



MINISTERO DELLO SVILUPPO ECONOMICO
DIREZIONE GENERALE PER LA LOTTA ALLA CONTRAFFAZIONE
UFFICIO ITALIANO BREVETTI E MARCHI

DOMANDA NUMERO	102006901477255
Data Deposito	19/12/2006
Data Pubblicazione	19/06/2008

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F		

Titolo

METODO DI CLASSIFICAZIONE DI PAGINE WEB E DI ORGANIZZAZIONE DEI CORRISPONDENTI CONTENUTI.
--



Domanda di brevetto per invenzione industriale dal titolo:

"Metodo di classificazione di pagine Web e di organizzazione dei corrispondenti contenuti"

a nome di: **REVAMPING S.r.l.**

5 con sede in: **Pavia**

* * * * *

001 2006 A 00 2436

DESCRIZIONE

Campo di applicazione

La presente invenzione si riferisce ad un metodo di
10 classificazione di pagine Web e di organizzazione dei corrispondenti
contenuti del tipo comprendente una fase di registrazione di una
pluralità di indirizzi Internet, comprendente esecuzioni di una pluralità
di procedure automatiche di registrazione, ed una fase di selezione per
l'impostazione di un corrispondente valore di pertinenza alla pluralità di
15 indirizzi Internet.

Arte nota

Come è ben noto, un personal computer connesso alla rete
Internet e dotato di un navigatore, permette di visualizzare pagine Web.

In particolare, con il termine personal computer si intende un
20 dispositivo elettronico dotato di una interfaccia di collegamento alla rete
Internet mentre con navigatore ci si riferisce ad un software, di per sé
noto, installato su tale dispositivo elettronico, per la visualizzazione
delle pagine Web. Un esempio di tale navigatore è il programma Internet
Explorer della società Microsoft®.

25 Ancor più in particolare, le pagine Web sono memorizzate su

una pluralità di server, a loro volta connessi alla rete Internet. Attraverso il navigatore si specifica un indirizzo Internet di una pagina Web, ad esempio in una porzione superiore o barra del navigatore, e si convalida tale indirizzo Internet per la visualizzazione della pagina Web
5 corrispondente, generalmente in una porzione principale del navigatore.

Sono altresì noti programmi o script di individuazione di una o più pagine Web funzionanti sulla base di un criterio di ricerca, impostato da un utente.

Tali programmi o script di individuazione sono generalmente
10 accessibili tramite una interfaccia grafica, ad esempio richiamabile dal navigatore, che comprende una porzione di inserimento del criterio di ricerca ed un pulsante di attivazione del programma o script di individuazione.

Dato il grande numero di pagine Web disponibili sulla rete
15 Internet, un programma o script di individuazione fornisce, per un predeterminato criterio di ricerca, un elevato numero di pagine Web.

Per guidare l'utente ad una efficace consultazione di tali pagine Web, il programma o script di individuazione è supportato da un metodo di classificazione, che assegna, a ciascuna pagina Web, una
20 corrispondente importanza. In altre parole l'esecuzione di un programma o script di individuazione fornisce all'utente un elenco di pagine Web, ordinate per valore di importanza stabilito dal metodo di classificazione.

Un noto metodo di classificazione, comunemente indicato con
25 il termine PageRank, prevede di associare ad una pagina Web un valore

di importanza comunemente indicato come rango [rank]. In particolare, il rango di una pagina Web A è determinato dal numero di ulteriori pagine Web B1, ..., Bx che fanno riferimento a tale pagina Web A, cioè che contengono un puntatore ad essa.

5 Più in particolare ancora, secondo il metodo di classificazione noto come PageRank, ogni pagina Web può esprimere un voto in merito ad altre pagine Web, attraverso puntatori, anche noti come collegamenti. In altre parole, attraverso i collegamenti disponibili su una pagina Web di partenza, un utente può collegarsi ad una o più pagine
10 Web destinazione, aumentandone il rango. Generalmente, maggiore è il grado di popolarità di una pagina Web, vale a dire il numero di collegamenti ad essa, maggiore è il valore dei voti che quella stessa pagina può esprimere.

I metodi di classificazione del tipo sopra indicato soffrono
15 tuttavia di svariate limitazioni che appariranno evidenti dalla descrizione che segue.

Con riferimento alla figura 1, si consideri una interfaccia 10 per l'attivazione di un programma o script di individuazione di pagine Web.

20 L'interfaccia 10 è ad esempio una pagina Web comprendente una porzione o barra di inserimento 1 di un criterio di ricerca ed un bottone di attivazione 2, per l'attivazione del programma o script di individuazione.

A scopo esemplificativo, in figura 1 è impostato un criterio di
25 ricerca per l'individuazione di pagine Web inerenti l'inventore e artista

italiano Leonardo da Vinci. In questo semplice esempio, il criterio di ricerca è dato dalla concatenazione delle tre parole chiave “Leonardo”, “da” e “Vinci”, specificate nella barra di inserimento 1 dell’interfaccia 10.

In tal caso, una porzione principale 4 dell’interfaccia 10 visualizza il risultato del programma o script di individuazione dopo una sua esecuzione, sulla base del criterio di ricerca “Leonardo da Vinci”. Il risultato della ricerca comprende un elenco 3 di indirizzi Internet di pagine Web, di cui solo le prime occorrenze 3a-3h sono visualizzate in tale porzione principale 4.

Infatti, essendo molto elevato il numero di pagine Web individuate, nell’esempio specificato nell’ordine di milioni di pagine Web, il programma o script di individuazione prevede un loro raggruppamento all’interno dell’interfaccia grafica 10 in successive pagine da visualizzare. Generalmente, una sequenza di numeri da 1 ad n, non rappresentato in figura perché di per sé convenzionale, è disponibile ad esempio in una porzione inferiore dell’interfaccia 10, al fine di visualizzare nella porzione principale 4, su richiesta dell’utente, una specifica pagina e quindi una porzione dell’elenco 3 di indirizzi Internet.

Ad esempio, supponendo che il programma o script di individuazione abbia individuato cento pagine Web, corrispondenti ad un elenco 3 di cento indirizzi Internet, e supponendo che ogni pagina di visualizzazione comprenda dieci indirizzi di tali pagine Web individuate, la porzione inferiore dell’interfaccia 10 potrebbe indicare numeri da uno a dieci. In tal modo, la selezione del numero uno nella porzione inferiore

consente di visualizzare, all'interno della porzione principale 4, gli indirizzi Internet da uno a dieci, cioè la prima pagina dell'elenco 3 di pagine Web. In particolare, negli attuali programmi o script di individuazione, gli indirizzi Internet dell'elenco 3 di pagine Web sono
5 messi in una successione corrispondente ad un ordine decrescente di importanza come individuato dal corrispondente metodo di classificazione e quindi la selezione della pagina uno dell'elenco 3 corrisponde alle pagine Web classificate come più importanti da tale metodo di classificazione.

10 La selezione del numero due consente di visualizzare, nella stessa porzione principale 4 dell'interfaccia 10, gli indirizzi Internet da undici a venti, corrispondenti ad una seconda pagina dell'elenco 3 di pagine Web. La selezione del numero dieci nella porzione inferiore, fa riferimento all'ultima pagina dell'elenco 3, associata quindi alle pagine
15 Web di minor rilievo, sempre secondo il metodo di classificazione utilizzato dal programma o script di individuazione.

In altre parole, l'ordine di visualizzazione dei risultati della ricerca deriva dal metodo di classificazione.

Ancor più in particolare, con riferimento alla figura 1, la
20 pagina Web 3a è quella di maggiore importanza tra tutte le pagine Web individuate dal metodo di classificazione utilizzato dal programma Google®, utilizzato come esempio per tale figura 1 e fa riferimento ad un servizio internet denominato Wikipedia, che fornisce la definizione dell'inventore ed artista Leonardo da Vinci.

25 La pagina Web 3b fa riferimento ad una pagina Web che

descrive l'eredità lasciata all'umanità dall'inventore ed artista Leonardo da Vinci; la pagina Web 3c si riferisce al museo nazionale della scienza e della tecnologia; la pagina Web 3e si riferisce all'aeroporto di Roma "Leonardo da Vinci".

5 Dai risultati sopra elencati del programma Google, che rappresentano le prime cinque occorrenze in ordine di importanza, si evince che il metodo di classificazione basato su PageRank soffre delle seguenti limitazioni.

10 Il numero di pagine Web fornite dal programma o script di classificazione è ridondante: in questo caso, nei primi otto risultati della prima pagina, ben tre pagine, in particolare le pagine 3d, 3f e 3h sono ridondanti, ripetono cioè indirizzi internet interni ad uno stesso sito precedentemente visualizzato, rispettivamente dagli indirizzi internet 3c, 3e e 3g. Inoltre, un utente non è legge generalmente più di qualche
15 decina di risultati forniti dal metodo di individuazione , risultato sovente ottenuto anche per criteri di ricerca piuttosto generici.

 Sebbene alcune pagine Web fornite dal programma o script di individuazione siano poco o per nulla pertinenti rispetto alle informazioni ricercate dall'utente, godono di un rango eccessivamente
20 elevato. Si consideri ad esempio la pagina Web 3e: è difficile pensare che un utente interessato all'aeroporto "Leonardo da Vinci", inserisca come criterio di ricerca "Leonardo da Vinci" senza specificare "Aeroporto". Tuttavia, inaspettatamente, la pagina Web 3e, che fa riferimento all'aeroporto di Roma, si trova come quinta occorrenza dei risultati del
25 programma Google. La pagina Web 3e costituisce un esempio tipico di

risultato affetto da rumore informativo e quindi altamente indesiderabile poiché non solo non pertinente ma anche inutile ai fini della ricerca svolta. E' possibile verificare che un tale risultato, vale a dire la presenza dell'indirizzo relativo all'Aeroporto di Roma in un posto di
5 rango elevato dell'elenco 3 di pagine Web risultante dall'impostazione del criterio di ricerca "Leonardo da Vinci" è ottenuto utilizzando molti dei programmi o script di ricerca attualmente disponibili.

Il metodo di classificazione basato su PageRank dunque, sebbene fornisca una organizzazione automatizzata delle pagine Web,
10 non consente all'utente di concentrarsi su un numero di risultati ridotto, limitatamente ai risultati per l'utente più pertinenti.

I metodi di classificazione di pagine Web sopra descritti soffrono inoltre di ulteriori problemi legati sostanzialmente alla affidabilità delle fonti intesa come la notorietà, la riconoscibilità e
15 credibilità della fonte rispetto al criterio di ricerca impostato. In altre parole, una pagina Web effettivamente correlata all'inventore ed artista Leonardo da Vinci, potrebbe godere di un rango elevato ma contenere informazioni errate: ad esempio informazioni erronee sulla sua data di nascita o prive di contenuti significativi, quali alcune delle sue opere più
20 famose od altre indicazioni fondamentali circa la sua vita ed il suo lavoro, che dovrebbero logicamente interessare un utente che ha impostato il criterio di ricerca "Leonardo da Vinci".

Quindi, una pagina Web di rango elevato potrebbe essere pertinente pur non godendo di sufficiente affidabilità della fonte, come
25 nel caso della pagina 3b relativa ad una pagina realizzata da un utente

privato non costituente una fonte autorevole.

Al contrario, una pagina Web correlata all'inventore Leonardo da Vinci potrebbe contenere cioè una pluralità di informazioni non strettamente correlate, sostanzialmente superflue, che rallentano
5 l'individuazione delle informazioni di interesse da parte dell'utente, che è un tipico esempio di informazione a bassa pertinenza. In sostanza, i metodi di classificazione di pagine web attualmente disponibili non consentono ad un programma o script di individuazione di effettuare una ricerca che abbia come risultato un pertinente, atteso e filtrato
10 elenco di pagine Web ma agiscono come collettori di una pluralità di informazioni con seppur limitata attinenza con un criterio di ricerca impostato dall'utente.

In tal modo, i metodi di classificazione noti costringono l'utente ad una laboriosa operazione di filtraggio e selezione,
15 sostanzialmente finalizzata ad individuare, tra tutte le pagine Web fornite dal programma o script di individuazione, le sole di interesse, con il rischio di non venire a conoscenza di pagine Web di considerevole importanza.

Il problema tecnico che sta alla base della presente invenzione
20 è quello di escogitare un metodo di classificazione di pagine Web che consenta ad un programma o script di individuazione di fornire un risultato in primo luogo privo di rumore informativo, pertinente al criterio di ricerca impostato dall'utente, non ridondante in quanto privo di pagine Web ripetute, ed affidabile per loro contenuto, al contempo
25 riducendo il tempo impiegato dall'utente per filtrare i risultati della

ricerca e risolvendo le limitazioni che ancora affliggono i metodi di classificazione di pagine Web secondo l'arte nota.

Sommario dell'invenzione

5 L'idea di soluzione che sta alla base della presente invenzione è quella di fornire un metodo di classificazione e di organizzazione dei corrispondenti contenuti in grado di determinare, a partire da una configurazione del Web per aree tematiche e lemmi, un insieme di pagine Web ritenute pertinenti, utilizzando come fonte sia differenti motori di ricerca, sia azioni di spidering diretto.

10 Le azioni di spidering sono eseguite verso siti ritenuti affidabili nella fase di configurazione del web.

Sulla base di questo insieme di pagine Web viene effettuata una selezione che elimini le pagine Web ridondanti, associate ad esempio a differenti traduzioni della stessa pagina Web o a uno stesso
15 indirizzo internet ripetuto più volte, e che elimini, attraverso un contributo umano, le pagine Web spurie, cioè contenenti rumore informativo e/o derivato da fonti poco affidabili.

In accordo con tale idea di soluzione, il metodo di classificazione di pagine Web e di organizzazione dei corrispondenti
20 contenuti secondo la presente invenzione comprende:

- una fase di registrazione di un prefissato numero di indirizzi Internet, comprendente esecuzioni di una pluralità di procedure automatiche di registrazione della pluralità di indirizzi Internet,
- una fase di selezione e di impostazione di un corrispondente
25 valore di pertinenza alla pluralità di indirizzi Internet,

caratterizzato dal fatto che

- la fase di selezione imposta il corrispondente valore di pertinenza ad un valore di pertinenza proporzionale ad una sua registrazione nella fase di registrazione e seleziona gli indirizzi internet con valore di pertinenza superiore ad un valore di soglia prefissato,

il metodo comprendendo ulteriormente

- una fase di riduzione che riduce, tra gli indirizzi internet a valore di pertinenza superiore a detto valore di soglia, agli indirizzi Internet rispondenti ad uno o più criteri di essenzialità, ed

- una fase di validazione per validare un sottoinsieme degli indirizzi Internet rispondenti ai criteri di essenzialità, la fase di validazione comprendendo una azione umana.

Vantaggiosamente, secondo il metodo di classificazione di pagine Web della presente invenzione un programma o script di individuazione è in grado di fornire un risultato pertinente ad un criterio di ricerca impostato da un utente, non ridondante in termini di ripetitività di pagine Web ed affidabile per loro contenuto, riducendo notevolmente il tempo impiegato dall'utente per filtrare i risultati della ricerca.

Ulteriori caratteristiche e vantaggi del metodo di classificazione secondo l'invenzione risulteranno dalla descrizione, fatta qui di seguito, di un suo esempio di realizzazione dato a titolo indicativo e non limitativo con riferimento ai disegni allegati.

Breve descrizione dei disegni

In tali disegni:

- la Figura 1 mostra un'interfaccia grafica per un programma o script di individuazione, secondo l'arte nota;

- la Figura 2 mostra un'interfaccia grafica per un programma o script di individuazione, secondo il metodo di classificazione della presente invenzione.

Descrizione dettagliata

Con riferimento alle figure allegate, la presente invenzione descrive un metodo di classificazione di pagine Web comprendente una pluralità di fasi.

In particolare, il metodo comprende una prima fase di configurazione che prevede la definizione di aree tematiche per la classificazione delle pagine Web.

A scopo esemplificativo e senza ridurre l'ambito di protezione dell'invenzione, per area tematica si intende una porzione del sapere, ad esempio la tecnica, l'arte, la letteratura, lo sport, l'attualità ed è definita da un insieme di parole descrittive di tale porzione del sapere.

La fase di configurazione comprende la definizione di una pluralità di lemmi per ciascuna area tematica. Inoltre, a ciascuna area tematica è associata una molteplicità di parole chiave di area, alcune delle quali vengono associate ai singoli lemmi. Un lemma appartenente a più di una area tematica è associato ad una pluralità di parole chiave di area appartenenti alle corrispondenti aree tematiche.

Ad esempio in fase di configurazione il criterio di ricerca "Leonardo da Vinci" è configurato come un lemma, al quale sono associate le parole chiave di area "Pittura", "Rinascimento", "Scultura"

appartenenti all'area tematica "Arte". Il lemma "Leonardo da Vinci" è altresì associato alle parole chiave di area "Canali", "Idraulica", appartenenti all'area tematica "Ingegneria".

5 Più in particolare ancora, in fase di configurazione è altresì definita una lingua principale, ad esempio l'italiano, per la creazione delle aree tematiche, dei lemmi e delle parole chiave di area ad esse correlate.

10 In fase di configurazione sono altresì impostate una pluralità di lingue ammesse, ad esempio contraddistinte da uno stesso alfabeto, ed utilizzate in una fase di riduzione come verrà spiegato successivamente. La lingua principale e le lingue ammesse permettono di effettuare ricerche semplici e complesse nella propria lingua, ottenendo anche pagine Web in un'altra lingua. In altre parole, impostato il lemma "Leonardo da Vinci", un risultato secondo il metodo
15 di classificazione permette di individuare non solo le pagine Web in lingua italiana, contenenti il lemma "Leonardo da Vinci" ma anche le pagine Web delle lingue ammesse in fase di configurazione.

Il metodo di classificazione secondo la presente invenzione comprende quindi, una fase di registrazione, per la memorizzazione di
20 un prefissato numero di indirizzi Internet di pagine Web, sulla base delle aree tematiche e dei corrispondenti lemmi definiti in fase di configurazione. Tale prefissato numero di indirizzi Internet di pagine Web è configurabile in fase di configurazione. In particolare, la fase di registrazione determina l'associazione di una pagina Web ad uno o più
25 lemmi, ad esempio tramite l'individuazione del lemma nella pagina Web.

Per ogni pagina Web associata ad un lemma, il suo indirizzo Internet è memorizzato.

5 Più in particolare ancora, l'esecuzione della fase di registrazione è eseguita tramite una pluralità di procedure automatizzate, ad esempio comprendenti una procedura di spidering.

10 Tale procedura di spidering, a partire da una pagina Web x, associata ad uno specifico lemma, registra gli indirizzi Internet della pagine Web z1, ..., zn, direttamente collegate a tale pagina Web x. La procedura di spidering è altresì in grado di registrare gli indirizzi Internet delle pagine Web y1, ..., yn, alle quali la pagina Web x è indirettamente collegata, tramite le pagine Web z1, ..., zn.

15 Vantaggiosamente, la procedura di spidering è configurabile per effettuare solo una registrazione diretta delle pagine Web, cioè delle pagine Web direttamente raggiungibili da una pagina Web x, o per effettuare una registrazione indiretta come sopra descritto.

20 La fase di registrazione prevede quindi di memorizzare gli indirizzi Internet, con riferimento alla specifico lemma, all'interno di un database, il cui contenuto rappresenta un giacimento informativo per il metodo di classificazione. In particolare, la pertinenza, la ridondanza, l'affidabilità ed il rumore informativo del giacimento informativo sono ottimizzati, secondo il metodo di classificazione dell'invenzione, come apparirà evidente dalla descrizione che segue.

25 Per chiarezza di esposizione, il giacimento informativo risultante dalla fase di registrazione verrà nel seguito indicato come primo giacimento informativo.

In particolare, la fase di registrazione comprende l'esecuzione di altre procedure automatizzate comprendenti, ad esempio, funzioni da metamatore di ricerca. Come noto, tali funzioni prevedono l'individuazione di una pluralità di pagine Web, appartenenti ad un lemma, tramite la consultazione di differenti motori di ricerca che classificano le pagine Web secondo un proprio metodo di classificazione.

In altre parole, le funzioni da metamatore di ricerca interrogano una pluralità di motori di ricerca per determinare, una pluralità di pagine Web comprendenti ad uno specifico lemma, già memorizzate da tali motori di ricerca, secondo una propria classificazione. Esempi di motori di ricerca noti sono Google, Yahoo, Altavista e simili.

Gli indirizzi internet delle pagine Web individuate dalle funzioni da metamatore di ricerca sono inserite nel primo giacimento informativo. Inoltre, la fase di registrazione memorizza nel primo giacimento informativo una copia della pagina Web. Tale copia è utilizzata sia per consultazione sia per confronto con una sua versione successiva, individuata da una successiva azione di spidering al fine di rendere disponibile un aggiornamento della pagina Web stessa all'interno del primo giacimento informativo.

In accordo con la presente invenzione, il metodo di classificazione comprende ulteriormente una fase di selezione che esegue il confronto degli indirizzi Internet memorizzati dalle procedure automatizzate in fase di registrazione.

In particolare, gli indirizzi Internet registrati tramite la

procedura di spidering, per un certo lemma, sono confrontati con gli indirizzi Internet registrati tramite le funzioni da metamatore di ricerca. Se un indirizzo Internet è stato individuato sia dalla procedura di spidering o nella fase di configurazione, sia dalle funzioni da metamatore di ricerca, è convalidato in tale fase di selezione. Infatti, la sua presenza in differenti motori di ricerca e la sua individuazione in fase di configurazione è indicativa di una alta probabilità di pertinenza del suo contenuto, in riferimento ad un lemma specifico.

La fase di registrazione comprende vantaggiosamente diverse tipologie di procedure automatizzate, non necessariamente basate su procedure di spidering o o nella fase di configurazione o su funzioni da metamatore di ricerca. Sulla base delle registrazioni effettuate da tutte le procedure automatizzate, la fase di selezione effettua il confronto degli indirizzi Internet individuati, effettuando un matching statistico, cioè assegnando agli indirizzi Internet un valore di probabile pertinenza e affidabilità tanto più alto quanto più è frequente la sua registrazione tramite le procedure automatizzate.

Secondo la fase di selezione della presente invenzione, il primo giacimento informativo è raffinato in un secondo giacimento informativo che comprende gli indirizzi Internet aventi un valore di probabile pertinenza ed affidabilità sopra una soglia prestabilita. Il secondo giacimento informativo è ad esempio contenuto nel database di cui sopra o in un database separato.

Il secondo giacimento informativo, sebbene comprenda indirizzi Internet a probabile alto valore di pertinenza, rappresenta un

contenuto potenzialmente ridondante. Proprio in virtù del probabile alto valore di pertinenza assegnato in fase di selezione, diversi indirizzi Internet potrebbero fare riferimento alla stessa pagina Web, ad esempio a più versioni della stessa pagina Web, tradotta in diverse lingue oppure a versioni non aggiornate, ma ancora pubbliche, della stessa pagina Web oppure allo stesso contenuto informativo memorizzato in pagine Web appartenenti a differenti domini.

Al fine di perfezionare il contenuto del secondo giacimento informativo, il metodo di classificazione secondo l'invenzione comprende quindi una fase di riduzione degli indirizzi Internet convalidati in fase di selezione.

Tale fase di riduzione prevede che ogni pagina Web, indirizzata da un indirizzo Internet del secondo giacimento informativo, rispetti un criterio di essenzialità.

Tale criterio di essenzialità prevede ad esempio l'eliminazione di tutte le pagine Web non corrispondenti ad una lingua ammessa ed impostata in fase di configurazione. Inoltre, il criterio di essenzialità comprende l'eliminazione di pagine Web provenienti per lo stesso lemma da un identico dominio, riportando alla pagina principale tra esse.

La fase di riduzione, verificando il soddisfacimento del criterio di essenzialità, stabilisce la presenza del corrispondente indirizzo Internet in un terzo giacimento informativo, ulteriormente raffinato.

Vantaggiosamente, un programma o script di individuazione eseguito sul terzo giacimento informativo è in grado di individuare indirizzi Internet di pagine Web ad alta probabilità di pertinenza,

eliminando sostanzialmente la ridondanza.

5 A scopo esemplificativo e non limitativo, un programma o script di individuazione basato sul criterio di ricerca “Leonardo da Vinci”, supportato dal metodo di classificazione della presente invenzione, è in grado di determinare già in questa fase poche centinaia di indirizzi Internet, ad alta probabilità di pertinenza e sostanzialmente non ridondanti, mentre un programma o script di individuazione supportato da un metodo di classificazione secondo l’arte nota individua un numero di indirizzi Internet nell’ordine dei milioni nella selezione
10 finale.

Secondo la presente invenzione, è prevista una ulteriore fase di validazione sul terzo giacimento informativo, finalizzata a determinare i soli indirizzi Internet che si riferiscono a pagine Web altamente affidabili e sostanzialmente prive di rumore informativo.

15 Tale fase di validazione comprende l’esecuzione di una interfaccia di validazione che elenca gli indirizzi Internet del terzo giacimento informativo. Più in particolare, l’interfaccia di validazione propone un ordinamento di tali indirizzi Internet secondo una valore di probabile pertinenza già assegnato dalle precedenti fasi di registrazione, di selezione e di riduzione.
20

Attraverso l’interfaccia di validazione, un operatore analizza le pagine Web associate agli indirizzi Internet e convalida quelli pertinenti, rispetto ad uno specifico lemma. Più specificatamente, l’operatore elimina, attraverso l’interfaccia grafica e con l’ausilio di procedure automatiche di confronto e segnalazione dei risultati, gli indirizzi
25

internet a bassa o nulla affidabilità, non pertinenti ed indesiderati, ad esempio evitando che l'indirizzo internet che fa riferimento all'aeroporto "Leonardo da Vinci" sia inserito nel quarto giacimento informativo.

5 Vantaggiosamente l'apporto umano è sostanziale per la definizione di un quarto giacimento informativo, derivante da una selezione manuale degli indirizzi Internet del terzo giacimento informativo.

10 Dunque, secondo il metodo di classificazione della presente invenzione, la classificazione delle pagine Web comprende almeno le seguenti fasi.

A) Una fase di registrazione di un prefissato numero di indirizzi Internet, comprendente esecuzioni di una pluralità di procedure automatiche di registrazione di detta pluralità di indirizzi Internet..

15 B) Una fase di selezione, per l'impostazione di un corrispondente valore di pertinenza alla pluralità di indirizzi Internet; tale fase di selezione imposta il corrispondente valore di probabile pertinenza ad un valore proporzionale ad una sua registrazione in detta fase di registrazione e seleziona gli indirizzi internet con valore di
20 pertinenza superiore ad un valore di soglia prefissato.

C) Una fase di riduzione che elimina, tra indirizzi internet a valore di pertinenza superiore al valore di soglia, gli indirizzi Internet non rispondenti ad uno o più criteri di essenzialità.

25 D) Una fase di validazione per validare un sottoinsieme degli indirizzi Internet rispondenti ai criteri di essenzialità. Tale fase di

validazione comprende una azione manuale, eseguita da un utente.

Ad esempio, sul criterio di ricerca “Leonardo da Vinci”, un programma o script di individuazione secondo la presente invenzione è in grado di determinare una alcune decine di indirizzi Internet, come
5 rappresentato in figura 2.

In particolare, la figura 2 rappresenta un programma o script di individuazione di pagine Web, richiamabile da una interfaccia 101, ad esempio da una pagina Web comprendente una porzione o barra di inserimento 11 di un criterio di ricerca ed un bottone di attivazione 21,
10 per l’attivazione del programma o script di individuazione.

In figura 2 è impostato un criterio di ricerca per l’individuazione di pagine Web inerenti l’inventore e artista italiano Leonardo da Vinci.

Una porzione principale 41 dell’interfaccia 101 visualizza il
15 risultato del programma o script di individuazione dopo una sua esecuzione, sulla base del criterio di ricerca “Leonardo da Vinci”. In particolare, il risultato della ricerca comprende un elenco 31 di indirizzi Internet di pagine Web, di cui solo le prime occorrenze 3a1-3c1 sono visualizzate.

Vantaggiosamente, il programma o script di individuazione è
20 in grado di fornire all’utente finale alcune decine di indirizzi Internet, che identificano pagine Web molto pertinenti, affidabili e senza rumore e ridondanza. L’indirizzo Internet 3a1 si riferisce ad una pagina Web della BBC, l’indirizzo Internet 3b1 al Metropolitan Museum e l’indirizzo
25 Internet 3c1 ad una rubrica dedicata dalla Rai, nota emittente televisiva

italiana, all'inventore Leonardo da Vinci.

Il programma o script di individuazione dunque scarta una pluralità di pagine Web che hanno bassa pertinenza con il criterio di ricerca impostato dall'utente, proponendo solo le pagine Web appartenenti al quarto giacimento informativo.

Per una ricerca di tipo classico, l'interfaccia 101 permette di eseguire il programma o lo script di individuazione su un giacimento informativo meno raffinato rispetto al quarto giacimento informativo, ad esempio sul secondo giacimento informativo.

Ad esempio, selezionando il pulsante 61 in figura 2, il programma o script di individuazione fornisce un elenco di indirizzi Internet contenuti nel secondo giacimento informativo, permettendo di consultare cioè anche le pagine Web sulle quali la fase di validazione umana non è stata eseguita.

Vantaggiosamente, l'interfaccia 101 comprende inoltre una porzione laterale 42 che propone informazioni di carattere enciclopedico, in formato ipertestuale, correlate ad un criterio di ricerca impostato. Ad esempio, per il criterio di ricerca "Leonardo da Vinci", la porzione laterale 42 propone la Biografia, la Cronologia della vita, la Cronologia del tempo, I codici di Leonardo, la Bibliografia di Leonardo.

Vantaggiosamente, secondo il metodo di classificazione di pagine Web della presente invenzione un programma o script di individuazione è in grado di fornire un risultato pertinente, quindi senza rumore, ad un criterio di ricerca impostato da un utente, non ridondante in termini di numerosità e ripetitività di pagine Web ed

affidabile per loro contenuto, riducendo notevolmente il tempo impiegato dall'utente per filtrare i risultati della ricerca.

RIVENDICAZIONI

1. Metodo di classificazione di pagine Web e di organizzazione dei corrispondenti contenuti del tipo comprendente

5 - una fase di registrazione di un prefissato numero di indirizzi Internet, comprendente esecuzioni di una pluralità di procedure automatiche di registrazione di detta pluralità di indirizzi Internet,

10 - una fase di selezione e di impostazione di un corrispondente valore di pertinenza a detta pluralità di indirizzi Internet,

caratterizzato dal fatto che

15 detta fase di selezione imposta detto corrispondente valore di pertinenza ad un valore di pertinenza proporzionale ad una sua registrazione in detta fase di registrazione e seleziona gli indirizzi internet con valore di pertinenza superiore ad un valore di soglia prefissato,

il metodo comprendendo ulteriormente

20 - una fase di riduzione che riduce, tra detti indirizzi internet a valore di pertinenza superiore a detto valore di soglia, agli indirizzi Internet rispondenti ad uno o più criteri di essenzialità, ed

 - una fase di validazione per validare un sottoinsieme di detti indirizzi Internet rispondenti ai criteri di essenzialità, detta fase di validazione comprendendo una azione umana.

25 2. Metodo di classificazione secondo la rivendicazione 1 caratterizzato dal fatto di comprendere una fase di configurazione per

una definizione di una pluralità di aree tematiche ed una definizione di una pluralità di parole chiave di area ad esse associate.

3. Metodo di classificazione secondo la rivendicazione 2 caratterizzato dal fatto che detta fase di configurazione comprende una
5 definizione di una pluralità di lemmi, per ciascuna di dette aree tematiche.

4. Metodo di classificazione secondo la rivendicazione 3 caratterizzato dal fatto che detta fase di configurazione comprende una associazione di uno di detti lemmi ad una o più di dette aree tematiche.

10 5. Metodo di classificazione secondo la rivendicazione 4 caratterizzato dal fatto che detta fase di configurazione comprende una definizione di una lingua principale, per dette definizioni di dette aree tematiche, di detti lemmi e di dette parole chiave di area.

6. Metodo di classificazione secondo la rivendicazione 5
15 caratterizzato dal fatto che detta fase di configurazione comprende una impostazione di detto prefissato numero di indirizzi internet.

7. Metodo di classificazione secondo la rivendicazione 6 caratterizzato dal fatto che detta fase di registrazione comprende una fase di associazione di detti lemmi a dette pagine Web.

20 8. Metodo di classificazione secondo la rivendicazione 7 caratterizzato dal fatto che detta fase di associazione comprende una individuazione di detti lemmi in dette pagine Web.

9. Metodo di classificazione secondo la rivendicazione 8
25 caratterizzato dal fatto che detta fase di registrazione comprende una memorizzazione di detti indirizzi Internet, con riferimento a detti lemmi,

in un primo giacimento informativo.

5 10. Metodo di classificazione secondo la rivendicazione 9
caratterizzato dal fatto che detta fase di registrazione comprende una
memorizzazione di una copia di dette pagina Web in detto primo
giacimento informativo.

 11. Metodo di classificazione secondo la rivendicazione 10
caratterizzato dal fatto che detto criterio di essenzialità comprende una
eliminazione di una pluralità di pagine Web, non comprendenti alfabeti
di dette lingue ammesse.

10 12. Metodo di classificazione secondo la rivendicazione 11
caratterizzato dal fatto che detto criterio di essenzialità comprende una
eliminazione di una pluralità di pagine Web provenienti da un identico
dominio.

15 13. Metodo di classificazione secondo la rivendicazione 12
caratterizzato dal fatto che detta fase di riduzione comprende esecuzioni
di una pluralità di query di selezione.

 14. Metodo di classificazione secondo la rivendicazione 13
caratterizzato dal fatto che detta fase di validazione comprende
un'esecuzione di una interfaccia di validazione.

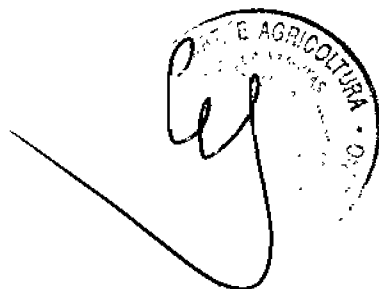
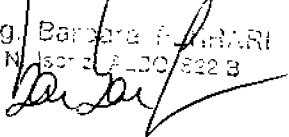
20 15. Metodo di classificazione secondo la rivendicazione 14
caratterizzato dal fatto che detta interfaccia di validazione elenca detti
indirizzi internet ridotti da detta fase di riduzione.

 16. Metodo di classificazione secondo la rivendicazione 15
caratterizzato dal fatto che detta fase di registrazione comprende
25 esecuzioni di almeno una procedura di spidering.

17. Metodo di classificazione secondo la rivendicazione 16 caratterizzato dal fatto che detta fase di registrazione comprende esecuzioni di almeno una funzione da metamotores di ricerca.

5 18. Metodo di esecuzione di un programma o script di individuazione per una individuazione di una pluralità di indirizzi Internet rispondenti ad un criterio di ricerca caratterizzato dal fatto che detta pluralità di indirizzi Internet è classificata tramite un metodo di classificazione secondo una delle precedenti rivendicazioni, detti indirizzi Internet avendo un valore di pertinenza sopra ad una prefissata
10 soglia.

Ing. Barbara Ferrari
N. Iscr. ALBO 822 B



1/2

Google

Web Immagini Gruppo News altro...

1

10

2

Cerca

Recherche avancée

Préférence

3

Web

3a **Leonardo da Vinci - Wikipedia**
Nel 1502 Leonardo da Vinci produsse il disegno di un ponte a campata unica di 300 metri, come parte di un progetto di ingegneria civile per il Sultano ...
it.wikipedia.org/wiki/Leonardo_da_Vinci - 165k - [Copia cache](#) - [Pagine simili](#)

3b **Leonardo da Vinci - Home page**
L'eredità di Leonardo da Vinci è un patrimonio universale; pensare come un genio.
freeweb.supereva.com/floben/index.htm - 10k - [Copia cache](#) - [Pagine simili](#)

3c **Museo di Vinci**
www.leonet.it/comuni/vinci/ - 1k - [Copia cache](#) - [Pagine simili](#)

3d **IL MUSEO LEONARDIANO DI VINCI**
Museo Leonardiano di Vinci Castello dei Conti Guidi I 50053 VINCI - FI - Tel: +39-571-56055
Direttore del Museo e della Biblioteca: Dr. Romano Nanni E-mail: ...
www.leonet.it/comuni/vincimus/vincimus.html - 5k - [Copia cache](#) - [Pagine simili](#)
[Altri risultati in www.leonet.it]

3e **Museo Nazionale della Scienza e della Tecnologia "Leonardo da Vinci"**
Nel museo si possono percorrere alcune visite "virtuali" in 3D sulle opere di Leonardo.
www.museoscienza.org/ - 13k - [Copia cache](#) - [Pagine simili](#)

3f **Leonardo da Vinci al Museo della Scienza**
Propone la cronologia della vita dell'artista, le invenzioni, i manoscritti, le macchine, le opere e alcune risorse.
www.museoscienza.org/leonardo/default.asp - 9k - [Copia cache](#) - [Pagine simili](#)
[Altri risultati in www.museoscienza.org]

3g **ADR - Sito ufficiale Aeroporti di Roma -**
Aeroporto Intercontinentale Leonardo da Vinci di Fiumicino e Giovan Battista Pastine di Ciampino. Le infrastrutture, la mappa, i voli e le tariffe.
www.adr.it/ - 53k - [Copia cache](#) - [Pagine simili](#)

3h **ADR - Sito ufficiale Aeroporti di Roma - Fiumicino**
Aeroporto Leonardo da Vinci di Fiumicino via dell'Aeroporto di Fiumicino, 00060 Fiumicino (RM) tel: +39 06 85951 Terminal A = Voli Nazionali ...
www.adr.it/content.asp?L=1&idmen=199 - 23k - 16 nov 2006 - [Copia cache](#) - [Pagine simili](#)

4

FIG.1

Ing. Barbara ...
N. ...

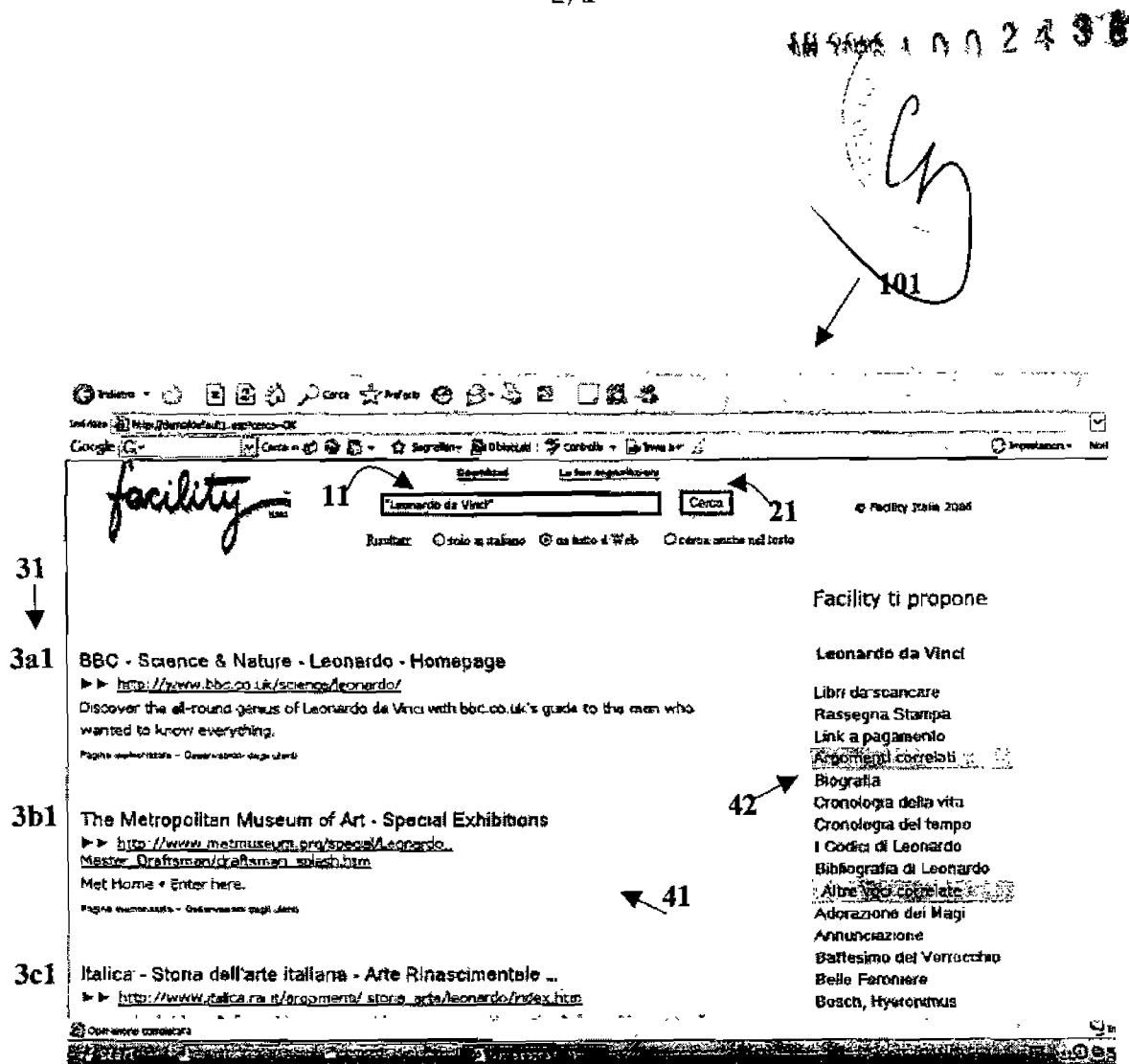


FIG.2

Ing. Barbara CERRARI
10/10/2004

Barbara Cerrari