(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0020649
(43) 공개일자 2011년03월03일

(51) Int. Cl.
G06Q 50/00 (2006.01) *G06F 17/20* (2006.01)
(21) 출원번호 10-2009-0078372
(22) 출원일자 2009년08월24일
심사청구일자 2009년08월24일

(71) 출원인
부산대학교 산학협력단
부산 금정구 장전동 산30 부산대학교 내
(72) 발명자
조환규
부산광역시 동래구 온천2동 럭키아파트 19동 202호
윤태진
부산광역시 금정구 구서동 롯데캐슬 603동 1601호
(74) 대리인
문춘오, 오위환, 정기택

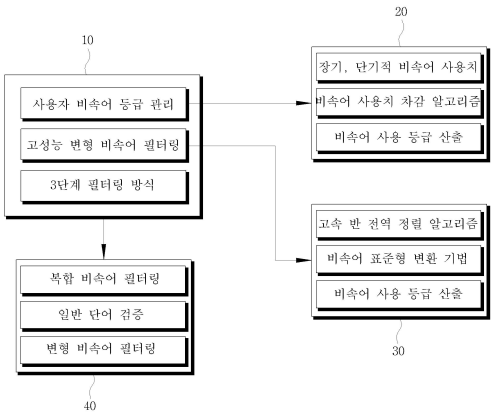
전체 청구항 수 : 총 9 항

(54) 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법

(57) 요 약

본 발명은 인터넷을 이용한 커뮤니케이션 과정에서 사용될 수 있는 비속어를 필터링 하기 위하여 해당 시스템에 부차적으로 사용될 수 있는 필터링 모듈에 관한 것으로, 변형 비속어 필터링을 위하여 사용자에게 유동적인 금칙어 필터링 레벨 적용을 위한 단기적, 장기적인 비속어 사용성향을 계산하여 수치화하여 관리할 수 있는 사용자 비속어 필터링 등급 관리 시스템과 필터링 레벨에 따른 유동적인 비속어 필터링이 가능하도록 반 전역 정렬 값을 이용한 유사도 측정을 통해서 수치화된 비속어 필터링 시스템으로 구성되어 일반적인 비속어 필터링 시스템이 가지는 문제를 해결함과 동시에 사용자에게 필터링 등급을 통한 합리적인 제재가 가해짐으로써 비속어 사용에 대한 경각심을 불러일으키는 것이 가능한 시스템이다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업
과제고유번호 2009-0070594
부처명 한국학술진흥재단
연구관리전문기관
연구사업명 2009 일반연구자지원사업(기본연구)
연구과제명 진화유전학 분석기법을 이용한 지능형 인터넷 비속어/악플 필터링 시스템
기여율
주관기관 부산대학교 산학협력단
연구기간 2009.5.1~2012.04.30

특허청구의 범위

청구항 1

문자 커뮤니케이션 과정에서 사용자의 비속어 필터링을 제어 관리하는 사용자 관리부;

사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 등급을 관리하는 비속어 등급 관리부;

사용자의 비속어 사용 등급에 따라 유동적으로 적용될 수 있는 반 전역 정렬 값을 단어 간의 유사도로 이용하는 비속어 필터링을 하는 변형 비속어 필터링부;

복합 비속어 필터링, 일반 단어 검증, 변형 비속어 필터링을 하는 필터링 관리부;를 포함하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 시스템.

청구항 2

제 1 항에 있어서, 상기 비속어 등급 관리부는,

사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 성향 등급을 설정하고, 비속어를 사용하지 않은 기간에 따라 장기적 수치와 단기적 수치를 차등적인 비율로 차감하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 시스템.

청구항 3

문자 커뮤니케이션 과정에서 사용자의 비속어 필터링을 위하여,

사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 등급을 설정하는 단계;

문장이 입력되면 복합 비속어 필터링, 일반 단어 검증, 변형 비속어 필터링을 하는 단계;

사용자의 비속어 사용 등급에 따라 필터링 기준을 다르게 적용하여 해당 비속어를 차단하거나 정상적인 문장을 출력하는 단계;를 포함하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 4

제 3 항에 있어서, 상기 사용자의 비속어 사용 등급을 설정하는 단계에서,

비속어를 사용하지 않은 기간에 따라 장기적 수치와 단기적 수치를 차등적인 비율로 차감하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 5

제 3 항에 있어서, 상기 변형 비속어 필터링은,

반 전역 정렬을 이용하여 비속어의 변형형태와 일반 비속어 간의 유사도를 계산하여 임계값 이상의 단어를 필터링하는 방식인 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 6

제 3 항에 있어서, 문장이 입력되면,

데이터베이스에 저장된 비속어를 필터링하는 단계와,

정상 단어가 필터링 되는 것을 방지하기 위하여 데이터베이스 내의 단어사전에 등록된 정상단어를 검증하는 단계와,

변형 비속어를 필터링 하기 위하여 반 전역 정렬의 유사도 값을 이용하는 필터링 단계를 순차적으로 진행하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 7

제 6 항에 있어서, 상기 정상단어를 검증하는 단계에서,

정상단어의 필터링을 방지하기 위하여 문장 내의 정상단어를 미리 필터링 대상에서 제외하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 8

제 6 항에 있어서, 데이터베이스에 저장된 비속어를 필터링하는 단계에서 정상단어를 포함하는 비속어가 필터링되는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

청구항 9

제 3 항에 있어서, 상기 변형 비속어 필터링은,

한글의 유사한 발음, 형태의 글자를 여러 집합으로 분류하여 각 집합의 대표 문자로 변환하여 단어를 표준형태(canonical form)로 변환하는 단계와,

상기 해당 표준형태를 색인으로 이용하여 데이터베이스내의 유사한 비속어를 검색하는 단계와,

검색하여 추출된 후보군과 입력된 단어 간의 반 전역 정렬 값을 측정하여 이 값이 일정 임계값을 넘을 경우 비속어로 판정하는 단계를 포함하는 것을 특징으로 하는 변형 비속어 필터링을 위한 비속어 필터링 방법.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 비속어 필터링 시스템에 관한 것으로, 구체적으로 반 전역 정렬(semi-global alignment) 알고리즘을 이용하여 변형 비속어 필터링을 가능하도록 한 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근에 인터넷이나 통신 등 가상의 사이버공간을 통한 커뮤니케이션이 활발하게 이루어져서 인터넷을 통한 언어생활은 현실에도 큰 영향을 미치고 있다. 인터넷을 통한 커뮤니케이션은 인터넷 BBS 혹은 채팅 등 대부분 문자를 통해 이루어지고 직접 상대의 얼굴과 목소리를 대면하지 않는 문자 대화의 특성상 익명성과 책임의식부족으로 인해 현실에서 보다 비속어 사용이 활발하게 이루어지게 되어 올바른 언어생활에 악영향을 주고 사용자들에게 불쾌감을 주게 된다.

[0003] 종래 기술의 비속어 필터링 방법은 입력된 문자를 보유하고 있는 데이터베이스 내에서 검색하여 해당 단어가 존재하면 필터링을 수행하는 방식으로 다음과 같은 문제가 있다.

[0004] 사용자가 단어의 변형을 통해서 얼마든지 비속어 필터링을 얼마든지 피해나갈 수 있다. 유사한 발음이나 형태의 문자로 일부분을 변형하는 것만으로도 간단하게 비속어 필터링을 피해 상대에게 비속어의 의미를 전달할 수 있다.

[0005] 대한민국 공개특허 10-2009-0001435호의 기술에서는 여러 라인에 걸쳐 비속어를 나눠서 쓰게 되는 경우를 상정하여 변형 비속어 필터링 방식을 구현하였으나 그 범위가 매우 한정되고 이 방법 역시 데이터 베이스에 등록되어 있지 않은 변형된 형태의 비속어를 이용할 경우 필터링하는 것이 불가능하다.

[0006] 대한민국 공개특허 특2001-0067853호의 기술에서는 음절 결합 정보를 이용하여 문자를 자소별로 나누거나 단어를 빈칸으로 분리하는 등의 변형 방법에 적용될 수 있으나 이 역시 매우 한정된 비속어 변형에 대응할 수 있고, 이와 같은 비속어 필터링은 정상단어를 필터링하는 부작용을 가지고 있다.

[0007] 예를 들어 비속어로 자주 사용되는 “시발”을 금칙어로 등록할 경우 “시발점”이라는 일반단어를 사용할 수 없게 된다. 이 문제는 scunthorpe problem이라 명명 되었으며 인터넷 검색 사이트 등에서 음란 비속어 필터링을 강화하는데 큰 어려움을 주고 있는 문제이다.

[0008] 현재 온라인 커뮤니케이션 시스템의 비속어 필터링 기능들은 상기의 문제점으로 인해서 사용자의 불만이 끊이지 않고 있으며 비속어 필터링이 역할을 수행하지 못하고 불편만 가중한다는 이유로 제거를 요구하는 경우도 자주 목격할 수 있다.

[0009] 그러나 올바른 국어를 지키고 선의 사용자의 피해를 막기 위해서도 고성능의 비속어 필터링 시스템의 개발은 반드시 이루어져야할 과제이다.

[0010] 이러한 요구를 만족시킬 수 있는 새로운 방식의 비속어 필터링 시스템이 요구되고 있다.

발명의 내용

해결 하고자하는 과제

[0011] 본 발명은 비속어 필터링 시스템에 반 전역 정렬(semi-global alignment) 알고리즘을 적용하여 변형 비속어 필터링을 가능하도록 한 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법을 제공하는데 그 목적이 있다.

[0012] 본 발명은 사용자의 비속어 사용성향을 분석하고 그 결과에 따라 유동적으로 적용할 수 있는 비속어 필터링 시스템을 개발하여 정상적인 커뮤니케이션에 영향을 최소한으로 줄이면서 높은 성능의 비속어 필터링 성능을 갖는 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법을 제공하는데 그 목적이 있다.

[0013] 본 발명은 DNA정보를 분석하는데 이용되는 반 전역 정렬(semi-global alignment) 알고리즘을 이용하여 변형 비속어 필터링에 대응할 수 있는 비속어 필터링 시스템을 구축함은 물론 문제에 대한 인식은 있었으나 해결되지 못했던 일반단어 필터링 문제(Scunthorpe Problem)에 대해 정상단어사전을 통해 해결할 수 있도록 한 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법을 제공하는데 그 목적이 있다.

과제 해결수단

[0014] 이와 같은 목적을 달성하기 위한 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템은 문자 커뮤니케이션 과정에서 사용자의 비속어 필터링을 제어 관리하는 사용자 관리부;사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 등급을 관리하는 비속어 등급 관리부;사용자의 비속어 사용 등급에 따라 유동적으로 적용될 수 있는 반 전역 정렬 값을 단어 간의 유사도로 이용하는 비속어 필터링을 하는 변형 비속어 필터링부;복합 비속어 필터링, 일반 단어 검증, 변형 비속어 필터링을 하는 필터링 관리부;를 포함하는 것을 특징으로 한다.

[0015] 여기서, 상기 비속어 등급 관리부는, 사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 성향 등급을 설정하고, 비속어를 사용하지 않은 기간에 따라 장기적 수치와 단기적 수치를 차등적인 비율로 차감하는 것을 특징으로 한다.

[0016] 다른 목적을 달성하기 위한 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 방법은 문자 커뮤니케이션 과정에서 사용자의 비속어 필터링을 위하여,사용자의 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 구하여 사용자의 비속어 사용 등급을 설정하는 단계;문장이 입력되면 복합 비속어 필터링, 일반 단어 검증, 변형 비속어 필터링을 하는 단계;사용자의 비속어 사용 등급에 따라 필터링 기준을 다르게 적용하여 해당 비속어를 차단하거나 정상적인 문장을 출력하는 단계;를 포함하는 것을 특징으로 한다.

[0017] 그리고 상기 사용자의 비속어 사용 등급을 설정하는 단계에서,비속어를 사용하지 않은 기간에 따라 장기적 수치와 단기적 수치를 차등적인 비율로 차감하는 것을 특징으로 한다.

[0018] 그리고 상기 변형 비속어 필터링은, 반 전역 정렬을 이용하여 비속어의 변형형태와 일반 비속어 간의 유사도를 계산하여 임계값 이상의 단어를 필터링하는 방식인 것을 특징으로 한다.

[0019] 그리고 문장이 입력되면, 데이터베이스에 저장된 비속어를 필터링하는 단계와,정상 단어가 필터링 되는 것을 방지하기 위하여 데이터베이스 내의 단어사전에 등록된 정상단어를 검증하는 단계와,변형 비속어를 필터링 하기 위하여 반 전역 정렬의 유사도 값을 이용하는 필터링 단계를 순차적으로 진행하는 것을 특징으로 한다.

[0020] 그리고 상기 정상단어를 검증하는 단계에서,정상단어의 필터링을 방지하기 위하여 문장 내의 정상단어를 미리 필터링 대상에서 제외하는 것을 특징으로 한다.

[0021] 그리고 데이터베이스에 저장된 비속어를 필터링하는 단계에서 정상단어를 포함하는 비속어가 필터링되는 것을

특징으로 한다.

[0022] 그리고 상기 변형 비속어 필터링은, 한글의 유사한 발음, 형태의 글자를 여러 집합으로 분류하여 각 집합의 대표 문자로 변환하여 단어를 표준형태(canonical form)로 변환하는 단계와,상기 해당 표준형태를 색인으로 이용하여 데이터베이스내의 유사한 비속어를 검색하는 단계와,검색하여 추출된 후보군과 입력된 단어 간의 반 전역 정렬 값을 측정하여 이 값이 일정 임계값을 넘을 경우 비속어로 판정하는 단계를 포함하는 것을 특징으로 한다.

효 과

[0023] 이와 같은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법은 다음과 같은 효과를 갖는다.

[0024] 첫째, 비속어 필터링 시스템으로 인한 불편함이 해소되어 원할한 커뮤니케이션을 수행할 수 있다.

[0025] 둘째, 사용자의 비속어 사용성향을 분석하고 그 결과에 따라 유동적으로 적용할 수 있는 비속어 필터링 시스템을 개발하여 정상적인 커뮤니케이션에 영향을 최소한으로 줄이면서 높은 비속어 필터링 성능을 구현한다.

[0026] 셋째, 효과적인 변형 비속어 필터링을 가능하도록 하여 사용자가 단어의 변형을 통해서 비속어 필터링을 얼마든지 피해나가는 것을 막는다.

[0027] 넷째, 서비스 제공자 입장에서 비속어의 수집 및 사용자의 제재를 위한 인력 및 비용을 절감할 수 있다.

[0028] 다섯째, 높은 비속어 필터링 성능을 구현하여 사용자의 건전한 언어활동으로 인한 이미지 제고에도 큰 역할을 수행할 수 있다.

발명의 실시를 위한 구체적인 내용

[0029] 이하, 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법의 바람직한 실시예에 관하여 상세히 설명하면 다음과 같다.

[0030] 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법의 특징 및 이점들은 이하에서의 각 실시예에 대한 상세한 설명을 통해 명백해질 것이다.

[0031] 도 1은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템의 구성도이다.

[0032] 본 발명은 인터넷을 이용한 커뮤니케이션 과정에서 사용될 수 있는 비속어를 필터링 하기 위하여 해당 시스템에 부차적으로 사용될 수 있는 필터링 모듈에 관한 것이다.

[0033] 변형 비속어 필터링을 위하여 사용자에게 유동적인 금칙어 필터링 레벨 적용을 위한 단기적, 장기적인 비속어 사용성향을 계산하여 수치화하여 관리할 수 있는 사용자 비속어 필터링 등급 관리 부분과, 필터링 레벨에 따른 유동적인 비속어 필터링이 가능하도록 반 전역 정렬 값을 이용한 유사도 측정을 통해서 수치화된 비속어 필터링 부분으로 구성된다.

[0034] 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템은 도 1에서와 같이, 인터넷을 이용한 문자 커뮤니케이션 과정에서 사용자에 의해 사용될 수 있는 비속어 필터링을 제어 관리하는 사용자 관리부(10)와, 사용자의 비속어 사용성향에 따라 차등적으로 필터링을 적용할 수 있도록 단기적, 장기적 비속어 사용빈도에 대하여 수치화된 값을 보여주고 관리하는 비속어 등급 관리부(20)와, 사용자의 비속어 사용 등급에 따라 유동적으로 적용될 수 있는 반 전역 정렬 값을 단어 간의 유사도로 이용하는 비속어 필터링을 하는 변형 비속어 필터링부(30)와, 복합 비속어 필터링, 일반 단어 검증, 변형 비속어 필터링의 3단계의 비속어 필터링을 하는 필터링 관리부(40)를 포함한다.

[0035] 여기서, 비속어 등급 관리부(20)는 비속어 사용치 차감 알고리즘을 이용하여 장기,단기적인 비속어 사용치를 구하고 사용자의 비속어 사용 등급을 산출한다.

[0036] 이와 같은 구성을 갖는 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템에서의 비속어 필터링 과정을 구체적으로 설명하면 다음과 같다.

[0037] 도 2a와 도 2b는 본 발명에 따른 사용자 비속어 등급 결정 과정을 나타낸 플로우차트이고, 도 3은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 과정을 나타낸 플로우차트이다.

[0038] 그리고 도 4는 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 과정에서의 표준형 변환 과정을 나타낸 구성도이다.

[0039] 도 2a는 사용자 비속어 등급 결정 과정을 나타낸 것으로, 비속어 입력이 검출되면(S201), 입력 위치가 비속어 필터링 시스템에 의한 것인지 사용자의 신고에 의한 것인지를 판단한다.(S202)

[0040] 입력 위치가 비속어 필터링 시스템이면, 단기 비속어 수치를 증가시키고(S203), 장기 비속어 수치를 증가시키고(S204) 현재의 비속어 등급을 변경한다.(S205)

[0041] 만약, 비속어 입력 위치가 사용자의 신고에 의한 것이면, 단기 비속어 수치 증가없이 장기 비속어 수치를 증가시키고(S204) 현재의 비속어 등급을 변경한다.(S205)

[0042] 그리고 도 2b는 차감 주기에 따라 단기,장기 비속어 수치를 차감하여 비속어 등급을 변경하는 과정을 나타낸 것이다.

[0043] 차감 주기에 따라(S211) 단기 비속어 수치를 차감하고(S212), 차감 주기에 따라(S213) 장기 비속어 수치를 차감하여(S214) 현재의 비속어 등급을 변경한다.(S215)

[0044] 이와 같이 본 발명에서는 사용자의 비속어 사용 등급을 위한 사용자의 비속어 사용빈도는 장기적인 수치와 단기적인 수치로 병렬적으로 관리된다.

[0045] 장기적인 수치의 경우 비속어를 자주 사용해왔던 사람은 앞으로 비속어를 자주 사용할 것이라는 선험적인 지식에 의한 척도이다. 단기적인 수치의 경우 비속어는 사용자간의 분쟁과 같은 상황에서 연속적으로 사용될 수 있기에 그러한 분쟁을 조기에 막기 위하여 일시적으로 비속어 필터링 강도를 높이는데 사용된다.

[0046] 그리고 비속어 빈도를 측정하기 위한 척도로는 2가지가 있다. 타 사용자의 신고의 경우 관리자에 의한 사후 처리적인 의미가 강하기 때문에 단기 수치에는 영향을 미치지 않고 장기 수치에만 값을 더하게 된다.

[0047] 그리고 비속어 필터링 시스템에 의해 필터링 되는 단어를 입력한 경우 단기적인 비속어 사용 수치에는 크게 더해지고 장기적인 비속어 사용 수치에는 적게 더해지게 된다.

[0048] 이 두 가지 수치는 시간이 지남에 따라 일정 비율로 차감되게 되는데 단기적인 수치는 장기적인 수치에 비해 약 1/100의 주기로 빠르게 감소된다. 감소되는 주기 전체는 해당 시스템의 목적과 성향에 따른다.

[0049] 그리고 도 3은 변형 비속어 필터링을 위한 과정을 나타낸 것으로, 사용자의 문장이 입력되면(S301), 입력된 문장을 단어 단위로 분리한다.(S302)

[0050] 그리고 복합 비속어인지를 판단하여(S303), 복합 비속어로 판단되면 해당 비속어를 차단한다.(S311) 복합 비속어가 아닌것으로 판단되면 문장의 끝인가를 판단한다.(S304)

[0051] 문장의 끝이 아닌 것으로 판단되면 문장을 단어 단위로 분리하는 단계를 다시 수행하고, 문장의 끝인 것으로 판단되면 일반 단어 검증을 한다.(S305)

[0052] 이와 같이 복합 비속어의 판단 및 일반 단어 검증이 문장 끝까지 이루어지면(S306), 검색 후보군 추출을 위한 단어의 표준형 변환을 한다.(S307)

[0053] 그리고 유사 표제어 검색 후보군을 추출하고(S308), 유사 문자군 매칭 매트릭스(matching matrix)를 이용한 반전역 정렬 과정을 수행한다.(S309)

[0054] 그리고 반 전역 정렬값이 사용자 등급에 따른 문턱값보다 높은지를 판단하여(S310), 비속어를 차단하거나(S311) 정상적인 문장을 출력한다.(S312)

[0055] 이와 같은 변형 비속어 필터링을 위한 과정을 더 구체적으로 설명하면 다음과 같다.

[0056] 해당 사용자의 비속어 등급은 단기적인 수치와 장기적인 수치의 합으로 결정된다. 비속어 사용등급이 높아질수록 낮은 수치의 유사도 임계값을 적용하여 엄격한 비속어 필터링이 적용되는데 이 경우 일반단어의 필터링 비율이 높아지므로 해당 사용자는 정상적인 커뮤니케이션에 어려움을 겪게 되어 비속어 사용에 대한 제재의 용도로도 활용될 수 있다.

[0057] 반대로 비속어 사용 수치가 낮은 사용자에게는 정상적인 대화에 영향을 미치지 않도록 최소한의 비속어만 필터링 될 수 있도록 높은 유사도 임계값을 적용하여 정상단어가 필터링 될 가능성을 최소화하여 정상단어 필터링

문제를 해결할 수 있다.

[0058] 표 1은 발음을 이용한 변형 방식에 대한 표준형 변환 규칙을 나타낸 것이고, 표 2는 외래어 및 특수문자에 대한 표준형 변환 규칙을 나타낸 것이다.

표 1

초 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㄱ, ㄲ, ㅋ	ㄱ	ㅅ, ㅆ	ㅅ
ㄷ, ㄸ, ㅌ	ㄷ	ㅈ, ㅉ, ㅊ	ㅈ
ㅂ, ㅃ, ㅍ	ㅂ		
중 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㅏ, ㅑ	ㅏ	ㅗ, ㅛ	ㅗ
ㅐ, ㅒ, ㅔ, ㅖ, ㅙ, ㅚ, ㅞ, ㅞ, ㅞ	ㅐ	ㅜ, ㅠ	ㅜ
ㅓ, ㅕ	ㅓ	ㅟ, ㅡ, ㅣ	ㅣ
종 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㄱ, ㄲ, ㅋ	ㄱ	ㅂ, ㅍ	ㅂ
ㄷ, ㅌ, ㅅ, ㅈ, ㅉ, ㅊ	ㄷ		

[0059]

표 2

자 음			
원래 문자	표준화 문자	원래 문자	표준화 문자
g, k, c	ㄱ	b, v, p, f	ㅂ
n, L	ㄴ	s, A	ㅅ
d, t, E	ㄷ	o, w	ㅇ
l, r	ㄹ	j, g, z	ㅈ
m, ㅁ	ㅁ	h	ㅎ
모 음			
원래 문자	표준화 문자	원래 문자	표준화 문자
a	ㅏ	u	ㅓ
@, H	ㅐ	l, i, y, !, ㅣ	ㅣ

[0060]

[0061] 이와 같은 반 전역 정렬 값을 이용한 단어 간의 유사도 측정을 통한 필터링의 경우 입력된 단어와 비교해볼 후보군을 선정하는 것이 중요하다. 반 전역 정렬 값을 계산하기 위해서는 알고리즘적으로 많은 연산량을 필요로 하기 때문에 데이터베이스 내에 있는 모든 단어와 비교할 경우 성능이 크게 저하되기 때문이다.

[0062] 도 4는 비속어 필터링 과정에서의 표준형 변환 과정을 나타낸 것이다.

[0063] 그리고 표 3은 변형 비속어의 표준형 변환 예를 나타낸 것이다.

표 3

변형비속어		비속어의 표준형
1	ㄱHㅅHㄲㅣ	ㄱㅐㅅㅐㄱㅣ
	깨쉐끼	ㄱㅐㅅㅐㄱㅣ
2	미친년	ㅁㅣㅈㅣㄴㅕㄴ
	미췬년	ㅁㅣㅈㅣㄴㅕㄴ
	ㅁ]친년	ㅁㅣㅈㅣㄴㅕㄴ
3	씨팔	ㅅㅣㅂㅏㄹ
	쉬팔	ㅅㅣㅂㅏㄹ
	쉬빨	ㅅㅣㅂㅏㄹ
4	쌰년	ㅅㅏㅇㄴㅕㄴ
	쐉ㄴㅕㄴ	ㅅㅏㅇㄴㅕㄴ
5	자지	ㅈㅏㅈㅣ
	자ji	ㅈㅏㅈㅣ

[0064]

[0065] 이 후보군을 선정하기 위해서 본 발명의 시스템은 표준형 변환 방식을 이용하여 적합한 비교대상을 데이터베이스에서 추출하는 방법을 사용하게 된다.

[0066] 표준형 변환 방식이란 변형비속어에 대해서 유사 발음 및 형태의 단어를 데이터베이스에서 추출하기 위한 방법으로 한글의 각 자소 및 그 자소와 발음과 형태가 유사한 특수문자, 영어의 글자들을 몇 개의 그룹으로 분류하여 해당 대표 글자로 모두 변환하는 방법이다.

[0067] 이 방식을 이용하면 일반적으로 자주 쓰이는 변형 방법을 검색방법을 단순화하여 변형비속어가 입력되더라도 문제없이 유사한 발음의 비속어 후보군을 데이터베이스에서 추출해 낼 수 있다.

표 4

기본 자소	매칭 자소	점수
ㄱ	ㄲ, ㅋ	0.8
	g, k, c	0.6
	>	0.4
ㄴ	n	0.6
	L	0.5
ㄷ	ㄸ, ㅌ	0.8
	d, t	0.6
ㄹ	l, r	0.6
ㅁ	m	0.6
	ㅁ	0.5

[0068]

표 5

	비속어	Alignment 결과
Source	ㅅㅣ입팔	ㅅ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Target	씨입팔	ㅆ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Score	7.8	0.8 1.0 1.0 1.0 1.0 1.0 1.0 1.0
Source	시이방	ㅅ ㅣ ㅇ ㅣ ㅂ ㅏ ㅇ
Target	쒸이방	ㅆ ㅟ ㅇ ㅣ ㅂ ㅏ ㅇ
Score	6.4	0.8 0.6 1.0 1.0 1.0 1.0 1.0
Source	ㅅㅣ입팔	ㅅ _ _ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Target	씨이이이팔	ㅅ ㅣ ㅇ ㅣ ㅇ ㅣ _ ㅂ ㅏ ㄹ
Score	3.6	1.0-1.0-1.0 1.0 1.0 1.0-1.0 0.6 1.0 1.0
Source	띠이발	_ ㄸ ㅣ ㅇ ㅣ ㅂ ㅏ ㄹ
Target	쉬이벌	ㅅ - ㅟ ㅇ ㅣ ㅂ ㅓ ㄹ
Score	3.0	-1.0-1.0 0.4 1.0 1.0 1.0 0.6 1.0
Source	띠이벌	_ _ _ ㄸ ㅣ ㅇ ㅣ ㅂ ㅓ ㄹ
Target	씨이이팔	ㅅ ㅣ ㅇ _ ㅣ ㅇ ㅣ ㅍ ㅏ ㄹ
Score	1.4	-1.0-1.0-1.0-1.0 1.0 1.0 1.0 0.8 0.6 1.0

[0069]

[0070] 표 4는 변형 자소의 매칭값의 일부를 나타낸 것이고, 표 5는 비속어의 Semi-Global Alignment 예를 나타낸 것이다.

[0071] 이 추출된 후보군과 입력된 단어 간의 반 전역 정렬 값을 측정하여 이 값이 일정 임계값을 넘을 경우 비속어로 판정하는 것이다.

[0072] 반 전역 정렬(semi-global alignment)은 유전공학에서 DNA분석을 위해 주로 사용되는 기법으로 이 임계값은 적용대상의 비속어 사용 등급에 따라 결정되며 임계값이 낮게 설정될 수록 정상단어를 필터링할 가능성이 높아지나 비속어에 대한 엄격한 필터링이 수행되는 유동적인 필터링이 가능해진다.

[0073] 변형 비속어에 대한 필터링 성능을 높이기 위해서 반 전역 정렬 값 측정시 유사 문자군 간의 matching값을 설정해둔 matching matrix가 사용되어 변형 비속어에 대한 대응력을 높이게 된다.

[0074] 유사도 측정을 통한 비속어 필터링 방식의 경우 기존의 필터링 시스템이 가지게 되는 정상단어 필터링 문제가 더 부각되게 된다. 비속어와 일치하지 않더라도 유사한 형태가 발견되면 비속어 필터링이 수행되기 때문이다.

[0075] 이러한 정상단어 필터링 문제를 해결하기 위하여 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템은 3단계 필터링 방식을 사용한다.

[0076] 먼저 정상단어의 사전을 미리 만들어서 데이터 베이스에 저장해둔 뒤에 미리 정상단어를 검증하여 비속어 필터에 정상단어가 필터링 되는 것을 최소화하는 방법이다.

[0077] 그러나 이 경우 반대로 비속어가 정상단어에 의해 필터링 될 수 있으므로 정상단어의 검증 전에 정상단어 검증으로 문제가 될 수 있는 비속어를 일반적인 비속어 필터링 방법으로 필터링하는 단계를 거치게 된다.

[0078] 그리고 마지막 3번째 단계로 변형 비속어를 필터링하기 위한 상기의 반 전역 정렬을 사용하는 비속어 필터링 알고리즘을 적용하는 것이다.

[0079] 이러한 복합적인 필터링 과정을 통해서 비속어 필터링 시스템의 변형 비속어 필터링 문제와 정상단어 필터링 문제를 같이 해결할 수 있다.

[0080] 이상에서와 같은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법은 변형 비속어 필터링을 위하여 사용자에게 유동적인 금칙어 필터링 레벨 적용을 위한 단기적, 장기적인 비속어 사용성향을 계산하여 수치화하여 관리할 수 있는 사용자 비속어 필터링 등급 관리 부분과, 필터링 레벨에 따른 유동적인 비속어 필터링이 가능하도록 반 전역 정렬 값을 이용한 유사도 측정을 통해서 수치화된 비속어 필터링 부분으로 구성되어 높은 비속어 필터링 성능을 구현한다.

[0081] 이상 설명한 내용을 통해 당업자라면 본 발명의 기술 사상을 일탈하지 아니하는 범위에서 다양한 변경 및 수정이 가능함을 알 수 있을 것이다.

[0082] 따라서, 본 발명의 기술적 범위는 실시예에 기재된 내용으로 한정되는 것이 아니라 특허 청구의 범위에 의하여 정해져야 한다.

도면의 간단한 설명

[0083] 도 1은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 시스템의 구성도

[0084] 도 2a와 도 2b는 본 발명에 따른 사용자 비속어 등급 결정 과정을 나타낸 플로우차트

[0085] 도 3은 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 과정을 나타낸 플로우차트

[0086] 도 4는 본 발명에 따른 변형 비속어 필터링을 위한 비속어 필터링 과정에서의 표준형 변환 과정을 나타낸 구성도

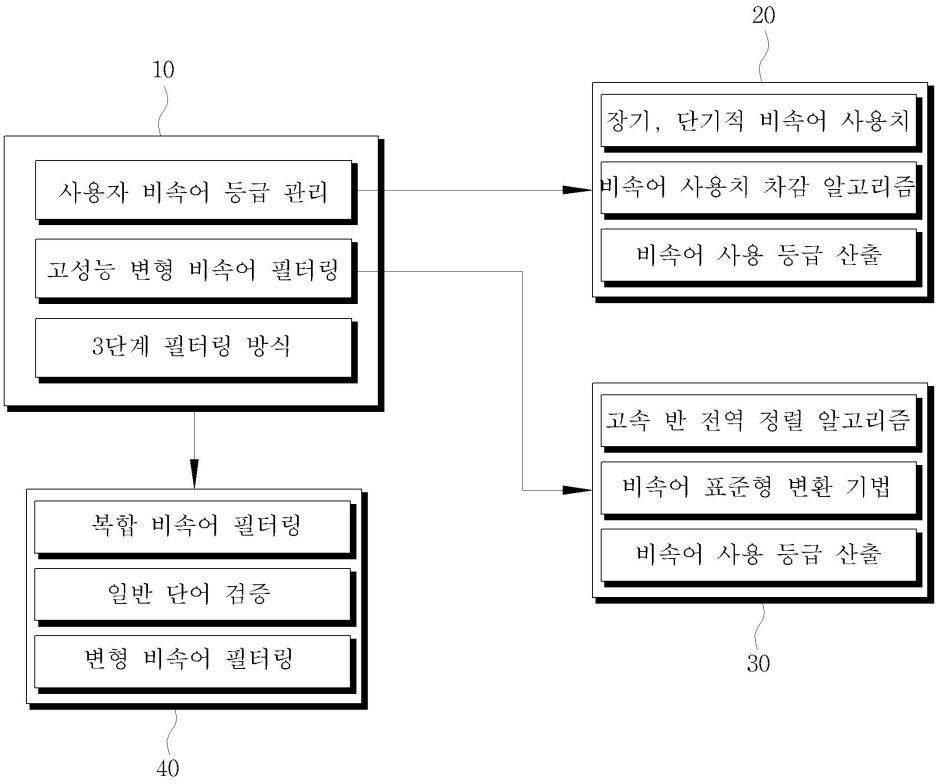
[0087] 도면의 주요 부분에 대한 부호의 설명

[0088] 10. 사용자 관리부 20. 비속어 등급 관리부

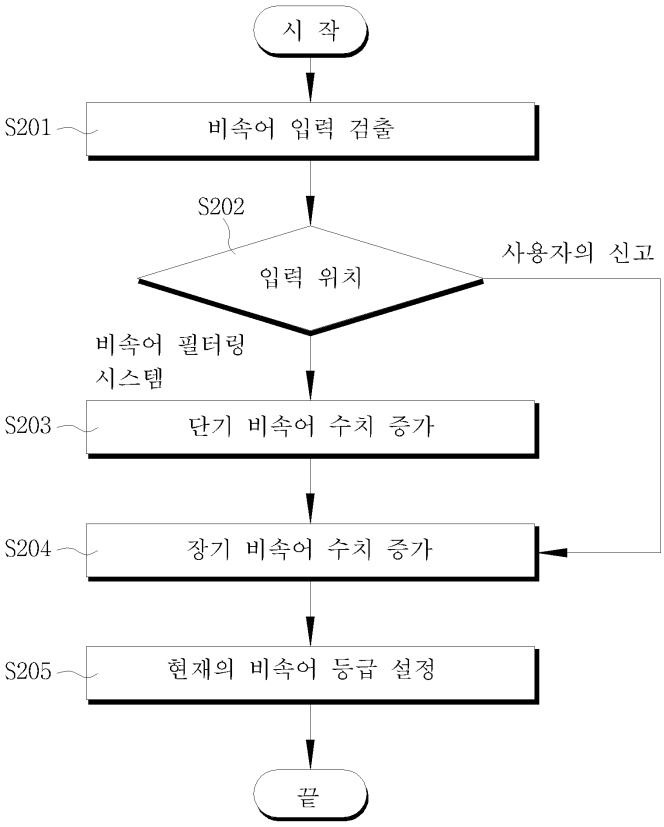
[0089] 30. 변형 비속어 필터링부 40. 필터링 관리부

도면

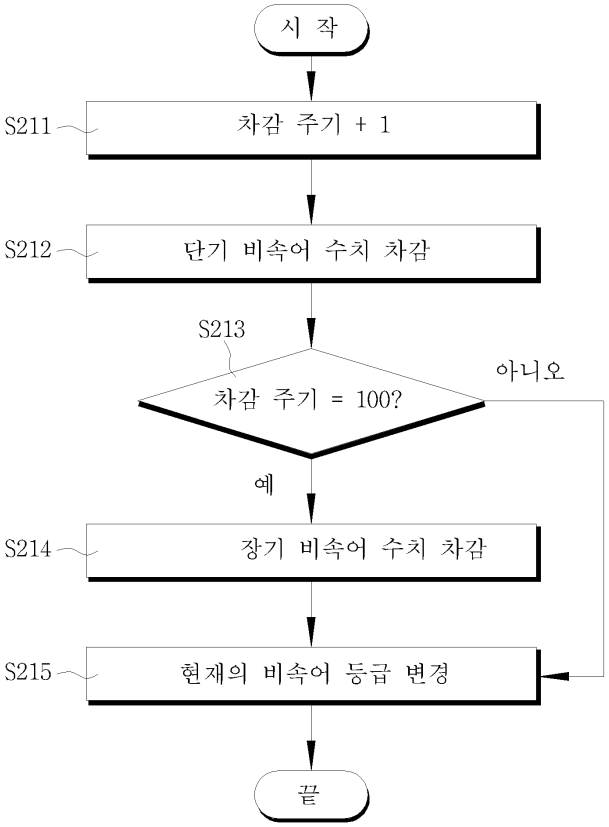
도면1



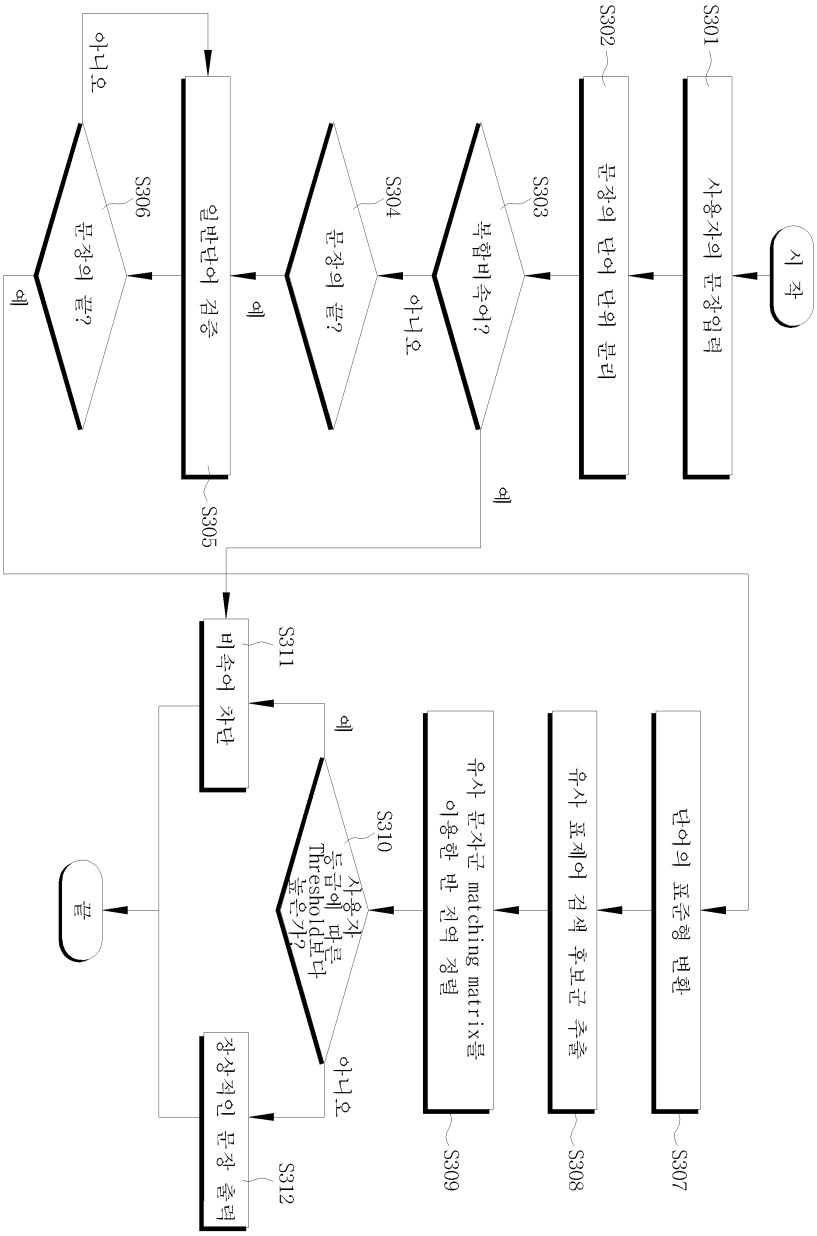
도면2a


시 작
S201
비속어 입력 검출
S202
입력 위치
사용자의 신고
비속어 필터링
시스템
S203
단기 비속어 수치 증가
S204
장기 비속어 수치 증가
S205
현재의 비속어 등급 설정
끝

도면2b


시 작
S211
차감 주기 + 1
S212
단기 비속어 수치 차감
S213
차감 주기 = 100?
아니오
예
S214
장기 비속어 수치 차감
S215
현재의 비속어 등급 변경
끝

도면3


시 작
S301
사용자의 문장입력
S302
문장의 단어 단위 분리
S303
복합비속어?
예
아니오
S304
문장의 끝?
예
아니오
S305
일반단어 검증
S306
문장의 끝?
예
아니오
S307
단어의 표준형 변환
S308
유사 표제어 검색 후보군 추출
S309
유사 문자군 matching matrix를 이용한 반 전역 정렬
S310
사용자 등급에 따른 Threshold보다 높은가?
예
아니오
S311
비속어 차단
S312
정상적인 문장 출력
끝

도면4