(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년07월18일
(11) 등록번호 10-1758669
(24) 등록일자 2017년07월11일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(21) 출원번호 10-2012-7019933
(22) 출원일자(국제) 2011년01월13일
심사청구일자 2016년01월12일
(85) 번역문제출일자 2012년07월27일
(65) 공개번호 10-2012-0135218
(43) 공개일자 2012년12월12일
(86) 국제출원번호 PCT/US2011/021108
(87) 국제공개번호 WO 2011/088195
국제공개일자 2011년07월21일
(30) 우선권주장
61/294,663 2010년01월13일 미국(US)
(56) 선행기술조사문헌
JP2006163941 A*
JP2003271656 A*
WO2009017158 A1*
JP2006039871 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
아브 이니티오 테크놀로지 엘엘시
미국 02421 매사추세츠주 렉싱턴 스프링 스트리트 201
(72) 발명자
숀 앤드류
미국 02465 매사츄세츠주 뉴튼 오티스 스트릿 291
(74) 대리인
유미특허법인

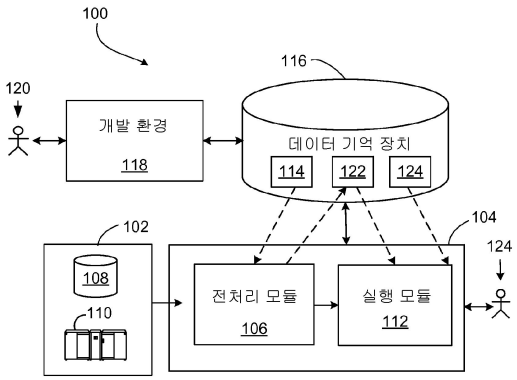
전체 청구항 수 : 총 33 항 심사관 : 홍경아

(54) 발명의 명칭 **매치를 특징화하는 규칙을 사용하는 메타데이터 소스의 매칭**

(57) 요 약

메타데이터를 처리하는 것은, 데이터 기억 시스템(116) 내에, 다수의 소스(102)의 각각에 대한 명세를 기억하는 과정으로서, 명세가 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 각각 포함하는, 명세를 기억하는 과정과, 데이터 기억 시스템에 접속된 데이터 처리 시스템(104) 내에서, 소스로부터 데이터 요소를 처리하는 과정과, 기억된 명세 중의 하나에 기초하여 각각의 소스에 대한 규칙 세트를 생성(204)하는 과정과, 여러 상이한 소스의 데이터 요소를 매칭(206)하고, 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치를 특징화하는 품질 메트릭을, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 기초해서 판정(208)하는 과정을 포함한다

대 표 도



명 세 서

청구범위

청구항 1

데이터 기억 시스템(data storage system) 내에, 다수의 소스(source)의 각각에 대한 명세(specification)를 기억하는 단계로서, 상기 명세는 대응하는 소스의 하나 이상의 데이터 요소(data element)를 식별하는 정보를 각각 포함하는, 상기 명세를 기억하는 단계; 및

상기 데이터 기억 시스템에 접속된 데이터 처리 시스템 내에서, 상기 소스로부터 데이터 요소를 처리하는 단계

를 포함하고,

상기 처리하는 단계는,

제1 소스 내에서 용어 및 관련된 용어 기술(description)을 식별하는 단계;

상기 제1 소스 내의 용어와 가장 유사한 제2 소스 내의 제1 미리 정해진 수의 용어를 식별하고, 상기 제1 소스 내의 용어 기술과 가장 유사한 제2 소스 내의 제2 미리 정해진 수의 용어 기술을 식별하는 단계; 및

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치의 품질을 평가하기 위해 하나 이상의 규칙을 적용하는 단계를 포함하고,

각각의 규칙은 (i) 상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 유사도(similarity) 및 (ii) 상기 제1 소스 내의 용어 기술과 상기 제2 소스 내의 각각의 식별된 용어 기술 간의 유사도를 평가하는 것을 특징으로 하는 방법.

청구항 2

제1항에 있어서,

각각의 규칙은 상기 제1 소스 내의 용어 및 관련된 용어 기술과 상기 제2 소스 내의 각 용어 및 관련된 용어 기술 간의 매치에 등급 레벨을 할당하는, 방법.

청구항 3

제2항에 있어서,

임계 레벨 미만의 등급 레벨을 갖는 매치를 식별하는 단계를 더 포함하는 방법.

청구항 4

삭제

청구항 5

삭제

청구항 6

제1항에 있어서,

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 식별된 용어를 하나 이상의 클래스(class)로 분류하는 단계;

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 각각의 식별된 용어에 대해 하나 이상의 클래스 단어(class word)를 할당하는 단계; 및

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치에 대한 품질을 평가하기 위해, 상기 제1 소스 내의 용어에 대한 하나 이상의 클래스 단어를 상기 제2 소스 내의 각각의 식별된 용어에 대한 하나 이상의 클래스 단어와 비교하는 단계를 더 포함하는 방법.

청구항 7

삭제

청구항 8

제1항에 있어서,

제1 규칙의 하나의 출력에 제1 등급이 할당되고, 제2 규칙의 다른 제2 출력에 제2 등급이 할당되며,

상기 제1 등급은 상기 제2 등급에 비해 상기 매치의 품질이 더 양호한 것을 나타내는, 방법.

청구항 9

삭제

청구항 10

삭제

청구항 11

제1항에 있어서,

상기 규칙 중 하나 이상의 규칙을 생성하기에 충분한 정보를 포함하는 입력을 수신하는 단계를 더 포함하는 방법.

청구항 12

삭제

청구항 13

삭제

청구항 14

제1항에 있어서,

상기 하나 이상의 규칙을 적용하는 단계는, 상기 제1 소스 내의 용어나 관련된 용어 기술에서의 단어 발견 횟수 및 상기 제2 소스 내의 식별된 용어나 용어 기술에서의 단어 발견 횟수의 측정치에 기초하여 품질 메트릭을 결정하는 단계를 포함하는, 방법.

청구항 15

제1항에 있어서,

상기 처리하는 단계는,

상기 제1 또는 제2 소스 내의 제1 빈도(frequency)로 발견되는 용어나 용어 기술에 대해 제1 가중치를 부여하는 단계, 및

상기 제1 또는 제2 소스 내의 제2 빈도로 발견되는 용어나 용어 기술에 대해 제2 가중치를 부여하는 단계를 포함하고,

상기 제1 가중치의 값은 상기 제2 가중치의 값보다 작은 것이며,

상기 하나 이상의 규칙을 적용하는 단계는 상기 용어와 관련되는 가중치를 고려하여 이루어지는, 방법.

청구항 16

제1항에 있어서,

상기 처리하는 단계는,

상기 제1 소스 내의 용어나 용어 기술의 제1 빈도와 상기 제2 소스 내의 용어나 용어 기술의 제2 빈도를 계산하는 단계, 및

상기 제1 빈도와 상기 제2 빈도의 수치 값의 곱(product)에 기초해서 품질 메트릭을 작성하는 단계를 포함하는, 방법.

청구항 17

삭제

청구항 18

삭제

청구항 19

제1항에 있어서,

상기 용어의 제1 미리 정해진 수 또는 상기 용어 기술의 제2 미리 정해진 수는 사용자에 의해 특정되는, 방법.

청구항 20

삭제

청구항 21

삭제

청구항 22

컴퓨터 프로그램이 저장된 컴퓨터로 판독가능한 매체에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

다수의 소스의 각각에 대한, 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 각각 포함하는 명세를 기억하도록 하고,

제1 소스 내에서 용어 및 관련된 용어 기술을 식별하도록 하고,

상기 제1 소스 내의 용어와 가장 유사한 제2 소스 내의 제1 미리 정해진 수의 용어를 식별하고, 상기 제1 소스 내의 용어 기술과 가장 유사한 제2 소스 내의 제2 미리 정해진 수의 용어 기술을 식별하도록 하고,

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치의 품질을 평가하기 위해 하나 이상의 규칙을 적용하도록 하기 위한 명령어를 포함하고,

각각의 규칙은 (i) 상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 유사도 및 (ii) 상기 제1 소스 내의 용어 기술과 상기 제2 소스 내의 각각의 식별된 용어 기술 간의 유사도를 평가하는 것을 특징으로 하는 컴퓨터로 판독가능한 매체.

청구항 23

다수의 소스의 각각에 대한, 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 각각 포함하는 명세를 기억하는 데이터 기억 시스템; 및

상기 데이터 기억 시스템에 접속된 데이터 처리 시스템

을 포함하고,

상기 데이터 처리 시스템은,

제1 소스 내에서 용어 및 관련된 용어 기술을 식별하고,

상기 제1 소스 내의 용어와 가장 유사한 제2 소스 내의 제1 미리 정해진 수의 용어를 식별하고, 상기 제1 소스 내의 용어 기술과 가장 유사한 제2 소스 내의 제2 미리 정해진 수의 용어 기술을 식별하고,

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치의 품질을 평가하기 위해 하나 이상의 규칙을 적용하도록 구성되고,

각각의 규칙은 (i) 상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 유사도 및 (ii) 상기 제1 소스 내의 용어 기술과 상기 제2 소스 내의 각각의 식별된 용어 기술 간의 유사도를 평가하는 것을 특징으로 하는 시스템.

청구항 24

데이터 기억 시스템 내에, 다수의 소스의 각각에 대한 명세를 기억하는 수단으로서, 상기 명세는 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 각각 포함하는, 상기 명세를 기억하기 위한 수단; 및

상기 소스로부터 데이터 요소를 처리하기 위한 수단

을 포함하고,

처리하는 것은,

제1 소스 내에서 용어 및 관련된 용어 기술을 식별하는 것,

상기 제1 소스 내의 용어와 가장 유사한 제2 소스 내의 제1 미리 정해진 수의 용어를 식별하고, 상기 제1 소스 내의 용어 기술과 가장 유사한 제2 소스 내의 제2 미리 정해진 수의 용어 기술을 식별하는 것,

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치의 품질을 평가하기 위해 하나 이상의 규칙을 적용하는 것을 포함하고,

각각의 규칙은 (i) 상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 유사도 및 (ii) 상기 제1 소스 내의 용어 기술과 상기 제2 소스 내의 각각의 식별된 용어 기술 간의 유사도를 평가하는 것을 특징으로 하는 시스템.

청구항 25

제22항에 있어서,

각각의 규칙은 상기 제1 소스 내의 용어 및 관련된 용어 기술과 상기 제2 소스 내의 각 용어 및 관련된 용어 기술 간의 매치에 등급 레벨을 할당하는, 컴퓨터로 판독가능한 매체.

청구항 26

제25항에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

임계 레벨 미만의 등급 레벨을 갖는 매치를 식별하도록 하는 명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 27

제25항에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 식별된 용어를 하나 이상의 클래스(class)로 분류하도록 하고,

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 각각의 식별된 용어에 대해 하나 이상의 클래스 단어(class word)를 할당하도록 하고,

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치에 대한 품질을 평가하기 위해, 상기 제1 소스 내의 용어에 대한 하나 이상의 클래스 단어를 상기 제2 소스 내의 각각의 식별된 용어에 대한 하나 이상의 클래스 단어와 비교하도록 하는 명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 28

제22항에 있어서,

제1 규칙의 하나의 출력에 제1 등급이 할당되고, 제2 규칙의 다른 제2 출력에 제2 등급이 할당되며,

상기 제1 등급은 상기 제2 등급에 비해 상기 매치의 품질이 더 양호한 것을 나타내는, 컴퓨터로 판독가능한 매체.

청구항 29

제22항에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

상기 규칙 중 하나 이상의 규칙을 생성하기에 충분한 정보를 포함하는 입력을 수신하도록 하는 명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 30

제22항에 있어서,

상기 하나 이상의 규칙을 적용하는 것은, 상기 제1 소스 내의 용어나 관련된 용어 기술에서의 단어 발견 횟수 및 상기 제2 소스 내의 식별된 용어나 용어 기술에서의 단어 발견 횟수의 측정치에 기초하여 품질 메트릭을 결정하는 것을 포함하는, 컴퓨터로 판독가능한 매체.

청구항 31

제22항에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

상기 제1 또는 제2 소스 내의 제1 빈도(frequency)로 발견되는 용어나 용어 기술에 대해 제1 가중치를 부여하도록 하고,

상기 제1 또는 제2 소스 내의 제2 빈도로 발견되는 용어나 용어 기술에 대해 제2 가중치를 부여하도록 하는 명령어를 더 포함하고,

상기 제1 가중치의 값은 상기 제2 가중치의 값보다 작은 것이며,

상기 하나 이상의 규칙을 적용하는 것은 상기 용어와 관련되는 가중치를 고려하여 이루어지는, 컴퓨터로 판독가능한 매체.

청구항 32

제22항에 있어서,

상기 컴퓨터 프로그램이, 상기 컴퓨터로 하여금,

상기 제1 소스 내의 용어나 용어 기술의 제1 빈도와 상기 제2 소스 내의 용어나 용어 기술의 제2 빈도를 계산하도록 하고,

상기 제1 빈도와 상기 제2 빈도의 수치 값의 곱(product)에 기초해서 품질 메트릭을 작성하도록 하는 명령어를 더 포함하는, 컴퓨터로 판독가능한 매체.

청구항 33

제22항에 있어서,

상기 용어의 제1 미리 정해진 수 또는 상기 용어 기술의 제2 미리 정해진 수는 사용자에 의해 특정되는, 컴퓨터로 판독가능한 매체.

청구항 34

제23항에 있어서,

각각의 규칙은 상기 제1 소스 내의 용어 및 관련된 용어 기술과 상기 제2 소스 내의 각 용어 및 관련된 용어 기술 간의 매치에 등급 레벨을 할당하는, 시스템.

청구항 35

제34항에 있어서,

상기 데이터 처리 시스템이 임계 레벨 미만의 등급 레벨을 갖는 매치를 식별하도록 구성되는, 시스템.

청구항 36

제23항에 있어서,

상기 데이터 처리 시스템이,

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 식별된 용어를 하나 이상의 클래스로 분류하고,

상기 제1 소스 내의 용어 및 상기 제2 소스 내의 각각의 식별된 용어에 대해 하나 이상의 클래스 단어를 할당하고,

상기 제1 소스 내의 용어와 상기 제2 소스 내의 각각의 식별된 용어 간의 매치에 대한 품질을 평가하기 위해, 상기 제1 소스 내의 용어에 대한 하나 이상의 클래스 단어를 상기 제2 소스 내의 각각의 식별된 용어에 대한 하나 이상의 클래스 단어와 비교하도록 구성되는, 시스템.

청구항 37

제23항에 있어서,

제1 규칙의 하나의 출력에 제1 등급이 할당되고, 제2 규칙의 다른 제2 출력에 제2 등급이 할당되며,

상기 제1 등급은 상기 제2 등급에 비해 상기 매치의 품질이 더 양호한 것을 나타내는, 시스템.

청구항 38

제23항에 있어서,

상기 규칙 중 하나 이상의 규칙을 생성하기에 충분한 정보를 포함하는 입력을 수신하도록 구성된 입력 장치를 더 포함하는 시스템.

청구항 39

제23항에 있어서,

상기 하나 이상의 규칙을 적용하는 것은, 상기 제1 소스 내의 용어나 관련된 용어 기술에서의 단어 발견 횟수 및 상기 제2 소스 내의 식별된 용어나 용어 기술에서의 단어 발견 횟수의 측정치에 기초하여 품질 메트릭을 결정하는 것을 포함하는, 시스템.

청구항 40

제23항에 있어서,

상기 데이터 처리 시스템이,

상기 제1 또는 제2 소스 내의 제1 빈도(frequency)로 발견되는 용어나 용어 기술에 대해 제1 가중치를 부여하고,

상기 제1 또는 제2 소스 내의 제2 빈도로 발견되는 용어나 용어 기술에 대해 제2 가중치를 부여하도록 구성되고,

상기 제1 가중치의 값은 상기 제2 가중치의 값보다 작은 것이며,

상기 하나 이상의 규칙을 적용하는 것은 상기 용어와 관련되는 가중치를 고려하여 이루어지는, 시스템.

청구항 41

제23항에 있어서,

상기 데이터 처리 시스템이,

상기 제1 소스 내의 용어나 용어 기술의 제1 빈도와 상기 제2 소스 내의 용어나 용어 기술의 제2 빈도를 계산하고,

상기 제1 빈도와 상기 제2 빈도의 수치 값의 곱(product)에 기초해서 품질 메트릭을 작성하도록 구성되는, 시스템.

청구항 42

제23항에 있어서,

상기 용어의 제1 미리 정해진 수 또는 상기 용어 기술의 제2 미리 정해진 수는 사용자에 의해 특정되는, 시스템.

청구항 43

데이터 기억 시스템 내에, 다수의 소스의 각각에 대한 명세를 기억하는 단계로서, 상기 명세는 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 각각 포함하는, 상기 명세를 기억하는 단계; 및

상기 데이터 기억 시스템과 통신하는 데이터 처리 시스템 내에서, 상기 다수의 소스 내의 각각의 요소 간의 매치를 발견하도록 소스의 데이터 요소를 처리하는 단계

를 포함하고,

상기 처리하는 단계는,

제1 소스의 적어도 몇몇의 데이터 요소 내의 이름이나 기술로부터의 용어를 공통의 레포지토리(repository)의 표준 속성의 이름이나 기술로부터의 용어에 매칭하는 단계,

갱신된 공통의 레포지토리를 생성하도록 상기 제1 소스의 매칭되지 않은 데이터 요소로부터의 용어를 상기 공통의 레포지토리에 추가하는 단계, 및

제2 소스의 적어도 몇몇의 데이터 요소 내의 이름이나 기술로부터의 용어를 상기 갱신된 공통의 레포지토리의 표준 속성의 이름이나 기술로부터의 용어에 매칭하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 44

제43항에 있어서,

상기 처리하는 단계는,

상기 제2 소스의 매칭되지 않은 데이터 요소로부터의 용어를 상기 갱신된 공통의 레포지토리에 추가하는 단계, 및

상기 제1 및 제2 소스가 아닌 적어도 몇몇의 소스의 각각에 대하여,

상기 소스의 적어도 몇몇의 데이터 요소 내의 이름이나 기술로부터의 용어를 상기 갱신된 공통의 레포지토리의 표준 속성의 이름이나 기술로부터의 용어에 매칭하고,

상기 소스의 매칭되지 않은 데이터 요소로부터의 용어를 상기 갱신된 공통의 레포지토리에 추가하는 단계를 포함하는, 방법.

발명의 설명

기 술 분 야

[0001] 본 설명은 매치를 특징화하기 위한 규칙을 사용하여 메타데이터 소스를 매칭하는 것에 관한 것이다.

배 경 기 술

[0002] 메타데이터 디스커버리(matadata discovery)(메타데이터 스캐닝이라고도 알려져 있음)는 데이터베이스 테이블 또는 스프레드시트의 필드 또는 컬럼의 이름 등과 같은, 데이터세트 내에 출현하는 값을 기술하는 메타데이터를 나타내는 데이터 요소 간의 관계를 발견하는 데에 사용될 수 있다. 어떤 경우, 소정의 데이터세트 내에 출현하는 데이터에 대한 메타데이터는 다양한 여러 소스에 기억된다. 메타데이터 디스커버리 과정 중에, 매치(match)

는 제1 소스 내의 데이터 요소와 제2 소스 내의 데이터 요소 사이에서 발견될 수 있다. 매치는, 예를 들어 테이블 내의 필드에 대한 메타데이터의 설명 및/또는 유사한 필드 명에 대응할 수 있다. 이러한 매치는 매칭 데이터 요소가 각각의 데이터세트 내의 동일 타입의 데이터 값에 대한 메타데이터를 표현하는 것을 나타낼 수 있다. 어떤 경우에, 사용자 전용 또는 사전 기반의 데이터베이스, 예를 들어 WordNet을 포함하는 데이터베이스는 유사한 시맨틱 의미를 갖는 데이터 요소들 간의 매치(예를 들어, "날"(day)과 "날짜"(date), 또는 "성별"(gender)과 "성"(sex)간의 매치)를 판정하는 데에 사용될 수 있다. 메타데이터의 마스터 콜렉션(master collection)("메타데이터 레지스트리"라고도 함)은 발견된 관계에 기초하여 메타데이터를 기억하도록 또는 메타데이터 디스커버리 프로세스에서 발견된 메타데이터에 링크하기 위해 생성 또는 갱신될 수 있다.

발명의 내용

[0003] 하나의 관점으로서, 일반적으로, 방법은 데이터 기억 시스템(data storage system) 내에, 다수의 소스(source)의 각각에 대한 명세(specification)를 기억하는 단계로서, 명세가 대응하는 소스의 하나 이상의 데이터 요소(data element)를 식별하는 정보를 각각 포함하는, 명세를 기억하는 단계; 데이터 기억 시스템에 접속된 데이터 처리 시스템 내에서, 소스로부터 데이터 요소를 처리하는 단계; 기억된 명세 중의 하나에 기초하여 각각의 소스에 대한 규칙 세트를 생성하는 단계; 및 여러 상이한 소스의 데이터 요소를 매칭하고, 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치(match)를 특징화하는 품질 메트릭(quality metric)을, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 기초해서 판정하는 단계를 포함한다.

[0004] 관점은 이하의 특징을 하나 이상 포함할 수 있다.

[0005] 각각의 소스에 대한 규칙 세트는 소정의 매치를 특징화하는 품질 메트릭에 대응하는 하나 이상의 등급(grade)을 작성할 수 있다. 본 방법은 하나 이상의 등급에 대응하는 설명 정보(explanatory information)를 제공하는 단계를 더 포함할 수 있다. 소정의 매치는 제1 데이터 요소 및 제2 데이터 요소에 대응하는 기술(description) 간의 매치를 포함할 수 있으며, 하나 이상의 등급은 소정의 매치를 특징화하는 품질 메트릭에 기초할 수 있다. 소정의 매치는 제1 데이터 요소 및 제2 데이터 요소에 대응하는 이름(name) 간의 매치를 포함할 수 있으며, 하나 이상의 등급은 소정의 매치를 특징화하는 품질 메트릭에 기초할 수 있다.

[0006] 방법은 제1 및 제2 데이터 요소에 출현하는 용어를 하나 이상의 클래스(class)로 분류하는 단계; 제1 및 제2 데이터 요소 내의 각각의 용어(term)에 대해 하나 이상의 클래스 단어(class word)를 할당하는 단계; 제1 및 제2 데이터 요소 내의 용어에 각각 대응하는 하나 이상의 클래스 단어를 비교해서 소정의 매치에 대한 품질 메트릭을 생성하는 단계; 및 소정의 매치를 특징화하는 품질 메트릭에 기초해서 하나 이상의 등급을 할당하는 단계를 더 포함할 수 있다. 소정의 매치를 특징화하는 품질 메트릭은 거리 측정 메트릭(distance measure metric)을 포함할 수 있다. 규칙 세트 중의 제1 규칙의 하나의 출력에 제1 등급이 할당될 수 있고, 규칙 세트 중의 제2 규칙의 다른 출력에 제2 등급이 할당될 수 있으며, 제1 등급은 제2 등급에 비해 소정의 매치를 특징화하는 품질 매트릭이 더 양호할 수 있다.

[0007] 규칙 세트는 제1 및 제2 데이터 요소 내에 출현하는 이름의 유사도(similarity)에 기초한다. 규칙 세트는 제1 및 제2 데이터 요소 내에 출현하는 기술의 유사도에 기초할 수 있다. 방법은 제1 데이터 요소와 제2 데이터 요소 간의 매치의 품질을 정량화하기 위한 규칙 세트 중의 하나 이상의 규칙을 생성하기 위한 입력을 제공할 수 있는 능력을 사용자에게 제공하는 단계를 더 포함할 수 있다. 규칙 세트 중의 각각의 규칙은 트리거 입력에 기초한 트리거 입력 및 출력을 포함할 수 있다. 규칙 세트 중의 각각의 규칙은 규칙 세트 중의 소정의 규칙의 모든 트리거 입력이 유효한 것으로 평가될 때까지 순차적으로 판독될 수 있다. 소정의 매치를 특징화하는 품질 메트릭은 제1 또는 제2 데이터 요소 내에서의 단어 발견 횟수 및 제1 또는 제2 소스로부터 일련의 용어 내에서의 단어 발견 횟수의 측정치에 기초할 수 있다.

[0008] 방법은, 제1 또는 제2 소스 내의 제1 빈도(frequency)로 발견되는 용어에 대해 제1 가중치를 부여하고, 제1 또는 제2 소스 내의 제2 빈도로 발견되는 용어에 대해 제2 가중치를 부여함으로써, 소정의 매치를 특징화하는 품질 메트릭을 계산하는 단계를 더 포함할 수 있으며, 제1 가중치의 값은 제2 가중치의 값보다 작다. 방법은 제1 소스 내의 용어의 제1 빈도와 제2 소스 내의 용어의 제2 빈도를 계산하고, 제1 빈도와 제2 빈도의 수치 값의 곱(product)에 기초해서 품질 메트릭을 작성함으로써, 소정의 매치를 특징화하는 품질 메트릭을 계산하는 단계를 더 포함할 수 있다. 방법은 품질 메트릭을 미리 정해진 상한 및 하한 범위(예를 들어, 0과 1 사이)가 되도록 정규화(normalize)하는 단계를 더 포함할 수 있다.

[0009] 방법은 제2 소스로부터 제1 소스 내의 용어에 대응하는 용어 세트를 생성하고, 용어와 용어 세트의 각각의 매치

를 특징화하는 미리 정해진 품질 메트릭을 갖는 단계를 더 포함할 수 있다. 용어 세트 내의 용어의 수는 사용자에 의해 특정된다. 용어와 용어 세트의 각각의 매치는 용어에 출현하는 매칭 이름에 기초한다. 용어와 용어 세트의 각각의 매치는 용어에 출현하는 매칭 기술에 기초한다.

[0010] 다른 관점으로서, 일반적으로, 컴퓨터 프로그램을 기억하는 컴퓨터로 판독가능한 매체는, 컴퓨터 프로그램이, 컴퓨터로 하여금, 다수의 소스의 각각에 대한, 대응하는 소스의 하나 이상의 데이터 요소(data element)를 식별하는 정보를 각각 포함하는 명세(specification)를 기억하도록 하고, 기억된 명세 중의 하나에 기초하여 각각의 소스에 대한 규칙 세트를 생성하도록 하며, 여러 상이한 소스의 데이터 요소를 매칭하고, 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치(match)를 특징화하는 품질 메트릭(quality metric)을, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 기초해서 판정하도록 하기 위한 명령어를 포함한다.

[0011] 다른 관점으로서, 일반적으로, 시스템은, 다수의 소스의 각각에 대한, 대응하는 소스의 하나 이상의 데이터 요소(data element)를 식별하는 정보를 각각 포함하는 명세(specification)를 기억하는 데이터 기억 시스템; 및 데이터 기억 시스템에 접속되어, 여러 상이한 소스의 데이터 요소를 매칭하고, 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치(match)를 특징화하는 품질 메트릭(quality metric)을, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 기초해서 판정하도록 구성된 데이터 처리 시스템을 포함한다.

[0012] 다른 관점으로서, 일반적으로, 시스템은, 데이터 기억 시스템(data storage system) 내에, 다수의 소스(source)의 각각에 대한 명세(specification)를 기억하는 수단으로서, 명세는 대응하는 소스의 하나 이상의 데이터 요소(data element)를 식별하는 정보를 각각 포함하는, 명세를 기억하기 위한 수단; 및 소스로부터 데이터 요소를 처리하고, 기억된 명세 중의 하나에 기초하여 각각의 소스에 대한 규칙 세트를 생성하며, 여러 상이한 소스의 데이터 요소를 매칭하고, 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치(match)를 특징화하는 품질 메트릭(quality metric)을, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 기초해서 판정하기 위한 수단을 포함한다.

[0013] 관점은 이하의 장점들 중 하나 이상을 포함할 수 있다.

[0014] 일반적으로, 하나 이상의 키 단어와 일부 텍스트(예를 들어, 웹 페이지) 간의 매치를 위한 검색을 할 때에, 검색 프로세스는 소정의 매치가 발생된 이유를 사용자에게 표시할 수 있는데, 예를 들어 텍스트 내에 키 단어의 출현을 하이라이트로(예를 들어, 키 단어를 굵게 표시) 나타냄으로써 표시할 수 있다. 일례로, 본원에 개시된 기술은 문서의 여러 버전의 변경을 식별하기 위해 사용될 수 있다. 또한, 소스 또는 키 용어는 매칭 소스 및 등급 등과 같은 상세를 포함할 수 있는 관계 도식에 의해 타겟 용어에 가시적으로 링크될 수 있다. 동일 타입의 데이터에 대한 메타데이터를 표현할 수 있는 2개의 데이터 요소 간의 매치를 수행할 때에, 매치가 일어난(또는 일어나지 않은) 이유가 각각의 키 단어들 간의 정확한 매치의 존재보다 더 복잡해질 수 있다. 예를 들어, 데이터 요소에 출현하는 용어는 확장 또는 변환(예를 들어, 스테밍(stemming)을 사용하여)되었을 수 있으며, 매칭 용어들 간의 관계가 동의어를 발견하거나 용어를 카테고리("클래스"라고 함)별로 분류하는 것에 기초할 수 있다. 매칭을 수행하는 데에 사용되는 과정은 각각의 매치에 등급을 할당함으로써 매치 품질을 특징화하기 위한 규칙을 사용할 수 있다. 이러한 등급은 매치 품질을 나타내기 위해 매치와 관련시켜서 기억될 수 있다.

[0015] 메타데이터의 다수의 소스가 존재하는 경우, 여러 소스들 간의 차이를 설명함으로써, 매칭 프로세스가 반복됨에 따라 소스가 임의의 횟수 효율적으로 처리될 수 있다. 전처리 과정에 의하면, 매치를 특징화하기 위한 규칙을 정의하는 데에 필요한 정보와 데이터 요소를 번역 및/또는 변환하기 위해 필요한 정보를 제공함으로써, 소스로부터 데이터 요소를 직접 처리할 수 있도록 하는 소스 처리 정보를 생성할 수 있다.

[0016] 본 발명의 다른 특성과 장점은 이하의 상세한 설명과 청구범위로부터 명백할 것이다.

도면의 간단한 설명

[0017] 도 1은 그래프 기반의 계산을 실행하기 위한 시스템의 블록도이다.

도 2는 메타데이터 처리 과정의 예를 나타내는 플로차트이다.

도 3은 자동화 매칭 프로세스의 단계를 나타낸다.

도 4는 자동화 매칭 프로세스의 그래프 기반의 구현의 예이다.

도 5는 도 4의 자동화 매칭 프로세서의 그래프 기반의 구현으로부터의 출력의 예이다.

도 6 내지 도 8은 규칙 및 규칙을 관리하기 위한 인터페이스의 스크린 샷이다.

도 9 내지 도 12는 메타데이터 인터페이스의 예를 나타내는 스크린 샷이다.

발명을 실시하기 위한 구체적인 내용

[0018] 비지니스 분석가(business analyst)는 많은 시스템에서의 데이터 요소의 비지니스 특징의 리스트를 포함하는 다수의 데이터 사전(data dictionary)을 유지할 수 있다. 데이터 사전(또는 메타데이터 레포지토리)는 의미, 다른 데이터와의 관계, 기원, 용례 및 포맷 등과 같은 데이터와 관련된 정보를 저장 장소이다. 이와 같이, 데이터 사전은 용어의 정의의 표준화 및 이들 용어의 사용의 일치성을 용이하게 한다. 일례로, 전사적 데이터 사전(enterprise wide data dictionary)은 기업 내에서 사용되는 데이터에 관한 메타데이터를 캡쳐하기 위해 유지될 수 있다.

[0019] 매칭될 데이터 요소는 하나 이상의 기술적 용어를 사용하는 데이터 요소를 식별하는 이름 부분(name portion)을 가질 수 있으며, 데이터 요소 또는 이러한 요소를 특징화하는 다양한 특성을 기술하는 기술 부분(description portion)을 선택적으로 가질 수 있다. 여러 사전에 포함된 이름 및 이에 대응하는 기술은 다양한 형태가 될 수 있다. 예를 들어, 데이터 사전은 여러 시점에서 그리고 독립적으로 유지될 수 있는 여러 시스템의 부분으로서 개발될 수 있다. 적어도 이러한 이유 때문에, 일반적으로 채택되는 이름 설정 표준이 되지 않을 수 있다. 이와 같이, 본 출원에 개시된 메타데이터 처리 기술의 장점은 여러 데이터 사전에서의 이름 및 기술의 조화이다. 추가로, 데이터 요소의 매치(match)를 정량화하는 품질 메트릭(quality metric) 또는 스코어를 제공함으로써, 자동화 메타데이터 처리(automated metadata processign)에 의해, 비지니스 분석가는 인간의 분석을 필요로 하는 매치의 일부분만 신경 쓰면 된다. 예를 들어, 비지니스 분석가는 매치에 가까운 메트릭에 의해 스코어링되는 매치만 관심을 가져도 된다.

[0020] 도 1은 메타데이터 처리 기술이 사용될 수 있는 데이터 처리 시스템(100)의 예를 나타낸다. 시스템(100)은 다양한 기억 포맷(예를 들어, 데이터베이스 테이블, 스프레드시트 파일, 플랫 텍스트 파일, 또는 메인프레임에 의해 사용되는 네이티브 포맷) 중의 하나로 데이터 및/또는 메타데이터를 기억할 수 있는, 기억 장치 또는 온라인 데이터 스트림에의 접속 등과 같은, 데이터 및/또는 메타데이터의 하나 이상의 소스를 포함할 수 있는 소스(102)를 포함한다. 일례로, 소스는 해당 메타데이터에 의해 기술되는 데이터로부터 독립적으로 메타데이터를 기억한다. 일례로, 메타데이터는 해당 메타데이터에 의해 기술되는 데이터와 동일한 데이터 구조 내에, 또는 예를 들어 링크 또는 포인터를 사용하여 데이터와 관련시켜 기억된다. 일례로, 소스(102)는 단일 마스터 데이터 기억 시스템을 형성하도록 통합되는 다수의 데이터 기억 시스템과 관련된다. 시스템을 통합하는 과정에서, 병합되는 대응 데이터를 기술하는 메타데이터 간의 매치를 판정하는 것이 필요할 수 있다. 예를 들어, 하나의 소스의 고객 리스트로부터 어떤 필드가 다른 소스의 고객 리스트로부터의 필드와 동일한 속성을 표현하는 데이터 값을 기억하는지를 판정하는 것이 필요할 수 있다(예를 들어, 어느 소스로부터의 "사회 보장 번호"(social security #) 필드는 다른 소스로부터의 "SSN" 필드와 동일한 속성임). 데이터 요소 간의 매치는 데이터 기억 시스템을 통합하는 데에 사용될 수 있다. 실행 환경(execution environment)(104)은 전처리 모듈(106)과 실행 모듈(112)을 포함하며, 전처리 모듈(106)은 소스(102)를 판독하고 소스 레지스트리(114)에 기초하여 메타데이터 소스에 대한 소스 처리 정보(122)를 생성한다. 실행 모듈(112)은 소스 처리 정보(122)와 참조 정보(124)에 기초하여 매치를 판정하고 품질을 기록하기 위해 메타데이터 처리를 수행한다. 데이터 기억 시스템(116)은 소스 레지스트리(114), 소스 처리 정보(122) 및 참조 정보(124)를 기억하는데, 이에 대해서는 나중에 상세히 설명한다. 실행 환경(104)은 유닉스(Unix) 운영 체제와 같은 적절한 운영 체제의 제어하에서 하나 이상의 범용 컴퓨터의 호스트가 될 수 있다. 예를 들어, 실행 환경(104)은

[0021] 다수의 중앙 처리 장치(CPU), 로컬(예를 들어, SMP 컴퓨터 등의 멀티프로세서 시스템) 또는 국부적으로 분산된(예를 들어, 클러스터 또는 MPP로서 결합된 다수의 프로세서), 또는 원격 또는 원격에 분산된(예를 들어, LAN 및/또는 WAN을 통해 접속된 다수의 프로세서), 또는 이들의 임의의 조합을 포함하는 다중 노드 병렬 컴퓨팅 환경을 포함할 수 있다. 소스(102)를 제공하는 기억 장치는 실행 환경(104)에 대해 로컬이 될 수 있으며, 실행 환경(104)을 구동하는 컴퓨터에 접속된 기억 매체(예를 들어, 하드 드라이브(108))에 기억되거나, 원격 접속에 의해 실행 환경(104)을 구동하는 컴퓨터와 통신하는 원격 시스템(예를 들어, 메인프레임(110))의 호스트가 되는, 실행 환경(104)에 대해 원격이 될 수 있다.

[0022] 데이터 기억 시스템(116)은 개발 환경(development environment)(118)에 액세스가능하다. 개발 환경(118)에서,

개발자(120)는 전처리 모듈(106)과 실행 모듈(112)을 구성할 수 있다. 일례로, 개발 환경(118)은 정점(vertex) 사이를 지향 링크(작업 요소의 흐름을 나타냄)에 의해 연결한 데이터플로우 그래프로서 애플리케이션을 개발하기 위한 시스템이다. 예를 들어, 이러한 환경은 "Managing Parameters for Graph-Based Applications"란 명칭으로 미국공개번호 2007/0011668호에 상세히 개시되어 있으며, 상기 문헌의 내용을 본원에 참조에 의해 원용한다. 전처리 모듈(106)과 실행 모듈(112)은 소스(102)로부터 입력 데이터의 흐름을 수신하는 데이터플로우 그래프로서 구현된 각각의 모듈과 병렬로 다수의 소스를 처리할 수 있는 능력을 갖도록 구성될 수 있으며, 소스(102) 내의 데이터 요소 간의 가능한 매치의 스트림으로서 출력 데이터의 흐름을 제공한다.

[0023] 전처리 모듈(106)은 소스 레지스트리(114)에 따라 소스로부터의 정보에 기초하여 소스 처리 정보(122)를 준비한다. 소스 레지스트리(114)는 소스에의 액세스 방법을 나타내는 액세스 정보, 소스 내의 데이터 요소의 포맷을 나타내는 포맷 정보, 및 매칭 프로세스 내에 포함되는 소스 내의 특정 데이터 요소의 표시를 특정하는, 처리할 각각의 소스에 대한 명세(specification)를 포함한다. 각각의 명세는, 예를 들어 테이블 내에 열(row)로서 기억될 수 있다. 전처리 모듈(106)은 소스 레지스트리(114)에 의해 식별된 소스로부터 데이터 요소를 판독하고, 소스 처리 정보(122)를 생성한다. 소스 처리 정보(122)는 매칭 프로세스에 사용되는 용어 및 기술을 추출하기 위해 데이터 요소를 번역 및/또는 변환하는 데에 필요한, 소스 레지스트리(114)로부터의 포맷 정보에 추가로 임의의 정보를 포함한다. 예를 들어, 다양한 포맷의 각각을 매칭 프로세스에서 사용될 공통의 포맷으로 변환하기 위한 여러 변환 함수(transformation function)가 기억될 수 있다.

[0024] 소스 처리 정보(122)는 매치를 특징화하기 위한 규칙(rule)을 정의하는 데에 필요한 정보를 포함한다. 등급(grade)을 판정하기 위한 규칙 중의 일부는 데이터 요소의 특징에 따라 달라질 수 있다. 그래서, 각각의 소스는 대응하는 규칙 세트를 가질 수 있으며, 다른 규칙 세트는 소정의 매치에 대한 등급을 판정하는 데에 사용될 수 있다.

[0025] 실행 모듈(112)은 실행 환경(104)에 액세스가능한 데이터 기억 시스템(116) 내에 기억된 참조 정보(124) 및 전처리 모듈(106)에 의해 생성된 소스 처리 정보(122)를 사용한다. 실행 모듈(122)은 데이터 요소로부터 추출된 용어 및 기술로부터 매칭할 단어를 생성하고, 데이터 요소 간의 매치를 제공하기 위해 매칭 프로세스를 수행한다. 매칭 프로세스는, 이하에 상세하게 설명하는 바와 같이, 매치의 품질을 특징화하는 데이터를 기억하는 과정을 포함한다. 일례로, 매칭은 소스 레지스트리(114) 내에 리스트된 각 소스와 데이터 기억 시스템(116) 내에 기억된 표준 메타데이터 레포지토리(canonical metadata repository: CMR) 간에 이루어진다. 예를 들어, CMR은 기업 환경에서 마스터 참조 사전으로서 기능하는 기업 데이터 사전을 나타낼 수 있다. 소스 내의 데이터 요소는 매치를 찾기 위해 CMR 내의 표준 속성과 비교된다.

[0026] 이름이나 기술 또는 소스의 데이터 요소 내에 기억된 다른 메타데이터로부터의 용어와 CMR 내에 나타낸 표준 속성의 이름이나 기술로부터의 용어 간의 매치는 매칭된 데이터 요소가 표준 속성과 동일한 의미를 가질 수 있다는 것을 나타낸다. 일례로, 매치는 표준 속성 이름을 데이터 요소 이름과 매칭하고 표준 속성 기술을 데이터 요소 기술과 매칭한 것의 조합에 기초하여 정해진다.

[0027] 일례로, 매칭은 각각의 데이터 요소들 사이에서 또는 데이터 요소와 표준 속성 간에 매치를 찾기 위해 CMR에 추가로 각각의 소스가 모든 다른 소스와 비교되도록 수행된다. 일례로, 매칭에 의해, 소스와의 이전의 비교로부터 매칭되지 않은 용어를 반복 사이의 CMR에 추가함으로써 소스 간의 비교가 가능하게 된다. 이러한 프로세스에 의해 "올투올"(all-to-all) 처리를 하지 않아도 된다. 예를 들어, 매칭 프로세스는 이하의 시퀀스를 사용하는데, 여기서 CMR(n)은 매칭되지 않은 이전의 소스 비교에서의 데이터 요소의 이름 또는 기술로부터 선택된 용어로 CMR을 갱신하는 n번째 반복이다.

[0028] • 소스1을 CMR(0)과 비교한다

[0029] • CMR(1)을 생성하는 CMR(0)에 매칭되지 않는 모든 소스1 용어를 추가한다

[0030] • 소스2를 CMR(1)과 비교한다

[0031] • CMR(2)을 생성하는 CMR(1)에 매칭되지 않는 모든 소스2 용어를 추가한다

[0032] • 소스3을 CMR(2)와 비교한다

[0033] • 이하 마찬가지

[0034] 실행 모듈(112)에 의해 수행되는 매칭 프로세스의 일례에서, 프로세스는 데이터 요소로부터 추출된 용어를 정규화, 확장 및 클리징해서 표준 형태로 하고, 데이터 요소 내의 메타데이터에 의해 정의된 속성의 이름에 대응하는 용어와 해당 속성의 기술에 대응하는 용어를 식별함으로써 개시한다. 클린징(cleansing)은 소정의 구두점(예를 들어, 밑줄, 대시 기호)을 선택적으로 필터링하는 과정, 격을 변환(예를 들어, 소문자로)하는 과정, 및 여분 공백을 제거하는 과정을 포함할 수 있다.

[0035] 표준 언어학의 "뉴슨스 단어"(nuisance words) 또는 "스톱 워드"(stop words)를 포함하는 미리 정해진 단어(예를 들어, "a, also, and" 등)의 리스트가 이러한 용어로부터 제거될 수 있다. 일례로, 참조 정보(124)는 스톱 워드, 두문자어, 및 가명의 리스트를 포함하는 조사 파일을 포함할 수 있다. 예를 들어, 스톱 워드의 리스트를 포함하는 스톱 워드 조사 파일은 클린징에 도움을 주는 데에 사용될 수 있다. 사용자는 조사 파일을 수정함으로써 이러한 리스트로부터 단어를 추가 또는 제거할 수 있다. 이 프로세스는 약자 및 두문자어를 완전한 단어로 된 구로 확장하는 과정과 이름 또는 기술 내의 용어를 일반적인 가명으로 확장하는 과정을 포함한다. 두문자어 조사 파일을 이러한 프로세스에 도움을 주는 데에 사용될 수 있다. 이와 같이, 사용자는 인터페이스를 통해 두문자어 조사 파일을 수정할 수 있다. 일례로, 인터페이스는 파일에 대한 임의의 변경의 승인 및 통지로 사용자 피드백을 요청하도록 하는 제어를 포함할 수 있다.

[0036] 일례로, 두문자어 조사 파일은 유사한 의미를 가질 수 있는 여러 단어를 지원하기 위해 용어 및 기술 내의 단어에 대한 동의어를 포함할 수 있다. 예를 들어, "기관"을 나타내는 여러 영어 단어, 즉 "agency", "authority", "bureau", "organization" 등은 특정의 용어 또는 기술과 관련해서 단어 "office"와 유사한 의미를 갖는다. 일례로, 이러한 동의어는 용어 및 기술 내의 소정의 단어와 국제적으로 등가인 표현을 포함할 수 있다. 예를 들어, "리터"(liter)는 영어 단어 "litre"와 동의어가 될 수 있다. 또한, 동의어 조사 파일은 여러 데이터 소스 내의 단순히 "address"에 대응하는 하나의 데이터 소스 내의 "address1" 및 "address2" 등의 달리 부르는 단어를 어드레싱하기 위한 지원을 제공할 수 있다. 또한, 용어 및 기술 내의 일부 단어는 단어들 간의 차이를 정규화하기 위한 노력으로 이들의 스템 형태로 변환될 수 있다. 일례로, 이러한 변환은 접미사를 조정함으로써, 활용, 시제 및/또는 복수형을 설명할 수 있다. 예를 들어, "acquisition"은 "acquisit"로 변환될 수 있으며, "parameters"는 "paramet"로 변환될 수 있다. 일례로, 클린징된 단어의 그룹이 이름에 대해 생성되고, 클린징된 단어의 그룹이 기술에 대해 생성된다.

[0037] 이 프로세스는 각각의 속성에 대해 "클래스 단어"(class word)를 판정하는 과정을 포함한다. 클래스 단어는 속성에 의해 기술된 데이터의 일부의 내용 및 역할을 정의하는 단어이다. 클래스 단어의 예로는, amount, code, date, time, date-time, class, description, identifier, image, indicator, name, address, number, quantity, percent, rate, sound, 및 text 등이 있다. 소정의 속성에 대하여 클래스 단어를 판정하기 위해, 속성 이름 내의 용어가 오른쪽에서 왼쪽으로 스캐닝되어 미리 정해진 일련의 클래스 단어 중의 하나에 대한 제1 매치를 식별할 수 있다. 예를 들어, 속성 이름 "start date"에 대응하는 클래스 단어는 "date"이다. 일부 클래스 단어는, 반드시 판정된 클래스 단어에 대한 매치(예를 들어, 속성 이름 "title"은, 속성 기술에 출현하는 용어에 따라, 클래스 단어 "text", "indicator" 또는 "name"에 대응할 수 있음)를 필요로 하지 않고도, 이름 및/또는 기술에 출현하는 단어에 기초하여 정해진다.

[0038] 소스 내의 용어 및 기술과 CMR 내의 용어 및 기술 간의 유사성의 계산은, 이하에 설명하는 수정된 TF-IDF 프로세스에 의해 수행될 수 있다. "TF-IDF"(Term Frequency-Inverse Document Frequency) 가중치는 속성 이름 또는 기술 내에 출현하는 용어 내의 단어가 소정의 데이터 요소에 대해 그리고 데이터 요소의 소스에 대해 얼마나 중요한지를 평가하기 위해 사용되는 통계적 측정치이다. 단어의 중요도는 데이터 요소 내에 출현하는 단어(예를 들어, 이름 및 기술을 포함)의 횟수에 비례해서 증가한다. 그러나, 단어의 중요도는 CMR 내에 표현된 속성 내의 단어의 빈도에 의해 오프셋된다.

[0039] TF-IDF 가중치는 CMR 내에 매우 빈번하게 나타나는 단어의 가중치를 감소시키고 거의 나타나지 않는 단어의 가중치를 증가시킨다. 예를 들어, 데이터 사전 용어 내의 상용어인 단어 "code"를 고려한다. 단어 "code"가 소스 및 타겟 용어에서 나타난다면, "code"가 일반적인 문자열이기 때문에, 출현 간의 매치는 적절하게 설명되지 않을 것이다. 그러나, "disputed"라는 용어가 소스 및 타겟 용어에 모두 포함되어 있으면, 용어들 간의 매치가 더 잘 설명되고, 따라서 이들 용어에 있는 "disputed"라는 단어는 2개의 용어 간의 매치를 용이하게 한다.

[0040] 일련의 문서 D(예를 들어, 소스 내의 일련의 데이터 요소를 나타냄)의 "document" d(예를 들어, 대표적인 용어가 들어간 데이터 요소의 적어도 일부를 나타냄)에 대한 가중치 벡터는,

[0041] $\mathbf{v}_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$ 이다.

$$w_{t,d} = \mathrm{tf}_t \cdot \log \frac{|D|}{|\{t \in d\}|}$$

[0042] 여기서, 이고,

[0043] tf_t 는 문서 d 내의 용어 t의 용어 빈도(로컬 파라미터)이고,

[0044] $\log \frac{|D|}{|\{t \in d\}|}$ 는 역 문서 빈도(글로벌 파라미터)이다.

[0045] $|D|$ 는 문서 세트 내의 문서의 총 개수이고, $|\{t \in d\}|$ 는 용어 t를 포함하는 문서의 개수이다.

[0046] 일례로, 속성 이름 및 기술은 8개의 단어를 포함하고, 단어 "branch"는 2번 출현한다. "branch"에 대한 용어 빈도(TF)는 0.25(8번에 2 단어)이다. CMR에서, 3,300개 이하의 속성이 있으며, "branch"는 이들 중 12번 출현한다. 이어서, 역 문서 빈도(IDF: inverse document frequency)는 ln(3,300/12)=5.61로 계산된다. TF-IDF 가중치는 이들 수량의 곱, 즉 0.25*5.61=1.4이다. 다른 예에서, 속성 이름 및 기술은 8개의 단어를 포함하며, "code"란 단어는 1번 출현한다. "code"에 대한 TF는 0.125(8번 중 1번)이다. CMR에는, 3,300개 이하의 속성이 있으며, "code"는 이들 중 900 내에 출현한다. 이어서, IDF는 in(3,300/900)=1.99로서 계산된다. TF-IDF 가중치는 이들 수량의 곱, 0.125*1.99=0.16이다. 따라서, 이들 예에서, 1.4의 가중치를 갖는 단어 "branch"는 가중치 0.16을 가진 단어 "code"보다 중요할 가능성이 높다.

[0047] 일례로, 용어와 기술 간의 유사도는 정규화되는 절대 수치가 될 수 있기 때문에, 예를 들어 0 내지 1의 범위가 된다. 이와 같이, 각각의 소스 용어에 대하여,

[0048] 수정된 TF-IDF 방식에 기초한 매칭 계산의 결과는 속성의 대응하는 소스 기술에 가장 잘 매치하는 N개의 CMR 및 속성의 이름 내의 소스 용어와 가장 잘 매치하는 N개의 CMR 용어의 세트가 될 수 있다. 개수 N은 매칭 시스템에 대한 입력 파라미터가 될 수 있다. 일례로, N의 값으로서 3이 사용될 수 있다.

[0049] 매칭 프로세스는 가장 높은 가중치를 가진 데이터 요소의 단어에 매치하도록 매치를 수행할 때의 단어의 TF-IDF 가중치를 고려할 수 있다. 이러한 매칭 프로세스는 단어가 CMR의 속성으로부터 추출된 이름 또는 기술 매치 단어 내의 용어로부터 추출된 경우를 판정하기 위한 다양한 매칭 기술 중의 임의의 것을 사용할 수 있다. 예를 들어, "MANAGING AN ARCHIVE FOR APPROXIMATE STRING MATCHING"란 명칭의 미국 공개 번호 2009/0182728에는 적절한 스트림 매칭을 위한 기술을 설명하고 있으며, 그 전체 내용을 본원에 참조에 의해 원용한다.

[0050] 매칭 프로세스의 출력은 이들 데이터 요소가 매치되는 CMR 내의 각각의 속성과 관련된 데이터 요소의 리스트를 포함한다. 일례로, 매치는 이름과 기술 내의 매칭 단어에 대응한다. 출력은 매칭되는 이름과 기술 내의 단어를 임의로 포함할 수 있으며, 이름이나 기술이 매칭되지 않는 데이터 요소의 리스트를 포함할 수 있다.

[0051] 각각의 매치와 관련해서, 실행 모듈(112)은 매치의 품질을 특징화하는 등급을 포함하는 설명 정보, 등급이 취득된 방식(예를 들어, 규칙 실행)의 설명, 및 매치가 생긴 이유에 대한 설명을 기억할 수 있다. 등급은 데이터 요소 이름에 대해 클린징된 단어와 표준 속성 이름에 대해 클린징된 단어 간의 "이름 매치", 데이터 요소 기술에 대해 클린징된 단어와 표준 속성 기술에 대해 클린징된 단어 간의 "기술 매치", 및 데이터 요소와 관련된 클래스 단어와 표준 속성과 관련된 클래스 단어 간의 "클래스 단어 매치"에 기초하여 판정될 수 있다. 이러한 매치에 대한 메트릭은 매치가 얼마나 유사한가(예를 들어, 거리 측정과 관련해서)를 나타낼 수 있다.

[0052] 예를 들어, 등급 "AA"는 정확한 이름 매치(예를 들어, 적어도 하나의 클린징된 단어가 정확하게 매치), 높은 기술 매치(예를 들어, 클린징된 단어 간의 매치에 대해 매트릭 > 75%), 및 동일한 클래스 단어가 있었는지를 판정하는 규칙에 기초하여 할당될 수 있다. 등급 "AB"는 높은 이름 매치(예를 들어, 클린징된 단어 간의 매치에 대해 메트릭 > 95%), 높은 기술 매치(예를 들어, 클린징된 단어 간의 매치에 대해 메트릭 > 70%), 및 동일한 클래스 단어가 있었는지에 대해 판정하는 규칙에 기초하여 할당될 수 있다. 클래스 단어가 동일하지 않으면, 등급은 일반적으로 더 낮게 설정되는데(예를 들어, "DA"급 이하), 데이터 요소에 의해 기술된 데이터가 나타내는 것에

서의 의미 있는 차이가 있을 가능성이 있기 때문이다.

[0053] 실행 환경(104)은 사용자(124)가 관련 등급 등의 정보에 기초하여 매치를 리뷰 및 허가하기 위해 매칭 프로세스의 출력과 상호작용하도록 하는 사용자 인터페이스를 포함한다. 이 사용자 인터페이스는 표준 속성 또는 다른 데이터 요소 내의 용어에 매칭되는 이름 및/또는 기술 내의 용어를 포함하는 데이터 요소의 리스트를 포함할 수 있으며, 이 리스트는 소스 내의 원본 데이터 요소 또는 원본 데이터 요소의 복제물에의 링크를 포함할 수 있다. 일례로, 매치는 사용자 입력을 필요로 하지 않고도, 임계치에 대한 등급의 비교에 기초하여 허가된다. 매칭되지 않은 데이터 요소는 사용자 입력에 기초하여 참조 정보(124)를 갱신하는 데에 리뷰 및 사용될 수 있다. 예를 들어, 사용자는 두문자어의 확장을 포함하는 참조 또는 동의어를 포함하는 참조에 추가하기 위한 확장되지 않은 두문자어 또는 기술의 용어를 리뷰할 수 있다.

[0054] 도 2는 소스(102)로부터 메타데이터를 전처리하고 메타데이터에 대해 매칭을 실행하기 위한 과정(200)에 대한 플로차트를 나타낸다. 이 과정(200)은 다수의 소스(102)의 각각에 대한 명세를 데이터 기억 시스템(116)에 기억하는 과정(202)을 포함하는데, 각각의 명세는 대응하는 소스의 하나 이상의 데이터 요소를 식별하는 정보를 포함한다. 이 과정(200)은 소스로부터의 데이터 요소를, 데이터 기억 시스템(116)에 접속된 실행 환경(104)을 제공하는 데이터 처리 시스템 내에서 처리하는 과정을 포함한다. 이 처리 과정은 기억된 명세 중의 대응하는 하나에 기초하여 각 소스에 대한 규칙 세트를 생성하는 과정(204)과, 여러 소스의 데이터 요소를 매칭하는 과정(206)과, 제1 소스에 대해 생성된 규칙 세트와 제2 소스에 대해 생성된 규칙 세트에 따라 제1 소스의 제1 데이터 요소와 제2 소스의 제2 데이터 요소 간의 소정의 매치를 특징화하는 품질 메트릭(예를 들어, 등급)을 판정하는 과정(208)을 포함한다. 소스를 처리한 후, 판정된 매치를 식별하는 결과를 저장한다(210). 소스가 추가됨에 따라, 이 과정(200)은 반복해서 추가의 소스를 처리할 수 있다.

[0055] 도 3은 소스(예를 들어, 소스 데이터 사전)가 CMR(예를 들어, 기업 사전)에 대해 매치되는 데이터 처리 시스템(예를 들어, 도 1의 시스템(100))에 의해 수행되는 자동화 매칭 프로세스의 단계를 나타낸다. 예를 들어, CMR은 일정 기간 동안 다양한 소스로부터 컴파일될 수 있다.

[0056] 준비 단계(310) 동안, 소스 데이터 사전 내의 정보는 상기 설명한 메타데이터 처리 기술에 호환가능한 포맷으로 변환될 수 있다. 예를 들어, 전처리 모듈(106, 도 1 참조)은 이 단계에서 소스 데이터 구조를 공통의 레코드 포맷에 매핑하는 데에 사용된다.

[0057] 준비 단계(310)의 예에서, 시스템은 특정 소스에 대한 데이터 구조를 생성하고 대응하는 소스 데이터 사전을 등록 형태로 등록하기 위해 사용자로부터의 입력을 허가할 수 있다. 이어서, 등록 형태가 판독될 수 있으며, 소스 데이터 사전은 매칭 시스템과 호환가능한 포맷으로 변환될 수 있다. 예를 들어, 그래프 기반의 시스템에서, "메타데이터 생성"(Generate Metadata) 데이터플로우 그래프는 소스 데이터 사전을 로딩하기 위해 실행될 수 있다. 그래프는 등록 형태를 판독하고 매칭 프로세스에서 사용되는 메타데이터를 생성할 수 있다. 일례로, 그래프는 추가의 데이터플로우 그래프를 설정하기 위한 파라미터 세트와 소스 데이터 사전에 대응하는 비지니스 용어, 기술, 두문자오 및 링크를 설정하기 위한 규칙 엔진을 설정하기 위한 규칙 파일을 생성할 수 있다. 메타데이터 및 파라미터 세트와 규칙 파일이 생성되면, 매칭 데이터플로우 그래프가 실행될 수 있다. 일례로, 하나 이상의 소스 사전이 "Generate graph"를 통해 동시에 실행될 수 있다.

[0058] 파싱 단계(320)에서, 소스 사전 내의 용어는 개별 단어를 추출하도록 처리될 수 있다. 상기 클린징 프로세스와 관련해서 설명한 바와 같이, 구문적 의미가 없는 구두점이 이 용어로부터 제거될 수 있다. 일례로, "$", "%" 등과 같은 특정의 문자가 용어 내에 남겨질 수 있는데, 이들은 구문적 의미를 가질 수 있기 때문이다. 일례로, 이들 용어는 상기 설명한 바와 같은 클래스 단어로 분류될 수 있다.

[0059] 표준화 단계(330)에서, 용어 및 기술의 변동성이 감소될 수 있다. 이 단계에서, 용어 및 기술은 스톱 단어를 제거하고, 약어를 확장하며, 가명을 매핑하도록 추가로 클린징될 수 있다. 승인 단계(340)에서, 소스 용어는 하나 이상의 매칭 기술을 사용하여 표준 용어와 매칭될 수 있다. 예를 들어, 상기 설명한 것과 같은 TF-IDF 가중치를 사용해서 용어 또는 기술 내의 단어가 데이터 요소의 소스에 대해 그리고 소정의 데이터 요소에 대해 얼마나 중요한지를 평가할 수 있다. 일례로, "퍼지 매칭"(fuzzy matching) 기술을 사용해서 매칭 프로세스를 수행할 수 있다(예를 들어, 미국 공개 2009/0182728에 설명되어 있음, 본원에 참조에 의해 원용함).

[0060] 가중치가 할당되었으면, 하나 이상의 사용자가 개발한 규칙을 사용해서 매치 단계(350) 동안 매치를 등급화할 수 있다. 예를 들어, 매치는 상기 설명한 바와 같이 등급화될 수 있다. 일례로, "A", "B", "C", 또는 "F" 등의 등급이 매치의 품질에 따라 매치에 할당될 수 있는데, "A"는 가장 높은 품질의 매치를 특정하는 등급이 될 수

있으며, "F"는 가장 불량한 품질의 매치를 특정하는 등급이 될 수 있다. 마지막으로, 스코어 단계(360)에서, 사용자는 매치에 할당된 등급에 기초해서 매치를 리뷰 및 허가할 수 있다. 일례로, 사용자는 CMR에 대한 새로운 표준 용어를 제안할 수 있다.

[0061] 도 4는 상기 설명한 매칭 기술을 구현하기 위한 그래프 기반의 방식의 예를 나타낸다. 매칭 그래프(400)는 소스 데이터가 매칭을 위한 준비가 되면, 상기 설명한 "Generate Metadata" 그래프에 의해 호출될 수 있다. 이와 같이, 판독된 타겟 성분(402)과 판독된 소스 성분(404)은 대응하는 타겟 및 소스 파일(406, 408)을 판독함으로써 매칭 프로세스를 시작한다. 타겟 파일(406)은 CMR로부터 기술 및 CMR 용어를 포함할 수 있다. 이어서, 매핑 성분(410, 412)은 타겟 및 소스 파일 내의 용어 및 기술에 대하여 타겟 및 소스 전용의 매핑 프로세스를 수행할 수 있다. 예를 들어, 소정 용어 및 기술의 다수의 예와 변형예가 매핑 성분(410, 412)에 의해 함께 매핑될 수 있다. 이와 같이, 이 프로세스는 소스 용어가 각각의 매치에 대한 매치 스코어를 다수의 타겟에 대해 매칭되도록 함으로써, 워크플로우를 사용하는 사용자가 "최적의" 매치를 판정하기 위한 프로세스를 지원할 수 있도록 한다.

[0062] 이어서, 상기 구체적으로 설명한 분류 프로세스를 사용하면, 분류 성분(414, 416)은 타겟 및 소스 파일(406, 408) 내의 용어 및 기술에 대한 클래스 단어를 판정할 수 있다. 일례로, 스트링 텍스트를 사용하는 매칭 용어가 계산적으로 느리게 될 수 있다. 이와 같이, 텍스트 용어는 수치 키(numeric key)로 변환, 즉 토큰화되고 매칭 프로세스의 속도를 크게 높일 수 있다. 예를 들어, 성분(415)은 이러한 변환을 소스 및 타겟 용어에 대해 수행할 수 있다. 매칭의 결론으로서, 키는 다시 원본 텍스트 용어로 복호화될 수 있다. 일단 소스 및 타겟 용어 및 기술이 매핑 및 표준화되면, 매칭 서브그래프(418)는 타겟 내의 기술 및 용어 내의 단어와 소스 내의 기술 및 용어 내의 단어의 매칭을 수행한다. 소스 및 타겟 내에서 매치하는 단어에 대해, 매칭 서브그래프(418)는 매치의 가장 근접도를 나타내는 수반하는 품질 메트릭에 의해 용어 또는 기술을 되돌릴 수 있다. 결합 성분(420)은 특정 용어 또는 기술의 원본 소스로부터 도출된 소스 속성 이름을 사용하여 결합을 수행할 수 있으며, 그 결과를 매칭된 출력(422)으로서 출력할 수 있다.

[0063] 도 4의 그래프(400)의 출력(500)을 도 5에 나타낸다. 도시한 바와 같이, 소스 용어 "milestone identifier"(마이스톤 식별자)(502)는 적어도 3개의 CMR 용어(또는 "최적의 매칭"의 임의의 사용가 지정한 수), 즉, "milestone name"(마일스톤 이름), "milestone identifier"(마일스톤 식별자), 및 "milestone date"(마일스톤 날짜)(504)에 대응한다. 품질 메트릭(506)은 CMR 용어의 각각으로 소스 용어의 유사도의 정도를 정량화한다. 예를 들어, 출력에서의 제2 항목 "milestone identifier"(마일스톤 식별자)에 대한 유사도 측정치는 1이며, 이 값은 완벽한 매칭을 나타낸다. 일례로, 용어 "milestone identifier"(502)의 최적의 3개의 용어 매칭은 최적의 3개의 기술 매치와 결합될 수 있으며, 9개의 용어/기술 통신이 소스 및 CMR 용어 간의 최적의 매치를 결정하기 위한 비지니스 규칙 세트로 전송될 수 있다.

[0064] 일례로, 비지니스 규칙은 상기 설명한 것과 같은 소스 및 CMR 용어에 대한 계산용 클래스 단어뿐만 아니라 용어 이름 및 기술 매치의 유사도에 기초할 수 있다. 일례로, 비지니스 규칙의 출력은 해당 매치에 대한 문자 등급일 뿐만 아니라 최적의 매치이다. 사용자는 문자 등급에 대응하도록 매치의 미리 정해진 품질을 설정할 수 있다. 또한, 사용자는 허가된 등급의 미리 정해진 범위를 특정할 수 있다. 예를 들어, 사용자는 등급 A에서 BC(또는 B 마이너스)에 대응하는 매치 품질만을 허가할 수 있다.

[0065] 도 6은 매칭 처리에서 사용되는 비지니스 규칙(600)의 예를 나타낸다. 트리거(602)는 비지니스 규칙(600)에 대한 입력으로서 작용하고, 도시된 바와 같이 대응하는 출력(604)을 생성한다. 소스 및 CMR 용어 간의 이름의 유사도 등의 유사도 측정치는 0 내지 1의 범위를 갖는 수치 값으로서 정량화될 수 있다. 이와 같이, 제1 비지니스 규칙(606)은 다음과 같이 번역될 수 있다. 소스 및 CMR 용어 간의 이름의 유사도가 0.95보다 크면, 소스 및 CMR 기술 간의 기술의 유사도는 0.70보다 크고, 2개의 용어의 이름은 동일하며, 2개의 용어에 대응하는 클래스 단어는 동일하게 됨으로써, 2개의 용어 간의 매치 등급은 "AA"(또는 "A 플러스")가 된다. 일례로, 상기 비지니스 규칙 중의 임의의 트리거(602)가 잘못 평가하게 되면, 제2 비지니스 규칙(608)이 판독되며,이것은 다음과 같은 상태를 나타낸다. 소스 및 CMR 용어 간의 이름의 유사도가 0.95보다 크면, 소스 및 CMR 기술 간의 기술의 유사도는 0.70보다 크고, 2개의 용어의 이름은 동일하지 않고, 2개의 용어에 대응하는 클래스 단어는 동일하게 됨으로써, 2개의 용어 간의 매치 등급은 "AB"(또는 "A 마이너스")가 된다. 일례로, 비지니스 규칙(600)은 모든 입력 트리거가 유효하다고 평가되는 비지니스 규칙이 판독될 때까지 하나씩 판독된다. 사용자는 비지니스 규칙(600)에 의해 트리거되는 최소 등급을 정의할 수 있다.

[0066] 일례로, 분석가는 테스트 데이터를 사용하여 비지니스 규칙(600)의 테스트 실행을 수행할 수 있다. 도 7은 테스

트 실행의 예를 나타내는 스크린 샷(700)이다. 나타낸 바와 같이, 테스트 데이터 항목(702)의 각각에 대하여, 매치 등급(704)이 생성되고 분석가에게 표시된다. 도한, 유사도 스코어(706) 및 클래스 단어 매치(708)가 각각의 항목(702)에 대응해서 표시된다. 분석가는 하이라이트된 테스트 항목(710)을 선택함으로써, 항목(710)에 대한 많은 정보를 볼 수 있게 된다.

[0067] 도 8은 특정의 테스트 데이터 항목(예를 들어, 도 7의 항목(702))에 대해 어떤 규칙이 실행되었는지를 정확하게 나타내는 스크린 샷(800)의 예이다. 또한, 분석가는 비지니스 규칙(예를 들어, 도 6의 규칙(600) 중의 하나)가 몇 번이나 실행되었는지에 대한 정보를 볼 수 있다. 도시된 바와 같이, 일례로, 하나 이상의 그래픽 버튼(802)을 사용하여 버튼(802)에 대응하는 트리거가 유효로 평가되었는지 여부를 나타낼 수 있다. 규칙 5에 대응하는 버튼이 모두 눌러져 있으면, 규칙 5는 특정의 테스트 데이터에 대해 실행된 것을 나타낸다. 또한, 각각의 규칙이 실행된 횟수가 표시될 수 있다. 예를 들어, 규칙 1은 77번 실행되었고 규칙 5는 303번 실행되었다. 신속한 "반복적 테스트, 수정 및 복귀"라고 부르는 이러한 방식은 매칭 규칙을 최적화하는 데에 사용될 수 있다. 분석가는 부적절하게 매칭된 항목이 있는지를 체크하기 위해 이러한 인터페이스를 사용할 수 있다.

[0068] 도 9 내지 도 12는 사용자에게 메타데이터 정보를 표시하기 위한 메타데이터 인터페이스(900)의 예를 나타내는 스크린 샷이다. 일례로, 매칭 프로세스의 결과(예를 들어, 도 4의 매칭된 결과(422))가 인터페이스(900)에 포함될 수 있다. 도 6에 나타낸 바와 같이, 인터페이스(900)는 메타데이터 레포지토리를 검색하기 위한 텍스트-필드(902)를 제공할 수 있다. 용어는 계층적 그룹(예를 들어, "Business"(904)) 및 차일드 그룹(예를 들어, "Baseline"(906))으로서 저장될 수 있다.

[0069] 도 10은 "Baseline"(906)에 속하는 용어 "허가 날짜"(acceptance date)의 상세한 내용을 표시하는 인터페이스(900)를 나타낸다. 일례로, 사용자는 용어 "허가 레이트"(acceptance rate)(1002)에 대해 오른쪽 클릭을 할 수 있으며, 용어 "허가 레이트"(acceptance rate)(1002)에 관한 관계를 체크하도록 요청할 수 있다.

[0070] 도 11은 Baseline 용어 "허가 레이트"(acceptance rate)(1002)에 대한 매치의 도식적 표현(1102)을 나타낸다. 일례로, 매치의 소스에 관한 정보가 사용자에게 표시될 수 있다.

[0071] 도 12를 참조하면, 일례로, 매치에 대한 표 형식의 화면을 사용자가 이용할 수 있다. 인터페이스(900)는 승인된 매치만 표시되도록 구성될 수 있다. 사용자는 용어에 대한 펜딩 및/또는 거부된 매치를 포함하는 매치를 리뷰하도록 "승인 워크플로우"(approval workflow) 탭(1202)을 사용할 수 있다.

[0072] 상기 설명한 메타데이터 처리 방식은 컴퓨터에서 실행되는 소프트웨어를 사용해서 구현될 수 있다. 일례로, 이러한 프로세스는 매우 짧은 실행 기간 내에 많은 사전에 대한 매칭 프로세스를 자동화할 수 있다. 예를 들어, 소프트웨어는 하나 이상의 프로그램된 또는 프로그램가능한 컴퓨터 시스템(분산형, 클라이언트/서버형, 또는 그리드 형 등의 다양한 구성이 가능함), 하나 이상의 입력 장치 또는 포트, 및 하나 이상의 출력 장치 및 포트에서 실행되는 하나 이상의 컴퓨터 프로그램 내에 절차를 형성한다. 소프트웨어는 프로그램 중의 하나 이상의 모듈을 형성할 수 있으며, 계산 그래프의 구성 및 설계에 관련된 다른 서비스를 제공할 수 있다. 그래프의 노드 및 요소는 데이터 판독가능 매체에 기억된 데이터 구조 또는 데이터 레포지토리 내에 기억된 데이터 모델에 부합되는 다른 데이터 구조로서 구현될 수 있다.

[0073] 소프트웨어는 범용 또는 전용 프로그램가능한 컴퓨터에 의해 판독가능한 CD-ROM 등의 기억 매체에 제공되거나 네트워크의 통신 매체를 통해(전파 신호로 부호화된 상태로) 컴퓨터에 제공될 수 있다. 모든 기능은 전용의 컴퓨터에서 수행되거나, 코프로세서 등의 전용 하드웨어를 사용하여 수행될 수 있다. 소프트웨어는 소프트웨어에 의해 특정된 계산의 여러 부분이 여러 컴퓨터에 의해 수행되는 분산형으로 구현될 수 있다. 이러한 컴퓨터 프로그램은 범용 또는 전용의 프로그램가능한 컴퓨터에 의해 판독가능한 기억 매체 또는 기억 장치(예를 들어, 고체 메모리 또는 매체, 또는 자기나 광 매체)에 기억되거나 이에 다운로드되는 것이 바람직하며, 본원에 개시된 과정을 수행하기 위해 컴퓨터 시스템에 의해 기억 매체 또는 장치가 판독될 때에 컴퓨터를 구성 및 동작시킨다. 본 발명의 시스템은 컴퓨터 프로그램에 의해 구성된, 컴퓨터로 판독가능한 기억 매체로서 구현되는 것으로 간주될 수 있으며, 기억 매체는 본원에 개시된 기능을 수행하도록 특정 및 미리 정해진 방식으로 컴퓨터 시스템을 동작시키도록 구성된다.

[0074] 본 발명의 여러 실시예를 개시하였지만, 본 발명의 범위를 벗어남이 없이 다양한 많은 변형이 가능하다는 것을 알 수 있을 것이다. 예를 들어, 상기 설명한 단계 중의 일부는 순서에 의존하지 않으며, 개시된 것과 다른 순서로 수행될 수 있다.

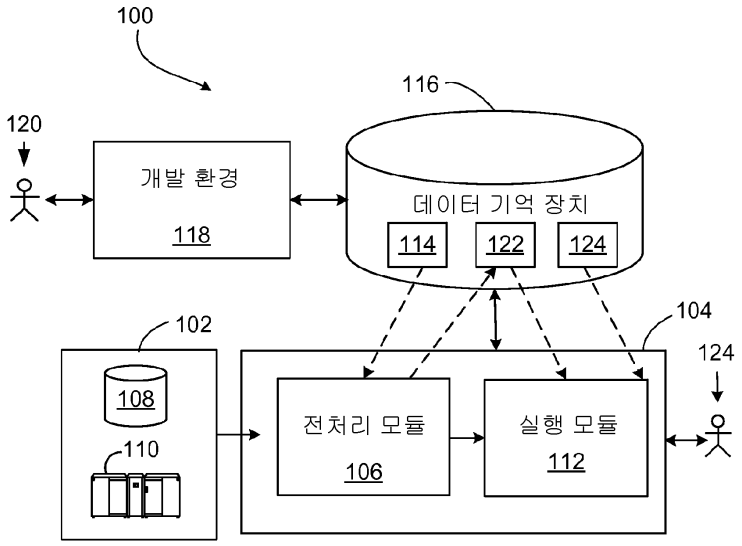
[0075] 상기 설명은 예시에 불과하며 본 발명의 범위를 제한하기 위한 것이 아니고, 청구범위에 의해서 정의된다는 것

을 알 수 있을 것이다. 예를 들어, 상기 설명한 많은 기능 단계는 전체적인 처리에 실질적으로 영향을 주지 않으면서 다른 순서로 수행될 수 있다. 다른 실시예도 청구범위의 기술적 범위 내에 포함된다.

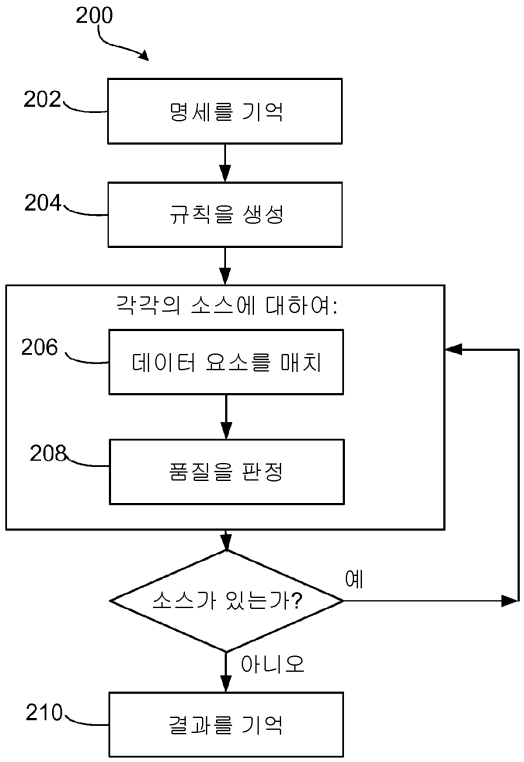
[0076] 본 출원은 2010년 1월 13일에 출원된 미국출원 제61/294,663호에 대하여 우선권을 주장하며, 그 전체 내용을 본원에 참조에 의해 원용한다.

도면

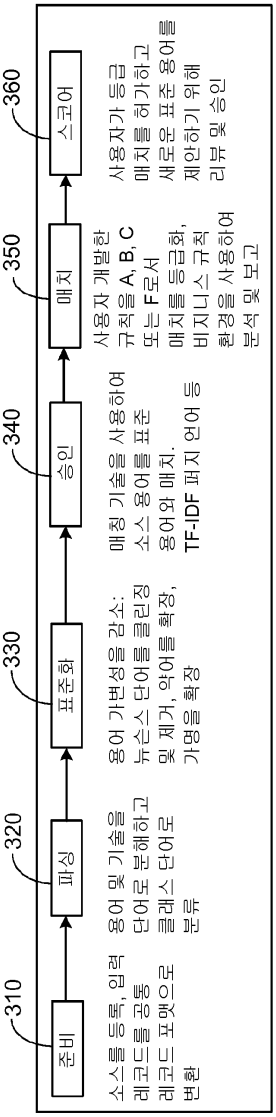
도면1



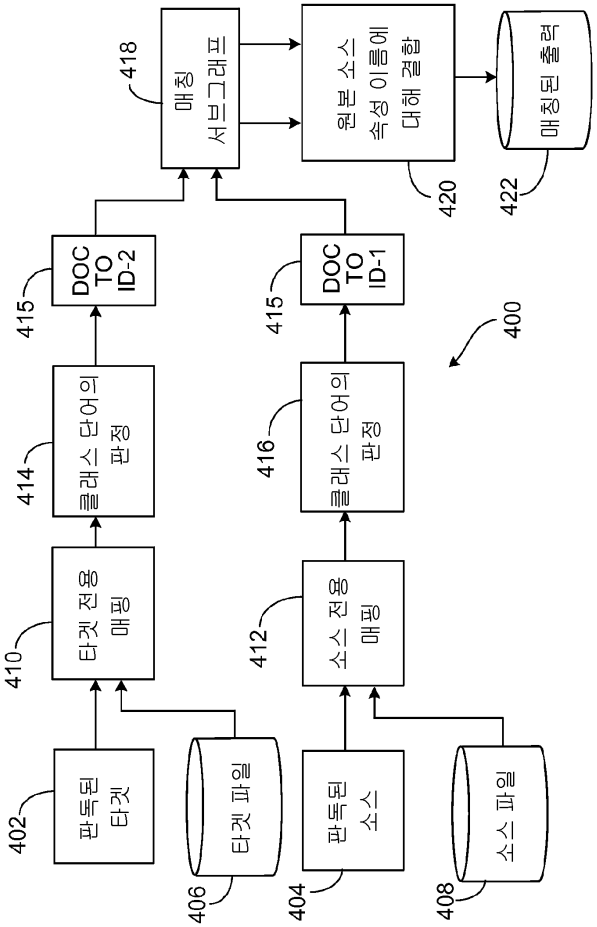
도면2



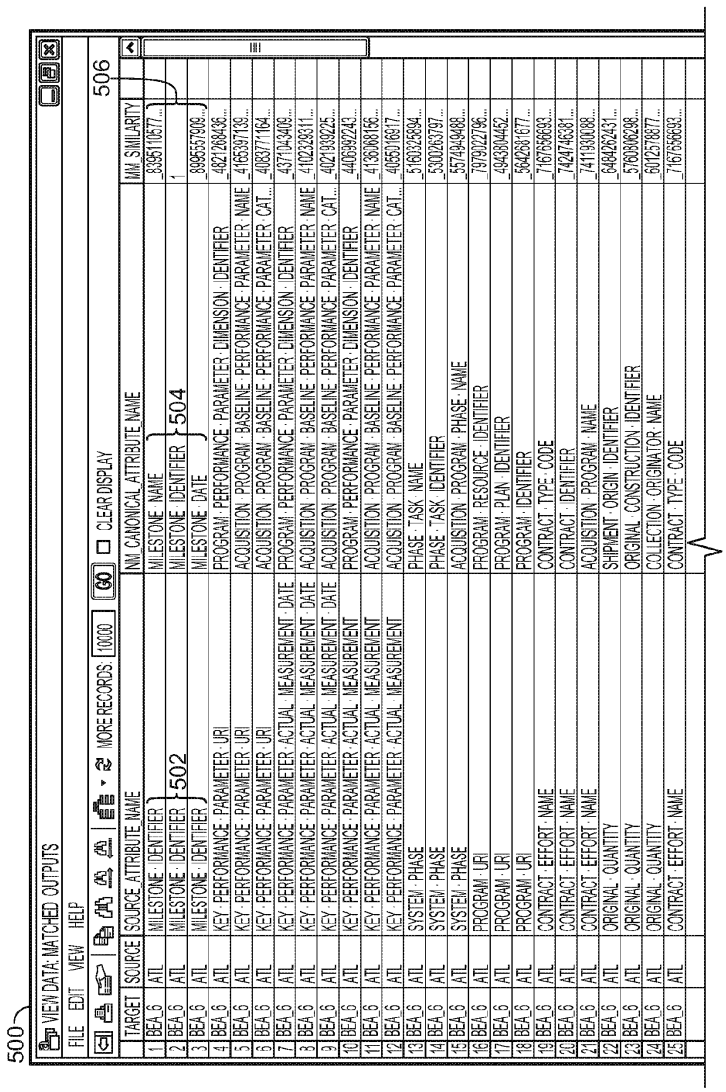
도면3


310
320
330
340
350
360
준비
파싱
표준화
승인
매치
스코어
소스를 등록, 입력
레코드를 공통
레코드 포맷으로
변환
용어 및 기술을
단어로 분해하고
클래스 단어로
분류
용어 가변성을 감소:
뉴슨스 단어를 클린징
및 제거, 약어를 확장,
가명을 확장
매칭 기술을 사용하여
소스 용어를 표준
용어와 매치.
TF-IDF 퍼지 언어 등
사용자 개발한
규칙을 A, B, C
또는 F로서
매치를 등급화,
비지니스 규칙
환경을 사용하여
분석 및 보고
사용자가 등급
매치를 허가하고
새로운 표준 용어를
제안하기 위해
리뷰 및 승인

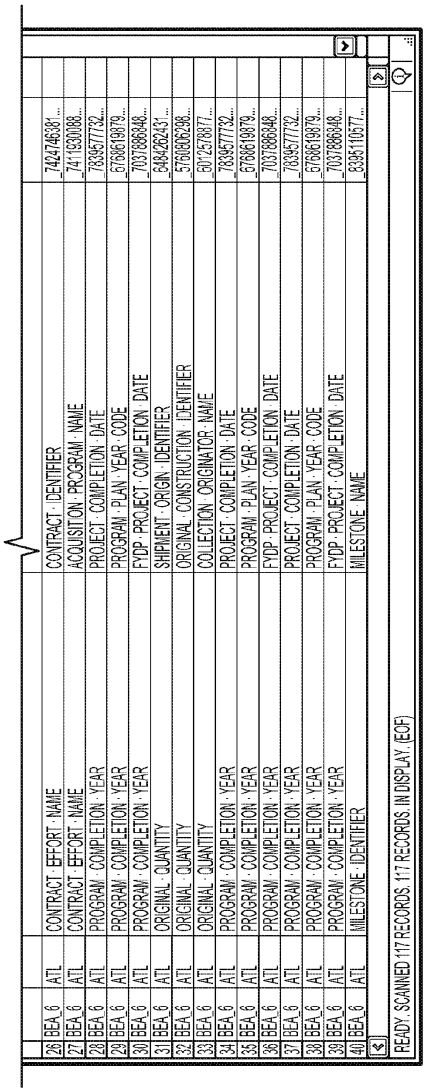
도면4


판독된 타겟
402
타겟 파일
406
판독된 소스
404
소스 파일
408
타겟 전용 매핑
410
소스 전용 매핑
412
클래스 단어의 판정
414
클래스 단어의 판정
416
DOC TO ID-2
415
DOC TO ID-1
415
400
매칭 서브그래프
418
원본 소스 속성 이름에 대해 결합
420
매칭된 출력
422

도면5a


500
VIEW DATA: MATCHED OUTPUTS
FILE EDIT VIEW HELP
MORE RECORDS 10000
CLEAR DISPLAY
506
502
504
TARGET
SOURCE
SOURCE ATTRIBUTE NAME
NM_CANONICAL_ATTRIBUTE_NAME
NM_SIMILARITY
1 BEA_6 ATL MILESTONE : IDENTIFIER MILESTONE : NAME 8395110677...
2 BEA_6 ATL MILESTONE : IDENTIFIER MILESTONE : IDENTIFIER 1
3 BEA_6 ATL MILESTONE : IDENTIFIER MILESTONE : DATE 8895537908...
4 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : URI PROGRAM : PERFORMANCE : PARAMETER : DIMENSION : IDENTIFIER -4621269436...
5 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : URI ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : NAME -4166397139...
6 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : URI ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : CAT... -4083771164...
7 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT : DATE PROGRAM : PERFORMANCE : PARAMETER : DIMENSION : IDENTIFIER -4071343408...
8 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT : DATE ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : NAME -4102228311...
9 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT : DATE ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : CAT... -4021939225...
10 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT PROGRAM : PERFORMANCE : PARAMETER : DIMENSION : IDENTIFIER 4406392743...
11 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : NAME 4136368156...
12 BEA_6 ATL KEY : PERFORMANCE : PARAMETER : ACTUAL : MEASUREMENT ACQUISITION : PROGRAM : BASELINE : PERFORMANCE : PARAMETER : CAT... 4055116317...
13 BEA_6 ATL SYSTEM : PHASE PHASE : TASK : NAME 5168326894...
14 BEA_6 ATL SYSTEM : PHASE PHASE : TASK : IDENTIFIER 5300263797...
15 BEA_6 ATL SYSTEM : PHASE ACQUISITION : PROGRAM : PHASE : NAME 3577494948...
16 BEA_6 ATL PROGRAM : URI PROGRAM : RESOURCE : IDENTIFIER 7979327796...
17 BEA_6 ATL PROGRAM : URI PROGRAM : PLAN : IDENTIFIER -4943504462...
18 BEA_6 ATL PROGRAM : URI PROGRAM : IDENTIFIER 5642931677...
19 BEA_6 ATL CONTRACT : EFFORT : NAME CONTRACT : TYPE : CODE 7167536693...
20 BEA_6 ATL CONTRACT : EFFORT : NAME CONTRACT : IDENTIFIER -7424746381...
21 BEA_6 ATL CONTRACT : EFFORT : NAME ACQUISITION : PROGRAM : NAME 7411930088...
22 BEA_6 ATL ORIGINAL : QUANTITY SHIPMENT : ORIGIN : IDENTIFIER -6484262431...
23 BEA_6 ATL ORIGINAL : QUANTITY ORIGINAL : CONSTRUCTION : IDENTIFIER 5760306238...
24 BEA_6 ATL ORIGINAL : QUANTITY COLLECTION : ORIGINATOR : NAME -8012570877...
25 BEA_6 ATL CONTRACT : EFFORT : NAME CONTRACT : TYPE : CODE 7167536693...

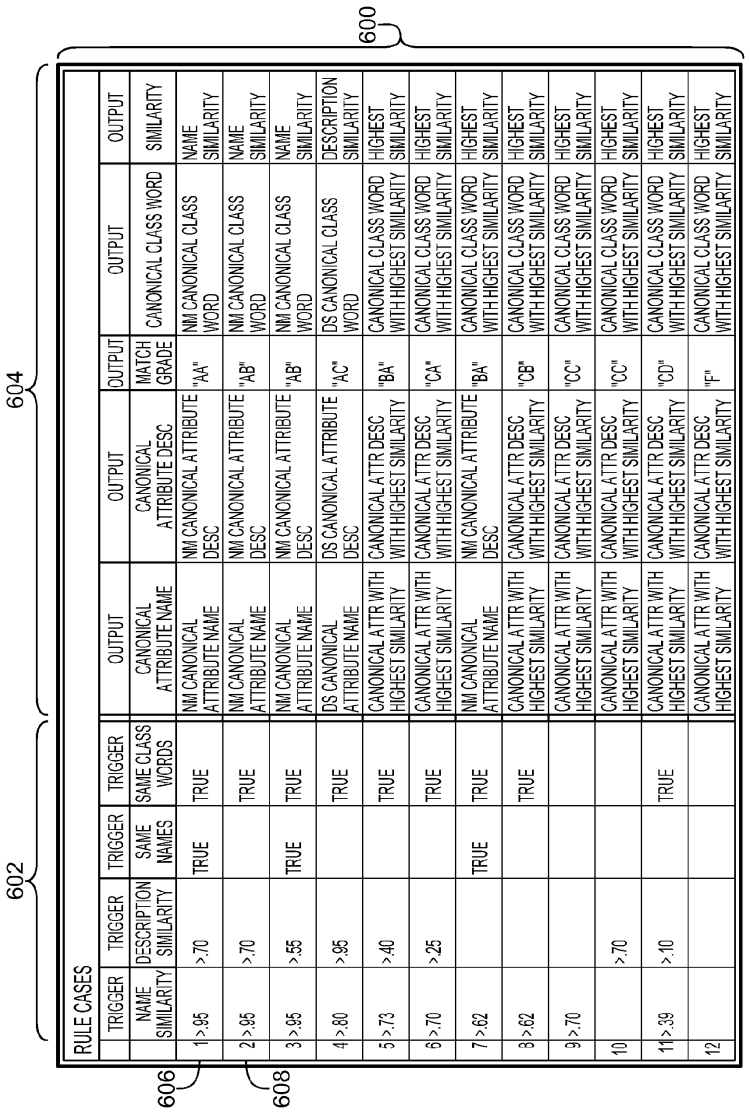
도면5b



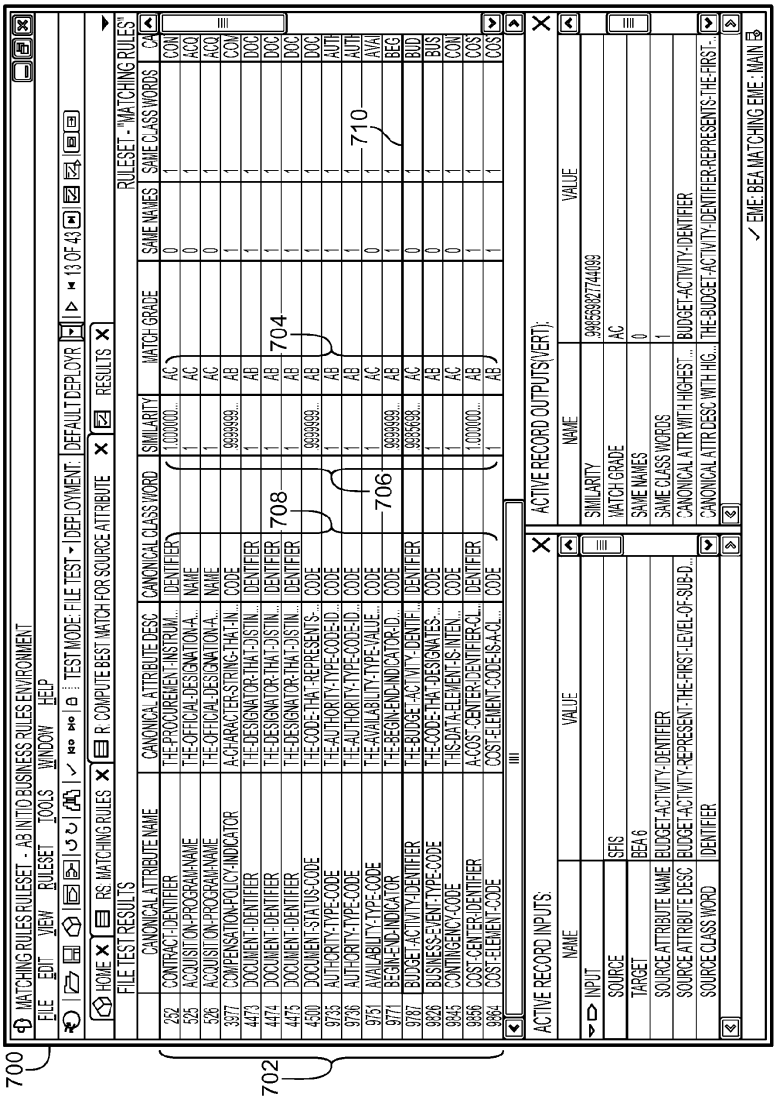
26	BEA_6	ATL	CONTRACT : EFFORT : NAME	CONTRACT : IDENTIFIER	_7424746381...
27	BEA_6	ATL	CONTRACT : EFFORT : NAME	ACQUISITION : PROGRAM : NAME	_7411930088...
28	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROJECT : COMPLETION : DATE	_7839577132...
29	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROGRAM : PLAN : YEAR : CODE	_6706619879...
30	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	FYDP : PROJECT : COMPLETION : DATE	_7037886848...
31	BEA_6	ATL	ORIGINAL : QUANTITY	SHIPMENT : ORIGIN : IDENTIFIER	_6494280243...
32	BEA_6	ATL	ORIGINAL : QUANTITY	ORIGINAL : CONSTRUCTION : IDENTIFIER	_5768066298...
33	BEA_6	ATL	ORIGINAL : QUANTITY	COLLECTION : ORIGINATOR : NAME	_8012578877...
34	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROJECT : COMPLETION : DATE	_7839577132...
35	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROGRAM : PLAN : YEAR : CODE	_6706619879...
36	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	FYDP : PROJECT : COMPLETION : DATE	_7037886848...
37	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROJECT : COMPLETION : DATE	_7839577132...
38	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	PROGRAM : PLAN : YEAR : CODE	_6706619879...
39	BEA_6	ATL	PROGRAM : COMPLETION : YEAR	FYDP : PROJECT : COMPLETION : DATE	_7037886848...
40	BEA_6	ATL	MILESTONE : IDENTIFIER	MILESTONE : NAME	_8036110677...

READY. SCANNED 117 RECORDS. 117 RECORDS IN DISPLAY. (EOF)

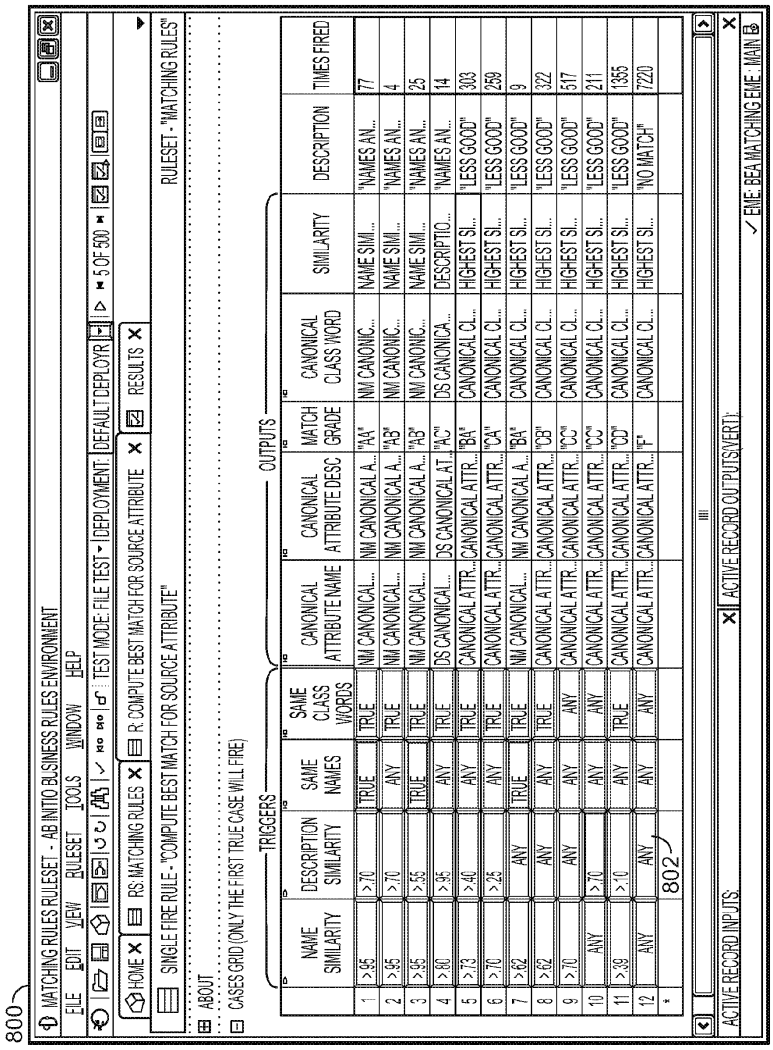
도면6


600
602
604
606
608
RULE CASES
TRIGGER
TRIGGER
OUTPUT
TRIGGER
OUTPUT
OUTPUT
OUTPUT
NAME SIMILARITY
DESCRIPTION SIMILARITY
SAME NAMES
SAME CLASS WORDS
CANONICAL ATTRIBUTE NAME
CANONICAL ATTRIBUTE DESC
MATCH GRADE
CANONICAL CLASS WORD
SIMILARITY
1 >.95 >.70 TRUE TRUE NM CANONICAL ATTRIBUTE NAME NM CANONICAL ATTRIBUTE DESC "AA" NM CANONICAL CLASS WORD NAME SIMILARITY
2 >.95 >.70 TRUE NM CANONICAL ATTRIBUTE NAME NM CANONICAL ATTRIBUTE DESC "AB" NM CANONICAL CLASS WORD NAME SIMILARITY
3 >.95 >.55 TRUE TRUE NM CANONICAL ATTRIBUTE NAME NM CANONICAL ATTRIBUTE DESC "AB" NM CANONICAL CLASS WORD NAME SIMILARITY
4 >.80 >.95 TRUE DS CANONICAL ATTRIBUTE NAME DS CANONICAL ATTRIBUTE DESC "AC" DS CANONICAL CLASS WORD DESCRIPTION SIMILARITY
5 >.73 >.40 TRUE CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "BA" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
6 >.70 >.25 TRUE CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "CA" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
7 >.62 TRUE TRUE NM CANONICAL ATTRIBUTE NAME NM CANONICAL ATTRIBUTE DESC "BA" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
8 >.62 TRUE CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "CB" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
9 >.70 CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "CC" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
10 >.70 CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "CC" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
11 >.39 >.10 TRUE CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "CD" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY
12 CANONICAL ATTR WITH HIGHEST SIMILARITY CANONICAL ATTR DESC WITH HIGHEST SIMILARITY "F" CANONICAL CLASS WORD WITH HIGHEST SIMILARITY HIGHEST SIMILARITY

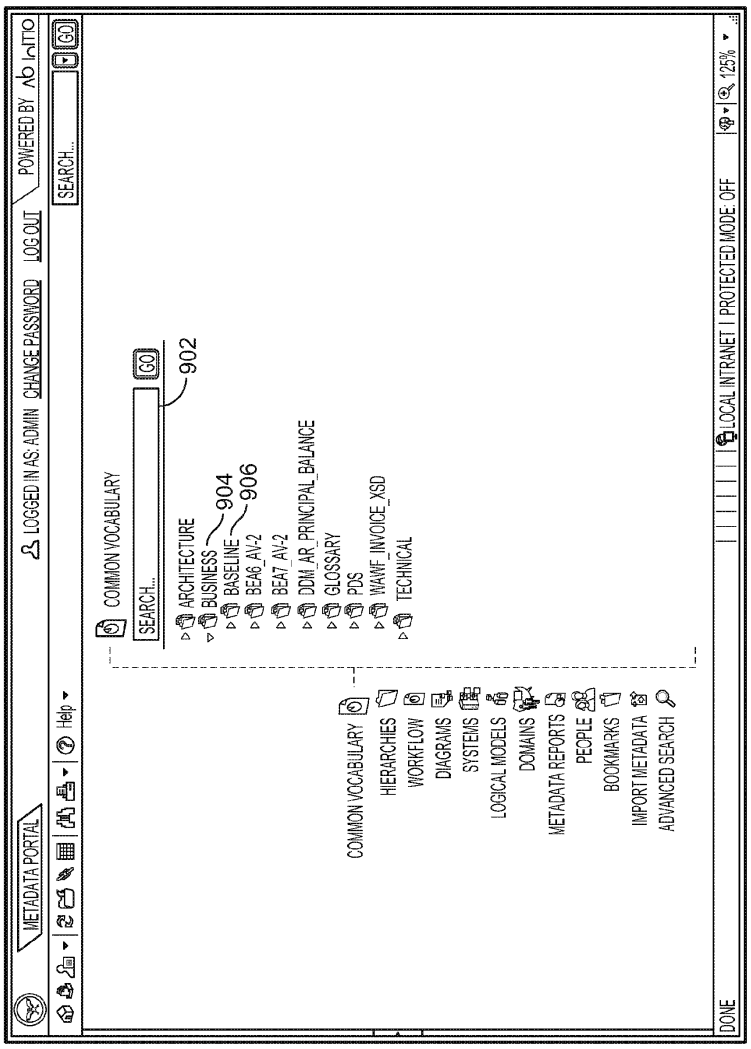
도면7


700
702
704
706
708
710
MATCHING RULES RULESET - AB INITIO BUSINESS RULES ENVIRONMENT
FILE EDIT VIEW RULESET TOOLS WINDOW HELP
TEST MODE: FILE TEST | DEPLOYMENT: DEFAULT DEPLOYR | 13 OF 43
HOME | RG: MATCHING RULES | R: COMPUTE BEST MATCH FOR SOURCE ATTRIBUTE | RESULTS
FILE TEST RESULTS
RULESET - "MATCHING RULES"
CANONICAL ATTRIBUTE NAME | CANONICAL ATTRIBUTE DESC | CANONICAL CLASS WORD | SIMILARITY | MATCH GRADE | SAME NAMES | SAME CLASS WORDS | CA
262 CONTRACT-IDENTIFIER | THE-PROCUREMENT-INSTRUM... | IDENTIFIER | 1.000000... | AC | 0 | 1 | CON
525 ACQUISITION-PROGRAM-NAME | THE-OFFICIAL-DESIGNATION-A... | NAME | 1 | AC | 0 | 1 | ACQ
526 ACQUISITION-PROGRAM-NAME | THE-OFFICIAL-DESIGNATION-A... | NAME | 1 | AC | 0 | 1 | ACQ
3277 COMPENSATION-POLICY-INDICATOR | A-CHARACTER-STRING-THAT-IN... | CODE | .9999999... | AB | 1 | 1 | COM
4473 DOCUMENT-IDENTIFIER | THE-DESIGNATOR-THAT-DISTIN... | IDENTIFIER | 1 | AB | 1 | 1 | DOC
4474 DOCUMENT-IDENTIFIER | THE-DESIGNATOR-THAT-DISTIN... | IDENTIFIER | 1 | AB | 1 | 1 | DOC
4475 DOCUMENT-IDENTIFIER | THE-DESIGNATOR-THAT-DISTIN... | IDENTIFIER | 1 | AB | 1 | 1 | DOC
4500 DOCUMENT-STATUS-CODE | THE-CODE-THAT-REPRESENTS-... | CODE | .9999999... | AB | 1 | 1 | DOC
9735 AUTHORITY-TYPE-CODE | THE-AUTHORITY-TYPE-CODE-ID... | CODE | 1 | AB | 1 | 1 | AUTH
9736 AUTHORITY-TYPE-CODE | THE-AUTHORITY-TYPE-CODE-ID... | CODE | 1 | AB | 1 | 1 | AUTH
9751 AVAILABILITY-TYPE-CODE | THE-AVAILABILITY-TYPE-VALUE... | CODE | 1 | AC | 0 | 1 | AVAI
9771 BEGIN-END-INDICATOR | THE-BEGIN-END-INDICATOR-D... | CODE | .9999999... | AB | 1 | 1 | BEG
9787 BUDGET-ACTIVITY-IDENTIFIER | THE-BUDGET-ACTIVITY-IDENTIF... | IDENTIFIER | .9995638... | AC | 0 | 1 | BUD
9926 BUSINESS-EVENT-TYPE-CODE | THE-CODE-THAT-DESIGNATES-... | CODE | 1 | AB | 0 | 1 | BUS
9945 CONTINGENCY-CODE | THIS-DATA-ELEMENT-IS-INTEN... | CODE | 1 | AC | 0 | 1 | CON
9956 COST-CENTER-IDENTIFIER | A-COST-CENTER-IDENTIFIER-CL... | IDENTIFIER | 1.000000... | AB | 1 | 1 | COS
9964 COST-ELEMENT-CODE | COST-ELEMENT-CODE-ISA-CL... | CODE | 1 | AB | 1 | 1 | COS
ACTIVE RECORD INPUTS:
NAME | VALUE
INPUT
SOURCE | SFIS
TARGET | BEA 6
SOURCE ATTRIBUTE NAME | BUDGET-ACTIVITY-IDENTIFIER
SOURCE ATTRIBUTE DESC | BUDGET-ACTIVITY-REPRESENT-THE-FIRST-LEVEL-OF-SUB-D...
SOURCE CLASS WORD | IDENTIFIER
ACTIVE RECORD OUTPUTS (VERT):
NAME | VALUE
SIMILARITY | 3985593827744099
MATCH GRADE | AC
SAME NAMES | 0
SAME CLASS WORDS | 1
CANONICAL ATTR WITH HIGHEST ... | BUDGET-ACTIVITY-IDENTIFIER
CANONICAL ATTR DESC WITH HIG... | THE-BUDGET-ACTIVITY-IDENTIFIER-REPRESENTS-THE-FIRST-...
EME: BEA MATCHING EME: MAIN

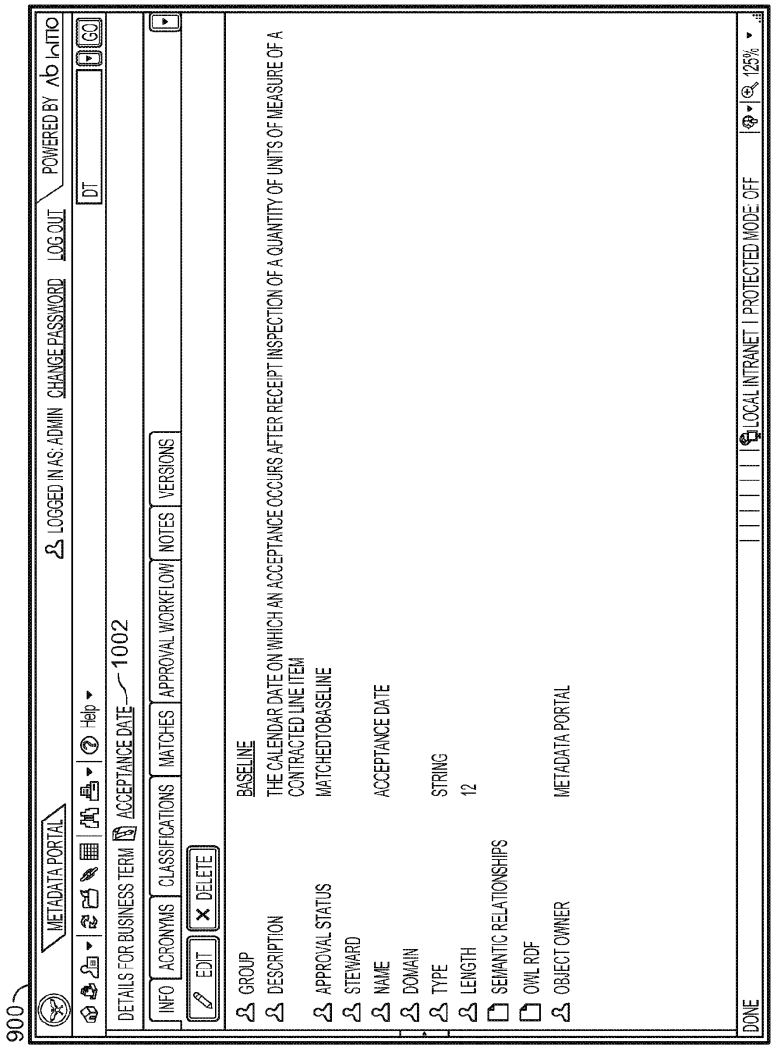
도면8


800
MATCHING RULES RULESET - AB INITIO BUSINESS RULES ENVIRONMENT
FILE EDIT VIEW RULESET TOOLS WINDOW HELP
HOME
RS: MATCHING RULES
TEST MODE: FILE TEST
DEPLOYMENT
DEFAULT DEPLOYR
5 OF 500
R: COMPUTE BEST MATCH FOR SOURCE ATTRIBUTE
RESULTS
SINGLE FIRE RULE - "COMPUTE BEST MATCH FOR SOURCE ATTRIBUTE"
RULESET - "MATCHING RULES"
ABOUT
CASES GRID (ONLY THE FIRST TRUE CASE WILL FIRE)
TRIGGERS
OUTPUTS
NAME SIMILARITY
DESCRIPTION SIMILARITY
SAME NAMES
SAME CLASS WORDS
CANONICAL ATTRIBUTE NAME
CANONICAL ATTRIBUTE DESC
MATCH GRADE
CANONICAL CLASS WORD
SIMILARITY
DESCRIPTION
TIMES FIRED
802
ACTIVE RECORD INPUTS:
ACTIVE RECORD OUTPUTS(VERT):
EME: BEA MATCHING EME: MAIN

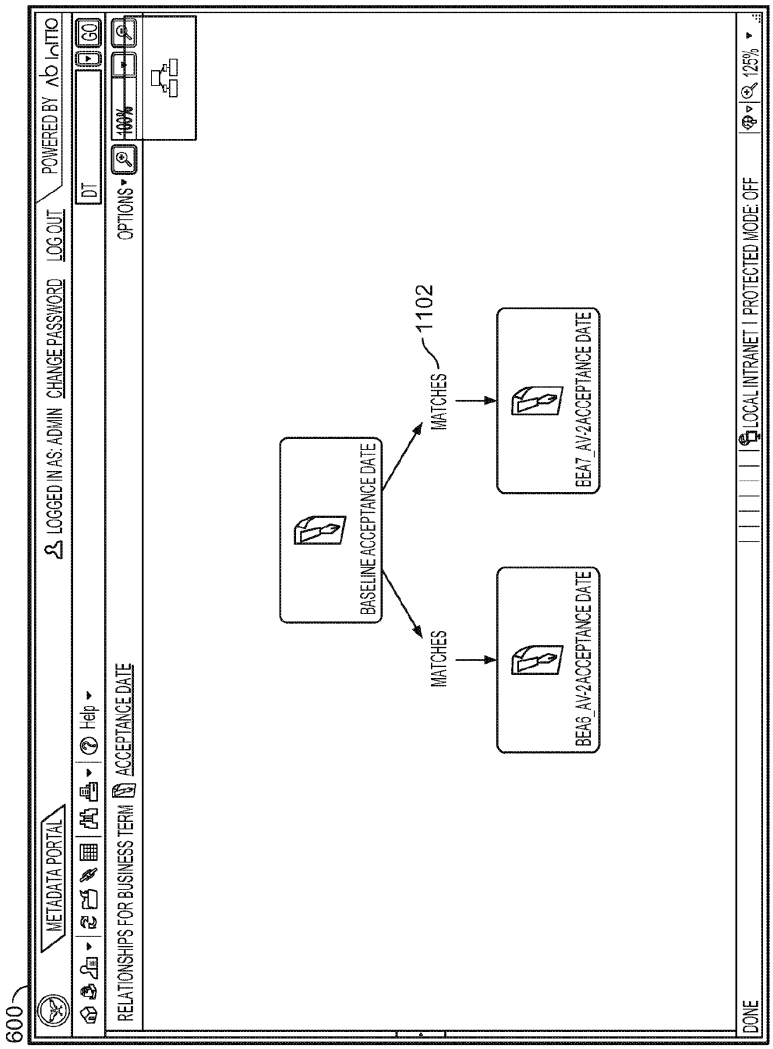
도면9


METADATA PORTAL
LOGGED IN AS: ADMIN
CHANGE PASSWORD
LOG OUT
POWERED BY
Help
SEARCH...
GO
COMMON VOCABULARY
SEARCH...
GO
902
ARCHITECTURE
BUSINESS
904
BASELINE
906
BEA6_AV-2
BEA7_AV-2
DDM_AR_PRINCIPAL_BALANCE
GLOSSARY
PDS
WAWF_INVOICE_XSD
TECHNICAL
COMMON VOCABULARY
HIERARCHIES
WORKFLOW
DIAGRAMS
SYSTEMS
LOGICAL MODELS
DOMAINS
METADATA REPORTS
PEOPLE
BOOKMARKS
IMPORT METADATA
ADVANCED SEARCH
DONE
LOCAL INTRANET | PROTECTED MODE: OFF
125%

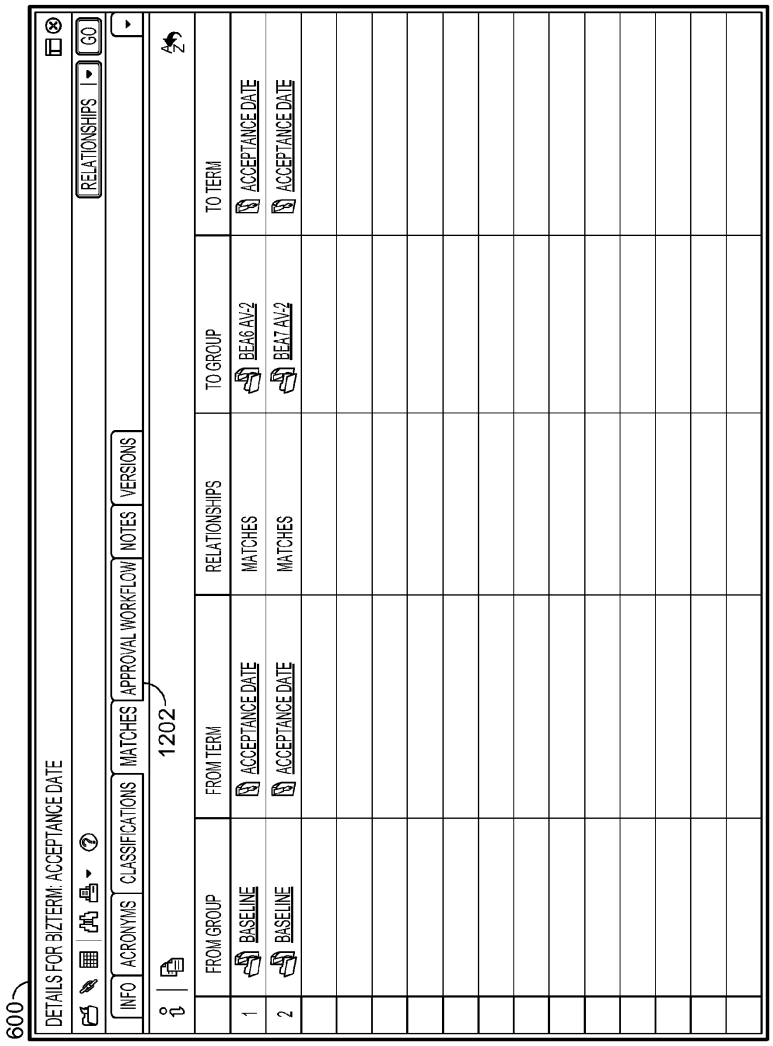
도면10


900
METADATA PORTAL
LOGGED IN AS: ADMIN
CHANGE PASSWORD
LOG OUT
POWERED BY
Help
DT
GO
DETAILS FOR BUSINESS TERM
ACCEPTANCE DATE
1002
INFO
ACRONYMS
CLASSIFICATIONS
MATCHES
APPROVAL WORKFLOW
NOTES
VERSIONS
EDIT
DELETE
GROUP
BASELINE
DESCRIPTION
THE CALENDAR DATE ON WHICH AN ACCEPTANCE OCCURS AFTER RECEIPT INSPECTION OF A QUANTITY OF UNITS OF MEASURE OF A CONTRACTED LINE ITEM
APPROVAL STATUS
MATCHEDTOBASELINE
STEWARD
NAME
ACCEPTANCE DATE
DOMAIN
TYPE
STRING
LENGTH
12
SEMANTIC RELATIONSHIPS
OWL RDF
OBJECT OWNER
METADATA PORTAL
DONE
LOCAL INTRANET | PROTECTED MODE: OFF
125%

도면11


600
METADATA PORTAL
LOGGED IN AS: ADMIN
CHANGE PASSWORD
LOG OUT
POWERED BY
Help
DT
GO
RELATIONSHIPS FOR BUSINESS TERM
ACCEPTANCE DATE
OPTIONS
100%
BASELINE ACCEPTANCE DATE
MATCHES
1102
MATCHES
BEA7_AV-2ACCEPTANCE DATE
BEA6_AV-2ACCEPTANCE DATE
DONE
LOCAL INTRANET | PROTECTED MODE: OFF
125%

도면12


600
DETAILS FOR BIZTERM: ACCEPTANCE DATE
RELATIONSHIPS
GO
INFO
ACRONYMS
CLASSIFICATIONS
MATCHES
APPROVAL WORKFLOW
NOTES
VERSIONS
1202
FROM GROUP
FROM TERM
RELATIONSHIPS
TO GROUP
TO TERM
1
BASELINE
ACCEPTANCE DATE
MATCHES
BEA6 AV-2
ACCEPTANCE DATE
2
BASELINE
ACCEPTANCE DATE
MATCHES
BEA7 AV-2
ACCEPTANCE DATE