Ministero dello Sviluppo Economico
DIREZIONE GENERALE PER LA TUTELA DELLA PROPRIETA' INDUSTRIALE
UFFICIO ITALIANO BREVETTI E MARCHI

UIBM

| | |
|---|---|
| **DOMANDA DI INVENZIONE NUMERO** | **102022000019902** |
| **Data Deposito** | **28/09/2022** |
| **Data Pubblicazione** | **13/06/2024** |

Classifiche IPC

| Sezione | Classe | Sottoclasse | Gruppo | Sottogruppo |
|---|---|---|---|---|
| G | 06 | Q | 10 | 08 |

| Sezione | Classe | Sottoclasse | Gruppo | Sottogruppo |
|---|---|---|---|---|
| G | 06 | Q | 10 | 087 |

| Sezione | Classe | Sottoclasse | Gruppo | Sottogruppo |
|---|---|---|---|---|
| G | 06 | Q | 50 | 04 |

Titolo

SISTEMA E METODO PER L?IDENTIFICAZIONE DI VOCI DUPLICATE, RELATIVE A MATERIALI IDENTICI O EQUIVALENTI, IN UN?ANAGRAFICA DI MATERIALI INDUSTRIALI.

# SISTEMA E METODO PER L'IDENTIFICAZIONE DI VOCI DUPLICATE, RELATIVE A MATERIALI IDENTICI O EQUIVALENTI, IN UN'ANAGRAFICA DI MATERIALI INDUSTRIALI

## DESCRIZIONE

La presente invenzione riguarda un sistema e un metodo per identificare le voci (dette anche record) di un'anagrafica di materiali (detti anche oggetti) industriali che sono relative a, ossia riguardano, materiali identici per natura oppure equivalenti per funzione nell'ambito di un processo industriale. I materiali industriali sono oggetti destinati all'utilizzo in un processo industriale.

Il sistema e il metodo secondo la presente invenzione sono particolarmente, seppur non esclusivamente, utili e pratici nell'ambito della manutenzione di anagrafiche di materiali industriali in aziende di medie e/o grandi dimensioni.

Si nota che nella presente descrizione il termine "voce" (o "record") indica ogni singolo elemento compreso in un registro o, più in generale, in un gruppo ordinato e omogeneo di

dati. Ogni voce comprende una pluralità di dati relativi a una rispettiva entità. Si nota anche che nella presente descrizione il termine "anagrafica" indica un registro comprendente una pluralità di voci (o record). Nell'ambito della presente invenzione, il registro comprendente le voci è l'anagrafica di materiali industriali e ogni voce comprende una pluralità di dati relativi ad un rispettivo materiale (o oggetto).

Comunemente, i processi organizzativi implementati e applicati nelle aziende di medie e/o grandi dimensioni prevedono la redazione e la manutenzione di documenti o registri strutturati per gestire le anagrafiche (in inglese *master data*) delle entità rilevanti per lo svolgimento delle funzioni aziendali. Ad esempio, queste entità possono essere clienti, fornitori, materie prime e materiali (o oggetti) industriali.

Più specificamente, nell'ambito del cosiddetto *procurement* - che comprende le attività di acquisto e approvvigionamento di beni e servizi diretti e indiretti, monitoraggio e selezione dei fornitori, negoziazione dei contratti, analisi dei dati di spesa, ottimizzazione dei costi di

acquisto -, un ruolo importante è rivestito dalle anagrafiche di materiali (in inglese *material master data*) che catalogano tutti gli oggetti utilizzati nei processi industriali, tipicamente relativi alla manifattura di prodotti industriali o alla fornitura di servizi.

In generale, un "materiale industriale" è un oggetto semplice, con peculiari informazioni tecniche (ad esempio marca, modello, attributi, parametri tecnici, specifiche tecniche, codici di distribuzione), che viene acquistato ripetutamente sul mercato da uno o più fornitori, conservato in modo organizzato all'interno dei magazzini, movimentato in modo coordinato dai processi logistici e infine utilizzato nei processi industriali, come detto tipicamente relativi alla manifattura di prodotti industriali o alla fornitura di servizi.

Oggigiorno, la gestione e la manutenzione delle anagrafiche di materiali industriali nelle aziende di medie e/o grandi dimensioni è affidata a sistemi software di tipo transazionale, nello specifico i cosiddetti sistemi ERP (acronimo dell'inglese *Enterprise Resource Planner*),

supervisionati da utenti umani. Questi sistemi ERP noti, oltre a gestire le stesse anagrafiche, gestiscono anche il ciclo di vita delle istanze delle entità, in questo caso dei materiali (o oggetti) industriali, descritte nelle anagrafiche nell'ambito più ampio dei sistemi informativi di impresa (EIS, acronimo dell'inglese *Enterprise Information Systems*).

In generale, il ciclo di vita delle istanze dei materiali comprende l'approvvigionamento, la logistica, l'immagazzinamento e la movimentazione di questi materiali nell'ambito dei processi industriali.

In questi sistemi ERP noti, ogni voce dell'anagrafica di materiali industriali comprende un codice identificativo del rispettivo materiale (in inglese *material code*), una descrizione testuale dello stesso materiale, ed altri campi strutturati relativi allo stesso materiale.

Il codice identificativo della voce di anagrafica è univoco all'interno del sistema ERP, e quindi del sistema informativo, ed è usato per identificare univocamente il rispettivo materiale nelle transazioni.

La descrizione testuale della voce di anagrafica è utilizzata solo dagli utenti umani del sistema ERP come forma di documentazione scritta relativa alle informazioni tecniche del rispettivo materiale (come detto, ad esempio marca, modello, attributi, parametri tecnici, specifiche tecniche, codici di distribuzione). Come tale, la descrizione testuale risulta "opaca", ossia non direttamente interpretabile dal sistema ERP che gestisce le transazioni del sistema informativo.

Ad esempio, gli altri campi strutturati della voce di anagrafica possono comprendere: codici di categorizzazione merceologica, afferenza a specifici centri di competenza o di costo, tipologia o nome del fornitore, e/o altri metadati.

Garantire un alto livello di qualità, o meglio accuratezza, dei dati (in inglese *data quality*) delle anagrafiche di materiali è un obiettivo importante per le aziende di medie e/o grandi dimensioni, perché l'efficacia e l'efficienza dei processi (ad esempio approvvigionamento, logistica, immagazzinamento,

produzione) collegati a queste anagrafiche dipende da questo livello di qualità, almeno in parte.

Tuttavia, questi sistemi ERP noti non sono scevri di inconvenienti, tra i quali va annoverata una delle tipologie di errori più comuni presenti nelle anagrafiche in generale, e nelle anagrafiche di materiali industriali in particolare, che consiste nella presenza di voci duplicate. Questa duplicazione si ha quando la stessa entità, in questo caso lo stesso materiale (o oggetto) industriale, è censita in due o più voci diverse, a cui sono associati due rispettivi codici identificativi diversi.

Questa tipologia di errori nelle anagrafiche gestite dai sistemi ERP noti può provocare perdite di efficacia ed efficienza in tutti i processi successivi dipendenti. Ad esempio, un utente umano che interroga il sistema ERP o informativo per recuperare e conoscere informazioni tecniche su una specifica entità, in questo caso uno specifico materiale industriale, potrebbe consultare una sola delle due o più voci relative alla stessa entità e quindi potrebbe ricevere informazioni parziali sull'approvvigionamento,

l'immagazzinamento, le scorte di magazzino, il consumo, e così via.

Le cause della presenza di questi errori di duplicazione nelle anagrafiche gestite dai sistemi ERP sono molteplici, e tutte legate in un modo o nell'altro al processo di redazione e manutenzione delle stesse anagrafiche all'interno delle aziende di medie e/o grandi dimensioni.

In teoria, il modo principale per capire se un'entità, in questo caso un materiale industriale, è già censita, e quindi evitare l'inserimento di una voce duplicata nell'anagrafica, consiste nel fare riferimento alle descrizioni testuali delle voci già presenti nell'anagrafica.

Tuttavia, questi testi possono essere inaccurati e/o incompleti, poiché tipicamente i sistemi ERP noti, e più in generali i sistemi informativi, mettono a disposizione uno spazio limitato (ad esempio di 80 caratteri al massimo) per la descrizione che spesso non è sufficiente per elencare tutte le informazioni tecniche indispensabili per una corretta identificazione del materiale industriale.

In altre parole, le informazioni tecniche dell'entità, in questo caso il materiale industriale, contenute all'interno della voce di anagrafica, sono potenzialmente incomplete. Quindi, non è detto che tutte le informazioni tecniche necessarie per individuare univocamente il materiale industriale siano specificate nella descrizione testuale e/o negli altri campi strutturati della voce di anagrafica.

Un'ulteriore causa degli errori di duplicazione nelle anagrafiche consiste nel fatto che spesso le descrizioni testuali dei materiali industriali sono scritte in lingue diverse. Questo aspetto è comune nel caso di aziende internazionali, dove le descrizioni testuali sono utilizzate per dialogare con i fornitori locali e dove globalmente si trovano anagrafiche con descrizioni in dieci o più lingue diverse. In questa situazione nessun utente umano all'interno dell'azienda è in grado di controllare le voci delle anagrafiche, tramite i sistemi ERP noti, e discriminare sulla totalità delle descrizioni testuali.

Inoltre, una causa degli errori di

duplicazione nelle anagrafiche consiste nel fatto che, all'interno delle aziende di medie e/o grandi dimensioni, è comune che il compito di manutenzione delle stesse anagrafiche sia distribuito tra decine o addirittura centinaia di utenti umani, con livelli di competenza e aree specifiche di esperienza diverse. In questa situazione un utente umano, all'atto di inserire una nuova voce nell'anagrafica, tramite il sistema ERP noto, non è in grado di decidere con un alto livello di confidenza se la relativa entità, in questo caso il materiale industriale, è effettivamente già censita.

Indipendentemente da queste sorgenti di errore, il numero di voci nelle anagrafiche in generale, e nelle anagrafiche di materiali industriali in particolare, è molto elevato (centinaia di migliaia o milioni, nel caso delle grandi aziende internazionali), e quindi il singolo utente umano non è in grado di vagliare, in una quantità di tempo economicamente giustificabile, un numero adeguato di voci dell'anagrafica che sono potenziali duplicati.

In alcuni contesti, l'inserimento delle voci

nelle anagrafiche è eseguito in modo automatico, ad esempio quando due sistemi informativi vengono fusi in uno solo, tipicamente a seguito della fusione e/o acquisizione di due aziende precedentemente separate. In questi casi di solito si ricorre opzionalmente ad un processo di armonizzazione manuale a posteriori delle voci, armonizzazione che però rischia di avere una qualità molto bassa per i motivi descritti in precedenza.

In altri contesti, può capitare che sistemi informativi separati debbano convivere all'interno della stessa azienda, e questi sistemi, anche se individualmente non hanno voci duplicate, globalmente assegnano codici identificativi diversi e/o usano sistemi di codifica non equipollenti per riferirsi alle stesse entità.

Attualmente sono noti sistemi software che affiancano i sistemi ERP e che implementano metodologie per orchestrare e ottimizzare il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani.

Tuttavia, questi sistemi noti non si avvalgono di una restrizione precisa del dominio dell'analisi, oppure propongono una restrizione del dominio dell'analisi basata su altri campi strutturati, ma non sulla descrizione testuale, delle voci di anagrafica. Però la descrizione testuale offre il contenuto informativo migliore, per completezza e utilità, in merito al materiale (o oggetto) industriale e alle relative informazioni tecniche.

Compito precipuo della presente invenzione è quello di superare i limiti dell'arte nota sopra esposti, escogitando un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali che consentano di ottenere effetti migliori rispetto a quelli ottenibili con le soluzioni note e/o effetti analoghi a minor costo e con prestazioni più elevate.

Nell'ambito di questo compito, uno scopo della presente invenzione è quello di concepire un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o

equivalenti, in un'anagrafica di materiali industriali che permettano di individuare, in modo automatico ed euristico, un sottoinsieme delle voci di anagrafica che potenzialmente contiene dei duplicati, questo sottoinsieme essendo sufficientemente piccolo e preciso da rendere economicamente sostenibile il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio.

Un altro scopo della presente invenzione è quello di escogitare un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali che consentano di supportare il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio, utilizzando l'analisi linguistica del testo della descrizione testuale delle voci di anagrafica.

Un ulteriore scopo della presente invenzione

è quello di concepire un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali che permettano di supportare il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio, indipendentemente dalla lingua in cui sono scritti questi dati, in particolare la descrizione testuale delle voci di anagrafica.

Ancora, scopo della presente invenzione è quello di escogitare un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali che consentano di manutenere agevolmente anagrafiche di materiali comprendenti un numero di voci molto elevato (centinaia di migliaia o milioni, nel caso delle grandi aziende internazionali).

Non ultimo scopo della presente invenzione è quello di realizzare un sistema e un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica

di materiali industriali che siano di elevata affidabilità, di relativamente semplice realizzazione, ed economicamente competitivi se paragonati alla tecnica nota.

Questo compito, nonché questi ed altri scopi che meglio appariranno in seguito, sono raggiunti da un sistema per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali, comprendente un'unità di memoria anagrafica configurata per memorizzare detta anagrafica di materiali industriali comprendente una pluralità di voci, ogni voce di anagrafica comprendendo una descrizione testuale di un rispettivo materiale industriale,

caratterizzato dal fatto che comprende:

- un modulo di categorizzazione configurato per associare detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi detta voce di anagrafica, ad una rispettiva categoria selezionata da una pluralità di categorie definite in una tassonomia standard e rappresentanti rispettive tipologie di materiale industriale;

- un modulo di ricerca configurato per scoprire ed estrarre almeno una informazione tecnica di detto materiale industriale da detta descrizione testuale compresa in ogni voce di anagrafica, tramite il riconoscimento di un rispettivo schema tra un gruppo di schemi di informazione tecnica associati a detta categoria selezionata da detto modulo di categorizzazione; e

- un'unità di memoria analitica configurata per memorizzare detta tassonomia standard comprendente detta pluralità di categorie che rappresentano rispettive tipologie di materiale industriale, e una pluralità di schemi di informazione tecnica raggruppati secondo detta pluralità di categorie di detta tassonomia standard.

Il compito e gli scopi prefissati sono altresì raggiunti da un metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali, tramite:

- un'unità di memoria anagrafica configurata per memorizzare detta anagrafica di materiali industriali comprendente una pluralità di voci,

ogni voce di anagrafica comprendendo una descrizione testuale di un rispettivo materiale industriale; e

- un'unità di memoria analitica configurata per memorizzare una tassonomia standard comprendente una pluralità di categorie che rappresentano rispettive tipologie di materiale industriale, e una pluralità di schemi di informazione tecnica raggruppati secondo detta pluralità di categorie di detta tassonomia standard;

caratterizzato dal fatto che comprende i passi che consistono nel:

- associare detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi detta voce di anagrafica, ad una rispettiva categoria selezionata da detta pluralità di categorie definite in detta tassonomia standard e rappresentanti rispettive tipologie di materiale industriale, tramite un modulo di categorizzazione; e

- scoprire ed estrarre almeno una informazione tecnica di detto materiale industriale da detta descrizione testuale compresa

in ogni voce di anagrafica, tramite il riconoscimento di un rispettivo schema tra un gruppo di schemi di informazione tecnica associati a detta categoria selezionata da detto modulo di categorizzazione, tramite un modulo di ricerca.

Ulteriori caratteristiche e vantaggi della presente invenzione risulteranno maggiormente dalla descrizione di una forma di realizzazione preferita, ma non esclusiva, del sistema e del metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione, illustrata a titolo indicativo e non limitativo con l'ausilio dei disegni allegati, in cui:

la figura 1 è uno schema a blocchi che illustra schematicamente una forma di realizzazione del sistema per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione;

la figura 2 è un diagramma di flusso che illustra schematicamente una forma di realizzazione del metodo per l'identificazione di

voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione.

Preliminarmente, si nota che la peculiarità del sistema e del metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione consiste nell'analisi della descrizione testuale delle voci dell'anagrafica di materiali industriali per mezzo della combinazione di tecniche di analisi e comprensione automatica del linguaggio naturale (in inglese *natural language processing* o *natural language understanding*), adattate al dominio di questo specifico tipo di dati, ossia dati relativi a materiali industriali, con tecniche di estrazione automatizzata di informazioni strutturate dal testo in linguaggio naturale (in inglese *text mining*), anch'esse adattate al dominio di questo specifico tipo di dati, ossia dati relativi a materiali industriali.

In breve, i moduli descritti di seguito, ossia modulo di pre-analisi 14, modulo di categorizzazione 15, modulo di ricerca 16 e modulo

di selezione 17, utilizzano le tecniche di *natural language processing* e *text mining*. Le tecniche di *natural language processing* e *text mining* sono studiate dalla branca dell'informatica comunemente chiamata linguistica computazionale.

Con particolare riferimento alla figura 1, il sistema per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione, indicato globalmente con il numero di riferimento 10, comprende sostanzialmente: un'unità elettronica di controllo 12, un modulo di categorizzazione 15, un modulo di ricerca 16, un'unità di memoria anagrafica 20 e un'unità di memoria analitica 22. Preferibilmente, il sistema 10 per l'identificazione di voci duplicate secondo l'invenzione comprende ulteriormente un modulo di pre-analisi 14. Preferibilmente, il sistema 10 per l'identificazione di voci duplicate secondo l'invenzione comprende ulteriormente un modulo di selezione 17.

L'unità elettronica di controllo 12 è l'elemento funzionale principale del sistema 10

per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione, e per questo motivo essa è operativamente collegata e in comunicazione con gli altri elementi compresi nel sistema 10 per l'identificazione di voci duplicate.

L'unità elettronica di controllo 12 del sistema 10 per l'identificazione di voci duplicate è dotata di opportune capacità di calcolo e di interfacciamento con gli altri elementi del sistema 10 per l'identificazione di voci duplicate, ed essa è configurata per comandare, controllare e coordinare il funzionamento degli elementi del sistema 10 per l'identificazione di voci duplicate con i quali essa è operativamente collegata e in comunicazione.

L'unità di memoria anagrafica 20 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, secondo l'invenzione è configurata per memorizzare, ossia registrare, un'anagrafica di materiali industriali comprendente una pluralità di voci, dove ogni voce di anagrafica comprende

una pluralità di dati relativi ad un rispettivo materiale (o oggetto) industriale. Ogni voce dell'anagrafica di materiali industriali, memorizzata nell'unità di memoria anagrafica 20, comprende una descrizione testuale del rispettivo materiale (o oggetto) industriale. Vantaggiosamente, ogni voce dell'anagrafica di materiali industriali, memorizzata nell'unità di memoria anagrafica 20, comprende un codice identificativo del rispettivo materiale (o oggetto) industriale.

Come accennato, l'uso della descrizione testuale della voce di anagrafica come sorgente delle informazioni tecniche del materiale (o oggetto) industriale è fondamentale, perché questa descrizione testuale risulta essere l'unico elemento della voce di anagrafica dove è presente un'informazione precisa relativamente alla natura del materiale (o oggetto) industriale censito.

Il modulo di pre-analisi 14 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione è configurato per scoprire ed

estrarre almeno una feature del materiale industriale (in inglese *feature engineering*) dalla descrizione testuale compresa in ogni voce dell'anagrafica di materiali industriali, memorizzata nell'unità di memoria anagrafica 20. Si nota che nella presente descrizione il termine "*feature*" indica sinteticamente una caratteristica, una proprietà e/o un attributo del materiale (o oggetto) industriale.

Vantaggiosamente, il modulo di pre-analisi 14 è configurato per operare in modo ottimale sulle descrizioni testuali dei materiali (o oggetti) industriali, come detto comprese nelle voci dell'anagrafica di materiali industriali, e caratterizzate da testi corti, multilingua, linguaggio tecnico, e molte informazioni tecniche di tipo numerico.

Preferibilmente, alla luce del dominio dell'analisi che comprende brevi descrizioni di materiali (o oggetti) industriali, il modulo di pre-analisi 14 è ulteriormente configurato per estrarre solo feature rappresentate da parole della descrizione testuale che sono sostantivi e/o aggettivi, e per ignorare (ossia non estrarre)

*feature* rappresentate da parole della descrizione testuale che sono verbi e/o avverbi.

Preferibilmente, alla luce del dominio dell'analisi che comprende brevi descrizioni di materiali (o oggetti) industriali, il modulo di pre-analisi 14 è ulteriormente configurato per ignorare (ossia non estrarre) *feature* rappresentate da parole della descrizione testuale che sono ripetizioni di parole precedenti.

Vantaggiosamente, il modulo di pre-analisi 14 è ulteriormente configurato per associare un peso a ogni feature del materiale (o oggetto) industriale, in modo che alcune *feature* (più "pesanti") siano valutate come più importanti rispetto ad altre *feature* (meno "pesanti").

In una forma di realizzazione, il modulo di pre-analisi 14 può dare più peso, e quindi più importanza, ai numeri "corti" (composti da poche cifre), che spesso identificano specifiche tecniche, rispetto ai numeri "lunghi" (composti da molte cifre), che invece spesso identificano codici specifici del produttore.

In una forma di realizzazione, il modulo di pre-analisi 14 può dare più peso, e quindi più

importanza, alle prime parole della descrizione testuale del materiale (o oggetto) industriale rispetto alle ultime della stessa descrizione testuale. Questa distribuzione di peso, e quindi di importanza, basata su un'analisi di tipo statistico, è peculiare della presente invenzione perché non è vera nelle frasi comuni.

Vantaggiosamente, il modulo di pre-analisi 14 è configurato per operare in modo ottimale sulle descrizioni testuali dei materiali (o oggetti) industriali in lingue diverse.

In una forma di realizzazione, il modulo di pre-analisi 14 può dare meno peso, e quindi meno importanza, alle *feature* del materiale (o oggetto) industriale collegate a forme linguistiche ambigue tra varie lingue, queste feature essendo rintracciate in base a un'analisi estensiva dei vocabolari delle varie lingue, in modo da ridurre le ambiguità tra lingue diverse.

Le *feature* del materiale (o oggetto) industriale, scoperte ed estratte dal modulo di pre-analisi 14, sono fornite in ingresso al modulo di categorizzazione 15, preferibilmente in forma strutturata.

Il modulo di categorizzazione 15 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione è configurato per associare la descrizione testuale del materiale (o oggetto) industriale compresa in ogni voce dell'anagrafica di materiali industriali, e quindi la stessa voce di anagrafica, ad una rispettiva categoria selezionata da una pluralità di categorie definite in una tassonomia standard. Ogni categoria della tassonomia standard rappresenta una rispettiva tipologia di materiale (o oggetto) industriale.

L'unità di memoria analitica 22 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione è configurata per memorizzare, ossia registrare, la tassonomia standard, estremamente granulare ed estesa (ad esempio un albero con più di 140.000 categorie), comprendente la pluralità di categorie che rappresentano le tipologie di materiale (o oggetto) industriale.

Vantaggiosamente, il modulo di

categorizzazione 15 è configurato per operare utilizzando la combinazione di una rete neurale multistrato e un classificatore di tipo *naive bayes*.

Preferibilmente, il modulo di categorizzazione 15 è configurato per associare la descrizione testuale del materiale (o oggetto) industriale, rappresentata sinteticamente dalle *feature* precedentemente scoperte ed estratte dal modulo di pre-analisi 14, ad una rispettiva categoria selezionata dalla pluralità di categorie definite nella tassonomia standard.

La categoria del materiale (o oggetto) industriale, selezionata dal modulo di categorizzazione 15, è fornita in ingresso al modulo di ricerca 16, preferibilmente in forma strutturata.

Il modulo di ricerca 16 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione è configurato per scoprire ed estrarre almeno una informazione tecnica del materiale industriale (come detto, ad esempio marca, modello, attributi,

parametri tecnici, specifiche tecniche, codici di distribuzione) dalla descrizione testuale compresa in ogni voce dell'anagrafica di materiali industriali, tramite il riconoscimento di un rispettivo schema (in inglese *pattern-matching*) tra un gruppo di schemi di informazione tecnica associati alla categoria precedentemente selezionata dal modulo di categorizzazione 15. Questi schemi sono predefiniti, e ognuno di essi è associato ad almeno una categoria della tassonomia standard.

L'unità di memoria analitica 22 del sistema 10 per l'identificazione di voci duplicate è ulteriormente configurata per memorizzare, ossia registrare, una pluralità di schemi di informazione tecnica, come detto predefiniti. Questi schemi di informazione tecnica sono raggruppati secondo le categorie della tassonomia standard. In pratica, la pluralità di schemi di informazione tecnica comprende vari gruppi di schemi di informazione tecnica, dove ogni gruppo è associato ad una rispettiva categoria della tassonomia standard. Si nota che uno stesso schema di informazione tecnica può appartenere ad più

gruppi, essendo quindi associato a più categoria della tassonomia standard.

Vantaggiosamente, il modulo di ricerca 16 è configurato per risolvere le possibili ambiguità nell'interpretazione della descrizione testuale del materiale (o oggetto) industriale compresa in ogni voce dell'anagrafica di materiali industriali, e quindi nel riconoscimento dello schema di informazione tecnica, questa risoluzione essendo basata sull'analisi statistica di un *corpus* di dati storici di una specifica tipologia.

Vantaggiosamente, l'unità di memoria analitica 22 del sistema 10 per l'identificazione di voci duplicate è ulteriormente configurata per memorizzare, ossia registrare, una pluralità di *corpus* di dati storici, ognuno relativo ad una specifica tipologia.

La categoria del materiale (o oggetto) industriale, selezionata dal modulo di categorizzazione 15, e le informazioni tecniche del materiale (o oggetto) industriale, scoperte ed estratte dal modulo di ricerca 16, sono fornite in ingresso al modulo di selezione 17, preferibilmente in forma strutturata.

Il modulo di selezione 17 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione è configurato per selezionare ed estrarre una pluralità di voci dell'anagrafica di materiali industriali, dove queste voci di anagrafica sono associate ad una comune categoria di materiale (o oggetto) industriale e dove le informazioni tecniche del materiale (o oggetto) industriale, di cui queste voci di anagrafica, sono identiche o equivalenti.

In una forma di realizzazione, la pluralità di voci dell'anagrafica di materiali industriali, selezionate ed estratte dal modulo di selezione 17, possono essere presentate ad un utente umano tramite opportuni mezzi di visualizzazione (non illustrati), come ad esempio uno schermo.

Vantaggiosamente, il modulo di selezione 17 è configurato per calcolare una metrica di valutazione della similarità tra ogni coppia di voci dell'anagrafica di materiali industriali, basata sulle rispettive categorie e soprattutto sulle rispettive informazioni tecniche dei

materiali (o oggetti) industriali, e per selezionare ed estrarre la pluralità di voci dell'anagrafica di materiali industriali, dove il valore della metrica di valutazione della similarità di queste voci di anagrafica è posizionato all'interno di un intervallo predefinito.

Questa metrica di valutazione della similarità permette di associare, ad ogni coppia di voci dell'anagrafica di materiali industriali, una misura di confidenza in merito alla possibilità che queste due voci si riferiscano ad un materiale (o oggetto) industriale identico o equivalente.

In una forma di realizzazione, la pluralità di voci dell'anagrafica di materiali industriali, selezionate ed estratte dal modulo di selezione 17, possono essere presentate ad un utente umano in ordine e/o raggruppate secondo il valore della metrica di valutazione della similarità.

Con particolare riferimento alla figura 2, il metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali, in particolari

materiali industriali, secondo la presente invenzione comprende i passi descritti di seguito.

Preferibilmente, al passo 32, il modulo di pre-analisi 14 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione scopre ed estrae almeno una feature del materiale industriale (in inglese *feature engineering*) dalla descrizione testuale compresa in ogni voce dell'anagrafica di materiali industriali.

Le *feature* del materiale (o oggetto) industriale, scoperte ed estratte al passo 32 dal modulo di pre-analisi 14, sono fornite in ingresso al modulo di categorizzazione 15, preferibilmente in forma strutturata.

Al passo 34, il modulo di categorizzazione 15 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione associa la descrizione testuale del materiale (o oggetto) industriale compresa in ogni voce dell'anagrafica di materiali industriali, e quindi la stessa voce

di anagrafica, ad una rispettiva categoria selezionata da una pluralità di categorie definite in una tassonomia standard. Ogni categoria della tassonomia standard rappresenta una rispettiva tipologia di materiale (o oggetto) industriale.

Preferibilmente, ancora al passo 34, il modulo di categorizzazione 15 associa la descrizione testuale del materiale (o oggetto) industriale, rappresentata sinteticamente dalle *feature* precedentemente scoperte ed estratte dal modulo di pre-analisi 14, ad una rispettiva categoria selezionata dalla pluralità di categorie definite nella tassonomia standard.

La categoria del materiale (o oggetto) industriale, selezionata al passo 34 dal modulo di categorizzazione 15, è fornita in ingresso al modulo di ricerca 16, preferibilmente in forma strutturata.

Al passo 36, il modulo di ricerca 16 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione scopre ed estrae almeno una informazione tecnica del materiale

industriale (come detto, ad esempio marca, modello, attributi, parametri tecnici, specifiche tecniche, codici di distribuzione) dalla descrizione testuale compresa in ogni voce dell'anagrafica di materiali industriali, tramite il riconoscimento di un rispettivo schema (in inglese *pattern-matching*) tra un gruppo di schemi di informazione tecnica associati alla categoria precedentemente selezionata dal modulo di categorizzazione 15. Questi schemi sono predefiniti, e ognuno di essi è associato ad almeno una categoria della tassonomia standard.

Vantaggiosamente, ancora al passo 36, il modulo di ricerca 16 risolve le possibili ambiguità nell'interpretazione della descrizione testuale del materiale (o oggetto) industriale compresa in ogni voce dell'anagrafica di materiali industriali, e quindi nel riconoscimento dello schema di informazione tecnica, questa risoluzione essendo basata sull'analisi statistica di un corpus di dati storici di una specifica tipologia.

La categoria del materiale (o oggetto) industriale, selezionata al passo 34 dal modulo di categorizzazione 15, e le informazioni tecniche

del materiale (o oggetto) industriale, scoperte ed estratte al passo 36 dal modulo di ricerca 16, sono fornite in ingresso al modulo di selezione 17, preferibilmente in forma strutturata.

Preferibilmente, al passo 38, il modulo di selezione 17 del sistema 10 per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo l'invenzione seleziona ed estrae una pluralità di voci dell'anagrafica di materiali industriali, dove queste voci di anagrafica sono associate ad una comune categoria di materiale (o oggetto) industriale e dove le informazioni tecniche del materiale (o oggetto) industriale, di cui queste voci di anagrafica, sono identiche o equivalenti.

Vantaggiosamente, ancora al passo 38, il modulo di selezione 17 calcola una metrica di valutazione della similarità tra ogni coppia di voci dell'anagrafica di materiali industriali, basata sulle rispettive categorie e soprattutto sulle rispettive informazioni tecniche dei materiali (o oggetti) industriali, e seleziona ed estrae la pluralità di voci dell'anagrafica di

materiali industriali, dove il valore della metrica di valutazione della similarità di queste voci di anagrafica è posizionato all'interno di un intervallo predefinito.

Questa metrica di valutazione della similarità permette di associare, ad ogni coppia di voci dell'anagrafica di materiali industriali, una misura di confidenza in merito alla possibilità che queste due voci si riferiscano ad un materiale (o oggetto) industriale identico o equivalente.

Si è in pratica constatato come la presente invenzione assolva pienamente il compito e gli scopi prefissati. In particolare, si è visto come il sistema e il metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali così concepiti permettono di superare i limiti qualitativi dell'arte nota, in quanto consentono di ottenere effetti migliori rispetto a quelli ottenibili con le soluzioni note e/o effetti analoghi a minor costo e con prestazioni più elevate.

Un vantaggio del sistema e del metodo per

l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione consiste nel fatto che essi permettono di individuare, in modo automatico ed euristico, un sottoinsieme delle voci di anagrafica che potenzialmente contiene dei duplicati, questo sottoinsieme essendo sufficientemente piccolo e preciso da rendere economicamente sostenibile il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio.

Un altro vantaggio del sistema e del metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione consiste nel fatto che essi consentono di supportare il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio, utilizzando

l'analisi linguistica del testo della descrizione testuale delle voci di anagrafica.

Un ulteriore vantaggio del sistema e del metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione consiste nel fatto che essi permettono di supportare il processo di valutazione manuale della qualità dei dati compresi nelle anagrafiche e di armonizzazione manuale delle eventuali voci duplicate, eseguito da utenti umani esperti di dominio, indipendentemente dalla lingua in cui sono scritti questi dati, in particolare la descrizione testuale delle voci di anagrafica.

Ancora, un vantaggio del sistema e del metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali secondo la presente invenzione consiste nel fatto che essi consentono di manutenere agevolmente anagrafiche di materiali comprendenti un numero di voci molto elevato (centinaia di migliaia o milioni, nel caso delle grandi aziende internazionali).

Benché il sistema e il metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali secondo l'invenzione sia stato concepito in particolare per la manutenzione di anagrafiche di materiali industriali in aziende di medie e/o grandi dimensioni, essi potranno comunque essere utilizzati, più generalmente, per la manutenzione di anagrafiche di materiali di qualsiasi tipologia in aziende di qualsiasi dimensione.

L'invenzione così concepita è suscettibile di numerose modifiche e varianti, tutte rientranti nell'ambito delle rivendicazioni allegate. Inoltre, tutti i dettagli potranno essere sostituiti da altri elementi tecnicamente equivalenti.

In pratica, i materiali impiegati, purché compatibili con l'uso specifico, nonché le dimensioni e le forme contingenti potranno essere qualsiasi a seconda delle esigenze e dello stato della tecnica.

In conclusione, l'ambito di protezione delle rivendicazioni non deve essere limitato dalle

illustrazioni o dalle forme di realizzazione preferite illustrate nella descrizione sotto forma di esempi, ma piuttosto le rivendicazioni devono comprendere tutte le caratteristiche di novità brevettabile che risiedono nella presente invenzione, incluse tutte le caratteristiche che sarebbero trattate come equivalenti dal tecnico del ramo.

# RIVENDICAZIONI

1. Sistema (10) per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali industriali, comprendente un'unità di memoria anagrafica (20) configurata per memorizzare detta anagrafica di materiali industriali comprendente una pluralità di voci, ogni voce di anagrafica comprendendo una descrizione testuale di un rispettivo materiale industriale,

caratterizzato dal fatto che comprende:

- un modulo di categorizzazione (15) configurato per associare detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi detta voce di anagrafica, ad una rispettiva categoria selezionata da una pluralità di categorie definite in una tassonomia standard e rappresentanti rispettive tipologie di materiale industriale;

- un modulo di ricerca (16) configurato per scoprire ed estrarre almeno una informazione tecnica di detto materiale industriale da detta descrizione testuale compresa in ogni voce di anagrafica, tramite il riconoscimento di un

rispettivo schema tra un gruppo di schemi di informazione tecnica associati a detta categoria selezionata da detto modulo di categorizzazione (15); e

- un'unità di memoria analitica (22) configurata per memorizzare detta tassonomia standard comprendente detta pluralità di categorie che rappresentano rispettive tipologie di materiale industriale, e una pluralità di schemi di informazione tecnica raggruppati secondo detta pluralità di categorie di detta tassonomia standard.

2. Sistema (10) per l'identificazione di voci duplicate secondo la rivendicazione 1, caratterizzato dal fatto che comprende ulteriormente un modulo di pre-analisi (14) configurato per scoprire ed estrarre almeno una *feature* di detto materiale industriale da detta descrizione testuale compresa in ogni voce di detta anagrafica di materiali industriali.

3. Sistema (10) per l'identificazione di voci duplicate secondo la rivendicazione 2, caratterizzato dal fatto che detto modulo di categorizzazione (15) è configurato per associare

detta descrizione testuale di detto materiale industriale, rappresentata sinteticamente da detta almeno una *feature* scoperta ed estratta da detto modulo di pre-analisi (14), a detta rispettiva categoria selezionata da detta pluralità di categorie di detta tassonomia standard.

4. Sistema (10) per l'identificazione di voci duplicate secondo una qualsiasi delle rivendicazioni precedenti, caratterizzato dal fatto che comprende ulteriormente un modulo di selezione (17) configurato per selezionare ed estrarre una pluralità di voci di detta anagrafica di materiali industriali, dove dette voci di anagrafica sono associate ad una comune categoria di materiale industriale e dove detta almeno una informazione tecnica di detto materiale industriale, di cui dette voci di anagrafica, è identica o equivalente.

5. Sistema (10) per l'identificazione di voci duplicate secondo la rivendicazione 4, caratterizzato dal fatto che detto modulo di selezione (17) è configurato per calcolare una metrica di valutazione della similarità tra ogni coppia di voci di detta anagrafica di materiali

industriali, e per selezionare ed estrarre detta pluralità di voci di detta anagrafica di materiali industriali dove il valore di detta metrica di valutazione della similarità di dette voci di anagrafica è posizionato in un intervallo predefinito.

6. Sistema (10) per l'identificazione di voci duplicate secondo una qualsiasi delle rivendicazioni precedenti, caratterizzato dal fatto che detto modulo di ricerca (16) è configurato per risolvere possibili ambiguità nell'interpretazione di detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi nel riconoscimento di detto schema di informazione tecnica, detta risoluzione essendo basata sull'analisi statistica di un *corpus* di dati storici di una specifica tipologia, detta unità di memoria analitica (22) essendo ulteriormente configurata per memorizzare una pluralità di *corpus* di dati storici, ognuno relativo ad una specifica tipologia.

7. Metodo per l'identificazione di voci duplicate, relative a materiali identici o equivalenti, in un'anagrafica di materiali

industriali, tramite:

- un'unità di memoria anagrafica (20) configurata per memorizzare detta anagrafica di materiali industriali comprendente una pluralità di voci, ogni voce di anagrafica comprendendo una descrizione testuale di un rispettivo materiale industriale; e

- un'unità di memoria analitica (22) configurata per memorizzare una tassonomia standard comprendente una pluralità di categorie che rappresentano rispettive tipologie di materiale industriale, e una pluralità di schemi di informazione tecnica raggruppati secondo detta pluralità di categorie di detta tassonomia standard;

caratterizzato dal fatto che comprende i passi che consistono nel:

- associare (34) detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi detta voce di anagrafica, ad una rispettiva categoria selezionata da detta pluralità di categorie definite in detta tassonomia standard e rappresentanti rispettive tipologie di materiale

industriale, tramite un modulo di categorizzazione (15); e

- scoprire ed estrarre (36) almeno una informazione tecnica di detto materiale industriale da detta descrizione testuale compresa in ogni voce di anagrafica, tramite il riconoscimento di un rispettivo schema tra un gruppo di schemi di informazione tecnica associati a detta categoria selezionata da detto modulo di categorizzazione (15), tramite un modulo di ricerca (16).

8. Metodo per l'identificazione di voci duplicate secondo la rivendicazione 7, caratterizzato dal fatto che comprende ulteriormente il passo che consiste nello scoprire ed estrarre (32) almeno una *feature* di detto materiale industriale da detta descrizione testuale compresa in ogni voce di detta anagrafica di materiali industriali, tramite un modulo di pre-analisi (14).

9. Metodo per l'identificazione di voci duplicate secondo la rivendicazione 8, caratterizzato dal fatto che, in detto passo di associare (34) detta descrizione testuale di detto

materiale industriale a detta rispettiva categoria selezionata da detta pluralità di categorie di detta tassonomia standard, detta descrizione testuale di detto materiale industriale è rappresentata sinteticamente da detta almeno una *feature* scoperta ed estratta da detto modulo di pre-analisi (14).

10. Metodo per l'identificazione di voci duplicate secondo una qualsiasi delle rivendicazioni precedenti, caratterizzato dal fatto che comprende ulteriormente il passo che consiste nel selezionare ed estrarre (38) una pluralità di voci di detta anagrafica di materiali industriali, dove dette voci di anagrafica sono associate ad una comune categoria di materiale industriale e dove detta almeno una informazione tecnica di detto materiale industriale, di cui dette voci di anagrafica, è identica o equivalente, tramite un modulo di selezione (17).

11. Metodo per l'identificazione di voci duplicate secondo la rivendicazione 10, caratterizzato dal fatto che detto passo di selezionare ed estrarre (38) detta pluralità di voci di detta anagrafica di materiali industriali

comprende ulteriormente il passo che consiste nel calcolare una metrica di valutazione della similarità tra ogni coppia di voci di detta anagrafica di materiali industriali, e selezionare ed estrarre detta pluralità di voci di detta anagrafica di materiali industriali dove il valore di detta metrica di valutazione della similarità di dette voci di anagrafica è posizionato in un intervallo predefinito, tramite detto modulo di selezione (17).

12. Metodo per l'identificazione di voci duplicate secondo una qualsiasi delle rivendicazioni precedenti, caratterizzato dal fatto che detto passo di scoprire ed estrarre (36) almeno una informazione tecnica di detto materiale industriale da detta descrizione testuale comprende ulteriormente il passo che consiste nel risolvere possibili ambiguità nell'interpretazione di detta descrizione testuale di detto materiale industriale compresa in ogni voce di anagrafica, e quindi nel riconoscimento di detto schema di informazione tecnica, tramite detto modulo di ricerca (16), detta risoluzione essendo basata sull'analisi statistica di un *corpus* di dati

storici di una specifica tipologia, detta unità di memoria analitica (22) essendo ulteriormente configurata per memorizzare una pluralità di *corpus* di dati storici, ognuno relativo ad una specifica tipologia.

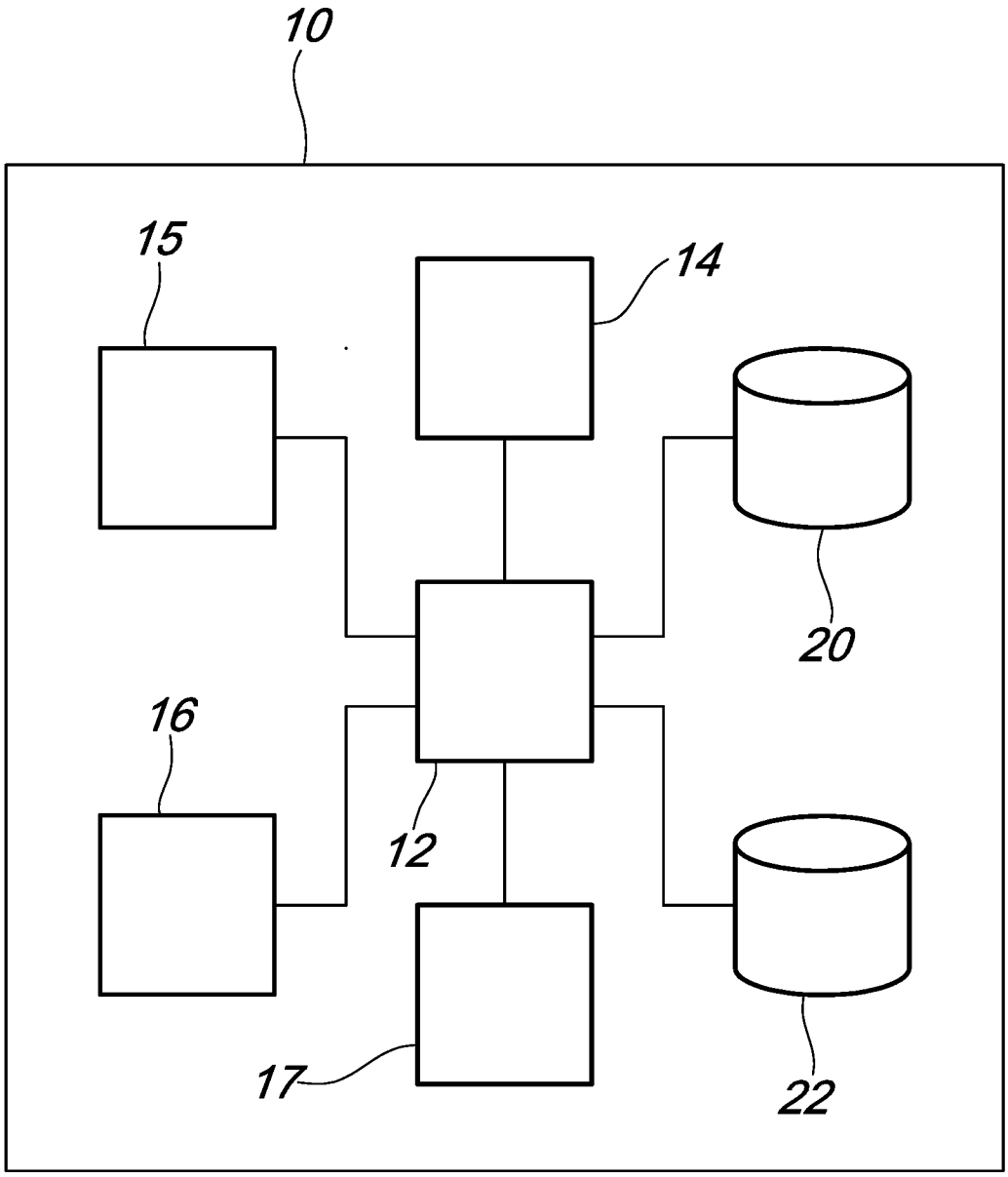


Fig. 1

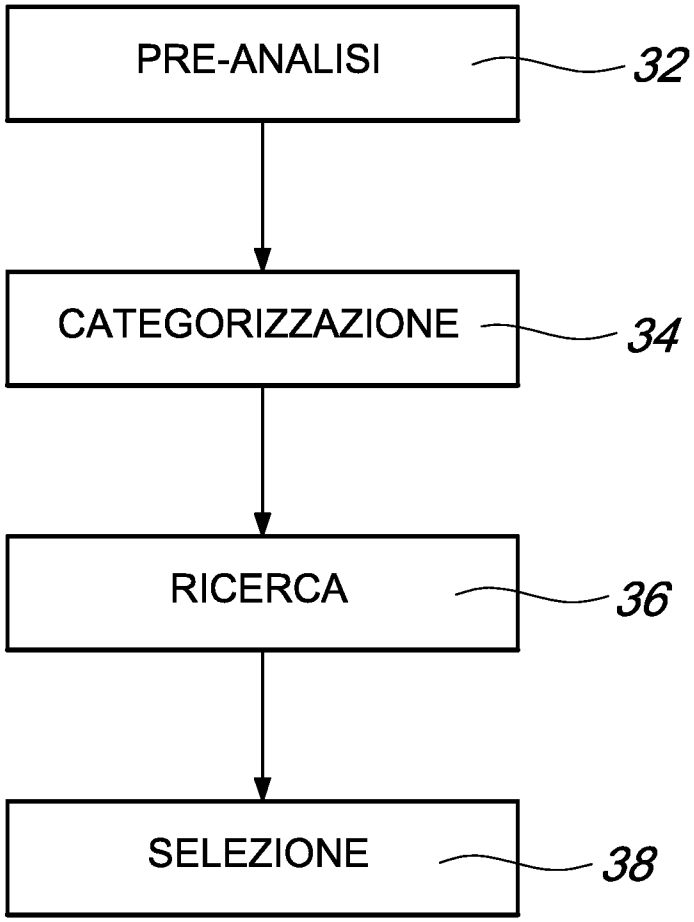
PRE-ANALISI
32
CATEGORIZZAZIONE
34
RICERCA
36
SELEZIONE
38

Fig. 2