

# Coordinamento semantico

L. Serafini<sup>1</sup> and S. Zanobini<sup>2</sup>

<sup>1</sup> serafini@itc.it <http://www.sra.itc.it/people/serafini>  
ITC-IRST, via Sommarive 18, 38050 Povo (TN), Italy

<sup>2</sup> zanobini@dit.unitn.it <http://dit.unitn.it/zanobini>  
Dipartimento di Informatica e Telecomunicazioni Università di Trento  
Via Sommarive, 10 – 38050 Trento (Italia)

**Abstract.** Semantic coordination, namely the problem of finding an agreement on the meaning of heterogeneous semantic models, is one of the key issues in the development of the Semantic Web. In this paper, we propose a new approach for discovering semantic mappings across these heterogeneous models, approach which shifts the problem of semantic coordination from the problem of computing linguistic or structural similarities (what most other proposed approaches do) to the problem of deducing relations between sets of logical formulae that represent the meaning of concepts belonging to different models. Finally, we show how to apply the approach and the algorithm to an interesting family of semantic models, namely hierarchical classifications.

**Sommario.** Il coordinamento semantico, cioè il problema di trovare un accordo sul significato da attribuire a modelli semantici eterogenei, è una delle questioni chiave nello sviluppo del Semantic Web. In questo articolo, proponiamo un nuovo approccio per scoprire relazioni semantiche tra tali modelli eterogenei, approccio che sposta il problema del coordinamento semantico dal problema di computare similarità linguistiche o strutturali (ciò che fa la maggior parte degli approcci proposti) a quello di dedurre relazioni tra insiemi di formule logiche che rappresentano il significato di concetti appartenenti a differenti modelli. Mostriamo infine, definendo un algoritmo, come applicare tale approccio ad un interessante famiglia di modelli semantici, cioè le classificazioni gerarchiche.

## 1 Introduzione

Internet e il web sono al momento fonte di numerosissime informazioni e risorse a livello mondiale: le pagine web superano infatti abbondantemente il miliardo. Si fa pertanto sempre più pressante l'esigenza di strutturare tale massa di informazioni in modo tale da permettere all'utente di trovare ciò che realmente cerca (e non ciò che gli assomiglia).

A seguito di tale esigenza stiamo assistendo in questi ultimi anni ad una evoluzione del web verso quello che viene definito 'Semantic Web'. Ciò che differenzia quest'ultimo dal semplice *web* è la possibilità di associare una semantica ai dati che sono contenuti nella rete, rendendone così espliciti (e quindi utilizzabili da applicazioni) i relativi significati. Per esempio, nella home-page di un utente dovrebbe essere rappresentato esplicitamente il fatto che una certa stringa di caratteri (ad esempio 'Paolo') costituisce il nome della persona, un'altra (ad esempio 'via Torino') l'indirizzo della sua abitazione

e un'altra ancora (ad esempio 'via Inama') l'indirizzo del suo ufficio. L'arricchimento dei dati con tali significati espliciti impone la necessità di definire precedentemente l'insieme dei significati utilizzabili e le loro proprietà, cioè un' *ontologia*. Un tipico esempio di ontologia è costituito da un insieme di assiomi che asseriscono che una persona ha un nome, un cognome, un indirizzo di residenza e un lavoro, che a sua volta è svolto in un ufficio che ha un indirizzo e così via. Per questo motivo, uno degli standard proposti dal W3C riguarda i linguaggi per la definizione di ontologie: DAML-OIL e OWL. Un secondo esempio di arricchimento semantico è costituito dalle *classificazioni*: spesso si trovano dei siti web il cui unico scopo è quello di organizzarne altri secondo un certo schema classificatorio (si pensi ad esempio alle web-directories di Google<sup>TM</sup>); in altri casi, invece, molti siti offrono prodotti o servizi, che possono essere individuati tramite la navigazione in apposite gerarchie. Nel resto dell'articolo indicheremo con il nome di *modelli* o *schemi* le infrastrutture, come le ontologie o le classificazioni, che vengono utilizzate per aggiungere informazione semantica al web.

Ovviamente, spesso accade che modelli che descrivono la stessa porzione del mondo, essendo sviluppati autonomamente da diversi utenti, non coincidano e presentino per contro molte disomogeneità. Una questione chiave è quindi quella di sviluppare procedure automatiche che consentano di utilizzare tali modelli eterogenei. Ad esempio, si considerino due agenti, software o umani, ognuno dei quali abbia una propria rappresentazione formale (e autonoma) di una qualche conoscenza. Si immagini, a titolo d'esempio, che due agenti possiedano le due diverse – autonomamente sviluppate – classificazioni di immagini descritte nella parte sinistra di figura 1. Si immagini inoltre che i due agenti abbiano la necessità di comunicare tra loro. Uno dei modi per rendere possibile la comunicazione è quello di trovare una sorta di accordo su ciò che le rappresentazioni locali riguardano e, conseguentemente, sulle relazioni intercorrenti tra le due rappresentazioni. Una possibile soluzione sarebbe riuscire a comprendere che i due nodi MONTAGNE nelle due classificazioni hanno un qualche tipo di relazione (descrivono cioè una comune porzione di mondo). A nostro avviso ciò può essere considerato un problema di *coordinamento semantico*<sup>3</sup>, e può essere definito come *il problema di trovare le 'giuste' relazioni semantiche esistenti tra i modelli locali degli agenti*.

In ambienti con domini e linguaggi ben definiti il problema del coordinamento semantico può essere risolto con la definizione e l'uso di modelli condivisi da tutti i membri<sup>4</sup> (ad esempio, ontologie). In tale approccio le 'giuste' relazioni semantiche sono definite per mezzo di una struttura mediana che è tale in due sensi: (i) rappresenta una visione del mondo comune (potrebbe essere l'unione delle due, oppure l'intersezione, o altro); (ii) i modelli locali non sono direttamente in relazione tra loro, ma passano attraverso i modelli definiti da tale struttura<sup>5</sup>. Quindi, in sostanza, si definirà una sorta di rappresentazione *intersoggettiva* della porzione di mondo sotto esame su cui le rappresentazioni locali saranno mappate.

<sup>3</sup> Vedi l'introduzione di Bouquet (2002) per questa nozione e la sua relazione con la nozione di *meaning negotiation*.

<sup>4</sup> Ma vedi Bonifacio (2002) per una discussione sulle conseguenze di questo approccio dal punto di vista delle applicazioni knowledge management.

<sup>5</sup> In letteratura tali approcci sono essenzialmente di due tipi: GAV (Global as View) e LAV (Local as View), dove la struttura globale è, rispettivamente, visibile e invisibile ai partecipanti.

Sebbene tale approccio ‘centralizzato’ abbia indubbiamente il vantaggio di una maggiore velocità nel reperire le informazioni, in un ambiente aperto, come lo è certamente il web, non è di fatto perseguibile. Per molte ragioni: la difficoltà di *negoziare* un modello condiviso che soddisfi le necessità di tutte le parti in gioco (considerato il gran numero di partecipanti), l’impossibilità pratica di mantenere un tale modello in un ambiente altamente dinamico (dove cioè ogni partecipante è libero di cambiare i propri modelli locali quando vuole), il problema di trovare soddisfacenti relazioni tra preesistenti modelli locali dentro il modello *globale*<sup>6</sup>. A causa di questi problemi, preferiamo cercare di risolvere il problema definendo una procedura automatica che permetta di mettere in relazione *diretta* i modelli locali degli agenti, cioè una forma di coordinamento semantico da agente ad agente (senza alcun modello globale).

In questo articolo non affrontiamo il problema generale del coordinamento semantico, ma una sua importante istanza, ovvero il coordinamento di *classificazioni gerarchiche* (HCs), strutture cioè che hanno l’esplicito scopo di organizzare/classificare qualche tipo di dato (ad esempio documenti). Tale problema di coordinamento è significativo per almeno due ragioni:

- innanzitutto, le HCs sono largamente usate in molte applicazioni<sup>7</sup>. Alcuni esempi sono: web directories (vedi ad esempio la directory di Google<sup>TM</sup> o di Yahoo!<sup>TM</sup>), portali (che spesso usano classificazioni gerarchiche per organizzare documenti e pagine web), marketplaces (i beni sono classificati in cataloghi gerarchici), i file systems dei PC (dove i files sono tipicamente classificati in cartelle gerarchicamente strutturate);
- inoltre, la maggior parte delle HCs disponibili sul web sono costruite usando strutture le cui etichette sono espressioni del linguaggio naturale, ma anche linguaggi tecnici, neologismi, nomi propri, etc., il cui significato è altamente condiviso. Questa conoscenza condivisa è essenziale per sfruttare il complesso grado di coordinamento semantico implicito nel modo in cui una comunità utilizza il linguaggio dal quale le etichette sono tratte.

Rispetto ad altri approcci proposti in letteratura (sotto altri nomi, come ad esempio schema matching, ontology mapping, integrazione semantica – vedi Zhang (1995); Wang (1994); Pelillo (1998); Milo (1998); Carroll (2002); Madhavan (2001); Bergamaschi (1999); Madhavan (2002)), l’approccio qui descritto si distingue dai precedenti almeno in tre aspetti: (1) la definizione di un nuovo metodo per rendere espliciti i significati dei nodi in una HC (ed in generale in modelli semantici strutturati) combinando tre diversi tipi di conoscenza, ognuna delle quali assume uno specifico ruolo; (2) il risultato dell’applicazione di tale metodo è che tutta la conoscenza rilevante riguardo ad un nodo è codificata in una formula logica; (3) le relazioni esistenti tra due nodi di due HCs sono dedotte per mezzo di un ragionamento logico, invece che derivate attraverso

<sup>6</sup> Quest’ultimo problema, in particolare, oltre ad esserlo da un punto di vista pratico, lo è anche dal punto di vista concettuale: il fatto che debba essere l’agente a dover cambiare i propri modelli locali in funzione di quello globale, e non viceversa, limita l’eterogeneità semantica che, a nostro avviso, rappresenta una ricchezza nel web.

<sup>7</sup> Per una interessante discussione del ruolo delle classificazioni nella cognizione umana si veda, ad esempio, Lakoff (1987); Bowker (1999).

più o meno complesse euristiche. Diversamente da altri approcci, cerchiamo di risolvere il problema del coordinamento per mezzo di una *procedura interamente semantica*: cerchiamo cioè di trasformare i nodi di una HC (che possiamo considerare elementi sintattici) in elementi semantici, capaci di approssimare il significato inteso di un certo nodo. Dopo aver costruito tali elementi semantici, il problema del coordinamento diviene il problema di dedurre relazioni tra formule che rappresentano il significato di ogni concetto associato ad un nodo in una HC, invece del problema di calcolare similarità linguistiche e strutturali (magari con l'aiuto di un dizionario o di altre informazioni sul tipo di archi tra i nodi).

Il maggior contributo dell'articolo è la definizione di un algoritmo denominato CT-XMATCH per coordinare HCs. Tale algoritmo prende in input due HCs (ad esempio  $H$  e  $H'$ ) e, per ogni coppia di nodi  $k \in H$  e  $k' \in H'$ , restituisce una *relazione semantica*. Una relazione semantica tra due nodi di un HC è una tra le seguenti relazioni:

- $k$  è più generale di  $k'$ : intuitivamente  $k$  corrisponde ad un concetto che è più generale di quello descritto da  $k'$ . Ad esempio “Vacanze” e “Vacanze economiche”;
- $k$  è più specifico di  $k'$ : intuitivamente  $k$  corrisponde ad un concetto che è più specifico di quello descritto da  $k'$ . Ad esempio “Vacanze economiche” e “Vacanze”;
- $k$  e  $k'$  sono equivalenti: intuitivamente  $k$  e  $k'$  descrivono lo stesso concetto (possibilmente in modo diverso). Ad esempio “Vacanze economiche” e “Ferie a basso costo”;
- $k$  e  $k'$  sono opposti (o disgiunti): intuitivamente  $k$  e  $k'$  descrivono due concetti, l'uno l'opposto dell'altro. Ad esempio “Vacanze costose” e “Vacanze economiche”;
- $k$  e  $k'$  sono compatibili: intuitivamente i concetti descritti da  $k$  e  $k'$  sono compatibili. Ad esempio “Vacanze costose” e “Alberghi a cinque stelle”.

L'articolo segue in questo modo: nella sezione 2 introduciamo le principali assunzioni del nuovo approccio che proponiamo per il coordinamento semantico. La sezione 3 mostra come questo approccio possa essere istanziato nel problema di coordinare HCs. Infine, presenteremo le principali caratteristiche dell'algoritmo (sezione 4).

## 2 L'approccio

L'approccio al coordinamento semantico descritto in questo articolo è basato sull'intuizione che esiste una differenza concettuale tra coordinare generiche strutture astratte (per esempio grafi etichettati arbitrariamente) e coordinare strutture le cui etichette sono prese dal linguaggio naturale. Quest'ultimo costituisce infatti il background comune ad una larga comunità di utenti del web. La semantica delle parole di tale linguaggio, inoltre, si trova codificata in alcuni artefatti, come ad esempio i dizionari e le ontologie, che forniscono, rispettivamente, l'insieme di concetti che sono esprimibili con una parola e le relazioni che intercorrono tra i concetti. Ad esempio, un dizionario ci dirà che la parola ‘cane’ in italiano può esprimere senz'altro almeno 2 concetti (‘cane’ nel senso di animale e ‘cane’ della pistola) e, inoltre, un'ontologia ci dirà che il concetto ‘cane’ nel senso di animale ha una relazione con il concetto di ‘mammifero’ (‘tutti i cani sono mammiferi’). Il nostro scopo è sfruttare questi artefatti come una sorgente di vincoli sulle possibili/accettabili relazioni tra i concetti espressi dai nodi di due HCs.

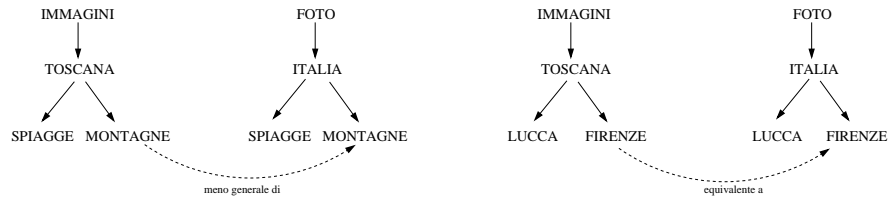


Figura 1. Coordinamento tra HCs

Per chiarire tale intuizione, si considerino le due HCs di figura 1. Tali strutture possono essere considerate come rappresentazioni locali ed autonome di due agenti su come classificare un insieme di foto. Si immagini inoltre che vogliamo scoprire la relazione semantica tra i concetti espressi dai nodi etichettati con MONTAGNE nei due HCs di sinistra e la relazione semantica esistente tra i concetti espressi dai due nodi FIRENZE delle HCs di destra. Utilizzando la conoscenza riguardo al significato delle parole usate nelle etichette e riguardo al mondo, noi comprendiamo immediatamente che la relazione tra la prima coppia di nodi è ‘meno generale di’ (intuitivamente, le foto che un agente classificherebbe come immagini di montagne toscane è un sottoinsieme delle foto che uno classificherebbe come foto di montagne italiane) e che la relazione tra la seconda coppia di nodi è ‘equivalente a’ (intuitivamente, le foto che un agente classificherebbe come immagini di Firenze in Toscana sono le medesime che classificherebbe come foto di Firenze in Italia). Da notare che la relazione trovata è diversa, a fronte di due strutture equivalenti (dal punto di vista strutturale/sintattico). Come possiamo definire una tecnica di coordinamento semantico che sfrutti queste conoscenze per raggiungere un tale risultato? L’approccio che proponiamo è basato su tre idee di fondo.

**1. Tre livelli di conoscenza** Per sfruttare il grado di coordinamento implicito nel fatto che le etichette sono prese dal linguaggio naturale è necessario rendere espliciti i significati (impliciti) espressi dalle etichette associate ad ogni nodo nella HC. Si possono individuare tre distinti livelli di conoscenza necessaria per costruire una adeguata approssimazione del significato associato ad un nodo di una struttura:

- conoscenza lessicale:** riguarda la competenza linguistica, ovvero la conoscenza del significato delle parole usate nelle etichette. Per esempio, la conoscenza che la parola ‘Italia’ può essere usata per denotare, ad esempio, la squadra di calcio o la nazione;
- conoscenza del mondo:** riguarda la conoscenza del mondo, ovvero le relazioni esistenti tra i concetti espressi nelle etichette nel mondo reale. Per esempio il fatto che Firenze e Lucca siano, nel mondo reale, due città della Toscana e dell’Italia;
- conoscenza strutturale:** riguarda la conoscenza che può essere derivata dal posizionamento delle etichette all’interno di una HCs. Per esempio, il fatto che l’etichetta FIRENZE sia posizionata sotto l’etichetta IMMAGINI ci aiuta a comprendere che, ovviamente, dentro il nodo FIRENZE molto probabilmente sono classificate immagini e non, ad esempio, libri.

Vediamo ora come questi tre livelli possano essere usati per spiegare il ragionamento intuitivo descritto precedentemente. Si considerino i due nodi MONTAGNE nelle strutture di sinistra. La conoscenza linguistica può essere usata per assumere che il concetto associato al nodo IMMAGINI e al nodo FOTO sia lo stesso. La conoscenza del mondo ci dice, tra l'altro, che la Toscana è una regione dell'Italia. Infine, la conoscenza strutturale ci dice che gli intesi significati dei due nodi MONTAGNE sono rispettivamente: 'immagini delle montagne toscane' (HC di sinistra) e 'foto delle montagne italiane' (HC di destra). Tutti questi fatti insieme ci permettono di concludere che il concetto espresso dal primo nodo è meno generale di quello espresso dall'altro nodo. Possiamo fare un ragionamento simile per i due nodi FIRENZE delle due HCs di destra, strutturalmente equivalenti. Ma esplorando la conoscenza del mondo, possiamo aggiungere che Firenze è situata in Toscana (relazione che ovviamente non esiste tra il concetto 'montagna' e il concetto 'Toscana' nel precedente caso). Questo ulteriore pezzo di conoscenza ci permette di concludere che, al di là della equivalenza strutturale, la relazione è diversa.

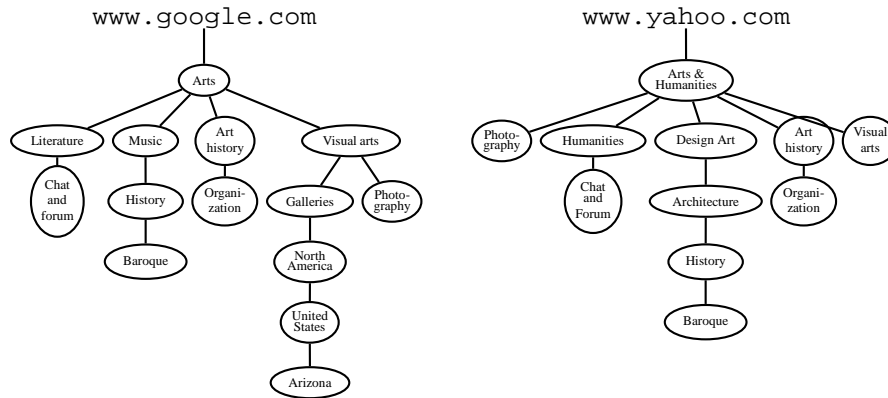
**2. Codifica** A questo punto, dobbiamo trovare un modo di sfruttare questi tre livelli di conoscenza. Gli approcci che conosciamo utilizzano talvolta qualcuna di queste conoscenze (in genere la conoscenza strutturale e lessicale) come euristiche per dare una sorta di 'punteggio' per migliorare o peggiorare un risultato, precedentemente definito per mezzo di procedure non semantiche.

Volendo, nel nostro approccio, rimanere in ambito interamente semantico, la soluzione che proponiamo è quella di combinare tutti e tre i livelli di conoscenza per costruire una nuova rappresentazione del problema: il significato scaturito da questi tre livelli è codificato in una formula logica, che deve approssimare il significato inteso del nodo, e in un insieme di assiomi logici, rappresentanti la conoscenza del mondo rilevante per i concetti che stiamo analizzando.

**3. Deduzione** A questo punto, possiamo facilmente introdurre la terza novità: una volta codificato il significato espresso da ogni nodo in una formula logica e in un insieme di assiomi, il problema di scoprire la relazione semantica esistente tra due nodi diviene un (relativamente semplice) problema di deduzione logica. Intuitivamente (i dettagli tecnici sono presentati nella sezione 4), determinare il tipo di relazione esistente tra due nodi viene formulato come un problema di soddisfacibilità. Un semplice esempio chiarirà meglio la questione. Siano  $\phi$  e  $\psi$  due formule che approssimano il significato inteso dei nodi  $k$  e  $k'$  rispettivamente. Sia inoltre  $B$  l'insieme degli assiomi derivanti dalla conoscenza sul mondo (relativa ai concetti espressi nei nodi  $k$  e  $k'$ ). Se vogliamo sapere, ad esempio, se il nodo  $k$  sia in una relazione di 'meno generale di' con il nodo  $k'$ , dobbiamo solo verificare se  $B \models \phi \rightarrow \psi$ . Tecnicamente utilizziamo per tale verifica un SAT solver standard.

### 3 Coordinamento semantico di classificazioni gerarchiche

In questa sezione mostriamo come applicare l'approccio generale descritto nella precedente sezione al problema di coordinare HCs. Intuitivamente, una classificazione è un



**Figura 2.** Esempi di classificazioni gerarchiche (Google e Yahoo)

raggruppamento di oggetti in classi o categorie. Se le classi sono arrangiate gerarchicamente, allora abbiamo una classificazione gerarchica. Di fatto una HC è una struttura ad *albero* che classifica oggetti (più frequentemente documenti): l'esempio più semplice di tale struttura è il file system dei nostri PC. Assumiamo che gli oggetti vengano classificati secondo il seguente principio di *specificità*: un oggetto è classificato sotto un nodo  $k$  dell'albero, se tale oggetto riguarda  $k$  (secondo certi criteri, come ad esempio l'intuizione semantica del creatore della classificazione) e non esiste un nodo  $k'$  più specifico sotto il quale potrebbe essere classificato<sup>8</sup>. Alcuni tipici esempi di HCs che si possono trovare su internet sono le web directories di molti motori di ricerca, come ad esempio la directory di Google<sup>TM</sup>, la directory di Yahoo!<sup>TM</sup> oppure la web directory di Looksmart<sup>TM</sup>. Una piccola frazione delle directories di Google<sup>TM</sup> e di Yahoo!<sup>TM</sup> è riportata nella figura 2.

Intuitivamente, il problema del coordinamento semantico sorge allorché un agente ha la necessità di trovare relazioni semantiche tra nodi appartenenti a diverse HCs le quali sono tipicamente eterogenee (dal punto di vista semantico). Si consideri la seguente situazione: un agente sta navigando nella directory di Google<sup>TM</sup> (lato sinistro figura 2) e trova che i documenti classificati sotto il nodo etichettato *Baroque* sono interessanti per il suo lavoro sulla musica barocca. Quello che noi vogliamo ottenere è un sistema automatico che scopra se esistono altre categorie in altre HCs (nell'esempio la directory di Yahoo!<sup>TM</sup>) che abbiano il medesimo contenuto, e per fare questo ci chiediamo come trovare nodi in altre HCs aventi lo stesso significato di quello espresso dal nodo *Baroque*. Formalmente, definiamo il coordinamento semantico come *il problema di scoprire la 'giusta' relazione semantica esistente tra due nodi di due HCs*.

L'insieme delle relazioni possibili tra due nodi dipende dall'uso inteso della struttura che vogliamo mappare. Infatti, nella nostra esperienza, al fine di determinare l'inter-

<sup>8</sup> Vedi per esempio le istruzioni di Yahoo!<sup>TM</sup> su come 'trovare una categoria appropriata' all'indirizzo <http://docs.yahoo.com/info/suggest/appropriate.html>.

pretazione di un nodo in una struttura, l'uso inteso della struttura (come ad esempio il classificare oggetti) è semanticamente molto più rilevante che il tipo di struttura astratta (come ad esempio grafo e albero). Poiché lo scopo di una classificazione gerarchica è quello di classificare oggetti, i nodi di tale struttura saranno intesi descrivere insiemi di oggetti. Conseguentemente, le relazioni possibili intercorrenti tra due nodi  $k_s$  e  $k_t$  appartenenti a differenti HCs saranno di tipo insiemistico:  $k_s \supseteq k_t$  ( $k_s$  descrive un insieme di oggetti più generale di quello descritto da  $k_t$ );  $k_s \subsetneq k_t$  ( $k_s$  descrive un insieme di oggetti meno generale di quello descritto da  $k_t$ );  $k_s \equiv k_t$  ( $k_s$  e  $k_t$  descrivono lo stesso insieme di oggetti);  $k_s \overset{*}{\rightarrow} k_t$  ( $k_s$  descrive una parte degli oggetti descritti da  $k_t$  – intersezione);  $k_s \perp k_t$  ( $k_s$  e  $k_t$  descrivono insiemi disgiunti di oggetti).

## 4 L'algoritmo CTXMATCH

CTXMATCH prende in input due classificazioni gerarchiche e restituisce un insieme di relazioni semantiche tra i nodi delle due strutture. L'algoritmo è composto da due passi principali:

**Esplicitazione semantica:** in questo passo si procede ad esplicitare il significato inteso di ogni nodo della struttura associandogli una formula logica che approssimi tale significato.

**Comparazione semantica:** il problema di trovare relazioni semantiche tra i nodi di due strutture viene trasformato nel problema di trovare le *relazioni logiche* valide tra le formule associate ai nodi stessi.

In questa versione dell'algoritmo la conoscenza linguistica e quella del mondo sono rappresentate entrambe da WORDNET (vedi Fellbaum (1998)). In WORDNET ogni possibile concetto denotato da una parola viene denominato *sensò*. Nel prosieguo dell'articolo useremo quindi la parola 'sensò' come sinonimo di 'concetto'.

### 4.1 Esplicitazione semantica

Per rendere più chiara questa fase, procederemo illustrando preliminarmente un semplice esempio. Si consideri il nodo etichettato *Arizona* nella struttura di sinistra di figura 2. Concentrandosi soltanto sull'etichetta, ignorando cioè il contesto in cui essa occorre, possiamo dedurre, per mezzo della *conoscenza linguistica*, che essa può esprimere almeno due concetti: (i) 'Arizona' nel senso di stato degli USA, o (ii) 'Arizona' nel senso di serpente<sup>9</sup>. La *conoscenza strutturale*, come abbiamo già detto, riguarda la collocazione contestuale di un nodo, ovvero, nel caso delle classificazioni gerarchiche, la posizione di un nodo all'interno della struttura. In questo caso, tale conoscenza ci informa che il nodo etichettato con *Arizona* giace sul seguente percorso: *Arts . Visual arts . Galleries . North America . United States . Arizona*. Infine la *conoscenza del mondo* ci informa inoltre che l'*Arizona* (intesa come stato) è parte degli

<sup>9</sup> Esiste un serpente denominato *Arizona*.

USA. L'insieme delle tre informazioni appena illustrate ci permette innanzitutto di concludere, con un certo margine di approssimazione, che il nodo etichettato con *Arizona* è da interpretarsi nel senso di 'stato degli USA', e non nel senso di 'serpente' (infatti il nodo *ARIZONA* è figlio del nodo *UNITED STATES*). Inoltre si può dedurre che il significato inteso da associare al nodo 'contestualizzato' *Arizona* sarà una formula logica approssimante il significato espresso dall'enunciato 'Gallerie di arti visuali localizzate in Arizona, stato del nord degli Stati Uniti'.

In particolare, la fase di esplicitazione semantica di un nodo avviene tramite due sottofasi: l'*interpretazione linguistica* del nodo e la sua *contestualizzazione*. Vediamole ora in dettaglio.

**Interpretazione linguistica.** Durante la fase di interpretazione linguistica, ogni singola etichetta viene analizzata al fine di individuare l'insieme dei concetti denotati dall'etichetta stessa, insieme che viene poi trasformato in una formula logica<sup>10</sup>. Riportiamo nella tabella 1 alcuni esempi dell'analisi linguistica fatta sulle etichette di figura 2.

Etichetta	forma logica	Spiegazione
Baroque	baroque#1	L'unico senso di 'baroque' presente in WORDNET
Arizona	arizona#1 ∨ arizona#2	WORDNET contiene due sensi per Arizona: tale ambiguità viene rappresentata logicamente con la disgiunzione
Chat and Forum	(chat#1 ∨ chat#2 ∨ chat#3) ∨ (forum#1 ∨ forum#2)	La congiunzione 'and' tra 'chat' e 'forum' viene interpretata come una disgiunzione logica invece che come congiunzione perché sotto tale nodo possiamo trovare documenti che riguardano chats, forums o entrambi
Classic Music	((classic#1 ∨ ...) ∧ (music#1 ∨ ...)) ∨ classic_music#1	WORDNET contiene, oltre ai diversi sensi di 'classical' e 'music', anche un senso (multiword) per 'classical music'. L'etichetta può essere interpretata sia come combinazione di 'classical' e 'music' sia come un singolo concetto 'classical music'

Tabella 1.

**Contestualizzazione.** L'obiettivo di questa fase è duplice: in prima istanza si procede, se la conoscenza del mondo lo permette, al raffinamento dell'interpretazione linguistica con l'eliminazione dei sensi che non sono coerenti con il contesto dell'etichetta; successivamente, si arricchisce ogni interpretazione linguistica con l'informazione contestuale, cioè si combina l'interpretazione linguistica di un nodo con quella dei nodi superiori che ne influenzano il significato. La contestualizzazione avviene in due sottofasi: selezione e composizione dei sensi.

*Selezione dei sensi.* La selezione dei sensi compatibili viene attuata interrogando la conoscenza del mondo sull'esistenza di relazioni ontologiche fra il concetto espresso

<sup>10</sup> La scelta della logica in cui esprimere il significato delle etichette dipende da quando sono sofisticati il parser e l'interpretazione del suo output. Più espressiva è la logica più complessa sarà l'analisi linguistica. In questa prima versione dell'algoritmo si è scelto di usare una logica estremamente semplice, cioè la logica proposizionale.

da un nodo e i concetti espressi dai nodi soprastanti. In particolare, vengono selezionati i concetti che hanno un maggior numero di relazioni ontologiche con i concetti associati ai nodi soprastanti. Ad esempio, il senso `arizona#1` ('stato degli USA') sarà preferito al senso `arizona#2` ('tipo di serpente') in quanto WORDNET contiene una relazione *Part-Of* tra `arizona#1` e `united_states#1`, mentre nessuna relazione tra `arizona#2` e `united_states#1` può essere trovata.

*Composizione dei sensi.* Come già detto precedentemente, il significato di un nodo non dipende solo dalla sua etichetta, ma anche dal significato scaturente dalla sua posizione all'interno della struttura e quindi, in ultima analisi, dai significati associati ai nodi soprastanti. Infatti, in una classificazione gerarchica, l'interpretazione di un nodo  $n$  figlio del nodo  $n'$  è la specializzazione dell'interpretazione di  $n'$  rispetto a quella dell'etichetta di  $n$  (o viceversa). Per esempio, considerando le strutture di figura 2, il nodo etichettato `United States` occorrente sotto il nodo etichettato `Galleries` deve essere interpretato come 'Gallerie degli Stati Uniti'. Viceversa, il nodo etichettato `History` che occorre sotto il nodo etichettato `Music` deve essere interpretato come 'Storia della Musica'. Tale specializzazione può essere espressa nella logica proposizionale mediante l'operazione di congiunzione<sup>11</sup>. In questa fase, quindi, la formula di ogni nodo viene messa in congiunzione con le formule di tutti i suoi nodi superiori. Ad esempio il significato 'contestualizzato' del nodo etichettato `Chat and Forum` della struttura di sinistra di figura 2 è codificato come

$$\text{art\#1} \wedge \text{literature\#2} \wedge (\text{chat\#1} \vee \text{forum\#1})$$

## 4.2 Calcolo delle relazioni semantiche *via* ragionamento logico

Una volta terminata l'esplicitazione semantica di ogni nodo, il problema di trovare le relazioni semantiche esistenti tra i nodi di due strutture può essere ricondotto al problema di trovare le relazioni logiche esistenti tra le formule associate ai nodi delle strutture. Quindi, dati due nodi  $m$  e  $n$  con le rispettive semantiche esplicitate  $\phi$  e  $\psi$ , il nostro obiettivo diventa quello di trovare le relazioni logiche tra queste due formule.

Per relazione logica fra due formule si intende una correlazione tra i loro valori di verità – correlazione esprimibile mediante una formula logica costruita a partire dalle due formule iniziali. Tutte le possibili formule proposizionali costruibili a partire da due formule  $\phi$  e  $\psi$  sono mostrate nella prima colonna della tabella 2. Si noti come le prime 10 formule non esprimono nessuna relazione logica tra  $\phi$  e  $\psi$ , in quanto i loro valori di verità sono indipendenti. Le ultime 6 formule invece esprimono delle correlazioni logiche la cui interpretazione in termini di relazioni semantiche tra i nodi è riportata nella seconda colonna della tabella. Ovviamente, una relazione semantica esistente tra i significati espressi da due nodi ha effetto anche sulla relazione tra gli

<sup>11</sup> La codifica della specializzazione con la congiunzione logica presenta sicuramente svariate controindicazioni. Ma tale limite deriva dalla logica proposizionale stessa. L'utilizzo di una logica più potente, come ad esempio la logica descrittiva, permetterebbe di codificare tale specializzazione in modo certamente più adeguato – per esempio attraverso l'utilizzo degli attributi.

	Formula	Relazione semantica tra i nodi	Relazione tra i documenti classificati sotto i nodi
1	$\top$	Nessuna correlazione tra $m$ and $n$	Nessuna correlazione tra i documenti classificati sotto $m$ e $n$
2	$\perp$		
3	$\phi \wedge \psi$		
4	$\neg\phi \wedge \psi$		
5	$\phi \wedge \neg\psi$		
6	$\neg\phi \wedge \neg\psi$		
7	$\phi$		
8	$\psi$		
9	$\neg\phi$		
10	$\neg\psi$		
11	$\phi \rightarrow \psi$	$m$ è più specifico di $n$	Ogni (esiste un) documento classificato sotto $m$ potrebbe essere classificato sotto $n$
12	$\psi \rightarrow \phi$	$m$ è più generale di $n$	Ogni documento classificato sotto $n$ potrebbe essere classificato sotto $m$
13	$\phi \equiv \psi$	$m$ e $n$ sono equivalenti	Ogni documento classificato sotto $m$ potrebbe essere classificato sotto $n$ e viceversa
14	$\phi \vee \psi$	$m$ e $n$ coprono tutto il dominio	Ogni documento può essere classificato sotto $m$ o sotto $n$
15	$\phi \rightarrow \neg\psi$	$n$ e $m$ sono disgiunti	Ogni documento classificato sotto $m$ non può essere classificato sotto $n$
16	$\neg\phi \equiv \psi$	$m$ e $n$ sono opposti	Ogni documento classificato sotto $n$ non può essere classificato sotto $m$ , e viceversa ogni documento che non è classificabile sotto $m$ è classificabile sotto $n$

**Tabella 2.**

insiemi di documenti classificati sotto gli stessi nodi. Tali relazioni tra i documenti sono riportate nella terza colonna della tabella<sup>12</sup>.

Verificare la validità delle formule 11–16 di tabella 2 potrebbe però non essere sufficiente per trovare le relazioni semantiche tra due nodi  $m$  e  $n$ . Per aumentare le nostre possibilità di successo, possiamo comunque utilizzare alcune informazioni derivanti dalla conoscenza del mondo: infatti nelle formule  $\phi$  e  $\psi$  approssimanti, rispettivamente, il significato inteso dei nodi  $m$  e  $n$ , potrebbero essere presenti atomi che hanno tra loro una qualche relazione ontologica. Si considerino, ad esempio, le seguenti due formule  $\phi$  e  $\psi$  associate rispettivamente ai due nodi *Chat* and *forum* della struttura di sinistra e di destra di figura 2:

$$\phi = (\text{art}\#1 \wedge \text{literature}\#2 \wedge (\text{chat}\#1 \vee \text{forum}\#1))$$

$$\psi = ((\text{art}\#1 \vee \text{humanities}\#1) \wedge \text{humanities}\#1 \wedge (\text{chat}\#1 \vee \text{forum}\#1))$$

Per tali formule nessuna delle relazioni semantiche espresse dai punti 12–16 vale. Ma analizzando la conoscenza del mondo, possiamo trovare alcune relazioni esistenti tra gli atomi delle due formule: ad esempio che ‘la letteratura è una disciplina umanistica’ e che ‘l’arte è una disciplina umanistica’. Tali informazioni possono essere codificate rispettivamente nei seguenti assiomi:  $\text{literature}\#2 \rightarrow \text{humanities}\#1 (= \alpha)$  e  $\text{art}\#1 \rightarrow \text{humanities}\#1 (= \beta)$ . Le relazioni logiche tra le formule associate a due nodi potranno perciò essere verificate tenendo presente un insieme di assiomi ( $\alpha$  e  $\beta$ ) che esprimono la conoscenza del mondo. A questo punto è facile vedere come la relazione 11 valga tra  $\phi$  e  $\psi$  ( $\alpha \wedge \beta \models \phi \rightarrow \psi$ ).

<sup>12</sup> In questa colonna si usa la frase ‘un documento classificato sotto un nodo  $n$ ’ per indicare più precisamente ‘un documento classificato nel nodo  $n$  o in qualche discendente di  $n$ ’.

Nella presente versione di CTXMATCH, gli assiomi della conoscenza del mondo rispetto a due formule  $\phi$  e  $\psi$  vengono estratti interrogando WORDNET sull'esistenza di relazioni di *sinonimia*, *iperonimia*, *iponimia*, e *meronimia* (essenzialmente le relazioni *Is-A* e *Part-Of*) tra i termini che occorrono in  $\phi$  e  $\psi$ , secondo il seguente criterio:

1.  $s\#k \equiv t\#h$ : se  $s\#k$  e  $t\#h$  sono sinonimi (i.e., sono contenuti nello stesso synset di WORDNET);
2.  $s\#k \rightarrow t\#h$ :  $s\#k$  è un iponimo (= termine più specifico) o un meronimo (= termine che indica una parte) di  $t\#h$ ;
3.  $t\#h \rightarrow s\#k$ :  $s\#k$  è un iperonimo (= termine più generale) o un olonimo (= termine che indica l'intero) di  $t\#h$ ;
4.  $\neg t\#k \equiv s\#h$ : se  $s\#k$  è opposto (= termine che indica un significato opposto) di  $t\#h$ .

## 5 Conclusioni

Nel presente articolo abbiamo presentato un nuovo approccio al coordinamento semantico tra modelli eterogenei. Allo scopo di trovare le relazioni semantiche tra i singoli componenti di due modelli, si è utilizzata una combinazione di conoscenza linguistica, conoscenza del mondo e conoscenza strutturale. In particolare, è stato definito l'algoritmo CTXMATCH, capace di trovare possibili relazioni semantiche tra classificazioni gerarchiche. CTXMATCH è stato integrato in una applicazione peer-to-peer per il Knowledge Management distribuito (come descritto in Bonifacio (2002)) e sarà implementato, in un prossimo futuro, anche in una applicazione peer-to-peer wireless-system per ambienti intelligenti (Busetta (2003)).

È importante notare come l'approccio risulti più generale dell'algoritmo, e che quindi possa essere applicato uniformemente per il coordinamento semantico di strutture diverse dalla gerarchiche, come ad esempio schemi di basi di dati, ontologie, work-flows, etc.

L'aspetto innovativo dell'approccio presentato, rispetto allo stato dell'arte degli algoritmi di schema matching, è costituito dall'importanza data al fatto che gli schemi sono etichettati con parole (o brevi frasi) prese dal linguaggio naturale, che hanno un significato proprio, indipendente dallo schema. Ricostruire il loro significato (esplicitazione semantica) e arricchirlo con della conoscenza di dominio costituisce un punto di forza di questo approccio. Questo, d'altra parte, non significa che il problema del coordinamento semantico sia unicamente un problema di *Natural Language Processing*: al contrario, la soluzione proposta è basata principalmente su sistemi di rappresentazione della conoscenza e di ragionamento automatico. Una soluzione soddisfacente al problema del coordinamento semantico si può quindi ottenere soltanto attraverso una adeguata combinazione delle due discipline

Il lavoro presentato in questo articolo<sup>13</sup> costituisce solo un primo passo verso un progetto di ricerca più ampio e ambizioso, che ha come obiettivo finale lo studio della configurazione minima necessaria per permettere la comunicazione tra agenti autonomi

<sup>13</sup> Una descrizione più approfondita e tecnicamente più completa di questo lavoro è contenuta in Bouquet (2003).

che non siano capaci di vedersi l'un l'altro dentro la 'testa', dove cioè la rappresentazione del mondo non sia direttamente accessibile (come del resto avviene per gli agenti umani) cosicché il coordinamento semantico può essere raggiunto solo tramite altri mezzi (significativi), come scambiarsi documenti, scambiarsi pezzi di conoscenza, indicare oggetti, congetturare su passate comunicazioni e così via. A tal fine, rimane molto lavoro da svolgere. Da parte nostra, i passi successivi saranno i seguenti: estendere l'algoritmo oltre le classificazioni (cioè a strutture utilizzate con scopi diversi dal classificare oggetti); generalizzare il tipo di struttura che può essere trattata (per esempio, strutture con relazioni non gerarchiche, come ad esempio ruoli); andare oltre WORDNET come sorgente di conoscenza lessicale e di dominio; permettere diverse sorgenti lessicali e di conoscenza del mondo per ognuna delle strutture locali che devono essere coordinate.

### Ringraziamenti

Gran parte del materiale di questo articolo è contenuto in Bouquet (2002). Ringraziamo Paolo Bouquet, Antonia Donà, Laura Gatti, Christian Girardi, Bernardo Magnini e Manuela Speranza, per aver contribuito alla definizione e all'implementazione dell'algoritmo CTXMATCH descritto in Sezione 4. Questo lavoro è stato parzialmente finanziato dal progetto EDAMOK (*Enabling Distributed and Autonomous Management of Knowledge*, <http://edamok.itc.it>) della Provincia Autonoma di Trento N° 1060 del 4/5/2001.

### Bibliografia

- Sonia Bergamaschi, Silvana Castano, and Maurizio Vincini. "Semantic integration of semistructured and structured data sources". *SIGMOD Record*, 28(1):54–59, 1999.
- M. Bonifacio, P. Bouquet, G. Mameli, and M. Nori. "Kex: a peer-to-peer solution for distributed knowledge management". In D. Karagiannis and U. Reimer, editors, *Fourth International Conference on Practical Aspects of Knowledge Management (PAKM-2002)*, Vienna (Austria), 2002.
- M. Bonifacio, P. Bouquet, and P. Traverso. "Enabling distributed knowledge management. Managerial and technological implications". *Novatica and Informatik/Informatique*, III(1), 2002.
- P. Bouquet, editor. *AAAI-02 Workshop on Meaning Negotiation*, Edmonton, Canada, July 2002. American Association for Artificial Intelligence (AAAI), AAAI Press.
- P. Bouquet, L. Serafini, S. Zanobini, "Semantic coordination: a new approach and an application", in *Proceedings of ISWC 2003 conference on Semantic Web, Sanibel Island, Florida, USA*.
- G. C. Bowker and S. L. Star. "Sorting things out: classification and its consequences". MIT Press., 1999.
- P. Busetta, P. Bouquet, G. Adami, M. Bonifacio, and F. Palmieri. *K-Trek: An approach to context awareness in large environments*. Technical report, Istituto per la Ricerca Scientifica e Tecnologica (ITC-IRST), Trento (Italy), April 2003. Submitted to UbiComp'2003.

- A. Doan, J. Madhavan, P. Domingos, and A. Halevy. “Learning to map between ontologies on the semantic web”. In *Proceedings of WWW-2002, 11th International WWW Conference, Hawaii*, 2002.
- Christiane Fellbaum, editor. “WordNet: An Electronic Lexical Database”. The MIT Press, Cambridge, US, 1998.
- Jeremy Carroll Hewlett-Packard. “Matching rdf graphs”. In *Proc. in the first International Semantic Web Conference - ISWC 2002*, pages 5–15, 2002.
- G. Lakoff. “Women, Fire, and Dangerous Things”. *Chicago University Press*, 1987.
- Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. “Generic schema matching with cupid”. In *The VLDB Journal*, pages 49–58, 2001.
- Tova Milo and Sagit Zohar. “Using schema matching to simplify heterogeneous data translation”. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 122–133, 24–27 1998.
- Marcello Pelillo, Kaleem Siddiqi, and Steven W. Zucker. “Matching hierarchical structures using association graphs”. *Lecture Notes in Computer Science*, 1998.
- Jason Tsong-Li Wang, Kaizhong Zhang, Karpjoo Jeong, and Dennis Shasha. “A system for approximate tree matching”. *Knowledge and Data Engineering*, 6(4):559–571, 1994.
- K. Zhang, J. T. L. Wang, and D. Shasha. “On the editing distance between undirected acyclic graphs and related problems”. In Z. Galil and E. Ukkonen, editors, *Proceedings of the 6th Annual Symposium on Combinatorial Pattern Matching*, volume 937, pages 395–407, Espoo, Finland, 1995. Springer-Verlag, Berlin.