http://dx.doi.org/10.5392/JKCA.2015.15.10.636

구문의미트리 비교기를 이용한 유사문서 판별기

Discriminator of Similar Documents Using the Syntactic-Semantic Tree Comparator

강원석
안동대학교 정보과학교육과

Won-Seog Kang(wskang@andong.ac.kr)

요약

정보사회에 문서 복제나 표절의 검출에 대한 필요성이 증대되고 있다. 그 필요성에 따라 많은 연구가 이루어지고 있으나 자연어 처리의 문제가 유사 문서 판별의 질 향상에 제약이 되었다. 최근 구문의미분석의 기술을 접목하여 유사문서 판별의 성능을 향상을 시도하였으나 구문의미분석의 결과인 구문의미트리를 비교하는 어려움이 있었다. 본 논문은 구문의미트리의 유사도를 계산하는 구문의미트리 비교기를 개발하고 이를 이용하여 유사문서를 판별하는 시스템을 설계, 구현한다. 본 시스템의 성능을 실험하기 위하여 휴먼 판별과 제안한 시스템의 판별과의 상관계수를 분석하였다. 실험결과, 구문의미트리 비교기를 이용한 유사문서 판별기의 성능을 검증할 수 있었다. 앞으로 문서 유형을 정의하고 각 유형에 맞는 판별 기법을 개발할 필요가 있다.

■ 중심어 : | 유사문서 판별 | 구문의미트리 비교기 |

Abstract

In information society, the need to detect document duplication and plagiarism is increasing. Many studies have progressed to meet such need, but there are limitations in increasing document duplication detection quality due to technological problem of natural language processing. Recently, some studies tried to increase the quality by applying syntatic-semantic analysis technique. But, the studies have the problem comparing syntactic-semantic trees. This paper develops a syntactic-semantic tree comparator, designs and implements a discriminator of similar documents using the comparator. To evaluate the system, we analyze the correlation between human discrimination and system discrimination with the comparator. This analysis shows that the proposed discrimination has good performance. We need to define the document type and improve the processing technique appropriate for each type.

■ keyword : | Similar Document Detection | Syntactic-Semantic Tree Comparator |

I. 서 론

수많은 정보들이 생성되는 정보사회에서는 창의적 아이디어를 내재한 정보는 그 가치가 매우 중요하다. 그러나 이와 같은 정보를 허가 없이 표절하거나 무단 복제하여 도용하는 사례가 늘어 사회적 문제가 되고 있다. 이러한 문제를 해결하기 위해 표절과 복제를 탐지하거나 방지하는 방법에 대한 연구가 진행되고 있다.

* 이 논문은 2012년도 안동대학교의 산학연구비에 의하여 연구되었음.
접수일자 : 2015년 05월 06일
수정일자 : 2015년 05월 27일
심사완료일 : 2015년 05월 27일
교신저자 : 강원석, e-mail : wskang@anu.ac.kr

그 연구들은 두 갈래로 나뉜다. 한 분야는 프로그램 소스 표절에 대한 연구이고 다른 분야는 일반문서 표절에 대한 연구이다. 프로그램 소스 분야는 비교적 많은 연구가 이루어졌다. 이것은 프로그램 소스가 형식언어로 기술되어 있기 때문이다. 그러나 일반문서 분야는 자연어인 한국어로 기술되어 있어 어려움이 있다. 한국어의 경우 어휘분석은 어느 정도 결과를 얻을 수 있으나 구문분석은 아직 실용할 정도는 되지 못한다.

이와 같은 형편에서 많은 연구들[1-3]이 문서가 나타내는 정보는 키워드에 담겨있다는 전제하에 키워드를 추출하는 어휘분석을 진행하고 추출된 키워드를 이용하여 문서의 유사성을 검사하는 방법을 적용하고 있다. 이 연구들은 문서가 가지고 있는 정보를 왜곡할 가능성도 포함하고 있다. 순서가 전혀 다른 어휘들이 출현 빈도가 같게 나올때 그 문서가 다르더라도 유사성은 높게 나올수 있다는 점에서 시스템의 질이 떨어질 수 있다. 이를 해결하기 위해서는 순서 정보나 구조 정보의 추출이 필요하다.

[4]의 연구는 구조정보를 얻기 위하여 구문의미 분석을 실시하고 그 결과를 이용하여 유사문서 식별에 사용한다. 그렇지만 처리의 결과로 얻어진 구문의미트리를 비교하는데 일대일 매칭의 비교만 진행하여 구문의미 분석의 효과를 제대로 거두지 못하였다. 이에 본 논문은 구문의미분석으로 얻어진 구문의미트리의 유사성을 검사하는 구문의미트리비교기를 개발하고 이를 이용하여 유사문서 판별에 적용한다. 그리고 [4]에서 진행한 저빈도와 특별한 패턴인 관용구에 대해 부여한 가중치를 적용하여 효과적인 유사문서 판별기를 구현하고자 한다.

본 연구의 2장에서는 유사문서 판별에 관련된 관련연구에 대해 서술하고 3장은 구문의미트리 비교기를 이용한 유사문서 판별 시스템을 서술한다. 4장에서는 제안한 시스템의 성능을 분석하기 위하여 사람이 판별한 휴먼 판별결과와 본 시스템의 판별 결과의 상관관계를 분석하고 검토한 후 5장에서 결론을 지었다.

II. 관련 연구

1. 프로그램 소스 표절 검사

[5]는 프로그램 소스의 토큰을 군집화하여 그 결과를 수형도로 나타내어 사람의 표절검사를 돕는 시스템을 개발하였다. [6]은 프로그램 소스 복제의 시간적 부담을 덜고자 구문트리에서 노드스트링을 비교하지 않고 키워드를 추출하여 이를 비교하는 방법을 적용하였다. [7]은 보안의 문제로 바이트코드의 유사성 검사를 시도하였다. [8]은 클래스의 멤버변수와 메소드간의 참조관계를 그래프로 나타내고 그래프 동형 검사를 실시하여 프로그램 소스의 복제를 찾아낸다. 위 연구들은 형식언어로 표기된 프로그램 소스 표절 검사에 초점을 두었다. 따라서 이 방법들은 형식언어가 아닌 자연어로 표기된 일반문서의 표절 검사에 적용할 수가 없다.

2. 영어 문서 표절 검사

[9]는 영어 문장을 중심으로 유사도를 찾아내고 이를 이용하여 문서의 표절을 찾아내는 방법을 적용하였다. [10]은 영어에서 관사와 같은 불용어의 중요성을 강조하여 가중치를 부가한 방법을 써서 유사문서 판별을 하였다. [11]은 워드넷을 이용하여 워드넷 추상화 수준에 따른 어휘의미 유사도를 계산하고 구문트리 분리정도와 트리깊이에 따른 구문적 유사도를 계산하여 유사문서 판별을 시도하였다. 위 연구들의 기법들은 영어 문서의 표절 검사에 대해 다루었기 때문에 언어적 특성이 다른 한국어 문서에는 적용할 수가 없다.

3. 한국어 문서 표절 검사

[12]는 한국어 문서 표절 검사를 위해 형태소 해석을 거치지 않고 어절을 추출하여 어절 일치도를 계산하여 표절검사를 시도하였다. [13]도 어절을 추출한 후 어절을 바이트별로 분리하여 스트링매칭 방식을 적용하였다. 위 두 방법은 형태소 해석을 시도하지 않아 어절의 변형이나 유사어절 등의 심층적 유사성은 판별할 수 없는 문제를 안고 있다.

[1]은 형태소 해석기법을 적용하여 용어를 찾아내고 찾아낸 용어에 대해 선택적으로 가중치를 부여하는 방

법을 적용하여 유사문서를 찾고자 하였다. [2]는 형태소 해석기법을 적용하여 추출한 용어에 대해 4가지 종류의 가중치를 정의하고 이를 신경망 시스템에 적용하여 유사문서를 판별하였다. [3]은 형태소 해석으로 용어를 추출하고 표절 유형에 따라 적합한 시스템을 찾기 위해 유사문서판별 시스템의 결과를 비교하는 연구를 하였다. 위 [1-3]의 연구들은 연구의 방향은 다르나 기본적으로 형태소 해석의 방법을 적용하고 있다. 형태소 해석의 결과는 구조적인 정보나 순서적인 정보를 포함하고 있지 않다. 따라서 이 방법은 문서나 문장에 들어있는 구조적 정보를 포착하지 못하는 문제를 안고 있다.

4. 구조적 관계를 이용한 표절검사

[6]은 구조적 정보를 나타내는 구문트리를 비교하여 다이스 계수와 같은 유사도 계산식으로 유사문서를 판별하였다. [8]도 클래스의 멤버 변수와 메소드간의 참조관계를 나타낸 그래프 구조의 동형검사를 하여 표절검사를 실시하였다. [6][8]의 연구는 용어 중심의 방법에서 탈피하여 구조적 정보를 이용하고자 프로그램 소스의 구문관계를 분석하고 그 관계를 나타내는 구문트리를 이용하여 유사성을 계산한다. 그렇지만 이 방법은 형식언어인 프로그램 소스에 대한 것이어서 형식언어가 아닌 일반문서에는 적용할 수가 없다.

[11]은 유사성을 구하기 위해 워드넷의 추상화 수준을 이용한 어휘구문유사도와 구문트리 분리와 트리깊이를 반영한 구문적 유사도를 복합한 식으로 유사문서 판별을 하였다. 이 연구의 구문트리는 자연어의 구문트리이지만 대상 언어가 한국어가 아닌 영어에 대한 것이다.

[14]는 온톨로지 트리를 대상으로 트리를 정렬하기 위해 트리의 요소를 비교하는 근사스트링매칭방식을 이용하였다. 스트링 매칭방식은 각 요소가 동등한 의미요소로 인식하여 진행하는 방식이다. 따라서 각기 다른 구문관계를 나타내는 구문요소의 구문트리 비교에는 적합하지 않다.

[15]는 트리를 비교하는 방식의 연구이다. 그렇지만 비교방식의 대상 트리가 XML 구문트리이다. 이 방식을 XML과 거리가 먼 한국어 문서에 적용할 수가 없다.

[4]는 한국어 문장의 구문관계를 찾는 구문의미해석을 적용하여 구문의미트리를 추출하고 이를 유사문서 판별에 사용하였다. 그렇지만 유사문서 판별에 추출한 구문의미트리를 비교하는데 정확히 일치하지는 않지만 구문의미적으로 유사한 패턴을 찾지 못하여 구문의미분석의 효과를 제대로 거두지 못하였다. 이에 본 논문은 구문의미트리의 유사도를 비교하는 비교기를 개발하고 이를 유사문서 판별에 활용하여 정확히 일치하지는 않지만 구문의미적으로 유사한 패턴까지 찾아낼 수 있도록 한다.

III. 구문의미트리 비교기를 이용한 유사문서 판별 시스템

본 연구의 유사문서 판별 시스템의 구조는 [그림 1]과 같다.

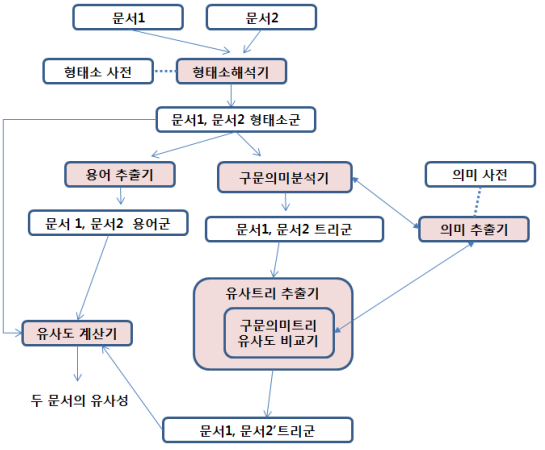

그림 1. 유사문서 판별 시스템 구조

유사문서 판별 시스템은 크게 형태소해석기, 용어추출기, 구문의미분석기, 의미추출기, 유사트리 추출기, 유사도 계산기로 구성된다. 형태소 해석기는 형태소 사전을 활용하여 형태소를 추출한다. 이 시스템의 형태소 해석기는 [16]을 이용하였다.

용어추출기는 형태소 해석의 결과에서 용어만 추출한다. 구문의미 분석기는 어절유형분석, 체언유형처리, 복합명사처리, 관형처리, 용언유형처리, 트리구조화의

단계로 구성된 것으로 [4]의 연구를 기초하여 구성되었다. 구문의미 분석기는 형태소 해석의 결과를 받아 구문의미트리를 생성한다.

유사트리 추출기는 생성된 구문의미트리를 비교하여 유사한 트리를 추출하는 것으로 구문의미트리 유사도를 계산하는 비교기를 이용한다. 구문의미트리 유사도 비교기는 대상이 되는 두 트리의 구성요소에 대해 어휘의미 유사도 비교와 구조적 유사도 비교를 종합하여 유사성을 계산한다. 의미사전과 의미 추출기는 구문의미분석기와 구문의미트리 유사도 비교기에서 필요한 조건 검사와 의미 추출에 활용한다. 유사도 계산기는 두 문서의 분석 결과물인 형태소군, 용어군, 그리고 유사트리 추출기를 통해 얻은 트리군을 입력으로 두 문서의 유사성을 판별한다. 각 단계의 분석결과물인 용어군, 형태소군, 트리군의 예는 다음 표와 같다.

표 1. 문서의 분석결과물의 예

구분	내용
문서	열거형을 정의하는 설비는 쉬운 방법이다 ...
용어군	열거 형 정의 설비 방법 ...
형태소군	열거/ncn+형/ncn+jco 정의/ncn+xsm+etm 설비/ncn+jxt 쉽/paa+etm 방법/ncn+jp+ef+sf ...
트리군	(ADNMEN 정의 (OBJ 열거형)) (ADNMEN 쉽 (TOP 설비)) (FINEN 방법) ...

유사트리 추출기를 통해 얻은 유사트리군과 유사도 비교의 과정에 대해 다음 절에 자세히 설명한다.

1. 유사트리 추출기

[4]의 연구에서는 형태소 해석과 구문의미해석의 결과로 나온 용어나 형태소, 구문의미트리에 대한 빈도수를 파악하여 그 값을 가중치로 정의하였다. 그리고 코사인 계수를 이용하여 두 문서의 유사도를 비교하였다. 이때 용어나 형태소 하나를 벡터의 한 축으로 삼은 것 같이 구문의미트리도 하나의 트리를 벡터의 한 축으로 삼아 유사도를 비교한다. 문제는 생성된 구문의미트리가 내용적으로 구조적으로 유사할지라도 구문의미트리의 한 구성요소가 생략되거나 격관계는 같지만 순서가 다른 트리는 다른 구문의미트리로 인식하여 벡터의 다른 축으로 전사된다는 점이다. 이로 인해 유사한 구문의미트리임을 파악하지 못하고 다른 구문의미트리로 인식하여 유사문서 판별에 구문의미분석의 효과를 제대로 거두지 못하였다.

본 연구는 [4]의 문제를 해결하기 위해 구문의미트리 비교기를 개발하고 이를 활용하여 유사한 구문의미트리를 찾아내어 유사문서 판별에 이용하였다. 유사트리 추출기는 구문의미트리 비교기를 활용하여 유사한 트리를 찾아내는 제어도구이다. 이 도구의 동작에 대해 단계적으로 설명한다.

유사트리 추출기의 입력은 비교대상이 되는 두 문서의 구문의미해석 결과인 두 개의 구문의미트리군이다. 이를 A와 B로 정의한다. A는 m 개의 구문의미트리들로 구성되고 각 구문의미트리는 AT1, AT2 등으로 표현된다. B는 비교대상 문서의 구문의미트리군을 나타낸다.

A = [AT1, AT2, AT3, ... ATm]
B = [BT1, BT2, BT3, ..., BTn]
B′ = [B′T1, B′T2, B′T3, ... , B′Tp]

유사트리 추출기는 B의 구문의미트리들을 기준으로 A의 모든 구문의미트리들과 비교하여 B′을 생성한다. B의 구문의미트리들은 A의 구문의미트리와 유사한 것이 있는 것과 없는 것으로 구분할 수 있다. 유사한 것이 있는 구문의미트리들은 유사성을 표현하기 위해 유사한 A의 구문의미트리를 복사하고 유사한 것이 없는 구문의미트리들은 그대로 복사하여 B′을 생성한다. 유사성의 여부는 구문의미트리 유사도비교기 CT(ATi, BTj)의 값이 한계값을 넘어서는지의 여부에 따라 결정된다. 예를 들어 설명한다.

A = [(COCO 읽 (OBJ 프로그램)), (FINEN 쉽 (SUBJ 이해))]
B = [(COCO 읽 (TOP 이것) (OBJ 프로그램)), (UNDEF 이해), (FINEN 하)]
B‘ = [(COCO 읽 (OBJ 프로그램)), (UNDEF 이해), (FINEN 하)]

A는 두 개의 구문의미트리로 구성되고 B는 3개의 구문의미트리로 구성된다. 유사트리추출기는 A와 B의 구

문의미트리들을 비교하여 유사한 구문의미트리들을 찾는다. A의 첫 번째 구문의미트리 '(COCO 읽 (OBJ 프로그램))'와 B의 첫 번째 구문의미트리 '(COCO 읽 (TOP 이것) (OBJ 프로그램))'는 유사성을 가졌다. 유사성을 계산하는 CT 계산의 예는 [표 2]와 [표 3]에 서술하였다.

유사트리 추출기는 B의 첫 번째 구문의미트리와 유사한 트리를 표현하기 위해 A의 첫 번째 구문의미트리를 복사하고 B의 나머지 유사하지 않은 트리들을 복사하여 B′을 생성한다. 이 결과를 이용하여 유사도 계산기는 A와 B′에 대한 유사도를 계산한다. 이것은 기존 유사도 계산기를 이용하여 유사문서를 판별할 수 있도록 하여 시스템의 단순성과 확장성을 제공하게 된다.

유사트리 추출기는 유사성을 검사하기 위해 구문의미트리 비교기를 활용하였다. 구문의미트리 비교기는 다음과 같은 원칙을 반영하여 정의하였다.

원칙 1. 트리의 루트 노드는 두 트리의 유사성에 아주 중요한 영향을 준다.

원칙 2. 트리의 자식 서브트리는 두 트리의 유사성에 중요한 영향을 준다.

원칙 2-1. 두 트리의 공통으로 나타나는 같은 격의 자식트리는 두 트리의 유사성에 중요하다.

원칙 2-2. 두 트리에서 한쪽에만 나타나는 격의 자식트리는 두 트리의 유사성에 역의 작용을 한다.

이 원칙에 따라 구문의미트리의 유사도 비교기 CT(T_1, T_2)는 다음과 같이 정의한다.

$$CT(T_1, T_2) = \frac{CL(ro(T_1), ro(T_2)) \times \alpha + \sum_{ca(T_{1i}) = ca(T_{2j})} CT(T_{1i}, T_{2j})) \times \beta}{\alpha + |T_1 \cap T_2| \times \beta + |C| + |D|} \quad (식1)$$

위 식의 T_1과 T_2는 대상이 되는 구문의미트리이고 $ro(T_1)$는 구문의미트리 T_1의 루트노드를 의미한다. $CL(ro(T_1), ro(T_2))$은 T_1과 T_2의 루트노드에 해당하는 어휘의 의미 유사성을 계산한다. T_{1i}와 T_{2j}는 구문의미트리 T_1와 T_2의 한 자식트리를 나타낸다. $ca(T_{1i})$, $ca(T_{2j})$는 구문의미트리 T_{1i}와 T_{2j}의 격관계를 나타내며, $ca(T_{1i}) = ca(T_{2j})$는 그 격관계가 같은 것을 뜻한다. $|T_1 \cap T_2|$는 격관계가 같은 자식트리의 수를 나타내며 |C|는 T_2의 자식트리에는 없는 T_1의 자식트리의 수를 나타내고 |D|는 반대로 T_1의 자식트리에 없는 T_2의 자식트리의 수를 말한다.

α와 β는 가중치를 나타낸다. 정의된 유사도 비교기 CT(T1, T2)의 가중치 α는 원칙 1을 반영하여 2로 정의하였고 β는 원칙 2-1을 반영하여 2로 정의하였다. 그리고 원칙 2-2를 반영하기 위해 |C|와 |D|의 값을 분모에만 더하였다. 예를 들어 설명하면 [표 2]와 같다.

트리 '(SUCO 대응 (OBJ 값))'과 트리 '(SUCO 대응 (OBJ 기법))'의 유사성을 보면 구조와 루트노드는 같으나 격관계 OBJ의 자식트리가 다르다. 계산식에 의하면 루트노드가 같아서 가중치 2를 곱하였고 자식노드도 격이 같아 가중치 2를 곱하게 되는데 '값'과 '기법'의 어휘의미유사도가 0이므로 그 결과는 결국 0.5의 계산값이 나오게 된다. T1와 T3를 비교해본다면 루트노드가 같아 가중치 2를 곱하고 같은 격노드의 자식노드가 있어 가중치 2를 곱하나 T3에는 T1에 없는 격의 자식노드가 하나 더 있으므로 이를 적용하기 위해 분모에 1을 더하는 효과가 오므로 0.4로 계산된다. T2와 T3는 '기법'과 '방법'의 어휘의미유사도가 0.8이므로 이를 반영하여 0.72로 계산된다.

표 2. 구문의미트리 유사도 계산 예

항목	값
T1	(SUCO 대응 (OBJ 값))
T2	(SUCO 대응 (OBJ 기법))
T3	(SUCO 대응 (SUBJ 조직) (OBJ 방법))
CT(T1,T2)	(CL(대응,대응) × 2 + CT(T11,T21) × 2) / (2 + 1 × 2 + 0 + 0) = (1 × 2 + 0 × 2) / 4 = 0.5
CT(T1,T3)	(CL(대응,대응) × 2 + CT(T11,T31) × 2) / (2 + 1 × 2 + 0 + 1) = (1 × 2 + 0 × 2) / 5 = 0.4
CT(T2,T3)	(CL(대응,대응) × 2 + CT(T21,T32) × 2) / (2 + 1 × 2 + 0 + 1) = (1 × 2 + 0.8 × 2) / 5 = 0.72

어휘의 의미유사성 비교기 CL(a, b)는 다이스계수를 인용하여 다음과 같이 정의하였다.

$$CL(a, b) = \frac{2 * |sem(a) \cap sem(b)|}{|sem(a)| + |sem(b)|} \quad (식2)$$

위 식의 a와 b는 단어를 나타내고 sem(a)는 단어 a의 의미속성집합을 나타낸다. |sem(a)|는 단어 a의 의미속성집합의 원소수를 뜻한다. 단어의 의미속성집합은 본 시스템의 의미추출기를 이용하여 얻는다. 의미 추출기는 상하위 의미 관계를 나타내는 시소러스 사전[17]을 참조한다. [17]의 시소러스는 크게 thing, event, feature로 구분된 트리로 구성된다. 그 일부는 [그림 2]와 같다.

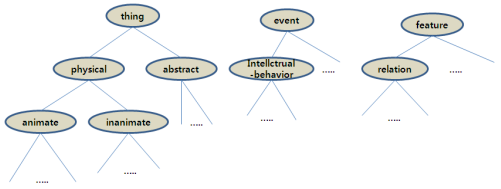

그림 2. 상하위 의미 체계의 일부

의미 추출기와 어휘의 의미유사성 비교의 예는 [표 3]과 같다.

표 3. 어휘의미유사도 계산 예

항목	값
sem(값)	{feature, measurement, 치수}
sem(기법)	{abstract-thing, intellectual-thing, means, thing, 기법}
sem(방법)	{abstract-thing, intellectual-thing, means, thing, 방법}
CL(값, 기법)	$\frac{2*0}{3+5}=0$
CL(기법, 방법)	$\frac{2\times4}{5+5}=0.8$

'값'의 의미속성집합과 '기법'의 의미속성집합의 교집합은 공집합이다. 따라서 '값'과 '기법'의 유사성 CL(값, 기법)은 0이 된다. CL(기법, 방법)은 의미속성의 교집합의 원소가 4개이므로 0.8이 된다.

2. 유사도 계산기

유사도 계산기는 비교대상의 두 문서에 대해 표현된 사상을 비교하여 유사도를 계산한다. 본 논문에서는 형태소 해석, 구문의미해석을 거쳐 나온 형태소군, 구문의미트리군과 용어추출기를 통해 나온 용어군을 기본으로 하고 유사트리 추출기를 통해 획득한 유사구문의미트리군을 벡터값으로 표현하고 유사도를 계산하였다.

추가로 분석을 통해 추출한 용어, 형태소, 구문의미트리에 대해 저빈도 여부와 관용표현 여부에 따른 가중치를 부가하여 유사도를 계산하였다. 이 과정에서 저빈도 여부와 관용표현 여부는 저빈도 DB와 관용표현 DB를 사용하였다. 이 DB들은 [18]의 말뭉치와 관용사전을 자료로 본 연구에서 구축된 용어추출기, 형태소해석기, 구문의미분석기를 이용하여 추출한 후 구축되었다.

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

$$
\begin{aligned}
&D_i=(a_{i1},a_{i2},a_{i3},\ldots\ldots,a_{in}),\\
&i\in\{\text{용어, 형태소, 트리}\}\\
&a_{ij}=\frac{freq_{ij}}{\max_{1\le k\le n}(freq_{ik})}\times idf_{ij}\ ,\quad (\text{식3})\\
&n:\ unique\text{한 용어의 수}\\
&idf_{ij}=\log(\frac{N}{n_{ij}})+1,\\
&N:\ \text{문서의 총수}\\
&n_{ij}:\ a_{ij}\text{가 출현한 문서수}
\end{aligned}
$$

D_i는 용어군이나 형태소군, 트리군 중 하나를 표현한다. (본 논문의 유사트리는 트리와 같은 형태이므로 트리군과 같이 표현된다.) [4]와 달리 본 논문에서는 용어들의 빈도가중치와 역문헌빈도수를 이용한 값으로 가중치를 표현하였다. max는 가장 빈도수가 높은 용어의 빈도수를 표현한다. idf는 역문헌빈도수를 뜻한다. 본 논문에서는 1을 기준으로 용어가 나타난 문서수가 적으면 그 용어의 가중치를 더 부가하는 방식으로 정의하였다.

분석 결과물 트리와 유사트리는 구조가 같으나 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.

$$sim(D_{1i}, D_{2i}) = \frac{D_{1i} \bullet D_{2i}}{|D_{1i}| \times |D_{2i}|}$$
$$= \frac{\sum_{j=1}^{n} a_{1ij} \times a_{2ij}}{\sqrt{\sum_{j=1}^{n} a_{1ij}^2} \times \sqrt{\sum_{j=1}^{n} a_{2ij}^2}}, \text{ (식4)}$$
$$i \in \{\text{용어, 형태소, 트리}\}$$

$$tsim(D_1, D_2) = \sum_{j \in \{\text{용어, 형태소, 트리}\}} c_j * sim(D_{1j}, D_{2j}), \text{ (식5)}$$
$$c_j = weight\ constant\ of\ each\ category(\text{용어, 형태소, 트리})$$

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 (식4) 와 (식5)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 (식4)와 (식5)을 적용한다. 그리고 (식6)과 같이 정의된 최종 유사도 비교식을 통해 두 문서의 유사성을 결과로 낸다.

$$fsim(D_1, D_2) = \sum_{k \in \{\text{기본, 저빈도, 관용}\}} f_k * tsim(D_{k1}, D_{k2})$$
$$= \sum_{k \in \{\text{기본, 저빈도, 관용}\}} f_k * \sum_{j \in \{\text{용어, 형태소, 트리}\}} c_j * sim(D_{k1j}, D_{k2j}), \text{ (식6)}$$
$$f_j = weight\ constant\ of\ each\ category(\text{기본, 저빈도, 관용})$$

두 문서가 유사한 지의 판단은 유사성 계산 값이 일정한 한계값을 초과하면 유사문서로 판단된다. 한계값은 사용자에 따라 달리 정의할 수 있다. 그렇지만 본 논문에서는 구문의미트리를 비교하는 비교기를 이용한 유사문서 식별기가 얼마나 효과가 있는지를 검사하기 위해 한계값으로 검사하지 않고 인간이 결정한 유사도의 값과 얼마나 일치하는지를 검사하였다.

IV. 실험 및 분석

본 논문에서는 구문의미트리 비교기를 이용하여 유사문서를 판별하는 시스템의 효과를 검사하였다. 검사를 위해 컴퓨터 분야의 전공과목 레포트에서 64쌍의 검사 문서를 발췌하였다. 검사 문서는 평균 40단어, 300바이트 정도의 문서로 이루어졌다. 검사 문서쌍의 유사도 값은 0.8이상 문서 33%, 0.4이상 문서 37% , 0.4미만 문서가 30%로 구성되었다.

본 논문에서는 시스템의 유사도 계산값이 사람이 계산한 것과 얼마나 유사한가로 시스템의 성능을 검사하였다. 이를 위해 시스템 유사도와 사람이 계산한 유사도의 상관계수를 이용하였다. 사람이 판별한 유사도 계산값은 오류를 줄이기 위해 3명의 유사도 계산값의 평균을 취하였다. 상관계수는 다음과 같이 정의된다.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \quad \text{(식7)}$$

위 식의 경우 x는 휴먼계산 유사도 값의 벡터이고 y는 시스템 유사도 값의 벡터이다. n은 검사문서쌍의 수이다. $\bar{x}$는 x 벡터값들의 평균을 의미하고 σ_x는 x 벡터값의 표준편차를 의미한다. [표 4]는 상관계수 계산 예를 보여준다.

표 4. 상관계수 계산 예

구분	비교문서쌍 1 유사도	비교문서쌍 2 유사도	...	피어선상관계수
휴먼계산	.95	.30		0.951
시스템	1.0	.15		

1. 실험 1 : 기본 유형별 시스템 비교

먼저 용어, 형태소, 트리별로 각 유형의 효과를 보기 위해 복합하지 않은 시스템에 대해 검사하였다. 그 결과는 [표 5]와 같다. 추가로 기본, 저빈도, 관용표현별로도 시스템을 검사하였다.

표 5. 각 유형별 시스템 유사도 상관계수

유형별 구분	기본(f1=1, f2=0, f3=0)	저빈도(f1=0, f2=1, f3=0)	관용(f1=0, f2=0, f3=1)
용어(c1=1, c2=0, c3=0)	0.951	0.952	0.877
형태소(c1=0, c2=1, c3=0)	0.922	0.941	0.774
트리(c1=0, c2=0, c3=1)	0.864	0.856	0.86

[4]의 연구에서는 가중치를 빈도수의 합과 빈도수의 비로 가중치를 선정하였다. 본 논문에서는 이를 수정하여 빈도 가중치와 역문헌빈도 가중치를 복합한 것으로 정의하였다. 그 결과 [4]의 시스템 상관계수 0.930보다 더 좋은 0.951의 값을 얻을 수 있어 역문헌빈도수의 반영이 효과적임을 알 수 있었다.

각 유형별 시스템을 비교해보면 형태소나 트리가 용어보다 더 좋지 않은 결과를 내고 있다. 이것은 형태소의 경우 형태소 분석에 따른 태그정보까지 정확히 일치해야 되기 때문이다. 주용어가 같고 태그유형이 같더라도 태그이름이 다르면 다른 것으로 판단한다. 그러나 용어는 용어를 분리하여 독립적인 축으로 간주하여 유사도 검사를 실시하므로 비록 토씨는 다르더라도 주용어가 같다면 이와 같은 사실이 반영되어 더 좋은 결과가 나온 것으로 분석된다. 그리고 트리의 경우도 트리구조와 격 유형 등이 조금이라도 다르면 다른 것으로 인식하여 같은 정보가 들어 있어도 이를 반영하지 못하여 좋지 않은 결과를 가져왔다. 따라서 이를 개선하기 위해서는 용어와 트리를 복합한 시스템으로 구성하거나 구문의미트리 비교기를 이용하는 것이 필요하다.

다음으로 기본과 저빈도, 관용에 대한 시스템을 비교해 본다. 용어나 형태소의 경우 저빈도 용어에 가중치를 부여한 방법이 가장 좋은 결과를 가져왔고 트리의 경우 관용표현에 가중치를 부여한 방법이 가장 좋은 결과를 가져왔다. 유사문서 비교의 특징상 특정하게 사용되지 않은 용어나 형태소들이 두 문서에 같이 출현한다면 두 문서의 유사성은 높다는 원칙을 확인할 수 있었다. 트리의 경우 비교문서에 나타난 트리가 저빈도 유형일때 가중치를 부여하는 방법보다는 그대로 비교하는 방법이 더 좋은 것으로 나왔다. 이것은 트리의 특징으로 인한 것으로 분석된다. 트리는 구조와 용어 그리고 격관계 등이 복합된 것으로 그 유형은 용어와 비교하여 트리의 크기에 따라 기하급수적으로 늘어난다. 따라서 대부분의 트리들이 저빈도유형인 것으로 간주된다고 볼 수 있어 가중치를 부여하는 것은 의미가 없는 것으로 보여진다. 그리고 관용표현에 대해 가중치를 부여하는 방법이 더 나은 결과를 보여주지 못하는 것도 트리의 특징에서 기인한 것으로 보인다.

2. 실험 2 : 가중치 복합 시스템의 비교

본 논문에서는 이와 같은 각 시스템의 특징을 고려하여 복합한 시스템을 구축하여 비교하였다. 그 실험 결과는 [표 6]과 같다.

표 6. 용어 형태소 트리 복합 시스템별 유사도 상관계수

가중치 복합(c1, c2, c3) 시스템 유형	기본(f1, f2,f3) =(1,0,0)	저빈도 (0,1,0)	관용 (0,0,1)
시스템1 (1,0,0)	0.951	0.952	0.877
시스템2 (0,1,0)	0.922	0.941	0.774
시스템3 (0,0,1)	0.864	0.856	0.86
시스템4 (1/2,1/2,0)	0.942	0.955	0.899
시스템5 (2/5,2/5,1/5)	0.934	0.945	0.912
시스템6 (1/2,0,1/2)	0.924	0.923	0.933
시스템7 (1/3,1/3,1/3)	0.927	0.935	0.912

[표 6]에서 저빈도 유형의 시스템 4가 가장 좋은 결과를 가져왔다. 구문의미트리 분석의 결과를 반영하는 시스템 5, 6, 7의 경우 구문의미트리만 반영하는 시스템 3보다는 더 좋은 결과를 가져왔다. 그렇지만 시스템 1보다는 효과가 좋지 않다. 이것은 구문의미트리의 유사도 비교기의 필요성을 보여주고 있다. 구문의미트리의 효과를 보기 위해서는 정확히 매칭하는 방식의 유사도 비교가 아닌 구문의미트리의 내부 구조와 어휘 의미 등의 유사성을 복합하여 계산하는 방식이 필요하다.

다음으로 기본, 저빈도, 관용의 복합 시스템에 대한 실험을 하였다. 그 결과는 [표 7]과 같다.

표 7. 기본 저빈도 관용 복합 시스템별 유사도 상관계수

가중치 복합(f1, f2, f3) 시스템 유형	용어(c1, c2,c3) =(1,0,0)	형태소 (0,1,0)	트리 (0,0,1)
시스템1 (1,0,0)	0.951	0.922	0.864
시스템2 (0,1,0)	0.952	0.941	0.856
시스템3 (0,0,1)	0.877	0.774	0.86
시스템4 (1/2,1/2,0)	0.953	0.933	0.861
시스템5 (2/5,2/5,1/5)	0.954	0.913	0.885
시스템6 (1/2,0,1/2)	0.94	0.864	0.894
시스템7 (1/3,2/3,0)	0.953	0.936	0.86
시스템8 (1/3,1/3,1/3)	0.95	0.896	0.893

이 실험에서는 시스템 5가 가장 좋은 결과를 가져왔다. 용어의 경우 저빈도와 관용표현에 대한 가중치를

복합하였을 때 더 좋은 결과를 가져옴을 알 수 있었다. 형태소에 대한 것만 비교해 볼 때는 저빈도 가중 시스템이 가장 좋은 결과를 가져왔고 트리의 경우 기본과 관용표현을 복합한 시스템 6이 가장 좋다. 그렇지만 전체 시스템을 비교해 볼 때는 용어에 대한 시스템 5가 가장 좋은 결과를 가져왔다.

3. 실험 3 : 구문의미트리 비교기를 이용한 시스템 실험

구문의미트리 분석의 효과를 볼 수 있는 구문의미트리 비교기 이용 실험을 한다. 실험 2의 자료를 근거로 가중치 c에 대한 것은 구문의미트리가 포함된 시스템 1,3,6,7을 선택하고 f에 대한 것은 결과값이 좋은 시스템 2,5,7을 선택하였다. 그 유형은 [표 8]과 같다.

표 8. 구문의미트리 유사도 비교기를 검사할 시스템 유형

용어,형태소,트리 가중치(c1,c2,c3) 복합 유형	기본,저빈도,관용 가중치 (f1,f2,f3) 복합 유형
시스템1 (1,0,0) 시스템3 (0,0,1) 시스템6 (1/2,0,1/2) 시스템7 (1/3,1/3,1/3)	시스템2 (0,1,0) 시스템5 (2/5,2/5,1/5) 시스템7 (1/3,2/3,0)

본 실험에 비교할 시스템은 12개가 된다. 용어형태소 트리의 가중치 유형 4가지와 기본, 저빈도, 관용 가중치 유형 3가지의 곱으로 시스템이 정의된다. 각 시스템에 대해 구문의미트리 유사도 비교기의 한계값에 따른 결과는 [표 9]와 같다.

표 9. 구문의미트리 비교기를 이용한 시스템별 유사도 상관계수

한계값 / 시스템	0.3	0.4	0.5	0.6	0.7	0.8
1X2	0.952	0.952	0.952	0.952	0.952	0.952
1X5	0.954	0.954	0.954	0.954	0.954	0.954
1X7	0.953	0.953	0.953	0.953	0.953	0.953
3X2	0.924	0.955	0.919	0.919	0.873	0.862
3X5	0.928	0.946	0.915	0.915	0.883	0.883
3X7	0.929	0.955	0.916	0.916	0.872	0.863
6X2	0.96	0.967	0.95	0.95	0.935	0.925
6X5	0.967	0.968	0.954	0.954	0.944	0.938
6X7	0.962	0.966	0.949	0.949	0.935	0.925
7X2	0.963	0.965	0.954	0.954	0.942	0.937
7X5	0.958	0.958	0.948	0.948	0.939	0.935
7X7	0.962	0.963	0.951	0.951	0.939	0.934

시스템 종류를 나타내는 행의 1X2의 표현에서 1은 가중치 c값에 따른 시스템 번호를 나타내고 2는 가중치 f값에 따른 시스템 번호를 나타낸다. 표의 열은 구문의미트리 비교기의 한계값을 나타낸다. 한계값이 1에 근접할수록 정확히 매칭되어야 유사트리로 판정되고 0에 근접할수록 느슨히 매칭되어도 유사트리로 인정됨을 의미한다. 표의 시스템의 각 값들은 상관계수 값으로 가중치 c값과 f값, 그리고 구문의미트리 비교기의 한계값을 반영한 시스템의 유사문서 판별의 정확성을 나타낸다. 1에 근접할수록 사람이 판별한 값과 유사하다.

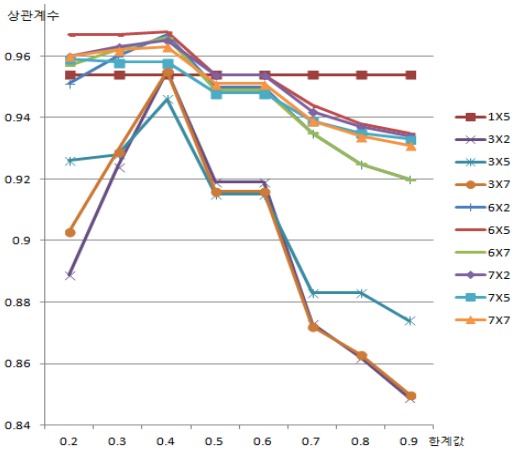

그림 3. 구문의미트리 비교기를 이용한 시스템별 유사도 상관계수

[그림 3]은 [표 9]의 내용을 그래프로 나타낸 것이다. 그래프를 보면 각 시스템은 구문의미트리 비교기의 한계값에 따라 상관계수 값이 변화됨을 볼 수 있다. 1X5의 시스템의 경우 용어형태소트리의 가중치가 (1,0,0)이므로 구문의미트리를 적용하지 않은 것이다. 따라서 구문의미트리 비교기의 한계값과 관계없이 동일한 상관계수값 0.954을 보인다. 3X2의 시스템의 경우는 용어형태소트리의 가중치가 (0,0,1)이므로 구문의미트리만 적용한 것이다. 이 시스템은 한계값의 변화에 따라 시스템의 상관계수가 변화됨을 보이고 있으며 한계값이 0.4일때 가장 높은 값 0.955를 나타내고 있다. 그 값은 구문의미분석을 반영하지 않은 1X2 시스템보다 미세하지만 더 나은 결과이다. 따라서 1X2 시스템에 구문의미분

석을 가중한 6X2 시스템은 보다 더 나은 결과를 보일 것으로 예상되고 실제로 그 시스템을 실험한 결과 월등히 나은 0.967의 상관계수 값을 보였다. 이 실험 결과로 본 논문의 구문의미트리 비교기를 이용한 유사문서 판별 시스템이 효과가 있음을 알 수 있었다.

가중치 복합의 실험으로 우리는 한계값이 0.4일때의 6X5 시스템이 가장 좋은 결과값을 가져옴을 알 수 있었다. 6X5 시스템은 용어와 형태소, 트리의 복합비율(1/2, 0, 1/2)과 기본, 저빈도, 관용표현의 복합비율 (2/5, 2/5, 1/5)로 가중된 시스템이다. 즉, 용어 추출과 함께 구문분석 방법이 반영되고 저빈도, 관용표현이 반영된 방법이 의미있음을 알려준다.

4. 개선점

실험결과로 본 시스템이 효과가 있음을 입증하였다. 그렇지만 시스템을 실험한 결과 개선의 여지가 있음을 알 수 있었다.

첫째가 구문의미트리 비교기의 개선이다. 구문의미트리 비교기를 적용한 결과 유사트리 추출기의 예를 든 것처럼 유사한 구문의미트리를 유사한 것으로 판단하여 유사문서 판별의 결과 상관계수가 0.935에서 0.968까지 개선되었다. 그렇지만 구문의미트리 비교기의 정의가 최적이냐 하는 점은 고려할 문제이다. 즉, 구문의미트리의 특징상 원칙 1과 2-1, 2-2를 반영한 가중치 α와 β, 그리고 분모에 반영한 배타적인 자식트리의 수에 대한 계수값을 무엇으로 할 것인지에 대한 연구가 필요하다.부가적으로 어휘의미 유사도 계산에 대한 부분도 개선을 해야 할 것으로 생각된다.

둘째는 구문의미분석의 질 문제이다. 본 시스템의 실험에서 구문의미분석의 오류로 인해 결과가 좋지 못함을 알 수 있었다. 본 논문의 구문의미분석은 단문 레벨의 분석을 시도하였다. 그렇지만 아직 처리하지 못하는 예가 있다. 예를 들면 '정수형은 값이 바뀔 수 있지만 ... 가능하다'에서 '(ADNMEN 가능 (TOP 정수형))'과 같은 결과를 내었다. 이것은 '바뀔 수 있지만'의 처리에서 있지만의 유형에 대한 처리를 하지 못하였다. 이와 같이 커버하지 못한 유형을 처리하도록 구문의미분석을 개선할 필요가 있다.

셋째는 어휘의 의미를 추출하는 의미추출기의 개선이 다. 본 논문에서 적용한 의미추출기가 커버할 수 있는 단어수는 8631개이다. 그렇지만 아직 의미를 추출하지 못하는 단어가 있음을 알 수 있었다. 구문의미트리 비교기의 효과를 제대로 살리기 위해서는 더욱 이를 확장할 필요가 있다.

마지막으로 더 다양하고 많은 문서에 대해서도 좋은 결과를 얻을 수 있도록 시스템을 튜닝하는 것이 필요하다. 다른 방법으로 문서 유형을 파악하고 그 유형에 맞는 시스템을 적용하는 적응형 문서판별 시스템으로 개선하는 것도 고려할 연구과제이다.

V. 결 론

문서 복제나 표절 검출의 필요성에 따라 많은 연구가 이루어지고 있다. 그러나 자연어 처리의 문제가 표절검사 시스템의 성능 향상의 발목을 잡고 있다. 본 논문은 구문의미분석을 이용한 시스템을 개발하고자 구문의미트리 비교기를 설계, 구현하고 이를 이용하여 유사문서 판별을 하였다. 시스템을 실험한 결과 구문의미트리 비교기를 이용한 유사문서 판별 시스템의 성능을 검증할 수 있었다. 따라서 본 시스템에서 개발한 구문의미트리 비교기는 유사문서 판별 분야 뿐 아니라 문서 분류, 문서 군집화 등의 다양한 분야에 응용할 수 있을 것으로 기대된다.

그렇지만 본 논문에서 개발한 구문의미트리 비교기는 개선의 여지가 남아 있다. 비교기에 들어있는 인수들의 값은 구문의미트리에 나타난 구문의미적 특징을 반영하여 설정하였지만 그 값이 최적인지는 알 수가 없다. 이를 찾기 위해서 많은 연구와 실험이 필요하다. 또한 어휘의 의미유사도 비교기도 개선할 필요가 있다.

그리고 올바른 구문의미분석과 어휘의 의미 추출은 본 시스템의 기초이므로 구문의미분석과 의미추출기의 성능을 향상할 필요가 있다. 또한 본 시스템이 일반성을 가지기 위해서는 다양한 종류의 검사문서를 실험하여 시스템을 개선할 필요가 있다.

참 고 문 헌

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

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

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

[4] 강원석, 황도삼, Jung H Kim, “구문의미분석을 이용한 유사문서판별기”, 한국콘텐츠학회논문지, 제14권, 제3호, pp.40-51, 2014.

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

[6] 김영철, 최재영, “구문트리에서 키워드 추출을 이용한 프로그램 유사도 평가”, 정보처리학회논문지A, 제12-A권, 제2호, pp.109-116, 2005.

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

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

[9] Daniel R. White and Mike S. Joy, "Sentence-Based Natural Language Plagiarism Detection," ACM Journal on Educational Resources in Computing, Vol.4, No.4, pp.1-20, 2004.

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

[11] 최성필, 정창후, 전홍우, 조현양, “시맨틱 구문 트리 커널을 이용한 생명공학 분야 전문용어간 관계 식별 및 분류 연구”, 한국문헌정보학회지, 제45권, 제2호, pp.251-275, 2011.

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

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

[14] 손정우, 박성배, “구조 및 의미 정보를 활용한 파스 트리 커널 기반의 온톨로지 정렬 방법”, 정보과학회논문지: 소프트웨어 및 응용, 제36권, 제4호, pp.329-334, 2009.

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

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

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

[18] 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