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

UIBM

DOMANDA DI INVENZIONE NUMERO	102021000027983
Data Deposito	03/11/2021
Data Pubblicazione	03/05/2023

Classifiche IPC

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F	16	332

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F	16	33

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F	16	9532

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F	16	9535

Titolo

METODO PER LA RICERCA E LA CLASSIFICAZIONE DI DOCUMENTI TECNICI MEMORIZZATI SU UNA PLURALITA? DI DATABASE ACCESSIBILI PER VIA TELEMATICA

TITOLO: METODO PER LA RICERCA E LA CLASSIFICAZIONE DI DOCUMENTI TECNICI MEMORIZZATI SU UNA PLURALITA' DI DATABASE ACCESSIBILI PER VIA TELEMATICA

DESCRIZIONE

Settore Tecnico dell'Invenzione

L'ambito di applicazione della presente invenzione riguarda la ricerca di documenti pubblicati sulla rete "*Internet*".

Da quando, nel 2001, Lawrence Page pubblicò l'algoritmo (cfr. US 6285999 B1 – "Method for node ranking in a linked database") su cui si basa il primo motore di ricerca di "*Google*", il problema di ricercare i contenuti di interesse di un utente nella imponderabile mole di documenti che sono pubblicati sulla rete "*Internet*", è diventato un problema sempre più difficile e complesso.

Infatti, è sempre più concreto il rischio di non riuscire ad individuare, e quindi reperire, i documenti che più si avvicinano a ciò che si cerca: cosicché documenti potenzialmente di grande interesse, pur presenti in qualche sito pubblicato in rete, è possibile che sfuggano alle ricerche.

Questo rischio è significativo, in modo particolare, anche nel settore scientifico e tecnologico. La documentazione che tratta di innovazioni tecnologiche, o comunque la documentazione che fornisce indicazioni di interesse sull'esistenza di eventuali innovazioni di cui si è alla ricerca, costituisce un ambito di ricerca estremamente vasto; e le informazioni che potrebbero essere più interessanti spesso non emergono con evidenza a seguito di ricerche documentali svolte con le tecniche note.

Tecnica Nota

Dai tempi delle prime ricerche di documentazione pubblicata sulla rete "*Internet*" che potevano essere eseguite vent'anni fa, fino ad arrivare ai giorni nostri, lo scenario di riferimento ha subito numerosi cambiamenti sostanziali, sia di tipo quantitativo che di tipo qualitativo.

I cambiamenti quantitativi sono evidenti. Il numero di documenti che vengono continuamente caricati in qualche sito della grande rete "*Internet*" (spesso si usa la locuzione anglofona "*Big Internet*") aumenta in modo assolutamente imponderabile. Qualsiasi argomento può essere citato o trattato, a molteplici livelli di approfondimento, in una quantità di documenti che non può essere realisticamente proposta al vaglio diretto di un operatore che esegue una specifica ricerca.

Oltre agli evidenti cambiamenti quantitativi, si registrano poi sostanziali cambiamenti qualitativi. Questi ultimi riguardano sia le tipologie di documenti, e sia i linguaggi utilizzati nella documentazione.

Informazioni di potenziale interesse possono essere sintetizzate in classici articoli o rapporti tecnici, oppure in documenti organizzati a schede, o in siti che ospitano dei "*blog*", oppure ancora in presentazioni multimediali in cui sono presenti anche contenuti audio e video.

Ma non è solo la tipologia di documento ad essere così varia da rendere difficili le ricerche automatiche; infatti, essendo possibile dedurre i criteri di ricerca utilizzati dai principali programmi di ricerca automatica (i così detti "motori di ricerca"), è sempre più diffusa la pratica di affidare la stesura finale dei documenti che vengono pubblicati in rete ad esperti in tecniche SEO (Search Engine Optimization).

Come il nome stesso lascia intuire, gli esperti in tecniche SEO introducono nei documenti pubblicati in rete delle particolari formulazioni (o accorgimenti di altro tipo) che intercettano in modo ottimizzato i criteri con cui i motori di ricerca assegnano un giudizio di pertinenza del documento analizzato, alla ricerca condotta.

Per effetto della larga diffusione delle tecniche SEO, i motori di ricerca tendono a far emergere i documenti redatti da chi si avvale dei migliori esperti SEO, ovvero i documenti che vengono pubblicati con fini commerciali.

Al contrario delle entità che pubblicano con fini commerciali, i soggetti che pubblicano con disinteressati fini divulgativi, o informativi in genere, non investono denaro ed energie per ottimizzazioni di tipo SEO, tipicamente pubblicano rivolgendosi a ristrette cerchie di contatti che appartengono a comunità che possono anche essere relativamente chiuse, precludendo, di fatto, la circolazione dei contenuti da loro pubblicati presso il pubblico più vasto.

Allo stato dell'arte, l'esecuzione di una ricerca di contenuti pubblicati sulla rete "*Internet*" è un'attività che può essere definita di tipo semi-professionale; infatti, viene spesso affidata a dei professionisti. Quando viene condotta da soggetti non professionali, o si tratta comunque di persone con competenze molto forti, animate da una passione per l'argomento (che, in molti casi, possono anche avere competenze migliori di quelle dei professionisti nel settore), oppure, quando le ricerche sono condotte da persone non esperte nell'ambito delle ricerche documentali, normalmente, finiscono per far emergere per lo più documenti promossi con fini commerciali o informazioni già molto note.

Difficilmente ricerche condotte da soggetti che non hanno una specifica competenza ed esperienza nella ricerca di informazioni in rete portano a scoprire informazioni

innovative ma poco conosciute, e quando ciò avviene, non è detto che si tratti delle informazioni che più rispondono alle esigenze che hanno ispirato la ricerca.

Questo genere di problema è molto penalizzante nel contesto delle ricerche di soluzioni tecnologiche perché, normalmente, ciò che è interessante scoprire, quanto più possibile precocemente, sono proprio le soluzioni e le idee nuove, possibilmente prima che queste diventino di comune dominio negli specifici settori.

Un possibile approccio al problema nella sua generalità consiste nell'ottimizzare sempre di più il funzionamento dei motori di ricerca.

Questo approccio mira a migliorare la capacità di tali motori di ricerca di analizzare velocemente e correttamente l'enorme quantità di dati pubblicati sulla rete "*Internet*". Oppure offre strumenti di interazione con l'operatore che esegue la ricerca in modo da guidarlo per rendere più efficaci e pertinenti le informazioni che egli fornisce al motore di ricerca stesso, focalizzando meglio il merito dell'oggetto della ricerca.

Un esempio, tra i molti, di questo approccio è indicato in US 11048756 B1 ["Search Engine Optimizer" – Paiz, Richard – 29 giugno 2021], in cui si insegna ad ottimizzare i pattern di ricerca, e ad organizzare meglio le informazioni reperite durante il vaglio dei documenti presenti nella rete "*Internet*", in modo da far lavorare meglio un motore di ricerca.

Tuttavia, benché si possano concepire motori di ricerca capaci di prestazioni sempre migliori, dotati di interfacce che facilitino l'operatore che esegue la ricerca, o che si avvalgono di supercomputer sempre più potenti, o che implementano algoritmi di ricerca che vengono continuamente arricchiti di accorgimenti che li rendono sempre più accurati, le ricerche della documentazione che meglio risponde ai motivi per cui ogni singola ricerca viene eseguita non sono quasi mai pienamente

soddisfacenti; ed è tutt'altro che improbabile che documenti di interesse reale, in relazione alla ricerca condotta, non vengano trovati.

Del resto, le ragioni, sopra espresse, per cui molte ricerche non sono soddisfacenti, permangono a prescindere dalla qualità dei motori di ricerca: infatti la documentazione disponibile è in continua crescita, si differenzia in tipologie sempre diverse e comprende documenti che sono costruiti appositamente per essere evidenziati dalle logiche implementate dai motori di ricerca.

Queste caratteristiche dell'ambiente di ricerca continuano a rappresentare ostacoli quasi insormontabili per giungere ad una piena automatizzazione delle ricerche, le quali, pertanto, continuano sempre a necessitare di un'attenta supervisione da parte di un esperto.

E tale difficoltà di ricerca, come già evidenziato, è certamente presente, e spesso in modo particolarmente significativo, anche nel caso in cui i documenti ricercati siano relativi ad innovazioni tecnologiche che contengano informazioni utili alla soluzione di particolari esigenze di tipo tecnico. Soprattutto, è molto difficile individuare notizie e documenti che si riferiscono alle soluzioni più innovative e meno note che provengono dai più svariati ambienti di studio o da piccole aziende.

L'abilità dell'operatore che esegue una ricerca, e la sua esperienza nelle ricerche, è pertanto alla base di ricerche di questo tipo, ed è tuttora essenziale.

Una questione aperta riguarda fino a che punto sia possibile automatizzare i processi di ricerca, e come si debbano impostare le ricerche usando i classici motori di ricerca, affinché gli automatismi eseguano efficacemente le ricerche volute. Infatti, per il solo fatto che la ricerca di documenti pubblicati sulla rete "*Internet*" avviene su una base di documenti il cui numero è dell'ordine dei miliardi implica il

fatto che sia necessario condurre queste ricerche avvalendosi di programmi di ricerca automatica.

La mole di documenti da analizzare, inoltre, implica anche che tali programmi di ricerca possano avvalersi di potenze di calcolo molto elevate: fatto, quest'ultimo, che suggerisce di usare motori di ricerca generalisti (come ad esempio *Google*), potendo, questi motori, contare su potenze di calcolo adatte ad analizzare velocemente, e con la dovuta complessità di analisi, tutti i documenti pubblicati nella grande rete "*Internet*".

L'abilità dell'operatore che esegue la ricerca è quella di impostare le chiavi di ricerca, combinandole in modo possibilmente intelligente, in modo da estrarre una piccola quantità di documenti, senza rischiare di perdere in tale estrazione selettiva i documenti più interessanti.

Benché l'esperienza degli operatori che eseguono le ricerche venga sicuramente in aiuto in questo compito, e benché la tecnica nota metta a disposizione soluzioni che facilitano la migliore definizione delle chiavi di ricerca (si pensi, a titolo d'esempio, alla già citata soluzione indicata in US 11048756 B1 – Paiz R. "Search Engine Optimizer"), la pratica nota non dispone di metodologie e strumenti che possano essere considerati pienamente soddisfacenti; e, quando qualcuno esegue una ricerca in "*Internet*", è sempre più diffusa la convinzione (o la sensazione) che, nascosto da qualche parte, vi sia un documento che tratta proprio quello che interessa maggiormente, ma che questo documento non sia stato trovato.

In definitiva, ogni metodo, o accorgimento, volti a migliorare i risultati di una ricerca documentale condotta sulla grande rete "*Internet*", rappresenta un risultato di grande interesse, con ricadute certamente positive anche nel contesto della circolazione virtuosa dell'innovazione.

Scopo e sintesi dell'invenzione

Lo scopo generale della presente invenzione è quello di indicare un metodo, e gli opportuni strumenti, che consentono di condurre una ricerca di documenti sulla grande rete "*Internet*", in cui sia ridotto il più possibile il rischio di trascurare documenti di interesse significativo.

In particolare, la presente invenzione concentra la propria attenzione nell'ambito della documentazione che fornisce indicazioni su innovazioni tecnologiche funzionali ad offrire soluzioni a particolari esigenze tecniche.

Un obiettivo intermedio, ma essenziale al conseguimento dell'obiettivo generale, consiste nel valorizzare al massimo il quesito della ricerca espresso con un linguaggio tipico del soggetto interessato, il quale, in genere è esperto dell'argomento tecnico che è l'oggetto della ricerca, ma non è esperto delle tecniche di ricerca documentale.

Tipicamente avviene che sia già la riduzione dell'oggetto di ricerca ad un insieme (o a pochi insiemi) di parole chiave (che dovrebbero sintetizzare ciò che si cerca) a costituire una prima causa per cui, nel processo di ricerca documentale, si perde informazione potenzialmente utile.

Gli scopi prefissati per questa invenzione sono raggiunti mediante il ricorso ad un metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*", che si avvale di un sistema di strumenti di tipo informatico che comprendono:

- almeno un motore di ricerca documentale generalista implementato attraverso un programma noto ed adeguati mezzi di calcolo idonei ad eseguirlo,

- almeno un programma che implementa un algoritmo NLP (Natural Language Processing) di tipo TF-IDF (Term Frequency – Inverse Document Frequency), con relativo programma di addestramento,
- un insieme preselezionato di librerie informatiche che contengono raccolte di documenti contenenti informazioni tecniche;

e detto metodo di ricerca comprende i seguenti passi:

- ✓ formulazione di una relazione di ingresso che sintetizza le informazioni tecniche che si intendono ricercare tra i documenti pubblicati nella grande rete "*Internet*", essendo detta relazione di ingresso comprendente almeno un testo redatto da un soggetto richiedente, esperto nella materia della ricerca, e scritto utilizzando il linguaggio ed il gergo di esposizione naturale di tale soggetto richiedente;
- ✓ esecuzione di una ricerca di documenti che trattano lo stesso argomento trattato in detta relazione di ingresso redatta da detto soggetto richiedente, essendo detta ricerca limitata ai documenti tecnici contenuti in dette librerie informatiche preselezionate;
- ✓ popolamento di un insieme di documenti di riferimento, costituito dai documenti selezionati attraverso la ricerca eseguita al punto precedente;
- ✓ utilizzo di detto insieme di documenti di riferimento, popolato al punto precedente, per istruire detto algoritmo NLP di tipo TF-IDF;
- ✓ individuazione di un insieme di parole chiave che rappresentano l'oggetto della ricerca richiesta da detto soggetto richiedente;
- ✓ esecuzione di una ricerca automatica e massiva, condotta su tutta la grande rete "*Internet*", utilizzando detto motore di ricerca documentale generalista, essendo detto motore di ricerca documentale generalista applicato

utilizzando una pluralità di diverse combinazioni di dette parole chiave individuate al passo precedente;

- ✓ raccolta di tutti i documenti selezionati attraverso la ricerca automatica e massiva eseguita al punto precedente, indipendentemente dalla valutazione di pertinenza assegnato da detto motore di ricerca documentale generalista a ciascun documento selezionato, e costruzione di un insieme di documenti genericamente focalizzato sull'argomento di ricerca;
- ✓ comparazione di ciascun documento raccolto in detto insieme di documenti genericamente focalizzato con detta relazione di ingresso, essendo detta comparazione effettuata mediante l'applicazione dell'algoritmo NLP (di tipo TF-IDF) precedentemente istruito;
- ✓ selezione di un sottoinsieme di detto insieme di documenti genericamente focalizzati, essendo detto sottoinsieme composto da documenti specificatamente rilevanti, ed essendo detta selezione effettuata in funzione di detta comparazione condotta al passo precedente.

Il principale vantaggio della presente invenzione consiste pertanto nel fatto che i suoi insegnamenti permettono di soddisfare i principali obiettivi per cui l'invenzione stessa è stata concepita.

Questa invenzione presenta anche ulteriori vantaggi, che risulteranno più evidenti dalla descrizione seguente, che illustra ulteriori dettagli dell'invenzione stessa attraverso alcune forme di implementazione, e dalle rivendicazioni allegate, che formano parte integrante della presente descrizione.

Descrizione dettagliata

Alla base dell'invenzione vi è l'intuizione di articolare la ricerca di documenti separando due funzioni che nelle tecniche di arte nota sono sempre intimamente accorpate: cioè la funzione di passare al vaglio tutti (o quasi tutti) i documenti pubblicati nella grande rete "*Internet*", e la funzione di analisi approfondita dei contenuti per verificarne l'interesse.

Il vaglio massivo di tutti i documenti pubblicati nella grande rete "Internet", viene condotto servendosi di uno strumento costituito da un motore di ricerca documentale generalista, sfruttando, di quest'ultimo, la grande potenza e velocità di calcolo.

In questa fase viene sostanzialmente ignorata un'effettiva analisi di pertinenza dettagliata, perché si tollera di selezionare un numero di documenti elevato e non gestibile da un operatore, costruendo però un insieme di documenti genericamente focalizzato sull'argomento di ricerca.

Tale insieme di documenti genericamente focalizzati, pur essendo un insieme troppo numeroso per costituire un primo esito di ricerca, è però di una dimensione assolutamente gestibile con un secondo procedimento di analisi automatizzata, basato su algoritmi, anche di una certa complessità computazionale, quali sono algoritmi di analisi di linguaggio naturale (detti algoritmi NLP – Natural Language Processing).

Un passo iniziale della ricerca documentale deve necessariamente essere condotto sulla totalità dei documenti pubblicati nella grande rete "*Internet*". Questa ricerca può essere condotta all'inizio, proprio come prima cosa, o anche parallelamente o immediatamente dopo altre fasi della metodologia, dalle quali non è comunque strettamente dipendente.

In questa fase iniziale non è importante isolare subito i documenti di interesse: ciò che conta è ridurre l'ambiente di ricerca focalizzandolo solo su documenti che attengono genericamente all'argomento della ricerca. Al tempo stesso, però, si deve scongiurare il rischio di escludere, in tale processo di focalizzazione, documenti che potrebbero essere interessanti, demandando la ricerca vera e propria a fasi successive, che però hanno il vantaggio di poter essere condotte su un insieme di documenti meno imponderabile, e certamente gestibile ricorrendo anche a programmi di analisi di una certa complessità.

Dunque la prima ricerca avviene con un classico motore di ricerca documentale generalista che agisce eseguendo ricerche sulla base del riscontro di combinazioni di parole chiave.

Quando la ricerca deve focalizzare rapidamente i documenti di maggior interesse, la scelta delle combinazioni di parole chiave è di grande rilevanza, e da tale scelta spesso dipende la qualità dell'esito della ricerca. Applicando il metodo secondo l'invenzione, invece, questa scelta non è così determinante, il solo scopo è quello di selezionare tutti i documenti potenzialmente di interesse, accettando anche di selezionare documenti ridondanti. Pertanto la scelta delle parole chiave può essere richiesta direttamente al soggetto cui interessa la ricerca, anche se costui non ha grande esperienza in ricerche documentali sulla grande rete "*Internet*". L'importante è che vengano fornite parole chiave attinenti all'argomento. La ricerca massiva viene poi condotta eventualmente arricchendo l'insieme di parole chiave con sinonimi in uso nel settore e lanciando una pluralità di analisi, ciascuna con una diversa combinazione di tali parole chiave.

Il numero di tali combinazioni è in genere molto elevato, dato che, come già detto, in questa fase si accetta di reperire un gran numero di documenti, anche dell'ordine dei milioni.

L'informazione più importante che viene chiesta al soggetto cui interessa la ricerca, tuttavia, non è una lista di parole chiave, bensì è una relazione scritta che sintetizza le informazioni tecniche che si intendono ricercare.

Tale relazione è detta relazione di ingresso. In essa, il soggetto richiedente si deve esprimere usando il linguaggio naturale, ed il gergo, tipico del settore in cui opera, ossia linguaggio e gergo condivisi anche nell'ambiente in cui i contenuti ricercati possono essere generati.

Sia i contenuti tecnici espressi nella relazione d'ingresso e sia il linguaggio usato costituiscono informazioni di ingresso preziose per automatizzare, come sarà chiarito nel seguito della presente descrizione, il processo di ricerca e per arrivare ad un numero di documenti relativamente ridotto, gestibile con analisi molto approfondite, ed ordinabile in funzione di effettiva pertinenza rispetto all'oggetto della ricerca: potendo lavorare solo su documenti su cui ne vale la pena.

In una forma di implementazione interessante, tale relazione di ingresso può anche costituire l'unico punto di partenza dell'intera ricerca, essendo possibile estrarre dalla relazione stessa anche le parole chiave.

In generale, è comunque raccomandabile che tale relazione di ingresso comprenda alcune informazioni che sono di grande importanza per la ricerca: tra queste, l'obiettivo della soluzione tecnica di interesse del soggetto richiedente e l'indicazione delle tecnologie sulle quali si ritiene che sia possibile basare la soluzione cercata. In una forma di implementazione preferita, queste ultime informazioni è opportuno che siano esplicitamente enunciate in parti di testo isolate

e chiaramente identificabili come recanti le suddette informazioni. Ciò allo scopo di facilitare il trattamento automatizzato del testo di detta relazione di ingresso.

Si sottolinea a questo punto che il metodo indicato nella presente invenzione si applica preferibilmente per ricercare documenti che illustrino, o che forniscano informazioni importanti, relativamente a soluzioni tecnologiche innovative; cioè soluzioni tecnologiche che possano interessare a soggetti richiedenti che sono alle prese con un problema tecnico, e che vogliono documentarsi in merito alla possibile presenza di soluzioni già individuate, ed utili alla soluzione del loro problema.

Contestualmente alla ricerca massiva effettuata passando al vaglio tutta la grande rete "*Internet*", o in sequenza a tale ricerca massiva, prima o dopo di essa, il metodo secondo l'invenzione prevede di selezionare un insieme ulteriore di documenti, detti documenti di riferimento sull'argomento.

Questa selezione è peculiare del metodo secondo l'invenzione, ha una funzione distintiva rispetto ad altri metodi, e presenta alcune caratteristiche che la contraddistinguono.

La prima caratteristica è che si tratta di una selezione condotta su un numero preselezionato di librerie informatiche contenenti raccolte di documenti il cui contenuto comprenda informazioni tecniche. Per dare un ordine di grandezza del numero di tali librerie preselezionate si può indicare il migliaio, anche se, evidentemente, molto dipende dalla disponibilità e dalla dimensione delle singole librerie. L'importante è che si tratti di archivi che contengono esempi di documenti di qualità, che trattano argomenti riguardanti soluzioni tecnologiche.

La limitazione di applicabilità del metodo secondo l'invenzione, alla ricerca di documenti di tipo tecnologico, permette di preselezionare queste librerie informatiche in modo tale che queste rappresentino una buona base di partenza

dove reperire documenti che trattano lo stesso argomento trattato nei documenti che si intendono ricercare e, soprattutto, documenti che fanno uso di linguaggi e gerghi tipicamente usati nei settori da cui provengono.

Un insieme di documenti di riferimento che trattino il medesimo argomento della ricerca è necessario per disporre di un buon numero di esempi di documenti redatti, per l'appunto, in un linguaggio usato nel settore della ricerca.

Tali documenti, forniscono quindi lo strumento essenziale per addestrare un algoritmo di elaborazione del linguaggio naturale (cioè un algoritmo NLP - Natural Language Processing) e per ricavare un glossario di parole specifiche usate nel contesto tecnico in cui la ricerca documentale deve essere condotta.

Un numero di documenti adeguato a questo scopo è dell'ordine delle varie decine, potendo anche arrivare all'ordine del centinaio.

In una forma di implementazione di particolare interesse, il glossario ricavato da detti documenti di riferimento può essere usato anche per determinare le parole chiave da impiegare per condurre la ricerca massiva di cui si è già detto. In tal caso, evidentemente, questa fase di determinazione dell'insieme dei documenti di riferimento precede la ricerca massiva.

Tuttavia la funzione essenziale di detto insieme di documenti di riferimento è quello di addestrare un algoritmo NLP per elaborare testi che trattano argomenti affini alla ricerca da condurre.

La selezione dei documenti di riferimento tratti da dette librerie predefinite può, anche questa, essere condotta con tecniche note per individuare documenti simili tra loro; tra le tecnologie che possono essere impiegate, si citano a titolo di esempio tecnologie basate sulla "*Keyword Density*" o sullo "*Indice di Jaccard*".

L'uso di un insieme di addestramento che contiene solo documenti che trattano l'argomento di interesse del soggetto richiedente, costituisce una garanzia convincente sul fatto che l'algoritmo NLP potrà essere applicato con buona affidabilità ai documenti che dovranno poi essere ricercati. Si ricorda che anche sugli algoritmi NLP, e sulle relative tecniche di addestramento a partire da esempi di testi redatti usando un particolare linguaggio naturale, o un particolare gergo, esiste un'ampia letteratura alla quale si rimanda, essendo questa una tematica scientifica che può ormai considerarsi matura.

A questo punto il metodo secondo l'invenzione dispone di due importanti risultati raggiunti.

Un primo risultato è quello di disporre di un insieme di documenti genericamente focalizzato sull'argomento di ricerca ottenuto facendo girare un motore di ricerca classico su tutta la grande rete "*Internet*". Questo insieme focalizzato è molto numeroso, infatti può contenere milioni di documenti, ma, al contempo, si può essere abbastanza certi del fatto che difficilmente sono stati trascurati documenti di potenziale interesse.

Il secondo risultato è quello di disporre di un algoritmo NLP opportunamente addestrato, che nelle forme di implementazione preferite è di tipo TF-IDF (Term Frequency – Inverse Document Frequency), con cui analizzare automaticamente tutti i documenti appartenenti all'insieme focalizzato. Questi ultimi, infatti, pur essendo molto numerosi, sono in numero trattabile effettuando comparazioni via software (ad esempio anche con il succitato algoritmo NLP/TF-IDF).

Il processo di selezione dei documenti effettivamente di interesse avviene quindi confrontando ciascun documento appartenente a detto insieme genericamente focalizzato con la relazione di ingresso che sintetizza le informazioni tecniche che

si intendono ricercare. Si ricorda, infatti, che anche detta relazione di ingresso è formulata nel linguaggio naturale usato dal soggetto interessato alla richiesta; costui, auspicabilmente, è una persona relativamente esperta dell'argomento, e pertanto, normalmente, usa un linguaggio naturale (cioè un lessico) affine a quello usato nei documenti su cui avviene la comparazione.

Il processo di comparazione può avvenire, al solito, con svariate tecniche note; in alcune forme di implementazione tra le preferite si possono ancora una volta indicare le tecniche basate sullo "*Indice di Jaccard*" (particolarmente significativo per questo genere di analisi).

Ciò che conta ai fini della presente invenzione è che attraverso tali comparazioni si giunga a selezionare, a partire dal grande insieme focalizzato, un sottoinsieme di poche migliaia di documenti, offrendo così un risultato di ricerca che può finalmente essere considerato gestibile dal soggetto richiedente.

Il processo di riduzione della documentazione selezionata può essere condotto in più passaggi, restringendo progressivamente i vincoli di contenuto che vengono imposti. In tali (uno o più) passaggi di comparazione è possibile far ricorso a tecniche di elaborazione del linguaggio naturale sempre più sofisticate.

Anche sul risultato della ricerca costituito dall'insieme dei documenti specificatamente rilevanti che emergono dalle comparazioni con la relazione di ingresso redatta dal soggetto richiedente, infine, possono essere applicati ulteriori metodi di classificazione per portare in ulteriore evidenza i risultati più pertinenti per il richiedente.

Una forma di classificazione interessante consiste nel produrre un ordinamento secondo un ordine di pertinenza alla esigenza espressa da detto soggetto richiedente nella relazione di ingresso. Alcune tecnologie a cui si può far ricorso per

questi passi di analisi finali fanno riferimento ai così detti algoritmi di "*Text classification*" e di "*URL classification*".

Varianti e Considerazioni conclusive

In sintesi, gli insegnamenti della presente invenzione hanno come finalità quella di definire in maniera innovativa ed efficace un metodo per reperire documenti pubblicati nella grande rete "*Internet*" che descrivono soluzioni tecnologiche di interesse in relazione a determinati problemi tecnici, ma che, ad oggi, sfuggono alle tecniche di ricerca note.

Dato che non si può prescindere da metodi di ricerca che siano in larga parte implementabili in modo automatico (a causa soprattutto delle grandissima mole di documenti pubblicati sulla rete "*Internet*"), l'invenzione indica come sia possibile ricorrere ad una pluralità di ricerche condotte applicando una pluralità di metodi informatici noti, diversi tra loro, concepiti con scopi e finalità diverse ed indipendenti tra loro, ma che possono funzionare sinergicamente nel contesto della metodologia insegnata.

In particolare, notevole esercizio di attività inventiva risiede nel fatto di aver definito un passo di ricerca finalizzato non tanto a ricercare contenuti di interesse, ma per raccogliere documenti caratterizzati dall'uso di un linguaggio ed un gergo particolari; e questo per disporre di uno strumento per addestrare un algoritmo adattivo (nella fattispecie un algoritmo NLP) da usare per effettuare analisi su una base dati costituita da documenti selezionati a loro volta in modo opportuno, ed usando un motore di ricerca in modo particolare (non badando cioè a ridurre la dimensione numerica dell'esito della ricerca, anzi, facendo il contrario).

Inoltre, l'invenzione può facilmente applicarsi anche a ricerche su documenti multimediali aggiungendo un'ulteriore fase metodologica, ad esempio ricorrendo a

strumenti di conversione da audio a testo. In questo modo è possibile riportare alla tipologia di documenti testuali anche tutti i documenti audio, rendendoli così trattabili secondo le metodiche sopra esposte.

A questo proposito, si osserva che i più moderni motori di ricerca allo stato dell'arte sono in grado di evidenziare anche pagine "*web*", pubblicate in rete, che comprendono contenuti multimediali: è pertanto giusto confrontare anche tali contenuti con la relazione descrittiva che riassume le esigenze del soggetto richiedente la ricerca.

È chiaro che il metodo insegnato attinge a piene mani ad algoritmi noti, sia per la ricerca documentale, che per l'analisi di testi, che per la loro classificazione. Tali algoritmi sono in continua evoluzione, cosicché nuovi algoritmi (anche non citati nella presente relazione descrittiva) con prestazioni sempre migliori possono essere adottati nell'implementazione della presente invenzione, dando così luogo a diverse varianti della medesima invenzione.

Del resto, l'esercizio di attività inventiva non è stato profuso nella messa a punto di tecniche elaborative o algoritmi di ricerca, bensì nel combinare in modo innovativo strumenti informatici noti che, usati secondo il metodo descritto portano a risultati di notevole interesse applicativo. Il metodo secondo l'invenzione, quindi, trae beneficio dall'eventuale (nonché auspicato) miglioramento delle prestazioni dei singoli strumenti informatici di cui si avvale.

In particolare, la comprensione del linguaggio naturale è una delle principali frontiere dell'intelligenza artificiale, cosicché il processo di comparazione tra la relazione di ingresso ed i documenti appartenenti all'insieme focalizzato, estratto massivamente dalla grande rete "*Internet*", è certamente passibile di molti affinamenti che riguardano sia la comprensione delle esigenze del soggetto richiedente

(producendo così così un impatto significativo sul processo di analisi della così detta relazione di ingresso) e sia la comprensione dei documenti appartenenti all'insieme genericamente focalizzato, così da proporre ricerche capaci di correlare anche gruppi di documenti, surrogando sempre più, e sempre meglio, tutte le analisi che ancora si pensa di lasciare all'analisi diretta del soggetto richiedente, così da proporgli documenti sempre più focalizzati e sempre meglio organizzati per favorire e velocizzare quella che alla fine dovrà essere la sua analisi personale.

È importante quindi ribadire, in tutta generalità, che la presente invenzione si presta a numerose varianti, pur mantenendo le prerogative rivendicate.
Di alcune di queste varianti si è anche già detto, e riguardano l'uso di differenti algoritmi, o di differenti ordini di esecuzione delle varie fasi del metodo; tuttavia, tali varianti, qualora rispettino le caratterizzazioni essenziali del metodo insegnato, devono essere considerate tutte varianti della medesima invenzione.

Soprattutto nel contesto di scenari evolutivi poi, l'invenzione si presta ad incorporare e supportare ulteriori sforzi di sviluppo e perfezionamento, capaci di migliorare, o aumentare, le prestazioni del metodo descritto.
Quindi, sviluppi ulteriori potrebbero essere apportati dall'uomo esperto del ramo senza per questo fuoriuscire dall'ambito dell'invenzione quale essa risulta dalla presente descrizione e dalle rivendicazioni qui allegate che costituiscono parte integrante della presente descrizione; oppure, qualora detti sviluppi non risultino compresi nella presente descrizione, possono essere oggetto di ulteriori domande di brevetto associate alla presente invenzione, o dipendenti da essa.

TITOLO: METODO PER LA RICERCA E LA CLASSIFICAZIONE DI DOCUMENTI TECNICI MEMORIZZATI SU UNA PLURALITA' DI DATABASE ACCESSIBILI PER VIA TELEMATICA

RIVENDICAZIONI

1. Un metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*", che si avvale di un sistema di strumenti di tipo informatico che comprendono:

   i. almeno un motore di ricerca documentale generalista implementato attraverso un programma noto ed adeguati mezzi di calcolo idonei ad eseguirlo,

   ii. almeno un programma che implementa un algoritmo NLP (Natural Language Processing) di tipo TF-IDF (Term Frequency – Inverse Document Frequency), con relativo programma di addestramento,

   iii. un insieme preselezionato di librerie informatiche che contengono raccolte di documenti contenenti informazioni tecniche;

   e detto metodo di ricerca comprende i seguenti passi:

   a. formulazione di una relazione di ingresso che sintetizza le informazioni tecniche che si intendono ricercare tra i documenti pubblicati nella grande rete "*Internet*", essendo detta relazione di ingresso comprendente almeno un testo redatto da un soggetto richiedente, esperto nella materia della ricerca, e scritto utilizzando il linguaggio ed il gergo di esposizione naturale di tale soggetto richiedente;

   b. esecuzione di una ricerca di documenti che trattano lo stesso argomento trattato in detta relazione di ingresso redatta da detto soggetto

richiedente, essendo detta ricerca limitata ai documenti tecnici contenuti in dette librerie informatiche preselezionate;

c. popolamento di un insieme di documenti di riferimento, costituito dai documenti selezionati attraverso la ricerca eseguita al precedente passo "b.";

d. utilizzo di detto insieme di documenti di riferimento, popolato come indicato al precedente passo "c", per istruire detto algoritmo NLP di tipo TF-IDF;

e. individuazione di un insieme di parole chiave che rappresentano l'oggetto della ricerca richiesta da detto soggetto richiedente;

f. esecuzione di una ricerca automatica e massiva, condotta su tutta la grande rete "*Internet*", utilizzando detto motore di ricerca documentale generalista, essendo detto motore di ricerca documentale generalista applicato utilizzando una pluralità di combinazioni di dette parole chiave individuate al precedente passo "e.";

g. raccolta di tutti i documenti selezionati attraverso la ricerca automatica e massiva eseguita al precedente passo "f.", indipendentemente dalla valutazione di pertinenza assegnato da detto motore di ricerca documentale generalista a ciascun documento selezionato, e costruzione di un insieme di documenti genericamente focalizzato sull'argomento di ricerca;

h. comparazione di ciascun documento raccolto in detto insieme di documenti genericamente focalizzato con detta relazione di ingresso, essendo detta comparazione effettuata mediante l'applicazione di detto

algoritmo NLP di tipo TF-IDF, istruito come indicato al precedente passo "d.";

i. selezione di un sottoinsieme di detto insieme di documenti genericamente focalizzati, essendo detto sottoinsieme composto da documenti specificatamente rilevanti, ed essendo detta selezione effettuata in funzione di detta comparazione condotta al precedente passo "h.".

2. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 1, in cui detta relazione di ingresso è strutturata in modo da isolare, in parti di testo identificabili,

- l'enunciazione dell'obiettivo della soluzione tecnica di interesse del soggetto richiedente e
- l'indicazione delle tecnologie sulle quali si ritiene che sia possibile basare la soluzione cercata.

3. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 1, in cui l'individuazione dell'insieme di parole chiave che rappresentano l'oggetto della ricerca richiesta da detto soggetto richiedente, ed usate per detta ricerca automatica e massiva, condotta su tutta la grande rete "*Internet*", sono ricavate elaborando detta relazione di ingresso.

4. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 1, in cui l'individuazione l'insieme di parole chiave che rappresentano

l'oggetto della ricerca richiesta da detto soggetto richiedente, ed usate per detta ricerca automatica e massiva, sono richieste esplicitamente a detto soggetto richiedente.

5. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 4, in cui detta ricerca automatica e massiva è condotta contemporaneamente alla, o prima della, ricerca di detti documenti di riferimento in dette librerie informatiche preselezionate, indicata ai punti "a.", b." e "c." della rivendicazione 1.

6. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 1, che si avvale anche di programmi di conversione di contenuti audio in contenuti testuali.

7. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 1, in cui detto sottoinsieme composto da documenti specificatamente rilevanti, selezionato come indicato al punto "i." della rivendicazione 1, è ordinato secondo un ordine di pertinenza alla esigenza espressa da detto soggetto richiedente nella relazione di ingresso.

8. Metodo di ricerca documentale per individuare informazioni a contenuto tecnologico pubblicate nella grande rete "*Internet*" secondo la rivendicazione 7, in cui detto ordinamento viene effettuato utilizzando algoritmi di "*Text classification*" e di "*URL classification*".