Document clustering e nuovi motori di ricerca

Una prospettiva basata sull'analisi per concetti e la categorizzazione

Barbara Fiorentini

Biblioteca dell'Università Cattolica
Piacenza
barbara.fiorentini@unicatt.it

I motori di ricerca sono uno strumento prezioso nelle mani dell'utente che si serve di Internet per cercare dati, informazioni e documenti. La nuova frontiera è rappresentata dai motori "intelligenti": oltre a selezionare le pagine web, ordinandole per rilevanza, le classificano per argomento, suggerendo i percorsi di ricerca e di approfondimento in base all'argomento desiderato e supportando il ricercatore come farebbe un esperto del settore in carne e ossa.
O quasi. A fronte di evidenti vantaggi, i motori di nuova generazione presentano infatti anche alcuni rischi che, se tenuti in debito conto, possono essere fronteggiati e superati, in modo da poter trarre il massimo beneficio da strumenti di ricerca innovativi e utili in vari campi, dall'economia alla finanza, dal marketing all'informazione e alla documentazione.

I motori di ricerca: presente e futuro prossimo

I motori di ricerca rappresentano lo strumento principale per l'utente che desidera cercare dati o informazioni in Internet. Dai dati di mercato alle analisi di settore, dalle informazioni varie su fatti, problemi e persone fino agli argomenti più futili e giocosi: in Internet c'è di tutto.[1] Il problema del ricercatore è come trovare le informazioni necessarie per soddisfare una determinata domanda, cercando di districarsi in un *mare magnum* di dati spesso fuorvianti, ridondanti o comunque troppo numerosi per essere gestiti con facilità.
I motori di ricerca svolgono due attività principali.
Prima di tutto "navigano" nella rete per trovare informazioni. Per fare questo si avvalgono di robot, detti in gergo *spiders* (ragni) oppure anche *crawlers* (cioè nuotatori). Questi non sono altro che programmi che partono da un insieme di URL e seguono la struttura ipertestuale del web per accedere ai documenti disponibili, generando poi un indice dei termini in essi contenuti.
Attraverso delle tecniche di indicizzazione[2] associano a ogni URL un coefficiente di rilevanza, ossia una specie di misuratore di quanto un particolare URL può essere importante rispetto a un dato termine.
Il secondo compito dei motori di ricerca consiste nel rispondere alle richieste da parte degli utilizzatori. Infatti propongono in ordine di importanza le risorse presenti on line che possono essere utili all'utente che ha introdotto nel motore un dato quesito.
In definitiva possiamo identificare due tipi di motori di ricerca: il primo è un semplice indice di argomenti che legge esclusivamente i titoli e le descrizioni; il secondo utilizza il sopracitato spider, cioè un programma di indicizzazione in grado di restituire risultati molto più selettivi.
I motori hanno contribuito a dare un ordine alle innumerevoli risorse della rete, creando vasti archivi di dati che comprendono un gran numero di pagine web. Alla complessità del metodo di archiviazione dei dati adottato dai motori, corrisponde la facilità d'uso da parte dell'utente.
Ai motori di ricerca si affiancano le directory,[3] che consistono in grandi archivi di siti, selezionati da personale specializzato e proposti al ricercatore in un indice di categorie. Ogni categoria si suddivide in sottocategorie, che a loro volta hanno altre sottocategorie. L'utente accede alla categoria d'argomenti di interesse e può affinare la ricerca selezionando le varie sottocategorie che gli vengono proposte.
Le risorse on line riportate nelle varie directory vengono scelte dagli operatori umani che danno vita alle directory stesse, e quindi non derivano da una scansione automatica e continua di tutto il contenuto della rete (come, invece, avviene per i motori di ricerca che si avvalgono, per questa attività, di robot).
Le directory sono utili soprattutto quando l'utente ha ben chiaro che cosa vuole chiedere e che cosa vuole ottenere dalla rete.
La sfida per il futuro di Internet è di dotare i motori di ricerca di un'intelligenza che ancora non hanno.[4] La strada aperta da alcuni

nuovi motori porta verso un perfezionamento della ricerca del documento giusto che soddisfi il più possibile le esigenze dell'utente. Per avere questo occorre una lettura dei documenti presenti in rete simile a quella che farebbe un essere umano: i documenti vengono analizzati non solo nella forma, ma soprattutto nel loro contenuto, tramite regole, inferenze e definizioni, utilizzando criteri semantici e concettuali. Il punto consiste proprio nel definire categorizzazioni, classificazioni, relazioni, schemi, associazioni, collegamenti fra dati e informazioni.
In questo modo il motore di ricerca del futuro (che, come vedremo, sta diventando già una realtà del presente) si allontana dalla ricerca per indirizzi e parole chiave, per andare nella direzione di una ricerca semantica basata su concetti e categorie.
Si tratta di un valore aggiunto offerto all'utente, che in questo modo viene supportato nel reperimento delle informazioni giuste adatte al quesito sottoposto al motore di ricerca.
Questi nuovi motori si basano sul *document clustering*, ossia sulla classificazione dei documenti, che vengono scandagliati nei contenuti e proposti suddivisi per argomento (appunto per classificazione) e per rilevanza. In questo modo l'utente ha un aiuto in più per capire se e come le pagine web trovate dal motore sono per lui più o meno interessanti.
L'utente di Internet può incontrare difficoltà nel reperire in tempi ragionevoli ciò che gli serve, perché spesso non è in grado di sfruttare al meglio gli strumenti che il web gli mette a disposizione. Se usati in modo corretto, i motori di ricerca guidano l'utente fino al risultato atteso.
Per ricerche più complesse, o per quesiti non ben definiti, sono necessari strumenti aggiuntivi: classificare i dati e i documenti presenti in rete come farebbe una persona esperta di tutto il contenuto del web è una caratteristica che può fare di un motore di ricerca uno strumento prezioso in mano al ricercatore più esigente (e magari anche per quello meno esperto nella navigazione in rete).
Tutto questo si inserisce nel più ampio discorso dell'intelligenza artificiale e del cosiddetto "machine learning", cioè l'apprendimento automatico. Da ciò derivano il *data mining*, il *text mining* e il *web mining*, che trovano applicazione proprio nel settore della classificazione dei documenti presenti in Internet. Chiarire questi concetti, aiuta anche a comprendere il quadro di riferimento dello stesso *document clustering*.

La nuova frontiera della ricerca in Internet

Iniziamo con il definire il *data mining*, che consiste nel processo di estrazione di conoscenza da banche dati di grandi dimensioni attraverso l'applicazione di algoritmi che individuano le associazioni "invisibili", o comunque nascoste, tra le informazioni e le rendono quindi visibili. In questo modo vengono esplorate grandi quantità di dati e le informazioni di maggiore rilievo e interesse vengono identificate, isolate e rese disponibili.
Questo procedimento è anche definito "estrazione di conoscenza" e avviene attraverso il reperimento di associazioni e di sequenze ripetute nei dati. Così queste associazioni indicano una struttura o, più in generale, una rappresentazione sintetica dei dati.
Questa procedura non è scevra da rischi impliciti, come ad esempio trovare correlazioni che nella realtà o non esistono o non sono effettivamente significative.
In definitiva il processo di *data mining* offre risultati apprezzabili solo in seguito a una attenta interpretazione dei risultati ottenuti.
Se integriamo il *data mining* nell'ambito della linguistica, parliamo allora di *text mining*. Questo procedimento consiste nell'estrazione e nella mappatura di informazioni direttamente dai testi. In questo modo si può realizzare una sorta di mappa cartografica delle informazioni.
Tale attività può essere messa a buon frutto nelle ricerche in Internet, e in particolare nei documenti presenti nel web: infatti si tratta di una specie di "filtraggio intelligente" di documenti in base alle esigenze specificate dall'utente.
Si stima che la maggior parte delle informazioni presenti in rete è rappresentata da testi: da ciò si comprende l'importanza strategica che il *text mining* può assumere, soprattutto in ambito economico-commerciale.
Se applichiamo insieme il *data mining* e il *text mining* abbiamo il cosiddetto *web mining*, che consiste nella ricerca di associazioni sul piano dei contenuti, della struttura e dell'uso delle informazioni. I contenuti vengono studiati prendendo in considerazione i dati raccolti dai motori di ricerca e dai *web crawlers*. La struttura viene esaminata partendo dai dati che riguardano la struttura stessa di una specifica pagina web. L'uso viene analizzato in base ai dati relativi a un determinato browser.
Una volta ottenute le informazioni con il *web mining*, si procede a un'ulteriore valutazione, spesso attraverso l'utilizzo di alcuni parametri del *data mining*, come il cosiddetto *clustering*, quindi ricercando e definendo le possibili aggregazioni, classificazioni e associazioni tra i dati.
Le possibili (ed effettive) applicazioni sono numerose, soprattutto nei settori del marketing, delle indagini di mercato e della gestione

aziendale. Nell'ambito dell'informazione si applica il processo di *text mining*. Alla base di alcuni motori di ricerca troviamo gli stessi algoritmi utilizzati per il *text mining*: permettono di ricercare i dati e proporli all'utente suddivisi per categorie.

La nuova generazione dei motori di ricerca: una rassegna

Come abbiamo già spiegato, per *document clustering* si intende il processo di raggruppamento delle pagine e dei documenti trovati nel web secondo pattern semantici, parole chiave e temi. Si tratta di una modalità di presentazione dei risultati di ricerca, utilizzata dai motori di ricerca di nuova generazione, come quelli che vengono illustrati di seguito.
Secondo questa modalità, il motore non offre solo, come risultato della ricerca, l'elenco delle pagine web più significative in base alla domanda inserita dall'utente, ma presenta anche un elenco di pagine web classificate per argomenti attinenti all'oggetto della ricerca dell'utente. Quest'ultimo viene così consigliato su come indirizzare la propria ricerca grazie a una prima classificazione delle pagine web effettuata dal motore di ricerca stesso.

Vivísimo
<http://vivisimo.com>

Fondato nel 2000 da alcuni ricercatori della Carnegie Mellon University, può essere definito un "motore per il raggruppamento di documenti". Utilizza un algoritmo di clustering per organizzare i risultati della ricerca in categorie e visualizzarli anche per gruppi tematici, oltre che in ordine di importanza e di argomento.
Vivísimo è utile soprattutto quando l'utente ha bisogno di farsi un'idea di un argomento o semplicemente di un termine di cui non conosce nulla. Inoltre è utile per trovare tutti i possibili termini e i concetti correlati all'argomento.
Quindi non è un semplice motore di ricerca e neppure un metamotore. La definizione corretta sarebbe *clustering engine*, cioè uno strumento che raggruppa le risorse della rete su un dato argomento, rendendole fruibili attraverso cartelle tematiche create in tempo reale.
Esempio. Desideriamo ottenere informazioni su Dante Alighieri. Inseriamo le parole "dante alighieri" come in un normale motore. I risultati ottenuti sono divisi in due sezioni. La prima, che occupa la parte centrale della videata, propone in ordine di importanza alcune pagine web scelte che trattano dell'argomento inserito. Per ogni pagina viene proposta una breve sintesi del contenuto e la possibilità di vedere un'anteprima della pagina stessa, senza doverla aprire.
La seconda sezione dei risultati è posta sulla sinistra della videata e propone una serie di cartelle, in cui i risultati della ricerca sono ordinati in base ad argomento e sotto-argomenti. Grazie a questa sezione possiamo decidere quale sotto-argomento visitare: là troveremo altre pagine web selezionate da Vivísimo e sempre relative a Dante Alighieri. (Figura 1)
Al momento Vivísimo è disponibile solo in lingua inglese.

Fig. 1 – Esito della ricerca su "Dante Alighieri" in Vivísimo

Teoma
<www.teoma.com>

In gaelico Teoma significa "esperto". Nato come progetto sperimentale nel 1998, è stato acquisito nel 2001 dalla Ask Jeeves Inc., la società che gestisce il più noto motore di ricerca Ask Jeeves.
Questo motore di ricerca raggruppa in tre sezioni i risultati della ricerca: un gruppo tematico di cartelle (*Web pages grouped by topic*); i singoli indirizzi web correlati di una breve descrizione, (*Web pages*); una serie di link a siti specializzati sull'argomento richiesto (*expert's links*).
Teoma è molto simile al più famoso Google, in particolare per quan-

Fig. 2 — La ricerca con Teoma

to riguarda l'algoritmo di ricerca. Il modello è chiamato Subject-Specific Poularity, e punta a definire i risultati di ricerca in base alla validità delle pagine web riguardanti lo stesso argomento.
A differenza di Google, come accennato sopra, viene offerta la possibilità di affinare ulteriormente il risultato della ricerca.
Teoma punta alla qualità, sia dei siti indicizzati sia dei risultati di ricerca. La stessa novità di proporre link di esperti (*expert's links*), messi in rete da persone o gruppi di appassionati o esperti in un determinato ambito, ne è una prova.
Esempio. Vogliamo trovare di nuovo informazioni su Dante Alighieri. Inseriamo le parole "dante alighieri" nella stringa di ricerca e otteniamo in pochi attimi i risultati suddivisi nelle tre sezioni. Il corpo centrale è occupato dai cosiddetti "Results – Relevant Web pages": Teoma ha scelto più di 200.000 pagine web che trattano del nostro argomento e le propone in ordine di rilevanza. Sulla destra della videata compaiono le altre due sezioni. La prima è "Refine – Suggestions to narrow your search". Si tratta di suggerimenti di argomenti correlati, con i quali è possibile affinare la ricerca su Dante Alighieri. La seconda è "Resources – Link collections from experts and enthusiasts": sono i contributi offerti da esperti e amanti del settore i quali suggeriscono alcuni link utili per approfondire la ricerca. (Figura 2)
Teoma offre anche la possibilità di effettuare una ricerca avanzata, compilando un form con varie specifiche per circoscrivere il più possibile l'ambito della ricerca.
Anche Teoma è disponibile solo in lingua inglese.

WiseNut
<www.wisenut.com>

È stato lanciato sul mercato insieme a Teoma, nel 2001. Ha un'interfaccia molto semplice, sulla falsariga di Google. Creato in realtà nel 1999, utilizza un sistema di ordinamento delle pagine web in base alla rilevanza. Questo algoritmo misura l'importanza sia delle pagine web in generale sia di quelle trovate all'interno della ricerca. Inoltre valuta il contenuto delle pagine web, analizzando i link riportati e la loro provenienza.
La tecnologia adottata è molto simile a quella impiegata da Teoma.
Offre la possibilità di effettuare la ricerca in 25 lingue, compreso l'italiano.
Il risultato della ricerca effettuata con WiseNut è duplice. Da un lato il motore propone la lista delle pagine web scelte: per ognuna viene indicato un breve riassunto del contenuto e, grazie alla funzione "Sneak-to-peek", viene consentito di "sbirciare" la pagina web proposta senza bisogno di aprirla completamente ma semplicemente visionandone un'anteprima. Inoltre WiseNnut mostra un breve elenco di categorie che riuniscono altri link ad altre pagine relative all'argomento della ricerca. In alcuni casi offre la possibilità di approfondire la ricerca proprio attraverso una di queste categorie.
Facciamo ancora l'esempio di "dante alighieri". Se effettuiamo la ricerca in lingua italiana, WiseNut trova oltre 13.000 documenti relativi all'argomento. Al centro della videata visualizza, in ordine di rilevanza, le pagine web scelte. Nella parte alta dello schermo vediamo invece le categorie proposte (a dire il vero sono poche e abbastanza imprecise; non risultano utili per approfondire in modo particolare la nostra ricerca). (Figura 3)
Se si utilizza WiseNut, la possibilità di effettuare le ricerche in lingua italiana su pagine web italiane è un valore aggiunto che manca agli altri motori di nuova generazione, anche se la funzione di clustering risulta meno accurata.

Clusty
<www.clusty.com>

È un metamotore sviluppato da Vivísimo e si basa sulla funzione

di clustering dei risultati di ricerca. Infatti il suo nome deriva proprio dal termine *cluster*.
È stato lanciato sul mercato con la versione beta nel settembre del 2004, dopo uno sviluppo durato quattro anni. Clusty aggiunge alcune nuove caratteristiche e una nuova interfaccia rispetto a Vivísimo. Ad esempio offre la possibilità di effettuare ricerche anche tra i blog e tra le news. Inoltre è interessante la possibilità di personalizzare, da parte dell'utente, le modalità di ricerca.
Ogni pagina web trovata può essere aperta in anteprima, senza dovere accedere al link per visionarne il contenuto. Inoltre è attiva una funzione che rimanda una determinata pagina web trovata nelle classificazioni proposte per la stessa ricerca. Infatti anche Clusty propone una serie di siti classificati per argomenti affini a quello della ricerca avviata dall'utente.
Ancora l'esempio di Dante Alighieri. Clusty ci segnala che ha trovato oltre 221.000 pagine web sull'argomento ma ne segnala, per rilevanza, in particolare 166. Queste pagine web vengono mostrate nella parte centrale della videata. A sinistra troviamo le classificazioni. Una nota interessante: queste classificazioni possono essere ordinate (e mostrate direttamente) in base all'argomento (ad es. Divine Comedy, Poet, Inferno, Società Dante Alighieri ecc.), alla fonte da cui sono state tratte le pagine web scelte (ad es. GigaBlast, Looksmart, MSN, Wikipedia ecc.), agli URL (ad es. .com, .org, .net ecc.). (Figura 4)
Possiamo scegliere di effettuare la nostra ricerca su Dante Alighieri in varie sezioni di Clusty: "Web", "News", "Images", "Shopping", "Encyclopedia", "Gossip". Basta inserire una volta sola nella stringa di interrogazione le parole "dante alighieri" e poi scegliere di volta in volta la sezione che ci interessa. Man mano cambieranno le pagine web trovate, l'ordine proposto e anche la classificazione riportata nella colonna di sinistra.
Anche Clusty è disponibile solo in lingua inglese, ma offre ampie possibilità di ricerca e di approfondimento. Si presenta come uno strumento agile, veloce e duttile.

Fig. 3 – La ricerca in WiseNut

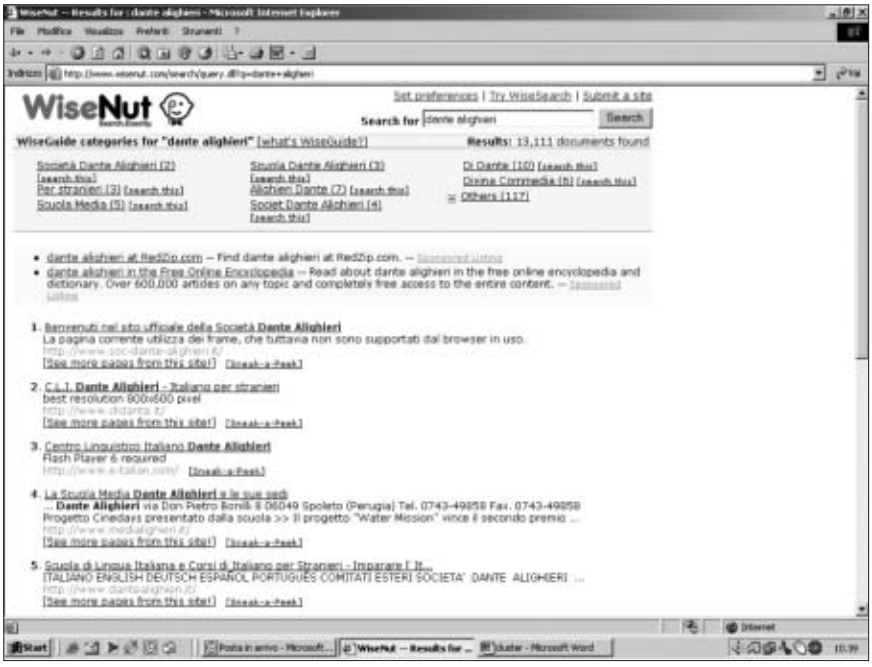
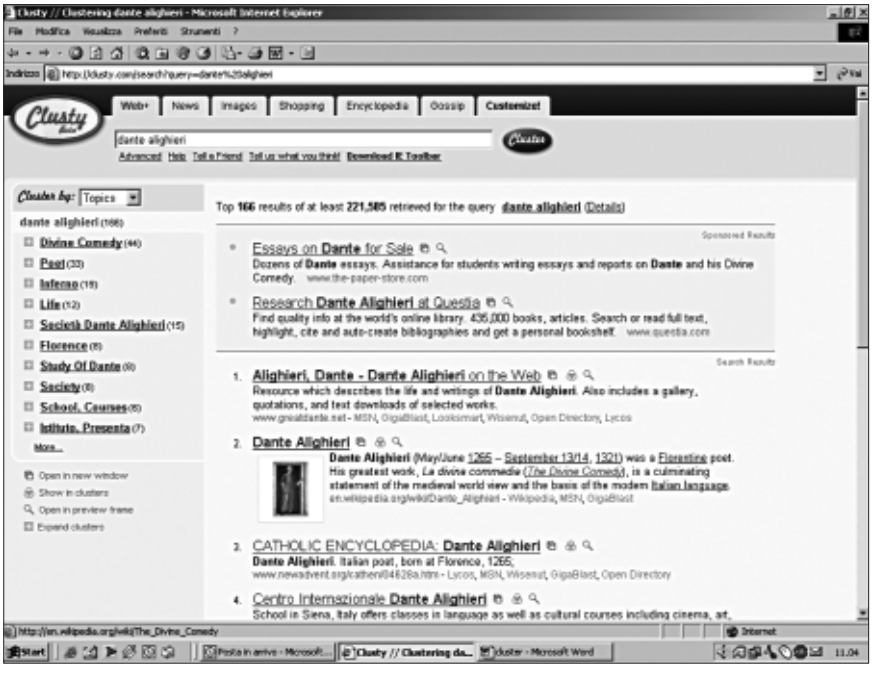


Fig. 4 – La ricerca con Clusty



Turbo10
<http://turbo10.com>

È un metamotore sviluppato da Fleetfoot Internet Solutions Limited (UK). Sfrutta un algoritmo di clustering: in base all'argomento, interroga la directory o il motore di ricerca che identifica come es-

Fig. 5 – La ricerca in Turbo

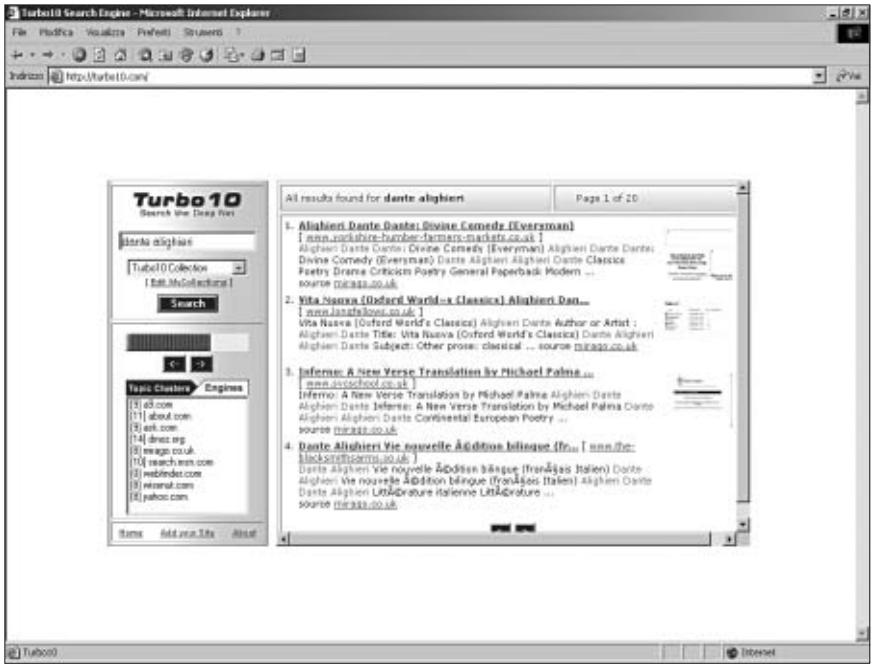
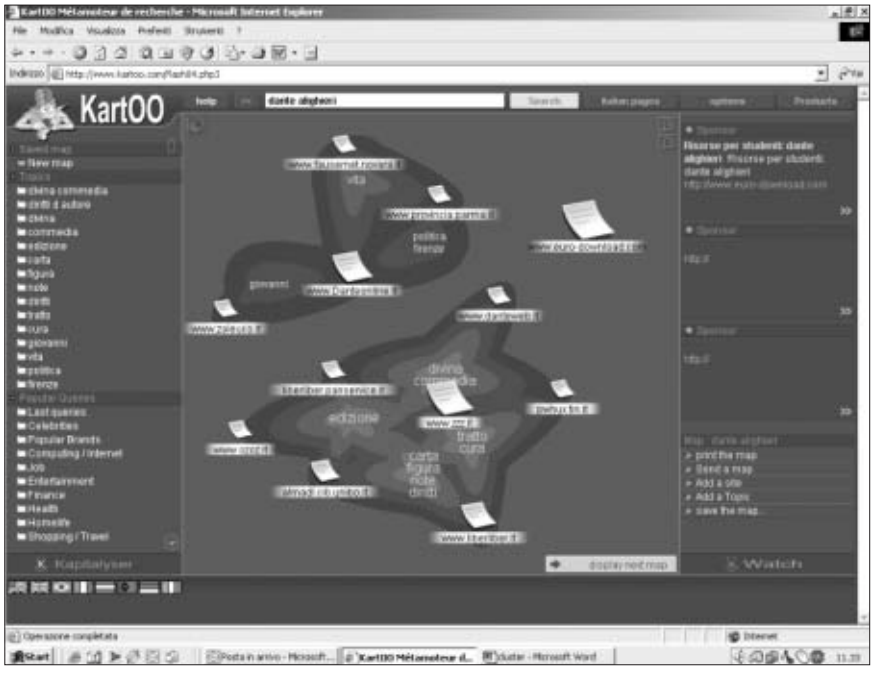


Fig. 6 – La ricerca con Kart00

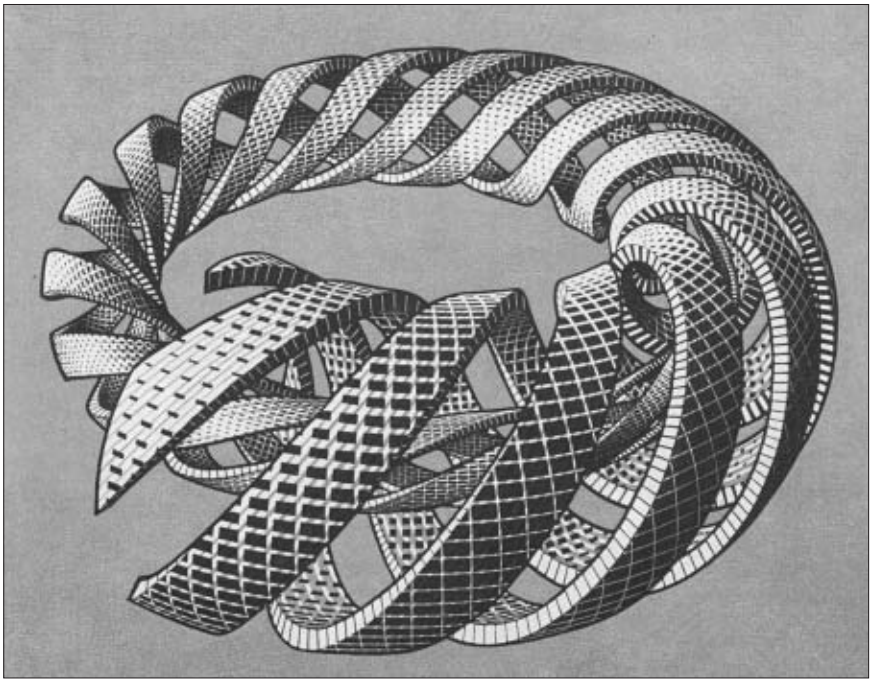


sere più adatto, e mostra in fondo alle classificazioni (*clusters*) la risorsa utilizzata per trovarli.
Quindi, oltre ai motori di ricerca, interroga anche numerose directory: il vantaggio consiste in una maggiore precisione della ricerca e nel reperimento di un maggior numero di documenti.
Inoltre Turbo10 offre la possibilità di creare una collezione personalizzata di directory o anche di motori di ricerca adatti alla esigenze dell'utente. In più suggerisce un motore da aggiungere alla lista nel caso in cui non fosse già presente.
La caratteristica interessante di Turbo10 è che permette all'utente di ordinare manualmente le voci di classificazione e i documenti in base al livello di pertinenza e di rilevanza.
È presente anche la funzione "Search-O-Meter", che consente di muoversi da una pagina all'altra, da un cluster all'altro, mettendo in evidenza i documenti già visionati.
Anche nel caso della nostra ricerca "dante alighieri", Turbo10 visualizza nella parte centrale della pagina dei risultati le pagine web scelte, ma più interessante risulta essere la classificazione per argomento e l'indicazione dei motori utilizzati per ottenere i risultati più pertinenti (in questo caso a9.com, about.com, search.msn.com ecc.). (Figura 5)
Lo svantaggio di Turbo10 consiste nel fatto che, ancora una volta, è disponibile solo la versione in inglese e la ricerca viene effettuata di preferenza su pagine web di lingua inglese.

Kart00
<www.kartoo.com>

È un metamotore di ricerca che ha la particolarità di proporre i risultati sotto forma di mappe grafiche bidimensionali o tridimensionali. I siti web trovati vengono infatti rappresentati con icone più o meno grandi in base al grado di rilevanza. Per affinare la ricerca, l'utente viene guidato dalla mappa stessa all'utilizzo di parole chiave.
I risultati della ricerca possono essere filtrati. Le mappe possono contenere dei siti cosiddetti "parasiti", cioè non concordanti con l'oggetto della ricerca. Per escluderli, esiste un apposito pulsante (a forma di gomma per cancellare) che serve per eliminarli dai propri risultati.
Kart00 offre la possibilità di scegliere la lingua.
I risultati della ricerca vengono disposti in una mappa: le icone che appaiono, quando si passa sopra con il cursore del mouse, mostrano le parole chiave corrispon-

denti e, a sinistra della pagina, appare una breve descrizione del sito. A questo punto è possibile affinare la ricerca, aggiungendo o escludendo dei temi.
Interessante è il barometro che rappresenta graficamente il numero di siti che corrispondono alla ricerca.
I cluster si presentano come parole bianche su sfondo blu, disseminate sulla superficie della mappa e collegate tra di loro.
La vera novità di Kart00 consiste nell'interattività della mappa: infatti spostandosi con il cursore del mouse sopra le varie zone della mappa, possiamo creare vari livelli di legami tra le informazioni che il metamotore ha selezionato nel web per noi.
Nel caso della ricerca su "dante alighieri", questo motore risulta essere molto ricco di collegamenti e di spunti di approfondimento. I legami tra le risorse trovate sono numerosi e tutti rilevanti. L'interfaccia visuale, davvero innovativa, rappresenta veramente una marcia in più per Kart00, che, inoltre, possiede tutte le caratteristiche dei migliori motori di nuova generazione, come ad esempio Vivísmo.


M.C. Escher

Considerazioni finali

I motori di ricerca si muovono da sempre nella rete mondiale, effettuando le loro ricerche per indirizzi e per parole chiave. La nuova generazione invece cambia prospettiva, e va verso un'analisi per concetti, per categorie, addentrandosi nella ricerca semantica.
Grazie a questi nuovi strumenti di ricerca, l'utente viene guidato verso un ampliamento della propria conoscenza, accompagnato attraverso proposte di navigazione, aiutato da interfacce grafiche facili, immediate e addirittura interattive.
Il lavoro dei motori di ricerca diventa sempre meno meccanico e sempre più simile al contributo che potrebbe offrire un esperto umano: con la classificazione delle pagine web, con l'ordinamento per rilevanza dei documenti trovati, con l'interattività tra l'utente e il motore stesso, l'algoritmo che sottostà a questi nuovi motori conduce al più ampio settore dell'intelligenza artificiale.
Proprio grazie a questi algoritmi i nuovi motori consentono di valutare un'enorme quantità di dati e di pagine web, estrapolandone le relazioni, le regolarità, i legami. L'utilità non riguarda solo il settore economico (ad esempio il commercio, il marketing, l'ambito economico-finanziario), ma anche quello dell'informazione e della ricerca bibliografica.
La ricerca così effettuata nel web apre la possibilità a risultati anche inaspettati su documenti prima quasi introvabili e su legami tra dati e informazioni che possono aiutare il ricercatore con notevole risparmio di tempo e di energie.
Nell'utilizzo dei motori di ricerca di nuova generazione si possono individuare due problemi.
Prima di tutto bisogna porre attenzione nella scelta delle pagine web rilevanti e delle classificazioni per argomento proposte dal motore: non sempre il grado di rilevanza è veritiero e non sempre i cluster sono veramente i più adatti per approfondire la ricerca. Non è raro scoprire che alcune delle pagine web presentate sono in realtà irrilevanti, se non addirittura fuorvianti. L'algoritmo, per quanto affinato e preciso, non può sostituire *in toto* il contributo della persona studiosa ed esperta di un dato argomento, anche se il motore di ricerca rappresenta un aiuto in più rispetto agli strumenti tradizionali presenti in rete.
Inoltre rappresenta un indubbio problema per l'utente italiano la disponibilità di motori che, per lo più, utilizzano esclusivamente la lingua inglese e che effettuano la loro ricerca a partire da pagine web scritte in inglese. I motori di nuova generazione (tranne alcune eccezioni) offrono il meglio delle loro potenzialità proprio se i termini della ricerca vengono inseriti in lingua inglese. In caso contrario, i risultati ottenuti sono limitati e devono essere presi in considerazione con cautela.
La funzione di clustering resta comunque un'importante innovazione, qualunque sia la lingua utiliz-

zata dal motore o dall'utente, perché va a cambiare concettualmente l'idea di ricerca nel web. La ricerca viene effettuata dai motori anche nel contenuto delle pagine web e questo rappresenta un tentativo (spesso ben riuscito, ma non sempre) di fare ordine e di analizzare la straordinaria mole di dati e documenti presenti in rete, resi spesso introvabili per motivi tecnici legati alla scrittura in codice delle pagine web stesse.

Fonti di riferimento

A road map to text and web mining (2005), URL: <http://intelligent-web.org/wsm/>.

Adding a new dimension to search: the Teoma difference is authority, URL: <http://sp.teoma.com/docs/teoma/about/searchwithauthority.html> (Ask Jeeves Inc., 2005).

Stephen Arnold, *Vivísimo: the next generation*, "Information World Review", (2003), 192, p. 23.

Mary Ellen Bates, *Online spotlight. Teoma.com*, "OnlineMag.net", September 2001, p. 96.

Tim Berners-Lee, *A road map to the Semantic Web*, URL: <http://www.w3.org/ DesignIssues/Semantic.html>.

Juri Bordani, *Metti l'intelligenza nel motore*, URL: <http://www.mytech.it/internet/articolo/idA028001050459.art>.

Francesco Caccavella, *Teoma: "noi sostituiremo Google"*, URL: <http://webnews.html.it/focus/190.htm>.

Daniela Canali, *La nuova generazione dei motori di ricerca*, "Biblioteche oggi", 20 (2002), 7, p. 8-12.

Clustering engine, URL: <http://vivisimo.com/products/Clustering_Engine/Introduction.html>.

Clusty, "Wikipedia", URL: <http://en.wikipedia.org/wiki/Clusty/>.

CRM Today – Data mining, URL: <http://www.crm2day.com/data_mining/>.

Extreme searcher's profile of WiseNut, "Extreme Searcher.com", URL: <http://extremesearcher.com/wisenut.htm>.

Weiguo Fan, *Text mining, web mining, information retrieval and extraction from the World WideWeb references*, URL: <http://filebox.vt.edu/users/wfan/text_mining.html> (2004).

Luca Fini, *Ricerca delle informazioni*, URL: <http://www.vialattea.net/esperti/inform/google.htm>.

Barbara Fiorentini, *Presente e futuro prossimo della ricerca in Internet*, URL: <http://villaggiovirtuale.splinder.com/archive/2005-02>.

Alessandra Guidoni, *Vivísimo, il motore di ricerca che organizza i risultati per categorie*, "Apogeo On Line", 8 agosto 2001, URL: <http://www.apogeonline.com/webzine/2001/08/08/01/200108080101>.

Id., *Teoma, il motore di ricerca di terza generazione*, "Apogeo On line", 15 febbraio 2002, URL: <http://www.apogeonline.com/webzine/2002/02/15/01/200202150102>.

David Hand – Heikki Mannila – Padhraic Smyth, *Principles of data mining*, MIT Press, Cambridge, MA, 2001.

Saman Haqqi – Jason Morris, *Vivísimo launches Clusty – Inique new search site brings the power of clustering to web search, shopping, people finding, Wikipedia and more*, URL: <http://vivisimo.org/html/clusty-20040930>.

Kart00, "Kart00.com", URL: <http://www.kartoo.com>, <http://www.kartoo.net>.

KD Nuggets, URL: <http://www.kdnuggets.com>.

Kmining, URL: <http://www.kmining.com/info_conferences.html>.

Cathleen Moore, *Ask Jeeves solves search query*, "www.inforworld.com", 01.07.02, p. 15.

Next-Generation search… today, "Clusty.com", URL: <http://news.clusty.com/about>.

Greg R. Notess, *Review of Teoma*, "Search Engine Showdown", April 2004, URL: <http://www.searchengineshowdown.com/features/ teoma/review.html>.

Id., *Review of WiseNut*, "Search Engine Showdown", URL: <http://searchengineshowdown.com/features/wisenut/review.html>.

Search engine WiseNut, "Metamend.com", URL: <http://www.metamend.com/search-engine-wisenut.html>.

Kevin Sweeney, *Search engines emphasize detail through technology*, "Employee Benefit News", December 2001, p. 35 e sg.

Turbo10, "Turbo10.com", URL: <http://turbo10.com>.

Richard W. Wiggins, *Teoma search engine goes live*, "Information Today", May 2002, 19/5, p. 28.

Wikipedia.org, URL: <http://www.wikipedia.org>.

Elisabetta Zoni, *Web mining: come trovare l'oro in una miniera di parole*, URL: <http://applicata.clifo.unibo.it/risorse_online/e-Mining.htm>.

Note

[1] Per conoscere i motori di ricerca presenti nel web, le loro caratteristiche e le modalità di utilizzo, è utile la consultazione del sito MotoriDiRicerca.IT (www.motoridiricerca.it), gestito dalla società Ad maiora.

[2] L'indicizzazione nei motori di ricerca è un'attività svolta dagli spider durante l'esplorazione del web. In questa fase, i motori di ricerca archiviano le pagine secondo regole specifiche e diverse per ciascun motore. Le pagine del sito vengono esaminate e inserite negli indici di ricerca in base a diversi criteri: anzianità del sito, struttura, contenuto, parole chiave scelte, link popularity ecc.

[3] Le directory più note in Italia sono, ad esempio, Virgilio (www.virgilio.it) e Yahoo! (http://it.yahoo.com/). In più offrono anche un motore di ricerca.

[4] I motori di ricerca "intelligenti" si inseriscono nell'ampio discorso del cosiddetto "web semantico". Per web semantico si intende, in sintesi, la trasformazione del World Wide Web in un ambiente dove è possibile pubblicare non più solo documenti (pagine HTML, file office, immagini, file multimediali,...) ma anche informazioni e dati in un formato adatto all'interrogazione, all'interpretazione e, più in generale, all'elaborazione automatica.
Per approfondire l'argomento si consiglia la lettura di: Paolo Bouquet – Roberta Ferrario, *Il Semantic Web*, <http://lgxserve.ciseca.uniba.it/lei/ai/networks/03-2/introduzione.pdf>; vedere anche la definizione sul sito del W3C: <http:// www.w3.org/2001/sw/>; T.B. Lee, *L'architettura del nuovo web*, Feltrinelli, Milano 2004; J. Allen, *Making a Semantic Web*, <http://www.netcrucible.com/ semantic.html>; T. Berners-Lee – J. Hendler – O. Lassila, *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.