MATHEMATICS SUBJECT CLASSIFICATION
e classificazioni connesse nel mondo digitale

Antonella De Robbio
e-mail: derobbio@math.unipd.it
Dario Maguolo
e-mail: dario@math.unipd.it
Biblioteca del Seminario Matematico, Università di Padova

Alberto Marini
e-mail: alberto@iami.mi.cnr.it
Istituto per le Applicazioni della Matematica e dell'Informatica,
Consiglio Nazionale delle Ricerche (IAMI-CNR), Milano

La connessione di classificazioni nel mondo digitale

Nel mondo sempre più pervasivo e interconnesso dell'informazione digitale, è una necessità per le attività in rete basate sulla conoscenza, disporre di connessioni affidabili fra strumenti per la rappresentazione della conoscenza, la ricerca di informazioni e documenti, il lessico, quali classificazioni, soggettari, thesauri, terminologie e ontologie

Utenti di diversi ambienti e con diverse richieste e aspettative desiderano soddisfare le loro necessità informative da qualunque parte l'informazione si trovi, tagliando costi e tempi se possibile, senza tener conto dell'eterogeneità delle fonti: da basi di dati specializzate e portali dedicati a cataloghi generali in linea di biblioteche o motori di ricerca nel Web, da basi di dati di riferimento (metadati) a biblioteche digitali di documenti a testo completo, da aggregatori di periodici elettronici a server di preprint e archivi personali degli autori (comunemente nominati *e-print system*).

L'organizzazione, le funzionalità e le modalità di interazione messe in atto dalle biblioteche digitali in rete sono completamente diverse da quelle in cui generalmente ci si imbatte nelle tradizionali biblioteche di documenti cartacei. Inoltre, sulla linea degli e-print system, con lo sviluppo di meccanismi tecnici e strutture organizzative per dare supporto alla loro interoperabilità, promosso dalla Open Archives Initiative (OAI) [ved. il sito Web www.openarchives.org], ci sarà un'evoluzione che farà di questi sistemi i mattoni della costruzione di un modello completamente trasformato di comunicazione scientifica, radicalmente differente dal modello tradizionale, dominato dalla pesante mediazione delle aziende editoriali scientifico-accademiche.

D'altra parte, gli utenti non gradiscono di dover reimpostare ragionamenti e modalità operative ogni volta che si trovano di fronte a diversi sistemi di registrazione, indicizzazione e presentazione delle informazioni: questo lavoro dovrebbe essere svolto automaticamente dal sistema. Non è una cosa banale. Per quanto riguarda l'indicizzazione per soggetto, succede spesso che differenti classificazioni, thesauri o terminologie altrimenti strutturate, pur interessando una stessa area, mantengano forti disaccordi sul piano linguistico (non si tratta di semplici questioni di traduzione), strutturale e semantico. Si evidenziano disaccordi particolarmente gravi quando si passa dal mondo specialistico delle classificazioni disciplinari alle classificazioni generali di largo uso nelle biblioteche pubbliche, scolastiche e anche

universitarie, quali la Classificazione Decimale Dewey, la Classificazione Decimale Universale, o la Classificazione della Library of Congress.
È facile incorrere in interpretazioni scorrette quando le stesse parole vengono usate in contesti o per scopi diversi. Ciononostante, è possibile realizzare connessioni effettivamente funzionanti fra classificazioni o strumenti simili, purché gli oggetti a cui ciascuna classificazione si riferisce siano identificati senza ambiguità per mezzo di un appropriato linguaggio di rappresentazione.

Classificazioni: funzioni, struttura e dinamica

Le classificazioni possono essere viste come spazi astratti strutturati, o modelli per organizzare spazi materiali, dove oggetti rispettivamente materiali o immateriali possono essere collocati, a seconda di caratteristiche selezionate, in modo che si possano trovare oggetti d'interesse semplicemente scegliendo e muovendosi nei percorsi predisposti dalla struttura dello spazio o dall'organizzazione definita dal modello. Tipici oggetti materiali che sono collocati tramite una classificazione sono i libri negli scaffali di una biblioteca, o anche le citazioni bibliografiche in indici a stampa; per quanto riguarda gli oggetti immateriali, possiamo pensare ai campi o alle discipline della conoscenza e dell'attività umana, ai concetti e agli oggetti di un certo campo o disciplina, o generalmente ai soggetti di documenti intesi in astratto come unità di informazione.
Anche se una classificazione non è implicata nell'organizzazione di uno spazio fisico, per esempio in registrazioni bibliografiche trattate dal computer, si può dare per questa classificazione una semantica adeguata nei termini di una certa nozione di spazio, eventualmente meno vincolato e più complesso degli spazi materiali a cui siamo abituati. Lo spazio della classificazione è comunque la forma di un contenitore, uno spazio operativo, relativamente grossolano, per l'impacchettamento di concetti e l'inserimento dei pacchetti in una catena di trasporto; è del tutto diverso dallo spazio degli oggetti come sono effettivamente intesi dagli utenti della classificazione, uno spazio che può essere fatto emergere dalla classificazione in una maniera più o meno definita come uno spazio concettuale di veri ed effettivi significati.
Visitando lo spazio di una classificazione, siamo aiutati da descrizioni strutturate, descrizioni testuali semistrutturate o non strutturate, immagini o altre forme di segnalazione, secondo specifiche convenzioni, a riconoscere luoghi e oggetti che vi trovano collocazione, o a prendere un'altra strada per spostarci da un luogo a un altro. Ad ogni luogo, e quindi ad ogni oggetto, vengono assegnati uno o più indirizzi, che sono numeri o sequenze di caratteri funzionali a identificare il luogo tramite una codificazione del/dei percorsi per raggiungerlo.
Le descrizioni fanno riferimento agli oggetti attraverso la mediazione dei luoghi in cui gli oggetti sono raccolti, o dei canali che li veicolano, in funzione di un'adeguatezza rispetto a specificazioni o vincoli esterni (leggibilità da umani, trattabilità per usi specifici). Perciò una singola descrizione può far riferimento a una collezione di oggetti che sono intesi distintamente dall'utente, ma sono raccolti in conformità all'organizzazione della classificazione. D'altra parte, un singolo oggetto o luogo può essere rappresentato in forme differenti, pure nell'osservanza delle convenzioni linguistiche o semiotiche della classificazione. I thesauri e i soggettari, al contrario, si preoccupano di mantenere una stretta corrispondenza fra oggetti e descrizioni, al prezzo di arrabattarsi su forme preferite e non preferite: ma questo significa costringere la varietà del linguaggio naturale a passare attraverso le ruote dentate degli identificatori di macchina. L'aggiunta di note d'ambito a testo libero o semistrutturato è un ulteriore segnale di questa confusione.
È ruolo degli indirizzi la guida della macchina viaggiante: per svolgere questa funzione non c'è bisogno di conoscere perché la persona che viaggia vuole raggiungere un certo luogo, e per

trovare che cosa. Nelle classificazioni di soggetto, quindi, gli indirizzi (comunemente chiamati *codici* o *numeri di classificazione*) sono fondamentali, nella loro forma esatta, per la collocazione di documenti materiali in biblioteche materiali, e liste di indirizzi sono i mezzi principalmente utilizzati per l'indicizzazione per soggetto nelle basi di dati bibliografici e nei cataloghi in linea di biblioteche; gli indirizzi codificano e rendono intelleggibile la struttura dello spazio, ma agiscono nient'altro che come elementi di collegamento, senza nessun reale contenuto semantico. I reali vettori del contenuto semantico sono le descrizioni; la classificazione organizza questi vettori in una struttura che esiste indipendentemente dalle effettive forme degli indirizzi, cioè dalle forme che sono fissate allo scopo di rappresentare la struttura dello spazio della classificazione in vista di riferimenti e connessioni esterne.
Inoltre, mentre tanto le descrizioni quanto gli indirizzi possono cambiare, nel tempo o attraverso differenti convenzioni linguistiche, semiotiche o di codificazione, non è necessario che essi cambino in dipendenza le une con gli altri, o con i cambiamenti le trasformazioni, le nascite e le morti fra gli oggetti, gli spazi e i modi in cui oggetti e spazi sono organizzati e percepiti. Gli indirizzi possono cambiare mentre le descrizioni rimangono le stesse, o la struttura dello spazio, almeno localmente, è conservata; le descrizioni possono cambiare mentre gli oggetti rimangono gli stessi; gli oggetti possono cambiare mentre gli indirizzi rimangono gli stessi, e così via.
Differenti classificazioni che coprono aree in parte sovrapponibili possono esercitare nel tempo un'influenza reciproca, specialmente sulla struttura e sulle descrizioni, allo scopo di raggiungere per gli stessi oggetti modi di vedere simili o compatibili, anche se I punti di vista o le estensioni di campo possono essere differenti, e ciascuna classificazione può continuare a raggruppare gli oggetti in maniera in tutto o in parte differente dalle altre

Visualizzazioni di schemi di classificazione: i nostri risultati

Mentre un certo numero di approcci alle questioni della connessione di classificazioni o thesauri utilizza metodi statistici o tecniche di reti neurali, una tendenza diversa è orientata all'analisi, modellazione e supporto dell'attività di organizzazione concettuale umana. I primi possono essere di grande utilità, anche per queste ultime; un'integrazione ben definita sembra la ricetta per il prossimo futuro [ved. D01].
Il nostro lavoro non utilizza metodi statistici o neurali; riguardo a questi problemi la nostra prima attenzione è stata rivolta alla generazione di ipertesti altamente portabili e di modalità di presentazione atte a facilitare la leggibilità e la scoperta del significato da parte di umani, in una generalità di complesse strutture di documentazione come schemi di classificazione, terminologie, collezioni di metadati, ecc. Stiamo utilizzando con particolare attenzione una modalità di presentazione che permette di muoversi avanti e indietro attraverso viste parallele su una stessa struttura o su strutture simili; questo si dimostra molto utile nel nostro ambito.
Tali ipertesti sono prodotti principalmente da un gruppo di programmi in linguaggio C standard, che trattano solo file ASCII sequenziali e sono dedicati all'analisi e alla trasformazione di testi specifici e alla generazione di gruppi di pagine HTML sintatticamente semplici, ma altamente connesse e arricchite di procedure Javascript (*H-volumi*).
Sono state realizzate in questo modo diverse presentazioni ipertestuali a frame dell'ultima versione di Mathematics Subject Classification, MSC2000; queste presentazioni sono raccolte nella *Pagina della classificazione matematica* presso il server del Dipartimento di Matematica pura e applicata dell'Università di Padova (indirizzo Web: www.math.unipd.it/~biblio/math/index.html ; un altro accesso presso il server dell'Istituto IAMI-CNR di Milano, indirizzo Web: www.iami.mi.cnr.it/~alberto/a0msc.htm). Ecco una lista dei

gruppi di pagine HTML, insieme con l'indicazione delle loro caratteristiche e dell'indirizzo di accesso:

Gruppo 1: Presentazione semplice a frame
www.math.unipd.it/~biblio/math/mainb/mhbmain.htm

Gruppo 2: Presentazione a doppia visione
www.math.unipd.it/~biblio/math/doppiaeng/mhdmain.htm

Gruppo 3: Presentazione semplice a frame, con l'indicazione dei cambiamenti rispetto a MSC 1991
www.math.unipd.it/~biblio/math/complexc/mhcmain.htm

Gruppo 4: Presentazione semplice a frame, con collegamenti a pagine specifiche di siti Web rilevanti
www.math.unipd.it/~biblio/math/complexc/mhcmain.htm

Gruppo 5: Presentazione semplice a frame, traduzione italiana
www.math.unipd.it/~biblio/math/italiana/mhimain.htm

Gruppo 6: Presentazione semplice a frame, testi inglese e italiano interlineati
www.math.unipd.it/~biblio/math/it+eng/mhlmain.htm

Siamo andati avanti su questa linea abbozzando delle connessioni fra i codici delle classificazioni Dewey, 21. ed. e MSC2000; è stata quindi prodotta una bozza di pagina a doppia visione, esaminabile all'indirizzo Web: www.math.unipd.it/~biblio/msc-cdd/index.html .
In vista della revisione della sezione 510 della CDD, *Matematica*, stiamo aggiornando questa bozza secondo la proposta presentata da Giles Martin, Assistant Editor della Classificazione Decimale Dewey, che si può vedere all'indirizzo Web:
www.oclc.org/dewey/updates/discussion/doc/request_for_comment.htm .

Bus nello spazio della classificazione

Storicamente, le classificazioni di soggetto per documenti devono la loro struttura moogerarchica, ossia ad albero, a un'abitudine consolidata nel collocare a scaffale e ricercare documenti materiali nelle biblioteche, dove uno spazio fissato è definitivamente, o quasi permanentemente, diviso in sezioni e sottosezioni, e cambiare collocazione costa, tenendo presente che un oggetto materiale non può trovarsi simultaneamente in diversi luoghi. Tale abitudine consiste nel procedere a scelte per passare da argomenti generali a argomenti specifici, e quindi da grandi divisioni dello spazio a divisioni sempre più piccole, comunque senza sovrapposizioni. Anche nel mondo digitale questa abitudine è mantenuta e i sistemi per la gestione dei file presentano organizzazioni ad albero.
Tuttavia, gli spazi e gli oggetti della conoscenza e dell'attività umana mettono in evidenza strutture che sono molto più complesse e dinamiche di un semplice albero fisso. All'incrocio di intelligenza artificiale, linguistica computazionale e teoria delle basi di dati, queste strutture possono essere rappresentate con buona efficacia nel quadro di riferimento dell'Ontologia formale [G98], tramite formalismi come i grafi concettuali (CG) [S84; per un'applicazione di una variante of CG, ved. GMV99], le Description Logic (DL) [JoLC99; CDLNR98; ved. *The DL Website* all'indirizzo: www.ida.liu.se/labs/iislab/people/patla/DL/index.html], e lo Unified Modeling Language (UML) [BJR98], che viene dal campo dell'ingegneria del software e viene proposto come un approccio per la modellazione di ontologie e la codificazione del contenuto di conoscitivo di pagine Web [C01].

Con questo orientamento si stanno giungendo progressivamente alla definizione di formati di metadati per la rappresentazione di documenti; la bozza dell'*Academic Metadata Format*, in via di definizione nell'ambito della Open Archives Initiative [visibile all'indirizzo Web: openlib.org/amf/doc/ebisu.html] è un chiaro esempio di questa tendenza. Ulteriore informazione su questo campo si può trovare in [H01]

Come ci imbattiamo nelle classificazioni di soggetto, tali rappresentazioni strutturate (concepibili anche in tempi in cui mancavano linguaggi formali per esprimerle) devono essere drasticamente ridotte per avere qualche compatibilità con le forme ad albero in cui le classificazioni di soggetto vincolano la loro operabilità. Sebbene questa riduzione comporta inevitabilmente serie perdite di informazione, le classificazioni di soggetto sono state fornite di artifici più o meno efficaci per mettere rimedio a questa difficoltà.

Dal lavoro pionieristico di Ranganathan dal 1933 con la Classificazione Colon, attraverso le elaborazioni del Classification Research Group britannico negli anni '50 e '60, l'aggiunta delle Tavole Ausiliarie alla Classificazione Decimale Dewey dalla 18. ed., pubblicata nel 1971 [CCMS96], lo sviluppo del Preserved Context Indexing System (PRECIS) negli anni '70, e la pubblicazione nel 1986 dello standard ISO 2788 (BS 5723) *Guidelines for the establishment and development of monolingual thesauri*, si è progressivamente consolidato, con il nome di *analisi a facette*, un approccio composizionale all'analisi di soggetto [F96]. Nell'analisi a faccette, i concetti complessi vengono decomposti in combinazioni specificate di elementi atomici, che appartengono a classi omogenee, reciprocamente esclusive, le *faccette* [AG87].

Ritornando alle classificazioni di soggetto, un'organizzazione dello spazio della classificazione (chiamata *precoordinazione*), la quale permette che oggetti complessi siano recuperati tramite indirizzi composti opportunamente, e un apparato più o meno organizzato di richiami e rinvii fra i luoghi, risultano mezzi utili specialmente se gli oggetti non possono essere collocati in più luoghi simultaneamente.

Se una classificazione di soggetto è usata in contesti che permettono l'uso simultaneo di differenti codici di classificazione per lo stesso oggetto, vengono forniti meccanismi e indicazioni per la *postcoordinazione*, allo scopo di recuperare parzialmente un significato complesso mettendo insieme in lista indirizzi in modi opportuni, sia nelle basi di dati che offrono informazioni, sia nelle interrogazioni che richiedono informazioni.

Senza illudersi di dare un taglio alle sfumature del linguaggio naturale, come i thesauri mostrano di fare, la base per identificare gli oggetti a cui una o più classificazioni fanno riferimento, nel contesto dell'organizzazione della conoscenza delle persone che usano quelle classificazioni, resta ancora un'analisi dei testi descrittivi, insieme con la ricognizione delle dinamiche strutturali nel tempo delle stesse classificazioni. Segnali importanti per queste operazioni di identificazione sono dati da identità o somiglianze nel testo delle descrizioni in una classificazione (scoperte dalla stessa versione – sincronicamente – or da diverse versioni nel tempo – diacronicamente –), o attraverso diverse classificazioni (nella stessa lingua), sebbene una parola o locuzione può avere significati diversi, anche all'interno della stessa classificazione.

Perciò il primo passo nel processo di tirar fuori gli oggetti dalla loro busta, cioè lo spazio della classificazione, è di riconoscere la busta come una struttura che si sviluppa nel tempo attraverso una sequenza di versioni successive della classificazione, muovendosi fra gli indirizzi che segnano i percorsi e i luoghi in tale spazio. Il prodotto di questa fase consiste di sequenze di descrizioni per i *bus* che corrono nello spazio-tempo della classificazione. Ogni bus durante la sua corsa attraversa uno o più luoghi; gli indirizzi di questi luoghi, insieme con l'indicazione del periodo di passaggio, costituiscono la tabella di corsa di quel bus.

Per produrre dei risultati su questa linea guida, abbiamo iniziato un'analisi dell'intero corpus della Mathematics Subject Classification, in tutto il corso della sua evoluzione dal 1959, così come è disponibile per le ricerche in linea nel database *MathSci*. Ne è venuto fuori un database relazionale, per il quale stiamo specificando una presentazione Web da realizzare con un adattamento dei programmi C già sviluppati.

L'identificazione e la descrizione degli oggetti fuori dello spazio dei bus

Anche se ogni sezione sincronica dello spazio dei bus è ad albero, l'intera struttura può non essere ad albero, dato che nodi o sottoalberi possono migrare da un ramo a un altro; oltre alla struttura gerarchica principale, richiami e rinvii, e i meccanismi esplicitamente dichiarati di precoordinazione e postcoordinazione, anch'essi considerati dinamicamente, danno contributi sostanziali alla definizione dello spazio-tempo della classificazione.
Il passo ulteriore del processo di identificazione degli oggetti è l'estrazione di elementi concettuali dalle descrizioni. Si dovrebbe quindi attuare un'approfondita analisi relazionale per mezzo di un linguaggio di rappresentazione adeguato.
Un aiuto efficace in questa fase viene dato da strumenti per l'analisi testuale. In questa direzione, abbiamo prodotto alcuni H-volumi che presentano le locuzioni estratte dalle descrizioni in coppie di classificazioni, come CDD, 21. ed., e MSC2000: le locuzioni sono permutate circolarmente su parole significative, cioè formano una lista KWIC. Questa presentazione, ridondante ma opportunamente impaginata, permette l'esplorazione rapida delle somiglianze lessicali fra le descrizioni, indirizzando all'individuazione delle affinità di contenuto. Proseguiremo nello sviluppo di questo tipo di supporto lessicale preliminare, investigando le connessioni fra diversi gruppi di schemi di classificazione.
Indagheremo inoltre su come si possano ottenere miglioramenti aggiungendo alle locuzioni la possibilità di discriminare omonimi, nonché sinonimi e termini secondari

Le descrizioni degli oggetti nella macchina dei metadati

Mathematics Subject Classification è uno dei sistemi di classificazione previsti dal formato di metadati Dublin Core (DC), ed è utilizzata nei metadati DC per il motore di ricerca sviluppato nel progetto dell'Unione Europea *European Libraries and Electronic Resources in Mathematical Science* (EULER) – indirizzo Web: www.emis.de/projects/EULER/ .
L'obiettivo principale di EULER è stato la realizzazione di un "one-stop shop" per la ricerca di risorse informative per la matematica quali libri, preprint, pagine Web, abstract, collezioni di articoli e recensioni, periodici, rapporti tecnici e tesi. Il risultato è una meta-interfaccia Web per interrogazioni parallele dirette a una collezione eterogenea di database.
Una strategia simile potrebbe essere messa in campo per la connessione delle classificazioni: le descrizioni degli oggetti identificati a partire da diverse classificazioni, opportunamente codificate, potrebbero entrare a far parte dei metadati gestiti dal motore di ricerca.

RIFERIMENTI BIBLIOGRAFICI

AG87 J. Aitchison, A. Gilchrist, "Thesaurus construction: a practical manual", ASLIB, 1987

BJR98 G. Booch, L. Jacobson, J. Rumbaugh, "The Unified Modeling Language User Guide", Addison-Wesley, 1998

C01 S. Cranefield, *Networked Knowledge Representation and Exchange using UML and RDF*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Cranefield/

CCMS96 L.M. Chan, J.P. Comaromi, J.S. Mitchell, M.P. Satija, "Dewey Decimal Classification: a practical guide. 2nd ed., revised for DDC 21", OCLC Online Computer Library Center, 1996

CDLNR98 D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati, *Description logic framework for information integration*, "Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98)", Morgan Kaufman, 1998. p. 2-13

D01 M. Doerr, *Semantic Problems of Thesaurus Mapping*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

F96 A.C. Foskett, "The Subject Approach to Information", 5th ed., Library Association Publishing, 1996

G98b N. Guarino, *Formal Ontology and Information Systems*, "Formal Ontology in Information Systems: proceedings of FOIS'98", N. Guarino (ed), IOS Press, 1998. p 3-15

GMV99 N. Guarino, C. Masolo, G. Vetere, *Ontoseek: Content-based Access to the Web*, "IEEE Intelligent Systems", 14(3), 1999. p. 70-80

H01 J. Hunter, *MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains*, "Journal of Digital Information", 1(8), 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/

JoLC99 "Journal of Logic and Computation" – Vol. 9, No. 3: "Special Issue on Description Logics"

S84 J. Sowa, "Conceptual Structures: Information Processing in Minds and Machines", Addison-Wesley, 1984