(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0032102
(43) 공개일자 2019년03월27일

(51) 국제특허분류(Int. Cl.)
*G06F 16/00* (2019.01) *G06F 17/27* (2006.01)
(52) CPC특허분류
*G06F 16/93* (2019.01)
*G06F 16/3347* (2019.01)
(21) 출원번호 10-2017-0120617
(22) 출원일자 2017년09월19일
심사청구일자 2017년09월19일

(71) 출원인
**조선대학교산학협력단**
광주광역시 동구 필문대로 309 (서석동)
(72) 발명자
**김판구**
광주광역시 북구 설죽로 600, 101동 904호 (일곡동, 삼호아파트)
**홍택은**
전라남도 장흥군 용산면 장흥대로 2212
*(뒷면에 계속)*
(74) 대리인
**특허법인아이엠**

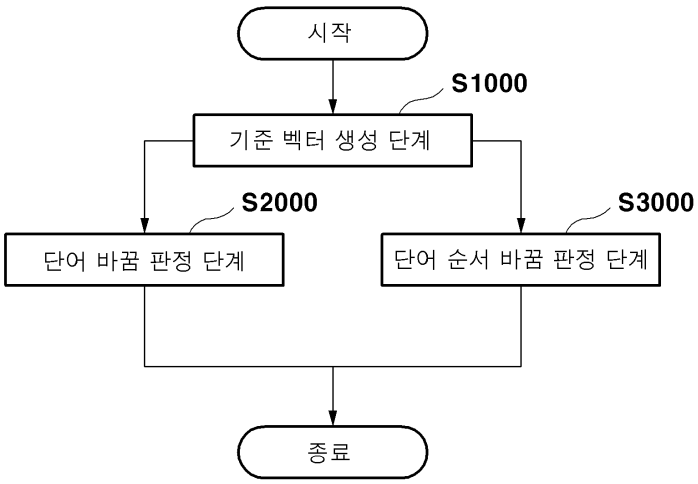
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **단어 바꿈과 순서 바꿈 검출을 통한 문서 표절 검사 방법**

## (57) 요 약

본 발명은 단어 바꿈과 순서 바꿈 검출을 통한 문서 표절 검사 방법에 관한 것으로, 보다 구체적으로는 검사하고자 하는 문서의 문장 내에서 다른 문서의 문장의 단어를 유사한 단어로 바꿔서 표절하는 단어 바꿈이나, 단어의 순서만을 바꿔서 표절하는 순서 바꿈을 검출하여 문서의 표절 여부를 검사할 수 있는 단어 바꿈과 순서 바꿈 검출을 통한 문서 표절 검사 방법에 관한 것이다.

*대 표 도* - 도1



(52) CPC특허분류
*G06F 17/2795* (2013.01)
(72) 발명자
**김정인**
광주광역시 광산구 첨단중앙로68번길 131 (산월동, 부영아파트)
**이은지**
서울특별시 송파구 중대로 24, 304동 501호 (문정동, 올림픽훼밀리타운)

이 발명을 지원한 국가연구개발사업
과제고유번호 1345257005
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 지역혁신창의인력양성사업
연구과제명 설화콘텐츠의 지능형 스토리텔링 지원 시스템 및 Beacon기반 스마트서비스 개발
기 여 율 1/1
주관기관 조선대학교 산학협력단
연구기간 2016.10.01 ~ 2017.09.30

# 명 세 서

## 청구범위

### 청구항 1

공지된 문서들을 수집하여 상기 수집된 문서들의 문장을 단어별로 벡터화한 기준 단어 벡터와 문장별로 벡터화한 기준 문장 벡터를 생성하는 기준 벡터 생성 단계; 및

검사 대상 문서를 입력받아 단어 벡터(이하'대상 단어 벡터'이라함)와 문장 벡터(이하'대상 문장 벡터'이라함)을 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 문장(이하'유사 문장'이라함)을 추출하며, 상기 기준 단어 벡터들 중 상기 대상 단어 벡터와 유사한 단어(이하'유의어'이라함)를 추출하여, 상기 유사 문장 내 상기 유의어의 존재 정도에 따라 단어 바꿈을 판정하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

### 청구항 2

제 1 항에 있어서,

상기 기준 단어 벡터를 생성 단계:는

상기 공지된 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계;

토큰화된 단어를 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계; 및

상기 태깅된 품사별 단어를 Word2Vec 알고리즘을 이용하여 벡터화된 기준 단어 벡터를 생성하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

### 청구항 3

제 2 항에 있어서,

상기 기준 문장 벡터를 생성 단계:는

상기 공지된 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계;

토큰화된 단어들을 Sent2Vec 알고리즘을 이용하여 벡터화된 기준 문장 벡터를 생성하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법.

### 청구항 4

제 3 항에 있어서,

상기 단어 바꿈을 판정하는 단계:는

상기 검사 대상 문서의 문장을 입력받는 단계;

상기 검사 대상 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계;

토큰화된 단어로부터 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계;및

상기 유의어가 상기 유사 문장 내에 포함되는 비율을 계산하여 단어 바꿈을 판정하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

### 청구항 5

제 4 항에 있어서,

상기 유사 문장을 검출하는 단계:는

상기 검사 대상 문서의 문장에서 토큰화된 단어들을 상기 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 추출

하는 단계;및

상기 기준 문장 벡터들 중 추출된 대상 문장 벡터와 유사한 유사 문장을 추출하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

**청구항 6**

제 5 항에 있어서,

상기 유의어 검출 단계:는

상기 검사 대상 문서의 문장에서 토큰화된 단어를 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계;

태깅된 단어를 상기 Word2Vec 알고리즘을 이용하여 대상 단어 벡터를 추출하는 단계;및

상기 기준 단어 벡터들 중 추출된 대상 단어 벡터와 유사한 유의어를 추출하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

**청구항 7**

제 6 항에 있어서,

상기 기준 벡터 생성 단계 이후에, 상기 단어 바꿈을 판정하는 단계와 병렬적으로 수행되고,

상기 검사 대상 문서를 입력받아 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 추출하며, 상기 대상 문서의 문장과 상기 유사 문장 내에 동일한 단어의 존재 정도에 따라 단어 순서 바꿈을 판정하는 단계;를 더 포함하는 것을 특징으로 하는 문서 표절 검사 방법

**청구항 8**

제 7 항에 있어서,

상기 단어 순서 바꿈을 판정하는 단계:는

상기 검사 대상 문서의 문장을 입력받는 단계;

상기 검사 대상 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계;

토큰화된 단어로부터 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계;

상기 유사 문장과 상기 검사 대상 문서의 문장을 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계;

태깅된 유사 문장과 검사 대상 문서의 문장에서 각각 어간을 추출하고 상기 추출된 어간을 기본형으로 변형하는 단계; 및

상기 기본형으로 변형된 유사 문장의 어간과 검사 대상 문서의 문장의 어간이 두 단어 이상 연속적으로 일치되는 비율을 계산하여 단어 순서 바꿈을 판정하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

**청구항 9**

제 8 항에 있어서,

상기 단어의 순서 바꿈을 판정하는 단계에서,

상기 유사 문장을 검출하는 단계:는

토큰화된 단어를 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 생성하는 단계;

상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법

**청구항 10**

컴퓨터와 결합하여 제 1 항 내지 제 9 항의 방법을 수행하는 기록 매체에 저장된 컴퓨터 프로그램

**청구항 11**

제 10 항의 컴퓨터 프로그램이 저장되며, 상기 문서 표절 검사를 수행하는 컴퓨터.

## 발명의 설명

### 기 술 분 야

[0001] 본 발명은 단어 바꿈과 순서 바꿈 검출을 통한 문서 표절 검사 방법에 관한 것으로, 보다 구체적으로는 검사하고자 하는 문서의 문장 내에서 다른 문서의 문장의 단어를 유사한 단어로 바꿔서 표절하는 단어 바꿈이나, 단어의 순서만을 바꿔서 표절하는 순서 바꿈을 검출하여 문서의 표절 여부를 검사할 수 있는 단어 바꿈과 순서 바꿈 검출을 통한 문서 표절 검사 방법에 관한 것이다.

### 배 경 기 술

[0002] 컴퓨터 관련 산업과 인터넷이 발달하면서 필기도구를 이용하여 종이 문서에 내용을 작성하던 시대가 최근에는 컴퓨터를 통해 일반인들이 용이하고 편리하게 내용을 작성할 수 있는 전자 문서로 대체되고 있다.

[0003] 전자 문서란 컴퓨터 등 정보처리능력을 가진 장치에 의하여 작성되어 송수신되거나 저장된 문서 형식의 자료로서 표준화된 것을 말한다.

[0004] 이러한 전자 문서는 작성과 편집이 용이할 뿐만 아니라 작성된 전자 문서는 다양한 경로로 유통 및 공유될 수 있으므로, 검색 엔진을 이용하여 일반인들이 손쉽게 얻을 수 있다는 장점이 있다.

[0005] 특히, 논문이나 리포터를 작성하는 학생 또는 연구원들에게는 원하는 문서를 검색하여 유용한 참고 자료로 사용될 수 있으나, 반면에, 손쉽게 문서들을 얻을 수 있게 되면서, 원본 문서로부터 내용을 그대로 인용하여 작성하는 표절 문제가 발생하기도 한다.

[0006] 또한, 최근에는 문서 내용을 그대로 인용하여 작성하는 표절을 넘어서, 인용된 문장 내의 단어를 유의어로 대체하여 작성하거나, 상기 인용된 문장 내의 단어들의 순서만을 바꿔서 표절하는 행위들이 증가하고 있다.

[0007] 종래에는 문서의 표절을 검출하기 위해서 단순히 문장 내에서 일정한 개수의 단어가 연쇄적으로 검출되는 경우에 표절이라고 판정하였으나 표절을 피하기 위해 인용된 문장 내의 단어를 유의어로 대체하거나, 상기 인용된 문장 내의 단어 순서를 바꾸는 표절 행위들에 대한 표절 검출이 어려운 문제점이 있다.

## 발명의 내용

### 해결하려는 과제

[0008] 본 발명은 상술한 문제점을 해결하기 위해 안출된 것으로 본 발명의 목적은 표절 행위 중 원본 문서의 문장 내 단어를 유사한 의미의 유의어로 대체하여 표절하는 단어 바꿈과 원본 문서의 문장 내 단어의 앞, 뒤 순서를 바꿔서 표절하는 단어 순서 바꿈을 검출하여 정밀한 표절 검사가 가능하게 할 수 있는 단어 바꿈과 단어 순서 바꿈을 통한 표절 검사 방법을 제공하는 데 있다.

### 과제의 해결 수단

[0009] 상술한 목적을 해결하기 위해 본 발명은 공지된 문서들을 수집하여 상기 수집된 문서들의 문장을 단어별로 벡터화한 기준 단어 벡터와 문장별로 벡터화한 기준 문장 벡터를 생성하는 기준 벡터 생성 단계; 및 검사 대상 문서를 입력받아 단어 벡터(이하'대상 단어 벡터'이라함)와 문장 벡터(이하'대상 문장 벡터'이라함)을 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 문장(이하'유사 문장'이라함)을 추출하며, 상기 기준 단어 벡터들 중 상기 대상 단어 벡터와 유사한 단어(이하'유의어'이라함)를 추출하여, 상기 유사 문장 내 상기 유의어의 존재 정도에 따라 단어 바꿈을 판정하는 단계;를 포함하는 것을 특징으로 하는 문서 표절 검사 방법을 제공하는 데 있다.

[0010] 바람직한 실시예에 있어서, 상기 기준 단어 벡터를 생성 단계:는 상기 공지된 문서의 문장을 단어 단위로 토큰

화하고 특수 문자를 제거하는 단계; 토큰화된 단어를 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계; 및 상기 태깅된 품사별 단어를 Word2Vec 알고리즘을 이용하여 벡터화된 기준 단어 벡터를 생성하는 단계;를 포함한다.

[0011] 바람직한 실시예에 있어서, 상기 기준 문장 벡터를 생성 단계:는 상기 공지된 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계; 토큰화된 단어들을 Sent2Vec 알고리즘을 이용하여 벡터화된 기준 문장 벡터를 생성하는 단계;를 포함한다.

[0012] 바람직한 실시예에 있어서, 상기 단어 바꿈을 판정하는 단계:는 상기 검사 대상 문서의 문장을 입력받는 단계; 상기 검사 대상 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계; 토큰화된 단어로부터 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계; 및 상기 유의어가 상기 유사 문장 내에 포함되는 비율을 계산하여 단어 바꿈을 판정하는 단계;를 포함한다.

[0013] 바람직한 실시예에 있어서, 상기 유사 문장을 검출하는 단계:는 상기 검사 대상 문서의 문장에서 토큰화된 단어들을 상기 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 추출하는 단계; 및 상기 기준 문장 벡터들 중 추출된 대상 문장 벡터와 유사한 유사 문장을 추출하는 단계;를 포함한다.

[0014] 바람직한 실시예에 있어서, 상기 유의어 검출 단계:는 상기 검사 대상 문서의 문장에서 토큰화된 단어를 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계; 태깅된 단어를 상기 Word2Vec 알고리즘을 이용하여 대상 단어 벡터를 추출하는 단계; 및 상기 기준 단어 벡터들 중 추출된 대상 단어 벡터와 유사한 유의어를 추출하는 단계;를 포함한다.

[0015] 바람직한 실시예에 있어서, 상기 기준 벡터 생성 단계 이후에, 상기 단어 바꿈을 판정하는 단계와 병렬적으로 수행되고, 상기 검사 대상 문서를 입력받아 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 추출하며, 상기 대상 문서의 문장과 상기 유사 문장 내에 동일한 단어의 존재 정도에 따라 단어 순서 바꿈을 판정하는 단계;를 더 포함한다.

[0016] 바람직한 실시예에 있어서, 상기 단어 순서 바꿈을 판정하는 단계:는 상기 검사 대상 문서의 문장을 입력받는 단계; 상기 검사 대상 문서의 문장을 단어 단위로 토큰화하고 특수 문자를 제거하는 단계; 토큰화된 단어로부터 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계; 상기 유사 문장과 상기 검사 대상 문서의 문장을 품사별로 태깅(POS,(Part of speech tagging))하고 조사를 제거하는 단계; 태깅된 유사 문장과 검사 대상 문서의 문장에서 각각 어간을 추출하고 상기 추출된 어간을 기본형으로 변형하는 단계; 및 상기 기본형으로 변형된 유사 문장의 어간과 검사 대상 문서의 문장의 어간이 두 단어 이상 연속적으로 일치되는 비율을 계산하여 단어 순서 바꿈을 판정하는 단계;를 포함한다.

[0017] 바람직한 실시예에 있어서, 상기 단어의 순서 바꿈을 판정하는 단계에서, 상기 유사 문장을 검출하는 단계:는 토큰화된 단어를 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 생성하는 단계; 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 검출하는 단계;를 포함한다.

[0018] 또한, 본 발명은 컴퓨터와 결합하여 상기 표절 검사 방법을 수행하는 기록 매체에 저장된 컴퓨터 프로그램을 더 제공한다.

[0019] 또한, 본 발명은 상기 컴퓨터 프로그램이 저장되며, 상기 문서 표절 검사를 수행하는 컴퓨터를 더 제공한다.

## 발명의 효과

[0021] 본 발명의 단어 바꿈과 순서 바꿈을 통한 표절 검사 방법에 의하면, 일반적인 단어의 연쇄적 검출을 이용한 표절 검사 방법을 넘어 원본 문서의 문장 내 단어를 유사한 의미의 유의어로 대체하여 표절하는 단어 바꿈과 원본 문서의 문장 내 단어의 순서를 앞, 뒤를 바꿔서 표절하는 단어 순서 바꿈과 같은 지능적인 표절 행위를 검출할 수 있으므로 정밀하고 정확한 표절 검사가 가능하다는 장점이 있다.

## 도면의 간단한 설명

[0022] 도 1은 본 발명의 일 실시예에 따른 표절 검사 방법을 설명하기 위한 흐름도,

도 2는 본 발명의 일 실시예에 따른 기준 백터 생성 단계를 설명하기 위한 흐름도,

도 3은 본 발명의 일 실시예에 따른 단어 바꿈을 판정하는 단계를 설명하기 위한 흐름도,

도 4는 본 발명의 일 실시예에 따른 단어 순서 바꿈을 판정하는 단계를 설명하기 위한 흐름도이다.

## *발명을 실시하기 위한 구체적인 내용*

[0023] 본 발명에서 사용되는 용어는 가능한 현재 널리 사용되는 일반적인 용어를 선택하였으나, 특정한 경우는 출원인이 임의로 선정한 용어도 있는데 이 경우에는 단순한 용어의 명칭이 아닌 발명의 상세한 설명 부분에 기재되거나 사용된 의미를 고려하여 그 의미가 파악되어야 할 것이다.

[0024] 이하, 첨부한 도면에 도시된 바람직한 실시예들을 참조하여 본 발명의 기술적 구성을 상세하게 설명한다.

[0025] 그러나 본 발명은 여기서 설명되는 실시예에 한정되지 않고 다른 형태로 구체화될 수도 있다. 명세서 전체에 걸쳐 동일한 참조번호는 동일한 구성요소를 나타낸다.

[0026] 본 발명의 일 실시예에 따른 표절 검사 방법은 단어 바꿈과 순서 바꿈의 표절을 검출함으로써 정밀한 표절 검사를 수행할 수 있는 방법이다.

[0027] 또한, 본 발명의 일 실시예에 따른 표절 검사 방법은 컴퓨터에 의해 수행되며, 상기 컴퓨터에는 상기 컴퓨터를 기능시켜 표절 검사를 수행하게 하는 컴퓨터 프로그램이 저장된다.

[0028] 또한, 상기 컴퓨터는 일반적인 퍼스널 컴퓨터뿐만 아니라, 스마트폰이나 태블릿 PC와 같은 스마트 기기를 포함하는 광의의 컴퓨팅 장치를 의미한다.

[0029] 또한, 상기 컴퓨터 프로그램은 별도의 기록 매체에 저장되어 제공될 수 있으며, 상기 기록매체는 본 발명을 위하여 특별히 설계되어 구성된 것들이거나 컴퓨터 소프트웨어 분야에서 통상의 지식을 가진 자에게 공지되어 사용 가능한 것일 수 있다.

[0030] 예를 들면, 상기 기록매체는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD, DVD와 같은 광 기록 매체, 자기 및 광 기록을 겸할 수 있는 자기-광 기록 매체, 롬, 램, 플래시 메모리 등 단독 또는 조합에 의해 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치일 수 있다.

[0031] 또한, 상기 컴퓨터 프로그램은 프로그램 명령, 로컬 데이터 파일, 로컬 데이터 구조 등이 단독 또는 조합으로 구성된 컴퓨터 프로그램일 수 있고, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라, 인터프리터 등을 사용하여 컴퓨터에 의해 실행될 수 있는 고급 언어 코드로 짜여진 컴퓨터 프로그램일 수 있다.

[0032] 이하에서는 도 1 내지 도 4를 참조하여 본 발명의 표절 검사 방법에 대해 설명한다.

[0033] 도 1은 본 발명의 일 실시예에 따른 표절 검사 방법을 설명하기 위한 흐름도로 도 1을 참조하면, 본 발명의 표절 검사 방법은 공지된 문서들을 수집하여 상기 수집된 문서들의 문장을 단어별로 벡터화한 기준 단어 벡터들과 문장별로 벡터화한 기준 문장 벡터들을 생성하는 기준 벡터 생성 단계(S1000)와 검사 대상 문서를 입력받아 단어 벡터(이하'대상 단어 벡터'이라함)와 문장 벡터(이하'대상 문장 벡터'이라함)를 생성하고, 상기 기준 문장 벡터들 중 상기 문장 벡터와 유사한 문장(이하'유사 문장'이라함)을 추출하며, 상기 기준 단어 벡터들 중 상기 대상 단어 벡터와 유사한 단어(이하'유의어'이라함)을 추출하여, 상기 유사 문장 내 상기 유의어의 존재 정도에 따라 단어 바꿈을 판정하는 단계(S2000)를 포함한다.

[0034] 또한, 상기 기준 벡터 생성 단계 이후에(S1000)는 상기 단어 바꿈 검출 단계(S2000)와 병렬적으로 상기 검사 대상 문서를 입력받아 상기 대상 문장 벡터를 생성하고, 상기 기준 문장 벡터들 중 상기 대상 문장 벡터와 유사한 유사 문장을 추출하며, 상기 대상 문서의 문장과 상기 유사 문장 내 동일한 단어의 존재 정도에 따라 단어 순서 바꿈을 판정하는 단계(S3000)를 더 포함한다.

[0035] 이하에서는 도 2를 참조하여 상기 기준 벡터 생성 단계를 상세히 설명한다.

[0036] 도 2는 본 발명의 일 실시예에 따른 기준 벡터 생성 단계의 흐름도를 보여주는 도면으로 도 2를 참조하면, 상기 기준 벡터 생성 단계(S1000)는 먼저, 다양한 분야의 공지된 문서들을 수집한다(S1100).

[0037] 다음, 수집된 문서들의 문장들을 단어 단위로 토큰화 작업을 수행하고, 포함된 특수 문자를 제거한다(S1200).

[0038] 다음, 토큰화된 단어에 품사별 태깅(POS.(Part of Speech tagging)) 작업을 수행하고, 태깅된 단어들 중 조사를 제거한다(S1300).

[0039] 또한, 상기 품사별 태깅은 단어에 품사 정보를 부여하여 분류하는 과정을 의미한다.

[0040] 다음, 태깅된 단어들을 Word2Vec 알고리즘을 이용하여 대상 단어 벡터를 생성한다(S1400).

[0041] 또한, 상기 Word2Vec 알고리즘은 상기 태깅된 단어들의 연관성을 숫자로 벡터화하는 알고리즘으로, 기계 학습 분야의 알고리즘 중 하나이다.

[0042] 또한, 상기 토큰화된 단어를 이용하여 Sent2Vec 알고리즘을 통해 대상 문장 벡터를 생성한다.

[0043] 또한, 상기 Sent2Vec 알고리즘은 단어를 벡터화하는 Word2Vec 알고리즘과 달리 토큰화된 단어들을 이용하여 문장과 문장 사이의 연관성을 숫자로 벡터화한다.

[0044] 즉, 상기 기준 벡터 생성 단계는 수집된 문서들의 문장으로부터 검사받고자 하는 문장과 비교하기 위한 기준 단어 벡터와 기준 문장 벡터를 생성하는 것이다.

[0045] 이하에서는 도 3을 참조하여 상기 단어 바꿈을 판정하는 단계를 상세히 설명한다.

[0046] 도 3은 본 발명의 일 실시예에 따른 단어 바꿈을 판정하는 단계의 흐름도를 보여주는 도면으로 도 3을 참조하면, 상기 단어 바꿈을 판정하는 단계(S2000)는 먼저, 검사받고자 하는 문서의 문장(이하'검사 대상 문장'이라함)을 입력받는다(S2100).

[0047] 다음, 상기 검사 대상 문장을 단어 단위로 토큰화하고 특수 문자를 제거한다(S2200).

[0048] 다음, 토큰화된 단어에 품사별 태깅(POS.(Part of Speech tagging)) 작업을 수행하고, 태깅된 단어들 중 조사를 제거한다(S2300).

[0049] 다음, 태깅된 단어를 Word2Vec 알고리즘을 이용하여 대상 단어 벡터를 생성한다(S2400).

[0050] 다음, 상기 대상 단어 벡터와 유사한 기준 단어 벡터를 갖는 단어들을 추출한다(S2500).

[0051] 여기서, 상기 대상 단어 벡터와 기준 단어 벡터의 유사도를 판단하는 방법은 벡터 간의 거리를 구하는 방법인 유클리디안(Euclidean) 거리 측정 방법, 코사인(Cosine) 거리 측정 방법, 멘하탄(Manhattan) 거리 측정 방법 등을 이용할 수 있고, CNN(Convolution Neural Network)와 같은 딥러닝(Deep learning)를 이용하여 판단할 수 있다.

[0052] 또한, 상기 토큰화된 단어를 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 생성한다(S2600).

[0053] 다음, 상기 대상 문장 벡터와 유사한 기준 문장 벡터를 갖는 유사 문장을 추출한다(S2700).

[0054] 여기서, 상기 대상 문장 벡터와 기준 문장 벡터의 유사도를 판단하는 기준은 벡터 간의 거리를 구하는 방법인 유클리디안(Euclidean) 거리 측정 방법, 코사인(Cosine) 거리 측정 방법, 멘하탄(Manhattan) 거리 측정 방법 등을 이용할 수 있고, CNN(Convolution Neural Network)와 같은 딥러닝(Deep learning)를 이용하여 판단할 수 있다.

[0055] 다음, 아래의 식 1을 이용하여 상기 유사 문장 내 상기 유의어가 포함되는 표절률을 계산한다(S2800).

*수학식 1*

[0056]
$$\text{표절률}(\%) = \frac{\text{유사문장에서유의어와일치하는단어수}}{\text{유사문장의단어수}}$$

[0058] 표 1은 상기 단어 바꿈을 검출하는 실시 예를 보여주는 것이다.

*표 1*

[0059]

| 단계 | 내용 |
|---|---|
| 검사 대상 문장 입력 | 사의가 있는 인간이라면 혁신의 문제들을 이렇게 제의할 것이다. |
| 토큰화 | '사의가', '있는' 인간이라면', '혁신의', '문제들을', '이렇게', '제의할', 것이다' |

| | | |
|---|---|---|
| 품사 태깅 및 조사 제거 | 품사 태깅 | 사의(명사), 가(조사), 있는(형용사), 인간(명사), 이(한정사), 라면(명사), 혁신(명사), 의(조사), 문제(명사), 들(접미사), 을(조사), 이렇게(부사), 제의할(동사), 거(명사), 이다(조사) |
| | 조사 제거 | 사의', '있는', '사람', '라면', '혁신', '문제', '들', '이렇게', '제의할', '것' |
| 유의어 검출 | 단어(10) | 유의어 |
| | 사의 | 생각, 각오, 사고, 의향 |
| | 있는 | 위치한, 존재하는 |
| | 인간 | 사람, 인성, 인물, 인류, 인품 |
| | 라면 | 라고, 하면 |
| | 혁신 | 개혁, 쇄신, 혁명, 개신, 변혁 |
| | 문제 | 물의, 사건, 물음, 문, 숙제 |
| | 들 | 을, 는, 이 |
| | 이렇게 | 이렇다, 요렇, 여사하게 |
| | 제의할 | 제안할, 제언할, 제기할 |
| | 것 | 거, 데, 차지 |
| 유사 문장 검출 | 생각이 있는 사람은 혁명의 문제를 이렇게 제기할 것이다. | |
| 표절 검출 | 일치하는 유의어(4) | 생각, 사람, 혁명, 제기할 |
| | 표절률 | $표절률(\%)=\frac{4}{10}\times 100=40\%$ |

[0061] 표 1을 참조하면 검사 대상 문장 내 단어들의 유의어로 검출된 단어들 중 상기 유사 문장에 포함된 단어와 동일한 단어는 '생각', '사람', '혁명', '제기할' 으로 조사가 제거된 10개의 태깅된 단어들 중 중 4개의 단어의 유의어가 상기 유사 문장의 단어와 동일함으로 40%의 표절률을 나타낸다.

[0062] 또한, 표 1에는 도시하지 않았으나 상기 유사 문장 검출 후에는 상기 유의어들 중 동일한 단어가 포함되어 있는지 비교하기 위해 상기 유사 문장은 토큰화 단계와 품사 태깅 및 조사 제거 단계를 수행한다.

[0063] 즉, 단어 바꿈을 판정하는 단계는 검사 대상 문장과 유사한 유사 문장을 추출하고, 상기 검사 대상 문장 내 단어와 유사한 의미를 갖는 유의어를 검출하여 상기 유사 문장 내 단어와 상기 유의어를 비교함으로써, 저자가 다른 문서의 문장을 유의어로 대체하는 표절을 하였는지 확인할 수 있다.

[0064] 이하에서는 도 4를 참조하여 단어 순서 바꿈을 판정하는 단계에 대해 상세히 설명한다.

[0065] 도 4는 본 발명의 일 실시예에 따른 단어 순서 바꿈을 판정하는 단계에 대한 흐름도를 보여주는 것으로, 도 4를 참조하면, 상기 단어 순서 바꿈을 판정하는 단계(S3000)는 먼저, 검사받고자 하는 검사 대상 문장을 입력받는다(S3100).

[0066] 다음, 상기 검사 대상 문장을 토큰화하고 특수 문자를 제거한다(S3200).

[0067] 다음, 토큰화된 단어들을 Sent2Vec 알고리즘을 이용하여 대상 문장 벡터를 생성한다(S3300).

[0068] 다음, 상기 대상 문장 벡터와 유사한 기준 문장 벡터를 갖는 유사 문장을 추출한다(S3400).

[0069] 여기서, 상기 대상 문장 벡터와 기준 문장 벡터의 유사도를 판단하는 기준은 벡터 간의 거리를 구하는 방법인 유클리디안(Euclidean) 거리 측정 방법, 코사인(Cosine) 거리 측정 방법, 멘하탄(Manhattan) 거리 측정 방법 등을 이용할 수 있고, CNN(Convolution Neural Network)와 같은 딥러닝(Deep learning)를 이용하여 판단할 수 있다.

[0070] 다음, 상기 유사 문장과 검사 대상 문장을 품사별 태깅하고, 조사를 제거한다(S3500).

[0071] 다음, 태깅된 유사 문장과 검사 대상 문서의 문장에서 어간을 추출하고 추출된 단어를 기본형으로 변형한다(S3600).

[0072] 본 발명에서는 상기 어간을 추출하고 기본형으로 변형하기 위해 트위터에서 오픈 소스로 제공하는 twitter-korean-text를 사용한다.

[0073] 다음, 기본형으로 변형된 유사 문장의 어간과 검사 대상 문장의 어간을 아래에 식 2를 이용하여 표절률을 도출한다(S3700).

*수학식 2*

[0074]
$$\text{표절률}(\%)=\frac{\text{두 단어 이상 연속적으로 일치하는 어간 수}}{\text{검사 대상 문장에서 어간 추출된 단어 수}}$$

[0075] 아래의 표 2는 단어 순서 바꿈을 판정하는 실시예를 보여주는 것이다.

*표 2*

[0076]

| 단계 | 내용 | |
|---|---|---|
| 검사 대상 문장 입력 | 농업문제는 노동계 정권이 애초에 해결했다. | |
| 토큰화 | '농업', '문제', '는', '노동계', '정권', '이', '애초', '에', '해결했', '다' | |
| 유사 문장 추출 | 노동계정권은 애초부터 농업문제를 해결했다. | |
| 품사 태깅 및 조사 제거 | 검사 대상 문장 | 농업(명사), 문제(명사), 노동계(명사), 정권(명사), 애초(명사), 해결했(동사), 다(어미) |
| | 유사 문장 | '노동계(명사), 정권(명사),, 애초(명사), , 농업(명사) 문제(명사), 해결했(동사), 다(어미) |
| 어간 추출 | 검사 대상 문장 | '농업', '문제', '노동계', '정권', '애초', '해결', '하다' |
| | 유사 문장 | '노동계', '정권', '애초;, '농업', '문제', '해결, '하다' |
| 표절 검출 | 두 단어 이상일치하는 단어(4) | (농업, 문제), (노동계, 정권) |
| | 표절률 | $\text{표절률}(\%)=\frac{4}{7}\times 100=57\%$ |

[0077] 표 2를 참조하면 태깅된 검사 대상 문장과 유사 문장에서 추출된 어간들 중 검사 대상 문장의 어간을 기준으로 서로 동일하며 연속적으로 일치하는 단어가 두 개 이상인 경우는 '농업'과 '문제' 및 '노동계'와 '정권'으로 검사 대상 문장 내 어간 추출된 7개의 단어 중 4개의 단어가 포함되므로 57%의 표절률을 나타낸다.

[0078] 또한, 본 발명에서는 상기 단어 순서 바꿈 표절의 기준을 두 단어 이상 일치한 경우로 정의하였으나, 사용자의 기준에 따라 이는 변경될 수 있다.

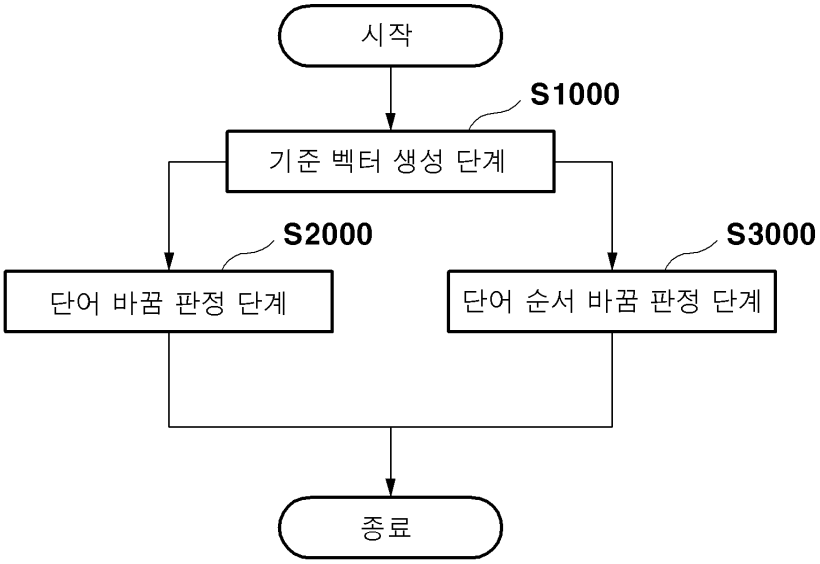
[0079] 또한, 표 2에는 도시하지는 않았으나 유사 문장 추출 후에는 태깅 단계를 수행하기 위해 상기 유사 문장을 토큰화하고 특수 문자를 제거하는 단계를 수행한다.

[0080] 즉, 단어 순서 바꿈 판정 단계는 검사 대상 문장과 유사한 유사 문장을 추출하고, 상기 검사 대상 문장 내 단어와 상기 유사 문장 내 단어들을 추출하여, 상기 유사 문장 내 단어들과 상기 검사 대상 문장 내 단어들의 유사성을 비교함으로써, 검사 대상 문서에서 다른 저자가 작성한 문서의 문장 내 단어의 순서만을 바꿔서 표절을 하였는지 확인할 수 있다.

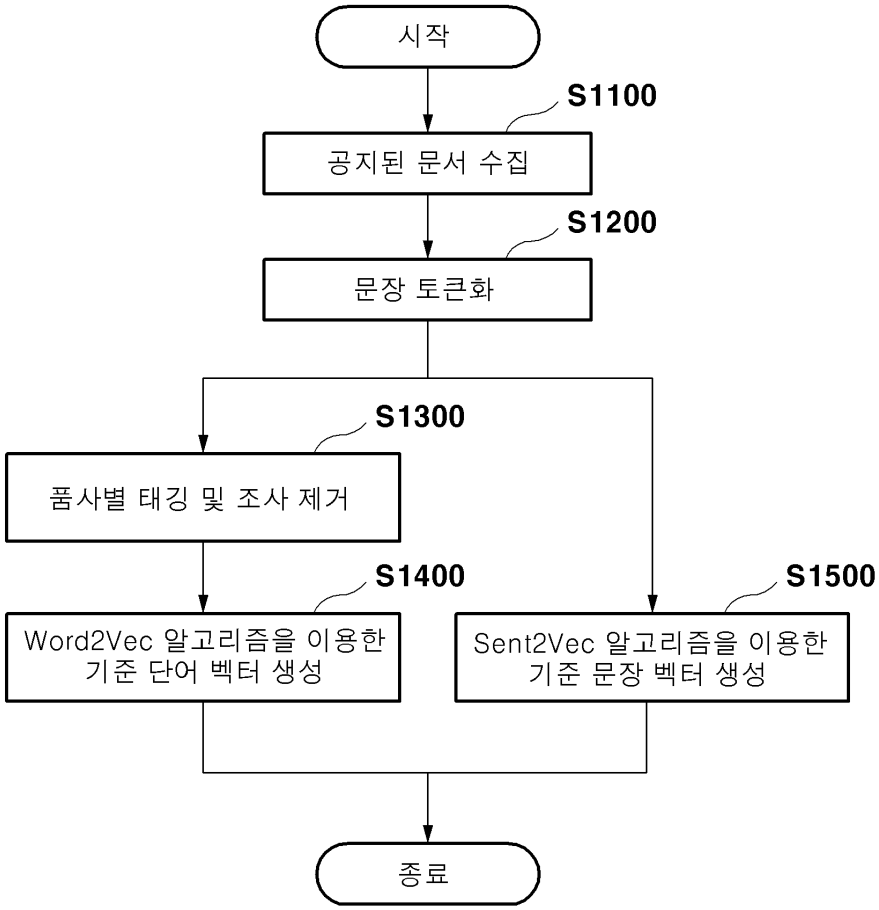
[0081] 이상에서 살펴본 바와 같이 본 발명은 바람직한 실시예를 들어 도시하고 설명하였으나, 상기한 실시예에 한정되지 아니하며 본 발명의 정신을 벗어나지 않는 범위 내에서 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변경과 수정이 가능할 것이다.

## 도면

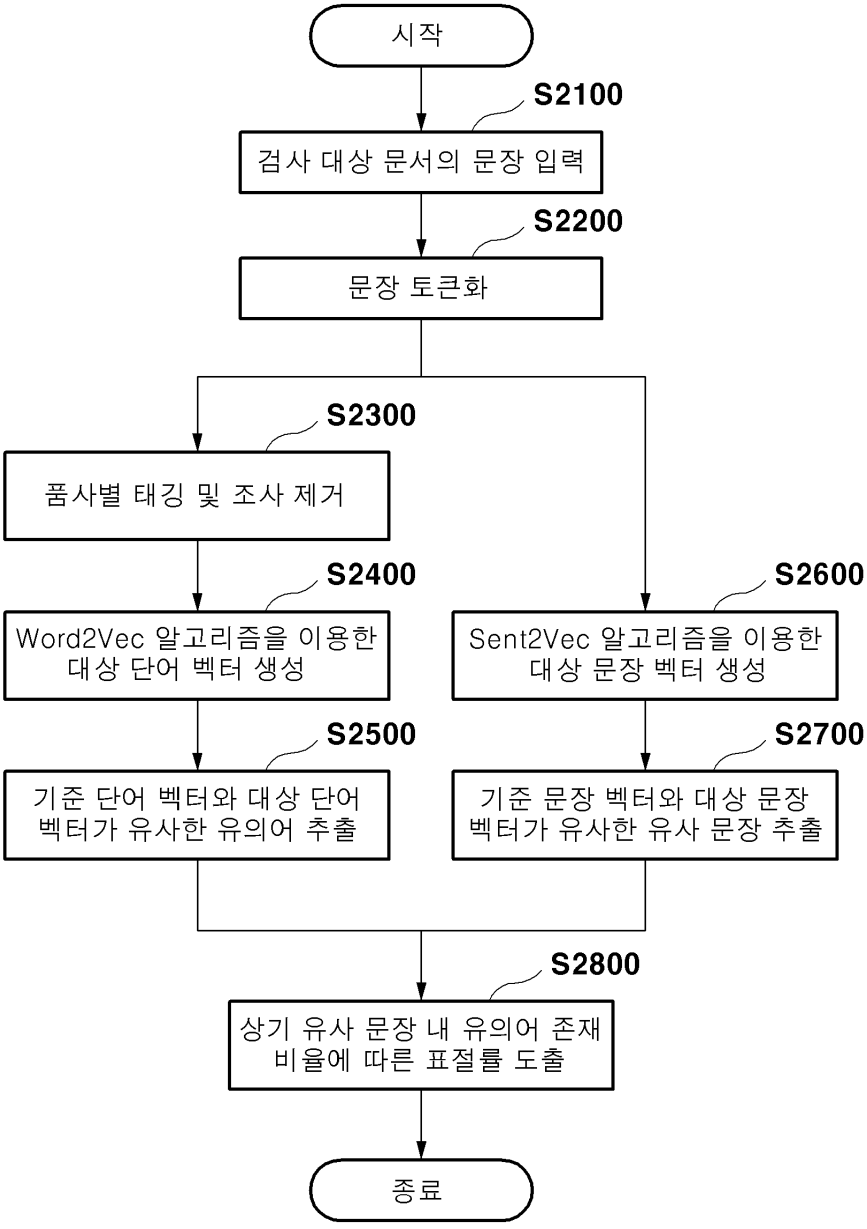
### 도면1


시작
S1000
기준 벡터 생성 단계
S2000
단어 바꿈 판정 단계
S3000
단어 순서 바꿈 판정 단계
종료

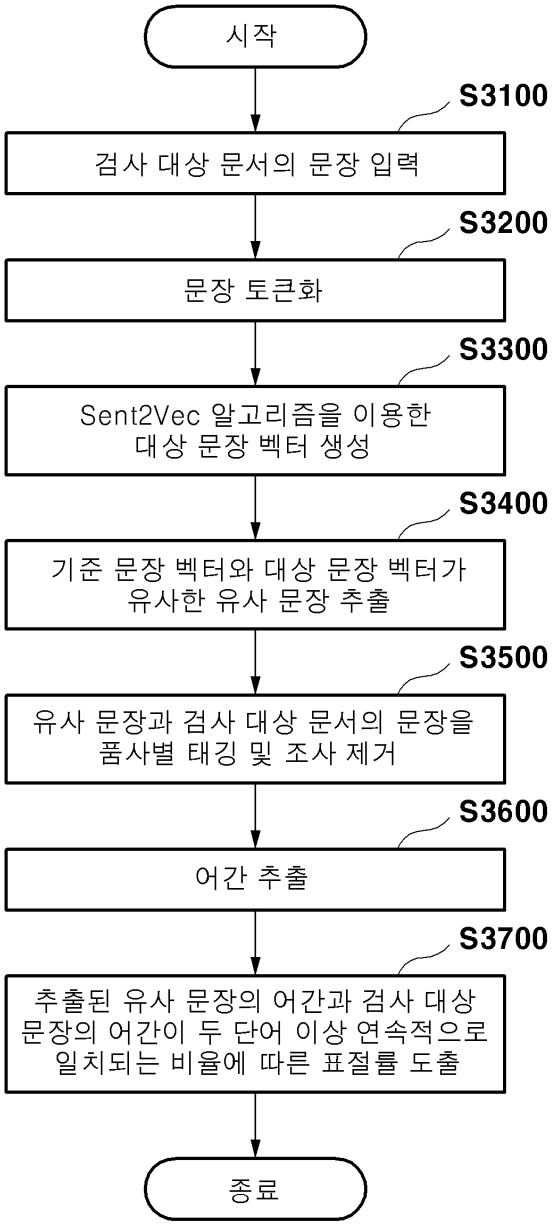
### 도면2


시작
S1100
공지된 문서 수집
S1200
문장 토큰화
S1300
품사별 태깅 및 조사 제거
S1400
Word2Vec 알고리즘을 이용한
기준 단어 벡터 생성
S1500
Sent2Vec 알고리즘을 이용한
기준 문장 벡터 생성
종료

도면3


시작
S2100
검사 대상 문서의 문장 입력
S2200
문장 토큰화
S2300
품사별 태깅 및 조사 제거
S2400
Word2Vec 알고리즘을 이용한 대상 단어 벡터 생성
S2500
기준 단어 벡터와 대상 단어 벡터가 유사한 유의어 추출
S2600
Sent2Vec 알고리즘을 이용한 대상 문장 벡터 생성
S2700
기준 문장 벡터와 대상 문장 벡터가 유사한 유사 문장 추출
S2800
상기 유사 문장 내 유의어 존재 비율에 따른 표절률 도출
종료

도면4


시작
S3100
검사 대상 문서의 문장 입력
S3200
문장 토큰화
S3300
Sent2Vec 알고리즘을 이용한
대상 문장 벡터 생성
S3400
기준 문장 벡터와 대상 문장 벡터가
유사한 유사 문장 추출
S3500
유사 문장과 검사 대상 문서의 문장을
품사별 태깅 및 조사 제거
S3600
어간 추출
S3700
추출된 유사 문장의 어간과 검사 대상
문장의 어간이 두 단어 이상 연속적으로
일치되는 비율에 따른 표절률 도출
종료