http://dx.doi.org/10.5392/JKCA.2014.14.03.040

구문의미분석를 이용한 유사문서 판별기

Discriminator of Similar Documents Using Syntactic and Semantic Analysis

강원석*, 황도삼**, Jung H Kim***
안동대학교 정보과학교육과*, 영남대학교 컴퓨터공학과**, University of Arkansas at Little Rock***

Won-Seog Kang(wskang@andong.ac.kr)*, Do-Sam Hwang(dshwang@yu.ac.kr)**,
Jung H Kim(jhkim@ualr.edu)***

요약

문서 저작권에 대한 관심과 중요성이 높아짐에 따라 문서 복제나 표절의 검출에 대한 필요성이 증대되고 있다. 이와 같은 이유로 많은 연구가 이루어지고 있으나 자연어 처리기술의 한계가 있어 문서의 심층적 표절 검출에 어려움이 있다. 본 논문은 자연어 분석의 기술을 적용한 유사문서 판별기를 설계, 구현한다. 이 시스템은 형태소 분석의 기술과 함께 구문의미 분석의 기술, 저빈도 및 관용표현 가중치을 이용하여 유사문서를 판별한다. 본 시스템의 성능을 실험하기 위하여 휴먼 판별과 기존 시스템, 그리고 휴먼 판별과 제안한 시스템의 판별과의 상관계수를 분석하였다. 실험결과, 구문의미 분석을 활용한 시스템의 개선점을 발견할 수 있었다. 앞으로 문서 유형을 정의하고 각 유형에 맞는 판별 기법을 개발할 필요가 있다.

■ 중심어 : | 유사문서 판별 | 자연어 처리 |

Abstract

Owing to importance of document copyright the need to detect document duplication and plagiarism is increasing. Many studies have sought to meet such need, but there are difficulties in document duplication detection due to technological limitations with the processing of natural language. This thesis designs and implements a discriminator of similar documents with natural language processing technique. This system discriminates similar documents using morphological analysis, syntactic analysis, and weight on low frequency and idiom. To evaluate the system, we analyze the correlation between human discrimination and term-based discrimination, and between human discrimination and proposed discrimination. This analysis shows that the proposed discrimination needs improving. Future research should work to define the document type and improve the processing technique appropriate for each type.

■ keyword : | Similar Document Detection | Natural Language Processing |

I. 서 론

최근 저작권의 보호를 위해 표절과 저작권 침해에 대한 탐지, 방지의 기술의 연구도 그 필요성에 따라 활발하게 진행되고 있다. 그러나 자연어 분석의 문제점을 안고 있어 질높은 문서표절탐지 기술 개발의 어려움이 있다. 본 연구는 이와 같은 문제를 해결하기 위하여 자연어 분석의 기술을 적용하고자 한다.

* 이 논문은 2011년 안동대학교의 학술연구조성비에 의하여 지원되었음.
접수일자 : 2013년 12월 18일
수정일자 : 2014년 02월 24일
심사완료일 : 2014년 02월 24일
교신저자 : 강원석, e-mail : wskang@andong.ac.kr

[1-4]는 프로그램 표절 검사에 대한 연구이다. 일반 문서와 달리 프로그램 코드는 형식 언어로 표기되고 용어 중심의 표절검사 기법이 사용된다. 따라서 일반 문서 표절 검사에 이 방법을 적용하기에는 어려움이 있다.

[5-8]의 연구는 영어권의 표절검사 방법에 대한 것이다. 한국어는 영어와 언어적 특징이 다르다. 따라서 한국어 문서에 대한 유사문서 판별에는 앞에서 언급한 방법을 그대로 적용할 수 없다. 한국어 유사문서 판별에는 한국어 고유의 특성을 반영한 판별 방법이 필요하다.

[9]는 미국 아이패러다임 회사에서 개발한 시스템으로 대용량의 데이터를 기반으로 문서표절을 검사하는 상용시스템이나 한국어가 지원되지 않는다.

[10][11]의 연구는 XML 기반의 문서 구조에 초점을 맞추어 유사문서를 식별한 것으로 일반 문서에 적용할 수 없고 변환을 통해 적용하더라도 문서 내용의 의미적 유사성의 식별에는 한계가 있다.

[12]의 연구는 한글의 형태소적인 특징을 이용하여 문서를 축약하는 기술을 적용하여 문서 표절 검사의 성능을 개선하고자 하였다. [13]의 연구는 신문기사에서 구문을 추출하고 그 구문을 질의로 웹검색을 실시하여 그 결과 값을 이용하여 문서의 표절여부를 검사하는 연구를 하였다. 이 연구는 문서의 내용에 근거하기보다 웹검색의 결과를 이용하므로 인용을 많이하는 신문기사가 아닌 일반 문서에 적용하기에는 어려움이 있다.

[14]의 KUREPOLS 시스템은 형태소 해석을 이용하여 문서에서 검색어를 추출한 후 웹검색을 통해 검색된 문서와 유사도를 계산하여 표절여부를 판정하였다. [15]의 DEVAC 시스템은 어절을 기본 단위로 문서간의 유사성을 검사하는 표절검사 시스템으로 어절의 변형이나 유사어절의 경우 표절검사가 쉽지 않다.

[16]은 형태소 분석을 이용하여 추출한 형태소를 기반으로 문서간 표절 정도를 나타내는 표절지수를 정의하고 이를 계산하여 표절을 검사하였다. [17]의 연구는 조사가 추가된 어절을 중심으로 벡터 유사도를 계산하여 문서 표절을 검사하였고 [18]은 형태소 분석을 이용하여 추출한 명사를 기반으로 신경망을 이용하여 문서간 유사도를 계산하였다. [19]의 연구는 형태소 해석을 통해 추출한 용어 가운데 특별한 용어를 선별하는 방법을 사용하여 문서 유사도를 개선하는 방법을 연구하였다. [20]의 연구는 표절 유형에 견고한 유사문서판별을 위하여 형태소 분석을 통해 추출한 명사를 기반으로 벡터 유사도를 계산하는 모델을 제안하였다. 이 연구들은 문서의 유사성을 검사하기 위하여 다양한 방법을 적용하고 있으나 기본적으로 형태소 해석 기술을 이용한 용어 중심의 방법을 적용한 것이다(여기에서 형태소란 의미를 지니는 가장 작은 단위소를 말한다).

형태소 해석 기술을 이용한 용어 중심의 방법은 문서가 가지고 있는 의미를 표현하는데 한계가 있다. 이 방법은 의미를 지니는 단어의 출현에 의존한다. 즉, 단어와 단어의 구문의미적 관계 등의 정보를 표현할 수가 없다(구문 의미적 관계는 단어와 단어가 결합하여 구를 이루는 관계이거나 단어와 단어가 결합하여 의미적 구를 이루는 관계를 말한다). 문서의 문장의 뜻은 단어들의 나열로 결정되는 것이 아니라 단어들간의 구문 의미관계에 의해 결정된다. 따라서 이 문제를 해결하기 위해서는 단어와 단어간의 구문의미관계를 분석할 수 있는 방법이 필요하다. 본 연구에서는 구문의미분석기를 이용하여 문장 속에 들어 있는 단어간의 구문 관계와 격의미 관계를 찾아내고 이를 표현하여 문서표절 검사의 성능을 향상하고자 하였다.

[18][19]에서는 저빈도 단어의 경우 가중치를 부여하여 유사문서 식별에 활용하였다. 본 연구에서는 저빈도 단어뿐 아니라 저빈도 구문의미구조에 대해서도 가중치를 부여하고 특별한 패턴인 관용구에 대해서도 가중치를 부여하여 시스템을 설계, 구현하였다.

본 연구의 2장에서는 본 시스템에서 사용한 용어 추출 방법과 구문격의미 관계 분석 방법에 대해 서술하고 3장은 구문관계를 이용한 유사문서 식별 시스템을 서술한다. 4장에서는 제안한 시스템의 성능을 분석하기 위하여 사람이 판별한 휴먼 판별결과와 본 시스템의 판별 결과의 상관관계를 분석하고 검토한 후 5장에서 결론을 지었다.

II. 문서 해석 기술

1. 형태소 해석

정보검색이나 문서 분류 시스템들은 용어 중심의 방법을 사용하고 있다[21]. 용어 중심의 방법은 용어를 기반으로 시스템이 구성되므로 문서의 표현 언어인 한국어의 형태소 레벨의 분석이 전제되어야 한다. 한국어는 어절 단위의 띄어쓰기를 하고 있기 때문에 어절에 대한 형태소 해석이 진행된다. 본 연구에서는 [22]의 형태소 해석 시스템을 이용하여 어절 단위의 형태소 해석을 실시한다. 처리하는 한국어 어절의 유형은 체언, 체언+조사, 체언+접미사, 체언+체언, 체언+체언+조사, 체언+접미사+조사, 체언+체언+접미사, 체언+체언+접미사+조사, 용언+어미, 용언형체언+용언화접미사+어미 등이다. 여기서 용언형체언은 '건설'과 같이 '하다'가 붙어 용언이 될 수 있는 체언을 말하고 용언화접미사는 '하다' 또는 '되다'와 같이 용언형체언에 부착되어 용언으로 사용할 수 있도록 하는 접사를 말한다. 형태소 해석의 결과의 예는 3장에 기술한다.

2. 구문의미 분석

문서의 유사성을 계산하기 위해서는 먼저 문서의 정보를 표현하고 그 표현된 것을 비교하여야 한다. 용어 중심의 표현 방법은 문서가 가지고 있는 정보를 표현하는데 너무 부분적이다. 즉, 용어 중심의 방법은 문서를 용어의 나열만으로 나타내기 때문에 문서 속에 들어 있는 많은 정보를 표현하기에 제약이 많다. 문서의 정보는 문장으로 구성되고 있고 문장은 단어들의 구문의미관계를 근거로 조합되어 뜻을 나타낸다. 따라서 문서의 정보를 표현하기 위해서는 단어 뿐 아니라 단어들의 구문의미관계 정보도 필요하다. 본 연구도 이와 같은 점에 초안하여 구문의미 분석을 하고 이를 유사문서 판별에 활용하고자 하였다.

'문장을 분석하여 구문 관계를 찾아낸다.'라는 문장의 예를 들어 구문의미분석을 설명한다. 예 문장에서 '분석하여'라는 구성성분은 '문장을'이라는 성분과 목적관계를 맺는다. 그리고 '찾아낸다'라는 문장성분은 '구문관계를'이라는 성분과 역시 목적관계를 맺는다. 그리고 '문장을 분석하여'라는 성분은 '구문관계를 찾아낸다'라는 성분과 종속관계를 맺는다. 예 문장의 분석결과는 다음과 같은 트리로 표현할 수 있다.

(찾아내다 (OBJ 관계 (BLA 구문)) (SUCO 분석하여 (OBJ 문장을))

구문의미 분석시에 각 구문의미 관계에 따른 제한조건을 검사하여 해당하는 구문의미 관계를 찾는다. 이 과정에서 애매성이 발생할 경우 다음 세 가지 규칙에 따라 애매성을 해결한다.

- one role

구문의미 트리에서 헤드(루트)의 부속성분인 각 문장성분의 구문의미관계는 자신을 특징짓는 구문의미관계로서 중복되는 두 개의 부속성분을 가질 수 없다는 규칙이다. 이것은 서술어가 주어, 목적어 등의 문장성분을 하나씩 가진다는 점에 근거한 것이다.

- locality

부속 성분이 여러 개의 헤드와 관계를 지을 수 있는 상황의 경우 근접한 것을 선택한다는 규칙이다. 일반적으로 문장을 표현하는데 연관된 내용은 가까운 것에 둔다는데 근거한 것이다.

- no crossing

부속성분과 관계를 짓는 헤드와의 구문의미관계가 다른 구문의미관계와 교차되지 않아야 한다는 규칙이다. 한 부속성분이 연관될 헤드를 찾을 때 이미 구성된 부속성분과 헤드와의 연관관계 사이로 교차되는 것을 피한다.

[표 1]은 본 연구에서 정의한 구문의미관계이다. 각 구문의미 관계에 대한 내용과 제한 조건은 다음과 같다.

(1) 주격(SUBJ) : 문장 성분이 주격 역할을 하는 것으로 체언에 격조사 '가', '이'가 첨부된 것으로 용언과 구문의미관계를 맺는다.

(2) 목적격(OBJ) : 문장 성분이 목적격 역할을 하는 것으로 체언에 격조사 '을', '를'이 첨부된 것으로 용언과 구문의미관계를 맺는다.

(3) 종속명사(BLA) : 명사와 명사가 만나 복합 명사가 되는 경우를 표현한 관계이다. 종속 명사 관계는 격조사 없이 단독으로 나타난 명사가 연이어 명사가 올 경우 해당한다. 예를 들면 '구문' '분석' '시스템이'가 연이어 나타난다면 '구문'과 '분석'은 종속명사로서 '시스

템이'에 종속된다. 종속명사는 결국 최종 연이은 명사 가운데 격조사가 붙은 명사에 종속관계를 갖는다.

표 1. 본 시스템의 구문/의미 관계

명칭	구문/의미관계	격조사, 어미
SUBJ	주격	가, 이
OBJ	목적격	을, 를
BLA	복합명사	blank
TARG	방향격	로, 으로, 에
ADNM	관형격	의, 은, ㄴ
TOP	주제	는, 은
STAT_AS	자격	로, 로서
LOC_IN	장소	에
SOUR_FROM	근원격	에서,부터
SRPT_FROM	장소출발	부터
SRTI_FROM	시간출발	부터
SUCO	종속	면, 여
COCO	연결	며, 고
QUCO	인용	라고, 고
TARG_TO	대상	에게
COMP_WITH	경쟁	와, 과
JUNC	접속	와, 과
VOC	호격	야
COMP	보격	가, 이, 와
TIME_IN	시간	에
INST_WITH	도구	로
PERI_DURING	기간	동안
MEAS_IN	측정	로
SITU_IN	상황	에서

(4) 방향(TARG) : 체언에 격조사 '로', '으로', '에'가 붙고 체언의 의미가 direction, location, structure 가운데 하나인 경우이다. 체언에 대한 의미의 검사는 [23]의 시소러스 사전을 이용하여 구한 후 해당하는 의미가 포함된다면 방향 구문의미관계로 결정한다.

(5) 관형(ADNM) : 관형은 두 가지 종류로 구분할 수 있다. 체언과 격조사 '의'가 결합된 어절의 구문의미관계인 경우와 용언과 관형형어미 'ㄴ', '는'이 결합된 어절의 구문의미관계인 경우이다. 두 경우 모두 격조사나 관형형어미의 구분태그 정보와 연이어 나온 문장성분이 체언인 것을 검사하여 결정한다.

(6) 주제(TOPIC) : 체언과 격조사 '은', '는'이 첨가된 성분의 구문의미관계이다.

(7) 자격(STAT_AS) : 격조사 '로', '로서'가 첨부된 문장 성분의 체언의 의미가 state, organization, situation 가운데 하나가 되는 경우이다.

(8) 장소(LOC_IN) : 격조사 '에', '서', '에서'가 첨부된 문장 성분의 체언의 의미가 location, structure 가운데 하나가 들어 있고 이 문장성분과 연결되는 용언의 의미가 moving이나 positioning 일 경우이다.

(9) 근원(SOUR_FROM) : 격조사 '부터', '에서'가 첨부된 문장 성분의 체언의 의미가 location이나 time의 의미가 포함되어 있지 않은 경우 근원 구문의미관계로 판단한다.

(10) 장소출발(SRPT_FROM) : 격조사 '부터', '에서'가 첨부된 문장 성분의 체언의 의미가 location이 포함된다면 장소출발의 구문의미관계이다.

(11) 시간출발(SRTP_TIME) : 격조사 '부터', '에서'가 첨부된 문장성분의 체언의 의미에 time이 포함되어 있다면 시간출발의 구문의미관계이다.

(12) 종속(SUCO) : 어미 '면', '여' 등이 첨부된 용언의 주성분이 되는 문장성분은 다른 용언의 문장성분과 종속관계를 맺는다. 본 시스템에서는 구문의미해석의 범위를 단문으로 한정하였기 때문에 용언과 어미 등의 정보만으로 문장성분의 구문의미역할만 파악하는 것으로 하였다.

(13) 연결(COCO) : 어미 '며', '고' 등이 첨부된 용언의 문장성분의 구문의미역할로 용언과 용언이 연결되는 관계를 말한다.

(14) 인용(QUCO) : 어미 '라고', '고' 등이 첨부된 용언의 문장성분의 구문의미역할로 인용하는 관계를 의미한다.

(15) 대상(TARG_TO) : 격조사 '에게'가 첨부된 문장성분의 체언의 의미가 human, organization의 의미가 포함되어 있을 경우 대상 구문의미관계로 판단한다.

(16) 경쟁(COMP_WITH) : 격조사 '와', '과'가 첨부된 문장성분의 체언이 용언과 종속적인 관계를 맺는 경우이다. 이때 주성분이 되는 용언의 의미에 competition의 의미가 포함되어야 한다.

(17) 접속(JUNC) : 격조사 '와', '과'가 첨부된 문장성분이 연이어 나온 문장성분이 명사인 경우 그 명사와 접속의 구문의미관계를 맺는다.

(18) 호격(VOC) : 격조사 '이야', '야'가 첨부된 문장성분으로 호칭에 사용하는 구문의미관계이다.

(19) 보격(COMP) : 격조사 '가', '이'가 첨부된 문장성분이 용언과 관계를 맺는 구문의미관계로 용언이 이

미 주격 문장성분을 가지고 있는 경우에 해당한다.

(20) 시간(TIME_IN) : 격조사 '에'가 첨부된 문장성분의 체언이 time의 의미를 포함하는 경우 시간 구문의미관계로 판단한다.

(21) 도구(INST_WITH) : 격조사 '로', '으로'가 첨부된 문장성분의 체언이 tool의 의미를 포함하는 경우 도구 구문의미관계로 판단한다.

(22) 기간(PERI_DURING) : 격조사 '에'가 첨부된 문장성분의 체언이 duration의 의미를 포함하는 경우 기간 구문의미관계로 판단한다.

(23) 측정(MEAS_IN) : 격조사 '로', '으로'가 첨부된 문장성분의 체언이 measurement의 의미를 포함하는 경우 측정 구문의미관계로 판단한다.

(24) 상황(SITU_IN) : 격조사 '서', '에서'가 첨부된 문장성분의 체언이 situation의 의미를 포함하는 경우 상황 구문의미관계로 판단한다.

위의 구문의미관계를 검사할 때 사용하는 의미는 [23]의 상하위의미구조와 시소러스 사전을 사용한다. 상하위 의미구조는 세 개의 트리로 구성된다. 그 일부는 [그림 1]과 같다.

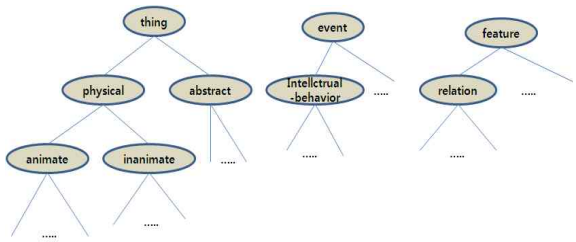

그림 1. 상하위 의미 체계의 일부

단어 '구문'의 예를 들어 의미 추출 과정을 보자. 먼저 의미사전을 검사하여 '구문' 단어의 의미를 추출한다. '구문' 단어는 word 의미를 가지고 있다. 다음으로 검색된 의미 word 에 대해 상하위 의미구조체계의 상위의미를 가져온다. 즉, word 의 상위의미는 intellectual-thing, 그 위의 상위의미는 abstract-thing, 그 위의 상위의미는 thing이다. 이를 차례로 가져온 후 '구문' 단어의 의미집합으로 결과를 낸다. 이 의미집합과 제한조건의 의미와의 교집합이 공집합이 아니면 해당하는 의미를 가지고 있는 것이 된다.

III. 구문의미 분석을 이용한 유사문서 식별 시스템

본 연구의 유사문서 식별 시스템의 구조는 [그림 2]와 같다.

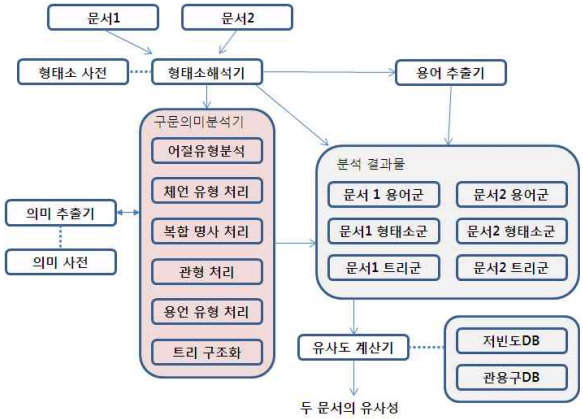

그림 2. 유사문서 식별 시스템 구조

유사문서 식별 시스템은 크게 형태소해석기, 구문의미분석기, 유사도계산기, 용어추출기, 의미추출기로 구성되고 형태소 사전과 의미 사전, 저빈도 DB, 관용 DB를 활용한다. 형태소 해석기와 형태소 사전은 [22]을 이용하였다. 구문의미 분석기는 어절유형분석, 체언유형처리, 복합명사처리, 관형처리, 용언유형처리, 트리구조화의 단계로 구성된다. 의미사전과 의미 추출기는 구문의미분석의 각 단계에서 필요한 조건을 검사하기 위하여 단어에 대한 의미를 찾을 때 사용한다. 분석 결과물은 형태소 해석과 구문의미분석에서 얻어진 것으로 용어추출기를 통해 얻은 용어군, 형태소 해석의 결과로 얻은 형태소군, 구문의미분석으로 얻은 트리군이다. 이것은 문서의 유사성을 검사하기 위한 두 문서에 대한 결과물이므로 문서1에 대한 것과 문서2에 대한 것으로 구성된다. 유사도계산기는 두 문서의 분석결과물을 비교하여 유사도를 구하는 것이다. 본 연구에서는 각 결과물에 대한 가중치를 반영하여 복합적인 유사도를 계산하여 유사문서를 식별한다. 이때 저빈도 DB와 관용구 DB를 이용하여 가중치를 반영한다. 각 내용에 대해 단계적으로 서술한다.

1. 구문의미분석기

형태소 해석기는 띄어쓰기로 구분된 어절에 대해 그 구성구조를 분석한다. 분석된 결과물은 다음과 같다.

입력 문장 예	시스템의 입력 장치가 자료를 입력한다.
형태소 해석 결과물	시스템/ncn+의/jcm 입력/ncn 장치/ncn+가/jcs 자료/ncn+를/jco 입력/ncpa+하/xsv+ㄴ다/ef+./sf

구문의미분석은 형태소 해석기가 분석한 어절을 대상으로 어절간의 구문의미 연관관계를 밝힌다. 그 첫 단계는 구문의미분석의 단위가 되는 어절에 대한 유형분석이다. 어절은 체언 또는 체언과 조사로 구성된 체언유형 또는 용언과 어미로 구성된 용언유형으로 구분된다. 이의 구분은 형태소 해석 결과물에 들어있는 태그로 구분한다.

구문의미분석의 다음 단계는 체언 유형 처리이다. 이 단계에서는 체언이 들어있는 어절의 형태소들 간의 연관관계를 고려하여 트리로 구조화한다. 구조화한 결과는 리스트로 표현한다. 리스트는 (헤드역할태그 단어 구문의미관계태그 관계사 부속리스트)로 정의한다. 헤드역할태그는 HNOUN, HNXS와 같이 체언유형인지 용언유형인지를 구분하는 것이고 구문의미관계태그는 [표 1]에 기록된 SUBJ, BLA, OBJ 등과 같이 구문의미관계를 나타낸다. HNOUN은 단어가 명사인 경우이고 HNXS는 용언형체언과 접사, 어미가 붙은 경우를 의미한다.

체언유형 처리 결과	(HNOUN 시스템 ADNM 의) (HNOUN 입력 BLA) (HNOUN 장치 SUBJ 가) (HNOUN 자료 OBJ 를) 입력/ncpa+하/xsv+ㄴ다/ef+./sf

다음 단계는 복합명사 처리단계이다. 이것은 연이은 두 명사가 하나의 복합명사로 사용될 경우를 처리한 것이다. 복합 명사는 그 특성상 수식받는 뒤 명사가 head 역할을 한다. 그래서 트리로 구성할 때 뒤 명사를 부모가 되게 하고 앞 명사를 자식이 되게 한다. 복합 명사의 앞 명사는 격조사가 붙지 않은 BLA 태그를 달고 있으므로 이를 찾아내어 트리로 구조화한다. 그 결과는 다음과 같다.

복합명사 처리결과	(HNOUN 시스템 ADNM 의) (HNOUN 장치 SUBJ 가 (HNOUN 입력 BLA)) (HNOUN 자료 OBJ 를) 입력/ncpa+하/xsv+ㄴ다/ef+./sf

다음 단계는 관형 관계 처리단계이다. 관형 관계는 두 경우가 있다. 관형격조사가 붙은 체언유형의 경우와 관형형어미가 붙은 용언유형의 경우이다. 본 연구에서는 구문의미분석의 범위를 단문으로 한정하였기 때문에 체언 유형의 경우만 처리하고 용언은 단지 그 역할만 표기하고 수식관계는 처리하지 않는다. 관형 관계는 연이어 나타나는 체언 유형의 문장성분을 수식하는 것이므로 해당하는 문장성분을 피수식 체언유형의 문장성분의 부속성분으로 구조화한다. 그 결과는 다음과 같다. 이때 어절 '시스템의'가 바로 뒤의 '입력'을 수식하는 것이 아니라 '장치가'를 수식하게 된다. 이것은 복합명사를 먼저 구성한 후이므로 복합명사를 하나의 단위로 인식하기 때문이다.

관형 처리 결과	(HNOUN 장치 SUBJ 가 (HNOUN 입력 BLA) (HNOUN 시스템 ADNM 의)) (HNOUN 자료 OBJ 를) 입력/ncpa+하/xsv+ㄴ다/ef+./sf

다음 단계는 용언 유형 처리 단계이다. 용언 유형은 크게 두 가지로 구분한다. 하다가 붙지 않는 일반동사 유형과 하다가 붙는 용언형체언 유형이다. 일반동사유형은 HVERB 태그를 부착하고 용언형체언은 HNXS 태그를 부착하여 리스트로 나타낸다. 그 결과는 다음과 같다.

용언 유형 처리 결과	(HNOUN 장치 SUBJ 가 (HNOUN 입력 BLA) (HNOUN 시스템 ADNM 의)) (HNOUN 자료 OBJ 를) (HNXS 입력 FINEN ㄴ다)

최종으로 리스트로 표현된 문장 성분이 어느 용언과 구문의미관계를 맺는지를 찾아내어 그 용언의 부속성분으로 구조화하는 단계이다. 이 결과가 최종 구문의미트리가 된다. 이전 단계에서는 형태소 해석의 결과를 입력받아 트리로 구조화하여 리스트로 표현하였다. 구성된 트리는 어절 단위로 구성된 것이고 복합 명사나 관형 유형의 경우 트리 간의 합성이 일어나서 하나의 트리로 구조화되었다. 이 단계에서는 각 체언 유형의 트리가 어느 용언유형의 트리와 합성할 것인지를 결정한다. 이때 반영되는 규칙은 2장에서 언급한 one role, locality, no crossing 규칙이다. 본 논문의 연구에서는 용언유형과 용언유형과의 구문의미관계에 대한 복문에 대한 범주는 다루지 않았다. 최종 처리 결과는 다음과 같다.

결과 구문트리	(HNXS 입력 FINEN ㄴ다 (HNOUN 장치 SUBJ 가 (HNOUN 입력 BLA) (HNOUN 시스템 A DNM 의)) (HNOUN 자료 OBJ 를))

2. 유사도 계산기

본 연구에서 추구하고자 한 것은 용어만으로 문서를 표현하는데 한계가 있으므로 구문의미분석을 통해 얻은 구문의미트리를 용어와 함께 문서를 표현하는 것으로 정의한다면 문서의 표현에 더 정확성을 도모할 수 있어 유사문서 식별의 성능을 향상할 수 있다고 본다. 또한 사용이 빈번하지 않는 단어가 두 문서에 동시에 사용되었다면 두 문서는 유사문서일 가능성이 높다는 점에서 착안하여 빈번하지 않는 단어의 사용에 가중치를 부가한다면 시스템의 성능을 향상할 수 있다고 본다. 덧붙여 특별한 것으로 간주되는 관용구 사용에 대해서도 가중치를 부가하여 유사도 계산기를 설계, 구현하였다.

따라서 본 연구에서는 형태소 해석과 구문의미분석의 결과로 용어군, 형태소군, 트리군을 얻고, 시스템의 성능을 향상하기 위하여 각 결과에 대한 가중치를 부여하고, 추가로 빈번하지 않은 단어 사용과 관용표현의 사용에 대한 가중치를 부가하여 유사도 계산기를 구현하였다.

이 과정에서 사용한 저빈도 DB와 관용구 DB는 [24]의 말뭉치와 관용사전을 자료로 본 연구에서 구축된 용어추출기, 형태소해석기, 구문의미분석기를 이용하여 추출한 후 구축되었다. 각 DB들은 종류별로 용어군, 형태소군, 구문/의미트리군 DB로 나뉘어진다.

본 연구에서는 입력되는 두 문서에 대해 결과로 얻은 용어군, 형태소군, 구문의미트리군에 대한 표현을 다음과 같은 벡터 형태로 정의한다.

$$D_i = (a_{i1}, a_{i2}, a_{i3}, \ldots\ldots, a_{in}),\ i \in \{\text{용어, 형태소, 트리}\}$$
$$a_{ij} = \frac{freq_{ij}}{\sum_{k=1}^{n} freq_{ik}},\ n = \#\ of\ unique\ term \qquad (1)$$

Di는 용어군이나 형태소군, 트리군의 결과 가운데 하나를 표현한다. 용어들이 반복해서 나타날 수 있으므로 그 빈도수를 이용하여 가중치를 정하였다. 예 문장 '시스템의 입력 장치가 자료를 입력한다'의 경우 용어들은 시스템, 입력, 장치, 자료가 되고 벡터표현 $D_{용어군}$는 (1/5, 2/5, 1/5, 1/5)로 된다. 형태소들은 (시스템, ncn), (입력, ncn), (장치, ncn), (자료, ncn), (입력, ncpa)가 되고 벡터표현 $D_{형태소군}$는 (1/5, 1/5, 1/5, 1/5, 1/5)가 된다. 트리는 (HNXS 입력 FINEN ㄴ다 (HNOUN 장치 SUBJ 가 (HNOUN 입력 BLA) (HNOUN 시스템 ADNM 의)) (HNOUN 자료 OBJ 를))가 되고 벡터표현 $D_{트리군}$는 (1/1)로 된다. 물론 문서들은 여러 문장으로 구성되므로 분모가 1보다 크게 된다.

분석 결과물 용어, 형태소, 트리는 서로 구조가 다르다. 따라서 두 문서 A, B의 유사성을 비교할 때 문서 A의 용어군 표현은 B의 용어군 표현과 비교를 하고 A의 형태소군 표현은 B의 형태소군 표현과 비교를 한다. 유사도 계산값을 얻으면 그 유사도 계산값을 복합한 식을 유도하여 최종 유사도를 결정짓는다. 같은 군 표현의

유사도 계산식 sim과 각 군의 가중치를 반영한 유사도 계산식 tsim은 다음과 같다.

$$sim(D_{1i}, D_{2i}) = \frac{D_{1i} \cdot D_{2i}}{|D_{1i}| \times |D_{2i}|} = \frac{\sum_{j=1}^{n} a_{1ij} \times a_{2ij}}{\sqrt{\sum_{j=1}^{n} a_{1ij}^2} \times \sqrt{\sum_{j=1}^{n} a_{2ij}^2}}, (2)$$

$i \in \{$용어, 형태소, 트리$\}$

$$tsim(D_1, D_2) = \sum_{j \in \{용어, 형태소, 트리\}} c_j * sim(D_{1j}, D_{2j}), (3)$$

$c_j = weight\ constant\ of\ each\ category$(용어, 형태소, 트리)

본 연구에서 제안한 저빈도의 용어와 관용구 표현에 대한 가중치를 구하는 과정은 다음과 같다. 먼저 문서1과 문서2에 대한 용어, 형태소, 트리를 구한다. 다음으로 용어, 형태소, 트리 군에서 저빈도 DB 검색을 통해 검색되는 용어, 형태소, 트리를 찾는다. 검색된 용어, 형태소, 트리에 대해 유사도 비교식 (2) 와 (3)을 적용한다. 다음으로 관용구 표현도 이와 같은 순서를 따른다. 즉 비교대상의 문서에 대한 용어, 형태소, 트리를 구한 후 관용구 DB 검색을 통해 검색된 용어, 형태소, 트리에 대해 유사도 비교식 (2)와 (3)을 적용한다. 그리고 (4)식과 같이 정의된 최종 유사도 비교식을 통해 두 문서의 유사성을 결과로 낸다.

$$fsim(D_1, D_2) = \sum_{k \in \{기본, 저빈도, 관용\}} f_k * tsim(D_{k1}, D_{k2})$$
$$= \sum_{k \in \{기본, 저빈도, 관용\}} f_k * \sum_{j \in \{용어, 형태소, 트리\}} c_j * sim(D_{k1j}, D_{k2j}), (4)$$

$f_j = weight\ constant\ of\ each\ category$(기본, 저빈도, 관용)

두 문서가 유사한 지의 판단은 유사성 계산 값이 일정한 한계값을 초과하면 유사문서로 판단된다. 한계값은 사용자에 따라 달리 정의할 수 있다. 본 논문에서는 구문해석을 사용한 유사문서 식별기가 질높은 유사문서 식별에 영향을 주는지에 대해 실험하기 위해 한계값을 정의하지 않고 유사성 계산값이 얼마나 인간이 결정한 값과 유사한지를 검사하였다.

IV. 실험 및 분석

유사문서 식별의 판정은 유사문서의 기준을 어떻게 잡느냐에 따라 달라진다. 따라서 다양한 유사문서 식별 시스템마다 기준이 다르므로 다른 시스템과 비교하기 곤란하다. 본 논문의 실험에서는 제안한 시스템이 기본 시스템보다 얼마나 질적 성능이 좋아졌는지를 검사하였다.

본 연구의 시스템의 실험을 위하여 검사 문서를 64쌍 정하였다. 검사 문서는 전산분야의 학생들의 전공과목 보고서에서 발췌한 것으로 그중 36(56%)쌍이 유사문서이고 나머지 28쌍(44%)이 비유사문서이다. 각 문서는 평균 350바이트 정도의 길이를 가졌다. 제안한 구문해석기의 효과를 확인하기 위하여 다음과 같이 가중치를 선정하여 실험을 하였다. 가중치는 각 시스템의 효과를 보기위해 점층적으로 선정하였다. 시스템 A는 용어만을 사용한 시스템이고 시스템 B는 용어와 형태소를 복합하여 사용한 것이고 시스템 C는 용어와 구문트리를 복합하여 사용한 것으로 정의하였고 마지막 시스템 D는 용어, 형태소, 구문트리를 복합하여 사용한 것이다. 즉, 기본 용어만을 사용한 시스템을 기준으로 각 요인의 추가에 따른 효과를 보기위해 가중치를 단계적으로 추가하여 가중치의 최대합이 10이 넘지 않도록 정의하였다. 그 각 경우에 대한 내용과 그 실험 결과는 표와 같다.

표 2. 형태소, 구문의미분석 이용한 실험 설명

	시스템 A	시스템 B	시스템 C	시스템 D
가중치 cj (용어, 형태소, 트리)	(4 0 0)	(4 4 0)	(4 0 4)	(4 4 2)
설명	용어만 이용한 유사문서 식별	용어와 형태소 이용한 유사문서 식별	용어와 구문의미해석 이용한 유사문서 식별	용어, 형태소, 구문의미해석 이용한 유사문서 식별

표 3. 시스템 실험 결과

시스템종류	1번쌍 유사도비교 결과	2번쌍 유사도 비교결과	...	피어선상관계수
휴먼계산	.95	.30		
시스템 A	.16	.07		0.930773
시스템 B	.32	.16		0.953939
시스템 C	.32	.19		0.939971
시스템 D	.40	.21		0.960382

각 시스템의 결과를 분석하기 위하여 피어선 적률 상관계수를 사용하였다. 상관계수는 시스템의 분석 결과가 기준과 얼마나 일치되는지를 보여주는 것이다. 기준의 정의는 앞에서 언급한 바와 같이 시스템마다 달리 정의될 수 있으므로 본 실험에서는 사람이 판별한 유사도 계산값을 기준으로 하여 시스템의 결과값이 얼마나 일치하는지의 상관계수를 구하여 검사하였다. 사람이 판별한 유사도 계산값은 3명의 유사도 계산값의 평균을 취하였다. 상관계수 식은 다음과 같다.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad (5)$$

이 식의 $\bar{x}$는 사람이 판별한 유사도 계산값의 평균을 나타내고 $\bar{y}$는 시스템의 유사도 계산값의 평균을 나타낸다. σ_x 와 σ_y는 표준편차를 나타낸다.

시스템 검사 결과 시스템 D가 가장 좋은 결과를 보였고 시스템 B와 C가 다음으로 나은 결과를 보였다. 본 연구에서 제안한 것과 같이 형태소와 구문/의미분석의 결과물에 대한 가중치를 부여한 것이 효과가 있음을 보여주고 있다. 두 시스템을 비교해보면 시스템 B가 시스템 C보다 나음을 보여준다. 시스템 B는 용어와 형태소에 가중치를 부여한 것이고 시스템 C는 용어와 구문트리에 가중치를 부여한 것이다. 이것은 구문의미분석의 효과가 형태소분석의 효과보다 좋지 않음을 보이고 있다. 이 문제는 시스템에 적용한 구문의미분석기의 성능의 문제가 될 수도 있으나 실험 결과를 분석해 본 결과 다음과 같은 원인이 중요한 영향을 미쳤음을 알 수 있었다.

유사도 비교를 위해 구문의미분석의 결과인 트리를 비교하기 위하여 본 연구에서는 단순한 일치기법을 사용하였다. 즉, 정확한 매치가 일어나는 경우만 같은 것으로 보고 조금이라도 다른 것은 다르게 해석하였다. 예를 들면 '표현해야 한다'와 '표현해야 된다'의 경우 의미적으로 미묘한 차이가 있지만 구문적으로 큰 의미 차이가 없다. 그렇지만 구문해석의 결과가 전자의 경우 (HNXS 표현 AUCO) (HVERB 하 FINEN)로 트리가 구성되고 후자는 (HNXS 표현 AUCO) (HVERB 되 FINEN)로 구성되어 이를 다르게 보았다. 이를 반영하려면 구문 트리의 심층적 비교가 필요하다. 다시 말하면 추상적인 트리로 변환한 후 비교하든지 아니면 트리에 대해 트리의 구조적 일치성, 노드 내용 또는 의미의 일치성, 각 노드의 구문관계의 일치성 등을 종합하여 트리의 유사도를 계산하는 방법을 설계, 적용하여 시스템을 개선할 필요가 있다. 또한 구문트리의 유형을 정의하고 각 유형에 따라 알맞은 가중치 계산 방법을 정의하여 시스템에 반영한다면 더 좋은 결과를 가져올 것으로 예상된다.

또 다른 문제는 복문의 처리이다. 예를 들면 '이해하기가 어렵다'와 '이해하기 어렵게 한다'의 경우 전자는 HVERB 어렵 FINEN (HNXS 이해 SUBJ))로 트리가 구성되었고 하나의 단문으로 처리되는 반면 후자는 (HNXS 이해 UNDEF) (HVERB 어렵 AUCO) (HVERB 하 FINEN)로 세 개의 트리로 결과가 나와 세 개의 단문으로 처리된다. 본 연구의 시스템은 단문의 처리에 범위를 정하였으므로 복문은 앞으로의 연구 과제이다.

다음은 저빈도 및 관용표현의 가중치 적용의 실험 이다. 이 실험은 형태소, 구문의미분석 실험에 대해 저빈도 및 관용구 가중치를 부여하여 결과를 얻었다.

표 4. 저빈도 및 관용구 가중치 이용한 실험 설명

	시스템 A	시스템 E	시스템 F	시스템 G
가중치 fk (기본, 저빈도, 관용구)	(4 0 0)	(4 4 0)	(4 0 4)	(4 4 2)
설명	저빈도 가중치와 관용구 가중치를 이용하지 않은 유사문서 식별	저빈도 가중치를 이용한 유사문서 식별	관용구 가중치를 이용한 유사문서 식별	저빈도 가중치와 관용구 가중치를 이용한 유사문서 식별

표 5. 시스템 실험결과

실험 1 \ 실험 2	A	B	C	D
A	0.930773	0.953939	0.939971	0.960382
E	0.975748	0.982291	0.938757	0.965258
F	0.884728	0.878517	0.910453	0.943550
G	0.954294	0.965258	0.921634	0.962702

이 실험 결과에서 우리는 저빈도에 대해 가중치를 부여한 시스템 E가 전반적으로 좋은 결과를 얻었음을 알 수 있다. 즉 저빈도의 가중치 부여가 효과 있음을 보여준다. 반면에 시스템 F는 관용구에 대해 가중치를 부여한 것으로 시스템 A 보다 오히려 좋지 않은 결과를 가져왔다. 이 문제를 분석한 결과 가중치를 부여하는 관용구의 표현에 문제가 있음을 알았다. 본 연구에서는 관용구에 사용되는 용어나 형태소에 대해 가중치를 부여하여 실험하였다. 관용구는 용어나 형태소 들이 여러 개 연관되어 하나의 관용표현으로 사용된 것이다. 따라서 관용구의 의미를 제대로 살리려면 관용 표현에 사용된 용어나 형태소에 가중치를 부여하는 것이 아니라 복합된 관용표현에 가중치를 부여하여야 한다. 차후 가중치를 반영하는 관용구의 복합표현을 적용하여 시스템을 개선할 필요가 있다.

그리고 유사문서 식별 시스템의 실험 분석과정에서 우리는 다음과 같은 추가 개선 사항이 있음을 깨닫게 되었다. 첫째로 문서의 표현에서 순서적인 정보가 누락됨에 따라 유사문서로 판별되는 사례가 있었다. 즉, 두 문서가 내용의 순서가 다르게 나타나도 내용이 같으면 높은 유사성의 결과가 도출되었다. 차후 문서 정보 표현에 순서적인 정보를 반영할 필요가 있다.

둘째로 문서의 유형이 다르면 같은 유사도 계산 방법을 적용하는 것이 아니라 유형에 맞는 유사도 계산 방법을 적용하면 더 좋은 결과를 얻을 수 있음을 알게 되었다. 차후 문서 유형을 분류, 정의하고 각 문서 유형에 따른 유사도 계산 방법을 설계하여 시스템을 개선할 필요가 있다.

V. 결론 및 논의

최근 문서 복제나 표절 등을 검사하기 위해 문서표절 검사 시스템이 개발되어졌으나 자연어 처리의 문제로 질높은 문서표절검사가 어려운 실정이다. 본 연구는 질높은 문서표절 검색을 위해 자연어처리 기법인 구문의미분석기법을 이용한 유사문서 식별 시스템을 개발하였다. 본 시스템을 실험한 결과 다음과 같은 결론을 얻었다.

첫째로 유사문서 식별을 위한 구문의미 분석의 문제점인 구문트리 비교법 개발, 복문 해결 등 문제점을 파악할 수 있었고 앞으로의 연구방향을 얻을 수 있었다.

둘째로 상용화 시스템을 위해서 대용량의 자료를 검사하기에 적합한 기법을 적용할 필요가 있다. 본 연구의 시스템은 학생들의 보고서간의 복제나 표절 검사를 위한 시스템으로 시작하였다. 차후 대량의 문서에 대한 표절여부를 검색하기 위한 시스템적인 효율성 문제 등을 고려할 필요가 있다.

셋째로 시스템을 개선하기 위해 다양한 종류와 대량의 검사문서를 검사할 필요성이 있다. 구문의미분석의 효과가 문서의 길이, 문서의 영역, 문서의 종류 등에 따른 효과가 어떠한지 분석할 필요가 있다.

본 연구의 목적대로 구문분석기를 활용한 유사문서 식별기의 효과가 나오지 못하였다. 그렇지만 질높은 유사문서 식별을 하기 위해서는 문서의 표층적인 비교가 아닌 심층적인 비교가 이루어져야 한다. 이를 위해서는 구문의미분석이 필수적이다. 본 연구는 이와 같은 취지로 시작되었고 실험 결과 파악된 구문의미분석 기술의 문제점, 즉 구문의미분석의 결과물인 트리의 유사도 계산기법, 복문 처리 기법, 관용표현에 대한 가중치 부여 기법 등을 개선점을 발견할 수 있었다. 본 연구에서 사용한 실험은 학생들의 보고서 표절에 국한되어 검사하였다. 제안된 시스템이 상용시스템으로의 활용을 위해서는 효율성 문제와 견고성(robustness) 문제를 해결해야 한다. 이를 위해 다양한 검사문서에 대한 실험과 효율성을 반영해야 함을 알 수 있었다. 본 시스템에 사용한 기술은 유사문서 검색 뿐 아니라 문서 분류, 문서 clustering 등의 자연어 처리 분야에도 활용할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 조정현, 김유섭, “웹 검색을 활용한 기사 표절 탐지 시스템”, 한국컴퓨터종합학술대회 발표논문집, 제35권, 제1호(C), pp.420-424, 2008.

[2] 손기락, 문승미, “계층적 군집화기법을 이용한 소스코드 표절검사”, 정보교육학회논문지, 제11권, 제1호, pp.91-98, 2007.

[3] 지정훈, 우균, 조환규, “바이트코드 분석을 이용한 자바프로그램 표절검사기법”, 정보과학회 논문지 : 소프트웨어및응용, 제35권, 제7호, pp.442-451, 2008.

[4] 김연어, 이윤정, 우균, “클래스 구조 그래프 비교를 통한 프로그램 표절 검사 방법”, 한국콘텐츠학회논문지, 제13권, 제11호, pp.37-47, 2013.

[5] S. Brin, J. Davis, and H. Garcia-Molina, "Copy Detection Mechanisms for Digital Documents," Proc. of the ACM SIGMOD international conference on management of Data, pp.398-409, 1995.

[6] A. Si, H. V. Leong, and R. W. H. Lau, "CHECK: A Document Plagiarism Detection System," Proc. of the 1997 ACM symposium on Applied Computing, pp.70-77, 1997.

[7] S. M. Eissen and B. Stein, “Intrinsic Plagiarism Detection," Proceedings of the 28th European Conference on Advanced Information Retrieval(ECIR'06), pp.565-569, 2006.

[8] 허원지, 정용규, "문서간 유사도 측정방법의 개선에 관한 연구", 한국정보과학회 2011년 가을 학술발표논문집, 제38권, 제2호(C), pp.122-124, 2011.

[9] http://www.turnitin.com

[10] 박우창, 서여진, “구조와 내용유사도에 기반한 XML 웹문서 검색시스템 구축”, 한국인터넷정보학회, 제6권, 제2호, pp.99-115, 2005.

[11] 신미애, 고방원, 김영철, 정진영, “문서구조정보 기반의 유사도 측정”, 2010년 한국컴퓨터정보학회 하계학술대회논문집, 제18권, 제2호, pp.499-502, 2010.

[12] 전명재, 박상돈, 박웅, 허진영, 조환규, “한글 구조특성과 지역정렬 알고리즘을 사용한 표절 판정 시스템의 개발”, 2004년 정보과학회 가을학술발표논문집, 제31권, 제2호, pp.727-729, 2004.

[13] 조동욱, 홍윤선, 조선옥, “효과적인 e-런닝 시스템 구축을 위한 과제물 표절 검사”, 한국콘텐츠학회 종합학술대회 논문집, 제1권, 제2호, pp.53-59, 2003.

[14] 임해창, 최성원, 우연문, *문서의 표절 검사 방법*, 특허출원, 2006.

[15] 류창건, 김형준, 조환규, “한글 말뭉치를 이용한 한글 표절 탐색 모델 개발”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제14권, 제2호, pp.231-235, 2008.

[16] 황인수, “인터넷 검색과 형태소분석을 이용한 표절검사시스템의 개발에 관한 연구”, J. of Information Technology Applications and Management, 제16권, 제1호, pp.21-36, 2009.

[17] 천승환, 김미영, 이귀상, “유사 어절트리와 비색인어 기반의 문서표절 유사도 분류 방법”, 한국컴퓨터산업교육학회 논문지, 제3권, 제8호, pp.1039-1048, 2002.

[18] 김혜숙, 박상철, 김수형, “단어가중치기반 문서간 유사도 측정에 관한 연구”, 2003년 한국멀티미디어학회 춘계학술발표논문집, pp.198-201, 2003.

[19] 장성호, 강승식, “용어 선별기법에 의한 유사문서 판별시스템”, 2003년도 정보과학회 봄학술발표논문집, 제30권, 제1호, pp.534-536, 2003.

[20] 지혜성, 조준희, 임희석, “한국어 문장 표절 유형을 고려한 유사 문장 판별”, 한국컴퓨터교육학회 논문지, 제13권, 제6호, pp.79-89, 2010.

[21] 김명철, 김덕봉, 이하규, 김유성, 김재훈, 박혁로 역, *최신정보검색론*, 홍릉과학출판사, 2001.

[22] 김재훈, 선충녕, 홍상욱, 이성욱, 서정연, 조정미, “KTAG99: 새로운 환경에 쉽게 적응하는 한국어 품사 태깅 시스템,", 제11회 한글 및 한국어정보처리 학술대회논문집, pp.99-105, 1999.

[23] 강원석, 노주환, 제환주, 조대흠, 황세연, 정부천, “검색엔진을 위한 키워드 관련어 추출기의 설계 및 구현”, 한국컴퓨터교육학회 2007년도 동계 학술대회 논문집, pp.241-246, 2007.

[24] 국립국어연구원, *21세기 세종계획 성과물*, 2008.

저 자 소 개

강 원 석(Won-Seog Kang) 정회원

- 1985년 2월 : 경북대학교 전자공학과(공학사)
- 1988년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 3월 ~ 현재 : 안동대학교 정보과학교육과 교수

<관심분야> : 자연어처리, 정보검색

황 도 삼(Do-Sam Hwang) 정회원

- 1980년 2월 : 홍익대학교 전자계산학과(이학사)
- 1983년 2월 : 연세대학교 전자계산학과(이학석사)
- 1995년 2월 : 교토대학교 전자정보통신과(공학박사)
- 1996년 3월 ~ 현재 : 영남대학교 컴퓨터공학과 교수

<관심분야> : 자연어처리, 온톨로지

Jung H Kim 정회원

- 1979년 2월 : 서울대학교 전기전자공학과(공학사)
- 1982년 : U. of IOWA, Electrical and Computer Eng.(공학석사)
- 1987년 : U. of IOWA, Electrical and Computer Eng.(공학박사)
- 2000년 7월 ~ 현재 : UALR, System Engineering Tenured Professor

<관심분야> : 시스템공학, 컴퓨터공학