Meta-informazioni, motori di ricerca, tesauri

Fabio Vitali

Introduzione

Le directory di siti nascono insieme al Web. Già dai primi mesi di esistenza del prototipo WWW di B-L, esistevano pagine di link ai server Web esistenti.

I motori di ricerca esistevano come prodotto locale (sui documenti contenuti in un unico server), specialmente con WAIS e il protocollo Z39-50.

L'ingrandirsi del Web ha reso di estremo successo i motori di ricerca come meccanismo principale di scoperta di risorse su Web.

Tipi di motori di ricerca

I motori di ricerca possono essere divisi in varie categorie a seconda del tipo di servizio che forniscono:

- **Crawlers** (o search engines semplici): un'applicazione scarica sistematicamente le pagine di un sito, ne indicizza il contenuto, applica vari meccanismi di prioritizzazione (keyword, autorevolezza, focalizzazione, ecc.). Il database viene consultato da un apposito form che permette di inserire qualunque stringa e vedere se da qualche parte esiste un documento con quella parola.
  Ad esempio Google, AltaVista, Inktomi (HotBot, MSN, AOL), Fast (Lycos).
- **Directories**: Una squadra di esseri umani (anche corposa) esamina a uno a uno decine di migliaia di siti, e li classifica e categorizza in un albero gerarchico di categorie. Ad esempio Yahoo, Open Directory Project, Looksmart (MSN)
- **Metacrawlers**: un motore sottopone la stessa query a molti crawlers diversi e filtra tutti i risultati ottenuti. Ad esempio www.savvysearch.com, www.metacrawlers.com, ecc.
- **Motori di ricerca specializzati**: Il Web invisibile, i nomi di dominio, le pagine appena inserite, ecc.

Strategie di posizionamento

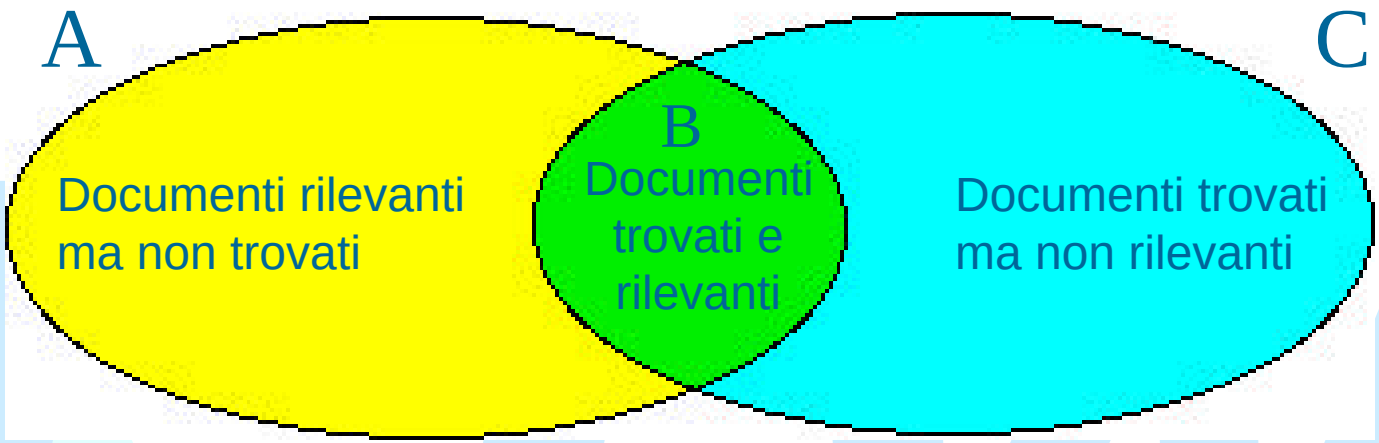
Con il successo del Web è diventato di enorme importanza la comprensione e lo sfruttamento dei meccanismi di categorizzazione e indicizzazione dei motori di ricerca.

Finire nelle prime righe di una categoria in una directory o nei primi risultati di una query in un crawler significa migliorare decisamente il successo del proprio sito, e spesso significa soldi in quantità.

I crawler hanno algoritmi noti di indicizzazione che possono essere sfruttati maliziosamente dai creatori di siti. Questo processo si chiama "Web positioning" ed è una vera e propria professione dei giorni nostri.

Le directory invece sono manipolate da esseri umani e l'evidenziazione di un sito rispetto ad un altro può essere decisa caso per caso dai redattori. Alcuni siti richiedono il pagamento di abbonamenti per essere indicizzati nelle prime posizioni delle varie categorie.

Concetti dei motori di ricerca



- Documenti perduti: documenti rilevanti ma non trovati
- Rumore: documenti trovati ma non rilevanti
- Recall (o richiamo): rapporto tra i documenti rilevanti trovati e i documenti rilevanti (B/A). Un sistema è equo verso gli autori se ha un recall alto
- Precision: rapporto tra i documenti rilevanti trovati e i documenti trovati (B/C). Un sistema è equo verso gli utenti se ha una precision alta.
- Grado di futilità: quantità massima di documenti che l'utente è disposto ad esaminare prima di abbandonare la ricerca (ad es. 30)

Indicizzazione

La differenza fondamentale tra directory e motori di ricerca è tra indicizzazione per concetti e indicizzazione per termini:

- **Indicizzazione per concetti** (o ***assegnata***): Viene definito un vocabolario controllato, su cui verrà fatta l'indicizzazione. Indipendentemente dal vocabolario usato dall'autore, è necessario ricondurre tutti i documenti a termini presenti nel vocabolario controllato. Questa operazione va fatta a mano documento per documento. E' lunga e manuale ma molto espressiva.
- **Indicizzazione per termini** (o ***derivata***): Viene indicizzato il vocabolario usato dall'autore, senza curarsi di ambiguità sinonimie, ecc. In fase di ricerca bisognerà pensare a tutte le possibili forme usate nei documenti rilevanti, e cercare anche esse. E' veloce, automatica ma non espressiva, lascia all'utente il compito di comporre la query più efficace per l'identificazione dei documenti cercati

Meta-informazioni

La creazione del vocabolario controllato e l'indicizzazione per concetti implicano la creazione autoriale di un nuovo oggetto, il ***catalogo***, che descrive le caratteristiche della collezione di documenti indicizzata.

Questo documento ***parla di*** documenti, e quindi costituisce una fonte di **meta-informazioni** sui documenti di cui parla.

Un vocabolario di meta-informazioni è caratterizzato da:

- Una limitazione nel numero di elementi (argomenti di meta-informazione)
- Un nome associato a ciascun elemento
- Un significato associato a ciascun elemento.

Il Dublin Core rappresenta senza dubbio il principale modello di vocabolario di meta-informazioni per documenti di rete.

Dublin Core

```
<HTML><HEAD><TITLE>Song of the Open Road</TITLE>
  <META NAME="DC.Title" CONTENT="Song of the Open Road">
  <META NAME="DC.Creator" CONTENT="Nash, Ogden">
  <META NAME="DC.Type" CONTENT="text">
  <META NAME="DC.Date" CONTENT="1939">
  <META NAME="DC.Format" CONTENT="text/html">
  <META NAME="DC.Identifier" CONTENT="http://www.site.com/nash/open.htm">
</HEAD>
<BODY><PRE>
I think that I shall never see
A billboard lovely as a tree.
Indeed, unless the billboards fall
I'll never see a tree at all.
</PRE></BODY>
</HTML>
```

I tesauri (o thesauri)

sing.: tesauro (o thesaurus)

Definizione di tesauro (ISO 2788-1986) «il thesaurus è il vocabolario di un "linguaggio di indicizzazione" controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni "a priori" fra i concetti»

Il concetto di **vocabolario controllato** indica l'esigenza di trovare un punto di incontro tra lessico dell'autore e lessico del ricercatore, una relazione biunivoca tra termine e concetto, così da ottenere ***univocità semantica***: un termine per ogni concetto, un concetto per ogni termine.

Questa condizione elimina i problemi connessi con l'uso del linguaggio naturale, in cui ridondanze, ambiguità, polisemie, omonimie, omografie ed altre caratteristiche che ne garantiscono ricchezza ed espressività, ma rendono difficile l'organizzazione funzionale dei motori di ricerca.

Le relazioni identificate nel tesauro debbono essere formalizzate e a priori, ovvero appartenenti alla sfera dei concetti (e non dei termini) e universali (cioè vere sempre).

Concetti per i tesauri

I concetti rappresentati dai termini di un th possono appartenere a diverse categorie:

- entità concrete
  - oggetti e loro parti fisiche
  - materiali
- entità astratte
  - azioni e avvenimenti
  - entità astratte e proprietà degli oggetti, dei materiali o delle azioni
  - discipline o scienze
  - unità di misura
- entità individuali o "classi di uno" analoghe a nomi propri.

Relazioni tra termini

Relazione preferenziale o sinonimica

- Identifica tra più termini per lo stesso concetto quello preferito. Identifica classi di equivalenza (sinonimi) ad un termine più importante
- Es.: regola/norma, week-end/finesettimana, mal di testa/cefalea.

Relazione gerarchica

- Identifica tra due termini una relazione di subordinazione all'interno di uno stesso albero gerarchico. E' ciò che effettivamente distingue un vocabolario controllato semplice da un tesauro propriamente detto.
- Es.: matematica/geometria, felini/gatti, veicoli/automobili

Relazione associativa

- Relazione residuale, volta ad identificare tra due termini una relazione né di equivalenza, né di subordinazione, ma comunque esistente ed innegabile.
- Es.: barca/nave, ecologia/inquinamento, ecc.

Relazione preferenziale (1)

Identifica un gruppo di equivalenza tra termini, tra i quali si sceglie il termine preferito. Gli altri vengono detti termini non preferiti o sinonimi.

La relazione tra termine non preferito (NPT) e termine preferito (PT) si chiama USE. La relazione inversa UF (Use For)

Thesaurus **Tesauro**

USE Tesauro **UF Thesaurus**

Rientrano in questa categoria:

- Sinonimia vera **regola e norma**
- Varianti ortografiche **psicoanalisi e psicanalisi**
- Sigle e acronimi **CNR e Centro Nazionale delle Ricerche**
- Preferenza linguistica
  - T. straniero e italiano **week-end e finesettimana**
  - T. attuale e antico **bicicletta e velocipede**
  - T. comune e scientifico **mal di testa e cefalea**
  - T. di origini diverse **poliglotta e multilingue, antologia e florilegio**
  - T. comuni e marche **penna a sfera e biro, fotocopiatrice e xerox**
  - Varianti molto recenti **telefonino, telefono cellulare, telefono portatile**

Relazione preferenziale (2)

Oltre alla sinonimia propria, con relazione preferenziale si possono mettere in relazione anche termini non strettamente sinonimici (***sinonimia convenzionale***), in cui i termini sono considerati sinonimi solo all'interno del contesto dei documenti gestiti da tesauro.

Possiamo distinguere:

- Quasi-sinonimia **punizione, ammenda, sanzione, pena**
- Upward posting **TIR e camion**
  (si parla di upward posting per termini in relazione gerarchica di cui non interessa gestire la specificità. Si usa il termine più generico).
- antinomia **guerra e pace, amore e odio, malattia e salute**

Relazione gerarchica

Descrive un albero di termini, tra i quali esiste un rapporto di subordinazione o sovraordinazione. I termini subordinati vengono anche detti iponimi, quelli sovraordinati vengono anche detti iperonimi.

La relazione tra termine e termine inferiore è NT (narrower term), tra termine e termine superiore è BT (broader term)

Geometria

NT1 Geometria euclidea

NT1 Geometria non euclidea

NT2 Geometria iperbolica

NT2 Geometria ellittica

Geometria ellittica

BT Geometria non euclidea

BT Geometria

BT Matematica

Rientrano in questa categoria:

- Relazione generica o genere/specie
- Relazione partitiva o parte/tutto
- Relazione esemplificativa o classe/istanza

Relazione generica

Detta anche relazione genere/specie o relazione **is-a** (**è-un**). Sigla specifica: BTG e NTG.

E' il legame che esiste tra una categoria e i suoi membri. Perché sia corretta, è necessario che tutte le istanze del termine subordinato siano istanze del termine sovraordinato.

Ad esempio, felino/gatto è una coppia di termini in relazione generica, mentre animale domestico/gatto non lo è, perché esistono gatti selvatici.

Questa differenza assoluta, però, può non essere vera nell'ambito dei documenti che vengono trattati (se non si parla di animali selvatici la relazione is-a vale anche per la coppia animale domestico/gatto.

Relazione partitiva

Detta anche relazione parte/tutto o relazione **has-a** (**ha-un**). Sigla specifica BTP e NTP.

E' il legame che esiste tra un concetto complesso e i suoi componenti. Perché sia corretta, è necessario che tutte le istanze del termine subordinato implichino il termine sovraordinato. Ovvero non possono esistere due esempi dello stesso termine all'interno di due gerarchie differenti.

In generale questo è possibile solo in quattro casi:

- Organi del corpo (**sistema circolatorio - vene**)
- Nomi geografici (**Italia - Emilia-Romagna - Bologna**)
- Discipline (**scienze - biologia - botanica**)
- Strutture sociali (**divisione - reggimento**)

Altrimenti è possibile solo per organizzazioni specifiche interne al tesauro.

Relazione esemplificativa

Detta anche relazione classe/istanza o specie/esempio.

E' il legame che esiste tra una classe ed un suo individuo (classe di uno).

Ad esempio

Pontefici

- **NT1 Giovanni XXIII**
- **NT1 Paolo VI**
- **NT1 Giovanni Paolo I**
- **NT1 Giovanni Paolo II**

Monogerarchie e poligerarchie

Come ben sappiamo, le relazioni gerarchiche possono assumere strutture complesse nel momento in cui assumiamo una classe specifica come derivato da più classi generiche.

E' importante allora mettere ben in chiaro se si adottano gerarchie multiple o semplici. Ad esempio:

Organo

BT1 Strumenti a fiato

BT1 Strumenti a tastiera

BT2 Strumenti

Strumenti a fiato

BT1 Strumenti

NT1 Organo

NT1 Flauto

Strumenti a tastiera

BT1 Strumenti

NT1 Organo

NT1 Pianoforte

Relazione associativa

Identifica una relazione non definibile né come sinonimica, né come gerarchica, e tuttavia innegabile. Relazione residuale.

Viene indicata con la sigla RT (related term) o "vedi anche".

- Termini appartenenti alla stessa categoria. Es.: barca e nave
- Termini appartenenti a categoria diverse
  - una disciplina e il suo oggetto di studio (zoologia e animali);
  - un processo od operazione e il suo agente o strumento (termometro e misurazione della temperatura);
  - una azione e il suo prodotto (scrittura e documenti);
  - una azione e chi o cosa la subisce (potatura e piante; pesca e pesci);
  - oggetti e fenomeni e loro proprietà (magneti e magnetismo);
  - concetti e loro origini (Tedeschi e Germania);
  - concetti legati da rapporti causali (inquinamento e sostanze inquinanti);
  - una cosa e il suo antidoto (piante ed erbicidi);
  - un concetto e la sua unità di misura (frequenza e hertz);

Riferimenti

- International Standard ISO-2788, Documentation -- ***Guidelines for the development of monolingual thesauri***, Second edition -- 1986-11-15
- Chris Taylor, ***An Introduction to Metadata***, http://www.library.uq.edu.au/iad/ctmeta4.html
- Serafina Spinelli, ***Introduzione all'indicizzazione,*** http://mail.biocfarm.unibo.it/~spinelli/indicizzazione/
- Serafina Spinelli, ***Introduzione ai thesauri***, http://mail.biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm