(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0049659
(43) 공개일자 2014년04월28일

(51) 국제특허분류(Int. Cl.)
G06F 17/21 (2006.01) *G06F 17/27* (2006.01)
(21) 출원번호 10-2012-0115730
(22) 출원일자 2012년10월18일
심사청구일자 2012년10월18일

(71) 출원인
조선대학교산학협력단
광주광역시 동구 필문대로 309 (서석동)
(72) 발명자
김판구
광주광역시 북구 설죽로 600 삼호아파트 101-904
최동진
광주광역시 남구 천변좌로370번길 13
(뒷면에 계속)
(74) 대리인
특허법인 누리

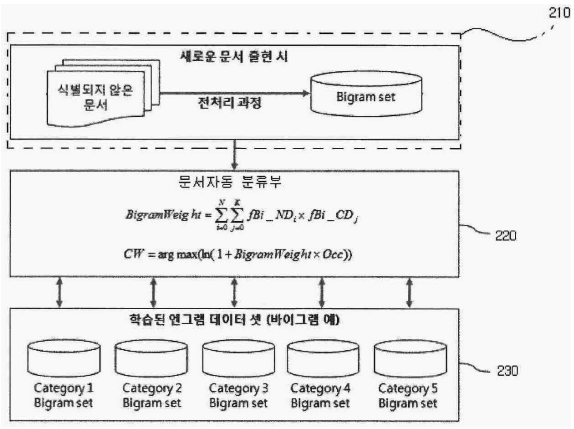
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법, 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법**

(57) 요 약

본 발명은 엔그램 기반의 문서 자동 학습 및 분류 과정을 통해 웹상의 대용량 문서들을 자동으로 학습 및 분류할 수 있도록 하는 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법, 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법에 관한 것으로서, 본 발명에 따른 문서의 자동 분류 장치는 카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀과, 상기 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부와, 상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하며, 상기 학습문서 풀을 통해 식별되지 않는 신규 문서 출현 시 상기 전처리부가 해당 신규문서를 전처리하여 바이그램 세트를 형성하는 문서 자동 학습부와, 상기 전처리부를 통해 형성되는 상기 신규문서의 바이그램 세트와 상기 엔그램 데이터 세트 풀의 바이그램 세트를 비교하여 상기 신규문서의 바이그램 세트를 상기 엔그램 데이터 세트 풀 중 어느 하나의 엔그램 데이터 세트에 할당하여 저장하는 문서 자동 분류부를 포함하여 형성된다.

대 표 도 - 도2



(72) 발명자

김정인

광주광역시 광산구 첨단중앙로 104 건영아파트 103동 903호

고미아

광주광역시 서구 하남대로710번길 20 우미린아파트 507동 2001호

이 발명을 지원한 국가연구개발사업

과제고유번호	2012-0230101
부처명	교육과학기술부
연구사업명	조선대학교 산학협력 선도대학(LINC) 육성사업
연구과제명	조선대학교 산학협력 선도대학(LINC) 육성사업지원 산학협력실험실
기 여 율	1/1
주관기관	조선대학교
연구기간	2012.06.18 ~ 2013.01.31

특허청구의 범위

청구항 1

카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀;

상기 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부;

상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하는 문서의 자동 학습 장치.

청구항 2

제 1 항에 있어서, 상기 전처리부는

상기 학습문서 그룹의 각 학습문서 별로 특수문자를 제거하는 특수문자 제거부;

상기 학습문서 그룹의 각 학습문서 별로 불용어(stopword)를 제거하는 불용어 제거부;

상기 학습문서 그룹의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 품사 태거부;

상기 품사 태거부를 거친 학습문서로부터 명사 및 동사를 추출하는 품사 추출부를 포함하는 것을 특징으로 하는 문서의 자동 학습 장치.

청구항 3

제 1 항에 있어서,

상기 엔그램 데이터 세트는 카테고리 별로 바이그램(Bigram) 세트를 생성하여 저장하는 것을 특징으로 하는 문서의 자동 학습 장치.

청구항 4

제 3 항에 있어서,

상기 바이그램 세트는 어절 단위 바이그램(Bigram)의 추출 및 구축을 통해 생성되는 것을 특징으로 하는 문서의 자동 학습 장치.

청구항 5

복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹이 형성되는 단계;

상기 복수의 학습문서 그룹이 개별적으로 전처리되는 단계;

상기 전처리 과정을 통해 상기 복수의 학습문서 그룹으로부터 엔그램 데이터 세트가 개별 생성되는 단계를 포함하는 문서의 자동 학습 방법.

청구항 6

제 5 항에 있어서,

상기 복수의 학습문서 그룹이 개별적으로 전처리되는 단계는

상기 학습문서 그룹의 각 학습문서 별로 특수문자가 제거되는 단계;

상기 학습문서 그룹의 각 학습문서 별로 불용어(stopword)가 제거되는 단계;

상기 학습문서 그룹의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 단계;

상기 형태소 분석 및 일부 품사를 가려내는 과정을 거친 학습문서로부터 명사 및 동사를 추출하여 해당 명사 및 동사의 동시출현 빈도의 수를 추출하는 단계;

상기 동시출현 빈도의 수를 기준으로 엔그램 데이터 세트를 구축하는 단계를 포함하는 것을 특징으로 하는 문서의 자동 학습 방법.

청구항 7

제 6 항에 있어서,

상기 동시출현 빈도의 수를 기준으로 엔그램 데이터 세트를 구축하는 단계는

상기 학습문서 그룹 별로 상기 전처리 과정을 통해 바이그램 데이터 세트가 구축되는 단계인 것을 특징으로 하는 문서의 자동 학습 방법.

청구항 8

카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀과, 상기 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부와, 상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하며, 상기 학습문서 풀을 통해 식별되지 않는 신규문서 출현 시 상기 전처리부가 해당 신규문서를 전처리하여 바이그램 세트를 형성하는 문서 자동 학습부;

상기 전처리부를 통해 형성되는 상기 신규문서의 바이그램 세트와 상기 엔그램 데이터 세트 풀의 바이그램 세트를 비교하여 상기 신규문서의 바이그램 세트를 상기 엔그램 데이터 세트 풀 중 어느 하나의 엔그램 데이터 세트에 할당하여 저장하는 문서 자동 분류부를 포함하는 문서의 자동 분류 장치.

청구항 9

제 8 항에 있어서,

상기 문서 자동 분류부는 아래의 식

$$Bigram\ Weight = \sum_{i=0}^{N}\sum_{j=0}^{K} fBi_ND_i \times fBi_CD_j$$

$$CW = \mathrm{argmax}(\ln(1 + Bigram\ Weight \times Occ))$$

『

Bigram Weight : 상기 신규문서의 바이그램 세트와 상기 엔그램 데이터 세트 풀의 각 엔그램 데이터 세트별 바이그램 세트 간의 유사도로써, 상기 신규문서에서 추출된 바이그램 데이터가 상기 엔그램 데이터 세트별 바이그램에 존재 시, 해당 바이그램 단어의 출현 빈도의 수를 곱한 값.

CW : 기 연산된 Bigram Weight에 상기 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값』

을 통해 상기 엔그램 데이터 세트 풀의 바이그램 세트와 상기 신규문서의 바이그램 세트 간 유사도를 측정하는 것을 특징으로 하는 문서의 자동 분류 장치.

청구항 10

제 9 항에 있어서,

상기 CW는 상기 기 연산된 Bigram Weight에 상기 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값 중 최대값에 할당되는 것을 특징으로 하는 문서의 자동 분류 장치.

청구항 11

제 8 항에 있어서, 상기 전처리부는

상기 학습문서 그룹의 각 학습문서 별로 특수문자를 제거하는 특수문자 제거부;

상기 학습문서 그룹의 각 학습문서 별로 불용어(stopword)를 제거하는 불용어 제거부;

상기 학습문서 그룹의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 품사 태거부;

상기 품사 태거부를 거친 학습문서로부터 명사 및 동사를 추출하는 품사 추출부를 포함하는 것을 특징으로 하는 문서의 자동 분류 장치.

청구항 12

제 8 항에 있어서,

상기 문서 자동 분류부는 단어의 공기정보와 출현빈도의 수 및 엔그램 출현빈도의 수를 기준으로 상기 신규문서의 학습 및 해당 학습을 통해 상기 엔그램 데이터 세트 풀 중 어느 한 엔그램 데이터 세트로의 분류를 결정하는 것을 특징으로 하는 문서의 자동 분류 장치.

청구항 13

복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹이 형성되는 단계;

상기 학습문서 그룹 별로 특수문자와 불용어(stopword) 제거 및 형태소 분석을 통해 명사 및 동사만을 추출하는 전처리 단계;

상기 전처리를 통해 상기 복수의 학습문서 그룹으로부터 엔그램 데이터 세트가 개별 생성되는 단계;

상기 엔그램 데이터 세트를 통해 식별되지 않는 신규문서가 출현하는 단계;

상기 신규문서를 상기 전처리와 동일한 과정으로 전처리하여 상기 신규문서의 해당 바이그램을 형성하는 단계;

상기 신규문서의 바이그램과 상기 엔그램 데이터 세트의 바이그램을 비교하여 해당 신규문서를 상기 학습문서 그룹들 중 어느 하나에 할당하여 분류하는 단계를 포함하는 것을 특징으로 하는 문서의 자동 분류 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법에 관한 것으로서, 특히 엔그램 기반의 문서 자동 학습 및 분류 과정을 통해 웹상의 대용량 문서들을 자동으로 학습 및 분류할 수 있도록 하는 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법에 관한 것이다.

배 경 기 술

[0002] 문서 분류는 과거 도서관 관리시스템에서 시작되어 현재까지 지속적으로 연구 개발되고 있는 분야이며, 대용량의 문서들을 효율적으로 관리하고 색인하기 위해 초기의 수작업을 시작으로 현재의 컴퓨터를 이용한 통계적 및 의미적 기법으로 발달하였다.

[0003] 특히 인터넷과 스마트폰의 급격한 발달로 인하여 웹상에 존재하는 정보는 대량으로 증가되었고, 따라서 이러한 웹상의 정보들을 그 사용자들이 효율적으로 관리 및 검색하는 것은 매우 중요한 문제가 되었다.

[0004] 예를 들면, 최근에 웹사이트 상에서 무수한 웹문서 정보를 처리하여 사용자의 요구에 해당되는 정보만을 추출 후 이를 사용자에게 제공하는 정보 검색 시스템이 널리 이용되고 있고, 또한 검색어로써 단어 형식이 아닌 일상적인 자연어로 인터넷 검색을 할 수 있는 인터넷 정보 검색 시스템도 이용되고 있다. 그리고 이와 같은 정보 검색 시스템의 기능 향상을 위해서는 상기와 같은 문서 분류 기술의 개발이 필수적 요건이라 할 수 있다.

[0005] 현재 널리 사용되고 있는 문서 분류 기술에 대해 간략히 설명하면, 문서에서 출현하는 단어의 빈도를 바탕으로 이를 벡터 공간에 매핑시켜 문서 간 거리를 측정 후 분류하는 방법, 그리고 나이브 베이즈(Naive Bayes) 이론을 적용한 확률적 모델 등이 있다.

[0006] 그러나 상기와 같은 종래의 문서 분류 기술은 자연어를 대상으로 할 경우, 단어 빈도의 수로 한정하는 기법이 자연어에 대해서는 그 신뢰성에 한계점을 보이게 된다. 이는 상술한 인터넷 정보 검색 시스템의 자연어를 이용한 검색 과정에서 검색 성능의 저하 요인이 되기도 한다.

[0007] 상기와 같은 이유로 WordNet과 같은 지식 데이터베이스를 활용한 방법이 제안되기도 하였지만, 이와 같은 지식 데이터베이스를 활용한 방법은 해당 지식 데이터베이스의 신뢰성에 따라 문서 분류의 정확성이 크게 좌우되는 문제가 있다.

[0008] 이에 본 출원인은 엔그램(N-gram)이 문서 내 공기정보와 출현 빈도의 수를 병합한 기법으로 음성 인식과 자연어 처리 분야에서 활발히 사용되고 있는 방법으로써, 동시 출현하는 단어와의 연관성과 중심어 및 핵심어, 대표어를 선정하는데 효과적으로 활용될 수 있는 점에 착안하여 본 발명을 제안하게 되었다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국등록특허 제10-0842080호, “문서의 그룹별 분류방법”

(특허문헌 0002) 한국등록특허 제10-1092352호, “문장 코퍼스에 대한 영역 자동분류 방법 및 장치”

발명의 내용

해결하려는 과제

[0010] 본 발명은 엔그램 기반의 문서 자동 학습 및 분류 과정을 통해 웹상의 대용량 문서들을 자동으로 학습 및 분류할 수 있도록 하는 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법을 제공하는데 목적이 있다.

[0011] 또한, 본 발명은 웹문서를 미리 정의된 분야별로 학습단계를 거친 후 식별되지 않은 신규문서 출현 시, 이를 자동 분류하는 방법을 통하여 웹문서의 의미적 분석과 함께 사용자의 질의어와 밀접한 검색결과를 제공할 수 있는 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법을 제공하는데 목적이 있다.

[0012] 또한, 본 발명은 동시에 출현하는 단어들의 집합인 엔그램과 이의 출현빈도수를 바탕으로 문서를 대표할 수 있는 엔그램을 구축하여 차후에 식별되지 않은 신규문서가 발견되었을 때 기 구축된 엔그램 정보를 바탕으로 신규문서를 자동으로 분류할 수 있도록 하는 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법을 제공하는데 목적이 있다.

과제의 해결 수단

[0013] 상기와 같은 목적을 달성하기 위해 본 발명에 따른 문서의 자동 학습 장치는, 카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀과, 상기 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부와, 상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하여 구성된다.

[0014] 또한, 상기 전처리부는 상기 학습문서 그룹의 각 학습문서 별로 특수문자를 제거하는 특수문자 제거부와, 상기 학습문서 그룹의 각 학습문서 별로 불용어(stopword)를 제거하는 불용어 제거부와, 상기 학습문서 그룹의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 품사 태거부와, 상기 품사 태거부를 거친 학습문서로부터 명사 및 동사를 추출하는 품사 추출부를 포함하는 것을 특징으로 한다.

[0015] 또한, 상기 엔그램 데이터 세트는 카테고리 별로 바이그램(Bigram) 세트를 생성하여 저장하는 것을 특징으로 한다.

[0016] 또한, 상기 바이그램 세트는 어절 단위 바이그램(Bigram)의 추출 및 구축을 통해 생성되는 것을 특징으로 한다.

[0017] 또한, 본 발명에 따른 문서 자동 학습 방법은 복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹이 형성되는 단계와, 상기 복수의 학습문서 그룹이 개별적으로 전처리되는 단계와, 상기 전처리 과정을 통해 상기 복수의 학습문서 그룹으로부터 엔그램 데이터 세트가 개별 생성되는 단계를 포함하여 구성된다.

[0018] 또한, 상기 복수의 학습문서 그룹이 개별적으로 전처리되는 단계는 상기 학습문서 그룹의 각 학습문서 별로 특수문자가 제거되는 단계와, 상기 학습문서 그룹의 각 학습문서 별로 불용어(stopword)가 제거되는 단계와, 상기

학습문서 그룹의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 단계와, 상기 형태소 분석 및 일부 품사를 가려내는 과정을 거친 학습문서로부터 명사 및 동사를 추출하여 해당 명사 및 동사의 동시출현 빈도의 수를 추출하는 단계와, 상기 동시출현 빈도의 수를 기준으로 엔그램 데이터 세트를 구축하는 단계를 포함하는 것을 특징으로 한다.

[0019] 또한, 상기 동시출현 빈도의 수를 기준으로 엔그램 데이터 세트를 구축하는 단계는, 상기 학습문서 그룹 별로 상기 전처리 과정을 통해 바이그램 데이터 세트가 구축되는 단계인 것을 특징으로 한다.

[0020] 또한, 본 발명에 따른 문서의 자동 분류 장치는, 카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀 상기 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부 상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하며 상기 학습문서 풀을 통해 식별되지 않는 신규문서 출현 시 상기 전처리부가 해당 신규문서를 전처리하여 바이그램 세트를 형성하는 문서 자동 학습부와, 상기 전처리부를 통해 형성되는 상기 신규문서의 바이그램 세트와 상기 엔그램 데이터 세트 풀의 바이그램 세트를 비교하여 상기 신규문서의 바이그램 세트를 상기 엔그램 데이터 세트 풀 중 어느 하나의 엔그램 데이터 세트에 할당하여 저장하는 문서 자동 분류부를 포함하여 구성된다.

[0021] 또한, 상기 문서 자동 분류부는 아래의 식

$$Bigram\ Weight = \sum_{i=0}^{N}\sum_{j=0}^{K} fBi_ND_i \times fBi_CD_j$$

$$CW = \mathrm{argmax}(\ln(1 + Bigram\ Weight \times Occ))$$

[0022] 『

[0023] Bigram Weight : 상기 신규문서의 바이그램 세트와 상기 엔그램 데이터 세트 풀의 각 엔그램 데이터 세트별 바이그램 세트 간의 유사도로써, 상기 신규문서에서 추출된 바이그램 데이터가 상기 엔그램 데이터 세트별 바이그램에 존재 시, 해당 바이그램 단어의 출현 빈도의 수를 곱한 값.

[0024] CW : 기 연산된 Bigram Weight에 상기 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값』

[0025] 을 통해 상기 엔그램 데이터 세트 풀의 바이그램 세트와 상기 신규문서의 바이그램 세트 간 유사도를 측정하는 것을 특징으로 한다.

[0026] 또한, 상기 CW는 상기 기 연산된 Bigram Weight에 상기 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값 중 최대값에 할당되는 것을 특징으로 한다.

[0027] 또한, 상기 문서 자동 분류부는 단어의 공기정보와 출현빈도의 수 및 엔그램 출현빈도의 수를 기준으로 상기 신규문서의 학습 및 해당 학습을 통해 상기 엔그램 데이터 세트 풀 중 어느 한 엔그램 데이터 세트로의 분류를 결정하는 것을 특징으로 한다.

[0028] 또한, 본 발명에 따른 문서 자동 분류 방법은 복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹이 형성되는 단계와, 상기 학습문서 그룹 별로 특수문자와 불용어(stopword) 제거 및 형태소 분석을 통해 명사 및 동사만을 추출하는 전처리 단계와, 상기 전처리를 통해 상기 복수의 학습문서 그룹으로부터 엔그램 데이터 세트가 개별 생성되는 단계와, 상기 엔그램 데이터 세트를 통해 식별되지 않는 신규문서가 출현하는 단계와, 상기 신규문서를 상기 전처리와 동일한 과정으로 전처리하여 상기 신규문서의 해당 바이그램을 형성하는 단계와, 상기 신규문서의 바이그램과 상기 엔그램 데이터 세트의 바이그램을 비교하여 해당 신규문서를 상기 학습문서 그룹들 중 어느 하나에 할당하여 분류하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0029] 본 발명에 따르면, 엔그램 기반의 문서 자동 학습 및 분류 과정을 통해 웹상의 대용량 문서들을 자동으로 학습 및 분류할 수 있게 된다.

[0030] 또한, 웹문서를 미리 정의된 분야별로 학습단계를 거친 후 식별되지 않은 신규문서 출현시, 이를 자동 분류하는 방법을 통하여 웹문서의 의미적 분석과 함께 사용자의 질의어와 밀접한 검색결과를 제공할 수 있다.

[0031] 또한, 동시에 출현하는 단어들의 집합인 엔그램과 이의 출현빈도수를 바탕으로 문서를 대표할 수 있는 엔그램을 구축하여 차후에 식별되지 않은 신규문서가 발견되었을 때 기 구축된 엔그램 정보를 바탕으로 신규문서를 자동으로 분류할 수 있다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 일 실시예에 따른 문서의 자동 학습 장치를 보인 블록도

도 2는 본 발명의 일 실시예에 따른 문서의 자동 분류 장치를 개념적으로 보인 블록도

도 3은 본 발명의 일 실시예에 따른 문서 자동 학습 방법을 보인 흐름도

도 4는 본 발명의 일 실시예에 따른 문서 자동 분류 방법을 보인 흐름도

발명을 실시하기 위한 구체적인 내용

[0033] 이하에서는, 첨부된 도면을 참조하여 본 발명의 일 실시예에 따른 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법을 상세하게 설명한다.

[0034] 도 1은 본 발명의 일 실시예에 따른 문서의 자동 학습 장치를 보인 블록도이다.

[0035] 도시된 바와 같이, 본 발명의 일 실시예에 따른 문서의 자동 학습 장치(100)는 학습문서 풀(110), 전처리부(120), 엔그램 데이터 세트 풀(130)을 포함하여 구성된다.

[0036] 학습문서 풀(110)은 카테고리 별로 분류된 복수의 학습문서 그룹(111)을 포함하여 형성된다.

[0037] 전처리부(120)는 학습문서 풀(110)의 각 학습문서 그룹에 대해 전처리 과정을 하며, 이와 같은 전처리부(120)는 특수문자 제거부(121), 불용어(stopword) 제거부(122), 품사 태거부(123), 품사 추출부(124)를 포함하여 형성된다.

[0038] 특수문자 제거부(121)는 학습문서 그룹(111)의 각 학습문서 별로 특수문자를 제거하며, 불용어(stopword) 제거부(122)는 학습문서 그룹(111)의 각 학습문서 별로 불용어(stopword)를 제거한다. 그리고 품사 태거부(123)는 학습문서 그룹(111)의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내며, 품사 추출부(124)는 품사 태거부(123)를 거친 학습문서로부터 명사 및 동사를 추출한다.

[0039] 엔그램 데이터 세트 풀(130)은 전처리부(120)의 전처리 과정을 통해 학습되어 형성된 학습문서 풀(110)의 엔그램 데이터 세트(131)가 저장된다. 그리고 엔그램 데이터 세트(131)는 카테고리 별로 바이그램(Bigram) 세트를 생성하여 저장하고, 여기서 바이그램 세트는 어절 단위 바이그램(Bigram)의 추출 및 구축을 통해 생성된다.

[0040] 다음은 도 2를 참조하여 본 발명의 일 실시예에 따른 문서의 자동 분류 장치에 대해 설명한다.

[0041] 도시된 바와 같이, 본 발명의 일 실시예에 따른 문서의 자동 분류 장치(200)는 문서 자동 학습부(210) 및 문서 자동 분류부(220)를 포함하여 구성된다.

[0042] 문서 자동 학습부(210)는 도 1을 참조하여 상술한 문서의 자동 학습 장치(100)에 해당하는 것으로서, 즉 문서 자동 학습부(210)는 카테고리 별로 분류된 복수의 학습문서 그룹이 포함되는 학습문서 풀과, 학습문서 풀의 각 학습문서 그룹에 대해 전처리 과정을 하는 전처리부와, 상기 전처리부의 전처리 과정을 통해 학습되어 형성된 상기 학습문서 풀의 엔그램 데이터 세트가 저장되는 엔그램 데이터 세트 풀을 포함하여 구성된다. 이와 같은 문서 자동 학습부(210)는 상기 구성을 통해 학습문서 풀을 통해 식별되지 않는 신규문서 출현 시 상기 전처리부가 해당 신규문서를 전처리하여 바이그램 세트를 형성한다.

[0043] 그리고 문서 자동 학습부(210)의 전처리부를 포함한 구성들 및 상기 전처리부의 구성은 도 1에 따른 문서의 자동 학습 장치의 전처리부와 동일한 구성이며, 따라서 본 실시예에서 이에 대한 상세 설명 및 도시는 생략한다.

[0044] 문서 자동 분류부(220)는 문서 자동 학습부(210)의 전처리부를 통해 형성되는 신규문서의 바이그램 세트와 엔그램 데이터 세트 풀의 바이그램 세트를 비교하여 신규문서의 바이그램 세트를 엔그램 데이터 세트 풀(230, 문서 자동 학습부(210)의 엔그램 데이터 세트 풀에 해당하며 이해를 위해 별도로 도시하였을 뿐 임) 중 어느 하나의 엔그램 데이터 세트에 할당하여 저장한다.

[0045] 여기서 문서 자동 분류부(220)는 아래의 식을 통해 엔그램 데이터 세트 풀(230)의 바이그램 세트와 신규문서의 바이그램 세트 간 유사도를 측정한다.

[0046] [식 1]

$$Bigram\ Weight = \sum_{i=0}^{N}\sum_{j=0}^{K} fBi_ND_i \times fBi_CD_j$$

$$CW = \mathrm{argmax}(\ln(1 + Bigram\ Weight \times Occ))$$

[0047]

[0048] Bigram Weight : 신규문서의 바이그램 세트와 엔그램 데이터 세트 풀의 각 엔그램 데이터 세트별 바이그램 세트 간의 유사도로써, 신규문서에서 추출된 바이그램 데이터가 엔그램 데이터 세트별 바이그램에 존재 시, 해당 바이그램 단어의 출현 빈도의 수를 곱한 값.

[0049] CW : 기 연산된 Bigram Weight에 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값.

[0050] 다시 말해, 문서 자동 분류부(220)는 단어의 공기정보와 출현빈도의 수 및 엔그램 출현빈도의 수를 기준으로 신규문서의 학습 및 해당 학습을 통해 엔그램 데이터 세트 풀(230) 중 어느 한 엔그램 데이터 세트로의 분류를 결정하는 것이다. 아래의 표 1은 이러한 예를 보인 것이다.

표 1

[0051]

	Bigram Data	*Bigram weight*	*CW*
New Document	chosun university 5, computer science 4, master students 5, ...		
Category 1	chosun university 15, computer science 40, gwangju korea 37 ...	ND*C1 = 5*15+4*40 = 235	1+235*2 = 471
Category 2	republic korea 64, chosun university 5, president korea 41, ...	ND*C2 = 5*5 = 25	1+25*1 = 26
Category 3	chosun university 71, artificial intelligence 41, international conference 13, ...	ND*C3= 5*71= 355	1+355*1 = 356

[0052] 그리고 상기 CW는 기 연산된 Bigram Weight에 신규문서 및 엔그램 데이터 세트 각각에 동시 출현한 바이그램의 횟수를 곱한 값 중 최대값에 할당되는 것일 수 있다.

[0053] 다음은 도 3을 참조하여 본 발명의 일 실시예에 따른 문서 자동 학습 방법에 대해 설명한다. 여기서 문서의 자동 학습 장치는 도 1의 실시예를 이용한 것이며, 따라서 이하의 설명에서 문서의 자동 학습 장치 구성은 도 1의 실시예를 따르는 동시에 동일 부호를 사용한다.

[0054] 먼저, 단계(S110)에서 복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹(111)이 형성된다.

[0055] 이어서, 단계(S120)에서 복수의 학습문서 그룹(111)이 개별적으로 전처리된다.

[0056] 여기서 전처리 과정은, 단계(S121)의 학습문서 그룹(111)의 각 학습문서 별로 특수문자가 제거되는 과정, 단계(S122)의 학습문서 그룹(111)의 각 학습문서 별로 불용어(stopword)가 제거되는 과정, 단계(S123)의 학습문서 그룹(111)의 각 학습문서 별로 형태소를 분석하여 명사 및 동사 이외의 품사들을 가려내는 과정, 단계(S124)의 형태소 분석 및 일부 품사를 가려내는 과정을 거친 학습문서로부터 명사 및 동사를 추출하여 해당 명사 및 동사

의 동시출현 빈도의 수를 추출하는 과정, 그리고 단계(S125)의 동시출현 빈도의 수를 기준으로 엔그램 데이터 세트를 구축하는 단계를 포함한다.

[0057] 그리고 이러한 일련의 전처리 과정 후, 단계(S130)에서 복수의 학습문서 그룹(111)으로부터 엔그램 데이터 세트(131)가 개별 생성되는 단계가 진행된다.

[0058] 또한, 단계(S130)은 학습문서 그룹(111) 별로 전처리 과정을 통해 바이그램 데이터 세트가 구축되는 단계에 해당된다.

[0059] 다음은 도 4를 참조하여 본 발명의 일 실시예에 따른 문서 자동 분류 방법에 대해 설명한다. 여기서 문서의 자동 분류 장치는 도 2의 실시예를 이용한 것이다.

[0060] 먼저, 단계(S210)에서 복수의 학습문서가 카테고리 별로 분류되어 복수의 학습문서 그룹이 형성된다.

[0061] 이어서, 단계(S220)에서 학습문서 그룹 별로 특수문자와 불용어(stopword) 제거 및 형태소 분석을 통해 명사 및 동사만을 추출하는 전처리가 이루어진다.

[0062] 이어서, 단계(S230)에서 단계(S220)의 전처리를 통해 복수의 학습문서 그룹으로부터 엔그램 데이터 세트가 개별 생성된다.

[0063] 이어서, 단계(S240)에서 엔그램 데이터 세트를 통해 식별되지 않는 신규문서가 출현한다.

[0064] 이어서, 단계(S250)에서 단계(S240)의 신규문서를 단계(S220)의 전처리와 동일한 과정으로 전처리하여 신규문서의 해당 바이그램을 형성한다.

[0065] 이어서, 단계(S260)에서 신규문서의 바이그램과 엔그램 데이터 세트의 바이그램을 비교하여 해당 신규문서를 학습문서 그룹들 중 어느 하나에 할당하여 분류한다.

[0066] 상술한 도 1 내지 도 4의 실시예를 통하여 알 수 있는 바와 같이, 본 발명에 따른 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법은, 엔그램 기반의 문서 자동 학습 및 분류 과정을 통해 웹상의 대용량 문서들을 자동으로 학습 및 분류할 수 있게 한다.

[0067] 또한, 웹문서를 미리 정의된 분야별로 학습단계를 거친 후 신규문서 출현 시, 이를 자동 분류하는 방법을 통하여 웹문서의 의미적 분석과 함께 사용자의 질의어와 밀접한 검색결과를 제공할 수 있게 한다.

[0068] 또한, 동시에 출현하는 단어들의 집합인 엔그램과 이의 출현빈도수를 바탕으로 문서를 대표할 수 있는 엔그램을 구축하여 차후에 식별되지 않은 신규문서가 발견되었을 때 기 구축된 엔그램 정보를 바탕으로 새로운 문서를 자동으로 분류할 수 있게 한다.

[0069] 또한, 기 정의된 분류항목에 걸맞은 학습문서를 대상으로 어절 단위 엔그램 통계적 분포치 집합을 구축한 후, 신규문서가 출현하였을 경우, 신규문서 내에 존재하는 엔그램 분포치와 기 학습된 엔그램 분포치와의 유사도 측정을 통하여 신규문서를 자동 분류한다. 이때 신규문서에서 생성된 엔그램 분포데이터는 기 학습 엔그램 분포치에 누적되어 점증적으로 확장되어 정보의 변화에 적응할 수 있는 데이터로 그 신뢰성을 향상시킬 수 있게 한다.

[0070] 이상에서 설명한 것은 본 발명에 따른 문서의 자동 학습 장치와 이를 이용한 문서 자동 학습 방법 및 문서의 자동 분류 장치와 이를 이용한 문서 자동 분류 방법을 실시하기 위한 하나의 실시예에 불과한 것으로서, 본 발명은 상기한 실시 예에 한정되지 않고, 이하의 특허청구범위에서 청구하는 바와 같이 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변경 실시가 가능한 범위까지 본 발명의 기술적 정신이 있다고 할 것이다.

부호의 설명

[0071]

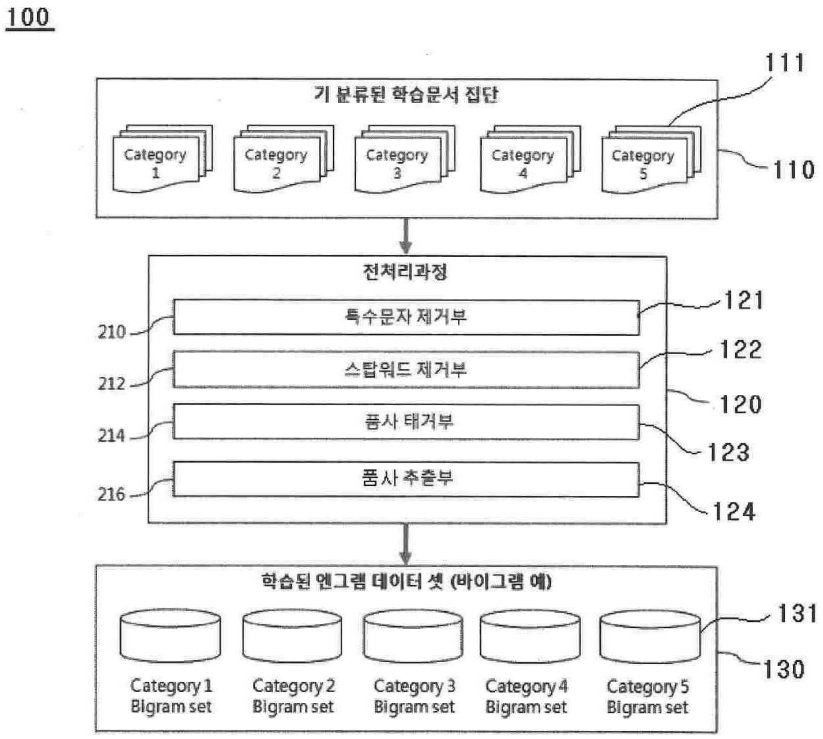
100 : 문서의 자동 학습 장치	110 : 학습문서 풀
111 : 학습문서 그룹	120 : 전처리부
121 : 특수문자 제거부	122 : 불용어 제거부
123 : 품사태거부	124 : 품사추출부
130 : 엔그램 데이터 세트 풀	131 : 엔그램 데이터 세트

200 : 문서의 자동 분류 장치　　210 : 문서 자동 학습부

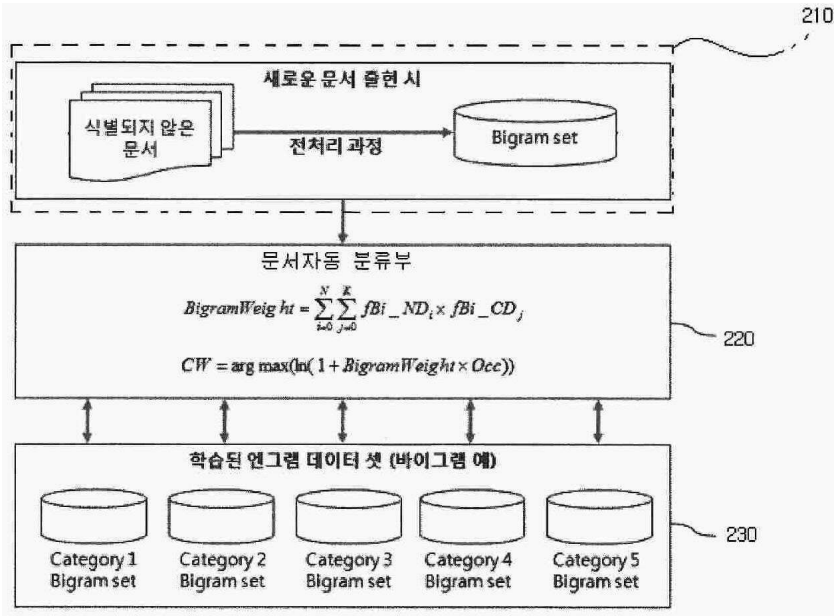
220 : 문서 자동 분류부

도면

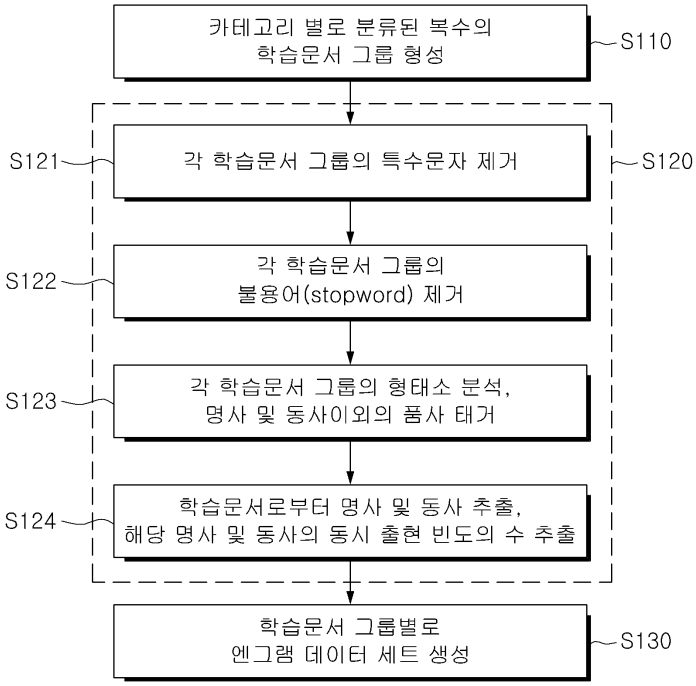
도면1



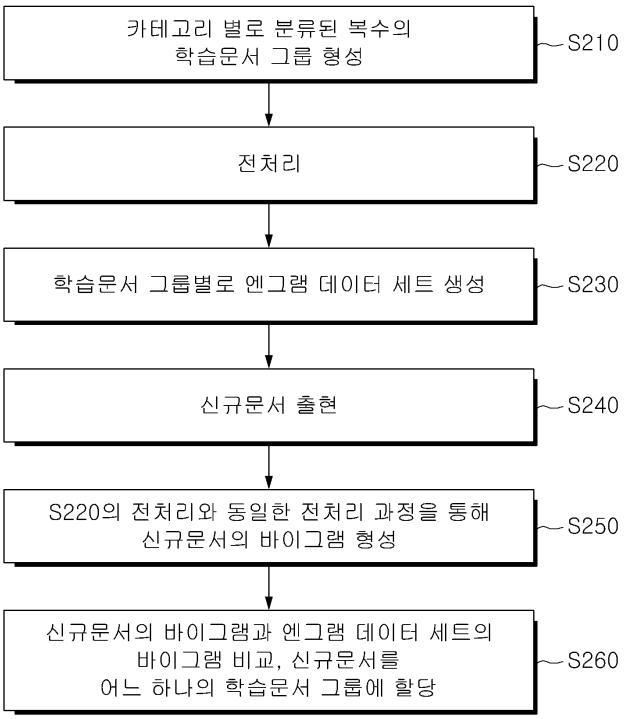
도면2



도면3


카테고리 별로 분류된 복수의
학습문서 그룹 형성
S110
S120
S121
각 학습문서 그룹의 특수문자 제거
S122
각 학습문서 그룹의
불용어(stopword) 제거
S123
각 학습문서 그룹의 형태소 분석,
명사 및 동사이외의 품사 태거
S124
학습문서로부터 명사 및 동사 추출,
해당 명사 및 동사의 동시 출현 빈도의 수 추출
학습문서 그룹별로
엔그램 데이터 세트 생성
S130

도면4


카테고리 별로 분류된 복수의
학습문서 그룹 형성
S210
전처리
S220
학습문서 그룹별로 엔그램 데이터 세트 생성
S230
신규문서 출현
S240
S220의 전처리와 동일한 전처리 과정을 통해
신규문서의 바이그램 형성
S250
신규문서의 바이그램과 엔그램 데이터 세트의
바이그램 비교, 신규문서를
어느 하나의 학습문서 그룹에 할당
S260