를 만들어주는 사건정보, 사건들이 일어나는 스토리 내 배경정보를 추출 할 수 있었다. 이를 토대로, 작가가 스토리를 작성하는데 있어 기존 스토리 내 존재하는 다양한 정보를 쉽게 볼 수 있으며, 창작을 위한 최소한의 수단이라고 볼 수 있다. 하지만, 전체적으로 자동화된 시스템이 아닌 반자동화 시스템인 것을 감안한다면 지속적인 개발 및 연구가 필요한 것으로 사료된다.

References

[1] 나성준, 김우중, 최이권, 전현택, "디지털 스토리텔링의 스토리 핵심 요소를 이용한 기획지원 온톨로지 검색 시스템 연구", 한국지능정보시스템학회, pp.338-242, 2010.

[2] 최향지, 류시천, "정보디자인에서 비주얼 스토리텔링의 이해", 스마트미디어저널, 제3권, 제2호, 29-36쪽, 2014년 6월.

[3] 남명희, 유은순, 이오준, 정재은, 황도삼, "디지털 콘텐츠 생태계에서의 미디어 간 변환에 대한 연구: 선행연구 중심으로", 한국스마트미디어학회 2015 추계학술대회, 1-3쪽, 조선대학교, 대한민국, 2015년 10월.

[4] Todd Cochrane, Hongwei Pan, Eddy Hui, "Three Little Pigs in a Sandwich: Towards Characteristics of a Sandwiched Storytelling based Tangible System for Chinese Primary School English", Proc. of 15th New Zealand Conference on Human-Computer Interaction (CHINZ), pp.5-8, Sep. 2015.

[5] Garbriele Kasper and Matthew T. Prior, "Analyzing Storytelling In TESOL Interview Research", TESOL Quarterly, Vol.49, No.2, pp.226 - 255, 2015.

[6] Frank van Gils, Potential applications of digital storytelling in education, Proc. of 3th Twente Student Conference on IT (TSConIT), pp.1 - 7, June 2005.

[7] 류철균, 서성은, "디지털 서사 창작 도구의 서사 알고리즘 연구 : <드라마티카 프로>를 중심으로", 현대소설연구, No.38, 117-152쪽, 2008년.

[8] 이은령, 김교정, "인터렉티브 스토리텔링 콘텐츠 저작지원도구 설계 및 구현에 관한 연구", 디지털융복합연구, 제11권, 제2호, 263-269쪽, 2013년.

[9]이주희, "한국어의 음절구조에 대한 연구 경향과 전망", 한국언어문학, 제61권, 57-81쪽, 2007년.

[10] http://semanticweb.kaist.ac.kr/home/index.php/HanNanum

[11] 홍승우, 이종연, 오상헌, "한글 어절 맞춤법 오류 검출을 위한 형태소 분석기", 한국정보과학회 1993년도 가을 학술발표논문집, 제20권, 제2호, 1143-1146쪽, 1993년 10월

[12] 백남주, "스토리텔링 저작을 위한 서사정보 추출 시스템 설계 및 구현", 조선대학교 산업대학원 소프트웨어융합공학과, 2015.

[13] 고병규, 최철웅, 나성희, 김판구, "Linked Data를 위한 한국어 자연어처리 플랫폼", 제24회 한글 및 한국어 정보처리 학술대회, 1-5쪽, 한국해양대학교, 대한민국, 2012년 10월.

[14] 유정훈, "워드넷 기반의 클러스터를 이용한 문서 요약 시스템 : 스포츠 뉴스 기사 헤드라인 추출", 단국대학교 석사학위논문, 2013.

[15] 고병규, 최철웅, 나성희, 김판구, "스토리텔링 저작 지원도구 개발에 관한 연구", 한국스마트미디어학회 2015 춘계학술대회, 1-3쪽, 숭실대학교, 대한민국, 2015년 4월.

저 자 소 개

고병규(학생회원)

2010년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2012년 조선대학교 컴퓨터공학과 석사 졸업(공학석사).
2015년 조선대학교 컴퓨터공학과 박사수료.
<주관심분야 : 시맨틱웹, 온톨로지, 자연어처리, IoT, 빅데이터 처리>

김정인(학생회원)

2011년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2014년 조선대학교 박사수료.
2015년 현재 조선대학교 석박사 연계과정
<주관심분문야 : 정보처리, 소셜네트워크, 시맨틱 웹, 온톨로지>

이은지(학생회원)

2012년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2015년 조선대학교 박사수료
2016년 현재 조선대학교 석박사 연계과정
<주관심분야 : 정보처리, 소셜네트워크, 시맨틱 웹, 온톨로지>

김판구(정회원)

1988년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
1990년 서울대학교 컴퓨터공학과 석사 졸업(공학석사).
1994년 서울대학교 컴퓨터공학과 박사 졸업(공학박사).
1994년 ~ 현재 조선대학교 교수
<주관심분야 : 정보검색, 시맨틱웹, 자연어처리, 빅데이터>