



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년06월07일
(11) 등록번호 10-1627592
(24) 등록일자 2016년05월31일

- | | |
|---|--|
| <p>(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) G06F 11/30 (2006.01)
G06F 21/55 (2013.01)</p> <p>(21) 출원번호 10-2011-7007587</p> <p>(22) 출원일자(국제) 2009년10월01일
심사청구일자 2014년09월26일</p> <p>(85) 번역문제출일자 2011년04월01일</p> <p>(65) 공개번호 10-2011-0081177</p> <p>(43) 공개일자 2011년07월13일</p> <p>(86) 국제출원번호 PCT/US2009/059240</p> <p>(87) 국제공개번호 WO 2010/042386
국제공개일자 2010년04월15일</p> <p>(30) 우선권주장
12/245,507 2008년10월03일 미국(US)</p> <p>(56) 선행기술조사문헌
US0652303 B1
US20080162225 A1
US20070028297 A1
US20080005106 A1</p> | <p>(73) 특허권자
아브 이니티오 테크놀로지 엘엘시
미국 02421 매사추세츠주 렉싱턴 스프링 스트리트 201</p> <p>(72) 발명자
푸니어 데이비드
미국 07030 뉴저지주 호보콘 #3 몬로 스트리트 318</p> <p>(74) 대리인
유미특허법인</p> |
|---|--|

전체 청구항 수 : 총 90 항

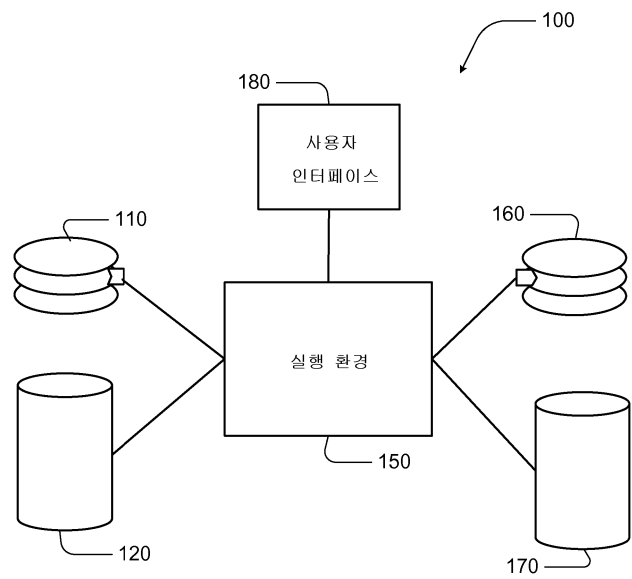
심사관 : 김병균

(54) 발명의 명칭 비밀 정보의 검출

(57) 요약

비밀 정보의 검출은, 저장된 데이터를 판독하여 저장된 데이터 내의 스트링들을 식별하며(210), 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함한다. 상기 스트링들 중 적어도 일부의 각각에 대해, 상기 스트링이 하나 이상의 포맷 일치를 나타내는 바이트들을 포함하는지를 결정하며(220-270), 포맷 일치는 비밀 정보와 연관된 미리 정해진 포맷과 일치하는 값의 세트를 포함한다. 각각의 포맷 일치에 대해, 상기 미리 정해진 포맷과 일치하는 값을 상기 비밀 정보와 연관된 규칙의 세트로 테스트하여 상기 포맷 일치가 하나 이상의 무효값을 포함하는 무효한 포맷 일치인지를 결정하고, 다른 포맷 일치의 카운트값에 대한 무효한 포맷 일치의 카운트값의 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어를 계산한다(280, 300).

대표도 - 도1



명세서

청구범위

청구항 1

비밀 정보를 검출하기 위한 방법으로서,

저장된 데이터를 판독하는 단계;

상기 저장된 데이터 내의 스트링을 식별하는 단계로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 단계;

상기 스트링에 기초하여 하나 이상의 포맷 일치를 식별하는 단계로서, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하는, 포맷 일치 식별 단계;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 포맷 일치가, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 무효한 하나 이상의 무효값이 포함되어 있는 무효한 포맷 일치인지 여부를 결정하는 단계; 및

상기 규칙에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 카운트값과 무효한 포맷 일치의 카운트값의 함수인 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제1 스코어를 계산하는 단계

를 포함하는 비밀 정보 검출 방법.

청구항 2

제1항에 있어서,

상기 비밀 정보는 신용 카드 번호인, 비밀 정보 검출 방법.

청구항 3

제2항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 신용 카드 번호의 표준 포맷에서의 숫자의 수와 동일할 때 발생하도록 결정되는, 비밀 정보 검출 방법.

청구항 4

제3항에 있어서,

상기 신용 카드 번호와 연관된 규칙들은 유효한 발행자 식별 번호의 리스트에 대한 명세를 포함하는, 비밀 정보 검출 방법.

청구항 5

제3항에 있어서,

상기 신용 카드 번호와 연관된 규칙들은 체크섬 알고리즘에 대한 명세를 포함하는, 비밀 정보 검출 방법.

청구항 6

제1항에 있어서,

상기 비밀 정보는 사회 보장 번호인, 비밀 정보 검출 방법.

청구항 7

제6항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 9와 동일할 때 발생하도록 결정되는, 비밀 정보 검출 방법.

청구항 8

제7항에 있어서,

상기 사회 보장 번호와 연관된 규칙들은 상기 사회 보장 번호의 처음 다섯 숫자가 나타내는 번호에 대한 값의 유효한 서브세트의 명세를 포함하는, 비밀 정보 검출 방법.

청구항 9

제1항에 있어서,

상기 비밀 정보는 전화 번호인, 비밀 정보 검출 방법.

청구항 10

제9항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 10과 동일하거나 상기 스트링 내에서 검출된 숫자의 수가 첫 번째 숫자가 "1"인 11개 숫자와 동일할 때 발생하도록 결정되는, 비밀 정보 검출 방법.

청구항 11

제10항에 있어서,

상기 전화 번호와 연관된 규칙들은 유효한 지역 번호의 리스트의 명세를 포함하는, 비밀 정보 검출 방법.

청구항 12

제10항에 있어서,

상기 전화 번호와 연관된 규칙들은 지역 번호 뒤의 첫 번째 숫자가 1 또는 0이 아니어야 한다는 명세를 포함하는, 비밀 정보 검출 방법.

청구항 13

제1항에 있어서,

상기 비밀 정보는 우편 번호인, 비밀 정보 검출 방법.

청구항 14

제13항에 있어서,

포맷 일치는 숫자를 나타내는 값을 가지는 5 바이트 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 9개의 숫자를 나타내는 값을 가지는 10 바이트로 구성되는 바이트의 시퀀스가 검출될 때 발생하도록 결정되는, 비밀 정보 검출 방법.

청구항 15

제14항에 있어서,

상기 우편 번호와 연관된 규칙들은 유효한 다섯 자릿수 우편 번호의 리스트에 대한 명세를 포함하는, 비밀 정보 검출 방법.

청구항 16

제1항에 있어서,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하는 단계로서, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스인, 단계; 및

상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 17

제16항에 있어서,

상기 성명의 리스트는 두 개의 서브세트: 이름과 성으로 나뉘어져 있는, 비밀 정보 검출 방법.

청구항 18

제17항에 있어서,

각 스트링에 대해, 상기 스트링이 하나 이상의 완전한 성명을 포함하는지를 결정하며, 완전한 성명은 공백이 뒤따르는 이름(first name)의 리스트로부터의 이름과, 이를 뒤따르는 성(last name)의 리스트로부터의 성으로 구성되는 문자의 시퀀스인, 단계; 및

검출된 완전한 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제3 스코어를 계산하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 19

제16항에 있어서,

상기 리스트 내의 성명 각각은 빈도수와 연관되고, 상기 저장된 데이터 내에 나타나는 성명들에 대한 평균 빈도수가 계산되며, 상기 저장된 데이터에 대한 제2 스코어는 적어도 부분적으로 상기 평균 빈도수에 기초하여 계산되는, 비밀 정보 검출 방법.

청구항 20

제19항에 있어서,

상기 평균 빈도수는 상기 저장된 데이터 내에서 검출된 성명의 수가 임계값 미만일 경우에 무시되는, 비밀 정보 검출 방법.

청구항 21

제1항에 있어서,

상기 방법은, 각 스트링에 대해, 두 글자로 이루어지는 단어의 수를 카운팅하는 단계를 더 포함하며, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스인, 단계를 더 포함하는 비밀 정보 검출 방법.

청구항 22

제21항에 있어서,

두 글자 단어 각각에 대해, 상기 두 글자 단어가 유효한 주의 약어인지를 결정하는 단계; 및

상기 유효한 주의 약어의 카운트값 및 상기 두 글자 단어의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 23

제1항에 있어서,

각 스트링에 대해, 상기 스트링이 하나 이상의 주/우편번호 쌍을 포함하는지를 결정하며, 주/우편번호 쌍은 공백이 뒤 따르고 그 뒤에 우편번호가 차례로 뒤따르는 주의 약어로 구성되는 문자의 시퀀스인, 단계; 및

검출된 주/우편번호 쌍의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 24

제1항에 있어서,

전자적으로 저장된 파일 시스템 내의 어떤 파일이 최근에 갱신되었는지를 검출하는 단계; 및

최근에 갱신되었던 각각의 파일에 대해 제1항의 방법을 적용하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 25

제1항에 있어서,

상기 스트링을 규정하는 바이트 값의 서브세트는 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백을 표현하는, 비밀 정보 검출 방법.

청구항 26

제1항에 있어서,

상기 제1 스코어를 임계값과 비교하는 단계; 및

상기 제1 스코어가 상기 임계값을 초과하는 경우, 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 단계

를 더 포함하는 비밀 정보 검출 방법.

청구항 27

비밀 정보를 검출하기 위한 시스템으로서,

데이터 스토리지 디바이스; 및

상기 데이터 스토리지 디바이스에 연결된 실행 환경

을 포함하고,

상기 실행 환경은,

상기 데이터 스토리지 디바이스로부터 저장된 데이터를 판독하도록 구성되고;

상기 저장된 데이터 내의 스트링을 식별하도록 구성되고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하며;

상기 스트링에 기초하여 하나 이상의 포맷 일치를 식별하도록 구성되고, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하며;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 포맷 일치가, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 무효한 하나 이상의 무효값이 포함되어 있는 무효한 포맷 일치인지 여부를 결정하도록 구성되고,

상기 규칙의 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 카운트값과 무효한 포맷 일치의 카운트값의 함수인 비율에 적어도 부분적으로 기초하여,

상기 저장된 데이터에 대한 제1 스코어를 계산하도록 구성되는,
비밀 정보 검출 시스템.

청구항 28

제27항에 있어서,

상기 비밀 정보는 신용 카드 정보이고, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 신용 카드 번호의 표준 포맷에서의 숫자의 수와 동일할 때 발생하도록 결정되어 있으며, 상기 신용 카드 번호와 연관된 규칙들은 유효한 발행자 식별 번호의 리스트에 대한 명세 및 체크섬 알고리즘에 대한 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 29

제27항에 있어서,

상기 비밀 정보는 사회 보장 번호이고, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 9와 동일할 때 발생하도록 결정되어 있으며, 상기 사회 보장 번호와 연관된 규칙들은 상기 사회 보장 번호의 처음 다섯 숫자가 나타내는 번호에 대한 값의 유효한 서브세트의 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 30

제27항에 있어서,

상기 비밀 정보는 전화 번호이고, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 10과 동일하거나 상기 스트링 내에서 검출된 숫자의 수가 첫 번째 숫자가 "1"인 11개 숫자와 동일할 때 발생하도록 결정되며, 상기 전화 번호와 연관된 규칙들은 유효한 지역 번호의 명세 및 상기 지역 번호 뒤의 첫 번째 숫자가 1 또는 0이 아니어야 한다는 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 31

제27항에 있어서,

상기 비밀 정보는 우편 번호이고, 포맷 일치는 숫자를 나타내는 값을 가지는 5 바이트 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 9개의 숫자를 나타내는 값을 가지는 10 바이트로 구성되는 바이트의 시퀀스가 검출될 때 발생하도록 결정되고, 상기 우편 번호와 연관된 규칙들은 유효한 다섯 자릿수 우편 번호의 리스트에 대한 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 32

제27항에 있어서,

상기 실행 환경은 또한,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하도록 구성되고, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이며, 상기 리스트는 두 개의 서브세트: 이름과 성으로 나뉘어져 있으며;

각 스트링에 대해, 상기 스트링이 하나 이상의 완전한 성명을 포함하는지를 결정하도록 구성되고, 완전한 성명은 공백이 뒤따르는 이름(first name)의 리스트로부터의 이름과, 이를 뒤따르는 성(last name)의 리스트로부터의 성으로 구성되는 문자의 시퀀스이며;

상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하고, 검출된 완전한 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하도록 구성되는, 비밀 정보 검출 시스템.

청구항 33

제27항에 있어서,

상기 실행 환경은 또한,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하도록 구성되고, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이며, 상기 리스트 내의 성명 각각은 빈도수와 연관되며;

상기 저장된 데이터 내에 나타나는 성명들에 대한 평균 빈도수를 계산하도록 구성되고;

상기 저장된 데이터에 대한 제2 스코어는 적어도 부분적으로 상기 저장된 데이터 내에서 검출된 성명의 카운트에 기초하고 적어도 부분적으로 상기 평균 빈도수에 기초하여, 계산되도록 구성되는, 비밀 정보 검출 시스템.

청구항 34

제27항에 있어서,

상기 실행 환경은 또한,

각 스트링에 대해, 상기 스트링이 하나 이상의 주/우편번호 쌍을 포함하는지를 결정하도록 구성되고, 주/우편번호 쌍은 공백이 뒤 따르고 그 뒤에 우편번호가 차례로 뒤따르는 주의 약어로 구성되는 문자의 시퀀스이고;

검출된 주/우편번호 쌍의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하도록 구성되는, 비밀 정보 검출 시스템.

청구항 35

제27항에 있어서,

상기 스트링을 규정하는 바이트 값의 서브세트는 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백을 표현하는, 비밀 정보 검출 시스템.

청구항 36

비밀 정보를 검출하기 위한 컴퓨터 프로그램을 저장한, 컴퓨터로 판독가능한 매체로서,

상기 컴퓨터 프로그램은 컴퓨터로 하여금,

저장된 데이터를 판독하게 하고;

상기 저장된 데이터 내의 스트링을 식별하게 하고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하며;

상기 스트링에 기초하여 하나 이상의 포맷 일치를 식별하게 하고, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하고;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 포맷 일치가, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 무효한 하나 이상의 무효값이 포함되어 있는 무효한 포맷 일치인지 여부를 결정하게 하며;

상기 규칙의 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 카운트값과 무효한 포맷 일치의 카운트값의 함수인 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제1 스코어를 계산하게 하는,

명령어를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 37

제36항에 있어서,

상기 비밀 정보는 신용 카드 번호인, 컴퓨터로 판독가능한 매체.

청구항 38

제37항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 신용 카드 번호의 표준 포맷에서의 숫자의 수와 동일할 때 발생하도록 결정되는, 컴퓨터로 판독가능한 매체.

청구항 39

제38항에 있어서,

상기 신용 카드 번호와 연관된 규칙들은 유효한 발행자 식별 번호의 리스트에 대한 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 40

제38항에 있어서,

상기 신용 카드 번호와 연관된 규칙들은 체크섬 알고리즘에 대한 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 41

제36항에 있어서,

상기 비밀 정보는 사회 보장 번호인, 컴퓨터로 판독가능한 매체.

청구항 42

제41항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 9와 동일할 때 발생하도록 결정되는, 컴퓨터로 판독가능한 매체,

청구항 43

제42항에 있어서,

상기 사회 보장 번호와 연관된 규칙들은 상기 사회 보장 번호의 처음 다섯 숫자가 나타내는 번호에 대한 값의 유효한 서브세트의 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 44

제36항에 있어서,

상기 비밀 정보는 전화 번호인, 컴퓨터로 판독가능한 매체.

청구항 45

제44항에 있어서,

포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 10과 동일하거나 상기 스트링 내에서 검출된 숫자의 수가 첫 번째 숫자가 "1"인 11개 숫자와 동일할 때 발생하도록 결정되는, 컴퓨터로 판독가능한 매체.

청구항 46

제45항에 있어서,

상기 전화 번호와 연관된 규칙들은 유효한 지역 번호의 리스트의 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 47

제45항에 있어서,

상기 전화 번호와 연관된 규칙들은 지역 번호 뒤의 첫 번째 숫자가 1 또는 0이 아니어야 한다는 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 48

제36항에 있어서,

상기 비밀 정보는 우편 번호인, 컴퓨터로 판독가능한 매체.

청구항 49

제48항에 있어서,

포맷 일치는 숫자를 나타내는 값을 가지는 5 바이트 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 9개의 숫자를 나타내는 값을 가지는 10 바이트로 구성되는 바이트의 시퀀스가 검출될 때 발생하도록 결정되는, 컴퓨터로 판독가능한 매체.

청구항 50

제49항에 있어서,

상기 우편 번호와 연관된 규칙들은 유효한 다섯 자릿수 우편 번호의 리스트에 대한 명세를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 51

제36항에 있어서,

상기 컴퓨터로 하여금,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하게 하고, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이며;

상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 52

제51항에 있어서,

상기 성명의 리스트는 두 개의 서브세트: 이름과 성으로 나뉘어져 있는, 컴퓨터로 판독가능한 매체.

청구항 53

제52항에 있어서,

상기 컴퓨터로 하여금,

각 스트링에 대해, 상기 스트링이 하나 이상의 완전한 성명을 포함하는지를 결정하게 하고, 완전한 성명은 공백이 뒤따르는 이름(first name)의 리스트로부터의 이름과, 이를 뒤따르는 성(last name)의 리스트로부터의 성으로 구성되는 문자의 시퀀스이며;

검출된 완전한 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제3 스코어를 계산하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 54

제51항에 있어서,

상기 리스트 내의 성명 각각은 빈도수와 연관되고, 상기 저장된 데이터 내에 나타나는 성명들에 대한 평균 빈도수가 계산되며, 상기 저장된 데이터에 대한 제2 스코어는 적어도 부분적으로 상기 평균 빈도수에 기초하여 계산되는, 컴퓨터로 판독가능한 매체.

청구항 55

제54항에 있어서,

상기 평균 빈도수는 상기 저장된 데이터 내에서 검출된 성명의 수가 임계값 미만일 경우에 무시되는, 컴퓨터로 판독가능한 매체.

청구항 56

제36항에 있어서,

상기 컴퓨터로 하여금,

각 스트링에 대해, 두 글자로 이루어지는 단어의 수를 카운팅하게 하고, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스인,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 57

제56항에 있어서,

상기 컴퓨터로 하여금,

두 글자 단어 각각에 대해, 상기 두 글자 단어가 유효한 주의 약어인지를 결정하게 하고;

상기 유효한 주의 약어의 카운트값 및 상기 두 글자 단어의 카운트값에 적어도 부분적으로 기초하여 상기 저장된 데이터에 대한 제2 스코어를 계산하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 58

제36항에 있어서,

상기 컴퓨터로 하여금,

각 스트링에 대해, 상기 스트링이 하나 이상의 주/우편번호 쌍을 포함하는지를 결정하게 하고, 주/우편번호 쌍은 공백이 뒤 따르고 그 뒤에 우편번호가 차례로 뒤따르는 주의 약어로 구성되는 문자의 시퀀스이며;

검출된 주/우편번호 쌍의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 59

제36항에 있어서,

상기 컴퓨터로 하여금,

전자적으로 저장된 파일 시스템 내의 어떤 파일이 최근에 갱신되었는지를 검출하게 하고;

최근에 갱신되었던 각각의 파일에 대해 제1항의 방법을 적용하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 60

제36항에 있어서,

상기 스트링을 규정하는 바이트 값의 서브세트는 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백을 표현하는, 컴퓨터로 판독가능한 매체.

청구항 61

제36항에 있어서,

상기 컴퓨터로 하여금,

상기 제1 스코어를 임계값과 비교하게 하고;

상기 제1 스코어가 상기 임계값을 초과하는 경우, 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하게 하는,

명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 62

비밀 정보를 검출하기 위한 방법으로서,

저장된 데이터를 판독하는 단계;

상기 저장된 데이터 내의 스트링을 식별하는 단계로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 단계;

상기 스트링에 기초하여 포맷 일치를 식별하는 단계로서, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하는, 포맷 일치 식별 단계;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부가 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값을 포함하는지 여부를 결정하는 단계;

각각의 포맷 일치에 대하여 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값이 포함되어 있는 포맷 일치의 수의 제1 카운트값을 결정하는 단계;

상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 무효값을 포함하지 않는 포맷 일치의 수의 제2 카운트값을 결정하는 단계; 및

상기 제1 카운트값과 상기 제2 카운트값의 비교에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 단계

를 포함하는 비밀 정보 검출 방법.

청구항 63

제62항에 있어서,

상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 단계는, 상기 제1 카운트값과 상기 제2 카운트값에 의존하는 스코어가 임계값을 초과할 정도로 상기 제1 카운트값이 상기 제2 카운트값에 비해 충분히 작은 것에 적어도 부분적으로 기초하는, 비밀 정보 검출 방법.

청구항 64

비밀 정보를 검출하기 위한 시스템으로서,

데이터 스토리지 디바이스; 및

상기 데이터 스토리지 디바이스에 연결된 실행 환경을 포함하고,

상기 실행 환경은,

상기 데이터 스토리지 디바이스로부터 저장된 데이터를 판독하도록 구성되고;

상기 저장된 데이터 내의 스트링을 식별하도록 구성되고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하며;

상기 스트링에 기초하여 포맷 일치를 식별하도록 구성되고, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하며;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부가 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값을 포함하는지 여부를 결정하도록 구성되며;

각각의 포맷 일치에 대하여 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값이 포함되어 있는 포맷 일치의 수의 제1 카운트값을 결정하도록 구성되고;

상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 무효값을 포함하지 않는 포맷 일치의 수의 제2 카운트값을 결정하도록 구성되며;

상기 제1 카운트값과 상기 제2 카운트값의 비교에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하도록 구성되는,

비밀 정보 검출 시스템.

청구항 65

제64항에 있어서,

상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 것은, 상기 제1 카운트값과 상기 제2 카운트값에 의존하는 스코어가 임계값을 초과할 정도로 상기 제1 카운트값이 상기 제2 카운트값에 비해 충분히 작은 것에 적어도 부분적으로 기초하는, 비밀 정보 검출 시스템.

청구항 66

비밀 정보를 검출하기 위한 컴퓨터 프로그램을 저장한, 컴퓨터로 판독가능한 매체로서,

상기 컴퓨터 프로그램은 컴퓨터로 하여금,

저장된 데이터를 판독하게 하고;

상기 저장된 데이터 내의 스트링을 식별하게 하고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하며;

상기 스트링에 기초하여 포맷 일치를 식별하게 하고, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하며;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부가 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값을 포함하는지 여부를 결정하게 하고;

각각의 포맷 일치에 대하여 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값이 포함되어 있는 포맷 일치의 수의 제1 카운트값을 결정하게 하며;

상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 무효값을 포함하지 않는 포맷 일치의 수의 제2 카운트값을 결정하게 하고;

상기 제1 카운트값과 상기 제2 카운트값의 비교에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하게 하는,

명령어를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 67

제66항에 있어서,

상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 것은, 상기 제1 카운트값과 상기 제2 카운트값에 의존하는 스코어가 임계값을 초과할 정도로 상기 제1 카운트값이 상기 제2 카운트값에 비해 충분히 작은 것에 적어도 부분적으로 기초하는, 컴퓨터로 판독가능한 매체.

청구항 68

비밀 정보를 검출하기 위한 시스템으로서,

저장된 데이터를 판독하기 위한 수단;

상기 저장된 데이터 내의 스트링을 식별하기 위한 수단으로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 수단;

상기 스트링에 기초하여 하나 이상의 포맷 일치를 식별하기 위한 수단으로서, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하는, 포맷 일치 식별 수단;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 포맷 일치가, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 무효한 하나 이상의 무효값이 포함되어 있는 무효한 포맷 일치인지 여부를 결정하기 위한 수단; 및

상기 규칙에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 카운트값과 무효한 포맷 일치의 카운트값의 함수인 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제1 스코어를 계산하기 위한 수단

을 포함하는, 비밀 정보 검출 시스템.

청구항 69

제68항에 있어서,

상기 비밀 정보는 신용 카드 번호이고, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 신용 카드 번호의 표준 포맷에서의 숫자의 수와 동일할 때 발생하도록 결정되며, 상기 신용 카드 번호와 연관된 규칙들은 유효한 발행자 식별 번호의 리스트에 대한 명세 및 체크섬 알고리즘에 대한 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 70

제68항에 있어서,

상기 비밀 정보는 사회 보장 번호이고, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 9와 동일할 때 발생하도록 결정되며, 상기 사회 보장 번호와 연관된 규칙들은 상기 사회 보장 번호의 처음 다섯 숫자가 나타내는 번호에 대한 값의 유효한 서브세트의 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 71

제68항에 있어서,

상기 비밀 정보는 전화 번호이며, 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 10과 동일하거나 상기 스트링 내에서 검출된 숫자의 수가 첫 번째 숫자가 "1"인 11개 숫자와 동일할 때 발생하도록 결정되고, 상기 전화 번호와 연관된 규칙들은 유효한 지역 번호의 리스트의 명세 및 상기 지역 번호 뒤의 첫 번째 숫자가 1 또는 0이 아니어야 한다는 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 72

제68항에 있어서,

상기 비밀 정보는 우편 번호이며, 포맷 일치는 숫자를 나타내는 값을 가지는 5 바이트 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 9개의 숫자를 나타내는 값을 가지는 10 바이트로 구성되는 바이트의 시퀀스가 검출될 때 발생하도록 결정되고, 상기 우편 번호와 연관된 규칙들은 유효한 다섯 자릿수 우편 번호의 리스트에 대한 명세를 포함하는, 비밀 정보 검출 시스템.

청구항 73

제68항에 있어서,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하기 위한 수단으로서, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이며, 상기 리스트는 두 개의 서브세트: 이름과 성으로 나뉘어져 있는, 수단;

각 스트링에 대해, 상기 스트링이 하나 이상의 완전한 성명을 포함하는지를 결정하기 위한 수단으로서, 완전한 성명은 공백이 뒤따르는 이름의 리스트로부터의 이름과, 이를 뒤따르는 성의 리스트로부터의 성으로 구성되는 문자의 시퀀스인, 수단; 및

상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하여, 그리고 검출된 완전한 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하기 위한 수단

을 더 포함하는, 비밀 정보 검출 시스템.

청구항 74

제68항에 있어서,

각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하기 위한 수단으로서, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이며, 상기 리스트 내의 성명 각각은 빈도수와 연관되는, 수단;

상기 저장된 데이터 내에 나타나는 성명들에 대한 평균 빈도수를 계산하기 위한 수단; 및

상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하여, 그리고 상기 평균 빈도수에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하기 위한 수단

을 더 포함하는, 비밀 정보 검출 시스템.

청구항 75

제68항에 있어서,

각 스트링에 대해, 상기 스트링이 하나 이상의 주/우편번호 쌍을 포함하는지를 결정하기 위한 수단으로서, 주/우편번호 쌍은 공백이 뒤 따르고 그 뒤에 우편번호가 차례로 뒤따르는 주의 약어로 구성되는 문자의 시퀀스인, 수단; 및

검출된 주/우편번호 쌍의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 제2 스코어를 계산하기 위한 수단

을 더 포함하는, 비밀 정보 검출 시스템.

청구항 76

제68항에 있어서,

상기 스트링을 규정하는 바이트 값의 서브세트는 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백을 표현하는, 비밀 정보 검출 시스템.

청구항 77

비밀 정보를 검출하기 위한 시스템으로서,

저장된 데이터를 판독하기 위한 수단;

상기 저장된 데이터 내의 스트링을 식별하기 위한 수단으로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 수단;

상기 스트링에 기초하여 포맷 일치를 식별하기 위한 수단으로서, 각각의 포맷 일치는 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 적어도 일부를 포함하는, 포맷 일치 식별 수단;

각각의 포맷 일치에 대해, 상기 비밀 정보의 제1 타입과 연관된 규칙의 세트를 이용하여 각각의 상기 포맷 일치를 테스트함으로써, 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부가 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값을 포함하는지 여부를 결정하기 위한 수단;

각각의 포맷 일치에 대하여 상기 미리 정해진 포맷과 일치하는 상기 스트링 중 하나의 스트링의 일부에 상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 하나 이상의 무효값이 포함되어 있는 포맷 일치의 수의 제1 카운트값을 결정하기 위한 수단;

상기 규칙에 따라 상기 비밀 정보의 제1 타입에 대해 무효값을 포함하지 않는 포맷 일치의 수의 제2 카운트값을 결정하기 위한 수단; 및

상기 제1 카운트값과 상기 제2 카운트값의 비교에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하기 위한 수단

을 포함하는, 비밀 정보 검출 시스템.

청구항 78

제77항에 있어서,

상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 것은, 상기 제1 카운트값과 상기 제2 카운트값에 의존하는 스코어가 임계값을 초과할 정도로 상기 제1 카운트값이 상기 제2 카운트값에 비해 충분히 작은 것에 적어도 부분적으로 기초하는, 비밀 정보 검출 시스템.

청구항 79

비밀 정보를 검출하기 위한 방법으로서,

저장된 데이터를 판독하는 단계;

상기 저장된 데이터 내의 스트링을 식별하는 단계로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 단계;

식별된 각각의 상기 스트링에 하나 이상의 규칙의 제1 세트를 적용하여, 상기 스트링의 적어도 일부가 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 포맷 일치가 있는지 여부를 결정하는 단계;

포맷 일치와 연관된 각각의 스트링에 대해, 상기 비밀 정보의 제1 타입과 연관된 하나 이상의 규칙의 제2 세트를 상기 스트링에 적용하여, 상기 포맷 일치가, 상기 비밀 정보의 제1 타입과 연관된 상기 미리 정해진 포맷과 일치하는 상기 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 유효한 값이 포함되지 않는 무효한 포맷 일치인지 여부를 결정하는 단계;

상기 무효한 포맷 일치의 제1 카운트값을 계산하는 단계;

상기 하나 이상의 규칙의 제2 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 제2 카운트값을 계산하는 단계; 및

상기 제1 카운트값과 상기 제2 카운트값의 함수에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하는 단계

를 포함하는, 비밀 정보 검출 방법.

청구항 80

제79항에 있어서,

상기 비밀 정보의 제1 타입은 사회 보장 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 9-자리 번호가 무효한 사회 보장 번호인 9-자리 번호를 각각 가지는 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 방법.

청구항 81

제79항에 있어서,

상기 비밀 정보의 제1 타입은 전화 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 첫 번째 숫자가 "1"인 11-자리 번호 또는 10-자리 번호를 각각 포함하면서 11-자리 번호 또는 10-자리 번호가 무효한 전화 번호인 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 방법.

청구항 82

비밀 정보를 검출하기 위한 시스템으로서,

데이터 스토리지 디바이스; 및

상기 데이터 스토리지 디바이스에 연결된 실행 환경

을 포함하고,

상기 실행 환경은,

저장된 데이터를 판독하도록 구성되고,

상기 저장된 데이터 내의 스트링을 식별하도록 구성되며, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하고;

식별된 각각의 상기 스트링에 하나 이상의 규칙의 제1 세트를 적용하여, 상기 스트링의 적어도 일부가 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 포맷 일치가 있는지 여부를 결정하도록 구성되고;

포맷 일치와 연관된 각각의 스트링에 대해, 상기 비밀 정보의 제1 타입과 연관된 하나 이상의 규칙의 제2 세트를 상기 스트링에 적용하여, 상기 포맷 일치가, 상기 비밀 정보의 제1 타입과 연관된 상기 미리 정해진 포맷과 일치하는 상기 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 유효한 값이 포함되지 않는 무효한 포맷 일치인지 여부를 결정하도록 구성되며;

상기 무효한 포맷 일치의 제1 카운트값을 계산하도록 구성되고;

상기 하나 이상의 규칙의 제2 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 제2 카운트값을 계산하도록 구성되며;

상기 제1 카운트값과 상기 제2 카운트값의 함수에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하도록 구성되는,

비밀 정보 검출 시스템.

청구항 83

제82항에 있어서,

상기 비밀 정보의 제1 타입은 사회 보장 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 9-자리 번호가 무효한 사회 보장 번호인 9-자리 번호를 각각 가지는 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 시스템.

청구항 84

제82항에 있어서,

상기 비밀 정보의 제1 타입은 전화 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 첫 번째 숫자가 "1"

인 11-자리 번호 또는 10-자리 번호를 각각 포함하면서 11-자리 번호 또는 10-자리 번호가 무효한 전화 번호인 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 시스템.

청구항 85

비밀 정보를 검출하기 위한 컴퓨터 프로그램을 저장한, 컴퓨터로 판독가능한 매체로서,

상기 컴퓨터 프로그램은 컴퓨터로 하여금,

저장된 데이터를 판독하게 하고,

상기 저장된 데이터 내의 스트링을 식별하게 하며, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하고;

식별된 각각의 상기 스트링에 하나 이상의 규칙의 제1 세트를 적용하여, 상기 스트링의 적어도 일부가 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 포맷 일치가 있는지 여부를 결정하게 하고;

포맷 일치와 연관된 각각의 스트링에 대해, 상기 비밀 정보의 제1 타입과 연관된 하나 이상의 규칙의 제2 세트를 상기 스트링에 적용하여, 상기 포맷 일치가, 상기 비밀 정보의 제1 타입과 연관된 상기 미리 정해진 포맷과 일치하는 상기 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 유효한 값이 포함되지 않는 무효한 포맷 일치인지 여부를 결정하게 하며;

상기 무효한 포맷 일치의 제1 카운트값을 계산하게 하고;

상기 하나 이상의 규칙의 제2 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 제2 카운트값을 계산하게 하며;

상기 제1 카운트값과 상기 제2 카운트값의 함수에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하게 하는,

명령어를 포함하는, 컴퓨터로 판독가능한 매체.

청구항 86

제85항에 있어서,

상기 비밀 정보의 제1 타입은 사회 보장 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 9-자리 번호가 무효한 사회 보장 번호인 9-자리 번호를 각각 가지는 스트링의 수를 결정하는 것을 포함하는, 컴퓨터로 판독가능한 매체.

청구항 87

제85항에 있어서,

상기 비밀 정보의 제1 타입은 전화 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 첫 번째 숫자가 "1"인 11-자리 번호 또는 10-자리 번호를 각각 포함하면서 11-자리 번호 또는 10-자리 번호가 무효한 전화 번호인 스트링의 수를 결정하는 것을 포함하는, 컴퓨터로 판독가능한 매체.

청구항 88

비밀 정보를 검출하기 위한 시스템으로서,

저장된 데이터를 판독하기 위한 수단;

상기 저장된 데이터 내의 스트링을 식별하기 위한 수단으로서, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 스트링 식별 수단;

식별된 각각의 상기 스트링에 하나 이상의 규칙의 제1 세트를 적용하여, 상기 스트링의 적어도 일부가 비밀 정보의 제1 타입과 연관된 미리 정해진 포맷과 일치하는 포맷 일치가 있는지 여부를 결정하기 위한 수단;

포맷 일치와 연관된 각각의 스트링에 대해, 상기 비밀 정보의 제1 타입과 연관된 하나 이상의 규칙의 제2 세트를 상기 스트링에 적용하여, 상기 포맷 일치가, 상기 비밀 정보의 제1 타입과 연관된 상기 미리 정해진 포맷과 일치하는 상기 스트링의 일부에 상기 비밀 정보의 제1 타입에 대해 유효한 값이 포함되지 않는 무효한 포맷 일

치인지 여부를 결정하기 위한 수단;

상기 무효한 포맷 일치의 제1 카운트값을 계산하기 위한 수단;

상기 하나 이상의 규칙의 제2 세트에 따라 결정된 바와 같은 상기 비밀 정보의 제1 타입에 대해 무효한 무효값을 포함하지 않는 포맷 일치의 제2 카운트값을 계산하기 위한 수단; 및

상기 제1 카운트값과 상기 제2 카운트값의 함수에 적어도 부분적으로 기초하여 상기 저장된 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정하기 위한 수단

을 포함하는, 비밀 정보 검출 시스템.

청구항 89

제88항에 있어서,

상기 비밀 정보의 제1 타입은 사회 보장 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 9-자리 번호가 무효한 사회 보장 번호인 9-자리 번호를 각각 가지는 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 시스템.

청구항 90

제88항에 있어서,

상기 비밀 정보의 제1 타입은 전화 번호를 포함하며, 상기 제1 카운트값을 결정하는 것은, 첫 번째 숫자가 "1"인 11-자리 번호 또는 10-자리 번호를 각각 포함하면서 11-자리 번호 또는 10-자리 번호가 무효한 전화 번호인 스트링의 수를 결정하는 것을 포함하는, 비밀 정보 검출 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 비밀 정보의 검출에 관한 것이다.

배경 기술

[0002] 어떤 데이터 처리 환경에서, 보안이 불충분한(insufficiently secured) 데이터 스토리지 디바이스 내의 전자 파일에는 비밀 정보가 나타날 수 있다. 안전하지 못한 파일(insecure file) 내의 이러한 비밀 정보의 존재는 개인의 보안 및 프라이버시를 위태롭게 할 수 있으며, 이것은 안전하지 못한 데이터 스토리지 시스템을 운영하는 엔티티와 연관되어 법적 책임을 야기한다. 비밀 정보는 민감한 금융 데이터(financial data) 또는 특정한 개인을 식별하고 파일의 콘텐츠에 그들을 관련짓는 데 사용될 수 있는 임의의 정보를 포함할 수 있다. 비밀 정보의 몇몇 예로는 성명, 주소, 전화 번호, 사회 보장 번호(social security number), 및 신용 카드 번호를 포함한다.

발명의 내용

[0003] 전반적인 측면에서, 저장된 데이터를 판독하는 방법은, 상기 저장된 데이터 내의 스트링을 식별하며, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트(subset) 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하는, 단계; 상기 스트링들의 적어도 일부의 각각에 대해, 상기 스트링이 하나 이상의 포맷 일치(format match)를 나타내는 바이트들을 포함하는지를 결정하며, 포맷 일치는 비밀 정보와 연관된 미리 정해진 포맷과 일치하는 값의 세트를 포함하는, 단계; 각각의 포맷 일치에 대해, 상기 미리 정해진 포맷과 일치하는 값을 상기 비밀 정보와 연관된 규칙들의 세트로 테스트하여 상기 포맷 일치가 하나 이상의 무효값(invaild value)을 포함하는 무효한 포맷 일치인지를 결정하는 단계; 및 다른 포맷 일치의 카운트값(count)에 대한 무효한 포맷 일치의 카운트값의 비율(ratio)에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어(score)를 계산하는 단계를 포함한다.

[0004] 다른 전반적인 측면에서, 비밀 정보를 검출하는 시스템은, 데이터 스토리지 디바이스(data storage device); 및 상기 데이터 스토리지 디바이스에 연결된 실행 환경(runtime environment)을 포함한다. 상기 실행 환경은 상기 데이터 스토리지 디바이스로부터 저장된 데이터를 판독하도록 구성되고; 상기 저장된 데이터 내의 스트링들을 식별하도록 구성되고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하

는 바이트의 시퀀스를 포함하며; 상기 스트링들의 적어도 일부의 각각에 대해, 상기 스트링이 하나 이상의 포맷 일치를 나타내는 바이트들을 포함하는지를 결정하도록 구성되며, 포맷 일치는 비밀 정보와 연관된 미리 정해진 포맷과 일치하는 값의 세트를 포함하고; 각각의 포맷 일치에 대해, 상기 미리 정해진 포맷과 일치하는 값을 상기 비밀 정보와 연관된 규칙들의 세트로 테스트하여 상기 포맷 일치가 하나 이상의 무효값을 포함하는 무효한 포맷 일치인지를 결정하도록 구성되고; 다른 포맷 일치의 카운트값에 대한 무효한 포맷 일치의 카운트값의 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어를 계산하도록 구성된다.

[0005] 다른 전반적인 측면에서, 컴퓨터로 판독가능한 매체(computer-readable medium)는 비밀 정보를 검출하는 컴퓨터 프로그램을 저장한다. 상기 컴퓨터 프로그램은 컴퓨터로 하여금, 저장된 데이터를 판독하게 하고; 상기 저장된 데이터 내의 스트링들을 식별하게 하고, 각 스트링은, 모두가 가능한 값들의 미리 정해진 서브세트 내에 있는 값을 가지는 연속하는 바이트의 시퀀스를 포함하며; 상기 스트링들의 적어도 일부의 각각에 대해, 상기 스트링이 하나 이상의 포맷 일치를 나타내는 바이트들을 포함하는지를 결정하게 하고, 포맷 일치는 비밀 정보와 연관된 미리 정해진 포맷과 일치하는 값의 세트를 포함하고; 각각의 포맷 일치에 대해, 상기 비밀 정보와 연관된 규칙들의 세트로 상기 미리 정해진 포맷과 일치하는 값을 테스트하여 상기 포맷 일치가 하나 이상의 무효값을 포함하는 무효한 포맷 일치인지를 결정하게 하고; 무효한 포맷 일치의 카운트값의 다른 포맷 일치의 카운트값에 대한 비율에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어를 계산하게 하는, 명령어를 포함한다.

[0006] 상기한 측면들은 다음 특징들 중 하나 이상을 포함할 수 있다:

[0007] 상기 비밀 정보는 신용 카드 번호일 수 있다. 포맷 일치는 상기 스트링 내에서 검출된 숫자(digit)를 나타내는 값을 가지는 바이트의 수가 신용 카드 번호의 표준 포맷 내의 숫자들의 수와 동일할 때 발생하도록 결정될 수 있다. 신용 카드 번호와 연관된 규칙들은 유효한 발행자 식별 번호의 리스트에 대한 명세(specification)를 포함할 수 있다. 상기 신용 카드 번호와 연관된 규칙들은 체크섬 알고리즘(check sum algorithm)에 대한 명세를 포함할 수 있다.

[0008] 상기 비밀 정보는 사회 보장 번호일 수 있다. 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 9와 동일할 때 발생하도록 결정될 수 있다. 사회 보장 번호와 연관된 규칙들은 사회 보장 번호의 처음 다섯 숫자가 나타내는 번호에 대한 값의 유효한 서브세트의 명세를 포함할 수 있다.

[0009] 상기 비밀 정보는 전화 번호일 수 있다. 포맷 일치는 상기 스트링 내에서 검출된 숫자를 나타내는 값을 가지는 바이트의 수가 10과 동일하거나 상기 스트링 내에서 검출된 숫자의 수가, 첫 번째 숫자가 "1"인 11개의 숫자와 동일할 때 발생하도록 결정될 수 있다. 전화 번호와 연관된 규칙들은 유효한 지역 번호(valid area code)의 명세를 포함할 수 있다. 상기 전화 번호와 연관된 규칙들은 상기 지역 번호 뒤의 첫 번째 숫자가 1 또는 0(영)이 아니어야 한다는 명세를 포함할 수 있다.

[0010] 상기 비밀 정보는 우편 번호일 수 있다. 포맷 일치는 숫자를 나타내는 값을 가지는 5 바이트, 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 9개의 숫자를 나타내는 값을 가지는 10 바이트로 구성되는 바이트의 시퀀스가 검출될 때 발생하도록 결정될 수 있다. 우편 번호와 연관된 규칙들은 유효한 다섯 자릿수(five digit) 우편 번호의 리스트에 대한 명세를 포함할 수 있다.

[0011] 각 스트링에 대해, 상기 스트링이 성명과 일치하는 하나 이상의 단어를 포함하는지를 결정하며, 단어는, 모두가 알파벳 및 숫자 문자(alpha-numeric character)를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이고, 성명은 보통 개인(individual people)을 지칭하는 데 사용되는 시퀀스의 리스트에 의거한 문자의 시퀀스이고; 상기 저장된 데이터 내에서 검출된 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어를 계산한다. 상기 성명의 리스트는 두 개의 서브세트: 이름과 성으로 나뉠 수 있다.

[0012] 각 스트링에 대해, 상기 스트링이 하나 이상의 완전한 성명(full name)을 포함하는지를 결정하며, 완전한 성명은 공백(space)이 뒤따르는 이름(first name)의 리스트로부터의 이름과, 이를 뒤따르는 성(last name)의 리스트로부터의 성으로 구성되는 문자의 시퀀스이고; 검출된 완전한 성명의 카운트값에 적어도 부분적으로 기초하여, 상기 저장된 데이터에 대한 스코어를 계산한다.

[0013] 상기 리스트 내의 성명은 각각 그들과 연관된 빈도수(frequency count)를 가질 수 있으며, 상기 저장된 데이터 내에 나타나는 이름에 대한 평균 빈도수는 계산될 수 있고, 상기 저장된 데이터에 대한 스코어는 적어도 부분적으로 상기 평균 빈도수에 기초하여 계산될 수 있다. 상기 평균 빈도수는 상기 저장된 데이터 내에서 검출된 이름의 수가 임계값 미만일 경우에 무시될 수 있다.

- [0014] 각 스트링에 대해, 두 글자(two letter)로 이루어지는 단어의 수를 카운팅하며, 단어는, 모두가 알파벳 및 숫자 문자를 나타내는 값을 가지는 스트링 내의 연속하는 바이트의 시퀀스이다. 두 글자 단어 각각에 대해, 상기 두 글자 단어가 유효한 주의 약어(valid state abbreviation)인지를 결정하고; 상기 유효한 주의 약어의 카운트값 및 상기 두 글자 단어의 카운트값에 적어도 부분적으로 기초하여 상기 저장된 데이터에 대한 스코어를 계산한다.
- [0015] 각 스트링에 대해, 상기 스트링이 하나 이상의 주/우편 번호 쌍을 포함하는지를 결정하며, 주/우편 번호 쌍은 공백이 뒤 따르고 그 뒤에 우편 번호가 차례로 뒤따르는 주의 약어로 구성되는 문자의 시퀀스이고; 상기 검출된 주/우편 번호 쌍의 카운트값에 적어도 부분적으로 기초하여 상기 저장된 데이터에 대한 스코어를 계산한다.
- [0016] 전자적으로 저장된 파일 시스템 내의 어떤 파일이 최근에 갱신되었는지를 검색하고; 비밀 정보가 최근에 갱신되었던 각각의 파일을 검색한다.
- [0017] 스트링들을 규정하는 바이트 값의 서브세트는 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백을 포함할 수 있다.
- [0018] 상기 스코어를 임계값과 비교하고; 상기 스코어가 상기 임계값을 초과하는 경우, 상기 저장 데이터가 비밀 정보를 포함할 가능성이 있는 것으로 플래그를 설정한다(flagging).
- [0019] 상기한 측면들은 다음의 이점 중 하나 이상을 포함할 수 있다:
- [0020] 비밀 정보 검색의 자동화를 가능하게 한다. 효율적으로 비밀 정보를 검출함으로써 보안 및 프라이버시 보호 수단을 가능 및 강화시킨다.
- [0021] 본 발명의 다른 특징들 및 이점들은 아래의 구체적인 설명 및 특허청구범위로부터 더욱 명백해질 것이다.

도면의 간단한 설명

- [0022] 도 1은 저장된 데이터 내의 비밀 정보를 검출하는 시스템의 블록도이다.
- 도 2는 저장된 데이터 내의 비밀 정보를 검출하는 데 사용된 소프트웨어의 블록도이다.
- 도 3은 파일이 비밀 정보를 포함할 가능성을 나타내는 스코어를 계산하는 프로세스의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0023] 대량의 데이터 세트 내의 비밀 정보의 출현(occurrence)을 검출할 수 있도록 하는 것이 바람직하며, 비밀 데이터를 검색하는 과정에서 데이터의 많은 부분을 사람 에이전트(human agent)의 검토 없이 비밀 정보를 검출하는 것이 특히 바람직하다. 비밀 정보를 검출하는 시스템은 잠재적인 비밀 데이터를 자동으로 검출할 수 있으며, 이것은 그후 사람 에이전트에 의해 전부 또는 부분적으로 검토될 수 있다. 어떤 실시예에서, 사람이 비밀 데이터를 검토하는 것은 시간이 부족한 허가받은 직원(cleared personnel)으로 제한되거나 완전히 회피될 수도 있으며, 따라서 비밀 정보를 잘못 취급함으로써 야기되는 프라이버시 침해를 줄이거나 없앤다.
- [0024] 도 1은 전자적으로 저장된 데이터 내의 비밀 정보를 검출하는 예시적인 시스템을 나타낸다. 관심 데이터(data of interest)는, 병렬 파일 시스템(예컨대, 인용에 의해 본 명세서에 포함되는 미국특허 제5,897,638호에 개시된 것) 내의 다수의 디바이스에 구현된 병렬 "멀티파일(multifile)"(110)과 같은, 하나 이상의 데이터 스토리지 디바이스에 저장될 수 있다. 비밀 정보 검출(confidential information detection, CID) 시스템(100)은 실행 환경(150)에서 실행되는 소프트웨어를 사용하여 데이터 스토리지 디바이스 또는 디바이스들에 저장된 데이터를 분석한다. 파일과 같은, 저장된 데이터의 유닛 각각에 대한 스코어, 및 어쩌면 저장된 데이터 중 어떤 유닛이 비밀 정보를 포함할 가능성이 있는지를 나타내는 플래그를 포함하는, 분석 결과가 동일한 데이터 스토리지 디바이스(110, 120) 또는 다른 데이터 스토리지 디바이스(160, 170)에 기록될 수 있다. 어떤 경우에는, CID 시스템(100)의 실행을 구성(configure) 및 제어하기 위해서는 물론 결과를 검토하기 위해 운영자에 의해 사용자 인터페이스(80)가 사용된다.
- [0025] 실행 환경(150)은 UNIX 운영 체제(operating system)와 같은 적당한 운영 체제의 제어 하의 하나 이상의 범용 컴퓨터 상에 호스팅(hosting)될 수 있다. 예를 들면, 실행 환경(150)은 로컬(예컨대, SMP 컴퓨터와 같은 멀티프로세서 시스템), 또는 로컬 분산형(예컨대 클러스터 또는 MPP로서 연결된 다수의 프로세서) 또는 원격, 또는 원격 분산형(예컨대, LAN이나 WAN 네트워크를 통해 연결된 다수의 프로세서), 또는 이들의 임의의 조합의, 다수의 중앙 처리 유닛(central processing unit, CPU)을 사용하는 컴퓨터 시스템의 구성을 포함하는 다중 노드 병렬 컴퓨팅 환경(multiple-node parallel computing environment)을 포함할 수 있다. 실행 환경(150)에 의해

엑세스되는 입력, 출력, 또는 중간 데이터 세트는 병렬 파일 시스템(예컨대, 통신 링크를 통해 CID 시스템(100)에 로컬 또는 원격으로 연결된 데이터 스토리지 디바이스(160) 또는 외부 데이터 스토리지 디바이스(170))에 저장된 병렬 "멀티파일"일 수 있다.

[0026] 도 2는 전자적으로 저장된 파일 내의 비밀 정보를 검출하는 시스템을 구현하기 위해 실행 환경에서 실행될 수 있는 소프트웨어의 구성을 나타낸다. CID 시스템(100)은 모든 파일 포맷을 미지의 것(unknown)으로 취급하여 비밀 데이터를 포함하는 문자 스트링을 검색한다. 파일은 데이터 스토리지 디바이스(201)로부터 판독된다. 스트링 추출 모듈(string extraction module)(210)은 파일을 일련의 데이터 바이트로 취급한다. 스트링을 인식하기 위한 이 접근법은 비밀 정보를 나타내는 데이터에 사용된 문자를 나타내는 바이트, 또는 비밀 정보를 나타내는 데이터를 일반 포맷팅(common formatting)하는 바이트를 제외한 모든 바이트를 제거하는 것이다. 보통 바이트는 8비트이지만, 검출될 문자 세트에 적합한 임의의 사이즈로 정의될 수 있다. 예를 들면, 바이트는 16비트 또는 32비트 길이로 규정될 수 있다. 나타난 예의 시스템은 8비트의 바이트 사이즈를 사용한다.

[0027] 관심 문자와 연관된 가능한 바이트 값의 서브세트를 사용하여 스트링을 식별한다. 이 서브세트 외의 값을 가지는 바이트는 스트링 구분자(string delimiter)로 취급된다. 이 예에서, 서브세트 내의 바이트 값은 알파벳 및 숫자 문자, 괄호, 하이픈, 및 공백의 ASCII 표현(representation)이다. 괄호, 하이픈, 및 공백은 보통 전화 번호, SSN, 신용 카드 등과, 주소의 단어들 사이에서 포맷을 맞추는데 사용되는 문자이기 때문에 포함된다. 스트링 추출 모듈(210)에 의해 식별된 스트링들은 여러 비밀 정보 검출 모듈(220, 230, 240, 250, 260, 및 270)에 전달된다.

[0028] 신용 카드 번호 검출 모듈(220)은, 번호가 있는지 각 스트링을 검색하고 그 번호를 신용 카드 번호와 연관된 규칙의 세트와 대조한다. 이 규칙들은 숫자(digit)의 수(number)로 판단한 하나 이상의 허용된 신용 카드 번호 길이에 대한 명세를 포함한다. 신용 카드 번호 검출 모듈(220)은 숫자를 나타내는 바이트의 스트링을 검색함으로써 시작한다. 스트링에서 검출된 숫자의 수가 신용 카드 번호의 표준 포맷의 숫자 길이와 동일한 경우, 기본 포맷 일치에 선언된다. 기본 포맷 일치는 숫자 그룹 사이의 공백과 같이, 번호(number)를 나타내는 숫자들 사이에 다른 문자를 나타내는 바이트가 배치되더라도 발생할 수 있다. 기본 포맷 일치 각각에 대해, 스트링 내의 숫자들에 의해 나타난 번호는 신용 카드 번호와 연관된 규칙의 완전한 세트를 사용하여 테스트된다. 완전한 세트 내의 다른 규칙들은, 예를 들면, 유효한 발행자 식별 번호 또는 유효한 체크섬의 명세를 포함할 수 있다. 번호가 어떠한 규칙도 충족시키지 못하면, 그 번호는 무효, 또는 신용 카드 번호와 유사한 것(look-a-like)으로 식별된다. 파일이 처리될 때, 신용 카드 번호 검출 모듈(220)은 기본 포맷 일치의 횟수(number)와 이것들이 무효인 것으로 결정된 횟수를 카운트한다. 다른 기본 포맷 일치의 카운트값에 대한 무효 번호의 카운트값의 비율은 다른 기본 포맷 일치가 실제 유효한 신용 카드 번호일 가능성에 관련있을 수 있다. 이 예의 시스템에서, 이 비율은 스코어 계산 시에 완전한 포맷 일치의 카운트값에 가중치를 부여하기 위해 사용된다.

[0029] 다른 실시예에서, 신용 카드 번호에 대한 기본 포맷 일치는, 스트링이 신용 카드 번호의 표준 포맷의 자릿수 길이와 동일한 길이를 가지는, 각각이 숫자를 나타내는 값을 가지는 연속된 바이트의 시퀀스를 포함할 때에만 선언될 수 있다.

[0030] 마찬가지로, 사회 보장 번호 검출 모듈(230)은, 번호가 있는지 스트링을 검색하고 그 번호를 사회 보장 번호와 연관된 규칙의 세트와 대조한다. 이 규칙들은 사회 보장 번호는 9 자릿수이어야 한다는 명세를 포함한다. 사회 보장 번호 검출 모듈(230)은 숫자를 나타내는 바이트의 스트링을 검색함으로써 시작한다. 스트링에서 검출된 숫자의 수가 9개이면, 기본 포맷 일치에 선언된다. 기본 포맷 일치 숫자 그룹들 사이의 공백과 같이, 번호를 나타내는 숫자들 사이에 다른 문자를 나타내는 바이트가 배치되더라도 발생할 수 있다. 기본 포맷 일치 각각에 대해, 스트링 내의 숫자들에 의해 나타난 번호는 사회 보장 번호와 연관된 규칙의 완전한 세트를 사용하여 테스트된다. 완전한 세트 내의 다른 규칙들은, 예를 들면, 사회 보장 번호의 처음 다섯자리로 표현된 번호에 대한 값의 유효한 서브세트의 명세를 포함할 수 있다. 번호가 어떠한 규칙도 충족시키지 못하면, 그 번호는 무효, 또는 사회 보장 번호와 유사한 것으로 식별된다. 파일이 처리될 때, 사회 보장 번호 검출 모듈(230)은 기본 포맷 일치의 횟수와 이것들이 무효인 것으로 결정된 횟수를 카운트한다. 다른 기본 포맷 일치의 카운트값에 대한 무효 횟수의 카운트값의 비율은 다른 기본 포맷 일치가 실제 유효한 사회 보장 번호일 가능성에 관련있을 수 있다. 이 예의 시스템에서, 이 비율은 스코어 계산 시에 완전한 포맷 일치의 카운트값에 가중치를 부여하기 위해 사용된다.

[0031] 마찬가지로, 전화 번호 검출 모듈(240)은, 번호가 있는지 각 스트링을 검색하고 그 번호를 전화 번호와 연관된 규칙의 세트와 대조한다. 이 규칙들은 전화 번호가 10 자릿수 또는 첫 번째 숫자가 1인 11 자릿수라는 명세를

포함한다. 전화 번호 검출 모듈(240)은 숫자를 나타내는 바이트가 있는지 스트링을 검색함으로써 시작한다. 스트링에서 검출된 숫자의 수가 10개이거나 11개이고 첫 번째 숫자가 1이면, 기본 포맷 일치가 선언된다. 기본 포맷 일치는 지역 코드 숫자 주위의 괄호 또는 숫자 그룹들 사이의 하이픈과 같이, 번호를 나타내는 숫자들 사이에 다른 문자를 나타내는 바이트가 배치되더라도 발생할 수 있다. 기본 포맷 일치 각각에 대해, 스트링 내의 숫자들에 의해 나타난 번호는 전화 번호와 연관된 규칙의 완전한 세트를 사용하여 테스트된다. 완전한 세트 내의 다른 규칙들은, 예를 들면, 유효한 지역 코드 또는 지역 코드 뒤의 첫 번째 숫자가 1 또는 0이 아니어야 한다는 명세를 포함할 수 있다. 번호가 어떠한 규칙도 충족시키지 못하면, 그 번호는 무효, 또는 전화 번호와 유사한 것으로 식별된다. 파일이 처리될 때, 전화 번호 검출 모듈(240)은 기본 포맷 일치의 횟수와 이것들이 무효인 것으로 결정된 횟수를 카운트한다. 다른 기본 포맷 일치의 카운트값에 대한 무효 횟수의 카운트값의 비율은 다른 기본 포맷 일치가 실제 유효한 전화 번호일 가능성에 관련될 수 있다. 이 예의 시스템에서, 이 비율은 스코어 계산 시에 완전한 포맷 일치의 카운트값에 가중치를 부여하기 위해 사용된다.

[0032] 다른 실시예에서, 전화 번호에 대한 기본 포맷 일치는, 다음의 시퀀스 중 하나를 나타낸 값을 가지는 연속된 바이트의 시퀀스를 포함할 때에만 선언될 수 있다:

```

*****
1*****
***_***_***
1-***_***_***
(***)***_***
1(***)***_***

```

[0033]

[0034] 위에서, *은 숫자 0,1,2,3,4,5,6,7,8, 또는 9 중에서 임의의 숫자를 나타내는 와이드 카드(wild card)이다.

[0035]

성명 검출 모듈(250)은 일반 성명(common name)의 리스트에 의거하여 성명과 일치하는 단어가 있는지 각 스트링을 검색한다. 여기서 단어는, 모두가 글자(letter)를 나타내는 값을 가지는 연속하는 바이트의 시퀀스이다. 이러한 일반 성명의 리스트는 정부의 인구조사(census)로부터 얻을 수 있다. 성명의 리스트는 이름과 성으로 나눌 수 있다. 보통 사용되는 단어인 에일리어스(alias, 가명)으로 알려진 어떤 특정한 성명은 거짓 긍정(false positive)의 확률을 낮추기 위해 배제될 수 있다. 예를 들면, 성명의 리스트는 월 및 요일의 성명을 배제하도록 맞춤 제작될 수도 있다. 거짓 긍정을 보상하는 다른 방법은 파일에 나타나는 성명의 평균 빈도를 모니터링하는 것이다. 예를 들면, 미국 인구조사는 각 성명에 대한 빈도수(frequency count)를 제공한다. 흔한 성명보다 흔치 않은 성명이 아주 더 많기 때문에, 대부분의 성명은 평균보다 훨씬 낮은 빈도를 가진다. 성명의 리스트는 대략 평균의 빈도를 가지는 것으로 예측될 수 있다. 몇 개의 성명에 히트(hit)가 발생하는 무작위 문자의 리스트는 훨씬 더 낮은 빈도를 가져야한다. 파일에서 검출된 성명의 수가 예컨대 10개의 성명과 같은, 최소 샘플 사이즈를 초과한 경우, 파일에서의 그러한 성명의 평균 빈도가 그 성명들을 테스트하기 위해 계산된다. 평균 빈도를 임계값과 비교하여 그 성명들이 거짓 긍정일 가능성이 큰지 또는 참 성명인지를 결정할 수 있다.

[0036]

예에서, 성명 검출 모듈(250)은 빈도와 연관된 일반 성명의 리스트로부터 이름과 성이 있는지 각 스트링을 검색한다. 또한 성명 검출 모듈(250)은 동일한 스트링 내에서 성 바로 앞에 이름이 나타나는 경우를 검출하고 그러한 완전한 성명으로서의 출현(occurrence)을 카운트한다. 성명 검출 모듈(250)은 파일에 나타나는 그 모든 성명에 대해 이름의 카운트값, 성의 카운트값, 성명의 카운트값, 및 평균 빈도를 출력한다.

[0037]

거리 주소 검출 모듈(260)은 차례로 간주되는 거리 약어(recognized street abbreviation)가 뒤 따르는 글자로 구성되는 하나 또는 두 단어가 뒤 따르는 번지(number)를 포함하는 일련의 단어가 있는지 각 스트링을 검색한다. 거리 주소 처음의 번지는 첫 번째 숫자가 0(영)이 아닌 연속하는 일련의 숫자로 시작되어야 한다. 이 번지는 거리명의 하나 또는 두 단어에 선행하는 공백 앞의 끝에 선택적인 글자를 가질 수 있다. 거리 주소 검출 모듈(260)은 파일 내에서 검출된 거리 주소 시퀀스의 수의 카운트값을 스코어 결정 모듈(scoring module)에 전달한다.

[0038]

주(state) 및 우편 번호 검출 모듈(270)은 우편 번호 및 간주되는 두 문자의 주 약어(state abbreviation)가 있는지 각 스트링을 검색한다. 또한 주 및 우편 번호 검출 모듈(270)은 유효한 우편 번호가 뒤따르는 주 약어로 구성되는 두 단어 시퀀스의 출현을 카운트한다. 주 및 우편 번호 검출 모듈(270)은 모든 두 글자 단어를 카운트하고 각 두 글자 단어가 미국 우정청(United States Post Office)이 명시한 유효한 주 명칭의 약어인지를

대조한다. 주 및 우편 번호 검출 모듈(270)은 또한 번호(number)가 있는지 각 스트링을 검색하여, 그 번호를 우편 번호와 연관된 규칙의 세트와 대조한다. 이 규칙들은, 우편 번호가 길이 다섯 자릿수인 연속하는 숫자들의 시퀀스 또는 다섯 번째 숫자와 여섯 번째 숫자 사이에 하이픈이 있는 길이 아홉 자릿수의 시퀀스인 명세를 포함한다. 이들 패턴 중 어느 것과 일치하는 바이트의 시퀀스가 검출되는 경우, 기본 포맷 일치가 선언된다. 각 기본 포맷 일치에 대해, 스트링 내에서 숫자(digit)로 표현된 번호(number)들은 우편 번호와 연관된 규칙의 전체 세트를 사용하여 테스트된다. 전체 세트 내의 다른 규칙들에는, 예를 들면, 모든 가능한 다섯 자릿수 번호(five digit number)의 서브세트인 유효한 다섯 자릿수 우편 번호의 명세를 포함할 수 있다. 그 번호가 규칙 중 어느 것도 충족시키지 못하면, 무효 또는 우편 번호와 유사한 것으로 식별된다. 파일이 처리될 때, 주 및 우편 번호 검출 모듈(270)은 기본 포맷 일치의 횟수와 이것들이 무효인 것으로 결정된 횟수를 카운트한다. 다른 기본 포맷 일치의 카운트값에 대한 무효 번호의 카운트값의 비율은, 다른 기본 포맷 일치가 실제 유효한 우편 번호일 가능성에 관련될 수 있다. 이 예의 시스템에서, 이 비율은 스코어 계산 시에 완전한 포맷 일치의 카운트값에 가중치를 부여하기 위해 사용된다. 주 및 우편 번호 검출 모듈(270)은 최종적으로, 유효한 우편 번호가 뒤따르는 공백이 뒤따르는 유효한 주 약어로 구성되는 주 및 우편 번호 시퀀스의 수를 카운트한다. 주 및 우편 번호 검출 모듈(270)은 두 글자 단어의 카운트값, 유효한 주 약어의 카운트값, 우편 번호 포맷 일치의 카운트값, 유효한 우편 번호의 카운트값, 그리고 주 및 우편 번호 시퀀스의 카운트값을 스코어 결정 모듈에 전달한다.

[0039] CID 시스템(100)에 의해 처리된 각 파일은 스코어를 할당받고, 그 스코어에 따라 잠재적으로 비밀 정보를 포함하는 것으로서 플래그가 설정될 수 있다.

[0040] 스코어 결정 모듈(280)은 비밀 정보를 검출하는 모듈들(220, 230, 240, 250, 260, 270)의 출력에 기초하여 파일에 대한 스코어를 계산한다. 이 스코어는 CID 시스템(100)에 저장되거나 출력될 수 있다. 이 스코어는 또한 스코어를 임계값과 비교하여 스코어가 임계값을 초과하는 경우에 플래그를 설정하는 임계 결정 모듈(thresholding module)(290)에도 전달된다.

[0041] 도 2에는 비밀 정보를 검출하는 모듈들이 스트링에 독립적으로 동작하는 것으로 도시되어 있지만, 어떤 경우에는 일부 검출 모듈들 사이에 스트링에 대한 중간 처리 결과를 공유함으로써 더 높은 효율을 달성할 수 있다는 것을 알아야 한다. 예를 들면, 스트링에서 숫자를 나타내는 바이트를 카운트하는 루틴의 결과는 신용 카드 번호 검출 모듈(220), 사회 보장 번호 검출 모듈(230), 및 전화 번호 검출 모듈(240)에 의해 공유될 수 있다. 또한, 많은 검출 모듈들은 검색하고 있는 비밀 정보의 포맷과 일치하기 위해 요구되는 최소 바이트의 수보다 적은 스트링을 무시함으로써 최적화될 수도 있다. 나타낸 모듈 경계(boundary)들은 구현되는 논리의 이해를 전달하기 위한 것이고 개시된 방법을 소프트웨어로 구현하는 코드의 구성에 엄격한 제약을 부과하려는 것은 아니다.

[0042] 스코어를 계산하는 스코어 결정 모듈(280)로 구현될 수 있는 예시적인 스코어 결정 방법(300)은 도 3에 나타나 있다. 스코어 결정 모듈(280)이 실행을 개시한 후(301), 비밀 정보를 검출하는 모듈들(220, 230, 240, 250, 260, 270) 중 하나에 의해 출력된 비밀 정보(Confidential Information, CI) 타입 중 하나에 관한 데이터를 검색한다(310). 그후 스코어 결정 모듈(280)은 CI 타입에 대한 히트 카운트값 및 거짓 긍정 표시자를 추출한다(320).

[0043] 히트 카운트값은 파일 내의 무효인 것으로 결정되지 않은 CI 타입 포맷과 일치하는 횟수이다. 예를 들면, 신용 카드 번호 검출 모듈에 의해 생성된 히트 카운트값은 무효인 것으로 결정되지 않은 기본 포맷 일치의 횟수이다. 히트는 CI 타입과 연관된 규칙의 전체 세트와 완전한 일치이다. CI 타입의 정보를 인코딩한 것이 아니라도 무작위 데이터가 완전히 일치하는 것을 포함할 수 있기 때문에, 히트는 여전히 거짓 긍정일 가능성이 있다. 거짓 긍정 표시자는 연관 히트 카운트값(associate hit count)의 신뢰성을 평가하기 위해 스코어 결정 모듈에 의해 사용되는 기준(metric)이다. 예를 들면, 신용 카드 번호 검출 모듈에 의해 생성된 거짓 긍정 표시자는 무효로 결정된 기본 포맷 일치의 카운트값이다. 이 무효의 기본 포맷 일치는 어느 정도까지는 '근접(near miss)' 또는 '유사한' 신용 카드 번호이고 이들의 존재는 높은 거짓 긍정의 확률을 나타낼 수 있다. 완전 일치와 구별하여, 기본 포맷 일치가 정의되어 있지 않은 CI 타입의 경우, 다른 기준이 거짓 긍정 표시자로서 사용될 수 있다. 예를 들면, 성명 검출 모듈에 의해 생성된 평균 빈도 카운트값은 성명 카운트값들에 대한 거짓 긍정 표시자로 사용될 수 있다.

[0044] 스코어 결정 모듈은 부분적으로 거짓 긍정 표시자에 기초하여 거짓 긍정 가중 계수(false positive weighting factor)를 계산한다(330). 이 예에서, 거짓 긍정 가중 계수는 거짓 긍정 표시자에 역비례한다. 더욱 정확하게 말하면, 거짓 긍정 가중 계수는 아래와 같은 형태를 취한다:

- [0045]
$$W = (H / (H + F))^n$$
- [0046] 위 식에서, H는 히트 카운트값이고, F는 거짓 긍정 표시자이고, n은 보통 1과 5 사이의 정수 지수(integer exponent)이다. 어떤 거짓 긍정 표시자들은 거짓 긍정 가중 계수를 계산하기 위해 다른 함수 형태를 필요로 할 수 있다. 예를 들면, 거짓 긍정 가중 계수는 성명의 평균 빈도 카운트값에 직접 비례한다. 또한 가중 계수가 취할 수 있는 값들의 경계를 정하기 위해 추가적인 단계들을 가지는 것이 유리할 수 있다. 예를 들면, 성명에 대한 가중 계수는 0.5과 1 사이의 값을 갖도록 불연속적으로 경계가 정해질 수 있다.
- [0047] 스코어를 결정할 때 거짓 긍정 표시자를 사용할 수 없는 것으로 간주된 CI 타입이 있을 수 있다. 예의 시스템에서, 거리 주소 검출 모듈은 거짓 긍정 표시자를 생성하지 않는다. 거짓 긍정 표시자가 없는 CI 타입은 거짓 긍정 가중 계수의 계산 단계 330을 건너뛰으로써, 또는 거짓 긍정 가중 계수를 1(unity) 또는 다른 디폴트값으로 등가적으로 설정함으로써 스코어 결정에 영향을 미칠 수 있다.
- [0048] 그후 스코어 결정 모듈이 스코어 결정 시에 고려된 각 CI 타입에 대한 서브스코어를 계산한다(340). 예에서, 서브스코어는 히트 카운트값, 거짓 긍정 가중치 계수, 및 파일 사이즈 표시자의 함수로서 계산된다. 더 정확히 말하면, 서브스코어는 아래와 같은 형태를 취한다:
- [0049]
$$S = W * k * (H * c / N)^{1/2}$$
- [0050] 위 식에, N은 파일 사이즈 표시자이고, k와 n은 서브스코어에서의 인자들을 정규화하기 위해 각 CI 타입에 대해 조정된 상수이다. 특수한 경우에, 서브스코어 계산의 형태는 단순화될 수 있다. 예를 들면, 완전한 성명의 검출은 성명의 서브스코어에 대한 다른 계산을 트리거할 수 있다. 설명한 것과 유사한 스코어 부여 알고리즘을 실행하는 코드의 예는 아래의 샘플 코드 목록 부분에 포함되어 있다.
- [0051] 서브스코어를 계산(340)한 후, 스코어 결정 모듈은 고려되어야 할 CI 타입이 더 있는지를 확인한다.(350) 고려되어야 할 모든 CI 타입에 대해 서브스코어가 계산되지 않은 경우, 스코어 결정 모듈은 다음 CI 타입에 대한 비밀 정보 검출 모듈에서 데이터를 검색하기 위해 루프를 되돌아간다. 모든 서브스코어가 계산되었으면, 그 서브스코어에 기초하여 파일에 대한 합성 스코어(composite score)를 계산(360)하기 위해 진행한다. 나타난 예에서, 합성 스코어는 모든 서브스코어의 합계(sum)이다. 이 서브스코어는 합성 스코어에 가산되기 전에 경계가 정해질 수 있다(bounded).
- [0052] 그후 결과 합성 스코어는, 예를 들면 데이터베이스 서버(170) 또는 병렬 "멀티파일" 시스템(160)과 같은 데이터 스토리지 디바이스의 휘발성 메모리에 기록함으로써 저장(37)될 수 있다. 또한 합성 스코어는 스코어 결정 모듈의 실행을 종료(395)하기 전에 임계 결정 모듈(290)에 전달될 수 있다.
- [0053] 어떤 구현예에서, CID 시스템(100)은 시스템의 사용자에게 의해 제공되는 하나 이상의 파일의 리스트를 처리하도록 구성될 수 있다. 다른 구현예에서, CID 시스템(100)은 파일 시스템 내의 모든 파일을 처리하도록 구성될 수 있다. CID 시스템(100)은 주기적으로 실행하도록 구성될 수 있으며, 어떤 구현예에서는 파일 시스템 내의 파일과 관련된 타임스탬프(time stamp)를 확인하여 최근에 갱신된(예컨대, CID 시스템이 실행되었는 최종 시각과 같이, 주어진 시간 후에 최종 편집된 것) 파일만을 처리하도록 구성될 수 있다.
- [0054] 진술한 비밀 정보 검출 방법은 컴퓨터 상에서 실행하는 소프트웨어를 사용하여 구현될 수 있다. 예를 들어, 이 소프트웨어는, 각각이 하나 이상의 프로세서, 하나 이상의 데이터 스토리지 시스템(휘발성 및 비휘발성 메모리 및/또는 스토리지 요소를 포함), 하나 이상의 입력 디바이스나 포트, 및 하나 이상의 출력 디바이스나 포트를 포함하는 하나 이상의 프로그램된 또는 프로그램 가능한 컴퓨터 시스템(분산형, 클라이언트/서버 또는 그리드와 같은 다양한 아키텍처를 가지는 것일 수 있음) 상에서 실행하는 하나 이상의 컴퓨터 프로그램 내의 절차들을 구성한다. 이 소프트웨어는 예를 들면, 컴퓨터 그래프의 디자인 및 구성과 관련된 다른 서비스들을 제공하는 더 큰 프로그램의 하나 이상의 모듈을 구성할 수 있다. 그래프의 노드 및 요소는 컴퓨터로 판독 가능한 매체에 저장된 데이터 구조 또는 데이터 저장소에 저장된 데이터 모델에 따르는(conforming) 다른 체계화된(organized) 데이터로서 구현될 수 있다.
- [0055] 소프트웨어는 범용 또는 특수용의 프로그램 가능한 컴퓨터에 의해 판독 가능한 CD-ROM과 같은, 저장 매체 상에 제공되거나, 네트워크의 통신 매체를 통해 소프트웨어가 실행될 컴퓨터에 전달(전파 신호로 인코딩)될 수 있다. 모든 기능은 특수 목적의 컴퓨터 상에서 또는 코프로세서(coprocessor)와 같은 특수한 목적의 하드웨어를 사용하여 수행될 수 있다. 이 소프트웨어는, 소프트웨어로 특정된 계산의 상이한 부분들이 상이한 컴퓨터에서 수행되는 분산 방식으로 구현될 수 있다. 이러한 각 컴퓨터 프로그램은 본 명세서에 기술한 절차를 수행하기 위해

컴퓨터 시스템에 의해 스토리지 매체 또는 디바이스가 판독될 때, 컴퓨터를 구성 및 동작 시키기 위해, 범용 또는 특수용의 프로그램 가능한 컴퓨터로 판독 가능한 스토리지 매체 또는 디바이스(예컨대, 고체 메모리나 매체, 또는 자기 또는 광학 매체)에 저장되거나 다운로드되는 것이 바람직하다. 본 발명의 시스템은 또한 컴퓨터 프로그램을 사용하여 구성된, 컴퓨터로 판독 가능한 스토리지 매체로서 구현되는 것으로 간주될 수 있으며, 스토리지 매체는 컴퓨터 시스템을 본 명세서에 기술된 기능들을 구체적이고 미리 정해진 방식으로 동작시키도록 구성된다.

[0056] 이상에서는 본 발명의 많은 실시예를 설명하였다. 그렇지만 본 발명의 사상 및 범위를 벗어 나지 않는 다양한 변경을 가할 수 있음을 알 것이다. 예를 들면, 전술한 단계들 중 일부는 순서에 무관하고, 따라서 설명된 것과 다른 순서로 실행될 수 있다.

[0057] 샘플 코드 리스트

```

out::reformat(in) =

begin

letreal(8) first_name_ratio = if(in.first_name_pct == 0 or in.first_name_count <= 9) 0.5

elseif(in.first_name_pct > 0.04) 1.0 else in.first_name_pct / 0.04;

letreal(8) last_name_ratio = if(in.last_name_pct == 0 or in.last_name_count <= 9) 0.5

elseif(in.last_name_pct > 0.006) 1.0 else in.last_name_pct / 0.006;

letreal(8) credit_card_ratio = if(in.credit_card_count > 0) (double) in.credit_card_count /

(in.credit_card_count + in.non_credit_card_count) else 1;

```

[0058]

```

letreal(8) ssn_ratio = if(in.ssn_count > 0) (double) in.ssn_count / (in.ssn_count +
in.non_ssn_count) else 1;

letreal(8) phone_ratio = if(in.phone_count > 0) (double) in.phone_count / (in.phone_count +
in.non_phone_count) else 1;

letreal(8) zip_ratio = if(in.zip_count > 0) (double) in.zip_count / (in.zip_count + in.non_zip_count)
else 1;

letreal(8) state_ratio = if(in.state_count > 0) (double) in.state_count / (in.state_count +
in.total_two_char_string_length / 2) else 1;

first_name_ratio = first_name_ratio * first_name_ratio;

ssn_ratio = ssn_ratio * ssn_ratio * ssn_ratio * ssn_ratio * ssn_ratio;

phone_ratio = phone_ratio * phone_ratio * phone_ratio * phone_ratio * phone_ratio;

zip_ratio = zip_ratio * zip_ratio * zip_ratio * zip_ratio * zip_ratio;

state_ratio = state_ratio * state_ratio * state_ratio * state_ratio * state_ratio;

out.* :: in.*;

out.name_score :: if(in.first_and_last_count > 0) math_sqrt(in.first_and_last_count) * 1000
else
75000.0 * (math_sqrt((double) in.first_name_count * 6 / in.total_string_length) +
math_sqrt(in.first_name_count * 6 / in.total_string_length)) * first_name_ratio * last_name_ratio;

out.credit_card_score :: if(in.credit_card_count == 0) 0
else if(in.non_credit_card_count == 0) math_sqrt(in.credit_card_count) * 1000
else
(75000.0 * math_sqrt((double) in.credit_card_count * 16 / in.total_string_length)) *
credit_card_ratio;

```

[0059]

```

out.ssn_score :: if(in.ssn_count == 0) 0

else if(in.non_ssn_count == 0) math_sqrt(in.ssn_count) * 1000

else

((75000.0 * math_sqrt((double) in.ssn_count * 9 / in.total_string_length)) *

ssn_ratio);

out.phone_score :: if(in.phone_count == 0) 0

else if(in.non_phone_count == 0) math_sqrt(in.phone_count) * 1000

else

((750000.0 * math_sqrt((double) in.phone_count * 10 / in.total_string_length)) *

phone_ratio);

out.zip_score :: if(in.zip_count == 0) 0

else if(in.non_zip_count == 0) math_sqrt(in.zip_count) * 1000

else

((75000.0 * math_sqrt((double) in.zip_count * 5 / in.total_string_length)) *

zip_ratio);

out.state_score :: if(in.state_count == 0) 0

else if(in.total_two_char_string_length / 2 <= in.state_count) math_sqrt(in.state_count) * 1000

else

((5000000.0 * math_sqrt((double) in.state_count * 2 / in.total_string_length)) *

state_ratio);

end;

out::reformat(in) =

begin

out.overall_score :: (if(in.first_and_last_count > 0) 1000 else min(in.name_score, 2000) / 3) +

(if(in.credit_card_score >= 1000) 1000 else min(in.credit_card_score, 2000) / 3) +

(if(in.ssn_score >= 1000) 1000 else min(in.ssn_score, 2000) / 3) +

(if(in.phone_score >= 1000) 1000 else min(in.phone_score, 2000) / 3) +

(if(in.address_count > 0) 1000 else 0) +

(if(in.state_and_zip_count > 0) 1000 else 0) +

(if(in.zip_score >= 1000 and in.name_score >= 1000) 1000 else min(in.zip_score, 2000) / 3) +

(if(in.state_score >= 1000 and in.name_score >= 1000) 1000 else min(in.state_score, 2000) / 3);

out.username :: file_information(in.filename).username;

out.* :: in.*;

end;

```

[0060]

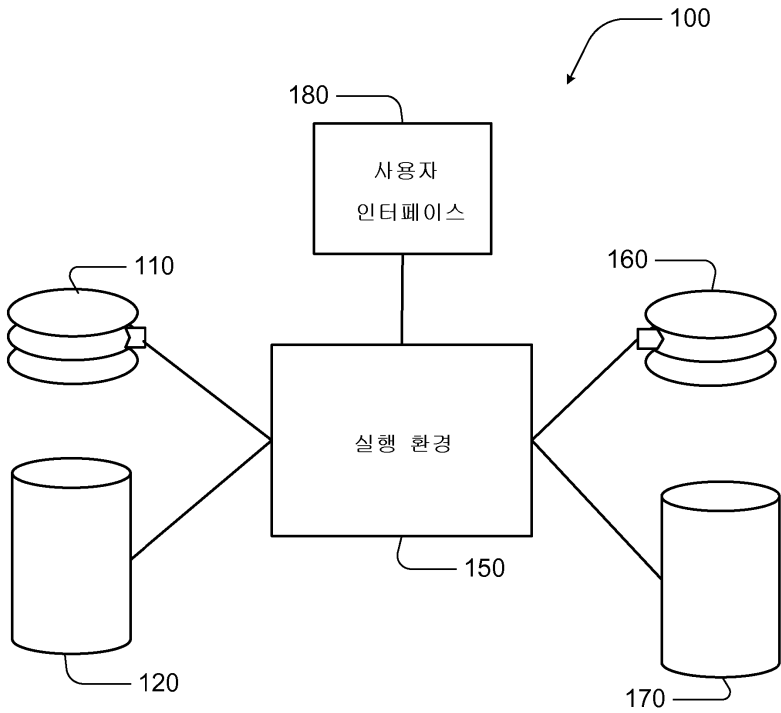
[0061]

[0062] /***** 코드 리스트의 끝 *****/

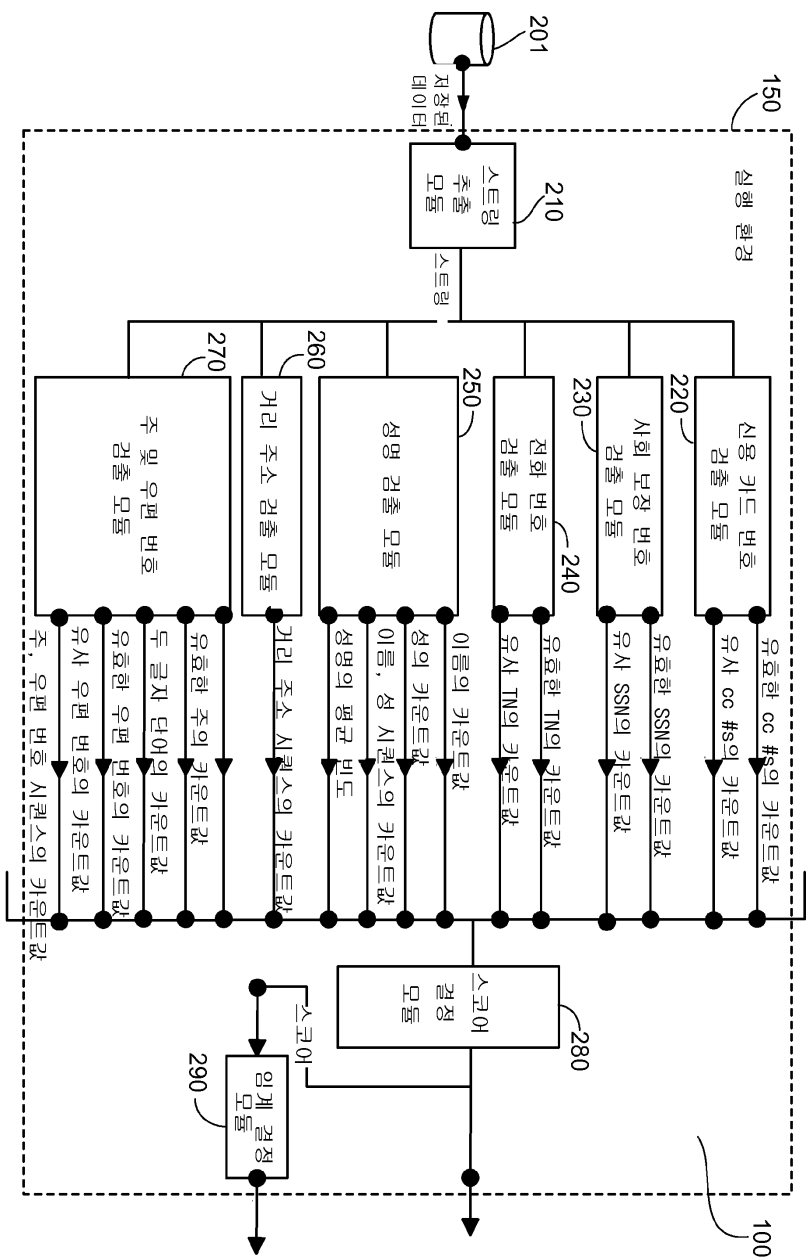
[0063] 다른 실시예들은 아래의 특허청구범위 내이다.

도면

도면1



도면2



도면3

