(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0050180
(43) 공개일자 2019년05월10일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) *G06F 17/10* (2006.01)
(52) CPC특허분류
G06F 17/2715 (2013.01)
G06F 17/10 (2013.01)
(21) 출원번호 10-2017-0145461
(22) 출원일자 2017년11월02일
심사청구일자 2017년11월02일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
서정연
서울특별시 서초구 서초중앙로 260 B동 203호 (반포동, 삼호가든3차)
고영중
부산광역시 해운대구 좌2동 벽산 1차 아파트 105동 1304호
염홍선
대전광역시 서구 둔산로 155, 크로바아파트 103동 103호(둔산동, 크로바아파트)
(74) 대리인
이지연

전체 청구항 수 : 총 6 항

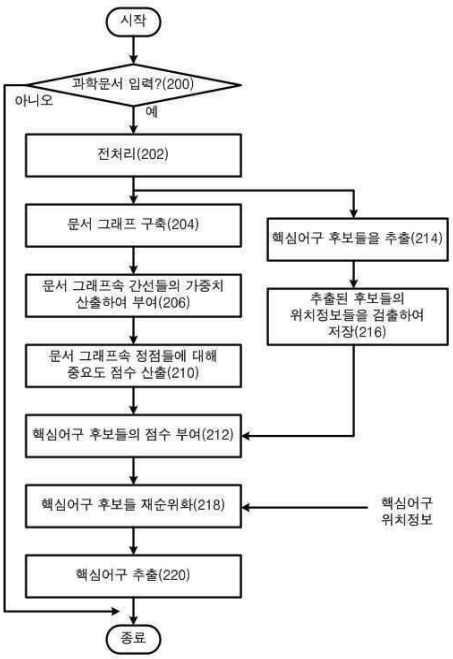
(54) 발명의 명칭 **과학문서의 핵심어구 추출방법 및 장치**

(57) 요 약

본 발명에 따르는 과학문서의 핵심어 추출방법은, 과학문서를 입력받아 형태소 분석을 통해 명사와 형용사, 동사로 분석된 단어로 구성되는 형태소 분석된 과학문서로 변환하는 단계; 상기 과학문서의 단어를 정점으로 하고, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 문서 그래프를 구축하는 단계; 상기 문서 그래프내의 정점

(뒷면에 계속)

대 표 도 - 도2



들에 대한 중요도 점수를 산출하는 단계; 상기 형태소 분석된 과학문서에서 핵심어구 후보들 및 추출된 핵심어구 후보들의 위치정보들을 검출하는 단계; 상기 핵심어구 후보들 각각에 대해 핵심어구 후보에 포함된 단어들의 점수와 핵심어구 후보의 길이에 따라 후보 핵심어구의 점수를 산출하는 단계; 상기 핵심어구 후보들 각각의 점수를 핵심어구 후보의 위치정보에 따라 변경하여 핵심어구 후보들의 순위를 재순위화하는 단계; 및 재순위화한 핵심어구 후보들 중 미리 정해둔 수의 상위 핵심어구 후보들을 과학문서의 핵심어로 결정하는 단계;를 포함하는 것을 특징으로 한다.

(52) CPC특허분류
G06F 17/2755 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711055377
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터
연구사업명 SW중심대학 지원사업
연구과제명 SW중심대학(서강대학교)
기 여 율 1/1
주관기관 서강대학교 산학협력단
연구기간 2017.03.01 ~ 2017.12.31

명 세 서

청구범위

청구항 1

과학문서의 핵심어구 추출방법에 있어서,

과학문서를 입력받아 형태소 분석을 통해 명사와 형용사, 동사로 분석된 단어로 구성되는 형태소 분석된 과학문서로 변환하는 단계;

상기 과학문서의 단어를 정점으로 하고, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 문서 그래프를 구축하는 단계;

상기 문서 그래프내의 정점들에 대한 중요도 점수를 산출하는 단계;

상기 형태소 분석된 과학문서에서 핵심어구 후보들 및 추출된 핵심어구 후보들의 위치정보들을 검출하는 단계;

상기 핵심어구 후보들 각각에 대해 핵심어구 후보에 포함된 단어들의 점수와 핵심어구 후보의 길이에 따라 후보 핵심어구의 점수를 산출하는 단계;

상기 핵심어구 후보들 각각의 점수를 핵심어구 후보의 위치에 따라 변경하여 핵심어구 후보들의 순위를 재순위화하는 단계; 및

재순위화한 핵심어구 후보들 중 미리 정해둔 수의 상위 핵심어구 후보들을 과학문서의 핵심어구로 결정하는 단계;를 포함하는 것을 특징으로 하는 과학문서의 핵심어구 추출방법.

청구항 2

제1항에 있어서,

상기 핵심어구 후보의 점수는 수학식 6에 따라 산출됨을 특징으로 하는 과학문서의 핵심어구 추출방법.

수학식 6

$$\text{핵심어구 후보의 점수}(a) = \frac{\text{핵심어구 길이}}{\sum_{i \in a} score(\text{핵심어구 후보에 속한 단어}_i)} * \log_2(\text{핵심어구 길이})$$

$$Score(\text{정점}_i) = \frac{1-d}{\text{정점개수}} + d * \sum_{\text{정점}_j \in \text{정점}_i\text{의주변정점들}} \left(\frac{\text{가중치}(\text{정점}_i, \text{정점}_j)}{\sum_{\text{정점}_k \in \text{정점}_j\text{의주변정점들}} \text{가중치}(\text{정점}_j, \text{정점}_k)} * Score(\text{정점}_j) \right)$$

$$\text{가중치}(\text{정점}_1, \text{정점}_2) = 1 - \left(\frac{\text{동시발생횟수}(\text{정점}_1, \text{정점}_2)}{\text{빈도수}(\text{정점}_1) * \text{빈도수}(\text{정점}_2)} - \text{가중치들의평균값} \right)$$

상기 수학식 6에서, 상기 Score(정점i)는 과학문서의 단어i의 중요도 점수이고, 상기 d는 댐핑 펙터(damping factor)이고, 상기 j는 정점i와 연결된 주변의 단어들의 식별번호이며, 상기 i는 계산하고자 하는 단어의 식별번호이며, 상기 k는 정점j와 연결된 단어들의 식별번호이며, 상기 가중치(정점1,정점2)는 두 개의 정점인 정점1과 정점2 사이를 연결하는 간선의 가중치이고, 상기 동시발생횟수(정점1,정점2)는 정점1,정점2가 미리 정해진 윈도우내에 존재하는 횟수를 나타내며, 빈도수(정점1),빈도수(정점2)는 과학문서에서 정점1,정점2가 나타내는 횟수를 나타냄.

청구항 3

제1항에 있어서,

상기 핵심어구 후보들의 순위는 수학식 7에 따라 산출됨을 특징으로 하는 과학문서의 핵심어구 추출방법.

수학식 7

$$\text{핵심어구 후보들의 재순위화를 위해 보정된 핵심어구 후보의 점수} =$$

$$\begin{cases} \log_2(\text{핵심어구 후보 길이}) * \text{핵심어구 후보의 점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합이 없을때} \\ \\ \log_2(\text{핵심어구 후보 길이}) * (\text{핵심어구 후보점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a) \\ -\frac{1}{N} * \sum_{i \in b}(\text{핵심어구 후보점수}(i) * \text{위치함수 결과값}(i) * \text{빈도수}(i)), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합 } b\text{가 있을때} \end{cases}$$

$$\text{위치함수}(i) = \frac{\text{핵심어구 후보 } i\text{가 나타난 위치구간에서 실제 핵심어구가 나타난 개수}}{\text{총 핵심어구의 개수}}$$

상기 수학식 7에서, a는 핵심어구 후보이고, b는 핵심어구 후보 a를 포함하는 더 큰 길이의 핵심어구 후보 집합이고, i는 b의 리스트 원소들이고, N은 리스트 b의 크기이고, 위치함수(i)는 과학문서 말뭉치를 이용하여 실제로 핵심어구들이 나타난 위치의 확률 분포에 따라 현재 핵심어구의 위치가 핵심어구일 확률을 산출하는 것임.

청구항 4

과학문서의 핵심어구 추출장치에 있어서,

사용자와의 인터페이스를 담당하는 사용자 인터페이스;

과학문서의 핵심어 추출을 위한 저장영역을 제공하는 메모리부; 및

과학문서를 입력받아 형태소 분석을 통해 명사와 형용사, 동사로 분석된 단어로 구성되는 형태소 분석된 과학문서로 변환하고,

상기 과학문서의 단어를 정점으로 하고, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 문서 그래프를 구축하고,

상기 문서 그래프내의 정점들에 대한 중요도 점수를 산출하고,

상기 형태소 분석된 과학문서에서 핵심어구 후보들 및 추출된 핵심어구 후보들의 위치정보들을 검출하고,

상기 핵심어구 후보들 각각에 대해 핵심어구 후보에 포함된 단어들의 점수와 핵심어구 후보의 길이에 따라 후보 핵심어구의 점수를 산출하고,

상기 핵심어구 후보들 각각의 점수를 핵심어구 후보의 위치정보에 따라 변경하여 핵심어구 후보들의 순위를 재순위화하고,

재순위화한 핵심어구 후보들 중 미리 정해둔 수의 상위 핵심어구 후보들을 과학문서의 핵심어로 결정하는 제어장치;를 포함하는 것을 특징으로 하는 과학문서의 핵심어구 추출장치.

청구항 5

제4항에 있어서,

상기 핵심어구의 점수들은 수학식 8에 따라 산출됨을 특징으로 하는 과학문서의 핵심어구 추출방법.

수학식 8

$$\text{핵심어구 후보의 점수}(a) = \frac{\text{핵심어구 길이}}{\sum_{i \in a} score(\text{핵심어구 후보에 속한 단어}_i)} * \log_2(\text{핵심어구 길이})$$

$$Score(\text{정점}_i) = \frac{1-d}{\text{정점개수}} + d * \sum_{\text{정점}_j \in \text{정점}_i\text{의주변정점들}} \left(\frac{\text{가중치}(\text{정점}_i, \text{정점}_j)}{\sum_{\text{정점}_k \in \text{정점}_j\text{의주변정점들}} \text{가중치}(\text{정점}_j, \text{정점}_k)} * Score(\text{정점}_j) \right)$$

$$\text{가중치}(\text{정점}_1, \text{정점}_2) = 1 - \left(\frac{\text{동시발생횟수}(\text{정점}_1, \text{정점}_2)}{\text{빈도수}(\text{정점}_1) * \text{빈도수}(\text{정점}_2)} - \text{가중치들의평균값} \right)$$

상기 수학식 8에서, 상기 Score(정점i)는 과학문서의 단어i의 중요도 점수이고, 상기 d는 댐핑 펙터(damping factor)이고, 상기 j는 정점i와 연결된 주변의 단어들의 식별번호이며, 상기 i는 계산하고자 하는 단어의 식별번호이며, 상기 k는 정점j와 연결된 단어들의 식별번호이며, 상기 가중치(정점1,정점2)는 두 개의 정점인 정점1과 정점2 사이를 연결하는 간선의 가중치이고, 상기 동시발생횟수(정점1,정점2)는 정점1,정점2가 미리 정해진 윈도우내에 존재하는 횟수를 나타내며, 빈도수(정점1),빈도수(정점2)는 과학문서에서 정점1,정점2가 나타내는 횟수를 나타냄.

청구항 6

제1항에 있어서,

상기 핵심어구 후보들의 순위는 수학식 9에 따라 산출됨을 특징으로 하는 과학문서의 핵심어구 추출방법.

수학식 9

$$\text{핵심어구 후보들의 재순위화를 위해 보정된 핵심어구 후보의 점수} = \begin{cases} \log_2(\text{핵심어구 후보 길이}) * \text{핵심어구 후보의 점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합이 없을때} \\ \log_2(\text{핵심어구 후보 길이}) * (\text{핵심어구 후보점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a) \\ -\frac{1}{N} * \sum_{i \in b} (\text{핵심어구 후보점수}(i) * \text{위치함수 결과값}(i) * \text{빈도수}(i)), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합 } b\text{가 있을때} \end{cases}$$

$$\text{위치함수}(i) = \frac{\text{핵심어구 후보 } i\text{가 나타난 위치구간에서 실제 핵심어구가 나타난 개수}}{\text{총 핵심어구의 개수}}$$

상기 수학식 9에서, a는 핵심어구 후보이고, b는 핵심어구 후보 a를 포함하는 더 큰 길이의 핵심어구 후보 집합이고, i는 b의 리스트 원소들이고, N은 리스트 b의 크기이고, 위치함수(i)는 과학문서 말뭉치를 이용하여 실제로 핵심어구들이 나타난 위치의 확률 분포에 따라 현재 핵심어구의 위치가 핵심어구일 확률을 산출하는 것임.

발명의 설명

기 술 분 야

[0001] 본 발명은 과학문서의 핵심어구 추출 기술에 관한 것으로, 더욱 상세하게는 과학문서의 핵심어구의 추출시에 핵심어구의 위치를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있는 과학문서의 핵심어구 추출방

법 및 장치에 관한 것이다.

배 경 기 술

[0002] 인터넷 등의 발전으로 정보가 대량으로 증가하고 있는 상황에서 문서 정보의 효율적인 관리 및 검색은 중요한 문제로 대두되었다. 이러한 문제에 대해 효과적인 해결책을 제시해줄 수 있는 방법으로 문서 요약이 요구된다.

[0003] 상기 문서 요약이란 문서의 기본적인 내용을 유지하면서 원문으로부터 가장 의미있는 내용만을 추출하여 문서의 길이를 줄이는 작업을 의미한다. 상기 문서 요약의 방식은 크게 두 가지로, 추출(extract)과 추상(abstract)이 있다.

[0004] 상기 추출(extract) 방식은 주어진 문서에서 중요한 문장 또는 구를 그대로 추출하는 방식으로, 그래프 구조를 이용하여 주요 문장을 추출하는 Rada Mihalcea와 Paul Tarau가 2005년에 발표한 논문 “A Language Independent Algorithm for Single and Multiple Document Summarization“과 문장의 문서내에서의 위치와 문장이 포함하고 있는 단어의 빈도수 등을 자질로 이용하여 주요 문장을 추출하는 Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, 그리고 Daniel Tam의 2004년 논문 ”Centroid-based summarization of multiple documents. Information Processing and Management “ 등이 있다.

[0005] 그리고 상기 추상(abstract) 방식은 문서의 일부분을 추출하지 않고, 문서의 내용으로부터 요약문을 완전히 새로이 작성해내는 방식이다. 이러한 추상 방식의 요약은 아직 현재의 기술 수준이 많이 부족하여 만족할 만한 수준의 결과를 만들어내기 힘들다.

[0006] 좀 더 설명하면, Mihalcea, R., & Tarau, P.(2004, July). TextRank: Bringing order into texts. Association for Computational Linguistics는 그래프 기반 랭킹 알고리즘으로 개별 문서들에서 중요한 핵심어를 추출하는 방법을 제시한다. 이는 그래프의 정점을 구성할 문서의 구성 단위들, 예를들어 문서의 명사, 형용사 단어들을 정점으로 구성하고, 문서의 구성단위들의 관계를 정의하여 정점들끼리 이어주며, 이어진 간선들은 방향성을 띄는 간선, 방향성을 띄지 않는 간선, 가중된 간선, 가중되지 않은 간선일 수 있다. 그 이후 그래프 기반 랭킹 알고리즘(PageRank)를 이용하여 정점의 중요도 점수를 계산하고, 최종 점수를 토대로 정렬한 후 상위 N개를 최종 핵심어로 추출하는 것이다.

[0007] 그리고 Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In AAAI (Vol. 8, pp. 855-860).는 Mihalcea and Tarau(2004)의 모델을 좀 더 발전시킨 모델이다. 전체 과정은 이전 모델과 같지만 다른 점이 2가지가 있다. 첫째로는 그래프의 간선들에 가중치를 주었다. 두 문서의 구성 단위(단어)들이 특정 길이 내에서 같이 발생한 횟수를 가중치로 주는 것이고, 둘째로는 문서에서 먼저 핵심어가 될 수 있는 후보들을 추출하여 그래프 랭킹 알고리즘 결과 점수들을 더하여 핵심어의 최종 점수를 계산하고, 최종 계산된 점수를 정렬하여 상위 N개를 최종 핵심어로 추출하는 것이다.

[0008] 그리고 Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries, 3(2), 115-130.는 의학문서에서 다단어(Multi-word)를 추출하기 위해 고안된 알고리즘이다. 이는 전체 문서의 정보를 이용하며 다음의 수학식 1을 통해 입력 문서에서 다단어를 추출하고, 모든 후보 다단어들의 점수를 계산한 후 상위 N개의 후보 다단어들을 문서에서 나타난 최종 다단어들로 추출한다.

수학식 1

[0009]
$$C\text{-}Value = \begin{cases} \log_2|a| * f(a) \\ \log_2|a| \left(f(a) - 1/P(T_a) \sum_{b \in T_a} f(b)\right) \end{cases}$$

[0010] 상기 수학식 1에서 a는 후보 다단어들이며 T_a는 후보 다단어 a를 포함하는 더 큰 후보 다단어들의 집합이다. f(a)는 후보 다단어 a가 말뭉치에서 나타난 빈도수이며 P(Ta)는 집합 Ta의 크기이다.

[0011] 상기한 종래 기술들의 문제점을 설명한다.

[0012] 먼저 Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. Association for Computational Linguistics는 그래프 상에서 정점(문서의 구성단위)과 정점을 잇는 간선(관계)의 가중치 값을 일정하게 하였고, 핵심어를 추출할 때 그래프 기반 랭킹 알고리즘으로 추출된 상위 후보 핵심어들로만 구성 및 추출 가능한 한계가 있었다.

[0013] 그리고 Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In AAAI (Vol. 8, pp. 855-860)는 그래프 상에서 간선의 가중치 값을 같이 발생한 빈도수로 주었지만 이 빈도수를 계산할 때 중복 계산의 문제점이 있다. 예를들어, “한국 정보 과학회” 단어와 “정보 과학회” 단어가 존재한다면 “정보”와 “과학회”는 두 후보 모두에서 빈도수가 세어지게 된다. 그리고 후보 핵심어들의 점수를 계산할 때 구성 단어들의 점수들의 합으로 계산하므로, 후보 핵심어의 길이가 길어질수록 점수가 높아지는 문제점이 있다. “엄격한 대한민국 헌법”, “대한민국 헌법”의 두 후보 핵심어가 있을 때 합산으로 점수를 책정하게 된다면 무조건 앞의 “엄격한 대한민국 헌법”이 점수가 더 높게 책정되는 문제점이 있었다.

[0014] 그리고 Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries, 3(2), 115-130은 단어를 추출하기 위해 같은 분야의 많은 문서들을 요구하고, 빈도수 정보만 사용하므로 중요도를 판단하기 어려운 문제가 있었다.

선행기술문헌

특허문헌

[0015] (특허문헌 0001) 대한한국특허등록 제101664711호

(특허문헌 0002) 대한한국특허공개 제1020100128007호

(특허문헌 0003) 대한민국특허등록 제1015365200000호

발명의 내용

해결하려는 과제

[0016] 본 발명은 과학문서의 핵심어구의 추출시에 핵심어구의 위치를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있는 과학문서의 핵심어구 추출방법 및 장치를 제공하는 것을 그 목적으로 한다.

[0017] 또한 본 발명의 다른 목적은 핵심어구의 추출시에 핵심어구의 길이를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있는 과학문서의 핵심어구 추출방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0018] 상기한 목적을 달성하기 위한 본 발명에 따르는 과학문서의 핵심어 추출방법은, 과학문서를 입력받아 형태소 분석을 통해 명사와 형용사, 동사로 분석된 단어로 구성되는 형태소 분석된 과학문서로 변환하는 단계; 상기 과학문서의 단어를 정점으로 하고, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 문서 그래프를 구축하는 단계; 상기 문서 그래프내의 정점들에 대한 중요도 점수를 산출하는 단계; 상기 형태소 분석된 과학문서에서 핵심어구 후보들 및 추출된 핵심어구 후보들의 위치정보들을 검출하는 단계; 상기 핵심어구 후보들 각각에 대해 핵심어구 후보에 포함된 단어들의 점수와 핵심어구 후보의 길이에 따라 후보 핵심어구의 점수를 산출하는 단계; 상기 핵심어구 후보들 각각의 점수를 핵심어구 후보의 위치정보에 따라 변경하여 핵심어구 후보들의 순위를 재순위화하는 단계; 및 재순위화한 핵심어구 후보들 중 미리 정해둔 수의 상위 핵심어구 후보들을 과학문서의 핵심어로 결정하는 단계;를 포함하는 것을 특징으로 한다.

발명의 효과

[0019] 상기한 본 발명은 과학문서의 핵심어구의 추출시에 핵심어구의 위치를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있는 효과를 야기한다.

[0020] 또한 본 발명은 핵심어구의 추출시에 핵심어구의 길이를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높

일 수 있는 효과를 야기한다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 바람직한 실시예에 따르는 과학문서의 핵심어구 추출장치의 구성도.

도 2 및 도 3은 본 발명의 바람직한 실시예에 따르는 과학문서의 핵심어구 추출방법의 흐름도.

발명을 실시하기 위한 구체적인 내용

[0022] 본 발명은 과학문서의 핵심어구의 추출시에 핵심어구의 위치를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있다.

[0023] 또한 본 발명은 핵심어구의 추출시에 핵심어구의 길이를 반영하여 핵심어구를 결정하여 핵심어구의 신뢰도를 높일 수 있다.

[0024] 이러한 본 발명의 바람직한 실시예를 도면을 참조하여 상세히 설명한다.

[0025] <과학문서의 핵심어구 추출장치의 구성>

[0026] 도 1은 본 발명의 바람직한 실시예에 따르는 과학문서의 핵심어구 추출장치의 구성을 도시한 것이다. 상기 도 1을 참조하면 본 발명의 바람직한 실시예에 따르는 과학문서의 핵심어 추출장치는 제어장치(100)와 메모리부(102)와 사용자 인터페이스부(104)와 디스플레이부(106)와 통신부(108)로 구성된다.

[0027] 상기 제어장치(100)는 본 발명의 바람직한 실시예에 따라 과학문서의 핵심어 추출을 이행한다. 좀더 설명하면 제어장치(100)는 사용자 인터페이스(104) 또는 통신부(108)를 통해 과학문서를 입력받아 형태소 분석을 통해 명사와 형용사, 동사로 분석된 단어로 구성되는 형태소 분석된 과학문서로 변환하고, 상기 과학문서의 단어를 정점으로 하고, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 문서 그래프를 구축하고, 상기 문서 그래프 내의 정점들에 대한 중요도 점수를 산출하고, 상기 형태소 분석된 과학문서에서 핵심어구 후보들 및 추출된 핵심어구 후보들의 위치정보들을 검출하고, 상기 핵심어구 후보들 각각에 대해 핵심어구 후보에 포함된 단어들의 점수와 핵심어구 후보의 길이에 따라 후보 핵심어구의 점수를 산출하고, 상기 핵심어구 후보들 각각의 점수를 핵심어구 후보의 위치정보에 따라 변경하여 핵심어구 후보들의 순위를 재순위화하고, 재순위화한 핵심어구 후보들 중 미리 정해둔 수의 상위 핵심어구 후보들을 과학문서의 핵심어로 결정한다.

[0028] 또한 상기 제어장치(100)는 다수의 말뭉치들로부터 상기 핵심어구 후보들의 재순위화를 위한 핵심어구 위치정보를 추출하여 저장한다. 이는 사용자 인터페이스(102)를 통해 핵심어구의 위치를 선정받아 결정될 수도 있다.

[0029] 상기 메모리부(102)는 본 발명의 바람직한 실시예에 따라 과학문서의 핵심어구 추출을 위한 각종 정보를 저장하기 위한 저장영역을 제공한다.

[0030] 상기 사용자 인터페이스부(104)는 사용자와 상기 제어장치(100) 사이의 인터페이스를 담당한다.

[0031] 상기 디스플레이부(106)는 상기 제어장치(100)의 제어에 따른 각종 정보를 저장한다.

[0032] 상기 통신부(108)는 외부의 네트워크 또는 외부 기기 등과 상기 제어장치(100) 사이의 통신을 담당한다.

[0033] <과학문서의 핵심어구 추출방법의 절차>

[0034] 이제 상기한 과학문서의 핵심어구 추출장치에 적용 가능한 본 발명에 따르는 과학문서 핵심어구 추출방법의 절차를 도 2를 참조하여 설명한다.

[0035] 상기 제어장치(100)는 사용자 인터페이스부(104) 또는 통신부(108)를 통해 과학문서가 입력되면(200단계), 전처리를 이행한다(202단계). 상기 전처리는 과학문서에 대해 형태소 분석을 통해 형태소 분석된 과학문서로 변환하는 것이며, 상기 형태소 분석된 과학문서는 명사와 형용사, 동사로 분석된 단어들로 구성된다.

[0036] 상기 제어장치(100)는 상기 형태소 분석된 과학문서에 포함된 명사와 형용사, 동사로 분석된 단어들을 이용하여 문서 그래프를 구축한다(204단계). 여기서, 상기 문서 그래프는 명사와 형용사, 동사로 구성된 단어들을 문서의 구성단위로 보고, 이 구성단위들을 정점으로 함과 아울러, 단어와 단어 사이의 관계를 나타내는 간선을 부여하여 구축한 것이다.

[0037] 상기 문서 그래프의 구축이 완료되면, 상기 제어장치(100)는 상기 문서 그래프의 간선들에 대해 가중치를 수학식 2에 따라 산출하여 부여한다(206단계). 여기서, 상기 가중치는 두 단어가 동시에 발생한 빈도수를 기준으로

확률론적으로 계산된다.

수학식 2

$$\text{가중치}(\text{정점}_1, \text{정점}_2) = 1 - \left(\frac{\text{동시발생횟수}(\text{정점}_1, \text{정점}_2)}{\text{빈도수}(\text{정점}_1) * \text{빈도수}(\text{정점}_2)} - \text{가중치들의평균값}\right)$$

[0038]

[0039] 상기 수학식 2는 두 개의 정점인 정점1과 정점2 사이를 연결하는 간선의 가중치를 산출하기 위한 것이다. 상기 동시발생횟수는 정점1과 정점2가 사용자 또는 개발자에 의해 설정된 윈도우 사이즈의 윈도우(앞뒤 단어 간격)내에 존재한다면 동시에 발생한 것으로 판단하여 카운트한 값이다. 상기 빈도수는 입력 문서에 대한 빈도수를 말한다. 상기 가중치들의 평균값은 가중치의 평균값을 고려하지 않은 채 산출된 가중치들 모두를 계산한 후에 평균을 취하여 구한 것이다.

[0040] 이후 상기 제어장치(100)는 그래프 기반 랭킹 알고리즘을 통해 과학 문서의 구성단위인 단어들, 즉 정점들 각각에 대한 중요도 점수를 계산한다(210단계). 여기서, 상기 그래프 기반 랭킹 알고리즘에 따르는 정점의 중요도 점수는 수학식 3에 따라 산출하여 부여한다.

수학식 3

$$Score(\text{정점}_i) = \frac{1-d}{\text{정점개수}} + d * \sum_{\text{정점}_j \in \text{정점}_i\text{의주변정점들}} \left(\frac{\text{가중치}(\text{정점}_i, \text{정점}_j)}{\sum_{\text{정점}_k \in \text{정점}_j\text{의주변정점들}} \text{가중치}(\text{정점}_j, \text{정점}_k)} * Score(\text{정점}_j)\right)$$

[0041]

[0042] 상기 수학식 3에서 d는 댐핑 펙터(damping factor)로써 0~1사이의 확률값으로 정하며, 이는 하이퍼 파라메터(hyperparameter)로써 사용자 또는 개발자에 의해 정해진다. 상기 j는 정점i와 연결된 주변의 단어들의 식별번호이며, 정점j로 사용된다. 상기 i는 계산하고자 하는 단어의 식별번호이며, 정점i로 사용된다. 상기 k는 정점j와 연결된 단어들의 식별번호이며, 정점k로 사용된다. 그리고 상기 주변 정점들은 어떤 한 단어를 현재 정점으로 보았을 때 그 기준으로부터 미리 정해진 윈도우 사이즈(window size)의 윈도우 내에 단어들이 존재한다면 그 단어들에 해당되는 정점과 현재 정점은 주변 정점의 관계를 가지는 것으로 판단한다.

[0043] 또한 상기 제어장치(100)는 형태소 분석된 과학문서에서 핵심어구 후보들을 추출하고, 상기 추출된 핵심어구 후보들의 위치정보를 검출하여 저장한다(214,216단계). 여기서 상기 핵심어구 후보 추출방법은 N-gram 기반 추출방법 또는 형태소를 이용한 규칙 기반의 후보 추출방법이 사용될 수 있다. 이때 핵심어구 후보들은 명사구들이며, 상기 위치정보는 해당 후보 핵심어들이 과학문서의 어느 위치인지를 나타내는 정보이다.

[0044] 여기서, 본 발명은 과학문서의 초반부에는 전체 문서에 대한 요약 및 핵심내용이 존재하는 특징을 이용하므로, 상기 과학문서를 위치별로 다수 구간으로 미리 분할하고, 해당 핵심어구 후보들이 과학문서의 어느 구간에 위치하지에 따라 핵심어구를 판별한다. 이를 위해 과학문서는 10개 단어마다 한 구간으로 나누어질 수 있다.

[0045] 상기한 바와 같이 핵심어구 후보들이 추출되면, 상기 제어장치(100)는 추출된 핵심어구 후보들을 구성하는 각 단어에 대해 각 단어들의 중요도 점수를 이용하여 수학식 4에 따라 핵심어구 후보들의 점수를 산출한다(212단계).

수학식 4

$$\text{핵심어구 후보의 점수}(a) = \frac{\text{핵심어구 길이}}{\sum^{i \in a} score(\text{핵심어구 후보에 속한 단어}_i)} * \log_2(\text{핵심어구 길이})$$

[0046]

[0047] 상기 수학식 4에서 score(핵심어구 후보에 속한 단어i)는 해당 단어의 중요도 점수이며, 이는 수학식 3에 따라 산출된다.

[0048] 그리고 상기 수학식 4에 따르는 핵심어구 후보의 점수는 단어들의 변형된 조화 평균을 이용하여 계산한 것이다. 이와 같이 변형된 조화 평균으로 점수를 계산하는 이유는 후보 핵심어구의 점수가 길이에 의존하는 것을 방지함과 동시에 길이 정보가 손실되는 것을 방지하기 위함이다.

[0049] 상기한 바와 같이 산출된 핵심어구 후보들의 점수들은, 그래프 구성시 단어와 단어 사이의 간선에 정보가 중첩되어 반영되는 문제가 있다. 예를들어 설명하면, "정보 과학회"와 "한국 정보 과학회"라는 후보 핵심어구들이 문서에서 나타난다면, "정보 과학회"의 빈도수를 측정할 때 "한국 정보 과학회"에 포함된 "정보 과학회"까지 측정하므로 빈도수 정보가 편향되는 경우가 있다.

[0050] 이에 제어장치(100)는 핵심어구 후보들에 대한 점수가 산출되면, 그래프의 간선에 정보가 중첩되는 문제를 해소하기 위해 수학식 5에 따라 점수를 보정하여 핵심어구 후보들을 재순위화한다(218단계).

수학식 5

$$\text{핵심어구 후보들의 재순위화를 위해 보정된 핵심어구 후보의 점수} = \begin{cases} \log_2(\text{핵심어구 후보 길이}) * \text{핵심어구 후보의 점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합이 없을때} \\ \log_2(\text{핵심어구 후보 길이}) * (\text{핵심어구 후보점수}(a) * \text{위치함수 결과값}(a) * \text{빈도수}(a) \\ -\frac{1}{N} * \sum_{i \in b}(\text{핵심어구 후보점수}(i) * \text{위치함수 결과값}(i) * \text{빈도수}(i)), \\ \text{if 핵심어구 } a\text{를 포함하는 핵심어구 후보 집합}b\text{가 있을때} \end{cases}$$

$$\text{위치함수}(i) = \frac{\text{핵심어구 후보}i\text{가 나타난 위치구간에서 실제 핵심어구가 나타난 개수}}{\text{총 핵심어구의 개수}}$$

[0051]

[0052] 상기 수학식 5는 재순위화를 위해 핵심어구 후보에 대한 점수를 보정하기 위한 수학식이며, a는 핵심어구 후보이고, b는 핵심어구 후보 a를 포함하는 더 큰 길이의 핵심어구 후보 집합이고, i는 b의 리스트 원소들(핵심어구 후보 a를 포함하는 더 큰 길이의 핵심어구 후보)이고, N은 리스트 b의 크기이다. 그리고 위치함수(i)는 과학문서 말뭉치를 이용하여 실제로 핵심어구들이 나타난 위치의 확률 분포에 따라 현재 핵심어구의 위치가 핵심어구일 확률을 산출하는 것이다.

[0053] 좀더 설명하면, 제어장치(100)는 대규모 과학문서 말뭉치를 다수개의 위치구간으로 분할하고, 위치구간별 실제 핵심어구가 발생할 확률을 구한다. 즉, 대규모 과학 문서 말뭉치에서 발생한 핵심어구들의 총 개수를 각 위치구간마다 실제 핵심어구가 발생한 개수로 나누어 각 위치구간별 핵심어구가 발생할 확률을 구하여 미리 저장해둘 수 있다.

[0054] 이 상태에서 제어장치(100)는 위치함수를 통해 과학문서가 입력되면, 입력된 과학문서에서 핵심어구 후보가 처음 발생한 위치를 기준으로 사전에 계산된 위치구간별 확률값을 독출하여 위치함수 결과값으로 출력하여 상기 수학식 5에 따르는 핵심어구 점수에 반영할 수 있다.

[0055] 즉, 본 발명은 그래프의 간선에 의해 중복된 정보를 없애기 위해 재순위화(Re-Ranking) 과정을 거치며, 이는 중

복된 빈도수 정보가 있기 때문이므로, 이를 해결하기 위해 Frantzi et al., (2000)에서 사용한 C-Value 수식을 변형하여 중복된 빈도수 정보를 해결한다. 이때 핵심어구 후보들의 위치까지도 포함하여 계산하게 된다.

[0056] 이후 제어장치(100)는 재순위화 과정까지 거친 핵심어구 후보들 중 최종 상위 N개의 핵심어구 후보를 최종 핵심어구로 결정한다(220단계).

[0057] 상기한 바와 같이 본 발명은 핵심어구 후보의 위치정보에 따라 핵심어구 후보의 점수를 보정하며 이를 위해 제어장치(100)는 도 3에 도시한 바와 같이 말뭉치들을 입력받아(300단계), 말뭉치에서의 핵심어구 위치정보를 미리 설정받아 저장한다(302단계). 이 핵심어구 위치정보는 과학문서내에서 핵심어구가 위치하는 위치를 지시하며, 이는 사용자 인터페이스를 통한 사용자 지정에 따를 수 있다.

부호의 설명

[0058] 100 : 제어장치

102 : 메모리부

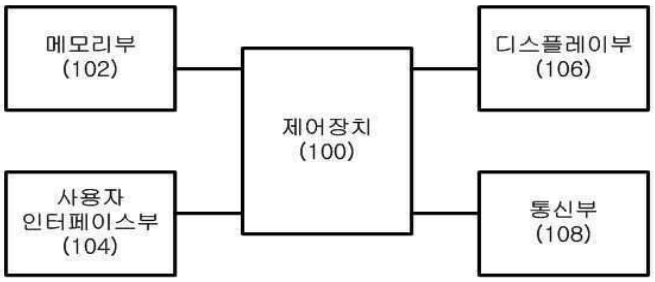
104 : 사용자 인터페이스부

106 : 디스플레이부

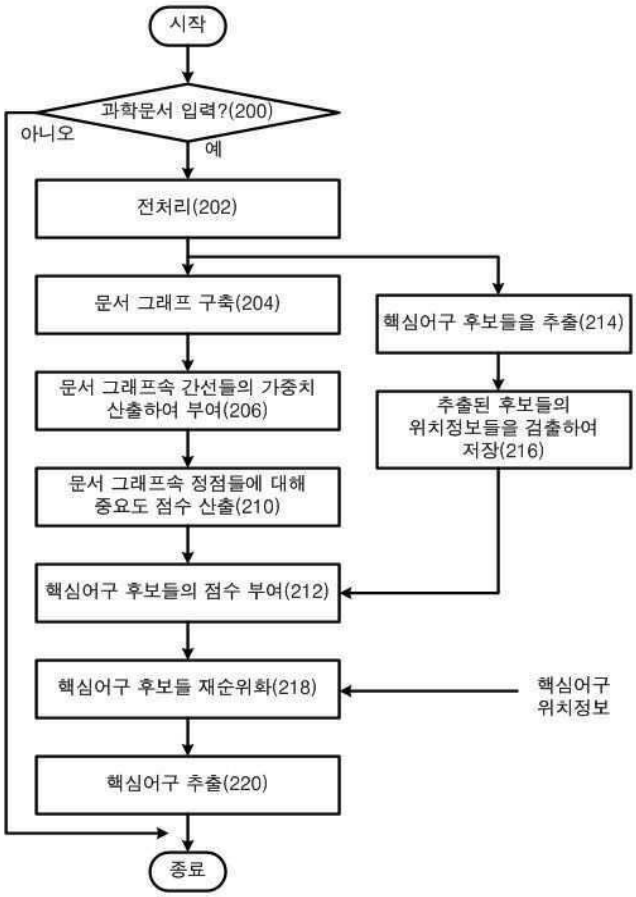
108 : 통신부

도면

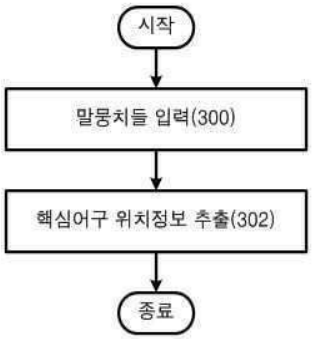
도면1



도면2


시작
과학문서 입력?(200)
아니오
예
전처리(202)
문서 그래프 구축(204)
문서 그래프속 간선들의 가중치 산출하여 부여(206)
문서 그래프속 정점들에 대해 중요도 점수 산출(210)
핵심어구 후보들을 추출(214)
추출된 후보들의 위치정보들을 검출하여 저장(216)
핵심어구 후보들의 점수 부여(212)
핵심어구 후보들 재순위화(218)
핵심어구 위치정보
핵심어구 추출(220)
종료

도면3


시작
말뭉치들 입력(300)
핵심어구 위치정보 추출(302)
종료