(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0042710
(43) 공개일자 2018년04월26일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01)
(52) CPC특허분류
G06F 17/2795 (2013.01)
G06F 17/2735 (2013.01)
(21) 출원번호 10-2016-0135209
(22) 출원일자 2016년10월18일
심사청구일자 없음

(71) 출원인
삼성에스디에스 주식회사
서울특별시 송파구 올림픽로35길 125 (신천동)
(72) 발명자
정동훈
서울특별시 송파구 올림픽로35길 125 (신천동, 삼성SDS West Campus)
(74) 대리인
특허법인가산

전체 청구항 수 : 총 14 항

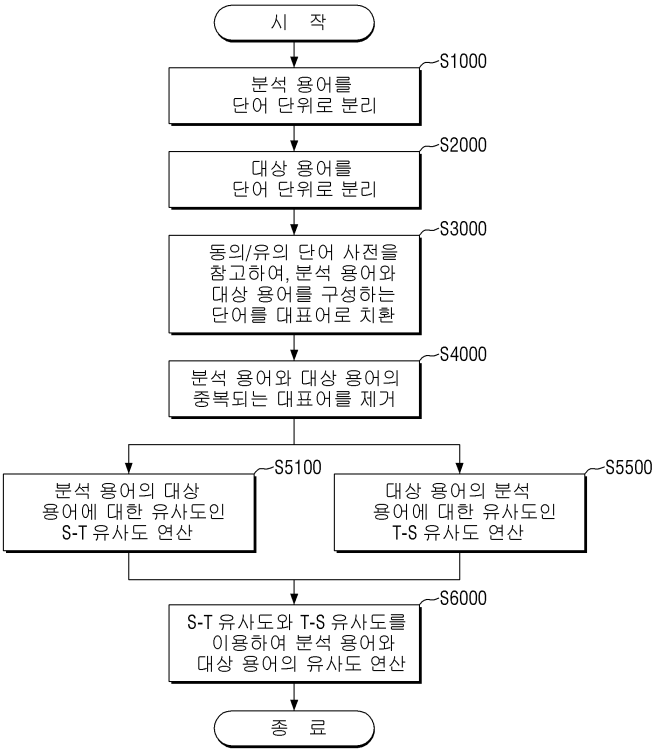
(54) 발명의 명칭 **유사도 분석 기반 이음 동의 항목 관리 방법 및 장치**

(57) 요 약

본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법은, 유사도 분석 장치가, 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계와 상기 유사도 분석 장치가, 제2 항목(Target)에서 상기 제2 항목의 하위 항목인 제2-1 항목 내지 제2-n 항목을 추출하는 단계와 상기

(뒷면에 계속)

대 표 도 - 도9



유사도 분석 장치가, 상기 제1-1 항목 내지 제1-m 항목의, 상기 제2 항목의 하위 항목에 대한 유사도를 이용하여, S-T 유사도를 연산하는 단계와 상기 유사도 분석 장치가, 상기 제2-1 항목 내지 제2-n 항목의, 상기 제1 항목의 하위 항목에 대한 유사도를 이용하여, T-S 유사도를 연산하는 단계 및 상기 유사도 분석 장치가, 상기 S-T 유사도와 상기 T-S 유사도를 이용하여, 상기 제1 항목과 상기 제2 항목의 유사도를 연산하는 단계를 포함하 되, 상기 S-T 유사도는 분석 항목(Source)을 구성하는 하위 항목이 대상 항목(Target)에 얼마나 포함되어 있는지를 기준으로 구한 유사도이고, 상기 T-S 유사도는 대상 항목(Target)을 구성하는 하위 항목이 분석 항목(Source)에 얼마나 포함되어 있는지를 기준으로 구한 유사도이다.

명 세 서

청구범위

청구항 1

유사도 분석 장치가, 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계;

상기 유사도 분석 장치가, 제2 항목(Target)에서 상기 제2 항목의 하위 항목인 제2-1 항목 내지 제2-n 항목을 추출하는 단계;

상기 유사도 분석 장치가, 상기 제1-1 항목 내지 제1-m 항목의, 상기 제2 항목의 하위 항목에 대한 유사도를 이용하여, S-T 유사도를 연산하는 단계;

상기 유사도 분석 장치가, 상기 제2-1 항목 내지 제2-n 항목의, 상기 제1 항목의 하위 항목에 대한 유사도를 이용하여, T-S 유사도를 연산하는 단계; 및

상기 유사도 분석 장치가, 상기 S-T 유사도와 상기 T-S 유사도를 이용하여, 상기 제1 항목과 상기 제2 항목의 유사도를 연산하는 단계를 포함하되,

상기 S-T 유사도는 분석 항목(Source)을 구성하는 하위 항목이 대상 항목(Target)에 얼마나 포함되어 있는지를 기준으로 구한 유사도이고,

상기 T-S 유사도는 대상 항목(Target)을 구성하는 하위 항목이 분석 항목(Source)에 얼마나 포함되어 있는지를 기준으로 구한 유사도인,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 2

제1항에 있어서,

상기 제1 항목과 상기 제2 항목의 유사도를 동의/유의 사전에 저장하는 단계를 더 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 3

제1항에 있어서,

상기 제1 항목과 상기 제2 항목의 유사도가 기 설정된 값 이상인 경우,

상기 제1 항목을 사용하려는 사용자에게, 상기 제1 항목 대신 상기 제2 항목과 상기 유사도를 제공하는 단계를 더 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 4

제1항에 있어서,

상기 제1 항목과 상기 제2 항목의 유사도가 기 설정된 값 이상인 경우,

상기 제1 항목이 상기 제2 항목의 표절인 것으로 판단하는 단계를 더 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계는,

상기 제1 항목의 어미 또는 조사를 제거하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 6

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계는,

상기 제1-1 항목 내지 제1-m 항목 중에서 임의의 두 항목을 선택하여, 상기 두 항목의 유사도가 기 설정된 값 이상인 경우에는 상기 두 항목 중에서 어느 한 항목을 제외하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 7

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1항목과 상기 제2 항목은 문서이고,

상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 문장이고,

상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는,

상기 문서에서 마침표를 기준으로 상기 문장을 추출하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 8

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1항목과 상기 제2 항목은 문장이고,

상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 용어이고,

상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는,

상기 문장에서 띄어 쓰기와 어미와 조사를 기준으로 상기 용어를 추출하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 9

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1항목과 상기 제2 항목은 용어이고,

상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 단어이고,

상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는,

상기 용어에서 형태소를 기준으로 의미를 가진 최소의 단위인 상기 단어를 추출하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 10

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제1-1 항목 내지 제1-m 항목의, 상기 제2 항목의 하위 항목에 대한 유사도를 이용하여, S-T 유사도를 연산하는 단계는,

상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 상기 제2 항목의 하위 항목에 대한 유사도를 동의/유의 사

전에서 조회하는 단계; 및

상기 조회 결과로 얻은, 상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 유사도의 값을 평균하여, 상기 S-T 유사도를 연산하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 11

제10항에 있어서,

상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 상기 제2 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계는,

상기 제1-1 항목 내지 제1-m 항목 중에서 특정 항목의 상기 제2 항목의 하위 항목에 대한 유사도가 상기 동의/유의 사전에 없는 경우,

상기 특정 항목의 하위 항목인 제3 항목을 추출하는 단계; 및

상기 제3항 항목의 상기 제2 항목의 하위 항목의 하위 항목에 대한 유사도를 상기 동의/유의 사전에 조회하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 12

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 제2-1 항목 내지 제2-n 항목의, 상기 제1 항목의 하위 항목에 대한 유사도를 이용하여, T-S 유사도를 연산하는 단계는,

상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 상기 제1 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계; 및

상기 조회 결과로 얻은, 상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 유사도의 값을 평균하여, 상기 T-S 유사도를 연산하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 13

제12항에 있어서,

상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 상기 제1 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계는,

상기 제2-1 항목 내지 제2-n 항목 중에서 특정 항목의 상기 제1 항목의 하위 항목에 대한 유사도가 상기 동의/유의 사전에 없는 경우,

상기 특정 항목의 하위 항목인 제4 항목을 추출하는 단계; 및

상기 제4항 항목의 상기 제1 항목의 하위 항목의 하위 항목에 대한 유사도를 상기 동의/유의 사전에 조회하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

청구항 14

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 S-T 유사도와 상기 T-S 유사도를 이용하여, 상기 제1 항목과 상기 제2 항목의 유사도를 연산하는 단계는,

상기 S-T 유사도와 상기 T-S 유사도의 최소값(min), 최대값(max), 평균값(avg) 중에서 어느 한 값을 상기 제1 항목과 상기 제2 항목의 유사도로 연산하는 단계를 포함하는,

유사도 분석 기반 이음 동의 항목 관리 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 유사도 분석에 기반을 두어 이음 동의 항목을 관리하는 방법 및 장치에 관한 것이다. 보다 자세하게는, 분석 항목을 최소 의미 단위인 단어로 분해하고, 분해된 단어의 유사도를 기준으로 분석 항목과 대상 항목 사이의 유사도를 연산하는 방법 및 그 방법을 수행하는 장치에 관한 발명이다.

배 경 기 술

[0002] 다양한 항목을 관리해야 하는 경우가 있다.

[0003] 예를 들면, 조직의 목표 달성의 정도를 평가하기 위한 목표 관리 시스템에서는 핵심 성과 지표(KPI; Key Performance Indicators)를 관리한다. A 조직이 등록한 "매출목표액 10% 증가"와 B 조직이 등록한 "가입 회원수 50% 증가"와 같은 핵심 성과 지표를 나타내는 항목을 관리해야 한다.

[0004] 다른 예를 들면, 일반 사용자를 대상으로 서비스를 제공하면서 각종 오류 상황을 대비한 안내 메시지를 관리한다. "아이디는 필수 입력 항목입니다." 또는 "입력한 메일 주소는 유효하지 않습니다."와 같은 안내 메시지를 나타내는 항목을 관리해야 한다.

[0005] 또 다른 예를 들면, 일반 사용자를 대상으로 서비스를 제공하는 대부분 시스템에서는 사용자의 편의를 강화하기 위한 FAQ(Frequently asked questions)를 관리한다. "누군가가 제 아이디로 접속을 시도했습니다. 해킹이 아닐까요"라는 질문에 대해서 "비밀번호를 변경한 후에 수사기관에 의뢰해야 합니다."와 같은 답변을 나타내는 항목을 관리해야 한다.

[0006] 또 다른 예를 들면, 시스템을 구축하기 위해서는 현실 세계의 개체를 분석하여 데이터베이스(database)의 논리적 구조를 모델링(modeling) 한다. 즉 개체(entity)를 나타내는 테이블(table)의 명칭과 개체(entity)의 속성(attribute)을 나타내는 칼럼(column)의 명칭을 나타내는 항목을 관리해야 한다. 대규모의 시스템은 테이블만 수만 개에 달하는 경우도 있다.

[0007] 이처럼 특정 정보를 나타내는 항목(Terminology)을 관리하기 위해서, 종래에는 동의/유의어 사전(dictionary)을 이용하였다. 즉 사람이 미리 A = B와 같이 제1 항목과 제2 항목이 같은 항목임을 사전(dictionary)에 등록하고, 이를 이용하여 이음 동의어를 찾아내는 방식을 이용하였다.

[0008] 다만 이와 같은 방식으로, 계속해서 생성되고 있는 신조어에서 이음 동의어를 선별해 내기에는 한계가 있다. 또한, 시스템의 규모가 커지고 복잡도가 증가함에 따라 관리 대상이 되는 항목의 수도 기하급수적으로 증가하고 있다. 이러한 상황에서는 새로운 항목이 생성될 때마다 사람이 인위적으로 개입해서 이음 동의어 항목을 관리하기란 불가능에 가깝다.

[0009] 이에 사람의 인위적인 개입 없이도 신조어로 생성되는 항목의 이음 동의어를 선별하고, 관리해야 하는 항목의 수가 많은 경우에도 자동으로 이음 동의어를 선별할 방법이 요구된다.

발명의 내용

해결하려는 과제

[0010] 본 발명이 해결하고자 하는 기술적 과제는 최소 의미 단위인 단어의 동의/유의 사전을 바탕으로 단어가 결합한 용어, 문장, 문서의 유사도를 자동으로 연산하는 방법 및 그 방법을 수행하는 장치를 제공하는 것이다.

[0011] 본 발명이 해결하고자 하는 다른 기술적 과제는 용어, 문장, 문서의 유사도를 연산하여 사용자에게 다른 용어, 다른 문장, 다른 문서를 추천하는 방법 및 그 방법을 수행하는 장치를 제공하는 것이다.

[0012] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0013] 상기 기술적 과제를 해결하고자 하는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법은, 유사도 분석 장치가, 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계와 상기 유사도 분석 장치가, 제2 항목(Target)에서 상기 제2 항목의 하위 항목인 제2-1 항목 내지 제2-n 항목을 추출하는 단계와 상기 유사도 분석 장치가, 상기 제1-1 항목 내지 제1-m 항목의, 상기 제2 항목의 하위 항목에 대한 유사도를 이용하여, S-T 유사도를 연산하는 단계와 상기 유사도 분석 장치가, 상기 제2-1 항목 내지 제2-n 항목의, 상기 제1 항목의 하위 항목에 대한 유사도를 이용하여, T-S 유사도를 연산하는 단계 및 상기 유사도 분석 장치가, 상기 S-T 유사도와 상기 T-S 유사도를 이용하여, 상기 제1 항목과 상기 제2 항목의 유사도를 연산하는 단계를 포함하되, 상기 S-T 유사도는 분석 항목(Source)을 구성하는 하위 항목이 대상 항목(Target)에 얼마나 포함되어 있는지를 기준으로 구한 유사도이고, 상기 T-S 유사도는 대상 항목(Target)을 구성하는 하위 항목이 분석 항목(Source)에 얼마나 포함되어 있는지를 기준으로 구한 유사도이다.

[0014] 일 실시 예에서, 상기 제1 항목과 상기 제2 항목의 유사도를 동의/유의 사전에 저장하는 단계를 더 포함할 수 있다.

[0015] 다른 실시 예에서, 상기 제1 항목과 상기 제2 항목의 유사도가 기 설정된 값 이상인 경우, 상기 제1 항목을 사용하려는 사용자에게, 상기 제1 항목 대신 상기 제2 항목과 상기 유사도를 제공하는 단계를 더 포함할 수 있다.

[0016] 또 다른 실시 예에서, 상기 제1 항목과 상기 제2 항목의 유사도가 기 설정된 값 이상인 경우, 상기 제1 항목이 상기 제2 항목의 표절인 것으로 판단하는 단계를 더 포함할 수 있다.

[0017] 또 다른 실시 예에서, 상기 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계는, 상기 제1 항목의 어미 또는 조사를 제거하는 단계를 포함할 수 있다.

[0018] 다른 실시 예에서, 상기 제1 항목(Source)에서 상기 제1 항목의 하위 항목인 제1-1 항목 내지 제1-m 항목을 추출하는 단계는, 상기 제1-1 항목 내지 제1-m 항목 중에서 임의의 두 항목을 선택하여, 상기 두 항목의 유사도가 기 설정된 값 이상인 경우에는 상기 두 항목 중에서 어느 한 항목을 제외하는 단계를 포함할 수 있다.

[0019] 또 다른 실시 예에서, 상기 제1항목과 상기 제2 항목은 문서이고, 상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 문장이고, 상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는, 상기 문서에서 마침표를 기준으로 상기 문장을 추출하는 단계를 포함할 수 있다.

[0020] 또 다른 실시 예에서, 상기 제1항목과 상기 제2 항목은 문장이고, 상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 용어이고, 상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는, 상기 문장에서 띄어 쓰기와 어미와 조사를 기준으로 상기 용어를 추출하는 단계를 포함할 수 있다.

[0021] 또 다른 실시 예에서, 상기 제1항목과 상기 제2 항목은 용어이고, 상기 제1-1 항목 내지 제1-m 항목과 상기 제2-1 항목 내지 제2-n 항목은 단어이고, 상기 제1-1 항목 내지 제1-m 항목을 추출하는 단계 또는 상기 제2-1 항목 내지 제2-n 항목을 추출하는 단계는, 상기 용어에서 형태소를 기준으로 의미를 가진 최소의 단위인 상기 단어를 추출하는 단계를 포함할 수 있다.

[0022] 또 다른 실시 예에서, 상기 제1-1 항목 내지 제1-m 항목의, 상기 제2 항목의 하위 항목에 대한 유사도를 이용하여, S-T 유사도를 연산하는 단계는, 상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 상기 제2 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계 및 상기 조회 결과로 얻은, 상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 유사도의 값을 평균하여, 상기 S-T 유사도를 연산하는 단계를 포함할 수 있다.

[0023] 또 다른 실시 예에서, 상기 제1-1 항목 내지 제1-m 항목에 속한 각 항목의 상기 제2 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계는, 상기 제1-1 항목 내지 제1-m 항목 중에서 특정 항목의 상기 제2 항목의 하위 항목에 대한 유사도가 상기 동의/유의 사전에 없는 경우, 상기 특정 항목의 하위 항목인 제3 항목을 추출하는 단계 및 상기 제3항 항목의 상기 제2 항목의 하위 항목의 하위 항목에 대한 유사도를 상기 동의/유의 사전에 조회하는 단계를 포함할 수 있다.

[0024] 또 다른 실시 예에서, 상기 제2-1 항목 내지 제2-n 항목의, 상기 제1 항목의 하위 항목에 대한 유사도를 이용하여, T-S 유사도를 연산하는 단계는, 상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 상기 제1 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계 및 상기 조회 결과로 얻은, 상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 유사도의 값을 평균하여, 상기 T-S 유사도를 연산하는 단계를 포함할 수 있다.

[0025] 또 다른 실시 예에서, 상기 제2-1 항목 내지 제2-n 항목에 속한 각 항목의 상기 제1 항목의 하위 항목에 대한 유사도를 동의/유의 사전에서 조회하는 단계는, 상기 제2-1 항목 내지 제2-n 항목 중에서 특정 항목의 상기 제1 항목의 하위 항목에 대한 유사도가 상기 동의/유의 사전에 없는 경우, 상기 특정 항목의 하위 항목인 제4 항목을 추출하는 단계 및 상기 제4항 항목의 상기 제1 항목의 하위 항목의 하위 항목에 대한 유사도를 상기 동의/유의 사전에 조회하는 단계를 포함할 수 있다.

[0026] 또 다른 실시 예에서, 상기 S-T 유사도와 상기 T-S 유사도를 이용하여, 상기 제1 항목과 상기 제2 항목의 유사도를 연산하는 단계는, 상기 S-T 유사도와 상기 T-S 유사도의 최소값(min), 최대값(max), 평균값(avg) 중에서 어느 한 값을 상기 제1 항목과 상기 제2 항목의 유사도로 연산하는 단계를 포함할 수 있다.

발명의 효과

[0027] 본 발명의 실시 예에 따른 효과는 다음과 같다.

[0028] 종래에는 사람이 미처 인식하지 못하거나 동의/유의 사전에 미리 등록하지 못한 이음 동의어가 시스템에 중복해서 등록되는 경우가 많았다. 실제로 금융, 제조 등 대규모 차 세대급 프로젝트를 수행하는 경우, 이음 동의어 성격의 정보 항목이 다수가 존재하여, 데이터 웨어하우스(DW; Data Warehouse) 시스템을 구축하거나 기간별 통계 정보를 생성하는 경우에 분석에 필요한 정보 항목을 찾아내는데 많은 시간과 비용이 소모되었다. 이로 인해 데이터 품질이 저하되는 악순환을 초래하게 되었다.

[0029] 이에 비해 본 발명에 따른 방법으로 정보 항목을 관리하는 경우, 의미의 최소 단위인 단어의 동의/유의 사전을 바탕으로 단어가 결합한 용어, 용어와 단어가 결합한 문장, 문장이 결합한 문서의 유사도를 자동으로 연산할 수 있다. 이를 통해 같은 의미의 용어, 같은 의미의 문장, 같은 의미의 문서를 선별하여 사용자에게 제공할 수 있다. 즉 동의/유의 사전(dictionary)에 새로운 용어, 새로운 문장, 새로운 문서가 등록되어 있지 않더라도 정보 항목의 유사도를 확인할 수 있다.

[0030] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0031] 도 1a 내지 도 1b는 종래의 항목 관리 방법과 본 발명의 일 실시 예에 따른 항목 관리 방법을 비교 설명하기 위한 예시도이다.

도 2는 본 발명의 몇몇 실시 예에서 사용되는 항목의 체계를 정의하기 위한 예시도이다.

도 3a 내지 도 3b는 본 발명의 몇몇 실시 예에서 사용되는 유사도를 정의하기 위한 예시도이다.

도 4a 내지 도 4b는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 전제가 되는 규칙들을 설명하기 위한 예시도이다.

도 5a 내지 도 5c는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 수식들을 설명하기 위한 예시도이다.

도 6 내지 도 7은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법을 설명하기 위한 예시도이다.

도 8은 본 발명의 일 실시 예에 따른 동의/유의 사전의 확장을 설명하기 위한 예시도이다.

도 9는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 순서도이다.

도 10은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 장치의 구성도이다.

도 11은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법을 설명하기 위한 예시도이다.

도 12a 내지 도 12b는 본 발명의 일 실시 예에 따른 상위 항목의 유사도를 연산하기 위해 하위 항목의 유사도를 이용하는 과정을 설명하기 위한 예시도이다.

도 13은 본 발명의 일 실시 예에 따른 전처리 과정을 설명하기 위한 예시도이다.

도 14a 내지 도 17b는 본 발명의 일 실시 예에 따른 항목 관리 방법을 설명하기 위한 구체적인 예시도이다.

도 18은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 장치의 하드웨어 구성도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시 예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 게시되는 실시 예에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시 예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0033] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않은 한 이상적으로 또는 과도하게 해석되지 않는다. 본 명세서에서 사용된 용어는 실시 예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다.

[0034] 명세서에서 사용되는 "포함한다 (comprises)" 및/또는 "포함하는 (comprising)"은 언급된 구성 요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성 요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0035] 이하, 본 발명에 대하여 첨부된 도면에 따라 더욱 상세히 설명한다.

[0036] 도 1a 내지 도 1b는 종래의 항목 관리 방법과 본 발명의 일 실시 예에 따른 항목 관리 방법을 비교 설명하기 위한 예시도이다.

[0037] 도 1a를 참고하면, 종래의 항목 관리 방법을 볼 수 있다. 대상 용어(Target Terminology)는 이미 기존에 등록한 항목이라 생각하면 된다. 예를 들면 기존에 생성한 테이블 명칭일 수 있다. 또는 기존에 등록한 안내 메시지일 수 있다. 동의어 사전은 사람이 미리 등록해 놓은 동의어의 목록이다. 도 1a를 참고하면 [사업부업무명]과 [사업부비즈니스명] 및 [디비전업무명]의 항목이 같은 의미를 가지는 것으로 사전(dictionary)에 등록되어 있다. 이하 항목을 표시하는 기호로 []를 사용하기로 한다.

[0038] 이러한 상황에서 사용자가 [디비전비즈니스명]이라는 새로운 항목을 생성한다고 가정해보자. 이때에는 새로운 항목을 등록해도 되는지 검증하는 과정이 필요하다. 즉 [디비전비즈니스명]이라는 항목을 분석 용어(Source Terminology)로 삼고, 해당 분석 용어를 새로 추가해도 될지 아니면 대신에 이미 기존에 등록한 대상 용어를 사용해야 할지 확인하는 과정이 필요하다. 여기서 분석 용어(Source Terminology)는 새로 등록하려는 항목이라 생각하면 된다.

[0039] 종래에는 사용자가 하나씩 확인하면서 새로운 용어를 항목으로 등록할지 결정했다. 도 1a의 예에서라면 [디비전비즈니스명]은 외래어로 기재된 항목으로 이를 사용자가 한글로 바꾸면 [사업부업무명]과 같은 형태로 변경이 가능할 것이다. 이때 [사업부업무명]은 대상 용어로 아직 등록되어 있지 않다. 즉, [사업부업무명]이라는 항목이 아직 생성되어 있지 않으므로, 사용자는 [디비전비즈니스명]이라는 항목을 생성하는 대신 [사업부업무명]이라는 항목을 생성할 수 있다.

[0040] 그러나 이 과정에서 새로운 항목을 생성하려는 사용자는 추가로 동의어 사전도 확인해 보아야 한다. 동의어 사전에는 [사업부업무명]의 동의어로 [사업부비즈니스명] 및 [디비전업무명]의 항목이 등록되어 있다. 여기서 [디비전업무명]은 대상 용어로 이미 등록이 되어 있으므로, 사용자는 최종적으로 [디비전비즈니스명]이라는 항목 대신에 [디비전업무명]을 사용하는 것으로 결론을 내릴 수 있다.

[0041] 이처럼 사용자의 수작업으로 각종 항목을 관리하다 보면, [디비전비즈니스명]을 [사업부업무명]으로 바꾸기 위한 인위적인 과정이 필요하고, 또한 [사업부업무명]을 다시 [디비전업무명]으로 바꾸기 위해서 동의어 사전을 확인하는 과정이 필요하다.

[0042] 하지만 사용자는 애초에 [디비전비즈니스명]을 등록하려던 사용자이기 때문에 이를 [사업부업무명]으로 바꾸기가 쉽지 않으며, 도 1a의 예시와 달리 동의어 사전에 등록된 항목이 수많은 경우에는 사용자가 동의어 사전을 일일이 확인하기도 쉽지가 않다.

[0043] 즉 사용자가 항목을 관리하게 되면 기존에 이미 대상 용어로 등록한 항목임에도 그와 같거나 유사한 의미의 항목을 또 등록하는 상황이 생길 수 있다. 이는 대상 용어나 동의어 사전을 참고하더라도, 사용자가 새로 등록하려는 [디비전비즈니스명]의 항목과 이음 동의어의 항목인 [디비전업무명]을 찾지 못하는 경우에 발생하게 된다.

[0044] 종래에는 사용자가 항목을 관리하는 방법 외에도 동의/유의어 사전을 이용하여 시스템에 의해서 자동으로 항목을 관리하는 방법을 사용하기도 한다. 그러나 시스템에 의해서 자동으로 항목을 관리하는 방법에 의하더라도 사용자가 새로 등록하려는 [디비전비즈니스명]이라는 항목이 동의어 사전에 등록되어 있지 않기 때문에, 시스템은 이음 동의어를 찾을 수 없다는 결론을 내리고 [디비전업무명]이라는 항목이 있음에도 불구하고 [디비전비즈니스명]이라는 항목을 생성하게 된다.

[0045] 이처럼 사람이 항목을 관리하는 방법이나 동의/유의어 사전을 이용하여 시스템이 항목을 관리하는 방법이나, 종래의 항목 관리 방법에 의하면 기존에 이미 등록한 이음 동의어를 찾지 못하고 새로운 항목을 다시 생성하는 경우가 비일비재하다.

[0046] 금융권 K 은행의 경우 시스템에 수십만 개의 용어가 등록되어 있다. 예를 들면, 통장을 개설하기 위해서 일반 사용자로부터 입력 받아야 하는 항목, 자동이체를 설정하기 위해서 일반 사용자로부터 입력 받아야 하는 항목, 예금이 만기가 된 경우 이를 받기 위해 일반 사용자로부터 입력 받아야 하는 항목 등이 다양하게 있을 수 있다. 그런데, 그 중에서 일부 항목은 같은 내용의 항목임에도 항목의 명칭이 다르게 표시되어 사용자의 혼선을 불러일으키는 경우가 있다.

[0047] 또는 정부의 행정 시스템에도 이음 동의어가 다수 등록된 경우가 있다. 예를 들면, 정부에서 사용하는 A 문서에서는 [사는 곳]이라고 표시하고, 또 다른 B 문서에서는 [거주지]라고 표시하고, 또 다른 C 문서에서는 [주소]라고 표시하기도 한다. 이러한 문서들을 바탕으로 시도별 인구 통계를 낸다고 가정해보자.

[0048] 이때, A 문서의 [사는 곳]을 사용해야 할지, B 문서의 [거주지]를 사용해야 할지, C 문서의 [주소]를 사용해야 할지 혼란스러울 수 있다. 이처럼 데이터 웨어하우스(DW; Data Warehouse) 시스템을 구축하거나 통계 정보를 생성하는 경우에, 정부에서 사용하는 문서의 각 항목이 의미하는 것이 무엇인지 파악하는 과정, 즉 메타 데이터(meta data)를 생성하는 과정에 상당한 비용과 시간이 소모될 수 있다.

[0049] 이처럼 관리해야 하는 항목이 수많은 경우, 될 수 있으면 새로운 항목(=분석 용어)을 새로 추가하지 않고, 기존의 이음 동의어의 항목(=대상 용어)을 재사용하도록 안내할 방법이 필요하다. 이를 위해서 기존 종래 방법의 문제점을 분석해보자.

[0050] 우선 기존 종래의 방법 중 첫 번째 방법인 사람이 항목을 관리하는 방법의 경우, 항목의 수가 적은 경우에는 나름 효율적이나 항목의 수가 기하급수적으로 증가하게 되면 그에 반비례해서 관리의 효율이 떨어지게 된다. 그러므로 이 방법은 개선의 여지가 없다.

[0051] 다음으로 기존 종래의 방법 중 두 번째 방법인 동의/유의어 사전을 이용하여 시스템에 의해서 자동으로 항목을 관리하는 방법의 경우, 자동으로 항목을 관리하므로 관리하는 항목의 수가 기하급수적으로 증가하더라도 시스템이 감당할 수 있다는 장점이 있다.

[0052] 여기서 항목의 수가 늘어나는 경우를 살펴보면, 대부분 하나의 단어로 이루어진 항목이 추가되는 경우보다 단어와 단어가 결합한 용어 형태의 항목이 추가되는 경우가 주를 이룬다. 즉 새로운 신조어가 자꾸 생기는 경우 그에 맞춰서 동의/유의어 사전을 업데이트(update)하지 않으면, 시스템에 의한 자동화된 방법이라고 하더라도 기존에 등록된 대상 용어 중에서 이음 동의어를 찾지 못하게 된다.

[0053] 하지만, 새롭게 등장하는 신조어의 동의/유의어를 찾아서 사전에 등록하는 작업은 사용자에 의해 진행되므로, 동의/유의어 사전을 이용한 관리 방법에도 한계가 있다. 예를 들어 10개의 단어 중에서 2개를 선택해서 만들 수 있는 새로운 용어의 수는 90개에 이른다. 이를 하나씩 사용자가 등록하기란 불가능에 가깝다.

[0054] 그러므로, 제안하고자 하는 방법은 단어와 단어가 결합해서 만들어진 신조어의 경우에도 이음 동의어를 찾아낼 방법이어야 한다. 그리고, 만약 단어와 단어가 결합한 신조어의 이음 동의어를 찾아낼 수 있다면, 이 방법을 확장해서 적용하여 단어와 단어가 결합한 용어뿐만 아니라, 용어와 단어 등이 결합한 문장, 나아가 문장과 문장이 결합한 문서의 유사도도 연산할 수 있다.

[0055] 예를 들면 복수의 뉴스 기사를 클러스터를 구성하여 중복된 뉴스를 제외하고, 다양한 뉴스를 사용자에게 제공하려고 할 때, 종래에는 단순히 키워드를 기반으로 뉴스의 유사도를 연산하여 클러스터링을 수행하였다. 하지만 이럴 때 유사한 내용의 문서임에도, TF-IDF (Term Frequency - Inverse Document Frequency) 등의 알고리즘을

적용하여 추출한 키워드가 서로 다른 경우에는 유사도가 낮게 결과가 나오게 되어 제대로 클러스터링이 수행되지 않는 단점이 있다.

[0056] 네이버의 출원 중에서 공개번호 10-2011-0117440 A 를 참고하면 두 논문의 유사도를 비교하기 위해 키워드를 추출하여 두 논문의 유사도를 연산하는 구성이 개시되어 있다. 그러나 이러한 방법은 단순히 해당 키워드만을 기준으로 유사도를 연산하기 때문에 유사도 연산이 불충분할 수 있다.

[0057] 물론 네이버의 출원은 이러한 문제점을 해결하기 위하여 해당 논문이 참조하고 있는 논문, 또는 해당 논문을 참조하고 있는 논문에서 추가로 키워드를 추출하여 키워드를 다양하게 선정함으로써 단점을 극복하고 있다. 이러한 선행 기술과 비교하면 본 발명에서 제안하고자 하는 방법은 문서 사이의 참조 관계가 없더라도 적용이 가능한, 단어의 동의어/유의어를 기준으로 문서의 유사도를 연산하는 방법이다.

[0058] 본 발명의 일 실시 예에 따른 이음 동의 항목 관리 방법의 구체적인 유사도 산출 방법은 다음에 보다 자세히 설명하도록 하고, 그 효과를 먼저 살펴보도록 하자. 본 발명의 일 실시 예에 따른 이음 동의 항목 관리 방법을 적용하면, 도 1b에서 보는 것과 같은 효과를 얻을 수 있다.

[0059] 즉 도 1a의 예시와 유사하게 사용자가 새로운 항목인 분석 용어 [디비전비즈니스명]을 등록하려고 할 때, 분석 용어 [디비전비즈니스명]이 동의/유의 사전에 등록되어 있지 않더라도, 기존에 등록된 항목인 대상 용어와 분석 용어 사이의 유사도를 연산하여 사용자에게 제공할 수 있다.

[0060] 도 1b의 예에서는 [디비전비즈니스명] 항목과 [사업부영문명] 항목은 66.7%의 유사도를, [디비전업무명] 항목은 100%의 유사도를, [사업부한글명] 항목은 66.7%의 유사도를 가진다. 그러므로 사용자는 [디비전비즈니스명] 항목을 추가하는 대신 [디비전업무명] 항목을 사용할 수 있다.

[0061] 본 발명의 동의/유의 사전에는 분석 용어 [디비전비즈니스명]이 등록되어 있지는 않음에도, 해당 분석 용어 [디비전비즈니스명]를 구성하고 있는 단어인 [디비전] 항목이나 [비즈니스] 항목이 동의/유의 사전에 (디비전, 사업부, 100%) 및 (업무, 비즈니스, 100%)의 형태로 등록되어 있기 때문에, 이들 단어의 조합으로 만들어진 새로운 용어의 유사도도 연산이 가능한 것이다.

[0062] 도 2는 본 발명의 몇몇 실시 예에서 사용되는 항목의 체계를 정의하기 위한 예시도이다.

[0063] 본 발명에서 사용하는 항목이란, 시스템을 통해 관리하고자 하는 데이터이다. 이는 앞서 설명한 것처럼 핵심 성과 지표가 항목이 될 수도 있고, 사용자를 위한 안내 메시지나 FAQ가 항목이 될 수도 있다. 또는 데이터베이스를 구성하는 테이블과 칼럼의 명칭이 항목이 될 수도 있다. 또는 논문이나 뉴스, 웹 페이지와 같은 문서나 공개공보 및 특허공보와 같은 특허 문서도 항목이 될 수 있다.

[0064] 본 발명에서는 이러한 항목들이 일정한 체계를 가지는 것으로 정의한다. 항목의 가장 작은 단위는 단어(111)이다. 단어(111)는 의미를 가지는 최소 단위이며, 단어(111)를 더 분해하면 그 의미는 사라지게 된다. 단어(111)는 마치 화학에서 원소에 대응되는 개념이라고 보면 충분하다.

[0065] 단어(111)와 단어(111)가 뭉치면 이는 용어(113)가 된다. 즉 용어(113)는 최소 2개 이상 단어(111)의 결합으로 이루어진다. 용어(113)는 마치 화학에서 원소와 원소가 결합한 분자에 대응되는 개념이라고 보면 충분하다.

[0066] 용어(113)나 단어(111)가 더 뭉치면 이는 문장(115)이 된다. 즉 문장(115)은 최소 2개 이상의 단어(111)나 용어(113)의 결합으로 이루어진다. 또한, 문장은 마침표라는 기호를 통해서도 구분할 수 있다.

[0067] 문장(115)과 문장(115)이 더 뭉치면 이는 문서(117)가 된다. 즉 문서(117)는 최소 2개 이상 문장(115)의 결합으로 이루어진다. 문장이나 문서는 마치 화학에서 고분자 화합물에 대응되는 개념이라고 보면 충분하다.

[0068] 단어(111), 용어(113), 문장(115), 문서(117)는 뒤로 갈수록 그 크기가 큰 상위 항목에 해당한다. 즉 단어(111)가 가장 단위의 하위 항목이며 문서(117)가 가장 큰 단위의 상위 항목이다. 문서(117)에 가까울수록 상위 항목, 단어(111)에 가까울수록 하위 항목으로 정의한다.

[0069] 물론 도 2에 도시된 항목의 체계는 발명의 이해를 돕고자 하는 일종의 예시일 뿐 발명을 제한하기 위한 것은 아니다. 예를 들면 문장(115)과 문장(115)이 모여서 단락(미도시)을 이루고, 다시 단락(미도시)과 단락(미도시)이 모여서 문서(117)를 이룰 수도 있다.

[0070] 도 2에 도시된 항목의 체계는 일종의 예시이나, 이하 발명에 대한 설명을 진행할 때에는 단어(111)-용어(113)-문장(115)-문서(117)의 체계를 기준으로 설명을 계속해 나가기로 한다.

[0071] 앞서 예시한 항목의 일 예와 도 2의 항목의 체계를 대응시켜보면 다음과 같이 비교할 수 있다. 데이터베이스의 테이블과 칼럼의 명칭과 같은 항목은 단어(111)와 용어(113)에 대응된다. 다음으로 핵심 성과 지표와 같은 항목은 용어(113)와 문장(115)에 대응된다. 다음으로 안내 메시지나 FAQ와 같은 항목은 문장(115)과 문서(117)에 대응된다. 마지막으로 논문, 뉴스, 웹 페이지, 특허 문서와 같은 항목은 문서(117)에 해당한다.

[0072] 우리가 시스템을 통해서 관리하고자 하는 항목은 단어(111)와 같은 하위 항목에서부터 문서(117)와 같은 상위 항목까지 다양하게 있다. 이처럼 하위 항목부터 상위 항목까지 다양한 데이터들의 유사도를 연산하는 과정을 이후의 도면들을 통해서 살펴보도록 하자.

[0073] 도 3a 내지 도 3b는 본 발명의 몇몇 실시 예에서 사용되는 유사도를 정의하기 위한 예시도이다.

[0074] 도 3a는 단어의 유사도를 정의하기 위한 표이다. 도 3a를 참고하면 의미를 기반으로 한 단어의 유사도가 예시되어 있다. 이해의 편의를 돕기 위해 유사도는 2가지 종류로 정하였다. 같은 의미를 가지는 동의어와 같지는 않지만 비슷한 의미를 가지는 유의어이다.

[0075] 도 3a의 예를 참고하면, [성공]과 [성취]는 같은 의미를 가진다. 동의어의 경우 두 단어(111)의 유사도는 100%라고 가정한다. 다음으로 [성공]의 유의어로 [달성]. [출세], [입신]의 단어들이 있다. 유의어의 경우 두 단어(11)의 유사도는 50%라고 가정한다.

[0076] 다음으로 [실패]의 경우에는 동의어는 없으나, [실수]. [실책], [낭패]의 유의어를 가진다. 또한 [사업]의 경우에는 [비즈니스]의 동의어와 [업무]. [일], [영업]의 유의어를 가진다. 마지막으로 [입력]의 경우에는 [등록]의 동의어와 [추가], [생성]의 유의어를 가진다.

[0077] 도 3a에 예시된 것과 같이 단어(111)의 동의/유의 사전(dictionary)은 이미 생성된 것으로 가정한다. 단어(111)는 의미의 최소 단위이며 단어(111)와 단어(111) 사이의 유사도는 동의어인지 유의어인지에 따라 사전(dictionary)에 이미 저장이 되어 있다.

[0078] 물론 단어(111)와 단어(111) 사이의 유사도가 사전(dictionary)에 저장되어 있지 않은 경우에 자동으로 단어(111) 사이의 유사도를 연산하여 저장하는 방법도 있다. 그러나, 이는 다음에 보다 자세히 설명하도록 한다. 일단 현재로써는 사전(dictionary)에 동의어인지 유의어인지에 따라 단어(111) 사이의 유사도는 이미 저장된 것을 전제로 설명을 계속해 나가기로 한다.

[0079] 도 3b를 참고하면 단어(111)보다 상위 항목인 용어(113), 문장(115), 문서(117)가 도시된 것을 볼 수 있다. 의미 기반으로 용어(113) 사이의 유사도를 연산하여 용어(113)의 유사도를 구하고, 문장(115) 사이의 유사도를 연산하여 문장(115)의 유사도를 구하고, 문서(117) 사이의 유사도를 연산하여 문서(117)의 유사도를 구한다.

[0080] 이때, 도 3a에 도시한 것과 같이 단어(111)의 유사도는 사전(dictionary)에 저장되어 있어 유사도를 쉽게 구할 수 있으나, 용어(113)와 용어(113)를 비교한 용어(113)의 유사도나 문장(115)과 문장(115)을 비교한 문장(115)의 유사도, 문서(117)와 문서(117)를 비교한 문서(117)의 유사도는 사전(dictionary)에 저장되어 있지 않은 경우가 많을 것이다.

[0081] 예를 들면, 단어(111)와 단어(111)를 결합하여 새로운 신조어인 용어(113)를 만들 때, 해당 용어(113)는 동의/유의 사전(dictionary)에 저장되어 있지 않고, 해당 용어(113)를 구성하는 단어(111)들만 동의/유의 사전(dictionary)에 저장된 경우가 일반적이다.

[0082] 본 발명은 이럴 때 해당 용어(113)를 구성하는 단어(111)들의 유사도를 이용하여 용어(113)의 유사도를 연산하는 방법이다. 용어(113)를 구성하는 단어(111)들의 유사도를 이용하여 용어(113)의 유사도를 연산하기 위해서는 몇 가지 전제가 필요하다.

[0083] 도 4a 내지 도 4b는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 전제가 되는 규칙들을 설명하기 위한 예시도이다.

[0084] 도 4a를 참고하면 첫 번째 규칙(Rule 1)을 확인할 수 있다. 첫 번째 규칙은 용어, 문장, 문서에서 의미를 가지는 것은 명사, 부사, 형용사, 동사와 같은 품사들이며, 조사나 어미는 의미에는 영향이 없다는 가정이다. 그러므로 용어의 유사도를 연산할 때, 문장의 유사도를 연산할 때, 문서의 유사도를 연산할 때 되도록 조사와 어미는 제거하도록 한다.

[0085] 또한, 비교의 편의를 위해서 명사형으로 변경하는 것을 원칙으로 한다. 다만, 명사형이 아니더라도 어미를 제외하고 어근 형태로만 단어를 추출해서 비교하는 것도 가능하다. 즉 도 4a에 예시된 것처럼 [들어가셨습니다.]를

비교하기 위해서는 [들어감]과 같은 명사형으로 변형하거나 [들어]와 같은 어근 형태로 변형하는 것이 바람직하다.

[0086] 도 4a를 참고하면 [목표매출액]이라는 용어의 유사도를 연산하기 위해서 [목표], [매출], [액]이라는 각각의 단어로 분해한 것을 볼 수 있다. 또한 [아버지가 방에 들어가셨습니다.]이라는 문장의 유사도를 연산하기 위해서 [아버지], [방], [들어감] (또는 [들어])로 분해한 것을 볼 수 있다.

[0087] 용어나 문장 및 문서와 같은 상위 항목들은 그 유사도를 연산하기 위해서 하위 항목으로 분해하는 과정이 필요하다. 이때에는 조사와 어미를 제거하는 전처리 과정과 동사의 경우에는 명사형으로 변형하거나 어근 형태로 변형하는 전처리 과정을 거치게 된다.

[0088] 다음으로 도 4b를 참고하면 두 번째 규칙(Rule 2)을 확인할 수 있다. 두 번째 규칙은 순서는 의미에 영향을 미치지 않는다는 가정이다. 도 4b에 예시된 것처럼 [목표매출액]이라는 용어와 [매출목표액]이라는 용어는 그 의미가 같다. 또한 [아버지가 방에 들어가셨습니다.]라는 문장과 [방에 아버지가 들어가셨습니다.]라는 문장은 그 의미가 같다.

[0089] 물론 순서에 따라 그 뉘앙스(nuance)가 미묘하게 달라지는 경우도 있지만, 대부분 경우에는 의미에 큰 차이가 없다. 이처럼 순서가 바뀌더라도 그 의미는 대부분 같으므로 유사도를 연산할 때 단어의 순서, 용어의 순서, 문장의 순서는 고려하지 않는 것으로 한다.

[0090] 순서를 반영해서 유사도를 연산하여 얻을 수 있는 정확한 유사도의 연산이라는 이득이, 순서를 반영하여 유사도를 연산하기 위해 추가되는 알고리즘의 복잡도라는 손실보다 크지 않다. 그러므로 유사도를 연산할 때 순서는 무시하여 더욱 빠른 연산이 가능하도록 한다.

[0091] 본 발명에서는 유사도를 연산할 때, 도 4b의 첫번째 문장에서 [아버지]라는 단어와 두번째 문장에서 [방], [아버지], [들어감]을 비교하고, 그 중에서 가장 유사도가 높은 단어를 첫번째 문장의 [아버지] 단어의 유사도로 사용한다. 그러므로 단어의 순서가 바뀌더라도 순서를 유사도에 반영하지 않는 이상, 유사도가 가장 높은 단어는 단어의 순서와 무관하다. 이에 본 발명에서는 항목들 사이의 순서는 무시하기로 한다.

[0092] 도 4a 내지 도 4b에서 살펴본 두 가지 전제 아래, 특정 항목 사이의 유사도를 연산하기 위한 구체적인 수식을 도 5a 내지 도 5b를 통해서 살펴보도록 하자.

[0093] 도 5a 내지 도 5c는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 수식들을 설명하기 위한 예시도이다.

[0094] 도 5a를 참고하면 수식 1(Equation 1)을 확인할 수 있다. 수식 1은 "유사도는 자기 자신을 100%로 봤을 때, 자신과 비교 대상과의 유사도를 0%~100%로 표시한 값"이다. 즉 비교하고자 하는 두 항목이 같다면 그 둘 사이의 유사도는 100%이다. 이는 당연한 수식이라고 할 수 있다.

[0095] 그리고 만약 비교하고자 하는 두 항목이 다르다면 유사도를 연산하여 0%부터 100%의 값으로 표시할 수 있다. 앞서 도 3a에서 단어의 동의어와 유의어를 설명한 예처럼, 두 단어가 다른 경우 동의어는 유사도를 100%라고 볼 수 있고 유의어는 유사도를 50%라고 볼 수 있다.

[0096] 물론 유의어라고 하더라도 의미가 유사한 정도의 차이가 있을 수 있으므로 실제로 유사도의 값은 50%가 아닌 다른 값을 가질 수 있다. 이에 관해서는 다음에 더욱 자세히 설명하기로 하고, 지금은 이해의 편의를 돕기 위해 동의어는 100%, 유의어는 50%의 유사도를 가진다고 가정한다.

[0097] 비교하고자 하는 분석 항목(source item)이 A이고 나머지 하나 다른 대상 항목(target item)이 A인 경우에는 식 1에 의해서 유사도가 100%이다. 그러나, 비교하고자 하는 분석 항목(source item)이 A이고 나머지 하나 다른 대상 항목(target item)이 B인 경우에는 A 항목과 B 항목 사이의 유사도 연산이 필요하다. 이때 식 2(Equation 2)와 식 3(Equation 3)이 사용될 수 있다.

[0098] 도 5b를 참고하면, 식 2(Equation 2)는 항목 A와 항목 B 사이의 유사도를 연산하기 위한 기준으로 2가지를 제시하고 있다.

[0099] 하나는 분석 항목(source item)인 항목 A를 기준으로 대상 항목(target item)인 항목 B와 비교한 결과로 이를 S-T 유사도로 정의한다. S-T 유사도는 분석 항목인 A를 구성하는 단어가 대상 항목인 B에 얼마나 포함되어 있는지를 기준으로 구하는 유사도이다.

[0100] 다른 하나는 대상 항목(target item)인 항목 B를 기준으로 분석 항목(source item)인 항목 A와 비교한 결과로 이를 T-S 유사도로 정의한다. T-S 유사도는 대상 항목인 B를 구성하는 단어가 분석 항목인 A에 얼마나 포함되어 있는지를 기준으로 구하는 유사도이다.

[0101] S-T 유사도와 T-S 유사도를 구하는 방법은 도 6 내지 도 7을 통해서 구체적인 예와 함께 살펴보기로 한다. 분석 항목 A와 대상 항목 B를 이용하여 S-T 유사도와 T-S 유사도를 구한 후에는 이 두 가지 유사도를 이용하여 A 항목과 B 항목 사이의 유사도를 구할 수 있다.

[0102] 도 5c를 참고하면, S-T 유사도와 T-S 유사도를 이용하여 A와 B 사이의 유사도를 구하는 식 3(Equation 3)이 예시되어 있다. 도 5c의 식 3(Equation 3)을 참고하면 A 항목과 B 항목 사이의 유사도는 S-T 유사도와 T-S 유사도의 최소값(min) 또는 최대값(max) 또는 평균값(avg)을 통해서 구할 수 있다.

[0103] 다만, 도 5c의 식 3(Equation 3)은 일종의 예시일 뿐 발명을 한정하고자 함은 아니다. 어떤 두 값을 가지는 수가 있고, 이 두 수를 연산하여 하나의 수를 만드는 방법이라면 무엇이든지 식 3(Equation 3)에 포함될 수 있다. 간단한 예를 들면, 두 수를 곱하거나 더하는 경우도 있다.

[0104] 식 2(Equation 2)를 통해서 S-T 유사도와 T-S 유사도를 구한 후, 식 3(Equation 3)을 이용하여 분석 항목(source item)과 대상 항목(target item) 사이의 유사도를 연산한다.

[0105] 즉 동의/유의 사전에 A 항목과 B 항목 사이의 유사도가 등록되어 있지 않은 경우에는 식 2와 식 3과 같이 A 항목을 구성하는 단어와 B 항목을 구성하는 단어의 유사도를 이용하여 A 항목과 B 항목의 유사도를 연산할 수 있다. 이를 확장하면 가장 작은 단위인 단어(111)의 유사도를 이용하여 용어(113)의 유사도를 구할 수 있고, 나아가 문장(115)의 유사도와 문서(117)의 유사도를 구하는 것도 가능하다.

[0106] 이를 통해 새로운 용어가 신조어로 등장해도, 새로운 문장이 등장해도 유사도의 연산이 가능하다는 장점이 있다. 물론, 본 발명에 의하더라도 최소한 단어(111) 사이의 유사도는 등록되어 있어야 한다는 전제가 필요하다. 지금은 단어(111) 사이의 유사도는 이미 등록되어 있는 것으로 가정하고, 단어(111) 사이의 유사도를 자동으로 등록하는 방법은 다음에 자세히 설명하도록 한다.

[0107] 도 6 내지 도 7은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법을 설명하기 위한 예시도이다.

[0108] 도 6을 참고하면 새로 등록하고자 하는 항목은 [사업목표등록]이고, 기존에 이미 생성한 항목은 [업무목표입력]이다. 즉 좌측의 [사업목표등록]은 분석 용어(source terminology)이고 우측의 [업무목표입력]은 대상 용어(target terminology)이다.

[0109] 도 6의 중앙에는 동의/유의 사전이 예시되어 있으나, [사업목표등록]과 [업무목표입력]은 사전(dictionary)에 등재되어 있지 않다. 이 경우 종래의 항목 관리 방법은 두 용어가 서로 다른 것으로 판단하고, [사업목표등록]을 등록해도 무방하다고 판단을 할 것이다. 그러나 본 발명에서는 동의/유의 사전에 [사업목표등록] 항목이 없더라도 유사도를 연산할 수 있다.

[0110] [사업목표등록] 용어와 [업무목표입력] 용어 사이의 유사도를 연산하기 위해서 각각을 가장 작은 의미의 단위인 단어로 분해한다. [사업목표등록]의 분석 용어는 [사업], [목표], [등록]의 세 개의 단어로 분리할 수 있다. 마찬가지로 [업무목표입력]의 대상 용어도 [업무], [목표], [입력]의 세 개의 단어로 분리할 수 있다.

[0111] 다음으로 S-T 유사도를 구해보면, 분석 용어의 [사업]은 동의/유의 사전에 대상 용어의 [업무]라는 단어와 유사도가 50%로 등록되어 있다. 즉 [사업]과 [업무]의 유사도는 50%로 두 단어는 유의어에 해당한다. 다음으로 분석 용어의 [목표]는 대상 용어의 [목표]와 같다. 이 경우에는 식 1에 의해서 유사도가 100%이다. 마지막으로 분석 용어의 [등록]은 동의/유의 사전에 대상 용어의 [입력]이라는 단어와 유사도가 100%로 등록되어 있다. 즉 [등록]과 [입력]의 유사도는 100%로 두 단어는 동의어에 해당한다.

[0112] S-T 유사도는 분석 용어(source terminology)를 구성하는 단어를 기준으로 대상 용어(target terminology)를 구성하는 단어와의 유사도 평균을 통해서 구할 수 있다. 그러므로, 도 6의 예에서 S-T 유사도는 avg(사업-업무, 목표-목표, 등록-입력) = avg(50%, 100%, 100%) = 83.3% 수식을 통해 83.3%의 값을 얻을 수 있다.

[0113] 마찬가지의 방법으로 T-S 유사도를 구하면, T-S 유사도는 avg(업무-사업, 목표-목표, 입력-등록) = avg(50%, 100%, 100%) = 83.3% 수식을 통해 83.3%의 값을 얻을 수 있다.

[0114] S-T 유사도와 T-S 유사도를 구한 후에는 이 두 값을 이용하여 [사업목표등록]과 [업무목표등록]의 유사도를 구

할 수 있다. 앞서 식 3을 설명하면서 최소값, 최대값, 평균값을 활용할 수 있다고 예시하였다. 도 6의 경우에는 S-T 유사도가 83.3%로, T-S 유사도도 83.3%로 값이 같으므로 최소값, 최대값, 평균값 모두 83.3%의 값을 가진다.

[0115] 도 6에서 볼 수 있듯이 [사업목표등록]과 [업무목표입력] 두 항목이 동의/유의 사전에 등록되어 있지 않더라도, 용어를 단어로 분리하고 단어 사이의 유사도를 이용하여 용어 사이의 유사도를 구할 수 있다. 만약 용어 사이의 유사도가 이미 설정된 값 이상이라면, 새로운 항목을 추가할 것이 아니라 기존의 항목을 사용하도록 사용자에게 제안할 수 있다.

[0116] 이렇게 S-T 유사도와 T-S 유사도를 이용하여 최종적으로 유사도를 구하는 것은 유사도의 정확도를 높이기 위해서이다. 물론 용어와 같은 하위 항목은 도 6에서 볼 수 있듯이 S-T 유사도와 T-S 유사도의 값이 같을 수 있다. 그러나 상위 항목으로 갈 수록 항목에 포함된 단어나 용어의 수가 많아지게 되므로, S-T 유사도와 T-S 유사도의 값이 차이가 나는 경우가 많아지게 된다. S-T 유사도와 T-S 유사도의 값이 차이가 나는 경우는 도 7에서 살펴보도록 한다.

[0117] 다음으로 도 7을 참고하면 새로 등록하고자 하는 항목은 [사업목표등록]이고, 기존에 이미 생성한 항목은 [광고업무목표입력]이다. 즉 좌측의 [사업목표등록]은 분석 용어(source terminology)이고 우측의 [광고업무목표입력]은 대상 용어(target terminology)이다.

[0118] 도 7의 중앙에는 동의/유의 사전이 예시되어 있으나, 도 6과 마찬가지로 [사업목표등록]과 [광고업무목표입력]은 사전(dictionary)에 등재되어 있지 않다. 이 경우 종래의 항목 관리 방법은 두 용어가 서로 다른 것으로 판단하고, [사업목표등록]을 등록해도 무방하다고 판단을 할 것이다. 그러나 본 발명에서는 동의/유의 사전에 [사업목표등록] 항목이 없더라도 유사도를 연산할 수 있다.

[0119] [사업목표등록] 용어와 [광고업무목표입력] 용어 사이의 유사도를 연산하기 위해서 각각을 가장 작은 의미의 단위인 단어로 분해한다. [사업목표등록]의 분석 용어는 [사업], [목표], [등록]의 세 개의 단어로 분리할 수 있다. 마찬가지로 [광고업무목표입력]의 대상 용어도 [광고], [업무], [목표], [입력]의 네 개의 단어로 분리할 수 있다.

[0120] 다음으로 S-T 유사도를 구하면, S-T 유사도는 avg(사업-업무, 목표-목표, 등록-입력) = avg(50%, 100%, 100%) = 83.3% 수식을 통해 83.3%의 값을 얻을 수 있다. 이는 앞서 살펴본 도 6의 예와 같다.

[0121] 다만, T-S 유사도를 구하면, [광고]에 대응되는 단어가 분석 용어에는 없는 것을 볼 수 있다. 그러므로 T-S 유사도를 구하면, T-S 유사도는 avg(광고-X, 업무-사업, 목표-목표, 입력-등록) = avg(0%, 50%, 100%, 100%) = 62.5% 수식을 통해 62.5%의 값을 얻을 수 있다.

[0122] 도 7의 예는 도 6의 예와는 달리 S-T 유사도와 T-S 유사도의 값이 다르다. 이때 두 값을 이용하여 [사업목표등록]과 [광고업무목표등록]의 유사도를 구하면 최소값은 62.5%, 최대값은 83.3%, 평균값은 71.4%의 값을 가진다. 필요에 따라 [사업목표등록]과 [광고업무목표등록]의 유사도로 62.5% 또는 83.3% 또는 71.4%의 값을 이용할 수 있다.

[0123] 물론 최소값, 최대값, 평균값 외에 다양한 수식을 사용하여, S-T 유사도 83.3%와 T-S 유사도 62.5%의 값을 연산하여 새로운 유사도를 연산할 수도 있다. 그리고 이렇게 S-T 유사도와 T-S 유사도를 이용하여 연산한 유사도의 값은 다시 동의/유의 사전에 저장할 수 있다.

[0124] 도 6 예에서 구한 [사업목표등록] 항목과 [업무목표입력] 사이의 유사도와 도 7의 예에서 구한 [사업목표등록] 항목과 [광고업무목표입력] 사이의 유사도는 다음에 해당 용어들을 포함하는 상위 항목, 예를 들면 문장이나 문서의 유사도를 구할 때 활용될 수 있다.

[0125] 도 8은 본 발명의 일 실시 예에 따른 동의/유의 사전의 확장을 설명하기 위한 예시도이다.

[0126] 도 8의 상단에는 동의/유의 단어 사전이 예시되어 있다. 앞서 도 6과 도 7에서는 동의/유의 단어 사전을 이용하여 두 용어 사이의 유사도를 연산하였다. 이렇게 연산한 유사도는 다시 사전에 저장할 수 있다. 도 8을 참고하면, 동의/유의 단어 사전 아래에 동의/유의 용어 사전이 예시되어 있다.

[0127] 도 8에 중단에는 [사업목표등록] 용어와 [업무목표입력] 용어의 유사도가 83.3%로 등록이 되어 있고, [사업목표등록] 용어와 [광고업무목표입력] 용어의 유사도가 62.5%로 등록이 되어 있다. 다시 동의/유의 단어 사전과 동의/유의 용어 사전을 활용하면 동의/유의 문장 사전을 만들 수도 있다. 또한, 동의/유의 문서 사전도 만들 수

있다.

[0128] 예를 들어 특허 문서를 검색한다고 가정해보자. 검색하고자 하는 A 발명의 특징이 광고를 디스플레이(display)하는 장치라고 할 때, 사용자는 우선 "(정보 or 영상 or 비디오 or 광고 or information or video or adverti*)"를 포함하는 검색식을 사용해서 특허 문서를 검색한다. 다음으로 검색된 특허 문서를 하나하나 확인하면서 수작업으로 노이즈(noise)를 제외하고, A 발명과 유사한 특허 문서를 찾아야 한다.

[0129] 이에 비해 본 발명을 이용하면 A 발명의 명칭이나 A 발명의 명세서를 선택하면 자동으로 A 발명의 명칭에 포함된 단어의 동의어나 유의어를 많이 포함한 다른 특허 문서 또는 A 발명의 명세서에 포함된 단어의 동의어나 유의어를 많이 포함한 다른 특허 문서를 검색할 수 있다.

[0130] 사람이 수작업으로 검색하고자 하는 A 발명의 특징을 나타내는 단어의 동의어나 유의어를 포함하는 검색식을 별도로 작성하지 않더라도 동의/유의 사전을 이용하여 간편하게 비슷한 기술 분야의 특허 문서를 검색할 수 있다.

[0131] 마찬가지로 특정 논문을 선택하고 비슷한 내용의 논문을 자동으로 검색한다거나, 특정 뉴스를 선택하고 비슷한 내용의 뉴스를 자동으로 취합하여 클러스터를 구성할 수 있다. 단순히 키워드를 기반으로 문서의 유사도를 구하는 종래의 기술에 비해, 본 발명은 사전(dictionary)에 구축된 키워드의 동의어/유의어를 더 활용하여 문서의 유사도를 구하므로 더욱 더 정확하게 유사한 문서의 검색이 가능하다.

[0132] 도 9는 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법의 순서도이다.

[0133] 도 9를 참고하면, 분석하고자 하는 분석 항목(source item)을 그보다 더 작은 단위의 항목으로 나눈다(S1000). 또한, 대상 항목(target item)도 그보다 더 작은 단위의 항목으로 나눈다(S2000). 만약 분석 항목이 문서라면 문장 단위로 나누고, 분석 항목이 문장이라면 용어 단위로 나눈다. 또한, 분석 항목이 용어라면 단어 단위로 나눈다.

[0134] 하위 항목으로 나눈 후에는 전처리 과정을 거친다. 앞서 도 4a 내지 도 4b에서 설명한 것처럼 조사와 어미를 제거하는 전처리 과정과 동사를 명사형으로 변형하거나 어근으로 변형하는 전처리 과정을 거친다. 그뿐만 아니라 하위 항목을 대표 항목으로 치환하고(S3000), 중복되는 대표어를 제거하는 과정을 거칠 수 있다(S4000). 대표어로 치환하고 중복된 대표어를 제거하는 전처리 과정에 대해서는 도 13에서 더욱 자세히 설명하기로 한다.

[0135] 조사와 어미를 제거하는 전처리 과정이나, 동사를 변형하는 전처리 과정, 중복된 대표어를 제거하는 전처리 과정(S3000, S4000)은 필수적인 과정은 아니며 선택적인 과정이다. 다만, 조사와 어미를 제거하는 전처리 과정이나 동사를 변형하는 전처리 과정은 유사도 연산의 편의를 위해 수행하면 바람직하며, 중복된 대표어를 제거하는 전처리 과정은 유사도 연산의 정확도를 높이기 위해 수행하면 바람직한 전처리 과정이다.

[0136] 분석 항목과 대상 항목의 전처리 과정이 끝나면 더욱 정확한 유사도의 연산을 위해서 두 가지 기준의 유사도를 연산한다. 즉 S-T 유사도를 연산하고(S5100), 또한 T-S 유사도를 연산한다(S5500). S-T 유사도와 T-S 유사도를 이용하여 최종적으로 분석 항목과 대상 항목 사이의 유사도를 연산한다(S6000). 최종적으로 분석 항목과 대상 항목 사이의 유사도를 연산하는 과정에서는 최소값, 최대값, 평균 등의 함수를 사용할 수 있다.

[0137] 도 10은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 장치의 구성도이다.

[0138] 도 10을 참고하면 분석 항목으로 가장 큰 단위인 문서(117)가 하단에 예시되어 있다. 문서를 분석하여 유사도를 연산하기 위해서는 문장 추출부(215), 용어 추출부(213), 단어 추출부(211)가 필요하다.

[0139] 우선 문서에서 문장을 추출하여야 한다. 문장 추출부(215)는 문서에서 마침표를 기준으로 문장을 추출한다. 이렇게 추출한 문장은 분석 문장(115a)이 되어 대상 문서에서 추출한 대상 문장(115b)과의 유사도를 비교한다. 만약 동의/유의 사전(129)에 해당 분석 문장(115a)과 대상 문장(115b)의 유사도가 등록되어 있지 않은 경우에는 분석 문장(115a)과 대상 문장(115b)을 더 작은 단위의 항목으로 분리하여야 한다.

[0140] 용어 추출부(213)는 분석 문장(115a)에서 용어를 분리한다. 이때 어미/조사 사전(123)을 이용하여 전처리 과정을 수행할 수 있다. 또한, 띄어쓰기를 이용하여 문장에서 용어를 추출할 수 있다. 분석 문장(115a)에서 추출된 용어는 분석 용어(113a)가 되어, 대상 문장(115b)에서 추출된 대상 용어(113b)와의 유사도를 비교한다. 만약 동의/유의 사전(129)에 해당 분석 용어(113a)와 대상 용어(113b)의 유사도가 등록되어 있지 않은 경우에는 마찬가지로 분석 용어(113a)와 대상 용어(113b)를 더 작은 단위의 항목으로 분리하여야 한다.

[0141] 단어 추출부(211)는 분석 용어(113a)에서 단어를 분리한다. 이때 형태소 사전(121)을 이용할 수 있다. 분석 용어(113a)에서 추출된 단어는 분석 단어(111a)가 되어, 대상 용어(113b)에서 추출된 대상 단어(111b)와의 유사도

를 비교한다.

[0142] 동의/유의 사전(129)에는 단어의 유사도는 이미 등록되어 있다고 가정했으므로, 분석하고자 하는 문서(117)가 동의/유의 사전에 없더라도, 또는 분석 문장(115a)가 동의/유의 사전에 없더라도, 또는 분석 용어(113a)가 동의/유의 사전에 없더라도, 가장 작은 의미 단위인 단어까지 문서를 쪼개면 유사도의 연산이 가능하다.

[0143] 도 10을 참고하면 분석 문장(115a), 분석 용어(113a), 분석 단어(111a)는 분석 항목(110)에 해당하며, 유사도 분석부(220)는 이를 동의/유의 사전에 등재된 대상 문장(115b), 대상 용어(113b), 대상 단어(111b)와 비교하여 유사도를 연산할 수 있다.

[0144] 도 11은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 방법을 설명하기 위한 예시도이다.

[0145] 도 11을 참고하면, 분석 용어(Source Term)와 시스템에 이미 등록된 대상 용어(Target Term) 사이의 유사도를 연산하기 위해 동의/유의 사전에 분석 용어와 대상 용어가 있는지 확인한다. 만약 없는 경우라면 형태소 사전을 이용한 단어 추출부를 이용하여 분석 용어를 단어 단위로 분리하고 대상 용어도 단어 단위로 분리한다.

[0146] 다음으로 동의/유의 사전에 등록된 분석 단어와 대상 단어의 유사도를 기준으로 용어의 S-T 유사도와 T-S 유사도를 구한다. 이 과정에서 중복된 대표어를 제거하는 전처리 과정을 수행할 수 있다.

[0147] 중복된 대표어를 제거하기 위해서는 동의/유의 사전에 등록된 대표어를 이용하여야 한다. 중복된 대표어를 제거해야 하는 이유나 그 제거 과정에 대해서는 도 13에서 더욱 자세히 설명하도록 한다. 다만, 중복된 대표어를 제거하는 과정은 선택적으로 수행할 수 있는 과정이다.

[0148] S-T 유사도와 T-S 유사도를 구한 후에는 이 두 유사도를 이용하여 분석 용어와 대상 용어의 최종적인 유사도를 구한다. 다양한 함수를 이용하여 유사도를 구할 수 있는데, 도 11의 예에서는 최소값인 min 함수를 사용하였다. 도 11의 최하단에는 분석 용어와 대상 용어 사이의 유사도를 구한 표가 도시되어 있다.

[0149] 도 11을 참고하면, 분석 용어와 제1 대상 용어 사이의 S-T 유사도는 100%, T-S 유사도는 100%, 최종적으로 두 용어 사이의 유사도는 100%의 값이 연산 되었다. 마찬가지로 분석 용어와 제2 대상 용어 사이의 S-T 유사도는 66.7%, T-S 유사도는 66.7%, 최종적으로 두 용어 사이의 유사도는 66.7%의 값이 연산 되었다. 마찬가지로 분석 용어와 제3 대상 용어 사이의 S-T 유사도는 50%, T-S 유사도는 66.7%, 최종적으로 두 용어 사이의 유사도는 50%의 값이 연산 되었다.

[0150] 이렇게 유사도가 연산된 경우 시스템에서는 분석 용어와 그 의미가 같은 제1 대상 용어가 있으므로 새로 분석 용어를 등록하기보다, 기존에 등록된 제1 대상 용어(Target Term 1)를 사용할 것을 제안할 수 있다. 이를 통해 시스템에 이음 동의어가 다수 등록되는 것을 예방할 수 있다.

[0151] 도 12a 내지 도 12b는 본 발명의 일 실시 예에 따른 상위 항목의 유사도를 연산하기 위해 하위 항목의 유사도를 이용하는 과정을 설명하기 위한 예시도이다.

[0152] 도 12a는 도 11을 간략하게 표시한 도면이다. 분석 용어와 대상 용어의 유사도를 연산하기 위해서 동의/유의 사전을 참고하는데, 해당 사전에 분석 용어와 대상 용어가 등록되어 있지 않은 경우에는 용어 상태로는 유사도의 연산이 어려우므로, 단어 추출부를 이용하여 분석 용어를 구성하는 단어와 대상 용어를 구성하는 단어를 추출한다.

[0153] 다음으로 분석 용어를 구성하는 단어의 대상 용어를 구성하는 단어에 대한 유사도인 S-T 유사도를 구하고, 반대로 대상 용어를 구성하는 단어의 분석 용어를 구성하는 단어에 대한 유사도인 T-S 유사도를 구해서 이 둘을 이용하여 최종적으로 분석 용어와 대상 용어 사이의 유사도를 연산한다.

[0154] 도 12b를 참고하면 도 12a를 확장해서 용어와 용어 사이의 유사도가 아닌 문장과 문장 사이의 유사도를 구하는 과정이 도시되어 있다. 문장과 문장 사이의 유사도를 구하는 경우에도 마찬가지로 동의/유의 사전을 참고한다. 만약 동의/유의 사전에 분석 문장과 대상 문장이 등록되어 있지 않은 경우에는 문장 상태로는 유사도의 연산이 어려우므로, 용어 추출부를 이용하여 분석 문장을 구성하는 용어와 대상 문장을 구성하는 용어를 추출한다.

[0155] 만약 이렇게 추출된 분석 용어와 대상 용어도 동의/유의 사전에 등록되어 있지 않은 경우에는 단어 추출부를 이용하여 더 하위 항목인 분석 단어와 대상 단어를 추출한다. 최소 의미 단위인 단어의 경우에는 동의/유의 사전에 등록되어 있으므로 이를 이용하면 최종적으로 분석 문장과 대상 문장 사이의 유사도도 연산할 수 있다.

[0156] 또한, 도 12a 내지 도 12b에 도시된 것과 마찬가지의 과정을 통해서 분석 문서와 대상 문서 사이의 유사도도 연산할 수 있다. 즉 문서에서 문장을 추출하고, 문장에서 용어를 추출하고, 용어에서 단어를 추출하여 단어의 유사도를 이용하여 문서의 유사도를 구할 수 있다.

[0157] 도 13은 본 발명의 일 실시 예에 따른 전처리 과정을 설명하기 위한 예시도이다.

[0158] 도 13을 참고하면, 본 발명의 전처리 과정 중에서 대표어를 치환하여 중복된 대표어를 제거하는 전처리 과정을 확인할 수 있다. 단어나 용어와 같은 하위 항목의 경우 중복된 단어나 용어가 나타나는 경우가 거의 없으나, 문장이나 문서와 같은 상위 항목은 중복된 표현이 나타나는 경우가 많다.

[0159] 도 13을 참고하면, [목표매출액을 반드시 입력하여야 하며, 매출목표액을 입력하지 않은 경우 에러가 발생할 수 있습니다.]의 문장으로 된 항목이 예시되어 있다. 사용자에게 안내 메시지를 제공하기 위한 문장으로 보인다. 여기서 다른 문장과의 유사도를 연산하기 위해 하위 항목을 추출하는 경우 [목표매출액], [반드시], [입력], [매출목표액], [입력], [에러], [발생]과 같은 용어와 단어를 추출할 수 있다.

[0160] 이때 [목표매출액]과 [매출목표액]은 같은 용어는 아니나, 같은 의미를 가진 동의어이다. 만약 이 둘을 그대로 놓고 S-T 유사도를 구하게 되면 유사도가 중복으로 반영될 수 있다. 그러므로 이 둘 중의 하나는 제거를 하여 정확한 유사도를 연산할 수 있도록 하는 것이 대표어로 치환하여 중복된 대표어를 제거하는 전처리 과정이다.

[0161] 앞서 도 8의 예에서는 [사업목표등록] 항목과 [업무목표입력] 항목이 유사도 83.3%의 값으로 동의/유의 사전에 등록되어 있다. 마찬가지로 [목표매출액] 항목과 [매출목표액] 항목이 유사도 100%의 값으로 동의/유의 사전에 저장이 될 수 있다.

[0162] 이 경우 실제 데이터베이스에 동의/유의 사전을 테이블로 구성할 때에는 분석 항목을 나타내는 source_item 칼럼과 대상 항목을 나타내는 target_item 칼럼과 유사도를 나타내는 similarity_index 칼럼 등을 이용하여 유사도를 관리할 수 있다. 이 경우 두 항목 사이의 유사도가 100%일 때, 두 항목은 동의어에 해당하고, 이때 분석 항목을 대표어로 정의할 수 있다.

[0163] 예를 들어, 동의/유의 사전에 유사도를 관리하기 위한 테이블이 다음과 같은 (source_item, target_item, similarity_index) 칼럼을 가지고 있고, (목표매출액, 매출목표액, 100%)와 같은 로우(row)가 있다면, [매출목표액]의 대표어로 [목표매출액]을 선정할 수 있다.

[0164] 도 13의 예에서는 앞부분의 [목표매출액]의 항목과 뒷부분의 [매출목표액]의 대표어로 선정된 [목표매출액]의 항목이 중복되므로 뒷부분의 [매출목표액]의 항목을 제거할 수 있다. 마찬가지로 앞부분의 [입력]과 뒷부분의 [입력]을 중복제거 하여 하나의 [입력] 항목만 남겨둘 수 있다.

[0165] 이렇게 동의어의 경우 중복된 항목을 제거하면 최종적으로 [목표매출액], [반드시], [입력], [에러], [발생]의 항목만을 분석 항목(110)으로 하여 다른 문장과의 유사도를 연산할 수 있다.

[0166] 다만 이렇게 중복된 항목을 제거하는 전처리 과정은 어디까지나 선택적인 과정이다. 예를 들어, 문서의 경우 TF-IDF 알고리즘은 문서에서 특정 단어의 빈도를 기준으로 키워드를 선정한다. 이런 경우 다른 문서와의 유사도를 연산할 때 중복된 단어를 제거할 것이 아니라, 오히려 분석 문서와 대상 문서에서 해당 단어가 얼마나 등장하였는지를 나타내는 빈도를 기준으로 유사도에 가중치를 두어 연산하는 구성도 가능할 것이다.

[0167] 도 14a 내지 도 17b는 본 발명의 일 실시 예에 따른 항목 관리 방법을 설명하기 위한 구체적인 예시도이다.

[0168] 도 14a를 참고하면, [영문사업부명]의 분석 용어와 [사업부영문명]의 대상 용어 사이의 유사도를 연산하는 과정을 볼 수 있다. 동의/유의 사전에 두 용어가 없으므로 바로 유사도를 비교할 수는 없고 하위 항목인 단어 단위로 분리한 후에 유사도를 연산할 수 있다.

[0169] [영문사업부명]의 분석 용어는 [영문], [사업부], [명]의 분석 단어를 가지고, [사업부영문명]의 대상 용어는 [사업부], [영문], [명]의 대상 단어를 가진다. 순서의 차이만 있을 뿐, 단어가 같으므로 두 용어 사이의 유사도는 100%가 연산 결과로 나올 것이다. 즉 [영문사업부명]와 [사업부영문명]는 동의어에 해당한다.

[0170] 실제로 유사도를 연산해보면, S-T 유사도는 avg(영문-영문, 사업부-사업부, 명-명) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가지며, 마찬가지로 T-S 유사도는 avg(사업부-사업부, 영문-영문, 명-명) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. S-T 유사도와 T-S 유사도가 똑같이 100%의 값을 가지므로 min, max, avg 모두 100%의 값을 가진다.

[0171] 즉 [영문사업부명]와 [사업부영문명]의 유사도는 100%라는 값을 가지며, 그 결과를 동의/유의 사전에 추가할 수 있다. 또한, 사용자가 새로운 항목으로 [영문사업부명]을 등록하려는 경우, [영문사업부명]을 등록하는 대신 [사업부영문명]을 사용할 것을 제안할 수 있다.

[0172] 도 14b를 참고하면, [업무영문명]의 분석 용어와 [사업부영문명]의 대상 용어 사이의 유사도를 연산하는 과정을 볼 수 있다. 동의/유의 사전에 두 용어가 없으므로 바로 유사도를 비교할 수는 없고 하위 항목인 단어 단위로 분리한 후에 유사도를 연산할 수 있다.

[0173] [업무영문명]의 분석 용어는 [업무], [영문], [명]의 분석 단어를 가지고, [사업부영문명]의 대상 용어는 [사업부], [영문], [명]의 대상 단어를 가진다. [영문], [명]은 공통되지만, [업무]와 [사업부]의 차이가 있으므로, 단어 [업무]와 [사업부]의 유사도에 따라 용어의 유사도가 결정될 것이다.

[0174] 동의/유의 사전에 [업무]와 [사업부]의 유사도가 등록되어 있지 않다. 즉 두 단어 사이의 유사도는 0%이다. 이때 실제로 유사도를 연산해보면, S-T 유사도는 avg(업무-X, 영문-영문, 명-명) = avg(0%, 100%, 100%) = 66.7% 수식을 통해 66.7%의 값을 가지며, 마찬가지로 T-S 유사도는 avg(사업부-X, 영문-영문, 명-명) = avg(0%, 100%, 100%) = 66.7% 수식을 통해 66.7%의 값을 가진다. S-T 유사도와 T-S 유사도가 똑같이 66.7%의 값을 가지므로 min, max, avg 모두 66.7%의 값을 가진다.

[0175] 즉 [업무영문명]와 [사업부영문명]의 유사도는 66.7%라는 값을 가지며, 그 결과를 동의/유의 사전에 추가할 수 있다. 또한, 사용자가 새로운 항목으로 [업무영문명]을 등록하려는 경우, 시스템에 이미 등록된 용어 중에서 [사업부영문명]은 66.7%의 유사도를 가짐을 안내할 수 있다.

[0176] 도 14a 내지 도 14b를 통해서 용어의 유사도를 연산하는 경우를 살펴보았다. 다음으로 도 15a 내지 도 15b를 통해서 문장의 유사도를 연산하는 경우를 살펴보도록 하자.

[0177] 도 15a를 참고하면, [디비전영문명을 꼭 입력해야 합니다.]의 분석 문장과 [사업부영문명을 반드시 등록해야 합니다.]의 대상 문장 사이의 유사도를 연산하는 과정을 볼 수 있다. 동의/유의 사전에 두 문장이 없으므로 바로 유사도를 비교할 수는 없고 하위 항목인 용어와 단어 단위로 분리한 후에 유사도를 연산할 수 있다.

[0178] [디비전영문명을 꼭 입력해야 합니다.]의 분석 문장은 [디비전영문명]의 분석 용어와 [꼭], [입력]의 분석 단어를 가지고, [사업부영문명을 반드시 등록해야 합니다.]의 대상 문장은 [사업부영문명]의 대상 용어와 [반드시], [등록]의 대상 단어를 가진다.

[0179] 이때 [디비전영문명]과 [사업부영문명]의 용어의 유사도가 동의/유의 사전에 등록되어 있으므로 이 용어들을 더 하위 항목인 단어로 분리할 필요는 없다. 두 문장의 유사도는 [디비전영문명]과 [사업부영문명]의 용어의 유사도와 [꼭]과 [반드시]의 단어의 유사도 및 [입력]과 [등록]의 단어의 유사도에 의해 결정될 것이다.

[0180] 실제로 유사도를 연산해보면, S-T 유사도는 avg(디비전영문명-사업부영문명, 꼭-반드시, 입력-등록) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가지며, 마찬가지로 T-S 유사도는 avg(사업부영문명-디비전영문명, 반드시-꼭, 등록-입력) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. S-T 유사도와 T-S 유사도가 똑같이 100%의 값을 가지므로 min, max, avg 모두 100%의 값을 가진다.

[0181] 즉 [디비전영문명을 꼭 입력해야 합니다.]와 [사업부영문명을 반드시 등록해야 합니다.]의 유사도는 100%라는 값을 가지며, 그 결과를 동의/유의 사전에 추가할 수 있다. 또한, 사용자가 새로운 안내메시지를 나타내기 위한 항목으로 [디비전영문명을 꼭 입력해야 합니다.]을 등록하려는 경우, [디비전영문명을 꼭 입력해야 합니다.]을 등록하는 대신 [사업부영문명을 반드시 등록해야 합니다.]을 사용할 것을 제안할 수 있다. 이를 통해 통일적인 사용자 환경을 제공할 수 있다.

[0182] 다음으로 도 15b에서는 도 15a와 같은 분석 문장과 대상 문장을 비교하되 동의/유의 사전에 [디비전영문명]과 [사업부영문명]의 용어의 유사도가 등록되어 있지 않은 경우에 유사도를 연산하는 과정을 확인할 수 있다. 도 15b를 참고하면 도 15a와는 같은데, 동의/유의 사전에 [디비전영문명]과 [사업부영문명]의 용어의 유사도가 등록되어 있지 않고 대신 [디비전]과 [사업부]의 단어의 유사도가 등록된 것을 볼 수 있다.

[0183] 이때에는 [디비전영문명]과 [사업부영문명]의 용어의 유사도를 바로 연산할 수 없으므로 이 두 용어를 하위 항목인 단어로 분리하는 과정이 필요하다. [디비전영문명]의 용어는 [디비전], [영문], [명]의 하위 항목을 가지고, [사업부영문명]의 용어는 [사업부], [영문], [명]의 하위 항목을 가진다. 보다시피 [디비전]과 [사업부]의 단어의 유사도에 따라 [디비전영문명]과 [사업부영문명]의 용어의 유사도가 결정될 것이다.

[0184] 실제로 유사도를 연산해보면, S-T 유사도는 avg(avg(디비전-사업부, 영문-영문, 명-명), 꼭-반드시, 입력-등록) = avg(avg(100%, 100%, 100%), 100%, 100%) = 100% 수식을 통해 100%의 값을 가지며, 마찬가지로 T-S 유사도는 avg(avg(사업부-디비전, 영문-영문, 명-명), 반드시-꼭, 등록-입력) = avg(avg(100%, 100%, 100%), 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. S-T 유사도와 T-S 유사도가 똑같이 100%의 값을 가지므로 min, max, avg 모두 100%의 값을 가진다.

[0185] 도 15b의 예에서 볼 수 있듯이, [디비전영문명]과 [사업부영문명]의 용어의 유사도가 동의/유의 사전에 등록되어 있지 않더라도, 두 용어를 그 하위 항목의 단어로 한 번 더 분해해서 문장의 유사도를 연산할 수 있으며, 그 결과 또한 도 15a의 경우와 같은 값을 얻을 수 있다.

[0186] 도 16a 내지 도 16b는 분석 문서와 대상 문서 사이의 유사도를 연산하기 위한 예시이다. 문서에 포함된 문장이 단 세 문장이어서 문서보다는 단락에 가깝지만 여러 개의 문장이 포함된 항목의 경우에도 유사도의 연산이 가능함을 도 16a 내지 도 16b를 통해서 참고하도록 하자. 지면의 한계상 하나의 도면을 도 16a와 도 16b에 나누어서 그렸으며 도 16b는 도 16a에 이어지는 도면이다.

[0187] 도 16a를 참고하면 분석 문서는 [디비전영문명은 필수 입력 항목입니다. 따라서 디비전영문명을 꼭 입력해야 합니다. 그렇지 않은 경우 모두 무효 처리 될 수 있습니다.]의 3개의 문장으로 이루어져 있다. 마찬가지로 대상 문서는 [사업부영문명은 필수 입력 항목입니다. 사업부영문명을 반드시 등록하세요. 그렇지 않은 경우 무효 처리됩니다.]의 3개의 문장으로 이루어져 있다.

[0188] 문서의 유사도를 연산하기 위해 문서를 마침표를 기준으로 문장으로 분리하고 각 문장의 용어와 단어를 다시 추출하면 다음과 갔다. 우선 분석 문서는 [디비전영문명은 필수 입력 항목입니다.]의 문장 1에서 [디비전영문명], [필수], [입력], [항목]의 항목들을 추출할 수 있다. 또한, 분석 문서는 [따라서 디비전영문명을 꼭 입력해야 합니다.]의 문장 2에서 [디비전영문명], [꼭], [입력]의 항목들을 추출할 수 있다. 마지막으로 분석 문서는 [그렇지 않은 경우 모두 무효 처리 될 수 있습니다.]의 문장 3에서 [모두], [무효], [처리]의 항목들을 추출할 수 있다.

[0189] 마찬가지로 대상 문서에서 문장을 추출하고 이를 다시 용어와 단어로 분리하면 다음과 같다. 우선 대상 문서는 [사업부영문명은 필수 입력 항목입니다.]의 문장 1에서 [사업부영문명], [필수], [입력], [항목]의 항목들을 추출할 수 있다. 또한, 대상 문서는 [사업부영문명을 반드시 등록하세요.]의 문장 2에서 [사업부영문명], [반드시], [등록]의 항목들을 추출할 수 있다. 마지막으로 대상 문서는 [그렇지 않은 경우 무효 처리됩니다.]의 문장 3에서 [무효], [처리]의 항목들을 추출할 수 있다.

[0190] 이렇게 분석 문서와 대상 문서의 유사도를 구하기 위한 준비를 마친 후 동의/유의 사전에 등록된 용어와 단어의 유사도를 참고로 각 문장의 유사도를 구하면 다음과 같다.

[0191] 우선 문장 1의 S-T 유사도를 구하면, avg(디비전영문명-사업부영문명, 필수-필수, 입력-입력, 항목-항목) = avg(100%, 100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. 마찬가지로 문장 1의 T-S 유사도를 구해보면, avg(사업부영문명-디비전영문명, 필수-필수, 입력-입력, 항목-항목) = avg(100%, 100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다.

[0192] 도 16a에 이어서 도 16b에서 문장 2와 문장 3의 유사도를 각각 구해보자. 도 16b를 참고하면, 문장 2의 S-T 유사도를 구하면, avg(디비전영문명-사업부영문명, 꼭-반드시, 입력-등록) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. 마찬가지로 문장 2의 T-S 유사도를 구해보면, avg(사업부영문명-디비전영문명, 반드시-꼭, 등록-입력) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다.

[0193] 다음으로 문장 3의 S-T 유사도를 구하면, avg(모두-X, 무효-무효, 처리-처리) = avg(0%, 100%, 100%) = 66.7% 수식을 통해 66.7%의 값을 가진다. 마찬가지로 문장 3의 T-S 유사도를 구해보면, avg(무효-무효, 처리-처리) = avg(100%, 100%) = 100% 수식을 통해 100%의 값을 가진다.

[0194] 각 문장의 S-T 유사도와 T-S 유사도를 통해서, 문서의 S-T 유사도와 T-S 유사도를 구해보자. 문서의 S-T 유사도를 구해보면, avg(문장1 S-T 유사도, 문장2 S-T 유사도, 문장3 S-T 유사도) = avg(100%, 100%, 66.7%) = 88.9% 수식을 통해 88.9%의 값을 가진다. 마찬가지로 문서의 T-S 유사도를 구해보면, avg(문장1 T-S 유사도, 문장2 T-S 유사도, 문장3 T-S 유사도) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다.

[0195] 문서의 S-T 유사도는 88.3%의 값이고, T-S 유사도는 100%의 값이므로 최소값은 88.9%, 최대값은 100%, 평균은 94.4%의 값을 가질 수 있다. 분석 문서와 대상 문서 사이의 유사도는 필요에 따라 이 중에서 어느 하나의 값으

로 정해서 사용할 수 있다.

[0196] 이렇게 문서의 유사도를 연산하면, 문서 작성이나 사전 조회할 때 유사 단어, 유사 용어, 유사 문장을 추천하여 활용할 수 있다. 또한, 이미 작성된 문서나 리포트가 있다면 유사도 분석을 통해서 표절 여부를 검사하는 데 활용할 수 있다.

[0197] 지금까지 도 14a 내지 도 16b를 통해서 용어, 문장, 문서의 유사도를 구하는 경우를 살펴보았다. 본 발명의 이음 동의어 항목 관리 방법은 한글 외에도 다른 언어에도 적용할 수 있다. 도 17a 내지 도 17b에서 영어를 대상으로 용어의 유사도를 구하는 예에 대해서 살펴보도록 하자.

[0198] 다른 언어 대부분도 품사 또는 형태소를 가지고 있으며, 배치 순서에 따라 의미가 달라지는 경우가 적다. 그러므로 본 발명에서 전제로 한 규칙 1과 규칙 2를 그대로 적용하여 단어, 용어, 문장, 문서의 유사도를 연산할 수 있다.

[0199] 도 17a를 참고하면, [DivisionEnglishName]의 분석 용어와 [DepartmentEnglishName]의 대상 용어 사이의 유사도를 연산하는 과정을 볼 수 있다. 동의/유의 사전에 두 용어가 없으므로 바로 유사도를 비교할 수는 없고 하위 항목인 단어 단위로 분리한 후에 유사도를 연산할 수 있다.

[0200] [DivisionEnglishName]의 분석 용어는 [Division], [English], [Name]의 분석 단어를 가지고, [DepartmentEnglishName]의 대상 용어는 [Department], [English], [Name]의 대상 단어를 가진다. [English], [Name]는 공통되나, [Division]과 [Department]의 차이가 있으므로 이 두 단어의 유사도에 따라 용어의 유사도가 결정될 것이다.

[0201] 실제로 유사도를 연산해보면, S-T 유사도는 avg(Division-Department, English-English, Name-Name) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가지며, 마찬가지로 T-S 유사도는 avg(Department-Division, English-English, Name-Name) = avg(100%, 100%, 100%) = 100% 수식을 통해 100%의 값을 가진다. S-T 유사도와 T-S 유사도가 똑같이 100%의 값을 가지므로 min, max, avg 모두 100%의 값을 가진다.

[0202] 즉 [DivisionEnglishName]와 [DepartmentEnglishName]의 유사도는 100%라는 값을 가지며, 그 결과를 동의/유의 사전에 추가할 수 있다. 또한 사용자가 새로운 항목으로 [DivisionEnglishName]을 등록하려는 경우, [DivisionEnglishName]을 등록하는 대신 [DepartmentEnglishName]을 사용할 것을 제안할 수 있다.

[0203] 도 17b를 참고하면, [WorkEnglishName]의 분석 용어와 [BusinessFieldEnglishName]의 대상 용어 사이의 유사도를 연산하는 과정을 볼 수 있다. 동의/유의 사전에 두 용어가 없으므로 바로 유사도를 비교할 수는 없고 하위 항목인 단어 단위로 분리한 후에 유사도를 연산할 수 있다.

[0204] [WorkEnglishName]의 분석 용어는 [Work], [English], [Name]의 분석 단어를 가지고, [BusinessFieldEnglishName]의 대상 용어는 [Business], [Field], [English], [Name]의 대상 단어를 가진다. [English], [Name]은 공통되지만, [Work]와 [Business], [Field]의 차이가 있으므로, 단어 [Work]와 [Business], [Field]의 유사도에 따라 용어의 유사도가 결정될 것이다.

[0205] 실제로 유사도를 연산해보면, S-T 유사도는 avg(Work-Business, English-English, Name-Name) = avg(100%, 100%, 100%) = 100% 수식을 통해 100의 값을 가지며, 마찬가지로 T-S 유사도는 avg(Business-Work, Field-Work, English-English, Name-Name) = avg(100%, 50%, 100%, 100%) = 87.5% 수식을 통해 87.5%의 값을 가진다. 따라서 [WorkEnglishName]의 분석 용어와 [BusinessFieldEnglishName]의 대상 용어 사이의 유사도는 최소값 87.5%, 최대값 100%, 평균 93.8%의 값을 가진다.

[0206] 지금까지 도면들을 통해서 동의/유의 사전에 등재된 단어 사이의 유사도를 바탕으로 단어보다 상위 항목인 용어의 유사도, 문장의 유사도, 문서의 유사도를 구하는 과정을 살펴보았다. 이때 단어 사이의 유사도는 동의어인 경우 100%, 유의어인 경우 50%로 가정하고 설명을 하였다. 또한, 단어의 동의/유의 사전은 이미 구축이 된 것으로 가정하고 설명을 하였다.

[0207] 하지만 용어에서 신조어가 생기는 것처럼 단어에서도 신조어가 생길 수 있다. 새로운 단어가 등장하면 기존 단어와 비교하여 새로운 단어와 기존 단어 사이의 유사도를 연산하고, 연산된 유사도를 동의/유의 사전에 등록하는 작업이 필요하다. 하지만 이를 수작업으로 진행하는 것은 매우 불편할 것이다.

[0208] 이럴 때 외부 API (Application Programming Interface)를 사용할 수 있다. 예를 들어 네이버 검색 오픈 API를 이용하여 새로 등장한 단어의 의미를 조회하고 기존 단어의 의미와의 유사도를 본 발명의 유사도 연산 방법을

적용하여 연산하면 동의/유의 사전에 자동으로 유사도를 등록할 수 있다.

[0209] 마찬가지로 영어의 경우에도 외부 API를 사용할 수 있다. 예를 들면, 영어 단어의 의미를 검색하기 위해 옥스포드 영어 사전의 오픈 API를 활용할 수 있다. http://public.oed.com/subscriber-services/sru-service/ 링크에서 옥스포드 영어 사전의 오픈 API에 관한 자세한 내용을 확인할 수 있다. 이처럼 새로 생성된 특정 단어의 의미는 외부의 다양한 API를 통해서 수집이 가능하다.

[0210] 예를 들어 네이버 사전을 이용하여 단어의 유사도 관리를 자동화한다고 가정해 보자. 이때 시스템에는 [성공]이라는 단어가 이미 등록되어 있다. 네이버 사전에서 "성공"을 검색해보면 다음과 같은 결과를 얻을 수 있다. "성공: 목적하는 바를 이룸." 이때 [성취]라는 단어가 새로 생성되었다고 가정해보자. 이 경우 사람이 인위적으로 [성공]과 [성취]의 유사도를 연산할 필요 없이 오픈 API를 통해 네이버 사전에서 "성취"를 검색하여 그 의미를 시스템에 저장하고, [성공]의 의미와 유사도를 연산하면 된다.

[0211] 네이버 사전에서 "성취"를 검색해보면 다음과 같은 결과를 얻을 수 있다. "성취: 목적한 바를 이룸." 다음으로 "성공"의 의미와 "성취"의 의미 사이의 유사도를 연산하여 이를 [성공]과 [성취]의 유사도로 사용하면 된다. 즉 avg(목적-목적, 바-바, 이룸-이룸) = avg(100%, 100%, 100%) = 100% 수식을 통해서 [성공]과 [성취]는 유사도가 100%인 동의어임을 확인할 수 있다.

[0212] 이와 같은 방식으로 오픈 API를 통해 외부 사전으로부터 단어의 의미를 조회하고, 단어의 의미로 조회된 문장의 유사도를 연산하면, 새로 등장한 단어의 유사도도 자동으로 동의/유의 사전으로 관리할 수 있다. 이 경우에는 앞서 가정한 것처럼 유의어의 유사도가 50%로 고정되어서 나오는 것이 아니라 각 단어의 의미에 포함된 단어들로 인해 다양한 값을 가지게 될 것이다.

[0213] 도 18은 본 발명의 일 실시 예에 따른 유사도 분석 기반 이음 동의 항목 관리 장치의 하드웨어 구성도이다.

[0214] 도 18를 참고하면 본 발명에서 제안하는 유사도 분석 기반 이음 동의 항목 관리 장치(10)는 하나 이상의 프로세서(510), 메모리(520), 스토리지(560) 및 인터페이스(570)을 포함할 수 있다. 프로세서(510), 메모리(520), 스토리지(560) 및 인터페이스(570)는 시스템 버스(550)를 통하여 데이터를 송수신한다.

[0215] 프로세서(510)는 메모리(520)에 로드(load)된 컴퓨터 프로그램을 실행하고, 메모리(520)는 상기 컴퓨터 프로그램을 스토리지(560)에서 로드(load) 한다. 상기 컴퓨터 프로그램은, 항목 추출 오퍼레이션(521), 유사도 분석 오퍼레이션(523) 및 동의/유사 추천 오퍼레이션(525)을 포함할 수 있다.

[0216] 항목 추출 오퍼레이션(521)은 스토리지(560)에서 문서(561)를 읽어서 시스템 버스(550)를 통해 메모리(520)에 로드(load)할 수 있다. 다음으로 문서(561)를 대상으로 마침표를 기준으로 문장을 추출하고, 스토리지(560)의 어미/조사 사전과 띄어쓰기를 기준으로 용어를 추출하고, 스토리지(560)의 형태소 사전(565)를 기준으로 단어를 추출할 수 있다.

[0217] 항목 추출 오퍼레이션(521)이 제1 문서와 제2 문서에서 각각 문장, 용어, 단어를 추출하면, 제1 문서와 제2 문서의 유사도를 직접적으로 연산할 수는 없어도, 제1 문서와 제2 문서를 구성하는 각각의 문장, 용어, 단어의 유사도를 이용하여 제1 문서와 제2 문서의 유사도를 간접적으로 연산할 수 있다.

[0218] 유사도 분석 오퍼레이션(523)은 스토리지(560)의 동의/유의 사전(567)를 참고하여, 제1 문서와 제2 문서의 유사도를 연산할 수 있다. 만약 동의/유의 사전(567)에 제1 문서와 제2 문서의 유사도가 등록되어 있다면 이를 이용할 수 있다. 그러나 동의/유의 사전(567)에 제1 문서와 제2 문서의 유사도가 등록되어 있지 않다면 제1 문서를 구성하는 제1 문장과 제2 문서를 구성하는 제2 문장의 유사도를 이용하여 제1 문서와 제2 문서의 유사도를 연산할 수 있다.

[0219] 만약 제1 문장과 제2 문장의 유사도가 동의/유의 사전(567)에 등록되어 있지 않다면 마찬가지로, 제1 문장을 구성하는 제1 용어와 제2 문장을 구성하는 제2 용어의 유사도를 이용하여 제1 문장과 제2 문장의 유사도를 구할 수 있다. 이때 만약 제1 용어와 제2 용어의 유사도가 동의/유의 사전(567)에 등록되어 있지 않다면, 제1 용어를 구성하는 제1 단어와 제2 용어를 구성하는 제2 단어의 유사도를 이용하여 제1 용어와 제2 용어의 유사도를 구할 수 있다.

[0220] 동의/유사 추천 오퍼레이션(525)는 유사도 분석 오퍼레이션(523)에서 분석한 결과를 활용하여 비슷한 의미의 문서나, 비슷한 의미의 문장이나, 비슷한 의미의 용어나, 비슷한 의미의 단어를 추천할 수 있다. 이는 사용자가 문서를 작성하거나 사전을 조회하는 데 활용될 수 있다. 또는 문서라 리포트의 유사도를 분석하여 표절 여부를

검사하는 데 활용될 수 있다.

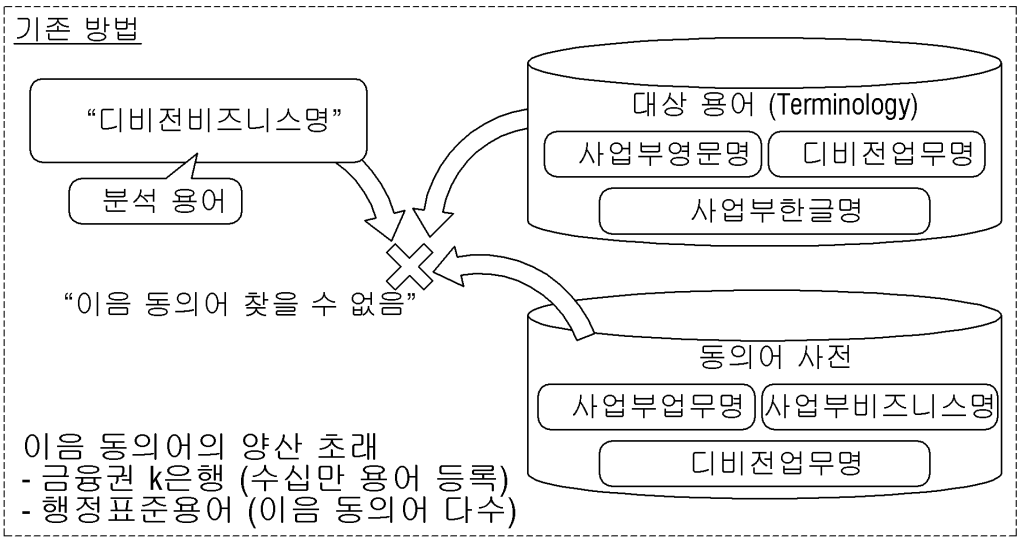
[0221] 또는 특정 논문과 관련성이 높은 논문을 검색하여 제공하거나, 특정 특허 문서와 관련성이 높은 특허 문서를 검색하여 제공할 수 있다. 이렇게 추천된 동의/유사 단어, 용어, 문장, 문서는 인터페이스(570)을 통해 네트워크(network)를 거쳐서 사용자에게 제공될 수 있다.

[0222] 도 18의 각 구성 요소는 소프트웨어(Software) 또는, FPGA(Field Programmable Gate Array)나 ASIC(Application-Specific Integrated Circuit)와 같은 하드웨어(Hardware)를 의미할 수 있다. 그렇지만, 상기 구성 요소들은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, 어드레싱(Addressing) 할 수 있는 저장 매체에 있도록 구성될 수도 있고, 하나 또는 그 이상의 프로세서들을 실행시키도록 구성될 수도 있다. 상기 구성 요소들 안에서 제공되는 기능은 더 세분된 구성 요소에 의하여 구현될 수 있으며, 복수의 구성 요소들을 합하여 특정한 기능을 수행하는 하나의 구성 요소로 구현될 수도 있다.

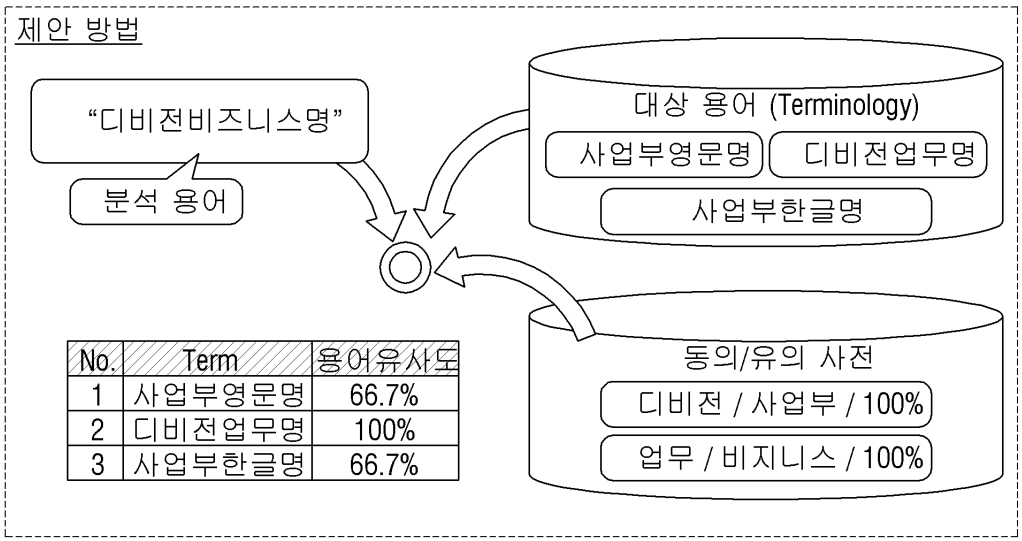
[0223] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

도면

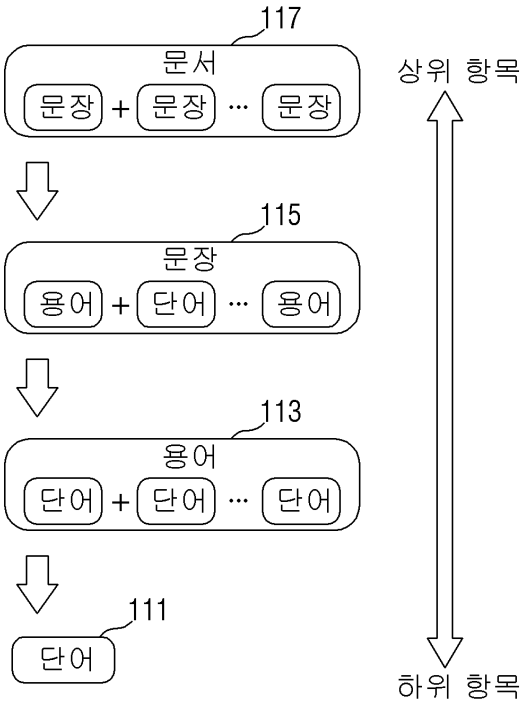
도면1a



도면1b



도면2

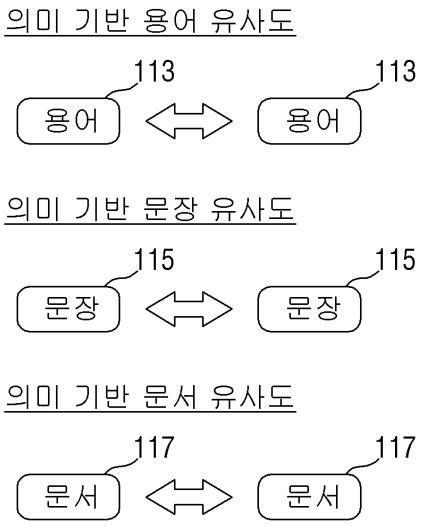


도면3a

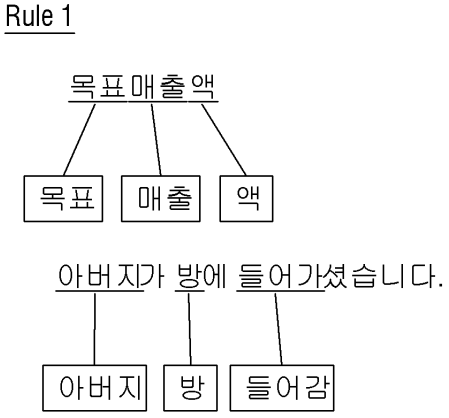
의미 기반 단어 유사도

구분	동의어 (100%)	유의어 (50%)
성공	성취	달성, 출세, 입신
실패	-	실수, 실책, 낭패
사업	비즈니스	업무, 일, 영업
입력	등록	추가, 생성

도면3b



도면4a



도면4b

Rule 2

목표매출액 ⇔ 매출목표액

아버지가 방에 들어가셨습니다.

⇔ 방에 아버지가 들어가셨습니다.

도면5a

Equation 1

유사도 = 자기 자신을 100%로 봤을 때,
비교 대상과의 유사도를 0%~100%로 표시

도면5b

Equation 2

항목의 S-T 유사도 = 분석 항목의 대상 항목에 대한 유사도
= 분석 항목을 구성하는 단어의 대상 항목의 구성 단어와의 유사도 평균

항목의 T-S 유사도 = 대상 항목의 분석 항목에 대한 유사도
= 대상 항목을 구성하는 단어의 분석 항목의 구성 단어와의 유사도 평균

도면5c

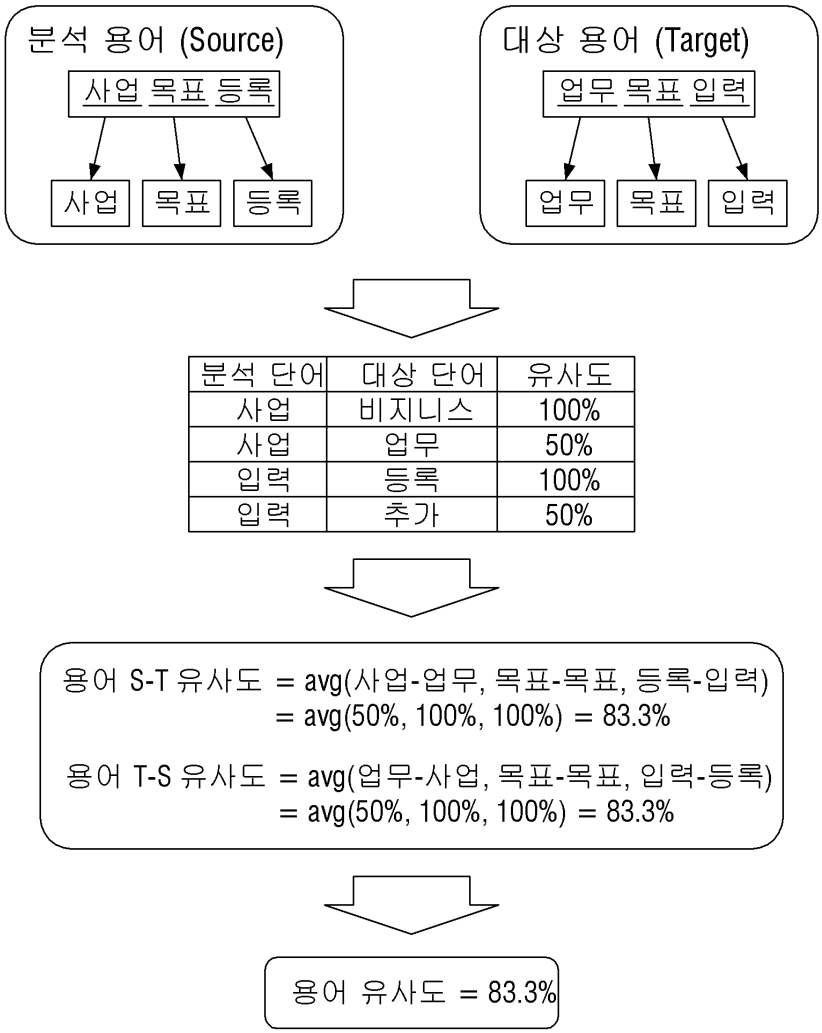
Equation 3

항목 유사도 = min(항목의 S-T 유사도, 항목의 T-S 유사도)

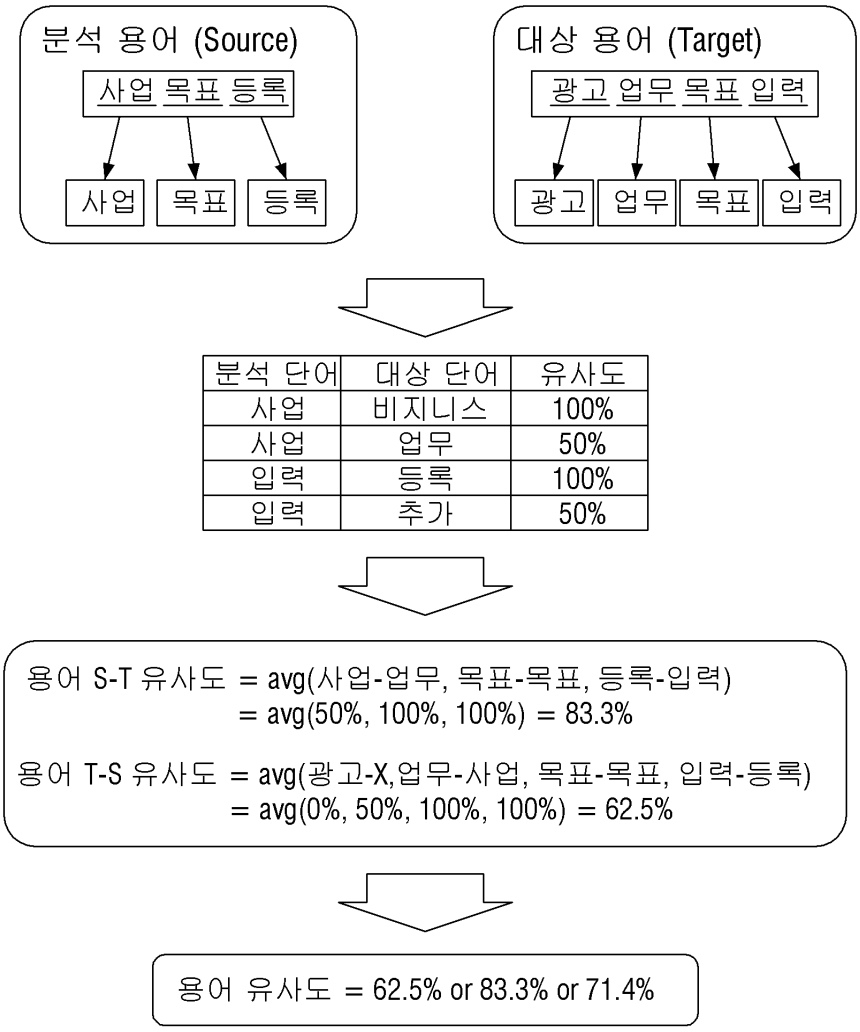
항목 유사도 = max(항목의 S-T 유사도, 항목의 T-S 유사도)

항목 유사도 = avg(항목의 S-T 유사도, 항목의 T-S 유사도)

도면6


분석 용어 (Source)
사업 목표 등록
사업
목표
등록
대상 용어 (Target)
업무 목표 입력
업무
목표
입력
분석 단어 | 대상 단어 | 유사도
사업 | 비지니스 | 100%
사업 | 업무 | 50%
입력 | 등록 | 100%
입력 | 추가 | 50%
용어 S-T 유사도 = avg(사업-업무, 목표-목표, 등록-입력)
= avg(50%, 100%, 100%) = 83.3%
용어 T-S 유사도 = avg(업무-사업, 목표-목표, 입력-등록)
= avg(50%, 100%, 100%) = 83.3%
용어 유사도 = 83.3%

도면7


분석 용어 (Source)
사업 목표 등록
사업
목표
등록
대상 용어 (Target)
광고 업무 목표 입력
광고
업무
목표
입력
분석 단어 | 대상 단어 | 유사도
사업 | 비지니스 | 100%
사업 | 업무 | 50%
입력 | 등록 | 100%
입력 | 추가 | 50%
용어 S-T 유사도 = avg(사업-업무, 목표-목표, 등록-입력)
= avg(50%, 100%, 100%) = 83.3%
용어 T-S 유사도 = avg(광고-X,업무-사업, 목표-목표, 입력-등록)
= avg(0%, 50%, 100%, 100%) = 62.5%
용어 유사도 = 62.5% or 83.3% or 71.4%

도면8

분석 단어	대상 단어	유사도
사업	비지니스	100%
사업	업무	50%
입력	등록	100%
입력	추가	50%

⇩

분석 용어	대상 용어	유사도
사업목표등록	업무목표입력	83.3%
사업목표등록	광고업무목표입력	62.5%

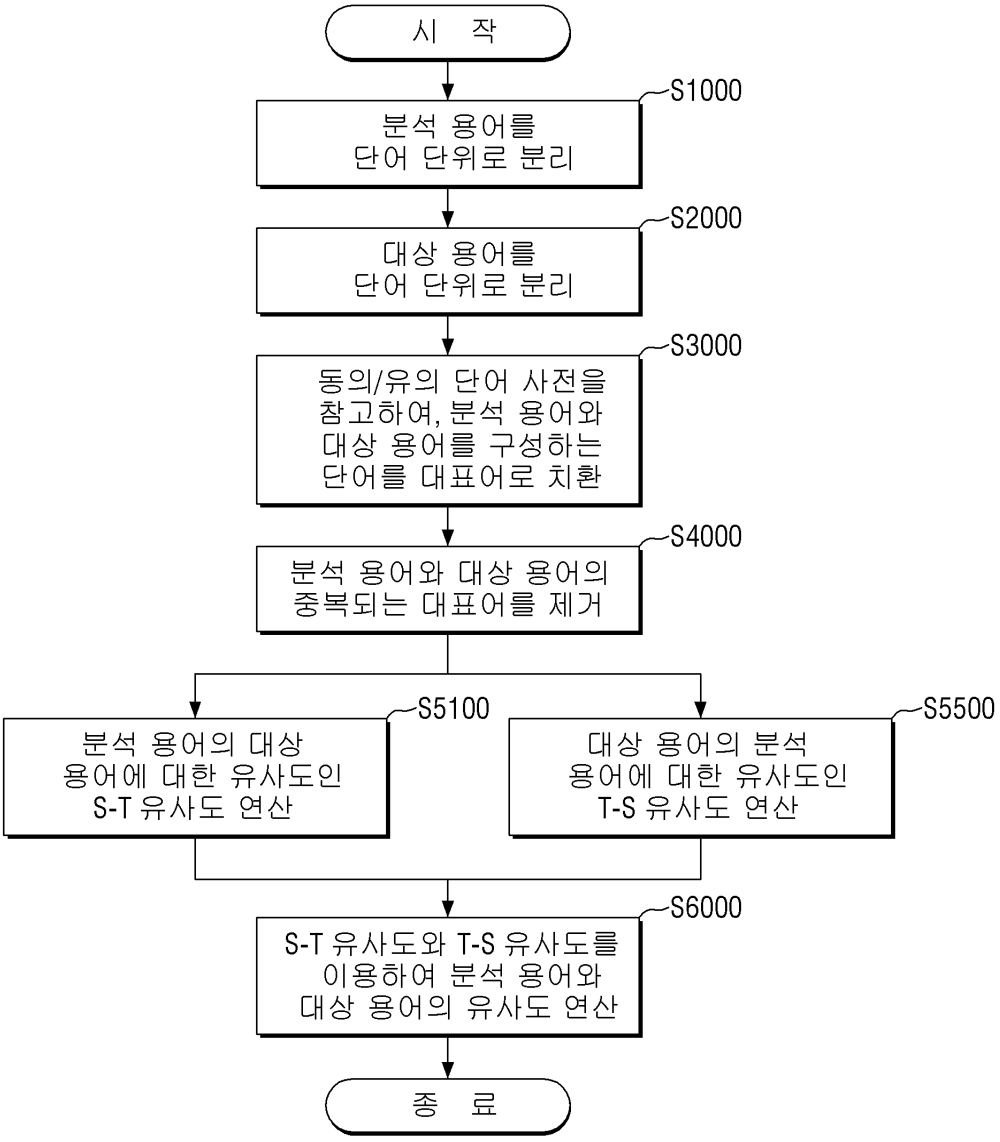
⇩

분석 문장	대상 문장	유사도
...	...	...

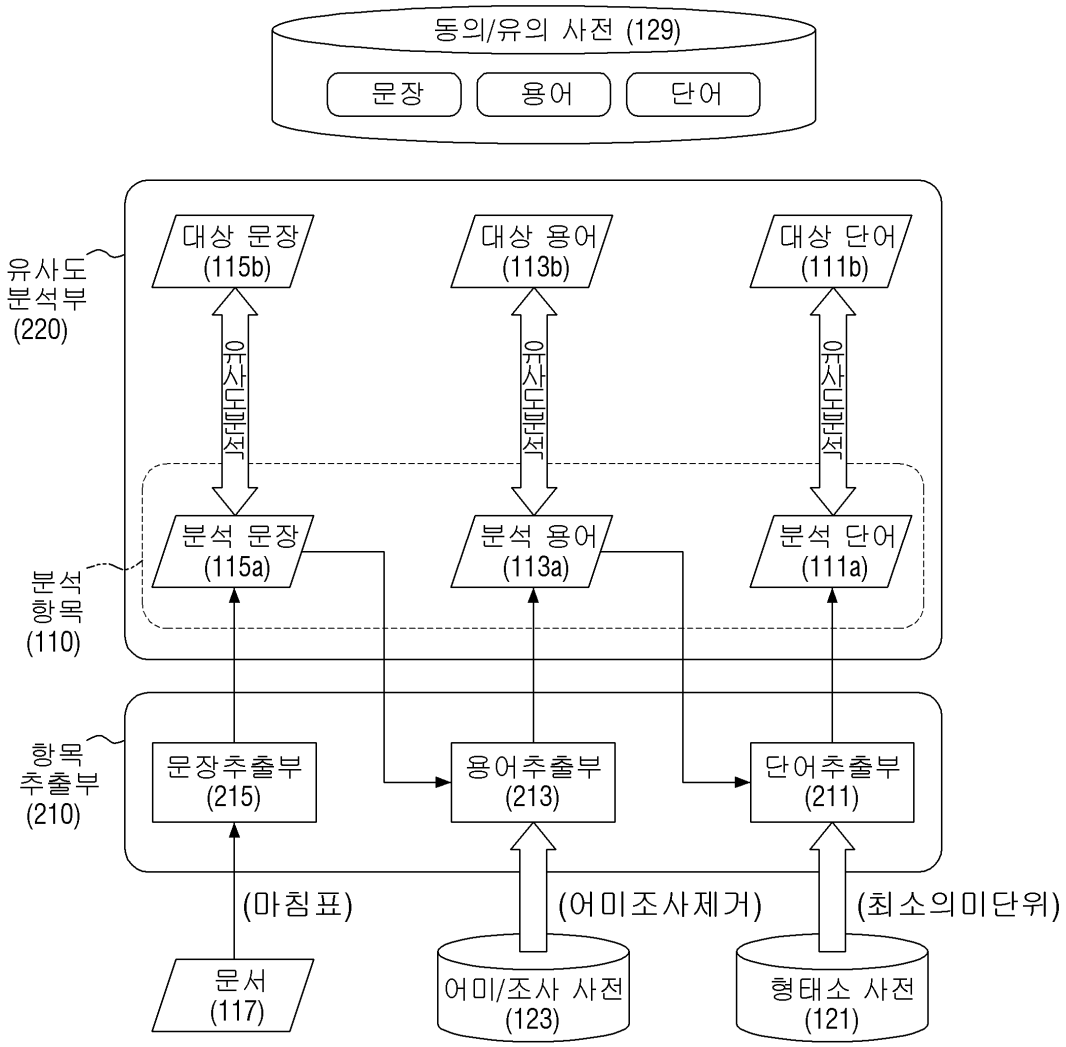
⇩

분석 문서	대상 문서	유사도
...	...	...

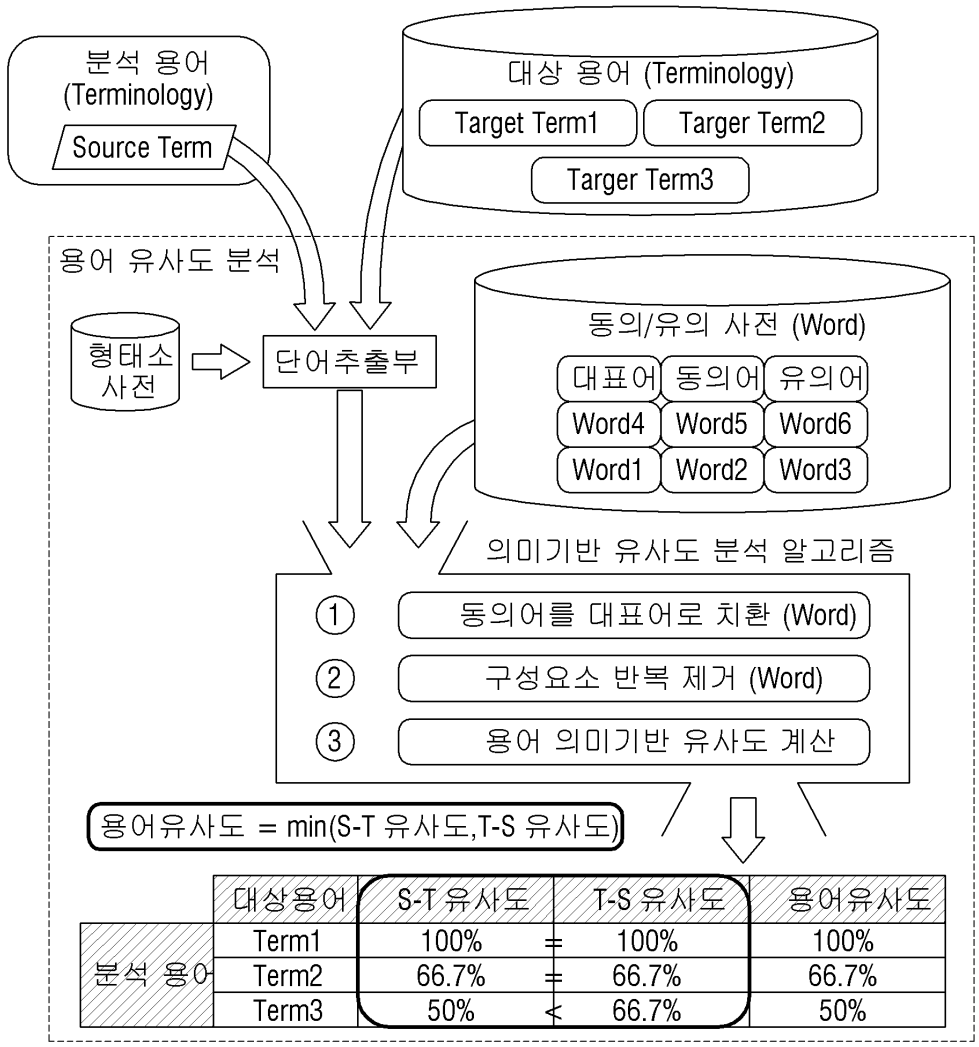
도면9


시 작
S1000
분석 용어를
단어 단위로 분리
S2000
대상 용어를
단어 단위로 분리
S3000
동의/유의 단어 사전을
참고하여, 분석 용어와
대상 용어를 구성하는
단어를 대표어로 치환
S4000
분석 용어와 대상 용어의
중복되는 대표어를 제거
S5100
분석 용어의 대상
용어에 대한 유사도인
S-T 유사도 연산
S5500
대상 용어의 분석
용어에 대한 유사도인
T-S 유사도 연산
S6000
S-T 유사도와 T-S 유사도를
이용하여 분석 용어와
대상 용어의 유사도 연산
종 료

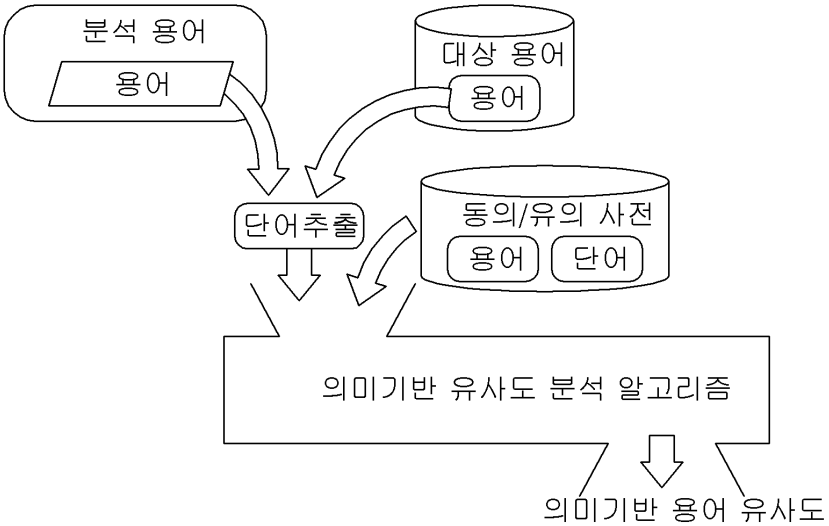
도면10


동의/유의 사전 (129)
문장
용어
단어
유사도 분석부 (220)
대상 문장 (115b)
대상 용어 (113b)
대상 단어 (111b)
유사도분석
유사도분석
유사도분석
분석 항목 (110)
분석 문장 (115a)
분석 용어 (113a)
분석 단어 (111a)
항목 추출부 (210)
문장추출부 (215)
용어추출부 (213)
단어추출부 (211)
(마침표)
(어미조사제거)
(최소의미단위)
문서 (117)
어미/조사 사전 (123)
형태소 사전 (121)

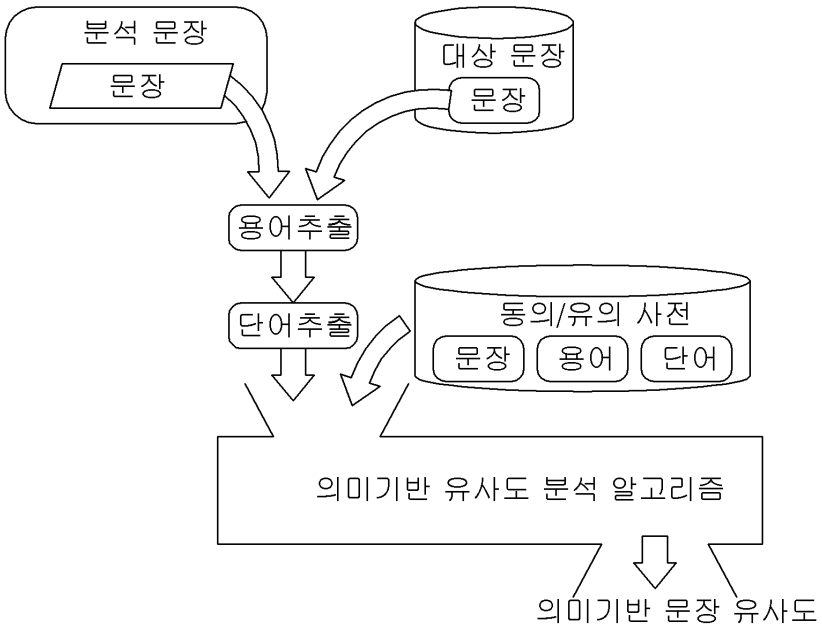
도면11


분석 용어
(Terminology)
Source Term
대상 용어 (Terminology)
Target Term1
Targer Term2
Targer Term3
용어 유사도 분석
형태소
사전
단어추출부
동의/유의 사전 (Word)
대표어
동의어
유의어
Word4
Word5
Word6
Word1
Word2
Word3
의미기반 유사도 분석 알고리즘
① 동의어를 대표어로 치환 (Word)
② 구성요소 반복 제거 (Word)
③ 용어 의미기반 유사도 계산
용어유사도 = min(S-T 유사도,T-S 유사도)
대상용어
S-T 유사도
T-S 유사도
용어유사도
분석 용어
Term1 100% = 100% 100%
Term2 66.7% = 66.7% 66.7%
Term3 50% < 66.7% 50%

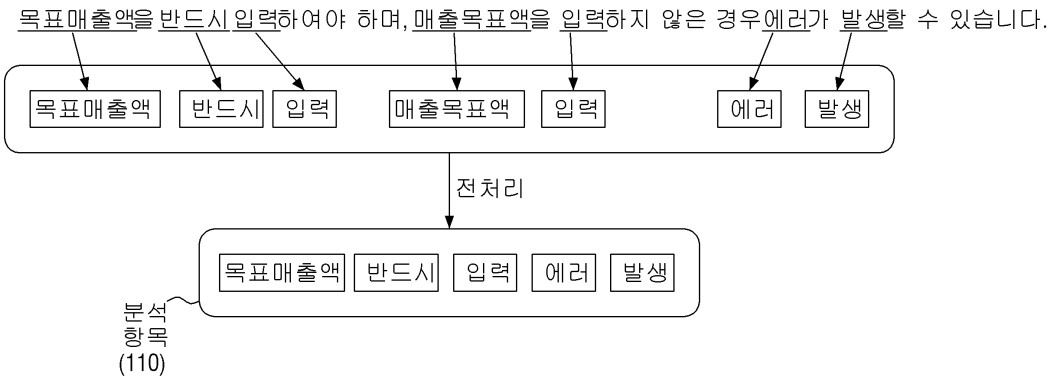
도면12a


분석 용어
용어
대상 용어
용어
단어추출
동의/유의 사전
용어
단어
의미기반 유사도 분석 알고리즘
의미기반 용어 유사도

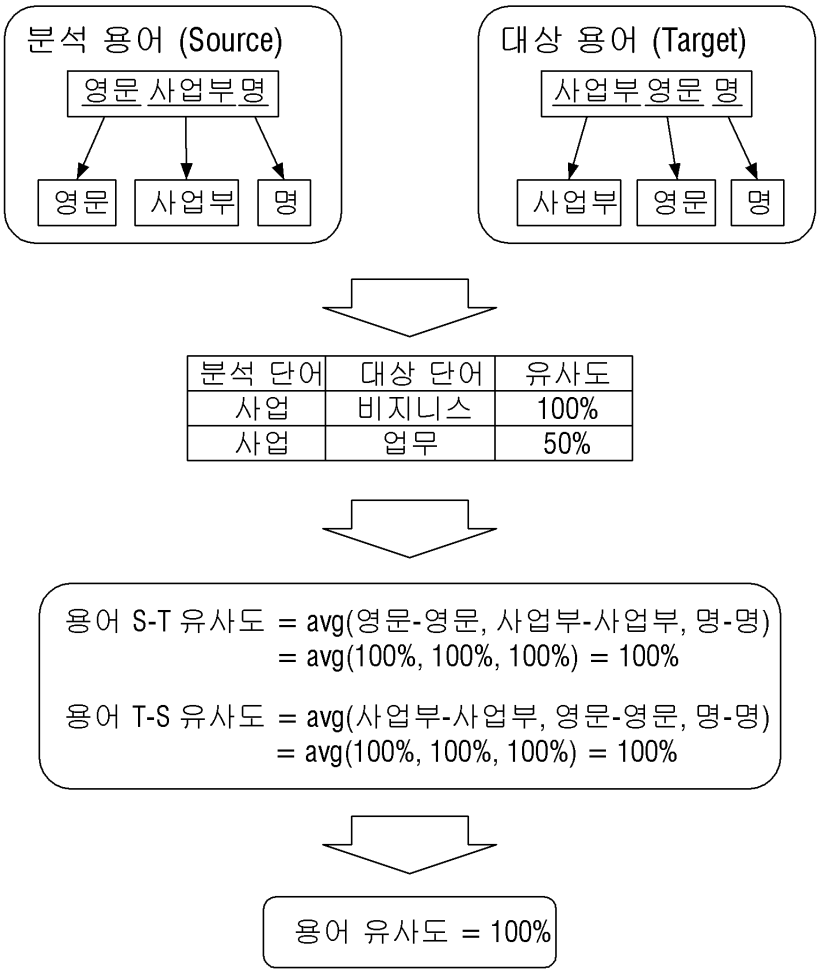
도면12b


분석 문장
문장
대상 문장
문장
용어추출
단어추출
동의/유의 사전
문장
용어
단어
의미기반 유사도 분석 알고리즘
의미기반 문장 유사도

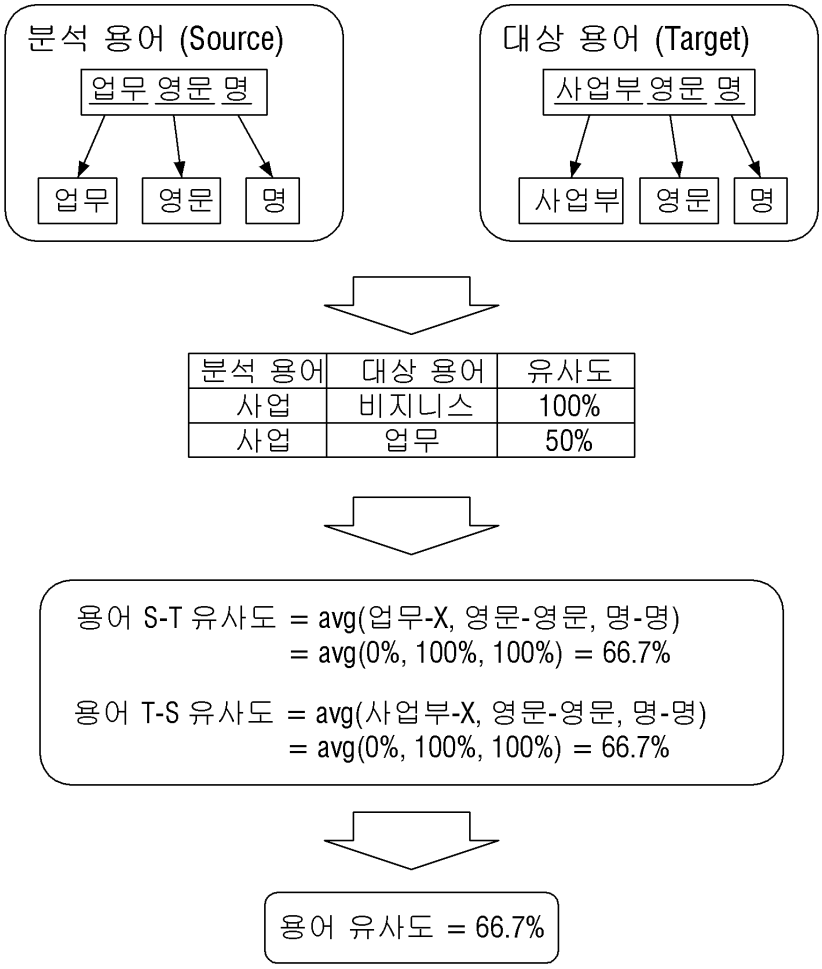
도면13


목표매출액을 반드시 입력하여야 하며, 매출목표액을 입력하지 않은 경우 에러가 발생할 수 있습니다.
목표매출액
반드시
입력
매출목표액
입력
에러
발생
전처리
목표매출액
반드시
입력
에러
발생
분석
항목
(110)

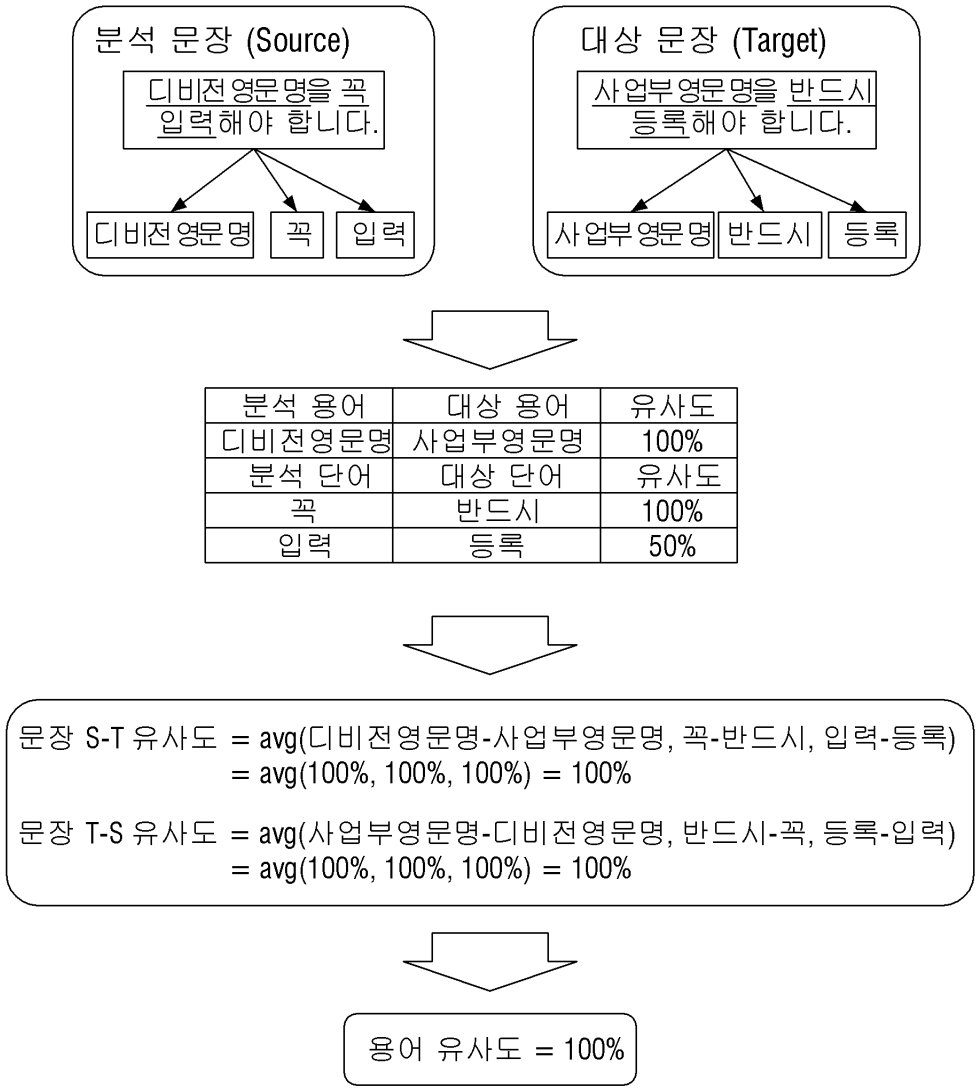
도면14a


분석 용어 (Source)
영문 사업부명
영문
사업부
명
대상 용어 (Target)
사업부 영문 명
사업부
영문
명
분석 단어 | 대상 단어 | 유사도
사업 | 비지니스 | 100%
사업 | 업무 | 50%
용어 S-T 유사도 = avg(영문-영문, 사업부-사업부, 명-명)
= avg(100%, 100%, 100%) = 100%
용어 T-S 유사도 = avg(사업부-사업부, 영문-영문, 명-명)
= avg(100%, 100%, 100%) = 100%
용어 유사도 = 100%

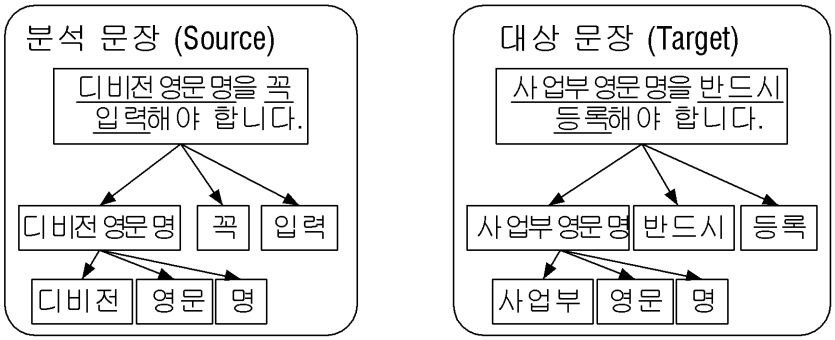
도면14b


분석 용어 (Source)
업무 영문 명
업무
영문
명
대상 용어 (Target)
사업부 영문 명
사업부
영문
명
분석 용어 | 대상 용어 | 유사도
사업 | 비지니스 | 100%
사업 | 업무 | 50%
용어 S-T 유사도 = avg(업무-X, 영문-영문, 명-명)
= avg(0%, 100%, 100%) = 66.7%
용어 T-S 유사도 = avg(사업부-X, 영문-영문, 명-명)
= avg(0%, 100%, 100%) = 66.7%
용어 유사도 = 66.7%

도면15a


분석 문장 (Source)
디비전영문명을 꼭 입력해야 합니다.
디비전영문명
꼭
입력
대상 문장 (Target)
사업부영문명을 반드시 등록해야 합니다.
사업부영문명
반드시
등록
분석 용어 | 대상 용어 | 유사도
디비전영문명 | 사업부영문명 | 100%
분석 단어 | 대상 단어 | 유사도
꼭 | 반드시 | 100%
입력 | 등록 | 50%
문장 S-T 유사도 = avg(디비전영문명-사업부영문명, 꼭-반드시, 입력-등록)
= avg(100%, 100%, 100%) = 100%
문장 T-S 유사도 = avg(사업부영문명-디비전영문명, 반드시-꼭, 등록-입력)
= avg(100%, 100%, 100%) = 100%
용어 유사도 = 100%

도면15b



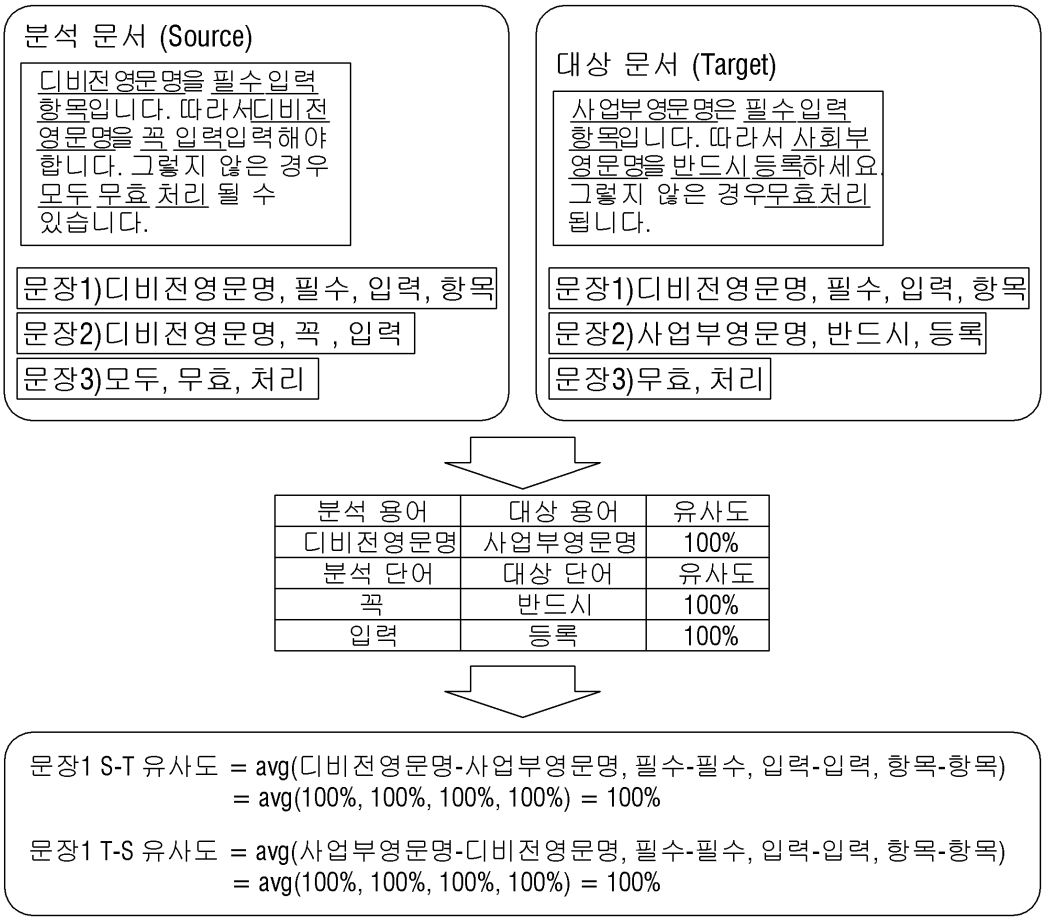
분석 단어	대상 단어	유사도
디비전	사업부	100%
꼭	반드시	100%
입력	등록	100%

문장 S-T 유사도 = avg(avg(디비전-사업부, 영문-영문, 명-명), 꼭-반드시, 입력-등록)
= avg(avg(100%, 100%, 100%), 100%, 100%) = 100%

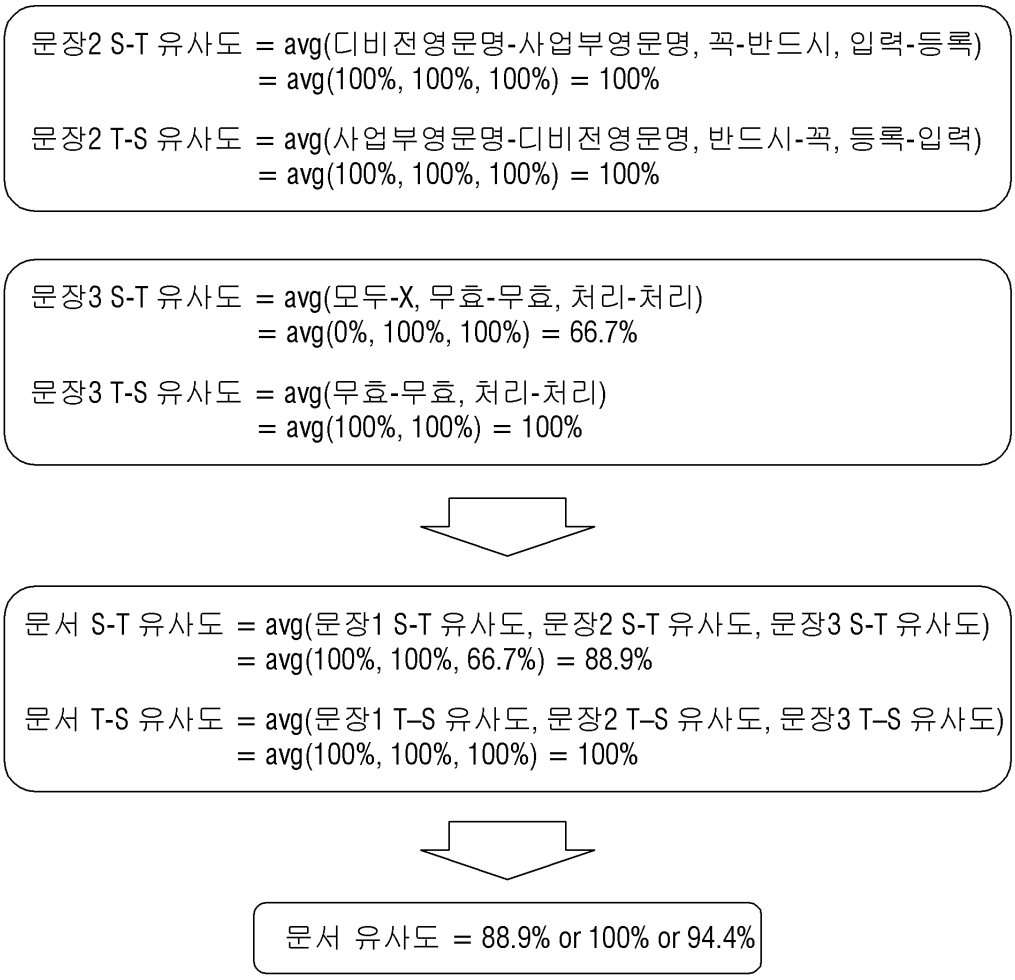
문장 T-S 유사도 = avg(avg(사업부-디비전, 영문-영문, 명-명), 반드시-꼭, 등록-입력)
= avg(avg(100%, 100%, 100%), 100%, 100%) = 100%

용어 유사도 = 100%

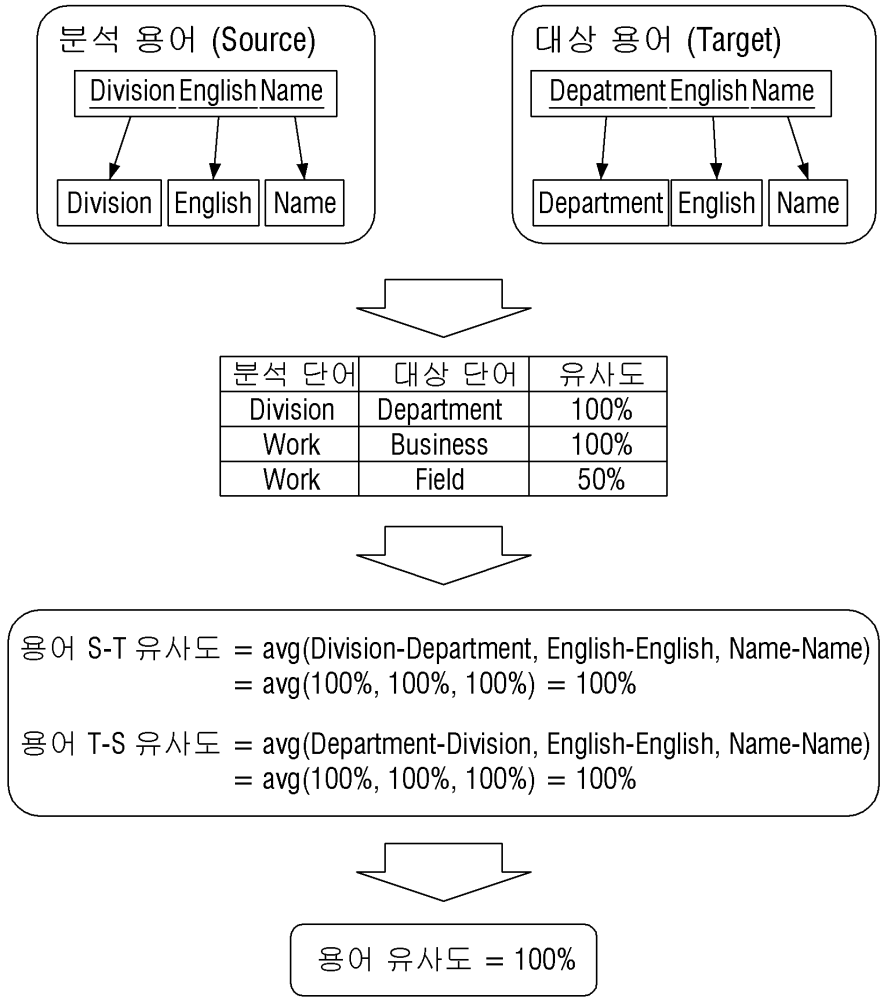
도면16a


분석 문서 (Source)
디비전영문명을 필수입력 항목입니다. 따라서디비전 영문명을 꼭 입력입력해야 합니다. 그렇지 않은 경우 모두 무효 처리 될 수 있습니다.
문장1)디비전영문명, 필수, 입력, 항목
문장2)디비전영문명, 꼭 , 입력
문장3)모두, 무효, 처리
대상 문서 (Target)
사업부영문명은 필수입력 항목입니다. 따라서 사회부 영문명을 반드시등록하세요. 그렇지 않은 경우무효처리 됩니다.
문장1)디비전영문명, 필수, 입력, 항목
문장2)사업부영문명, 반드시, 등록
문장3)무효, 처리
분석 용어 대상 용어 유사도
디비전영문명 사업부영문명 100%
분석 단어 대상 단어 유사도
꼭 반드시 100%
입력 등록 100%
문장1 S-T 유사도 = avg(디비전영문명-사업부영문명, 필수-필수, 입력-입력, 항목-항목)
= avg(100%, 100%, 100%, 100%) = 100%
문장1 T-S 유사도 = avg(사업부영문명-디비전영문명, 필수-필수, 입력-입력, 항목-항목)
= avg(100%, 100%, 100%, 100%) = 100%

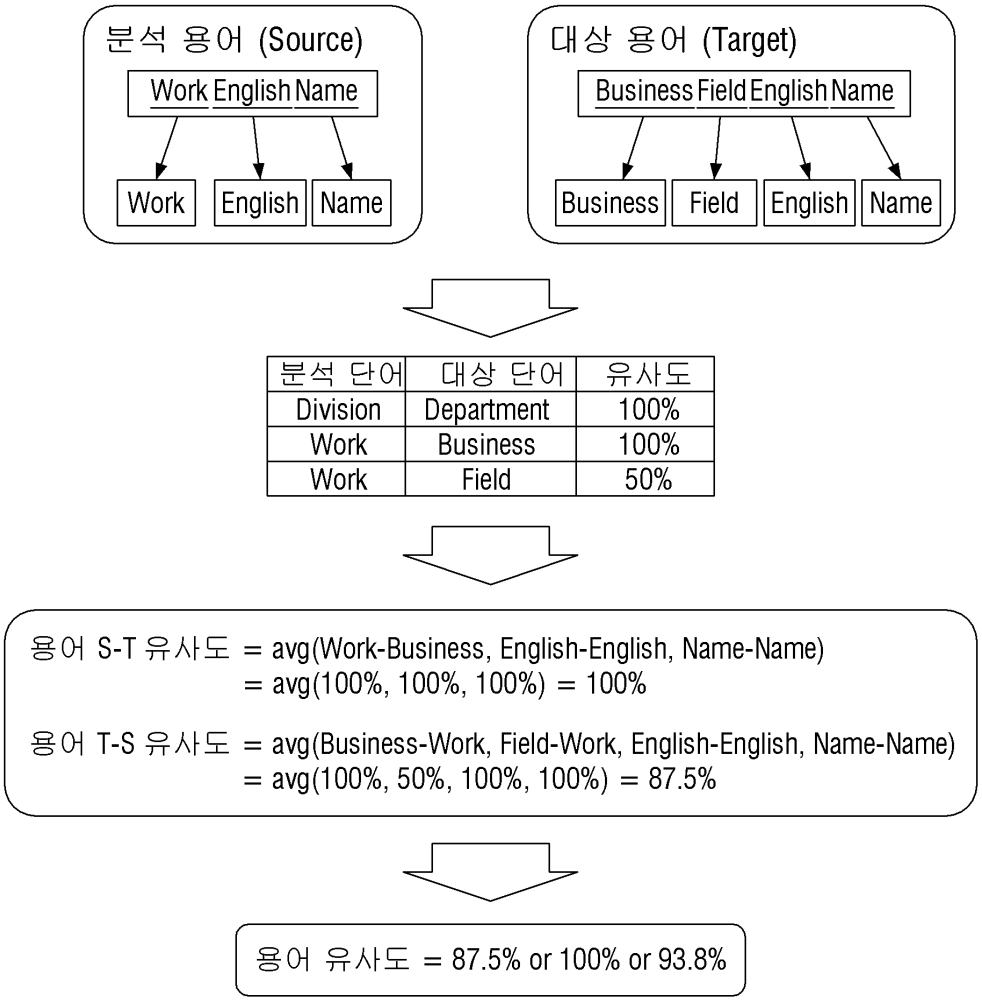
도면16b


문장2 S-T 유사도 = avg(디비전영문명-사업부영문명, 꼭-반드시, 입력-등록)
= avg(100%, 100%, 100%) = 100%
문장2 T-S 유사도 = avg(사업부영문명-디비전영문명, 반드시-꼭, 등록-입력)
= avg(100%, 100%, 100%) = 100%
문장3 S-T 유사도 = avg(모두-X, 무효-무효, 처리-처리)
= avg(0%, 100%, 100%) = 66.7%
문장3 T-S 유사도 = avg(무효-무효, 처리-처리)
= avg(100%, 100%) = 100%
문서 S-T 유사도 = avg(문장1 S-T 유사도, 문장2 S-T 유사도, 문장3 S-T 유사도)
= avg(100%, 100%, 66.7%) = 88.9%
문서 T-S 유사도 = avg(문장1 T-S 유사도, 문장2 T-S 유사도, 문장3 T-S 유사도)
= avg(100%, 100%, 100%) = 100%
문서 유사도 = 88.9% or 100% or 94.4%

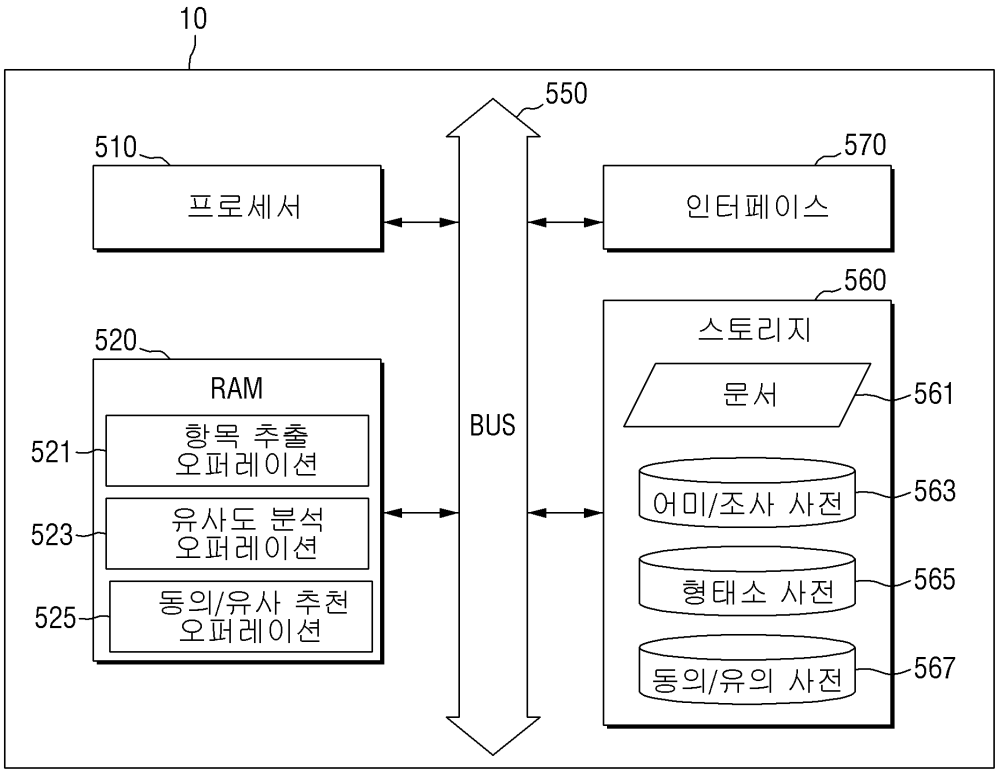
도면17a


분석 용어 (Source)
Division English Name
Division
English
Name
대상 용어 (Target)
Depatment English Name
Department
English
Name
분석 단어 | 대상 단어 | 유사도
Division | Department | 100%
Work | Business | 100%
Work | Field | 50%
용어 S-T 유사도 = avg(Division-Department, English-English, Name-Name)
= avg(100%, 100%, 100%) = 100%
용어 T-S 유사도 = avg(Department-Division, English-English, Name-Name)
= avg(100%, 100%, 100%) = 100%
용어 유사도 = 100%

도면17b


분석 용어 (Source)
Work English Name
Work
English
Name
대상 용어 (Target)
Business Field English Name
Business
Field
English
Name
분석 단어 대상 단어 유사도
Division Department 100%
Work Business 100%
Work Field 50%
용어 S-T 유사도 = avg(Work-Business, English-English, Name-Name)
= avg(100%, 100%, 100%) = 100%
용어 T-S 유사도 = avg(Business-Work, Field-Work, English-English, Name-Name)
= avg(100%, 50%, 100%, 100%) = 87.5%
용어 유사도 = 87.5% or 100% or 93.8%

도면18


10
550
510
프로세서
570
인터페이스
560
스토리지
520
RAM
문서
561
521
항목 추출
오퍼레이션
BUS
어미/조사 사전
563
523
유사도 분석
오퍼레이션
형태소 사전
565
525
동의/유사 추천
오퍼레이션
동의/유의 사전
567