(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2019년02월11일
(11) 등록번호 10-1946275
(24) 등록일자 2019년01월31일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01)
(52) CPC특허분류
G06F 16/951 (2019.01)
G06F 16/901 (2019.01)
(21) 출원번호 10-2017-0105853
(22) 출원일자 2017년08월22일
심사청구일자 2017년08월22일
(56) 선행기술조사문헌
KR1020030075219 A

(73) 특허권자
인천대학교 산학협력단
인천광역시 연수구 아카데미로 119 (송도동)
(72) 발명자
김관호
인천광역시 연수구 아카데미로 119 인천대학교 공과대학 443호
이수경
인천광역시 서구 봉오재2로 37, LH웨스턴블루힐 206동 1702호
(74) 대리인
김효성

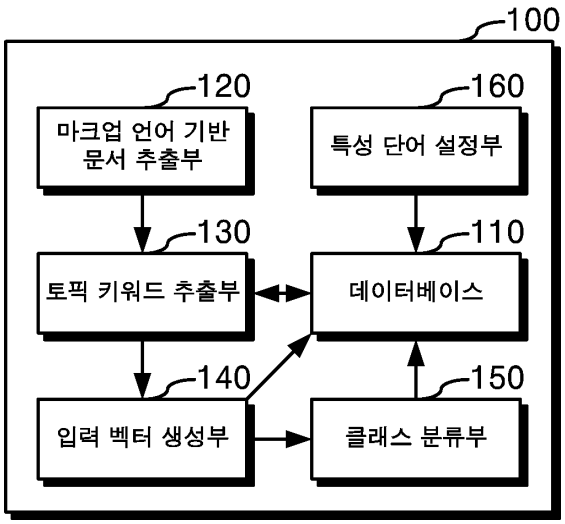
전체 청구항 수 : 총 20 항 심사관 : 이복현

(54) 발명의 명칭 **웹 페이지의 마크업 언어 기반 문서에서 토픽 키워드를 추출하여 웹 페이지를 분류하는 장치 및 그의 동작 방법**

(57) 요 약

본 발명에 따른 웹 페이지 분류 장치 및 그의 동작 방법은 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하고, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 입력 벡터를 생성하며, 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류함으로써, 자동으로 웹 페이지를 주제별로 분류하는 것을 가능하게 할 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업
과제고유번호 NRF-2017R1D1A1B03035639
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 개인기초연구(교육부)
연구과제명 인간-공간 상호작용의 전체최적화를 위한 딥러닝 기반의 지능공간 의사결정 모델연구
기 여 율 1/1
주관기관 인천대학교
연구기간 2017.06.01 ~ 2020.05.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

사전 설정된 L - 상기 L은 2 이상의 정수임 - 개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들 - 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수임 - 이 저장되어 있는 데이터베이스;

상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하는 토픽 키워드 추출부;

상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터 - 상기 입력 벡터의 원소들의 개수는 i개임 - 를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성하는 입력 벡터 생성부; 및

상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류하는 클래스 분류부

를 포함하는 웹 페이지 분류 장치.

청구항 2

제1항에 있어서,

상기 토픽 키워드 추출부는 상기 콘텐츠들의 태그 정보에 기초하여 상기 콘텐츠들에서 상기 텍스트 콘텐츠들을 추출하는

웹 페이지 분류 장치.

청구항 3

제2항에 있어서,

상기 토픽 키워드 추출부는 상기 텍스트 콘텐츠들에 포함된 워드들 간의 유사도 점수를 연산한 후 워드들 각각에 대하여 해당 워드와 다른 워드들 간에 연산된 유사도 점수들 중 가장 높은 유사도 점수를 해당 워드의 최종 유사도 점수로 할당하고, 최종 유사도 점수가 높은 j - 상기 j는 1 이상의 정수임 - 개의 워드를 토픽 키워드로 추출하는

웹 페이지 분류 장치.

청구항 4

제3항에 있어서,

상기 토픽 키워드 추출부는 상기 텍스트 콘텐츠들을 제1 텍스트 콘텐츠들과 제2 텍스트 콘텐츠들로 분류하고, 상기 제1 텍스트 콘텐츠들에 포함된 제1 워드들과 상기 제1 텍스트 콘텐츠들에 포함된 제2 워드들 간의 유사도 점수를 연산하며,

상기 제1 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들이고,

상기 제2 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들 이외의 콘텐츠들인

웹 페이지 분류 장치.

청구항 5

제4항에 있어서,

상기 토픽 키워드 추출부는 상기 콘텐츠들 중 상기 태그 정보가 텍스트 태그, 테이블 태그, 헤더 태그 및 목록 태그 중 어느 하나의 태그 정보를 가지는 콘텐츠를 텍스트 콘텐츠로 추출하는

웹 페이지 분류 장치.

청구항 6

제1항에 있어서,

상기 데이터베이스는 제1 내지 제i 강화 계수를 저장하고 있고,

상기 입력 벡터 생성부는 상기 워드들 중 제k - 상기 k는 1 이상 상기 i 이하의 정수임 - 워드가 상기 토픽 키워드에 포함되면, 제k 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산하고, 상기 제k 워드가 상기 토픽 키워드에 포함되지 않으면, 1의 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산하는

웹 페이지 분류 장치.

청구항 7

제6항에 있어서,

상기 입력 벡터 생성부는 상기 워드들 중 상기 제k 워드의 출현 빈도, 상기 제k 워드의 역문서 빈도 및 상기 제k 워드가 상기 토픽 키워드에 포함되는지 여부에 따라 결정되는 강화 계수를 곱함으로써 상기 제k 워드에 대한 원소 값을 연산하는

웹 페이지 분류 장치.

청구항 8

제1항에 있어서,

상기 클래스 분류부는 상기 연산된 분류 값들 중 제s - 상기 s는 1 이상 상기 L 이하의 정수임 - 클래스와 상기 제s 클래스 이외의 클래스들 각각을 분류하기 위한 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제s 분류 값들을 합산하여 제s 합산 값을 연산함으로써 제1 내지 제L 합산 값들을 생성하며, 상기 제1 내지 제L 합산 값들 중 가장 큰 값을 가지는 합산 값에 상응하는 클래스를 상기 웹 페이지의 클래스로 분류하는

웹 페이지 분류 장치.

청구항 9

제8항에 있어서,

상기 데이터베이스는 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값들을 더 저장하고 있고,

상기 클래스 분류부는 상기 분류 기준 벡터들 각각과 상기 입력 벡터 간의 내적 곱을 수행한 후 상기 분류 기준 조절 값들 중 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값을 가산함으로써 분류 값들을 연산하는

웹 페이지 분류 장치.

청구항 10

사전 설정된 L - 상기 L은 2 이상의 정수임 - 개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들 - 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수임 - 이 저장되어 있는 데이터베이스를 유지하는 단계;

상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하는 단계;

상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터 - 상기 입력 벡터의 원소들의 개수는 i개임 - 를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상

기 입력 벡터를 생성하는 단계; 및

상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류하는 단계

를 포함하는 웹 페이지 분류 장치의 동작 방법.

청구항 11

제10항에 있어서,

상기 토픽 키워드를 추출하는 단계는 상기 콘텐츠들의 태그 정보에 기초하여 상기 콘텐츠들에서 상기 텍스트 콘텐츠들을 추출하는

웹 페이지 분류 장치의 동작 방법.

청구항 12

제11항에 있어서,

상기 토픽 키워드를 추출하는 단계는 상기 텍스트 콘텐츠들에 포함된 워드들 간의 유사도 점수를 연산한 후 워드들 각각에 대하여 해당 워드와 다른 워드들 간에 연산된 유사도 점수들 중 가장 높은 유사도 점수를 해당 워드의 최종 유사도 점수로 할당하고, 최종 유사도 점수가 높은 j - 상기 j는 1 이상의 정수임 - 개의 워드를 토픽 키워드로 추출하는

웹 페이지 분류 장치의 동작 방법.

청구항 13

제12항에 있어서,

상기 토픽 키워드를 추출하는 단계는 상기 텍스트 콘텐츠들을 제1 텍스트 콘텐츠들과 제2 텍스트 콘텐츠들로 분류하고, 상기 제1 텍스트 콘텐츠들에 포함된 제1 워드들과 상기 제1 텍스트 콘텐츠들에 포함된 제2 워드들 간의 유사도 점수를 연산하며,

상기 제1 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들이고,

상기 제2 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들 이외의 콘텐츠들인

웹 페이지 분류 장치의 동작 방법.

청구항 14

제13항에 있어서,

상기 토픽 키워드를 추출하는 단계는 상기 콘텐츠들 중 상기 태그 정보가 텍스트 태그, 테이블 태그, 헤더 태그 및 목록 태그 중 어느 하나의 태그 정보를 가지는 콘텐츠를 텍스트 콘텐츠로 추출하는

웹 페이지 분류 장치의 동작 방법.

청구항 15

제10항에 있어서,

상기 데이터베이스를 유지하는 단계는 제1 내지 제i 강화 계수를 저장하고 있는 상기 데이터베이스를 유지하고,

상기 입력 벡터를 생성하는 단계는 상기 워드들 중 제k - 상기 k는 1 이상 상기 i 이하의 정수임 - 워드가 상기 토픽 키워드에 포함되면, 제k 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산하고, 상기 제k 워드가 상기 토픽 키워드에 포함되지 않으면, 1의 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산하는

웹 페이지 분류 장치의 동작 방법.

청구항 16

제15항에 있어서,

상기 입력 벡터를 생성하는 단계는 상기 워드들 중 상기 제k 워드의 출현 빈도, 상기 제k 워드의 역문서 빈도 및 상기 제k 워드가 상기 토픽 키워드에 포함되는지 여부에 따라 결정되는 강화 계수를 곱함으로써 상기 제k 워드에 대한 원소 값을 연산하는

웹 페이지 분류 장치의 동작 방법.

청구항 17

제10항에 있어서,

상기 웹 페이지의 클래스를 분류하는 단계는 상기 연산된 분류 값들 중 제s - 상기 s는 1 이상 상기 L 이하의 정수임 - 클래스와 상기 제s 클래스 이외의 클래스들 각각을 분류하기 위한 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제s 분류 값들을 합산하여 제s 합산 값을 연산함으로써 제1 내지 제L 합산 값들을 생성하며, 상기 제1 내지 제L 합산 값들 중 가장 큰 값을 가지는 합산 값에 상응하는 클래스를 상기 웹 페이지의 클래스로 분류하는

웹 페이지 분류 장치의 동작 방법.

청구항 18

제17항에 있어서,

상기 데이터베이스를 유지하는 단계는 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값들을 더 저장하고 있는 상기 데이터베이스를 유지하고,

상기 웹 페이지의 클래스를 분류하는 단계는 상기 분류 기준 벡터들 각각과 상기 입력 벡터 간의 내적 곱을 수행한 후 상기 분류 기준 조절 값들 중 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값을 가산함으로써 분류 값들을 연산하는

웹 페이지 분류 장치의 동작 방법.

청구항 19

제10항 내지 제18항 중 어느 한 항의 방법을 컴퓨터로 하여금 수행하도록 하는 프로그램을 기록한 컴퓨터 판독 가능 기록 매체.

청구항 20

제10항 내지 제18항 중 어느 한 항의 방법을 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 웹 페이지의 마크업 언어 기반 문서에서 토픽 키워드를 추출하여 웹 페이지를 분류하는 장치 및 그의 동작 방법에 대한 것이다.

배 경 기 술

[0002] 사용자들은 웹 페이지에 게시된 정보로부터 자기가 원하는 콘텐츠와 관련이 높은 웹 페이지를 탐색하므로, 온라인상에 웹 페이지가 급증함에 따라 사용자들의 웹 페이지 탐색 시간이 증가하고 있다.

[0003] 이에 따라, 주제 별로 웹 페이지를 분류하고 제공하는 서비스가 등장하였는데, 주제 별로 웹 페이지를 분류하는 작업은 대부분 수작업으로 이루어짐에 따라 올바른 웹 페이지 분류가 용이하지 않다는 문제점이 있으며, 또 분류 체계가 시간에 따라 변화하므로, 기존에 분류된 웹 페이지도 다시 분류해야 하는 문제점이 있다. 이에 따라, 웹 페이지를 분류하는데 많은 시간과 비용이 소모되며, 정확도가 떨어지는 문제점이 있다.

[0004] 따라서, 자동으로 웹 페이지를 주제 별로 분류함으로써 적은 시간과 비용이 소모되고, 정확도가 높은 웹 페이지

분류 알고리즘이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명에 따른 웹 페이지 분류 장치 및 그의 동작 방법은 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하고, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 입력 벡터를 생성하며, 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류함으로써, 자동으로 웹 페이지를 주제 별로 분류하는 것을 가능하게 하고자 한다.

과제의 해결 수단

[0006] 본 발명의 일실시예에 따른 웹 페이지 분류 장치는 사전 설정된 L - 상기 L은 2 이상의 정수임 - 개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들 - 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수임 - 이 저장되어 있는 데이터베이스, 상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하는 토픽 키워드 추출부, 상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터 - 상기 입력 벡터의 원소들의 개수는 i개임 - 를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성하는 입력 벡터 생성부 및 상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류하는 클래스 분류부를 포함한다.

[0007] 또한, 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법은 사전 설정된 L - 상기 L은 2 이상의 정수임 - 개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들 - 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수임 - 이 저장되어 있는 데이터베이스를 유지하는 단계, 상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하는 단계, 상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터 - 상기 입력 벡터의 원소들의 개수는 i개임 - 를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성하는 단계 및 상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류하는 단계를 포함한다.

발명의 효과

[0008] 본 발명에 따른 웹 페이지 분류 장치 및 그의 동작 방법은 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하고, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 입력 벡터를 생성하며, 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류함으로써, 자동으로 웹 페이지를 주제 별로 분류하는 것을 가능하게 할 수 있다.

도면의 간단한 설명

[0009] 도 1은 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 구조를 도시한 도면이다.

도 2는 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법을 도시한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0010] 이하에서는 본 발명에 따른 실시예들을 첨부된 도면을 참조하여 상세하게 설명하기로 한다. 이러한 설명은 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였으며, 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 본 명세서 상에서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 사람에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다.

[0011] 이하에서, 본 발명에 따른 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다.

[0012] 도 1은 본 발명의 일실시예에 따른 웹 페이지 분류 장치(100)의 구조를 도시한 도면이다.

[0013] 도 1에 도시된 바와 같이, 본 발명의 일실시예에 따른 웹 페이지 분류 장치(100)는 데이터베이스(110), 마크업 언어 기반 문서 추출부(120), 토픽 키워드 추출부(130), 입력 벡터 생성부(140) 및 클래스 분류부(150)를 포함한다.

[0014] 먼저, 데이터베이스(110)는 사전 설정된 L개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들을 저장할 수 있다. 여기에서, 상기 L은 2 이상의 정수이고, 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수이며, 특성 단어들은 존재하는 모든 단어들 중 웹 페이지의 클래스를 분류하기 위해 사전 설정된 단어들이며, 클래스는 웹 페이지를 분류하기 위한 것으로, 예를 들어, 학회와 관련된 웹 페이지를 분류하기 위한 클래스, 스포츠와 관련된 웹 페이지를 분류하기 위한 클래스 및 게임과 관련된 웹 페이지를 분류하기 위한 클래스 등으로 세분화될 수 있다.

[0015] 예를 들어, 클래스가 4개(즉, L이 4)인 경우, 데이터베이스(110)에는 상기 분류 기준 벡터들이 하기 표 1과 같이 저장되어 있을 수 있다.

표 1

[0016]

분류 기준 벡터	벡터 값
V_{12}	(1, 0, 0, 0)
V_{13}	(0, 1, 0, 0)
V_{14}	(0, 0, 1, 0)
V_{23}	(0, 0, 0, 1)
V_{24}	(1, 0, 1, 0)
V_{34}	(0, 1, 0, 1)
V_{21}	(-1, 0, 0, 0)
V_{31}	(0, -1, 0, 0)
V_{41}	(0, 0, -1, 0)
V_{32}	(0, 0, 0, -1)
V_{42}	(-1, 0, -1, 0)
V_{43}	(0, -1, 0, -1)

[0017] 표 1을 참조하면, 분류 기준 벡터 1(V_{12})은 벡터 값 (1, 0, 0, 0)을 가지고, 분류 기준 벡터 2(V_{13})는 벡터 값 (0, 1, 0, 0)을 가지고, 분류 기준 벡터 3(V_{14})은 벡터 값 (0, 0, 1, 0)을 가지고, 분류 기준 벡터 4(V_{23})는 벡터 값 (0, 0, 0, 1)을 가지고, 분류 기준 벡터 5(V_{24})는 벡터 값 (1, 0, 1, 0)을 가지고, 분류 기준 벡터 6(V_{34})은 벡터 값 (0, 1, 0, 1)을 가지는 것을 알 수 있다. 한편, V_{21}의 벡터 값은 분류 기준 벡터 1(V_{12})의 벡터 값 (1, 0, 0, 0)에 -1을 곱한 값이고, V_{31}의 벡터 값은 분류 기준 벡터 1(V_{13})의 벡터 값 (0, 1, 0, 0)에 -1을 곱한 값이고, 마찬가지로, V_{41}, V_{32}, V_{42}, V_{43}의 벡터 값은 분류 기준 벡터 3 내지 6(V_{14}, V_{23}, V_{24}, V_{34})의 벡터 값에 -1을 곱한 값이다.

[0018] 아울러, 데이터베이스(110)는 상기 분류 기준 벡터들(V_{12}, V_{13}, V_{14}, V_{23}, V_{24}, V_{34}, V_{21}, V_{31}, V_{41}, V_{32}, V_{42} 및 V_{43}) 각각에 매칭되어 있는 분류 기준 조절 값들(B_{12}, B_{13}, B_{14}, B_{23}, B_{24}, B_{34}, B_{21}, B_{31}, B_{41}, B_{32}, B_{42} 및 B_{43})을 저장할 수 있다.

[0019] 또한, 데이터베이스(110)는 하기 표 2와 같이 사전 설정된 N-1개의 웹 페이지에 출현하는 워드들의 출현 횟수를 저장할 수 있다. 여기에서, 상기 N은 2 이상의 정수이다.

표 2

[0020]

	w_1	w_2	w_3	...	w_i
p_1	1	0	0	...	1
p_2	2	2	1	...	1
p_3	3	5	0	...	3
...	...	...	...	...	...
p_{N-1}	3	1	0	...	2

[0021] 여기에서, p_1 내지 p_{n-1}은 사전 설정된 N-1개의 웹 페이지에 각각 대응하는 필드이고, w_1 내지 w_i는 사전 설정된 i개의 특성 단어들에 각각 대응하는 필드이며, 각 엔트리들은 각 웹 페이지에서 각 특성 단어들이 출현하는 횟수를 나타낸다. 예를 들어, 표 2를 참조하면, 웹 페이지 1(p_1)에는 워드 1(w_1)이 1회 출현하며 워드 2 및 3(w_2 및 w_3)이 0회 출현하며, 워드 i(w_i)가 1회 출현함을 알 수 있고, 웹 페이지 N-1(p_{N-1})에는 워드 1(w_1)이 3회 출현하며 워드 2(w_2)가 1회 출현하고 워드 3(w_3)이 0회 출현하며, 워드 i(w_i)가 2회 출현함을 알 수 있다.

[0022] 또한, 데이터베이스(110)는 상기 워드 1 내지 워드 i에 각각 대응하는 제1 내지 제i 강화 계수들을 저장한다.

[0023] 마크업 언어 기반 문서 추출부(120)는 월드 와이드 웹(WWW : World Wide Web)에 존재하는 웹 페이지 중 클래스가 분류되지 않은 웹 페이지를 선택하고, 선택된 웹 페이지의 마크업 언어 기반 문서(Markup Language Based Document)를 추출한다. 구체적으로, 마크업 언어 기반 문서 추출부(120)는 사전 설정된 유알엘(URL: Uniform Resource Locator) 리스트에 포함된 웹 페이지들 중 클래스가 분류되지 않은 웹 페이지를 선택한 후 상기 선택된 웹 페이지를 방문하고, 방문한 웹 페이지의 마크업 언어 기반 문서를 추출한다. 이후, 마크업 언어 기반 문서 추출부(120)는 상기 마크업 언어 기반 문서에 기재된 모든 하이퍼링크의 유알엘을 추출하고, 추출된 유알엘을 유알엘 리스트에 포함시킴으로써 유알엘 리스트를 갱신할 수 있다. 여기에서, 마크업 언어 기반 문서는 예를 들어, HTML(HyperText Markup Language) 및 XML(eXtensible Markup Language) 등과 같은 마크업 언어에 기반한 문서이다.

[0024] 토픽 키워드 추출부(130)는 상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출한다.

[0025] 구체적으로, 토픽 키워드 추출부(130)는 마크업 언어 기반 문서를 파싱(parsing)함으로써 마크업 언어 기반 문서에 포함된 콘텐츠들을 추출하고, 추출된 콘텐츠들의 태그 정보에 기초하여 텍스트 콘텐츠들을 추출할 수 있다. 보다 구체적으로, 상기 토픽 키워드 추출부(130)는 상기 콘텐츠들 중 상기 태그 정보가 텍스트 태그, 테이블 태그, 헤더 태그 및 목록 태그 중 어느 하나의 태그 정보를 가지는 콘텐츠를 텍스트 콘텐츠로 추출할 수 있다.

[0026] 이때, 토픽 키워드 추출부(130)는 상기 웹 페이지의 추출된 텍스트 콘텐츠들에 포함된 워드들에 대한 정보를 이용하여 상기 데이터베이스(110)에 저장되어 있는 웹 페이지에 출현하는 워드들의 출현 횟수를 갱신할 수 있다. 이에 따라, 상기 데이터베이스(110)에는 N-1개의 웹 페이지에 출현하는 워드들의 출현 횟수에서 N개의 웹 페이지에 출현하는 워드들의 출현 횟수로 갱신될 수 있다. 여기에서, 예를 들어, 마크업 언어 기반 문서가 HTML 기반 문서인 경우, 텍스트 태그는 , <p>, , <title> 등이고, 테이블 태그는 <table> 이며, 헤더 태그는 <h1>, <h2>, <h3>, <h4>, <h5> 및 <h6> 이고, 목록 태그는 <ul> 및 <li> 등일 수 있다.

[0027] 예를 들어, 마크업 언어 기반 문서가 HTML 기반 문서인 경우, 상기 토픽 키워드 추출부(130)는 추출된 콘텐츠들의 태그 정보에 기초하여 , , <title>, <p>, <ul>, <li>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6> 및

<table>를 태그 정보로 가지는 콘텐츠들을 텍스트 콘텐츠들로 추출할 수 있다. 여기에서, 예를 들어, 추출된 텍스트 콘텐츠에 포함된 텍스트는 "학회 웹사이트의 토픽 정보추출을 이용한 주제에 따른 학회 자동분류 기법"(이하, '텍스트 1'이라고 지칭함), "학회의 주요 정보추출과 관련된 기존연구는 연구목적, 추출 기법, 추출된 정보에 따라 <Table 1>과 같이 요약할 수 있다."(이하, '텍스트 2'라고 지칭함) 등일 수 있다. 또한, 상기 텍스트들에는 "학회", "웹사이트", "토픽", "정보추출", "이용", "주제", "따른", "학회", "자동분류", "기법", "학회", "주요", "정보추출", "관련", "기존연구", "연구목적", "추출", "기법", "추출", "정보", "Table" 및 "요약"과 같은 워드들이 포함되어 있으므로, 토픽 키워드 추출부(130)는 해당 워드들의 출현 횟수에 기초하여 상기 데이터베이스(110)에 저장되어 있는 웹 페이지에 출현하는 워드들의 출현 횟수를 갱신할 수 있다.

[0028] 예를 들어, 데이터베이스(110)에는 하기 표 3과 같은 갱신된 웹 페이지에 출현하는 워드들의 출현 횟수가 저장되어 있을 수 있다.

표 3

[0029]

	w_1	w_2	w_3	...	w_i
p_1	1	0	0	...	1
p_2	2	2	1	...	1
p_3	3	3	0	...	3
...	...	...	...	...	...
p_{N-1}	3	1	0	...	2
p_N	2	0	1	...	1

[0030] 이후, 토픽 키워드 추출부(130)는 기 텍스트 콘텐츠들에 포함된 워드들 간의 유사도 점수를 연산한 후 워드들 각각에 대하여 해당 워드와 다른 워드들 간에 연산된 유사도 점수들 중 가장 높은 유사도 점수를 해당 워드의 최종 유사도 점수로 할당하고, 최종 유사도 점수가 높은 j개의 워드를 토픽 키워드로 추출할 수 있다. 여기에서, 상기 j는 1 이상의 정수이다.

[0031] 예를 들어, 토픽 키워드 추출부(130)는 제1 텍스트 콘텐츠로 분류된 텍스트 콘텐츠 1에 포함된 텍스트 1에 포함된 제1 워드들(즉, "학회", "웹사이트", "토픽", "정보추출", "이용", "주제", "따른", "학회", "자동분류" 및 "기법",)과 제2 텍스트 콘텐츠로 분류된 텍스트 콘텐츠 2에 포함된 텍스트 2에 포함된 제2 워드들(즉, "학회", "주요", "정보추출", "관련", "기존연구", "연구목적", "추출", "기법", "추출", "정보", "Table" 및 "요약") 간의 유사도 점수를 연산한 후 워드들 각각에 대하여 해당 워드와 다른 워드들 간에 연산된 유사도 점수들 중 가장 높은 유사도 점수를 해당 워드의 최종 유사도 점수로 할당하고, 최종 유사도 점수가 높은 j개의 워드를 토픽 키워드로 추출할 수 있다. 예를 들어, 토픽 키워드 추출부(130)는 워드 "학회"와 워드들 "웹사이트", "토픽", "정보추출", ..., "요약" 간에 연산된 워드 "학회"의 유사도 점수들 중 가장 높은 유사도 점수를 워드 "학회"의 최종 유사도 점수로 할당할 수 있다. 토픽 키워드 추출부(130)는 이와 같이 할당된 최종 유사도 점수들에 기초하여 최종 유사도 점수가 높은 j개의 워드를 토픽 키워드로 추출할 수 있다.

[0032] 구체적으로, 토픽 키워드 추출부(130)는 하기 수학식 1을 이용하여 유사도 점수를 연산할 수 있다.

수학식 1

[0033]
$$S(w_a, w_b) = \frac{2\sum_{x=1}^{N}\delta(f_{ax}) \cdot \delta(f_{bx})}{\sum_{x=1}^{N}\delta(f_{ax}) + \sum_{x=1}^{N}\delta(f_{bx})}$$

[0034] 여기에서, $S(w_a, w_b)$는 워드 a(w_a)와 워드 b(w_b) 간의 유사도 점수이고, a는 1 이상 i 이하의 정수이고, b는 1

이상 i 이하의 정수이며, a 및 b는 서로 다른 정수이며, f_{ax}는 웹 페이지 x(p_x)에서 워드 a(w_a)의 출현 횟수이고, f_{bx}는 웹 페이지 x(p_x)에서 워드 b(w_b)의 출현 횟수이고, x는 1 이상 N 이하의 정수이며, $\delta(f)$는 출현 횟수(f)가 0을 초과하면 1의 값을 가지고, 0이면 0의 값을 가지는 함수이다. 또한, $\sum_{x=1}^{N}\delta(f_{ax})$는 워드 a($w_a$)가 출현한 웹 페이지의 개수이며, $\sum_{x=1}^{N}\delta(f_{bx})$는 워드 b($w_b$)가 출현한 웹 페이지의 개수이고, $\sum_{x=1}^{N}\delta(f_{ax})\cdot\delta(f_{bx})$는 워드 a 및 b($w_a$ 및 w_b)가 공동으로 출현한 웹 페이지의 개수이다. 이에 따라, S(w_a, w_b)는 워드 a(w_a) 또는 워드 b(w_b)가 출현하는 웹 페이지 중 워드 a(w_a) 및 워드 b(w_b)가 공동으로 출현하는 웹 페이지의 비율을 측정하는 함수이다.

[0035] 예를 들어, 데이터베이스(110)에 웹 페이지 1 내지 3, N-1 및 N에 출현하는 워드들의 출현 횟수가 저장되어 있고, 토픽 키워드 추출부(130)가 워드 1(w_1)과 워드 2(w_2) 간의 유사도 점수(S(w_1, w_2))를 연산하는 경우, 워드 1(w_1)이 출현하는 웹 페이지의 개수가 5이고, 워드 2(w_2)가 출현하는 웹 페이지의 개수가 3이고, 워드 1(w_1)과 워드 2(w_2)가 공동으로 출현하는 웹 페이지의 개수가 3이므로, 유사도 점수(S(w_1, w_2))는 0.75(즉, 2*3/(5+3))로 연산될 수 있다. 마찬가지로, 데이터베이스(110)에 웹 페이지 1 내지 3, N-1 및 N에 출현하는 워드들의 출현 횟수가 저장되어 있고, 토픽 키워드 추출부(130)가 워드 2(w_2)와 워드 3(w_3) 간의 유사도 점수(S(w_2, w_3))를 연산하는 경우, 워드 2(w_2)가 출현하는 웹 페이지의 개수가 3이고, 워드 3(w_3)이 출현하는 웹 페이지의 개수가 2이고, 워드 2(w_2)와 워드 3(w_3)이 공동으로 출현하는 웹 페이지의 개수가 1이므로, 유사도 점수(S(w_2, w_3))는 0.4(즉, 2*1/(3+2))로 연산될 수 있다. 이와 같은 방식으로, 워드 1(w_1)의 유사도 점수들은 0.75, 0.57 및 1로 연산되고, 워드 2(w_2)의 유사도 점수들은 0.75, 0.4 및 0.75로 연산되고, 워드 3(w_3)의 유사도 점수들은 0.57, 0.4 및 0.57로 연산되고, 워드 i(w_i)의 유사도 점수들은 1, 0.75 및 0.57로 연산되며, 이에 따라, 워드 1(w_1)의 최종 유사도 점수는 1, 워드 2(w_2)의 최종 유사도 점수는 0.75, 워드 3(w_3)의 최종 유사도 점수는 0.57 및 워드 i(w_i)의 최종 유사도 점수는 1로 연산된다. 예컨대, 위와 같은 예시에서, j가 2인 경우, 2개의 가장 높은 최종 유사도 점수를 가지는 워드 1(w_1) 및 워드 i(w_i)가 토픽 키워드로 추출될 수 있다.

[0036] 이후, 토픽 키워드 추출부(130)는 각 워드들에 대한 유사도 점수들 중 가장 높은 유사도 점수를 가지는 유사도 점수를 각 워드들에 대한 최종 유사도 점수로 할당하고, 를 연산할 수 있다.

[0037] 한편, 상기 토픽 키워드 추출부(130)는 상기 텍스트 콘텐츠들을 제1 텍스트 콘텐츠들과 제2 텍스트 콘텐츠들로 분류하고, 상기 제1 텍스트 콘텐츠들에 포함된 제1 워드들과 상기 제1 텍스트 콘텐츠들에 포함된 제2 워드들 간의 유사도 점수를 연산할 수 있다. 여기에서, 상기 제1 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들이고, 상기 제2 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들 이외의 콘텐츠들일 수 있다. 여기에서, 타이틀 콘텐츠는 태그 정보가 <title>인 콘텐츠이거나 <head>안에 포함된 콘텐츠일 수 있으며, 타이틀 콘텐츠들 이외의 콘텐츠들은 <body>안에 포함된 콘텐츠일 수 있다.

[0038] 예를 들어, 텍스트 1을 포함하는 텍스트 콘텐츠 1이 <head>안에 포함된 콘텐츠이고, 텍스트 2를 포함하는 텍스트 콘텐츠 2가 <body>안에 포함된 콘텐츠인 경우, 토픽 키워드 추출부(130)는 상기 텍스트 콘텐츠 1을 제1 텍스트 콘텐츠로 분류하고, 상기 텍스트 콘텐츠 2를 제2 텍스트 콘텐츠로 분류할 수 있다. 이에 따라, 토픽 키워드 추출부(130)는 텍스트 1에 포함된 워드들과 텍스트 2에 포함된 워드들 간의 유사도 점수를 연산할 수 있다. 이에 의해, 토픽 키워드 추출부(130)의 유사도 점수의 연산량이 감소할 수 있다.

[0039] 입력 벡터 생성부(140)는 상기 웹 페이지에 포함된 상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터를 생성할 수 있다. 이때, 입력 벡터 생성부(140)는 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성할 수 있다. 여기에서, 상기 입력 벡터의 원소들의 개수는 i개일 수 있다.

[0040] 구체적으로, 입력 벡터 생성부(140)는 상기 웹 페이지에 포함된 상기 워드들 중 워드 k가 상기 토픽 키워드에

포함되면, 제k 강화 계수를 이용하여 상기 워드 k에 대한 원소 값을 연산하고, 상기 제k 워드가 상기 토픽 키워드에 포함되지 않으면, 1의 강화 계수를 이용하여 상기 워드 k에 대한 원소 값을 연산할 수 있다. 여기에서, 상기 k는 1 이상 상기 i 이하의 정수이다. 상기 워드 k는 상기 사전 설정된 특성 단어들 중 특성 단어와 동일한 단어인

[0041] 예를 들어, 워드 1 및 워드 i가 토픽 키워드인 경우, 입력 벡터 생성부(140)는 제1 및 제i 강화 계수를 이용하여 워드 1 및 워드 i에 대한 원소 값을 연산하고, 1의 강화 계수를 이용하여 워드 2 및 워드 3에 대한 원소 값을 연산할 수 있다.

[0042] 이때, 입력 벡터 생성부(140)는 하기 수학식 2를 이용하여 상기 워드 k에 대한 원소 값을 연산할 수 있다.

수학식 2

[0043]
$$z_{ky} = \lambda_k \cdot \frac{f_{ky}}{\sum_{k=1}^{i} f_{ky}} \cdot \log\left(\frac{N}{1+\sum_{y=1}^{N}\delta(f_{ky})}\right)$$

[0044] 여기에서, z_{ky}는 웹 페이지 y(p_y)의 상기 워드 k(w_k)에 대한 원소 값, 즉, 웹 페이지 y(p_y)의 k번째 원소 값이고, k는 1 이상 i 이하의 정수이고, y는 1 이상 N 이하의 정수이며, λ_k는 워드 k(w_k)의 강화 계수로서, 워드 k(w_k)가 토픽 키워드인 경우 제k 강화 계수의 값을 가지고, 토픽 키워드가 아닌 경우 1의 값을 가지며, f_{ky}는 웹 페이지 y(p_y)에서 워드 k(w_k)의 출현 횟수이고, $\sum_{k=1}^{i} f_{ky}$는 웹 페이지 y(p_y)에서 워드의 출현 횟수(f_{ky})의 총 합이고, N은 웹 페이지의 개수이고, $\sum_{y=1}^{N}\delta(f_{ky})$는 워드 k($w_k$)가 출현한 웹 페이지의 개수이다. 또한, $\frac{f_{ky}}{\sum_{k=1}^{i} f_{ky}}$는 상기 워드 k($w_k$)의 출현 빈도이고, $\log\left(\frac{N}{1+\sum_{y=1}^{N}\delta(f_{ky})}\right)$는 상기 워드 k($w_k$)의 역문서 빈도이다.

[0045] 예를 들어, 데이터베이스(110)에 웹 페이지 1 내지 3, N-1 및 N에 출현하는 워드들의 출현 횟수가 저장되어 있고, 제1 및 제i 강화 계수가 각각 2 및 3으로 저장되어 있는 경우, 수학식 2를 참조하면, 현재 선택된 웹 페이지, 즉, 웹 페이지 N(p_N)의 상기 워드 1(w_1)에 대한 1번째 원소 값은 -0.029(2 * 2/11 * log (5 / (1 + 5)))이고, 상기 워드 2(w_2)에 대한 2번째 원소 값은 0(1 * 0/11 * log (5 / (1 + 3))), 상기 워드 3(w_3)에 대한 3번째 원소 값은 0.020(1 * 1/11 * log (5 / (1 + 2)))이며, 상기 워드 1(w_1)에 대한 1번째 원소 값은 -0.021(3 * 1/11 * log (5 / (1 + 5)))로 연산될 수 있다(소수 4번째 자리에서 반올림함).

[0046] 이후, 입력 벡터 생성부(140)는 상기 연산된 원소들의 값을 이용하여 상기 입력 벡터를 생성한다. 예를 들어, 현재 선택된 웹 페이지, 즉, 웹 페이지 N(p_N)의 입력 벡터는 (-0.029, 0, 0.020, -0.021)로 생성될 수 있다.

[0047] 다음으로, 클래스 분류부(150)는 상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류할 수 있다.

[0048] 구체적으로, 클래스 분류부(150)는 상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 분류 값들 중 제s 클래스와 상기 제s 클래스 이외의 클래스들 각각을 분류하기 위한 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제s 분류 값들을 합산하여 제s 합산 값을 연산함으로써 제1 내지 제L 합산 값들을 생성하며, 상기 제1 내지 제L 합산 값들 중 가장 큰 값을 가지는 합산 값에 상응하는 클래스를 상기 웹 페이지의 클래스로 분류할 수 있다. 여기에서, 상기 s는 1 이상 상기 L 이하의 정수이다.

[0049] 예를 들어, 클래스 분류부(150)는 상기 분류 기준 벡터들(V_{12}, V_{13}, V_{14}, V_{23}, V_{24}, V_{34}, V_{21}, V_{31}, V_{41}, V_{32}, V_{42}, V_{43}) 각각과 상기 입력 벡터 간의 내적 곱을 수행함으로써 -0.029, 0, 0.020, -0.021, -0.009, -0.021, 0.029, 0, -0.020, 0.021, 0.009 및 0.021의 분류 값들이 연산될 수 있다. 이후, 클래스 분류부(150)는 제1 클래스와 상기 제1 클래스 이외의 클래스들(즉, 제2 내지 4 클래스들) 각각을 분류하기 위한 분류 기준 벡터들(V_{12}, V_{13}, V_{14})과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제S 분류 값들(즉, -0.029, 0, 0.020)을 합산하여 제1 합산 값(즉, -0.009)을 생성하며, 제2 클래스와 상기 제2 클래스 이외의 클래스들(즉, 제1, 3 및 4 클래스들) 각각을 분류하기 위한 분류 기준 벡터들(V_{21}, V_{23}, V_{24})과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제S 분류 값들(즉, 0.029, -0.021, -0.009)을 합산하여 제2 합산 값(즉, -0.001)을 생성하며, 마찬가지의 방식으로, 제3 및 제4 합산 값들(0 및 0.009)을 생성한다. 이후, 클래스 분류부(150)는 제1 내지 제4 합산 값들 중 가장 큰 값을 가지는 합산 값(즉, 제4 합산 값)에 상응하는 제4 클래스를 상기 웹 페이지의 클래스로 분류할 수 있다.

[0050] 상기 클래스 분류부(150)는 상기 분류 기준 벡터들 각각과 상기 입력 벡터 간의 내적 곱을 수행한 후 상기 분류 기준 조절 값들(B_{12}, B_{13}, B_{14}, B_{23}, B_{24}, B_{34}, B_{21}, B_{31}, B_{41}, B_{32}, B_{42} 및 B_{43}) 중 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값을 가산함으로써 분류 값들을 연산할 수 있다. 이후, 상기 클래스 분류부(150)는 연산된 제s 분류 값들을 합산하여 제s 합산 값을 연산함으로써 제1 내지 제L 합산 값들을 생성할 수 있다.

[0051] 예를 들어, 래스 분류부(150)는 제1 클래스와 상기 제1 클래스 이외의 클래스들(즉, 제2 내지 4 클래스들) 각각을 분류하기 위한 분류 기준 벡터들(V_{12}, V_{13}, V_{14})과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 값에 제1 내지 제3 분류 기준 조절 값들(B_{12}, B_{13}, B_{14})을 각각 가산함으로써 분류 값들을 연산할 수 있다. 이후, 상기 클래스 분류부(150)는 분류 기준 조절 값들이 가산되어 연산된 분류 값들에 기초하여 상기 웹 페이지의 클래스를 분류할 수 있다.

[0052] 상술한 바와 같은 본 발명의 일실시예에 따른 웹 페이지 분류 장치(100)는 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출하고, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 입력 벡터를 생성하며, 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류함으로써, 자동으로 웹 페이지를 주제 별로 분류하는 것을 가능하게 할 수 있다.

[0053] 또한, 상술한 바와 같은 본 발명의 일실시예에 따른 웹 페이지 분류 장치(100)는 자동으로 웹 페이지를 주제 별로 분류함으로써 웹 페이지를 분류하는데 적은 시간과 비용이 소모되며, 정확도가 증가하며, 이에 따라 분류된 웹 페이지를 이용하여 사용자들은 웹 페이지 탐색 시간을 감소시킬 수 있다.

[0055] 또한, 본 발명의 일실시예에 따른 웹 페이지 분류 장치(100)는 특성 단어 설정부(160)를 더 포함할 수 있다.

[0056] 특성 단어 설정부(160)는 현존하는 단어들, 예컨대, 사전에 등록된 단어들 중 i개의 특성 단어를 설정할 수 있다. 예를 들어, 특성 단어 설정부(160)는 사과를 특성 단어로 설정하고, 배를 특성 단어로 설정하지 아니할 수 있다.

[0057] 구체적으로, 특성 단어 설정부(160)는 단어들과 클래스들 간의 상관관계 추론 기법에 기초하여 i개의 특성 단어를 설정할 수 있다. 여기에서, 상관관계 추론 기법은 임의 단어와 클래스들 각각 간의 상관관계를 측정하여 상기 임의 단어와 클래스들 각각 간의 연관 정도를 판단하는 기법으로서, 상관관계 추론 기법에 의해 획득되는 상기 임의 단어와 클래스들 각각 간의 상관계수는 -1에서 1 사이의 값을 가지고, 상관계수의 절대 값이 1에 가까울수록 클래스 간의 상관관계가 높다고 판단할 수 있다.

[0058] 구체적으로, 특성 단어 설정부(160)는 하기 수학식 3을 이용하여 임의 단어와 임의 클래스 간의 상관계수를 연산할 수 있다.

수학식 3

[0059]
$$\rho(w_g, c_h) = \frac{N \cdot (\sum_{y=1}^{N} z_{gy} \cdot c_{hy}) - (\sum_{y=1}^{N} z_{gy}) \cdot (\sum_{y=1}^{N} c_{hy})}{\sqrt{\left[N \cdot \sum_{y=1}^{N} z_{gy}^2 - (\sum_{y=1}^{N} z_{gy})^2\right] \cdot \left[N \cdot \sum_{y=1}^{N} c_{hy}^2 - (\sum_{y=1}^{N} c_{hy})^2\right]}}$$

[0060] 여기에서, $\rho(w_g, c_h)$는 상기 워드 g(w_g)와 클래스 h(c_h) 간의 상관계수이고, g는 1 이상의 정수이고, h는 1 이상 L 이하의 정수이며, z_{gy}는 웹 페이지 y(p_y)의 상기 워드 g(w_g)에 대한 원소 값이고, c_{hy}는 웹 페이지 y(p_y)가 클래스 h(c_h)에 속하면 1의 값을 가지고, 속하지 않으면 0의 값을 가진다.

[0061] 이후, 특성 단어 설정부(160)는 상기 수학식 3을 이용하여 임의 단어와 임의 클래스 간에 연산된 L개의 상관계수들 중 가장 높은 값을 가지는 상관계스를 해당 임의 단어의 최종 상관계수로 설정한다.

[0062] 이후, 특성 단어 설정부(160)는 각 단어들의 최종 상관계수에 기초하여 최종 상관계수의 값이 높은 순서대로 i개의 단어들을 특성 단어로 설정할 수 있다.

[0063] 도 2는 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법을 도시한 순서도이다.

[0064] 도 2에 도시된 바와 같이, 단계(S210)에서는 사전 설정된 L개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들이 저장되어 있는 데이터베이스를 유지한다.

[0065] 단계(S210)에서는 제1 내지 제i 강화 계수를 저장하고 있는 상기 데이터베이스를 유지할 수 있다.

[0066] 단계(S210)에서는 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값들을 더 저장하고 있는 상기 데이터베이스를 유지할 수 있다.

[0067] 단계(S220)에서는 상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출한다.

[0068] 단계(S220)에서는 상기 콘텐츠들의 태그 정보에 기초하여 상기 콘텐츠들에서 상기 텍스트 콘텐츠들을 추출할 수 있다.

[0069] 단계(S220)에서는 상기 텍스트 콘텐츠들에 포함된 워드들 간의 유사도 점수를 연산한 후 워드들 각각에 대하여 해당 워드와 다른 워드들 간에 연산된 유사도 점수들 중 가장 높은 유사도 점수를 해당 워드의 최종 유사도 점수로 할당하고, 최종 유사도 점수가 높은 j개의 워드를 토픽 키워드로 추출할 수 있다.

[0070] 단계(S220)에서는 상기 텍스트 콘텐츠들을 제1 텍스트 콘텐츠들과 제2 텍스트 콘텐츠들로 분류하고, 상기 제1 텍스트 콘텐츠들에 포함된 제1 워드들과 상기 제1 텍스트 콘텐츠들에 포함된 제2 워드들 간의 유사도 점수를 연산할 수 있다. 여기에서, 상기 제1 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들이고, 상기 제2 텍스트 콘텐츠들은 상기 웹 페이지의 타이틀 콘텐츠들 이외의 콘텐츠들일 수 있다.

[0071] 단계(S220)에서는 상기 콘텐츠들 중 상기 태그 정보가 텍스트 태그, 테이블 태그, 헤더 태그 및 목록 태그 중 어느 하나의 태그 정보를 가지는 콘텐츠를 텍스트 콘텐츠로 추출할 수 있다.

[0072] 단계(S230)에서는 상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성한다.

[0073] 단계(S230)에서는 상기 워드들 중 제k 워드가 상기 토픽 키워드에 포함되면, 제k 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산하고, 상기 제k 워드가 상기 토픽 키워드에 포함되지 않으면, 1의 강화 계수를 이용하여 상기 제k 워드에 대한 원소 값을 연산할 수 있다.

[0074] 단계(S230)에서는 상기 워드들 중 상기 제k 워드의 출현 빈도, 상기 제k 워드의 역문서 빈도 및 상기 제k 워드가 상기 토픽 키워드에 포함되는지 여부에 따라 결정되는 강화 계수를 곱함으로써 상기 제k 워드에 대한 원소 값을 연산할 수 있다.

[0075] 단계(S240)에서는 상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류한다.

[0076] 단계(S240)에서는 상기 연산된 분류 값들 중 제s 클래스와 상기 제s 클래스 이외의 클래스들 각각을 분류하기 위한 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 연산된 제s 분류 값들을 합산하여 제s 합산 값을 연산함으로써 제1 내지 제L 합산 값들을 생성하며, 상기 제1 내지 제L 합산 값들 중 가장 큰 값을 가지는 합산 값에 상응하는 클래스를 상기 웹 페이지의 클래스로 분류할 수 있다.

[0077] 단계(S240)에서는 상기 분류 기준 벡터들 각각과 상기 입력 벡터 간의 내적 곱을 수행한 후 상기 분류 기준 조절 값들 중 상기 분류 기준 벡터들 각각에 매칭되어 있는 분류 기준 조절 값을 가산함으로써 분류 값들을 연산할 수 있다.

[0078] 이상, 도 2를 참조하여 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법에 대해 설명하였다. 여기서, 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법은 도 1을 이용하여 설명한 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작에 대한 구성과 대응될 수 있으므로, 이에 대한 보다 상세한 설명은 생략하기로 한다.

[0079] 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법은 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다.

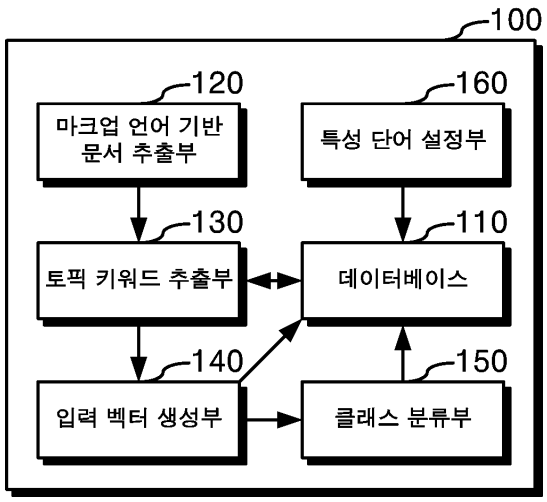
[0080] 또한, 본 발명의 일실시예에 따른 웹 페이지 분류 장치의 동작 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, 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.

[0081] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

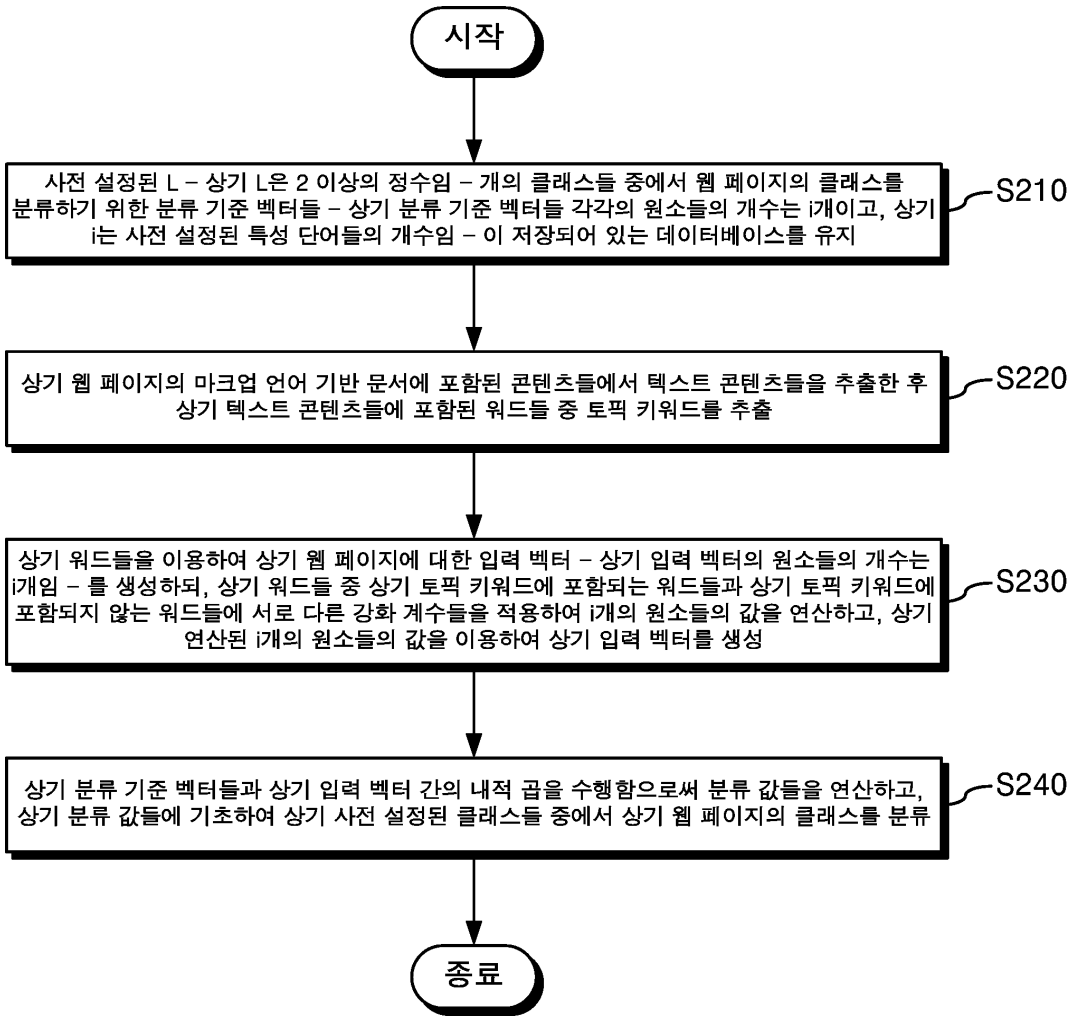
[0082] 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

도면

도면1


100
120
마크업 언어 기반 문서 추출부
160
특성 단어 설정부
130
토픽 키워드 추출부
110
데이터베이스
140
입력 벡터 생성부
150
클래스 분류부

도면2


시작
사전 설정된 L – 상기 L은 2 이상의 정수임 – 개의 클래스들 중에서 웹 페이지의 클래스를 분류하기 위한 분류 기준 벡터들 – 상기 분류 기준 벡터들 각각의 원소들의 개수는 i개이고, 상기 i는 사전 설정된 특성 단어들의 개수임 – 이 저장되어 있는 데이터베이스를 유지
S210
상기 웹 페이지의 마크업 언어 기반 문서에 포함된 콘텐츠들에서 텍스트 콘텐츠들을 추출한 후 상기 텍스트 콘텐츠들에 포함된 워드들 중 토픽 키워드를 추출
S220
상기 워드들을 이용하여 상기 웹 페이지에 대한 입력 벡터 – 상기 입력 벡터의 원소들의 개수는 i개임 – 를 생성하되, 상기 워드들 중 상기 토픽 키워드에 포함되는 워드들과 상기 토픽 키워드에 포함되지 않는 워드들에 서로 다른 강화 계수들을 적용하여 i개의 원소들의 값을 연산하고, 상기 연산된 i개의 원소들의 값을 이용하여 상기 입력 벡터를 생성
S230
상기 분류 기준 벡터들과 상기 입력 벡터 간의 내적 곱을 수행함으로써 분류 값들을 연산하고, 상기 분류 값들에 기초하여 상기 사전 설정된 클래스들 중에서 상기 웹 페이지의 클래스를 분류
S240
종료