(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년05월13일
(11) 등록번호 10-2110523
(24) 등록일자 2020년05월07일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01)
(52) CPC특허분류
G06F 40/289 (2020.01)
G06F 16/3344 (2019.01)
(21) 출원번호 10-2018-0115849
(22) 출원일자 2018년09월28일
심사청구일자 2018년09월28일
(65) 공개번호 10-2020-0036333
(43) 공개일자 2020년04월07일
(56) 선행기술조사문헌
KR1020110046098 A*
(뒷면에 계속)
전체 청구항 수 : 총 5 항

(73) 특허권자
배재대학교 산학협력단
대전광역시 서구 배재로 155-40 (도마동)
(72) 발명자
정희경
대전광역시 서구 둔산로 155, 112동 1303호(둔산동, 크로바아파트)
이종원
서울특별시 광진구 능동대로 27나길7호 202호
(74) 대리인
유병욱, 한승범

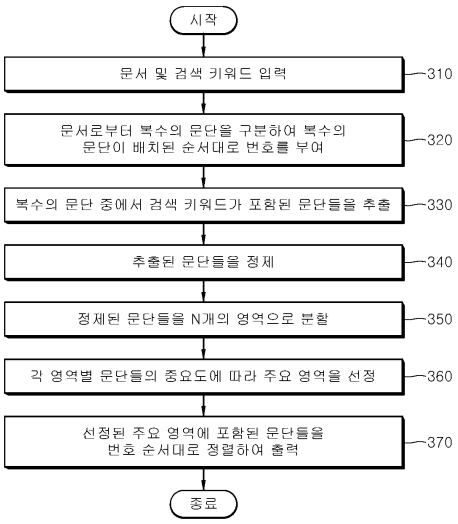
심사관 : 경연정

(54) 발명의 명칭 **문서 분석 기반 주요 요소 추출 시스템 및 방법**

(57) 요 약

본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템은 문서 및 검색 키워드를 입력받는 인터페이스부; 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출하는 추출부; 및 상기 추출된 문단들을 정제하고, 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정하고, 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력하는 프로세서를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06F 16/335 (2019.01)
G06F 40/30 (2020.01)

(56) 선행기술조사문헌
KR1020110050106 A*
KR1020160106984 A
KR1020070089449 A
KR1020190043857 A
KR1020070050305 A
KR1020100059516 A
*는 심사관에 의하여 인용된 문헌

명 세 서

청구범위

청구항 1

문서 및 검색 키워드를 입력받는 인터페이스부;

상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출하는 추출부; 및

상기 추출된 문단들 내 상기 검색 키워드의 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산하고, 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거하여 상기 추출된 문단들을 정제하고, 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 상기 N개의 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산하고, 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산하며, 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 주요 영역으로 선정하며, 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력하는 프로세서

를 포함하고,

상기 프로세서는

각각의 검색 키워드가 포함된 문단의 개수를 합하여 총 개수를 산출하고, 총 개수 대비 특정 검색 키워드가 포함된 문단의 개수에 대한 비율을, 상기 특정 검색 키워드에 대한 가중치로서 산출하며,

상기 정제된 문단들의 개수를 상기 N으로 나눈 값으로 상기 각 영역별 문단들의 개수를 정하되, 나머지 값이 발생하는 경우 상기 나머지 값을 맨 뒤에 배치된 영역에서부터 상기 나머지 값이 소진될 때까지 각 영역에 균등하게 순차적으로 가산하여 상기 각 영역별 문단들의 개수를 정하는 것을 특징으로 하는 문서 분석 기반 주요 요소 추출 시스템.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 프로세서는

하나의 문단 내에 상기 가중치가 가장 낮은 검색 키워드 이외의 다른 검색 키워드가 존재하는 경우에는 예외 처리하여 해당 문단의 제거 기능을 수행하지 않는 것을 특징으로 하는 문서 분석 기반 주요 요소 추출 시스템.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 인터페이스부는

상기 문서의 키워드 태그의 태그 값에 해당하는 키워드들을 불러온 뒤 해당 키워드들을 출력하여 사용자에게 보여주고, 상기 출력된 키워드들 중 상기 사용자에 의해 입력된 키워드를 상기 검색 키워드로서 입력받는 것을 특징으로 하는 문서 분석 기반 주요 요소 추출 시스템.

청구항 8

제1항에 있어서,

상기 추출부는

상기 검색 키워드가 복수 개일 경우, 상기 추출된 문단들 중 동일한 번호가 부여된 문단이 복수 개 존재하면 상기 복수의 문단 중 하나의 문단 이외의 나머지 문단을 상기 추출된 문단들에서 제거하는 것을 특징으로 하는 문서 분석 기반 주요 요소 추출 시스템.

청구항 9

문서 분석 기반 주요 요소 추출 시스템의 인터페이스부가 문서 및 검색 키워드를 입력받는 단계;

상기 문서 분석 기반 주요 요소 추출 시스템의 추출부가 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출하는 단계;

상기 문서 분석 기반 주요 요소 추출 시스템의 프로세서가 상기 추출된 문단들 내 상기 검색 키워드의 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산하는 단계;

상기 프로세서가 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거하여 상기 추출된 문단들을 정제하는 단계;

상기 프로세서가 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 상기 N개의 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산하고, 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산하는 단계;

상기 프로세서가 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 주요 영역으로 선정하는 단계; 및

상기 프로세서가 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력하는 단계

를 포함하고,

상기 가중치를 계산하는 단계는

각각의 검색 키워드가 포함된 문단의 개수를 합하여 총 개수를 산출하고, 총 개수 대비 특정 검색 키워드가 포함된 문단의 개수에 대한 비율을, 상기 특정 검색 키워드에 대한 가중치로서 산출하는 단계

를 포함하며,

상기 중요도를 계산하는 단계는

상기 정제된 문단들의 개수를 상기 N으로 나눈 값으로 상기 각 영역별 문단들의 개수를 정하되, 나머지 값이 발생하는 경우 상기 나머지 값을 맨 뒤에 배치된 영역에서부터 상기 나머지 값이 소진될 때까지 각 영역에 균등하게 순차적으로 가산하여 상기 각 영역별 문단들의 개수를 정하는 단계

를 포함하는 것을 특징으로 하는 문서 분석 기반 주요 요소 추출 방법.

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

발명의 설명

기 술 분 야

[0001] 본 발명의 실시예들은 XML 형태나 PDF 파일 형태의 논문이나 보고서로 작성된 문서를 분석하는 시스템에 관한 것으로, 더욱 상세하게는 문서의 주요 문단들을 추출하여 압축률을 향상시킴과 동시에, 추출된 문단들을 복수의 영역으로 분할하고 각 영역의 중요도를 계산하여 주요 영역을 알려줌으로써 문서의 이해도를 향상시켜 문서를 이해하는 데 필요한 시간을 줄일 수 있는 문서 분석 기반 주요 요소 추출 시스템 및 방법에 관한 것이다.

배 경 기 술

[0003] 다양한 종류의 문서 중에서도 보고서나 논문들은 일반적으로 XML 문서 형태나 PDF 파일 형태로 보관한다. 해당 문서들을 분석하기 위해 형태소 분석기를 기반으로 개발된 문서 분석 시스템들은 문서의 내용을 분석하여 문서 작성에 사용된 단어들과 단어들의 빈도수를 정렬하고 해당 결과를 사용자에게 보여준다. 그러나 사용자가 문서 작성에 사용된 주요한 단어들에 대해 시스템들을 통해 알 수 있다고 해서 해당 문서를 분석하는 시간이 문서 내용 전체를 읽는 것에 비해 단축되는 것은 아니다.

[0004] 이와 다른 방식의 문서 분석 시스템들은 해당 문서에 사용자가 입력한 검색어가 포함되어 있는지 여부를 판단한다. 그리고 검색어가 포함되어 있는 문서를 검색하여 정렬한 뒤 사용자에게 이를 보여주는 기능을 수행한다. 그러나 이러한 시스템의 기능이 사용자의 문서 이해에 대한 시간을 줄여주거나 문서 이해의 효율성을 높이지는 못한다.

[0005] 이에, 사용자가 XML 형태나 PDF 파일 형태의 보고서나 논문 등의 문서를 효율적으로 이해할 수 있도록 해당 문서의 내용을 압축하고 정렬하고 추출하는 시스템을 제안한다.

[0006] 관련 선행기술로는 대한민국 등록특허공보 제10-1060594호(발명의 명칭: 문서 데이터의 키워드 추출 및 연관어 네트워크 구성 장치 및 방법, 등록일자: 2011.08.24)가 있다.

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 실시예는 문서의 주요 문단들을 추출하여 압축률을 향상시킴과 동시에, 추출된 문단들을 복수의 영역으로 분할하고 각 영역의 중요도를 계산하여 주요 영역을 알려줌으로써 문서의 이해도를 향상시켜 문서를 이해하는 데 필요한 시간을 줄일 수 있는 문서 분석 기반 주요 요소 추출 시스템 및 방법을 제공한다.

[0010] 본 발명이 해결하고자 하는 과제는 이상에서 언급한 과제(들)로 제한되지 않으며, 언급되지 않은 또 다른 과제(들)은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0012] 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템은 문서 및 검색 키워드를 입력받는 인터페

이스부; 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출하는 추출부; 및 상기 추출된 문단들을 정제하고, 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정하고, 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력하는 프로세서를 포함한다.

[0013] 상기 프로세서는 상기 추출된 문단들 내 상기 검색 키워드의 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산하고, 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거하여 상기 추출된 문단들을 정제할 수 있다.

[0014] 상기 프로세서는 하나의 문단 내에 상기 가중치가 가장 낮은 검색 키워드 이외의 다른 검색 키워드가 존재하는 경우에는 예외 처리하여 해당 문단의 제거 기능을 수행하지 않는 것이 바람직하다.

[0015] 상기 프로세서는 상기 N개의 각 영역별 문단들의 중요도를 계산하고, 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 상기 주요 영역으로 선정할 수 있다.

[0016] 상기 프로세서는 상기 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산하고, 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산할 수 있다.

[0017] 상기 프로세서는 상기 정제된 문단들의 개수를 상기 N으로 나눈 값으로 상기 각 영역별 문단들의 개수를 정하되, 나머지 값이 발생하는 경우 상기 나머지 값을 맨 뒤에 배치된 영역에서부터 상기 나머지 값이 소진될 때까지 각 영역에 균등하게 순차적으로 가산하여 상기 각 영역별 문단들의 개수를 정하는 것이 바람직하다.

[0018] 상기 인터페이스부는 상기 문서의 키워드 태그의 태그 값에 해당하는 키워드들을 불러온 뒤 해당 키워드들을 출력하여 사용자에게 보여주고, 상기 출력된 키워드들 중 상기 사용자에 의해 입력된 키워드를 상기 검색 키워드로서 입력받을 수 있다.

[0019] 상기 추출부는 상기 검색 키워드가 복수 개일 경우, 상기 추출된 문단들 중 동일한 번호가 부여된 문단이 복수 개 존재하면 상기 복수의 문단 중 하나의 문단 이외의 나머지 문단을 상기 추출된 문단들에서 제거할 수 있다.

[0020] 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 방법은 문서 분석 기반 주요 요소 추출 시스템의 인터페이스부가 문서 및 검색 키워드를 입력받는 단계; 상기 문서 분석 기반 주요 요소 추출 시스템의 추출부가 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출하는 단계; 상기 문서 분석 기반 주요 요소 추출 시스템의 프로세서가 상기 추출된 문단들을 정제하는 단계; 상기 프로세서가 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정하는 단계; 및 상기 프로세서가 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력하는 단계를 포함한다.

[0021] 상기 추출된 문단들을 정제하는 단계는 상기 추출된 문단들 내 상기 검색 키워드의 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산하는 단계; 및 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거하여 상기 추출된 문단들을 정제하는 단계를 포함할 수 있다.

[0022] 상기 주요 영역을 선정하는 단계는 상기 N개의 각 영역별 문단들의 중요도를 계산하는 단계; 및 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 상기 주요 영역으로 선정하는 단계를 포함할 수 있다.

[0023] 상기 중요도를 계산하는 단계는 상기 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산하는 단계; 및 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산하는 단계를 포함할 수 있다.

[0024] 상기 중요도를 계산하는 단계는 상기 정제된 문단들의 개수를 상기 N으로 나눈 값으로 상기 각 영역별 문단들의 개수를 정하되, 나머지 값이 발생하는 경우 상기 나머지 값을 맨 뒤에 배치된 영역에서부터 상기 나머지 값이 소진될 때까지 각 영역에 균등하게 순차적으로 가산하여 상기 각 영역별 문단들의 개수를 정하는 단계를 더 포함할 수 있다.

[0026] 기타 실시예들의 구체적인 사항들은 상세한 설명 및 첨부 도면들에 포함되어 있다.

발명의 효과

[0028] 본 발명의 일 실시예에 따르면, 문서의 주요 문단들을 추출하여 압축률을 향상시킴과 동시에, 추출된 문단들을 복수의 영역으로 분할하고 각 영역의 중요도를 계산하여 주요 영역을 알려줌으로써 문서의 이해도를 향상시켜 문서를 이해하는 데 필요한 시간을 줄일 수 있다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템의 네트워크 구성을 도시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템의 상세 구성을 설명하기 위해 도시한 블록도이다.

도 3 내지 도 5는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 방법을 설명하기 위해 도시한 흐름도이다.

도 6 내지 도 15는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템을 구현하고 검증하는 일례를 설명하기 위해 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명의 이점 및/또는 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 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 구성요소를 지칭한다.

[0032] 또한, 이하 실시되는 본 발명의 바람직한 실시예는 본 발명을 이루는 기술적 구성요소를 효율적으로 설명하기 위해 각각의 시스템 기능구성에 기 구비되어 있거나, 또는 본 발명이 속하는 기술분야에서 통상적으로 구비되는 시스템 기능 구성은 가능한 생략하고, 본 발명을 위해 추가적으로 구비되어야 하는 기능 구성을 위주로 설명한다. 만약 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면, 하기에 도시하지 않고 생략된 기능 구성 중에서 종래에 기 사용되고 있는 구성요소의 기능을 용이하게 이해할 수 있을 것이며, 또한 상기와 같이 생략된 구성 요소와 본 발명을 위해 추가된 구성 요소 사이의 관계도 명백하게 이해할 수 있을 것이다.

[0033] 또한, 이하의 설명에 있어서, 신호 또는 정보의 "전송", "통신", "송신", "수신" 기타 이와 유사한 의미의 용어는 일 구성요소에서 다른 구성요소로 신호 또는 정보가 직접 전달되는 것뿐만이 아니라 다른 구성요소를 거쳐 전달되는 것도 포함한다. 특히 신호 또는 정보를 일 구성요소로 "전송" 또는 "송신"한다는 것은 그 신호 또는 정보의 최종 목적지를 지시하는 것이고 직접적인 목적지를 의미하는 것이 아니다. 이는 신호 또는 정보의 "수신"에 있어서도 동일하다.

[0035] 이하에서는 첨부된 도면을 참조하여 본 발명의 실시예들을 상세히 설명하기로 한다.

[0036] 도 1은 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템의 네트워크 구성을 도시한 도면이다.

[0037] 도 1을 참조하면, 단말(101)은 예컨대, 스마트폰, 태블릿 PC일 수 있으며, 유무선 통신을 통해 문서(예컨대, XML 문서)에 대한 분석 요청을 문서 분석 기반 주요 요소 추출 시스템(100)으로 전송할 수 있다.

[0038] 상기 단말(101)은 상기 분석 요청에 대한 응답으로서, 상기 문서 분석 기반 주요 요소 추출 시스템(100)으로부터 원래 문서의 압축본, 즉 압축된 문서를 수신하여 출력할 수 있다.

[0039] 상기 문서 분석 기반 주요 요소 추출 시스템(100)은 상기 단말(101)로부터의 문서에 대한 분석 요청에 연동하여, 상기 문서를 압축한 후 압축된 문서를 상기 단말(101)로 제공할 수 있다.

[0040] 이때, 상기 문서 분석 기반 주요 요소 추출 시스템(100)은 상기 분석 요청으로부터 검색 키워드를 추출하고, 상기 추출된 검색 키워드를 이용하여 상기 문서를 압축할 수 있다.

[0041] 구체적으로, 상기 문서 분석 기반 주요 요소 추출 시스템(100)은 상기 문서로부터 복수의 문단을 구분하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함되는 문단을 추출할 수 있다. 이어서, 상기 문서 분석 기반 주요 요소 추출 시스템(100)은 상기 추출된 문단들을 정제하고, 상기 정제된 문단들을 N(상기 N은 자연수)개의 영역으로 나눈 후, 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정하고, 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬함으로써, 상기 문서를 압축할 수 있다.

[0042] 한편, 상기 문서 분석 기반 주요 요소 추출 시스템(100)은 상기 단말(101)로부터 수신된 문서에 대한 분석 요청에 대해, 압축된 문서를 상기 단말(101)로 제공할 수 있으나, 이에 한정되지 않고, 사용자로부터 직접 문서를 입력받을 수 있으며, 입력된 문서에 대해, 압축된 문서를 출력할 수 있다.

[0044] 도 2는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템의 상세 구성을 설명하기 위해 도시한 블록도이다.

[0045] 도 2를 참조하면, 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템(100)은 인터페이스부(210), 추출부(220), 프로세서(230), 및 데이터베이스(240)를 포함할 수 있다.

[0046] 상기 인터페이스부(210)는 문서(예: XML 문서, PDF 문서 등) 및 검색 키워드를 입력받을 수 있다. 이때, 상기 인터페이스부(210)는 상기 문서의 키워드 태그(keyword tag)의 태그 값에 해당하는 키워드들을 불러온 뒤 해당 키워드들을 출력하여 사용자에게 보여주고, 상기 출력된 키워드들 중 상기 사용자에 의해 입력된 키워드를 상기 검색 키워드로서 입력받을 수 있다.

[0047] 상기 인터페이스부(210)는 상기 검색 키워드의 개수를 더 입력받거나, 또는 상기 사용자에 의해 입력된 키워드를 카운트하여, 상기 검색 키워드의 개수를 확인할 수 있다.

[0048] 상기 인터페이스부(210)는 상기 검색 키워드와 같은 의미를 갖는 유사 키워드(즉, 동의어)가 외부 서버(도시하지 않음)로부터 획득되는 경우, 상기 유사 키워드를 제공할 수 있다. 이로써, 사용자가 검색하고자 하는 내용이, 최종적으로 출력되는 문단에서 제거되는 것을 방지할 수 있다.

[0049] 예컨대, 상기 인터페이스부(210)는 '사람'이라는 검색 키워드가 입력된 경우, '사람'과 같은 의미를 갖는 유사 키워드로서, '인간'을 외부 서버로부터 획득하여 제공할 수 있다.

[0050] 상기 추출부(220)는 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여하고, 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출할 수 있다.

[0051] 이때, 상기 추출부(220)는 상기 인터페이스부(210)를 통해 제공된 상기 유사 키워드에 대해 추가 검색 요청이 입력되면, 상기 복수의 문단 중에서 상기 유사 키워드가 포함되는 문단을 더 추출할 수 있다.

[0052] 상기 추출부(220)는 상기 검색 키워드가 복수 개일 경우, 상기 추출된 문단들 중 동일한 번호가 부여된 문단이 복수 개 존재하면 상기 복수의 문단 중 하나의 문단 이외의 나머지 문단을 상기 추출된 문단들에서 제거할 수 있다.

[0053] 예를 들어, 제1 검색 키워드가 포함되어 추출된 문단이, 제1 문단(번호 1이 부여된 문단), 제3 문단, 제4 문단이고, 제2 검색 키워드가 포함되어 추출된 문단이, 제3 문단, 제5 문단일 경우, 상기 추출부(220)는 중복하여 추출된 2개의 제3 문단 중 하나의 문단을 제거할 수 있다.

[0054] 상기 프로세서(230)는 상기 추출된 문단들을 정제할 수 있다. 이를 위해, 상기 프로세서(230)는 상기 추출된 문단들 내 상기 검색 키워드의 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산하고, 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거할 수 있다.

[0055] 이때, 상기 프로세서(230)는 각각의 검색 키워드가 포함된 문단의 개수를 합하여, 총 개수(총 빈도수)를 산출하고, 총 개수 대비 특정 검색 키워드가 포함된 문단의 개수(특정 검색 키워드의 빈도수)에 대한 비율을, 상기 특정 검색 키워드에 대한 가중치로서 산출할 수 있다.

[0056] 구체적으로, 상기 검색 키워드가, 제1 검색 키워드 및 제2 검색 키워드를 포함하는 경우, 상기 프로세서(230)는 상기 제1 검색 키워드를 포함하여 추출된 문단의 제1 개수와 상기 제2 검색 키워드를 포함하여 추출된 문단의 제2 개수를 합하여, 총 개수를 산출할 수 있다. 이후, 상기 프로세서(230)는 상기 총 개수 대비 상기 제1 개수의 비율을, 상기 제1 검색 키워드에 대한 가중치로서 산출하고, 상기 총 개수 대비 상기 제2 개수의 비율을, 상

기 제2 검색 키워드에 대한 가중치로서 산출할 수 있다.

[0057] 한편, 상기 프로세서(230)는 하나의 문단 내에 상기 가중치가 가장 낮은 검색 키워드 이외의 다른 검색 키워드가 존재하는 경우에는 예외 처리하여 해당 문단의 제거 기능을 수행하지 않는 것이 바람직하다.

[0058] 상기 프로세서(230)는 상기 정제된 문단들을 N(상기 N은 자연수, 예컨대 10)개의 영역으로 나눈 후, 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정할 수 있다.

[0059] 이를 위해, 상기 프로세서(230)는 상기 N개의 각 영역별 문단들의 중요도를 계산하고, 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 상기 주요 영역으로 선정할 수 있다.

[0060] 이때, 상기 프로세서(230)는 상기 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산하고, 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산할 수 있다.

[0061] 여기서, 상기 프로세서(230)는 상기 정제된 문단들의 개수를 상기 N으로 나눈 값으로 상기 각 영역별 문단들의 개수를 정할 수 있다. 다만, 나머지 값이 발생하는 경우, 상기 프로세서(230)는 상기 나머지 값을 맨 뒤에 배치된 영역에서부터 상기 나머지 값이 소진될 때가지 각 영역에 균등하게 순차적으로 가산하여 상기 각 영역별 문단들의 개수를 정할 수 있다.

[0062] 예를 들면, 상기 정제된 문단들의 개수가 83개이고 상기 N이 10이라고 가정한다. 이러한 경우, 상기 프로세서(230)는 83을 10으로 나눈 결과 값에서 몫에 해당하는 8을 제1 영역부터 제10 영역까지 각 영역별 문단들의 개수로 정한 후, 나머지 값에 해당하는 3에 대해서 맨 뒤에 배치된 제10 영역에서부터 제8 영역까지 1씩 더해줌으로써, 제1 내지 제7 영역까지는 8, 제8 내지 제10 영역까지는 9로 각 영역별 문단들의 개수를 정할 수 있다.

[0063] 상기의 예와 같은 경우, 상기 프로세서(230)는 상기 제1 내지 제7 영역까지는 각 영역별 검색 키워드의 빈도수를 8로 나누어 상기 각 영역별 검색 키워드의 평균 빈도수를 계산함으로써 상기 각 영역별 문단들의 중요도를 계산할 수 있다. 그리고, 상기 프로세서(230)는 상기 제8 내지 제10 영역까지는 각 영역별 검색 키워드의 빈도수를 9로 나누어 상기 각 영역별 검색 키워드의 평균 빈도수를 계산함으로써 상기 각 영역별 문단들의 중요도를 계산할 수 있다.

[0064] 상기 프로세서(230)는 상기 선정된 주요 영역에 포함된 문단들을 상기 추출부(220)에 의해 부여된 번호 순서대로 정렬하여 출력할 수 있다. 이처럼 상기 프로세서(230)는 상기 문서의 주요 영역에 해당하는 문단들을 추출하여 압축시켜 제공할 수 있으며, 이때 상기 추출된 문단을, 상기 추출된 문단에 각각 부여된 번호에 따라, 정렬하여 출력함(예컨대, 부여된 번호가 작은 순서대로 문단을 정렬함)으로써, 상기 문서 상에서의 문단 간 순서가 뒤바뀌지 않고, 배치 순서를 유지할 수 있게 한다.

[0065] 예컨대, 제1 검색 키워드가 포함되어 추출된 문단이, 제1 문단, 제3 문단, 제4 문단이고, 제2 검색 키워드가 포함되어 추출된 문단이, 제3 문단, 제5 문단일 경우, 상기 프로세서(230)는 각 문단에 부여된 번호에 따라, 제1 문단, 제3 문단, 제4 문단, 제5 문단 순서대로 정렬하여 출력할 수 있다.

[0066] 한편, 상기 프로세서(230)는 상기 추출된 문단 내 검색 키워드를, 상기 추출된 문단 내 다른 문자와 구별하여 출력할 수 있다. 이때, 상기 프로세서(230)는 상기 검색 키워드가 복수일 경우, 각각의 검색 키워드별로 상이한 형태, 예컨대 상이한 색상, 글꼴, 크기 등으로 구별하여 출력함으로써, 문단 내 복수의 검색 키워드를 쉽게 인식할 수 있게 한다.

[0067] 또한, 상기 프로세서(230)는 연속적인 번호가 부여된 문단이, 설정된 개수 이상 추출되는 경우, 해당 문단에 연속 식별표시를 출력할 수 있다. 예컨대, 추출된 문단이, 제1 문단, 제3 문단, 제4 문단, 제5 문단이고, 설정된 개수가 '3'일 경우, 상기 프로세서(230)는 3개의 연속적인 번호가 부여된 문단 즉, 제3 문단, 제4 문단, 제5 문단 각각에 대해, 연속 식별표시를 출력할 수 있다.

[0068] 다른 일례로서, 상기 프로세서(230)는 상기 검색 키워드가 복수일 경우, 상기 복수의 검색 키워드를 모두 포함하는 문단에 중요 식별표시를 출력할 수 있다. 예컨대, 검색 키워드가 제1 검색 키워드, 제2 검색 키워드 및 제3 검색 키워드를 포함하고, 제3 문단에 3개의 검색 키워드가 모두 포함될 경우, 상기 프로세서(230)는 제3 문단에, 중요 식별표시를 출력함으로써, 사용자로 하여금 중요한 문단을 쉽게 인지할 수 있게 한다.

[0069] 상기 데이터베이스(240)는 문서로부터 추출된 키워드를 저장할 수 있다.

[0071] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨

어 구성 요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0072] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0074] 도 3 내지 도 5는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 방법을 설명하기 위해 도시한 흐름도이다.

[0075] 여기서 설명하는 문서 분석 기반 주요 요소 추출 방법은 본 발명의 하나의 실시예에 불과하며, 그 이외에 필요에 따라 다양한 단계들이 부가될 수 있고, 하기의 단계들도 순서를 변경하여 실시될 수 있으므로, 본 발명이 하기에 설명하는 각 단계 및 그 순서에 한정되는 것은 아니다.

[0076] 먼저 도 3을 참조하면, 단계(310)에서 문서 분석 기반 주요 요소 추출 시스템(100)의 인터페이스부(210)는 문서 및 검색 키워드를 입력받을 수 있다.

[0077] 다음으로, 단계(320)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 추출부(220)는 상기 문서로부터 복수의 문단을 구분하여 상기 복수의 문단이 배치된 순서대로 번호를 부여할 수 있다.

[0078] 다음으로, 단계(330)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 추출부(220)는 상기 복수의 문단 중에서 상기 검색 키워드가 포함된 문단들을 추출할 수 있다.

[0079] 다음으로, 단계(340)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 프로세서(230)는 상기 추출된 문단들을 정제할 수 있다. 이에 대해 도 4를 참조하여 구체적으로 설명하면 다음과 같다.

[0080] 즉, 도 4를 참조하면, 단계(410)에서 상기 프로세서(230)는 상기 추출된 문단들 내 상기 검색 키워드의 빈도수를 산출할 수 있다. 이후, 단계(420)에서 상기 프로세서(230)는 상기 산출된 빈도수에 기초하여 상기 검색 키워드에 대한 가중치를 계산할 수 있다. 이후, 단계(430)에서 상기 프로세서(230)는 상기 추출된 문단들 중에서 상기 가중치가 가장 낮은 검색 키워드만을 포함한 문단을 제거할 수 있다.

[0081] 다시 도 3을 참조하면, 단계(350)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 프로세서(230)는 상기 정제된 문단들을 N개의 영역으로 분할할 수 있다.

[0082] 다음으로, 단계(360)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 프로세서(230)는 각 영역별 문단들의 중요도에 따라, 주요 요소를 포함하는 영역을 나타내는 주요 영역을 선정할 수 있다. 이에 대해 도 5를 참조하여 구체적으로 설명하면 다음과 같다.

[0083] 즉, 도 5를 참조하면, 단계(510)에서 상기 프로세서(230)는 상기 N개의 각 영역별 문단들에 포함된 상기 검색 키워드의 빈도수를 상기 각 영역별 문단들의 개수로 나누어, 각 영역별로 상기 검색 키워드의 평균 빈도수를 계산할 수 있다. 이후, 단계(520)에서 상기 프로세서(230)는 상기 계산된 평균 빈도수에 기초하여 상기 각 영역별 문단들의 중요도를 계산할 수 있다. 이후, 단계(530)에서 상기 프로세서(230)는 상기 각 영역별 문단들의 중요도를 비교하여 가장 높은 중요도를 갖는 영역을 상기 주요 영역으로 선정할 수 있다.

[0084] 다시 도 3을 참조하면, 단계(370)에서 상기 문서 분석 기반 주요 요소 추출 시스템(100)의 프로세서(230)는 상기 선정된 주요 영역에 포함된 문단들을 상기 부여된 번호 순서대로 정렬하여 출력할 수 있다.

[0086] 이하에서는 본 발명의 일 실시예에 따른 문서 분석 기반 주요 요소 추출 시스템을 구현하고 검증하는 일례를 상세히 설명하기로 한다.

[0088] 시스템 구현

[0090] 본 발명에서 제안하는 시스템의 구현을 다루고 효율성을 검증한다. 구현 및 실험에는 Windows 운영체제 기반의 CPU - Intel i5-4690, RAM - 8의 PC 1대를 사용하였다.

[0091] 사용자는 시스템이 시작되면 도 7에 도시된 바와 같이, 원하는 XML 형태 문서의 파일명을 입력하게 된다. 이를 위해, 시스템은 도 6에 도시된 바와 같이 검색을 원하는 XML 문서의 입력을 요구하는 화면을 먼저 출력할 수 있다. 그리고 사용자가 입력한 문서의 파일명과 일치하는 파일을 FileInputStream 메소드의 기능을 사용하여 불러온다. 그리고 Buffer 메소드의 기능을 활용하여 시스템이 해당 파일의 내용을 읽기 시작한다.

[0092] 해당 문서의 키워드 태그를 찾은 뒤 키워드 태그의 태그 값인 워드들을 추출하여 도 8에 도시된 바와 같이 사용자에게 보여준다. 그리고 도 8에 도시된 바와 같이 시스템은 추출된 워드들 중에서 검색을 원하는 키워드 3개의 입력을 요구하는 화면을 출력할 수 있다. 시스템의 키워드 추출이 완료되면, 도 9에 도시된 바와 같이 사용자는 검색을 하고자 하는 키워드 3개를 입력하게 된다. 시스템은 사용자가 입력한 3개의 키워드가 포함되어 있는 문단들을 검색하고 이를 추출한다.

[0093] 시스템은 도 10에 도시된 바와 같이 키워드들이 포함되어 있는 문단들의 수를 보여준 뒤, 도 11에 도시된 바와 같이 문단의 중복을 제거하는 기능을 진행한다. 이때, 시스템은 문단의 순서를 유지하는 기능을 함께 진행할 수 있다. 중복 제거 기능을 수행한 뒤, 시스템은 도 12에 도시된 바와 같이 키워드들의 빈도수를 확인하고 해당 키워드들의 빈도수를 퍼센트로 표현하여 출력하며, 아울러 키워드들의 가중치를 비교하여 가장 낮은 가중치의 키워드가 포함되어 있는 문단들을 제거한다.

[0094] 해당 과정에서 가장 낮은 가중치의 키워드가 1개가 아닐 경우 문단 제거 기능을 수행하지 않는다. 그 이유는 2개 이상의 키워드가 포함되어 있는 문단들을 제거할 경우 문서 이해에 대해 도움을 주는 시스템의 본래 목적을 달성하기 어려워지기 때문이다. 압축률을 높이는 것은 사용자가 읽어야 하는 문단의 수를 줄이기 때문에 사용자가 문서를 이해하는데 필요한 시간을 줄일 수 있다. 그러나 문서의 내용을 온전히 사용자에게 전달한다고 보기는 어렵기 때문이다. 이러한 이유로 인해 가장 낮은 키워드 가중치의 키워드가 1개일 경우 해당 작업을 수행한다.

[0095] 압축률을 높이기 위한 시스템의 작업이 완료되면 도출된 문단들을 10개의 영역으로 분할한다. 그리고 영역 내 문단들이 포함하고 있는 키워드의 빈도수를 계산하여 해당 영역의 중요도를 계산한다. 그리고 각 영역별 중요도를 비교하여 가장 높은 중요도의 영역(예: 3 영역)을 해당 문서의 주요 영역으로 설정하고, 도 13에 도시된 바와 같이 상기 주요(핵심) 영역을 사용자에게 보여준다.

[0097] 시스템 검증

[0099] 기존에 사용되었던 문서 분석 시스템들은 대부분이 형태소 분석기를 기반으로 구현되었다. 이러한 구조적인 이유로 인해 해당 시스템들의 주 기능은 문서 작성에 사용된 단어들을 분류하고 해당 단어들의 빈도수를 확인하는 것이었다. 또한 사용자가 입력한 검색어가 해당 문서 작성에 사용되었는지 확인하기 위해 사용하였다.

[0100] 기존 시스템들은 검색어의 사용 여부를 확인하는 것이 주목적이었기 때문에 구조적인 이유로 인해 기능적 한계가 발생할 수 밖에 없었다. 기능적 한계로 인해 발생할 수 있는 문제점으로는 사용자가 문서를 이해하는데 효율적으로 도움을 주기 어려운 문제점이 있었다. 또한 문서를 이해하는 필요한 시간을 줄이지 못하는 문제점이 발생할 수 있었다.

[0101] 이러한 문제점들을 해결하기 위해 본 발명에서 제안하는 시스템은 사용자가 입력한 키워드가 포함되어 있는 문단들을 추출하고 추출된 문단들의 순서를 유지하며 중복된 문단들을 제거한다. 그리고 키워드들의 빈도수를 확인하고 가중치를 계산하여 사용자에게 이를 보여준다.

[0102] 제안하는 시스템의 효율성을 검증하기 위한 실험으로 정규화된 XML 형태 문서로 진행하였다. 그리고 실험은 주제 구분 없이 선정한 XML 형태 문서 6개와 주제별로 정리된 XML 형태 문서 7개로 진행하였다.

[0103] 도 14는 주제 구분 없이 선정한 XML 형태 문서들로 진행한 실험의 결과를 그래프로 정리한 것이다. 총 6차례의

실험을 진행하였으며 기존 시스템은 총 437개의 문단을 도출하였다. 제안하는 시스템은 총 302개의 문단을 도출하였고 이는 기존 시스템에 비해 135개의 문단을 적게 도출하여 사용자에게 제공했다는 뜻이 된다. 수치로 나타낼 경우 제안하는 시스템은 기존 시스템이 도출한 문단에 비해 69.1 퍼센트 정도의 문단을 도출하였고 사용자는 기존 시스템이 비해 약 30퍼센트 적은 수의 문단으로 문서를 이해할 수 있게 된다.

[0104] 도 15는 주제별로 정리된 XML 형태 문서들로 진행한 실험의 결과를 그래프로 정리한 것이다. 총 7차례의 실험을 진행하였으며 기존 시스템은 총 361개의 문단을 도출하였다. 제안하는 시스템은 총 269개의 문단을 도출하였고 이는 기존 시스템에 비해 92개의 문단을 적게 도출하여 사용자에게 제공했다는 뜻이 된다. 수치로 나타낼 경우 제안하는 시스템은 기존 시스템이 도출한 문단에 비해 74.5 퍼센트 정도의 문단을 도출하였고 사용자는 기존 시스템에 비해 약 25퍼센트 적은 수의 문단으로 문서를 이해할 수 있게 된다.

[0105] 총 13차례 실험 데이터를 기반으로 사용자는 기존 시스템에 비해 보다 높은 압축률과 다양한 정보(주요 핵심 정보)들을 제안하는 시스템으로부터 제공받을 수 있었다. 이로 인해, 제안하는 시스템은 기존 시스템에 비해 사용자의 효율적인 문서 이해에 보다 많은 도움을 줄 수 있다.

[0107] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CDROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0108] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0109] 그러므로, 다른 구현들, 다른 실시예들 및 청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

부호의 설명

[0111] 100: 문서 분석 기반 주요 요소 추출 시스템

101: 단말

210: 인터페이스부

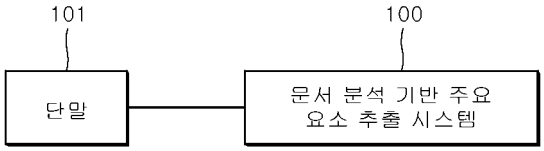
220: 추출부

230: 프로세서

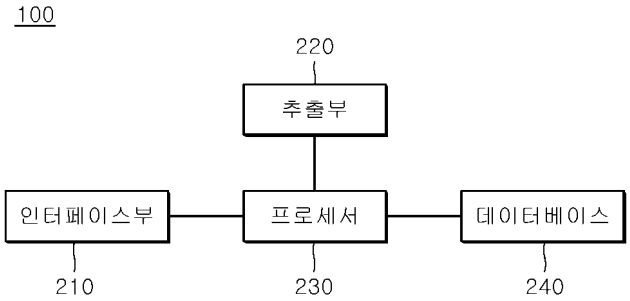
240: 데이터베이스

도면

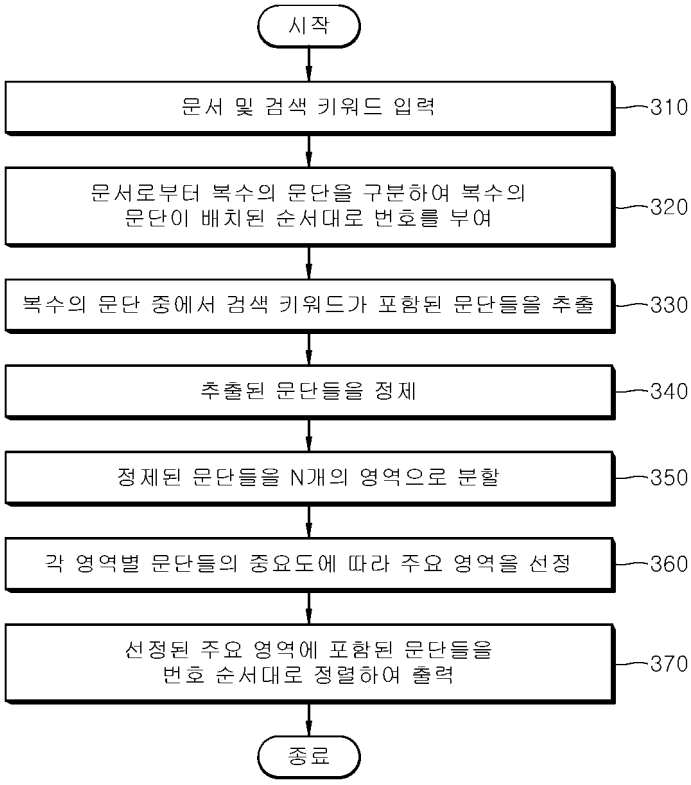
도면1



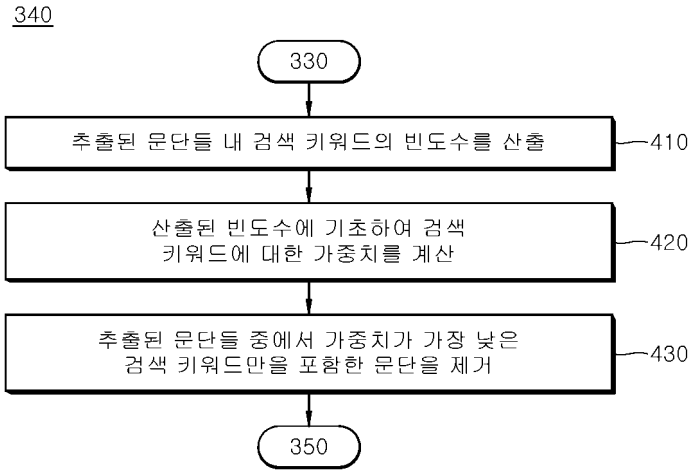
도면2


100
220
추출부
인터페이스부
프로세서
데이터베이스
210
230
240

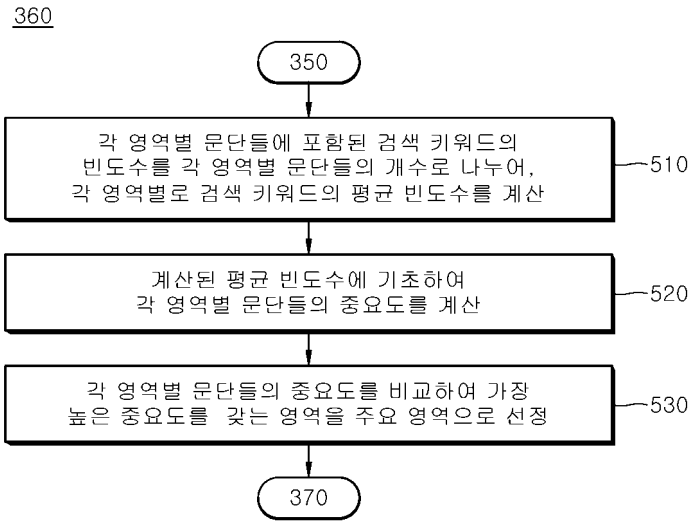
도면3


시작
문서 및 검색 키워드 입력
310
문서로부터 복수의 문단을 구분하여 복수의 문단이 배치된 순서대로 번호를 부여
320
복수의 문단 중에서 검색 키워드가 포함된 문단들을 추출
330
추출된 문단들을 정제
340
정제된 문단들을 N개의 영역으로 분할
350
각 영역별 문단들의 중요도에 따라 주요 영역을 선정
360
선정된 주요 영역에 포함된 문단들을 번호 순서대로 정렬하여 출력
370
종료

도면4

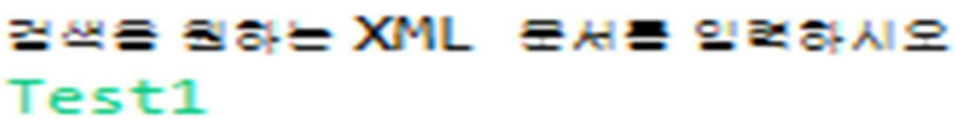


도면5

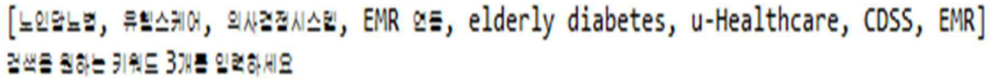


도면6

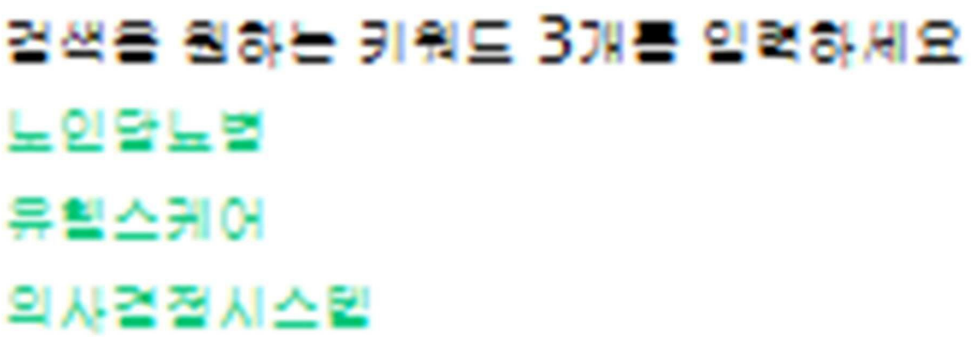
도면7



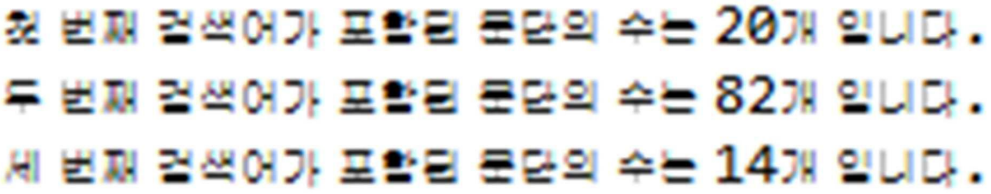
도면8



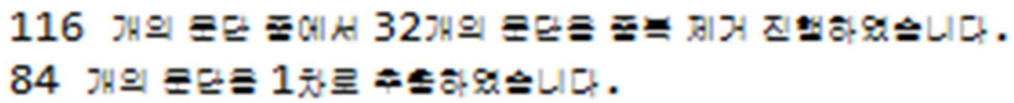
도면9



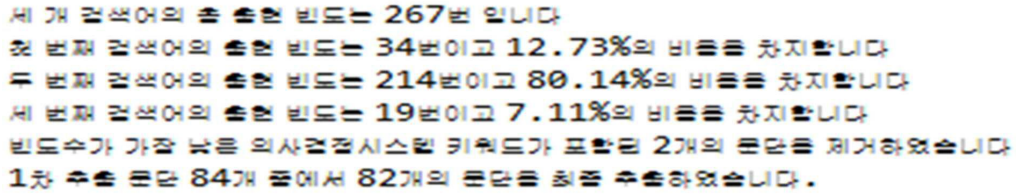
도면10



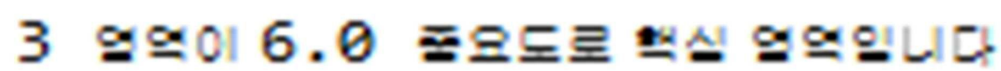
도면11



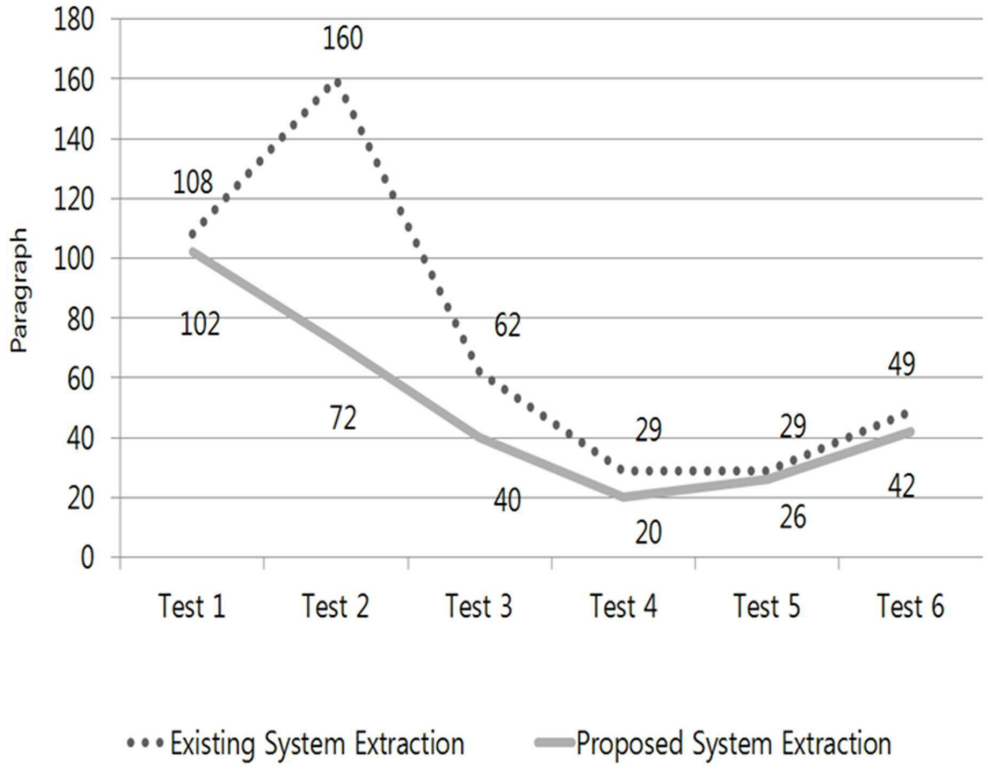
도면12



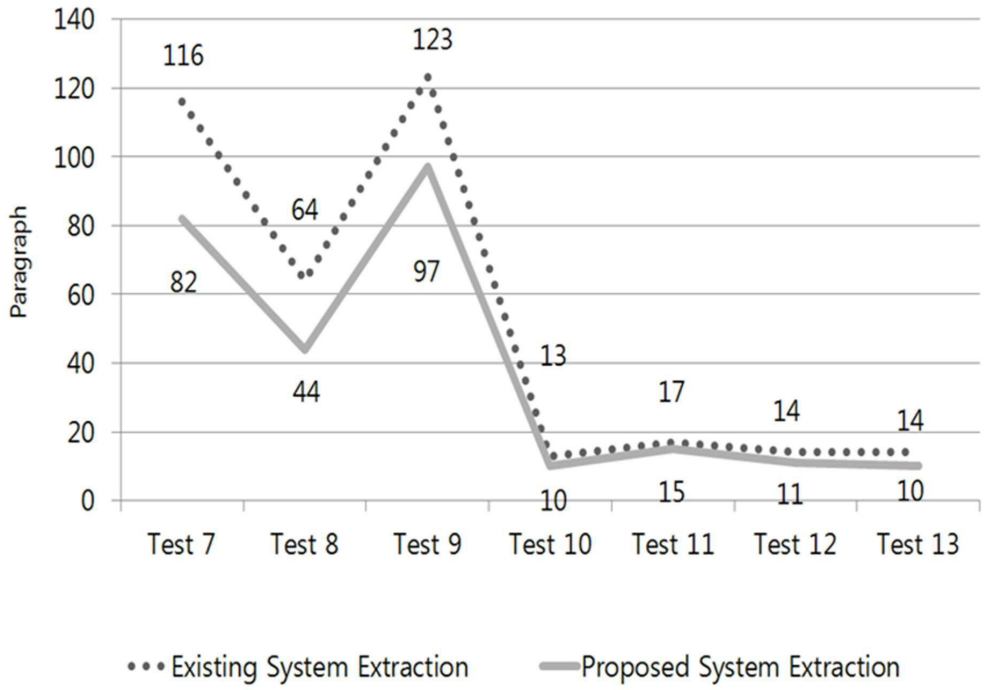
도면13



도면14


180
160
140
120
100
80
60
40
20
0
Paragraph
160
108
102
62
72
49
29
29
40
42
20
26
Test 1
Test 2
Test 3
Test 4
Test 5
Test 6
Existing System Extraction
Proposed System Extraction

도면15


140
120
100
80
60
40
20
0
Paragraph
116
123
64
82
97
13
44
17
14
14
10
15
11
10
Test 7
Test 8
Test 9
Test 10
Test 11
Test 12
Test 13
Existing System Extraction
Proposed System Extraction