Università Degli Studi di Modena e Reggio Emilia

Facoltà di Ingegneria – Sede di Modena
Corso di Laurea in Ingegneria Informatica
Nuovo Ordinamento Didattico

TESI DI LAUREA DI PRIMO LIVELLO
Anno accademico 2003 – 2004

Sistema Momis: analisi sperimentale dell'integrazione di una nuova sorgente

Candidato:
Matteo Generali

Relatore:
Prof. Sonia Bergamaschi

Correlatore:
Ing. Francesco Guerra

Parole Chiave:

Integrazione Dati
Vista Virtuale Globale
Aggiunta Sorgenti
Approccio di Integrazione
Scenario di Integrazione

Ringraziamenti

Desidero ringraziare anzitutto la Professoressa Sonia Bergamaschi per il patrocinio e i consigli che ha saputo darmi nella stesura del presente documento. Un ringraziamento particolare va al Dottore Ingegnere Francesco Guerra per la costante disponibilità e per l'aiuto determinante fornitomi in tutto il periodo di ricerca. Ringrazio inoltre tutto il Data Base Group del dipartimento di Ingegneria Dell'Informazione di questa università per l'impegno nello sviluppo del materiale su cui verte questo documento.

Esprimo tutta la mia gratitudine ai miei famigliari per il loro costante supporto nell'intero arco dei miei studi universitari e la mia fidanzata Maria Giovanna per aver sopportato ogni giorno noiosi discorsi di ambito informatico.

Ringrazio infine tutti i miei colleghi e compagni di corso senza i quali non avrei potuto portare a termine questa esperienza di studio.

Sommario

Indice delle Figure

Indice delle Tabelle

1 Il sistema Momis

1.1 Introduzione

Momis è un sistema basato su un'architettura a mediatore che sfrutta l'estensione ODL_{i3} dello standard *Object Definition Language* per generare una GVV (vista virtuale globale) che esprime un'integrazione dello schema di sorgenti di dato eterogenee. Esso provvede a tradurre in una base comune la struttura di varie sorgenti indipendentemente dalla loro localizzazione ed espressione. Successivamente accompagna il progettista nella procedura di integrazione di queste strutture per generare una singola espressione delle stesse (GVV), in cui ogni elemento globale corrisponde ad una vista dell'insieme di elementi associati appartenenti alle sorgenti locali.

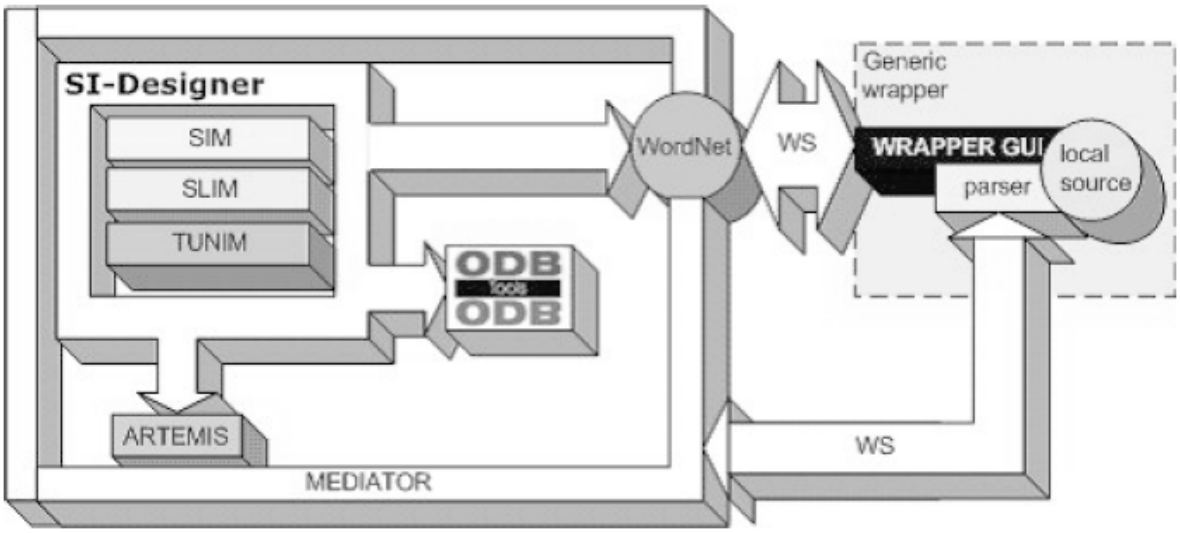

Figura 1.1: Architettura del sistema MOMIS

Il procedimento prevede diverse fasi:

- *Estrazione delle sorgenti locali*: wrapper dedicati per ciascuna struttura delle sorgenti ne estraggono lo schema e lo traducono in ODL_{i3};
- *Annotazione delle sorgenti locali*: il progettista sceglie un significato per ciascun elemento dello schema delle sorgenti locali, utilizzando il database lessicale WordNet;

- *Generazione del Common Thesaurus*: usufruendo dall'annotazione lessicale precedentemente realizzata, il sistema genera un insieme di relazioni inter e intra-schema tra le classi e gli attributi delle sorgenti locali;
- *Generazione della GVV*: MOMIS genera una vista virtuale globale e un insieme di mapping tra lo schema globale e le sorgenti locali basato sulle relazioni del Thesaurus;
- *Annotazione della GVV* : in modo semiautomatico, un significato viene affidato agli elementi globali che compongono la GVV. Il procedimento è analogo a quello utilizzato per l'annotazione delle sorgenti locali.

In ciascuna fase del processo è possibile l'interazione da parte del progettista tramite lo strumento grafico Momis Ontology Builder, un'interfaccia sviluppata per seguire il procedimento di integrazione delle sorgenti. Le fasi citate saranno ora descritte più nel dettaglio. La seguente immagine riassume la procedura di generazione della GVV adottata dal sistema.

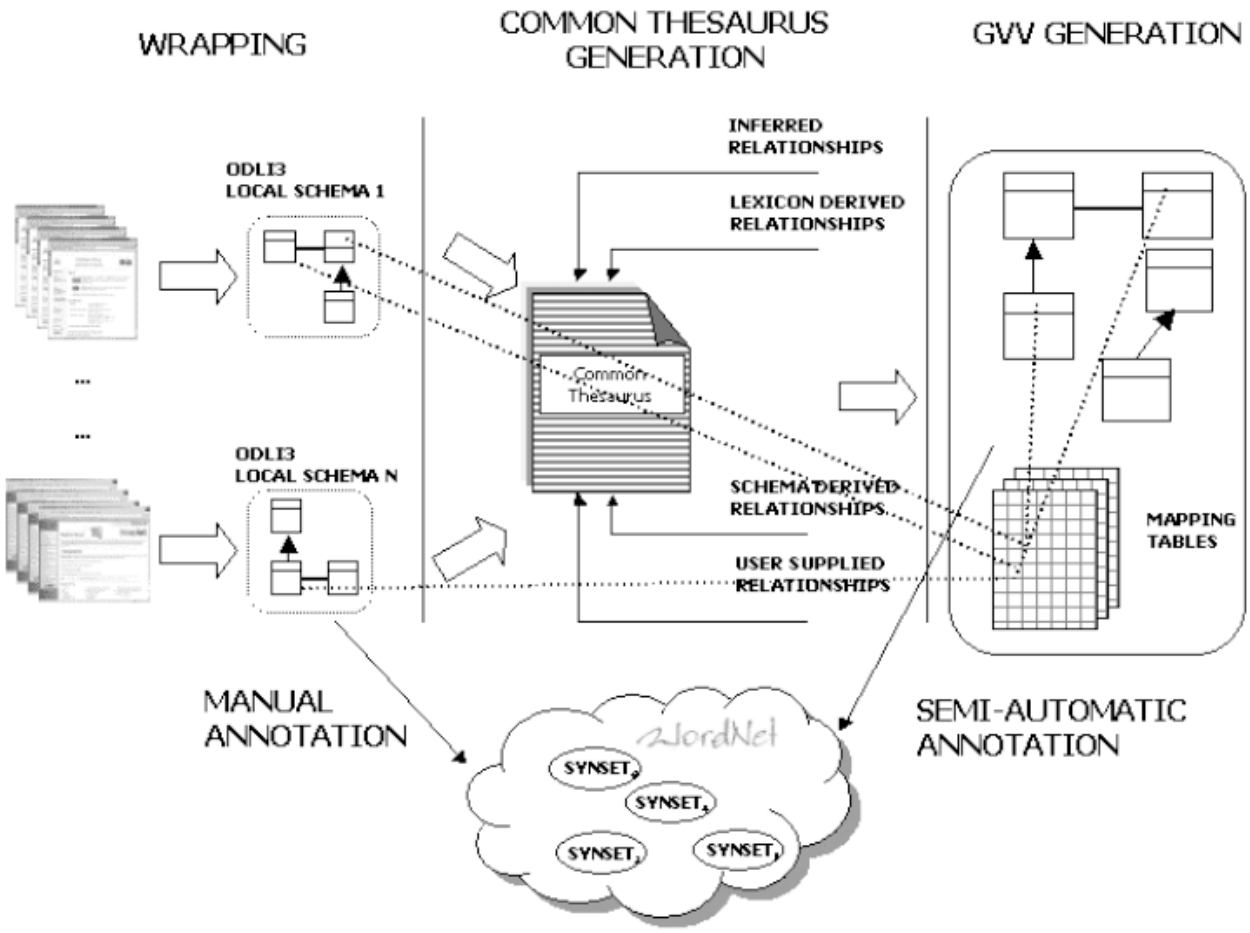

Figura 1.2: Architettura funzionale e metodologia applicata dal sistema MOMIS

1.2 Costruzione di una GVV

1.2.1 Estrazione delle sorgenti locali

Il primo passo nella generazione di una GVV è l'estrazione delle sorgenti locali. A ciascuna struttura di sorgente conosciuta è affidato un wrapper che ha lo scopo di tradurre tale struttura nel modello informativo offerto dal linguaggio ODL_{i3}. I wrapper divengono quindi elementi fondamentali nel gestire l'eterogeneità delle sergenti da integrare.

Per sorgenti di informazione strutturate come database relazionali o object-oriented è sempre disponibile una descrizione dello schema la cui traduzione è immediata. Per sorgenti semistrutturate (documenti XML e pagine web) non è direttamente disponibile una descrizione dello schema, in quanto caratteristica principale di questi tipi di dato è di essere autodescriventi. In altre parole la descrizione dello schema è specificata all'interno dei dati. Per affrontare questa situazione è stato sviluppato un wrapper in grado di estrarre lo schema di documenti XML validi, cioè conformi allo standard DTD (Document Type Definition).

Lo schema dei documenti presentati in linguaggio HTML, che non separa la struttura dal layout, viene estratto da un wrapper esterno di nome Lixto. Esso traduce il contenuto della pagina web in un file XML compatibile DTD, dal quale è possibile estrarre la struttura tramite il wrapper XML

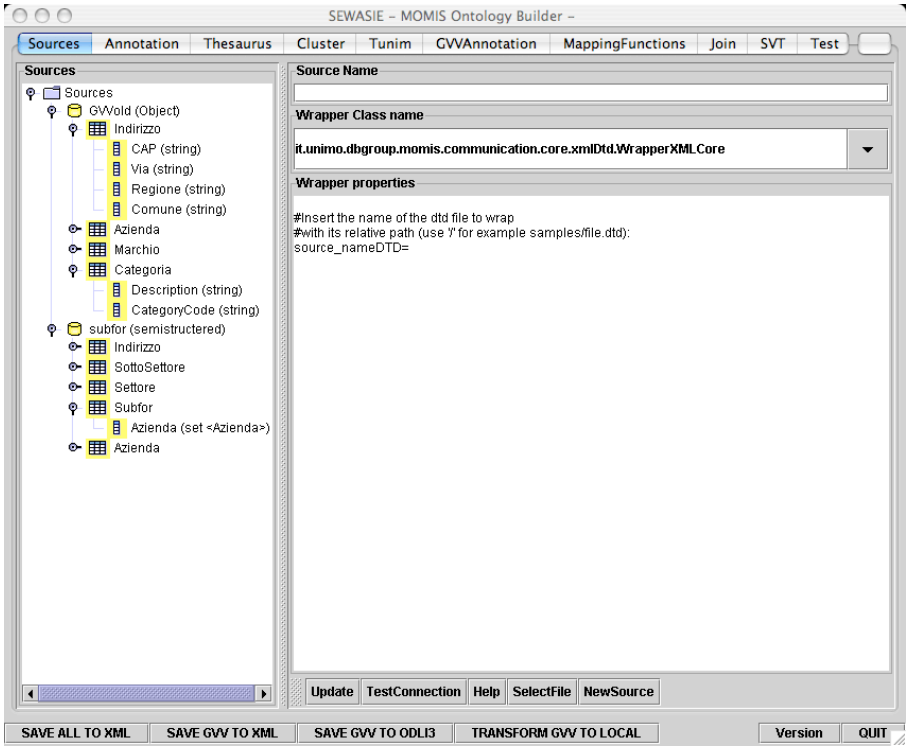

Figura 1.3: Momis Ontology Builder, estrazione delle sorgenti locali

1.2.2 Annotazione delle sorgenti locali con WordNet

Nella fase di annotazione a ciascun elemento (classe o attributo) delle sorgenti importate sono affidate una forma (o significante) ed un significato. Per fare questo il sistema guida il progettista nella scelta dei dati dal database WordNet. Questo database è organizzato a matrice: ciascuna riga rappresenta una forma (il modo in cui una parola è pronunciata, una radice) e ciascuna colonna rappresenta un significato. E' immediato rilevare relazioni di sinonimia tra parole presenti sulla stessa colonna e relazioni di polisemia (più significati affidati a un solo termine) per parole sulla stessa riga. All'interno del sistema il progettista può scegliere una forma per ciascun nome di classe o attributo delle sorgenti locali, successivamente può scegliere uno o più significati da affidare al nome tra quelli proposti da WordNet per la forma selezionata. Il sistema può eseguire l'affidamento di una forma a un nome di classe o attributo in modo automatico, basando la ricerca sul nome stesso. La funzione della fase di annotazione è quella di generare rapporti semantici tra classi e attributi delle sorgenti, da usare come materiale per la generazione del Common Thesaurus, pertanto se non viene scelto alcun significato per una classe o un attributo nessuna relazione potrà derivare dal valore semantico del suo nome.

Si è scelta la via dell'annotazione per sfruttare il più possibile il contenuto semantico intrinseco dei termini usati per descrivere gli schemi importati e per facilitare l'uso dell'ontologia risultante del processo tramite la sua identificazione in significati conosciuti.

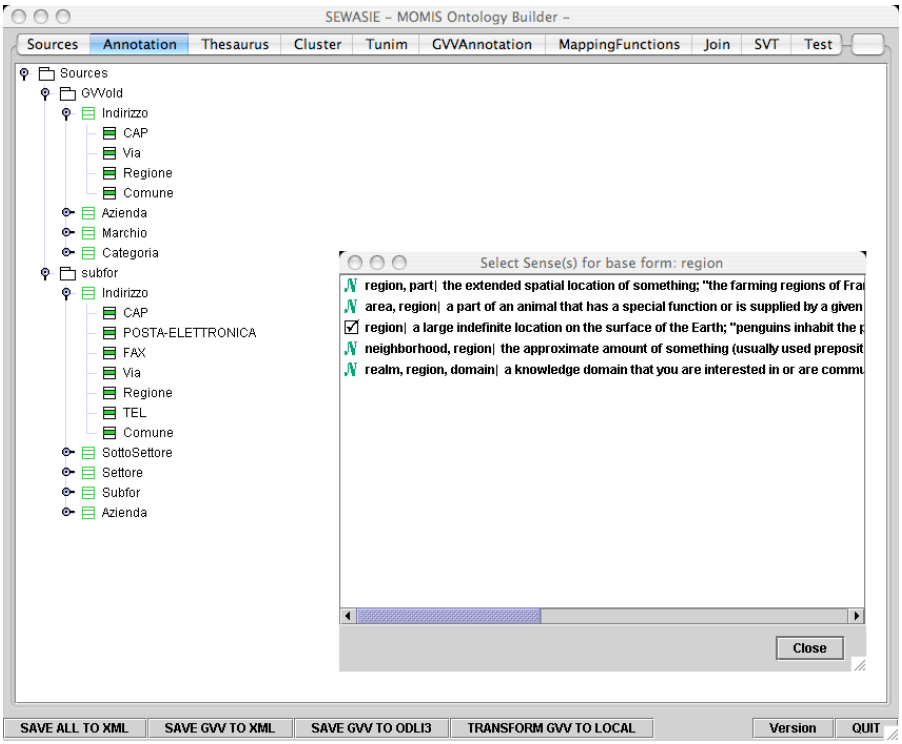

Figura 1.4: Momis Ontology Builder, annotazione delle sorgenti locali con WordNet

1.2.3 Generazione del Common Thesaurus

Una volta consolidate le conoscenze semantiche sugli schemi importati, il sistema procede a creare una libreria di relazioni (il Common Thesaurus) che guideranno la creazione dei cluster per generare la GVV. Il linguaggio ODL_{i3} supporta quattro tipi fondamentali di relazioni intensionali utilizzati dal sistema MOMIS, queste relazioni esprimono la conoscenza inter e intra schema delle sorgenti. Esse sono definite tra classi o attributi e sono specificate considerando i nomi delle classi o degli attributi. Le relazioni possibili sono:

- *SYN (Synonim):* la relazione di sinonimia indica che i due termini da essa coinvolti hanno lo stesso significato;
- *BT (Broader Term):* questa relazione, stabilita tra due termini t1 e t2, indica che il termine t1 ha un significato più generale del termine t2. La relazione BT non è simmetrica;
- *NT (Narrower Term):* si tratta dell'opposto della relazione BT;
- *RT (Related Term):* è una relazione simmetrica che lega due termini usati nello stesso contesto o tra i quali esiste un legame generico.

Le relazioni possibili sono affidate alle classi e agli attributi degli schemi importati secondo quattro procedimenti, ciascuno dei quali aggiunge relazioni di diversa origine:

Relazioni derivate dallo schema

Le relazioni derivare dallo schema sono estratte direttamente dalla struttura delle sorgenti, analizzando separatamente ciascuno schema. Il un documento XML è possibile derivare relazioni BT e NT dai tag ID e IDREF e relazioni RT dagli elementi innestati. Altre relazioni RT possono essere derivare dai vincoli di foreign key di sorgenti relazionali, esse sono arricchite da relazioni BT e NT nel caso le foreign key siano anche primary key.

Relazioni derivate dal lessico

Il sistema mette a frutto l'opera di annotazione effettuata in precedenza per generare relazioni lessicali tra gli elementi delle sorgenti. Esse sono derivate dalla grande quantità di relazioni tra i significati dei termini espresse nel database WordNet secondo le seguenti indicazioni:

- *Sinonimia:* unendo due significati che esprimono lo stesso concetto, la sinonimia genera una relazione SYN tra due elementi;

- *Ipernimia/Iponimia:* l'ipernimia lega un significato con il corrispondente più generale, l'iponimia con il corrispondente più specifico. Esse possono generare relazioni rispettivamente di BT ed NT tra gli elementi;
- *Omonimia/Meronimia:* la meronimia lega un significato con un suo componente, la meronimia fa l'opposto. La relazione derivata è RT;
- *Correlazione:* se due significati condividono un significato più generalizzante sono in correlazione. La relazione che li lega in ODL_{i3} è di tipo RT.

I termini per cui l'annotazione non è stata completata, cioè a cui non è stato affidato alcun significato, non sono legati da relazioni derivate dal lessico.

Relazioni fornite dal progettista

Al progettista è consentito eliminare relazioni generate automaticamente come pure inserire nuove relazioni direttamente nel Common Thesaurus. Questo procedimento consente di massimizzare il potere espressivo del Common Thesaurus ma è molto delicato e suscettibile a errori.

Validazione delle relazioni e inferenza di nuove relazioni

Con l'ausilio del modulo ODB-Tools, l'intero Contenuto del Common Thesaurus e delle sorgenti locali viene tradotto nel linguaggio OLCD (Object Language with Complements allowing Descriptive Cycles). Successivamente il sistema analizza la compatibilità dei domini associati con gli attributi tradotti in OLCD. Inoltre nuove relazioni possono essere derivate della chiusura transitiva di quelle già presenti.

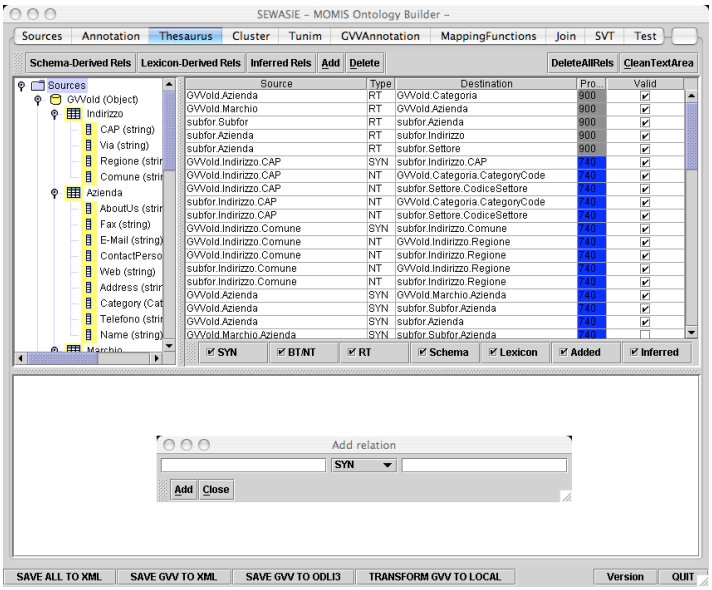

Figura 1.5: Momis Ontology Builder: generazione del Common Thesaurus

1.2.4 Generazione della GVV

La GVV è un insieme di classi globali (GC), ciascuna delle quali è associata tramite una Mapping Table a una o più classi locali (LC). Per determinare queste associazioni, Momis utilizza un algoritmo di clustering gerarchico, basato sulla valutazione del peso di ciascuna relazione presente nel Common Thesaurus. Per peso delle relazioni si intende il coefficiente che quantifica il grado di affinità tra i due termini che la relazione lega. Questi coefficienti sono suggeriti dal sistema e possono essere modificati dal progettista prima di procedere alla generazione dei cluster:

SYN = 1
NT/BT = 0,8
RT = 0,5

In base al peso e al numero delle relazioni, il sistema determina per ciascuna classe un coefficiente di affinità basato sul nome (Name Affinity Coefficient) e uno basato sul numero di relazioni tra gli attributi (Structural Affinity Coefficient), successivamente ne calcola la combinazione lineare (Global Affinity Coefficient):

Naming Affinity Coefficient

Questo coefficiente è calcolato considerando le relazioni tra i nomi delle classi. Il sistema ne valuta il peso, scegliendo il valore più alto, che indica un percorso "semantico" inferiore tra i due nomi, quindi una maggiore affinità. Il percorso viene scelto nell'insieme di quelli possibili, di cui fanno parte il percorso determinato dalla relazione diretta tra i due nomi (se esiste) e tutti quelli ottenuti seguendo relazioni che hanno elementi intermedi nel Common Thesaurus. Ad esempio se due termini sono NT di uno stesso termine sono almeno RT tra di loro, quindi i percorsi da considerare sono almeno 2, uno ha il peso della relazione RT, un altro ha il peso del prodotto delle due relazioni NT, cioè gli "step" semantici che separano i due significati passando per il terzo.

Structural Affinity Coefficient

Il coefficiente di affinità strutturale è calcolato eseguendo la divisione tra il numero degli attributi in relazione tra due classi e il numero degli attributi totali che esse possiedono. Non importa se un attributo di una classe ha più di una relazione con attributi dell'altra classe, è importante che ne abbia almeno una. Ad esempio una classe di 3 attributi e una classe di 5 attributi che hanno una relazione tra un solo attributo sono legate da un valore di 1/(5+3) =

0,125. In termini più precisi il coefficiente di affinità strutturale è determinato seguendo la seguente procedura:

$$SA(c,c^l) = \frac{\left|\left\{a_t \mid a_t \in A(c), a_q \in A(c^l), n_t \sim n_q\right\}\right| + \left|\left\{a_q \mid a_t \in A(c), a_q \in A(c^l), n_t \sim n_q\right\}\right|}{\left|A(c)\right| + \left|A(c^l)\right|} \cdot \frac{\left|\left\{x \in C \mid flag(x) = 1\right\}\right|}{\left|C\right|}$$

Dove c e c^l sono le due classi tra cui stabilire il coefficiente, $A(c)$ e $A(c^l)$ sono i due insiemi di attributi ad esse associati, a_t e a_q sono due attributi appartenenti ai due insiemi e n_t e n_q sono i loro nomi. Il simbolo ~ indica una relazione tra di essi. Inoltre

$$C = \left\{a_t, a_q) \mid a_t \in A(c), a_q \in A(c^l), n_t \sim n_q\right\}$$

E il termine moltiplicativo della prima espressione indica la presenza di relazioni valide tra i due attributi (*flag = 1* indica un risultato positivo).

Global Affinity Coeffcient

Esso non è altro che la somma dei due coefficienti ottenuti, ciascuno moltiplicato per un coefficiente di normalizzazione. I due coefficienti di normalizzazione sono suggeriti dal sistema come uguali tra di loro e pari al valore di soglia con cui sarà confrontato il Global Affinity Coefficient. La loro personalizzazione consente al progettista di variare il peso che ha l'affinità strutturale rispetto all'affinità di naming nella realizzazione dei cluster. Il valore predefinito per questi coefficienti è 0,5. Il Global Affinity Coefficient è pertanto determinato da:

$$GA = (NA \times w_{NA}) + (SA \times w_{SA})$$

Dove *NA* e *SA* sono Naming Affinity Coefficient e Structural Affinity Coefficient e w_{NA} e w_{SA} sono i coefficienti di normalizzazione ad essi associati. Nelle analisi alla base di questo documento si è fatto uso dei coefficienti, pesi e valore di soglia predefiniti.

Il risultato di questa procedura è un albero di affinità, dove le classi sono le foglie a ogni nodo corrisponde un coefficiente di affinità. Il sistema fa un'operazione di sintesi creando le classi globali da questo albero con un meccanismo a soglia, decretando l'unione di due classi tra di loro se il coefficiente di affinità globale supera la soglia definita (suggerita pari a 0,5).

Una volta determinate le mappature delle classi locali in classi globali, il sistema procede con l'affidamento di un set di attributi globali (GA) a ciascuna classe. Questi attributi sono il risultato

di una mappatura degli attributi locali (LA) delle classi coinvolte che opera in base alla relazioni definite nel Common Thesaurus. Il progettista può visualizzare e modificare questa mapping table la cui struttura a colonne elenca nella prima colonna il nome affidato all'attributo locale e nelle seguenti gli attributi locali in esso mappati, una colonna per ciascuna classe locale della mapping table. La funzione di mappatura può essere rappresentata come MT[GA][LC] e può assumere varie conformazioni:

- *Identità:* il valore di GA è uguale a quello di LA. La mappatura è del tipo: MT[GA][LC]=LA;
- *Congiunzione:* il valore GA è ottenuto dalla congiunzione di un insieme di valori LA della classe locale LC. Questo caso è espresso da MT[GA][LC]=LA1 and … and LAn;
- *Costante:* il GA assume come valore corrispondente alla classe locale LC un valore costante scelto dal progettista. Il caso in questione si esprime come MT[GA][L]=k;
- *Indefinito:* il GA non è definito per la classe LC. MT[GA][L]=null.

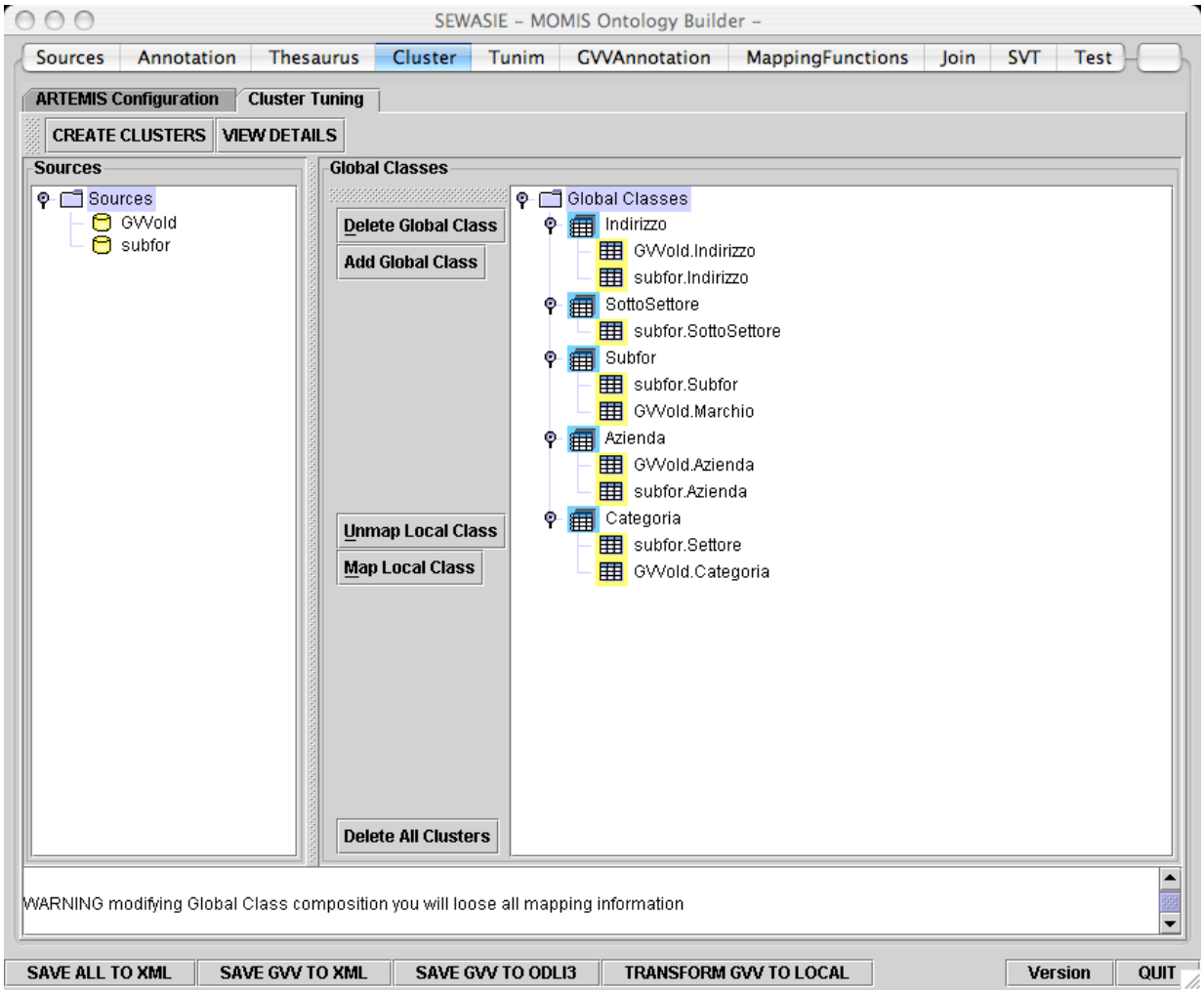

Figura 1.6: Momis Ontology Builder: generazione dei Cluster

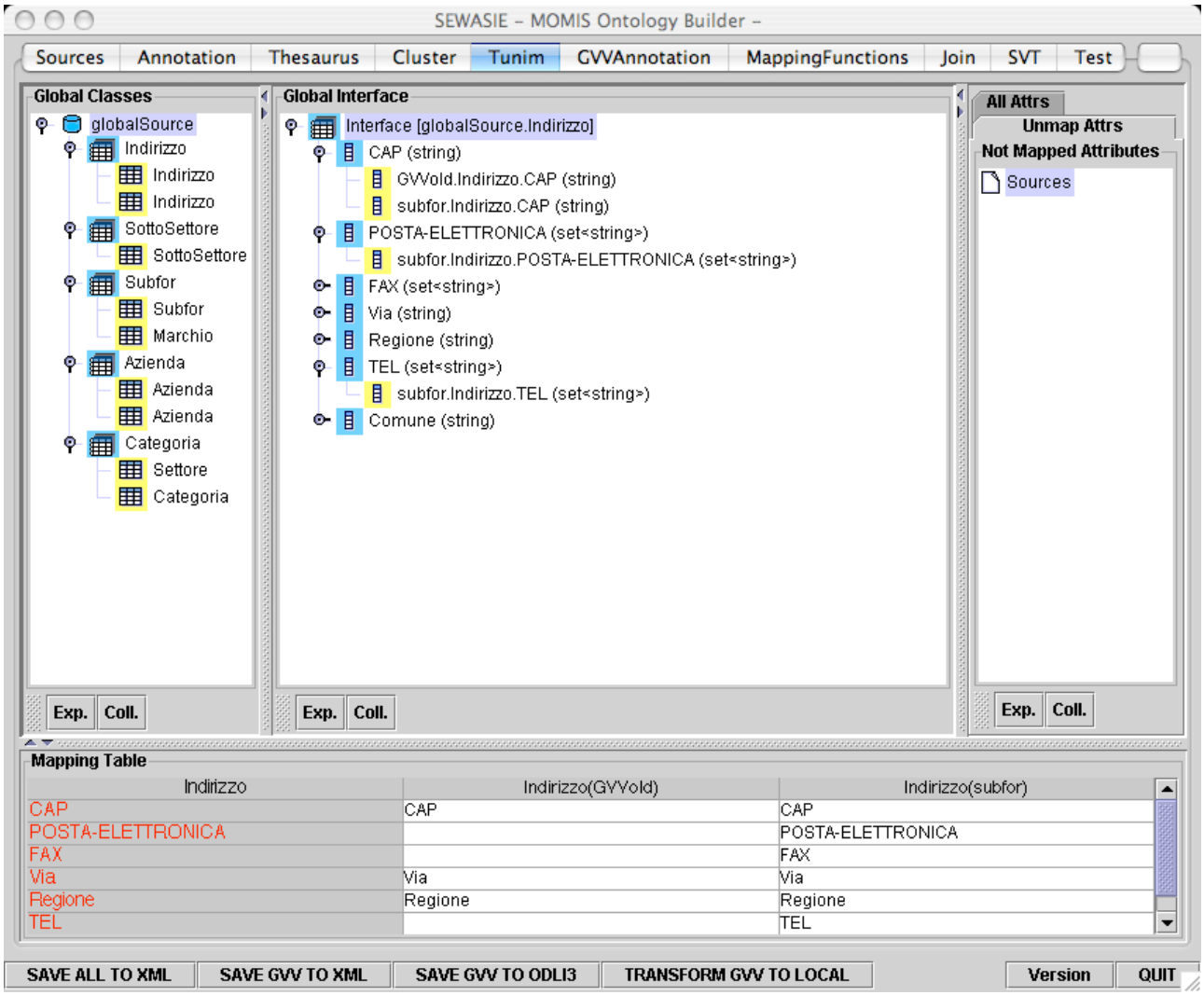

Figura 1.7: Momis Ontology Builder, elaborazione della GVV

1.2.5 Annotazione della GVV

Questa fase consente di affidare un nome e un significato a ogni elemento della GVV. Si tratta di un'operazione particolarmente utile nella prospettiva di utilizzo della GVV come sorgente locale per successive integrazioni, facilitata se ogni elemento ha un significato universalmente riconoscibile. Come per quanto riguarda l'annotazione delle sorgenti locali, anche per la GVV ogni elemento ha un nome e un'annotazione (significante + significato). Il nome è solo un'etichetta che identifica l'elemento all'interno di Momis, mentre l'altro dato è significativo per l'identificazione dell'elemento. Essi sono affidati dal sistema e modificabili dal progettista.

Annotazione di una classe

Per l'annotazione di una classe globale è opportuno considerare il peso delle classi locali che la compongono. Se la classe globale è costituita da una sola classe locale, il nome da affidare alla classe globale sarà lo stesso di quella locale. Lo stesso vale per l'annotazione da affidare al nome. Se una classe globale è composta da più classi locali si considera l'insieme delle classi locali più generali rispetto alle relazioni del Thesaurus (BLC_{GC}). Esso è definito come:

$$BLC_{GC} = \{LC \in GC \mid \neg \exists y \in GC, (LC\ NT\ y) \vee (y\ BT\ LC)\}$$

Dove *GC* è una classe globale e *LC* è una classe locale in essa mappata.

Il sistema propone al progettista una scelta tra i nomi di classe che sono presenti nell'insieme BLC_{GC}. Per quel che riguarda l'annotazione, il sistema affida alla classe globale tutti i significati delle classi locali appartenenti all'insieme.

Annotazione di un attributo

Nell'annotazione degli attributi il sistema segue una procedura simile. Anche gli attributi globali possono essere composti da uno o più attributi locali, quindi nel primo caso l'annotazione è immediatamente ereditata da quella dell'attributo locale corrispondente. Nel secondo caso si considera l'insieme degli attributi locali più generali BLA_{GA} (esso è un sottoinsieme dell'insieme degli attributi locali mappati in un attributo globale). In particolare LA_{GA} è l'insieme degli attributi locali mappati in un attributo globale, definito nel seguente modo:

$$LA_{GA} = \{LA \mid \exists LC \in GC, LA \in LC \vee MT[GA][LA] = null\}$$

Mentre BLA_{GA} è l'insieme degli attributi locali più generali mappati in GA, definito come seugue:

$$BLA_{GA} = \{LA \in LA_{GA} \mid \neg \exists y \in LA_{GA}, (LA\ NT\ y) \vee (y\ BT\ LA)\}$$

Dove *LA* è un attributo locale.
Nome e significato per l'annotazione sono scelti tra quelli degli attributi appartenenti a BLA_{GA}.

2 Aggiunta di una nuova sorgente ad una GVV: teoria

2.1 Procedura e approcci di integrazione

La GVV creata dal sistema è una rappresentazione statica degli schemi delle sorgenti importate. Essa è stata derivata dagli schemi singoli una volta e rappresenta la loro struttura solo finché essa è mantenuta invariata. Il problema principale che riguarda l'usabilità della GVV realizzata verte sulla possibilità di aggiornarla (e quindi mantenerla aggiornata) per reagire a un cambiamento delle sorgenti locali; cambiamento che può essere modificazione, cancellazione o inserimento di una sorgente. La GVV deve essere modificata per rispondere a qualunque variazione. In questo senso è opportuna una distinzione tra aggiunta di una sorgente e eliminazione/modifica di una sorgente. Se infatti della prima il mediatore può avere conoscenza diretta, in quanto operazione instaurata tra il sistema e la sorgente, per la seconda è opportuno definire dei trigger all'interno del wrapper che presidia la sorgente, il quale deve notificare ogni operazione di modifica della struttura o eliminazione di una sorgente locale. In questa sede sarà presa in considerazione l'operazione di inserimento di una nuova sorgente, con particolare attenzione al modo in cui il sistema deve reagire.

La procedura di aggiunta è particolarmente onerosa per il progettista e per il sistema e si svolge seguendo un processo di integrazione del tutto corrispondente a quello visto in precedenza. Tale processo può essere affrontato seguendo due approcci differenti. E' possibile realizzare un processo di integrazione che coinvolge le nuove sorgenti e le vecchie sorgenti locali indipendentemente dai cluster creati oppure è possibile utilizzare un processo in cui sono integrate le nuove sorgenti e la GVV precedentemente ottenuta, considerata come una sorgente unificata. Nel primo approccio è possibile fare uso dei dati ottenuti dall'annotazione delle vecchie sorgenti svolta in precedenza e nel secondo si trae vantaggio dall'annotazione operata sulla GVV. In altre parole la procedura differisce nei due approcci soltanto ad annotazione avvenuta e si svolge nel modo espresso in figura.

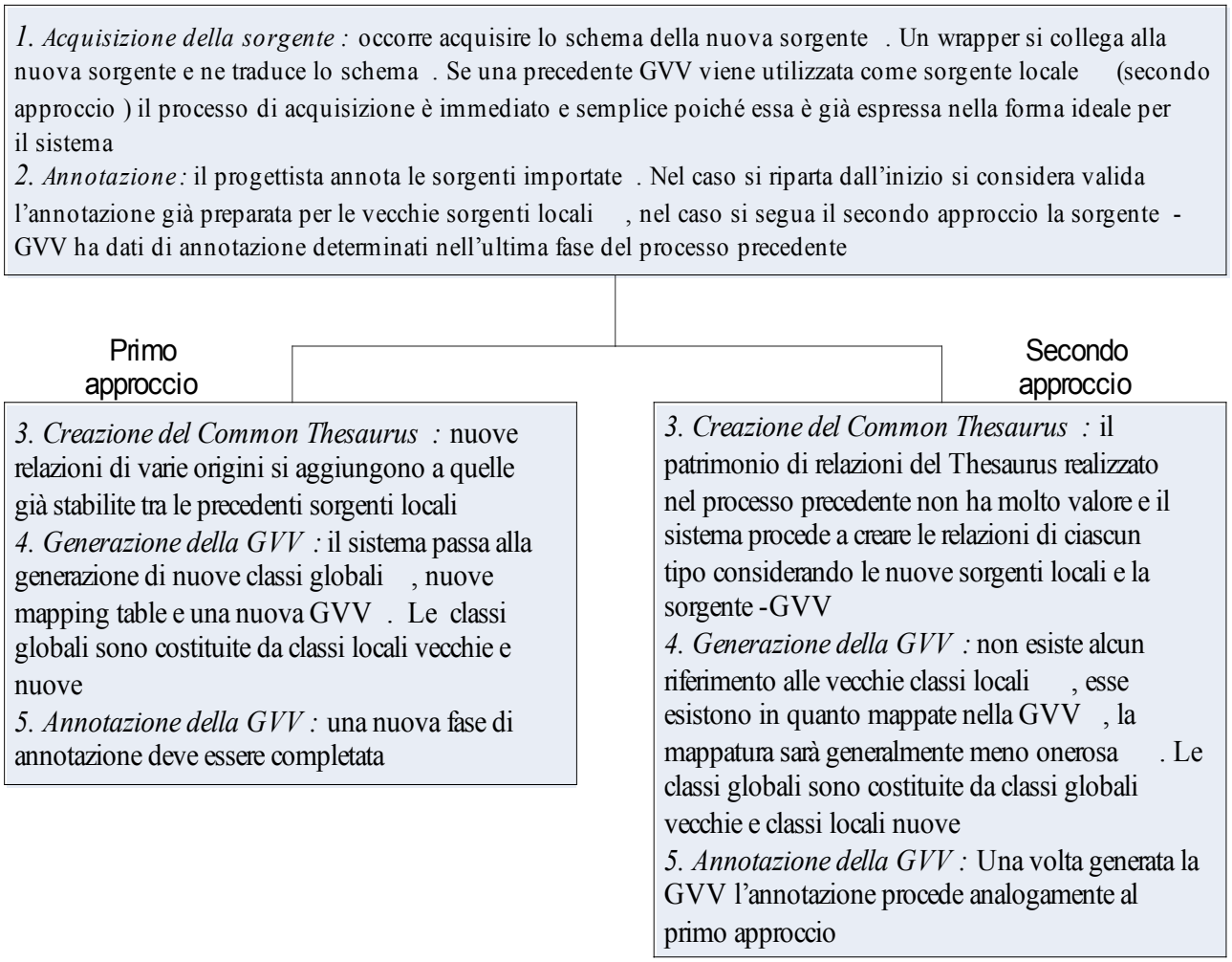

Figura 2.1: Due approcci di integrazione

Il primo approccio opera dunque in modo indistinto dal metodo generale di integrazione precedentemente descritto. Il secondo approccio merita una descrizione più dettagliata. I due metodi saranno oggetto di confronto nel corso degli esperimenti.

2.2 Ricalcolo della GVV sfruttando la GVV precedente

Questo procedimento, basato sul secondo approccio, prevede la creazione di una nuova GVV costituita da una nuova sorgente locale e una vecchia GVV, trattata anch'essa come sorgente locale. Il funzionamento corretto di applicativi basati sulla struttura della GVV è facilitato dalla somiglianza della nuova vista con quella ottenuta prima dell'integrazione. Si introduce la seguente notazione utilizzata nel resto della sezione:

gcNew rappresenta una classe globale nel nuovo schema integrato che ha nome ***gcNewName*** e un insieme di attributi globali $\boldsymbol{gcNewAtt_i}$

gcOld rappresenta una classe globale nel vecchio schema integrato che ha nome ***gcOldName*** e un insieme di attributi globali $\boldsymbol{gcOldAtt_j}$

lcNew rappresenta una classe locale della nuova sorgente che ha nome ***lcNewName*** e un insieme di attributi locali $\boldsymbol{lcNewAtt_k}$

Seguendo al procedura proposta, il sistema calcola le relazioni del Common Thesaurus. Queste saranno composte da relazioni derivare dallo schema e relazioni lessicali intraschema derivate dall'annotazione della nuova sorgente. L'apporto della GVV precedente comprende relazioni intra e inter-schema che ne costituivano il Common Thesaurus. Come risultato della fase di clustering, una nuova classe globale ***gcNew*** sarà generalmente definita sulle vecchie classi globali della GVV e sulle nuove classi locali. Il procedimento richiede in questa fase un passo aggiuntivo, in cui le vecchie classi globali vengono sostituite dalle classi locali che le componevano nella GVV, mantenendo le regole di mapping stabilite nel precedente processo di integrazione. In questo modo una nuova classe globale

$\boldsymbol{gcNew=\{gcOld_1,\ldots,gcOld_p,lcNew_1,\ldots,lcNew_n\}}$

è una classe del tipo

$\boldsymbol{gcNew = \{lcOld_{11},\ldots,lcOld_{1z},\ldots,lcOld_{p1},\ldots, lcOld_{pm},lcNew_1,\ldots,lcNew_n\}}$

Considerata questa composizione della nuova classe globale si possono distinguere tre casi:

1. Una nuova classe globale si compone di una vecchia classe globale e di una o più classi locali;
2. Una nuova classe globale si compone unicamente di nuove classi locali;
3. Una classe globale del nuovo schema si compone di più classi globali della vecchia GVV e di una o più nuove classi locali.

E' frequente il verificarsi di varie situazioni differenti nello stesso processo di integrazione. Ciascuna situazione sarà ora descritta più nel dettaglio.

Caso 1: $\boldsymbol{gcNew = \{gcOld,lcNew_1,\ldots,lcNew_i,\ldots,lcNew_n\}}$

In questo caso l'integrazione delle nuove classi non causa cambiamenti alla GVV che possono determinare problemi di dialogo con eventuali applicativi che operano su di essa. In questa situazione si ottiene una nuova classe globale con nome ***gcNewName*** e un insieme di attributi in parte comuni alle vecchie classi globali e in parte generati dall'apporto delle nuove classi locali. La tabella che segue illustra la situazione più generale della mappatura degli attributi in questo caso.

Il gruppo di vecchi attributi globali (da $gcOldAtt_1$ a $gcOldAtt_m$) è mantenuto anche nella nuova classe globale e nuove mappature sono definite in risposta alle corrispondenze di questo insieme di attributi con quelli appartenenti alle classi locali ***lcNew***. Il gruppo di nuovi attributi globali (da $gcNewAtt_1$ a $gcNewAtt_p$) non ha mappature in corrispondenza dei vecchi attributi locali, in quanto la sua introduzione è dovuta unicamente all'apporto degli attributi delle nuove sorgenti locali.

<table>
<tr><th>gcOldAtt</th><th>lcOld₁</th><th>...</th><th>lcOld_k</th><th>lcNew₁</th><th>lcNew_t</th><th>lcNew_n</th></tr>
<tr><td>gcOldAtt₁</td><td colspan="3" rowspan="3">Mappature di gcOld</td><td colspan="3" rowspan="6">Nuove mappature</td></tr>
<tr><td>...</td></tr>
<tr><td>gcOldAtt_m</td></tr>
<tr><td>gcNewAtt₁</td><td colspan="3" rowspan="3">Nessuna mappatura</td></tr>
<tr><td>...</td></tr>
<tr><td>gcNewAtt_p</td></tr>
</table>

Tabella 2.1

Nel caso in esame, il significato da affidare alla nuova classe globale sarà ottenuto da quelli affidati alla vecchia classe globale in essa mappata, arricchito di quelli affidati alle nuove classi locali aggiunte alla mappatura.

Caso 2: gcNew = {lcNew$_1$,...,lcNew$_i$,...,lcNew$_n$}

Si tratta di un caso semplice in cui non vi è corrispondenza alcuna tra le vecchie classi globali derivate dalla GVV e le nuove classi locali da integrare. Semplicemente viene creata una nuova classe globale ***gcNew*** con nome ***gcNewName*** la quale mappa una o più nuove lassi locali. La procedura di generazione dei cluster per questo caso non è dissimile a quanto visto per il primo processo di integrazione, lo stesso vale per l'affidamento dei significati alla classe globale ***gcNew***. Applicazioni che basavano il loro funzionamento sulla struttura della GVV non sono influenzate dall'aggiunta di questa nuova classe tuttavia non ne sfruttano l'apporto.

Caso 3: gcNew = {gcOld$_1$,...,gcOld$_p$,lcNew$_1$,...,lcNew$_i$,...,lcNew$_n$}

Si tratta della situazione più critica in quanto la struttura della GVV è modificata pesantemente attraverso il processo di integrazione. La distinzione tra le classi viene modificata poiché più vecchie classi globali vengono agglomerate in una nuova classe globale. Questo si può verificare se una delle nuove classi locali che si desidera integrare ha forti legami con le vecchie classi globali. Se un'applicazione basa il suo funzionamento sulla struttura della GVV, questo scenario potrebbe costituire problemi.

L'analisi delle procedure messe in atto dal sistema hanno fatto emergere questi tre casi, è tuttavia possibile che si verifichino situazioni differenti. Esse sono in parte dovute a un numero insufficiente di relazioni nel Common Thesaurus a creare opportunamente i cluster per descrivere la nuova GVV, e richiedono maggiori interazioni da parte del progettista.

3 Aggiunta di una nuova sorgente ad una GVV: esperimento

Gli esperimenti sono svolti tramite l'utilizzo del componente del sistema MOMIS per la costruzione della GVV: Ontology Builder. Ogni riferimento a tipi di dato e caratteristiche delle sorgenti è in relazione all'output di questo tool. I casi in esame prevedono l'integrazione di tre sorgenti in una GVV, partendo dall'ipotesi che due delle sorgenti considerate siano già state precedentemente integrate e la terza sia aggiunta in seguito alla GVV. Il procedimento di integrazione è eseguito in entrambi gli approcci descritti nel capitolo 1. Una prima fase vede l'integrazione delle tre sorgenti separate (primo approccio), una seconda fase prevede l'integrazione della terza sorgente e della GVV che esprime la struttura delle altre due (secondo approccio). In riferimento a quanto riportato in merito ai tre differenti scenari di integrazione, si ricordi che nel ricreare i primi due scenari trattati è rilevante soffermarsi sul modo in cui devono essere affrontati nei due approcci descritti. Il terzo scenario, per la sua criticità di analisi e occorrenza, richiede una trattazione più accurata che non riguarda l'approccio di integrazione, ma le casistiche in cui la situazione si può proporre. La trattazione sarà svolta nel quarto capitolo. Nel resto del documento si identificherà con il nome GVVold la GVV ottenuta con l'integrazione delle prime due sorgenti da utilizzare come sorgente locale nel secondo approccio, mentre i termini GVV_total e GVV_merge identificheranno il risultato finale rispettivamente seguendo il primo e il secondo approccio.

3.1 Le sorgenti dell'esempio

Di seguito sono presentati i contenuti delle sorgenti scelte per il primo esperimento di integrazione. Esse sono complete del significante (word form) scelto per ciascun elemento dello schema. Le sorgenti, di origine semistrutturata, sono state separatamente importate nel sistema e l'annotazione è stata svolta nel modo più completo possibile prima di procedere alla valutazione dei processi di integrazione poiché essa non è un elemento di particolare rilievo nel confronto degli approcci. Per leggere correttamente le seguenti tabelle si consideri la sintassi puntata *sorgente.classe.attributo*. Nelle tabelle il contenuto della colonna di sinistra è il nome dell'attributo completo del tipo di dato ODL_{i3} presente nel sistema, il contenuto della colonna di destra esprime la word form affidata al termine e il numero tra parentesi è abbinato al significato

del termine nel database WordNet. Infatti nell'accedere al database di WordNet ciascun significato affidato ad un termine è numerato, il layer di interazione con Wordent di Momis Ontology Builder utilizza la stessa numerazione, presentando i significati identificati dal loro numero, consentendo di visualizzarli e sceglierli. Ad esempio l'attributo *prontocomune.Indirizzo.CAP* è di tipo *string* nel sistema e ha una word form *zip_code* a cui è affidato il primo significato tra quelli proposti da WordNet. Tramite Momis Ontology Builder è possibile risalire al significato: *a code of letters and digits added to a postal address to aid the sorting of mails*. Il nome delle classi è affiancato dalla word form ad esso associata e dal numero corrispondete al significato. Ad esempio *prontocomune.Indirizzo* ha word form *address* ed è affidato al secondo dei significati disponibili per *address*. E' possibile che più di un significato sia affidato alla stessa word form ma questo caso non comparirà nell'esperimento.

Sorgente 1: prontocomune

Nome e tipo attributo	Forma e senso attributo
Indirizzo (address, 2)	
CAP (string)	zip_code (1)
Via (string)	street (1)
Regione (string)	region (3)
Comune (String)	town (1)
Prontocomune (trade_name, 1)	
Azienda (set <Azienda>)	business_organization (1)
Azienda (business_organization, 1)	
Indirizzo (Indirizzo)	address (2)
Fax (string)	fax (1)
Web (string)	web (5)
E-mail (string)	e-mail (1)
Telefono (string)	telephone (1)
Categoria (Categoria)	business_sector (1)
Nome (string)	name (1)
Categoria (business_sector, 1)	
CodiceCategoria (string)	code (2)
Descrizione (string)	description (1)

Tabella 3.1

Sorgente 2: fibre2fashion

Nome e tipo attributo	Forma e senso attributo

Category (sector, 2)	
Description (string)	description (1)
CategoryCode (string)	code (2)

Company (business_organization, 1)	
AboutUS (string)	description (1)
Fax (string)	fax (1)
E-Mail (string)	e-mail (1)
ContactPerson (string)	contact (5)
Web (string)	web (5)
Address (string)	address (2)
Tel (string)	telephone (1)
Category (Category)	sector (2)
Name (string)	name (1)

fibre2fashion (trade_name, 1)	
Company (set <Company>)	business_organization (1)

Tabella 3.2

Queste due sorgenti saranno coinvolte nel processo di integrazione iniziale per poi essere unite a una terza sorgente. E' previsto che dalla loro integrazione sia generata una GVV (*GVVold*) caratterizzata da una classe globale "indirizzo" che mappa *prontocomune.indirizzo*, una classe globale "azienda" che mappa *prontocomune.Azienda* e *fibre2fashion.Company*, una classe globale "marchio" (nome affidato dal progettista) che mappa *prontocomune.Prontocomune* e *fibre2fashion.fibre2fashion* ed infine una classe globale "categoria" che mappa *prontocomune.Categoria* e *fibre2fashion.Category*.

La terza sorgente scelta è *subfor*, si tratta di una sorgente semistrutturata descritta di seguito allo stesso modo delle precedenti.

Sorgente 3: subfor

Nome e tipo attributo	Forma e senso attributo

Indirizzo (address, 2)	
CAP (string)	zip_code (1)
POSTA-ELETTRONICA (set <string>)	e-mail (1)
FAX (set <string>)	fax (1)
Via (string)	street (1)
Regione (string)	region (3)
TEL (set <string>)	telehpone (1)
Comune (string)	town (1)

SottoSettore (business_sector, 1)	
Garante (string)	guarantor (1)
DataIstituzione (string)	date (2)
TitoloSottoSettore (string)	title (3)

Settore (sector, 1)	
SottoSettore (SottoSettore)	business_sector (1)
DescrizioneSettore (string)	description (1)
CodiceSettore (string)	code (2)

Subfor (trade_name, 1)	
Company (set <Company>)	business_organization (1)

Azienda (business_organization, 1)	
TECNOLOGIE (set <string>)	technology (1)
ADDETTI (string)	employee (1)
Indirizzo (Indirizzo)	address (2)
CONTROLLI-QUALITA (set < string>)	quality_control (1)
SITO-INTERNET (string)	web (5)
CONTATTI (set <string>)	contact (5)
FATTURATO (string)	turnover (3)
Settore (set <Settore>)	sector (2)
DescrizioneAttivita (string)	description (1)
ORDINI (string)	order (7)
CAPITALE-SOCIALE (string)	capital (2)
Nome (string)	name (1)

Tabella 3.3

subfor, costruita in questo modo, ricrea agevolmente i primi due scenari di integrazione. Infatti viene aggiunta *subfor.SottoSettore*, una nuova classe locale che non ha collegamenti forti

con alcuna classe locale precedente; di conseguenza sarà mappata in una classe globale della GVV finale a sé stante. Questo evento ricrea il secondo scenario di integrazione, in cui una nuova classe globale è composta solo da nuove classi locali. Il primo scenario, molto frequente e semplice da ricreare, ha un suo esempio evidente negli innumerevoli attributi che la classe *subfor.Azienda* ha in più rispetto alle classi "azienda" delle due sorgenti presentate in precedenza e rispetto alla classe "azienda" di *GVVold*. Considerati gli attributi comuni e la sinonimia tra i nomi, *subfor.Azienda* sarà mappata all'interno della classe globale "azienda" della GVV finale (in entrambi gli approcci) ma aggiungerà nuovi attributi rispetto al gruppo fornito da *prontocomune* e *fibre2fashion* (quindi dalla classe "azienda" di *GVVold* nel secondo approccio), ricreando in questo modo il primo scenario. Infatti la classe globale "azienda" della GVV finale conterrà attributi globali che mappano vecchi attributi locali e nuovi attributi locali, ad esempio *prontocomune.Azienda.Nome*, *fibre2fashion.Company.Name* e *subfor.Azienda.Nome*; essa conterrà inoltre attributi globali che mappano i soli attributi aggiunti da *subfor.Azienda*, come ad esempio *subfor.Azienda.CAPITALE-SOCIALE*.

3.2 Svolgimento dell'esperimento

3.2.1 Generazione della GVV tra le prime due sorgenti

Entrambi i procedimenti prevedono che un processo di integrazione anteriore all'aggiunta della terza sorgente abbia unito *prontocomune* e *fibre2fashion*. La creazione della GVV è descritta in questa fase. Le due sorgenti sono state importate nel sistema utilizzando un wrapper adeguato alla loro struttura (sorgenti semistrutturate) ad opera della quale ci si attendono alcune relazioni derivate dallo schema a completare l'insieme di relazioni derivate dal lessico. L'annotazione è stata trattata nella descrizione delle sorgenti, quindi si procede direttamente alla creazione del Common Thesaurus.

Generazione del Common Thesaurus

La seguente tabella descrive le relazioni valide (quindi considerate nel clustering) emerse dall'analisi della struttura e dei significati degli elementi.

Elemento 1	Relazione	Elemento 2
prontocomune.Prontocomune	RT	prontocomune.Azienda
prontocomune.Azienda	RT	prontocomune.Categoria
prontocomune.Azienda	RT	prontocomune.Indirizzo
fibre2fashion.Company	RT	fibre2fashion.Category
fibre2fashion.fibre2fashion	RT	fibre2fashion.Company
fibre2fashion.Ctagory	SYN	prontocomune.Azienda.Categoria
fibre2fashion.Category	SYN	fibre2fashion.Company.Category
fibre2fashion.Category.Description	SYN	fibre2fashion.Company.AboutUs
fibre2fashion.Company	SYN	prontocomune.Prontocomune.Azienda
fibre2fashion.Company	SYN	fibre2fashion.fibre2fashion.Company
fibre2fashion.fibre2fashion	NT	prontocomune.Azienda.Nome
fibre2fashion.fibre2fashion	NT	fibre2fashion.Company.Name
prontocomune.Azienda	SYN	prontocomune.Prontocomune.Azienda
prontocomune.Azienda	SYN	fibre2fashion.Company
prontocomune.Azienda	SYN	fibre2fashion.fibre2fashion.Company
prontocomune.Azienda.E-mail	SYN	fibre2fashion.Company.E-Mail
prontocomune.Azienda.Fax	SYN	fibre2fashion.Company.Fax
prontocomune.Azienda.Nome	SYN	fibre2fashion.Company.Name
prontocomune.Azienda.Telefono	SYN	fibre2fashion.Company.Tel
prontocomune.Azienda.Web	SYN	fibre2fashion.Company.Web
prontocomune.Categoria	SYN	fibre2fashion.Category
prontocomune.Categoria	SYN	fibre2fashion.Company.Category
prontocomune.Categoria.CodiceCat..	SYN	fibre2fashion.Category.CategoryCode
prontocomune.Categoria.Descrizione	SYN	fibre2fashion.Category.Description
prontocomune.Categoria.Descrizione	SYN	fibre2fashion.Company.AboutUs
prontocomune.Indirizzo	SYN	prontocomune.Azienda.Indirizzo
prontocomune.Indirizzo	SYN	fibre2fashion.Company.Address
prontocomune.Indirizzo.CAP	NT	prontocomune.Categoria.CodiceCat..
prontocomune.Indirizzo.CAP	NT	fibre2fashion.Category.CategoryCode
prontocomune.Indirizzo.Comune	NT	prontocomune.Indirizzo.Regione
prontocomune.Prontocomune	SYN	fibre2fashion.fibre2fashion
prontocomune.Prontocomune	NT	prontocomune.Azienda.Nome
prontocomune.Prontocomune	NT	fibre2fashion.Company.Name

Tabella 3.4

Le relazioni presenti nella prima parte della tabella, con lo sfondo più scuro, sono relazioni derivate dalla struttura delle sorgenti. Le altre relazioni sono derivare dal lessico. Si nota la presenza di relazioni tra il nome di una classe e il nome di attributi di altre classi della stessa sorgente. Ad esempio: *fibre2fashion.Company SYN fibre2fashion.fibre2fashion.Company*. La

presenza di relazioni come questa, dovuta alla natura gerarchica della sorgente semistrutturata, induce ad una osservazione delle relazioni tra classe e attributo. In base al procedimento di clustering espresso in precedenza relazioni di questo tipo vengono utilizzate soltanto nel calcolo del Naming Affinity Coefficient in un modo non evidente nell'esperimento e trattato più in dettaglio nelle conclusioni. Lo Structural Affinity Coefficint riguarda le relazioni tra attributi e non trae vantaggio dal tipo di relazioni evidenziato.

Generazione dei Cluster

Il sistema può procedere alla creazione dei cluster per la GVV. Le sorgenti presentate denotano forti affinità tra le classi, tale affinità ha espressione nei cluster presentati di seguito. Nelle seguenti tabelle i cluster sono rappresentati dalla classe globale nella colonna all'estrema sinistra e le classi locali in essa mappate nelle altre colonne. La prima riga contiene il nome delle classi, quella globale e quelle mappate, le altre righe presentano gli attributi della clase globale e i corrispondenti attributi delle classi locali che li hanno definiti. Queste tabelle sono dette Mapping Tables. Sotto ciascuna tabella sono indicati i valori dei coefficienti di affinità di naming (NA) e strutturale (SA) tra le classi mappate in ciascun cluster. I coefficienti sono presenti nella forma già computata: per facilità di confronto con la soglia sono stati moltiplicati per i rispettivi valori di normalizzazione. La trattazione che segue considera i coefficienti normalizzati come SA e NA.

Marchio (rinominata)	fibre2fashion(fibre2fashion)	Prontocomune(prontocomune)
Company	Company	Azienda

NA = 0,5
SA = 0

Category	Categoria(prontocomune)	Category(fibre2fashion)
CategoryCode	CodiceCategoria	CategoryCode
Descrizione	Descrizione	Description

NA = 0,5
SA = 0,5

Indirizzo	Indirizzo(prontocomune)
Via	Via
CAP	CAP
Regione	Regione
Comune	Comune

Azienda	Company(fibre2fashion)	Azienda(prontocomune)
ContactPerson	ContactPerson	
Indirizzo	Address	Indirizzo

AboutUs	AboutUs	
Fax	Fax	Fax
E-Mail	E-Mail	E-mail
Web	Web	Web
Tel	Tel	Telefono
Category	Category	Categoria
Nome	Name	Nome

NA = 0,5
SA = 0,155

Tabella 3.5

Il primo punto di analisi del risultato riguarda il coefficiente di affinità strutturale che ha legato le classi *fibre2fashion.Company* e *prontocomune.Azienda*. Esso è coerente con le relazioni presentate come parte del Common Thesaurus ma pare errato a uno studio più attento delle relazioni che dovrebbero esistere tra gli attributi. Il valore di 0,31 è stato ottenuto considerando cinque attributi in relazione: *Fax*, *E-Mail*, *Web*, *Tel*, *Nome*. Il calcolo eseguito è quindi

$$(5/16) \times 0,5 = 0,155$$

Riguardando però le Mapping Tables generate e le tabelle proposte per presentare l'annotazione degli elementi si evidenzia che gli attributi *Indirizzo* e *Categoria*, comuni a entrambe le classi agglomerate, sono stati trascurati nel computo del SA che in tal caso sarebbe stato

$$(7/16) \times 0,5 = 0,21$$

L'evento si è verificato perché le relazioni trascurate esistono ugualmente nel Common Thesaurus, ma non sono state considerate valide. La validazione delle relazioni si basa infatti sulla loro coerenza nel dominio degli elementi coinvolti. Si può notare che l'attributo *fibre2fashion.Company.Category* non è di tipo *string* come *prontocomune.Azienda.Categoria* ma istanzia un elemento della classe *fibre2fashion.category*, quindi è un dato complesso. Questa incongruenza rende le relazione che lega i due attributi citati non valida. Essa potrebbe essere valida se entrambi gli attributi fossero di tipo *string* quindi appartenenti a domini compatibili. Analogamente *fibre2fashion.Company.Address* è di tipo *string*, mentre *prontocomune.Azienda.Indirizzo* è di tipo *Indirizzo*, istanziando l'omonima classe della stessa sorgente. Questo evento influisce sull'accuratezza nello stabilire il coefficiente di affinità strutturale, in quanto le relazioni rimosse lo avrebbero aumentato. Nel caso specifico non sarebbe cambiato nulla considerando le relazioni scartate e aumentando il coefficiente di affinità

strutturale. Infatti il NA è senza alcuna aggiunta pari alla soglia e determina comunque l'unione delle due classi. Se tale coefficiente fosse stato più basso avrebbe potuto essere pari a 0,4 (NT o BT tra i nomi delle classi), a 0,34 (NT x NT) o a 0,25 (RT). Si Nota quindi come una relazione NT o BT avrebbe unito ugualmente le classi e una RT non l'avrebbe fatto, indipendentemente dalla scelta fatta per il SA (0,155 o 0,21). Tuttavia possono presentarsi situazioni analoghe in cui una differente scelta nelle relazioni da utilizzare comporta un differente risultato, come quello in cui il NA è determinato dal prodotto di due relazioni NT: un valore di 0,32 non sarebbe stato sufficiente a unire le due classi senza considerare le relazioni scartate (0,32 + 0,22 = 0,54 > 0,5 mentre 0,32 + 0,155 = 0,475 < 0,5).

Infine, sempre in riguardo al caso presentato, si può osservare che un SA più basso tra le due classi non avrebbe influito sulla loro integrazione. Infatti il NA esistente tra i nomi delle due classi (determinato dalla relazione SYN) è sufficiente per superare la soglia e determinare l'integrazione. Si ricordi che una relazione lessicale tra i nomi, per quanto debole, deve comunque esistere in ogni caso in cui non vi sia SA = 0,5, cioè una corrispondenza strutturale 1 a 1 tra le due classi. Il caso in questione è quello dei legami tra *fibre2fashion.Category* e *prontocomune.Categoria*. Come si può notare l'elevato SA (determinato dalla identità strutturale delle classi) non necessita dell'apporto del NA per superare la soglia. Si può affermare che le due classi "sono strutturalmente identiche", costituendo un caso di integrazione eccezionalmente fortunato.

Per quanto riguarda le due classi globali rimaste, *Indirizzo* è stata generata dal solo apporto di *prontocomune.Indirizzo* che non ha relazioni con altre classi sufficienti a determinarne l'integrazione.

La classe *Marchio* è poco rilevante ai fini dell'esperimento ma è stata mantenuta perché particolarmente significativa nell'organizzazione delle sorgenti separate. Essa è l'espressione dell'integrazione della root delle due sorgenti XML. In ogni sorgente XML (semistrutturataè presente una radice per l'albero di classi, da cui la struttura si sviluppa in modo che le classi annidate possano essere istanze di attributi di altre classi. La classe root ha un valore scarso in ODL_{I3} ma è stata mantenuta per mostrare come le relazioni di origine strutturale che legano la root alla prima classe figlia si mantengano nella GVV. Il coefficiente di affinità strutturale tra le due classi che compongono *Marchio* è nullo per lo stesso motivo esposto in precedenza riguardo alle relazioni trascurate: la relazione che legava *fibre2fashion.fibre2fashion* e *prontocomune.Prontocomune* è stata scartata in sede di validazione perché instaurata tra attributi appartenenti a domini non compatibili. Ancora una volta il SYN tra i nomi delle classi non richiede altra affinità, ma una relazione più debole non sarebbe stata sufficiente.

Tramite il procedimento descritto è stata ottenuta quella che sarà indicata come GVVold nel resto dell'esperimento.

3.2.2 Generazione della GVV da tutte e tre le sorgenti

Le due sorgenti già trattate si considerano importate nel sistema e con annotazione completata. Si importa e si annota nel modo più completo possibile la sorgente *subfor* per procedere a un'integrazione delle tre secondo il primo approccio, quindi senza considerare il lavoro fatto nel precedente processo di integrazione.

Generazione del Common Thesaurus

In generale la presenza di una terza sorgente non influisce sulle relazioni tra le altre due. Essa può al limite essere protagonista di nuove relazioni tra le altre due sorgenti quando esse sono generate per chiusura transitiva di relazioni strutturali nuove e semantiche vecchie. Si tratta di un'evenienza che si verifica molto raramente nel secondo approccio dove le due sorgenti iniziali si presentano nella forma già integrata in cluster (GVV), senza trasportare il pacchetto di relazioni elaborate per generarli eccezion fatta per alcune relazioni di origine strutturale in casi specifici (root XML). Nell'esempio, anche partendo dalle tre sorgenti separate, nessuna relazione del tipo descritto viene aggiunta. La tabella seguente riporta le relazioni valide ottenute in più rispetto all'integrazione vista in precedenza (quelle che coinvolgono *subfor* e una delle altre due classi). Ancora una volta sono evidenziate con un fondo più scuro le relazioni di origine strutturale.

Elemento 1	Relazione	Elemento 2
subfor.Subfor	RT	subfor.Azienda
subfor.Azienda	RT	subfor.Indirizzo
subfor.Settore	RT	subfor.SottoSettore
subfor.Azienda	RT	subfor.Settore
subfor.Azienda	SYN	prontocomune.Prontocomune.Azienda
subfor.Azienda	SYN	subfor.Subfor.Azienda
subfor.Azienda	SYN	fibre2fashion.Company
subfor.Azienda	SYN	fibre2fashion. fibre2fashion.Company
subfor.Azienda.DescrizioneAttivita	SYN	fibre2fashion.Category.Description
subfor.Azienda.DescrizioneAttivita	SYN	fibre2fashion.Company.AboutUs

subfor.Azienda.Nome	SYN	fibre2fashion.Company.Name
subfor.Azienda.SITO-INTERNET	SYN	fibre2fashion.Company.Web
subfor.indirizzo	SYN	prontocomune.Azienda.Indirizzo
subfor.indirizzo	SYN	subfor.Azienda.Indirizzo
subfor.indirizzo	SYN	fibre2fashion.Company.Address
subfor.indirizzo.CAP	NT	prontocomune.Categoria.CodiceCate..
subfor.indirizzo.CAP	NT	subfor.Settore.CodiceSettore
subfor.indirizzo.CAP	NT	fibre2fashion.Category.CategoryCode
subfor.indirizzo.Comune	NT	prontocomune.Indirizzo.Regione
subfor.indirizzo.Comune	NT	subfor.Indirizzo.Regione
subfor.Settore	BT	prontocomune.Azienda.Categoria
subfor.Settore	BT	subfor.Settore.SottoSettore
subfor.Settore	SYN	subfor.Azienda.Settore
subfor.Settore	SYN	fibre2fashion.Category
subfor.Settore	SYN	fibre2fashion.Company.Category
subfor.Settore.CodiceSettore	SYN	fibre2fashion.Category.CategoryCode
subfor.Settore.DescrizioneSettore	SYN	subfor.Azienda.DescrizioneAttivita
subfor.Settore.DescrizioneSettore	SYN	fibre2fashion.Category.Description
subfor.Settore.DescrizioneSettore	SYN	fibre2fashion.Company.AboutUs
subfor.SottoSettore	SYN	prontocomune.Azienda.Categoria
subfor.SottoSettore	SYN	subfor.Settore.SottoSettore
subfor.SottoSettore	NT	prontocomune.Categoria
subfor.SottoSettore	NT	subfor.Settore
subfor.SottoSettore	NT	subfor.Azienda.Settore
subfor.SottoSettore	NT	fibre2fashion.category
subfor.SottoSettore	NT	fibre2fashion.Company.Category
subfor.Subfor	SYN	fibre2fashion. fibre2fashion
subfor.Subfor	NT	prontocomune.Azienda.Nome
subfor.Subfor	NT	subfor.Azienda.Nome
subfor.Subfor	NT	fibre2fashion.Company.Name
prontocomune.Azienda	SYN	subfor.Azienda
prontocomune.Categoria.Descrizione	SYN	subfor.Azienda.DescrizioneAttivita
prontocomune.Indirizzo	SYN	subfor.Azienda.Indirizzo
prontocomune.Prontocomune	NT	subfor.Azienda.Nome
fibre2fashion.fibre2fashion	NT	subfor.Azienda.Nome
prontocomune.Azienda.Nome	SYN	subfor.Azienda.Nome
prontocomune.Azienda.Web	SYN	subfor.Azienda.SITO-INTERNET
prontocomune.Categoria	SYN	subfor.Azienda.Settore
fibre2fashion.Category	SYN	subfor.Azienda.Settore
prontocomune.Indirizzo	SYN	subfor.Indirizzo
prontocomune.Indirizzo.CAP	SYN	subfor.Indirizzo.CAP
prontocomune.Indirizzo.Comune	SYN	subfor.Indirizzo.Comune
prontocomune.Indirizzo.Comune	NT	subfor.Indirizzo.Regione
prontocomune.Indirizzo.Regione	SYN	subfor.Indirizzo.Regione
prontocomune.Indirizzo.Via	SYN	subfor.Indirizzo.Via
prontocomune.Categoria	SYN	subfor.Settore
prontocomune.Indirizzo.CAP	NT	subfor.Settore.CodiceSettore
prontocomune.Categoria.CodiceCat..	SYN	subfor.Settore.CodiceSettore
prontocomune.Categoria.Descrizione	SYN	subfor.Settore.DescrizioneSettore

prontocomune.Categoria	BT	subfor.Settore.SottoSettore
fibre2fashion.Category	BT	subfor.Settore.SottoSettore
prontocomune.Prontocomune	SYN	subfor.Subfor
prontocomune.Azienda	SYN	subfor.Subfor.Azienda
fibre2fashion.Company	SYN	subfor.Subfor.Azienda

Tabella 3.6

Risulta immediato notare come l'aggiunta di una terza sorgente non particolarmente complessa generi ugualmente una grande quantità di relazioni. Questo evento coinvolge soltanto il primo approccio dove la terza sorgente ha relazioni separate con le altre due. Seguendo il secondo approccio molte di queste relazioni sono sostituite dalle loro equivalenti tra le classi della sorgente e i cluster della GVV. In altre parole dove vi sono due relazioni che coinvolgono tre elementi di tre sorgenti ve ne sarà solo una che coinvolge un elemento e l'elemento globale che mappa gli altri due.

E' opportuna una ulteriore considerazione sulle relazioni strutturali. Data la struttura proposta per *subfor*, dovrebbe esistere una relazione RT di origine strutturale tra *subfor.Settore* e *Subfor.sottoSettore*. Infatti nella prima delle due classi esiste un attributo che fa riferimento alla seconda. Questa relazione viene scartata la momento della realizzazione delle relazioni di origine lessicale e rimpiazzata con una relazione NT (più forte) che lega il termine più stretto *subfor.SottoSettore* con il termine più generale *subfor.Settore*.

Generazione della GVV

La situazione trattata prevede che esistano più di due classi mappate nello stesso cluster, perciò i coefficienti di affinità strutturale e di naming più significativi possono essere tre per ciascun cluster. Essi esprimono l'affinità di ogni classe con ciascuna altra classe mappata nello stesso cluster. Allo scopo di evidenziare i coefficienti di affinità che legano le classi dell'esperimento si riportano le matrici di affinità. Queste matrici hanno un numero di righe e colonne pari al numero di classi e in ogni cella è presente il coefficiente di affinità che lega una classe (in colonna) con un'altra classe (riga). I coefficienti non sono presentati nella forma normalizzata (cioè moltiplicati per i valori predefiniti).

Matrice di affinità per il coefficiente di naming

e0 = fibre2fashion.Category
e1 = fibre2fashion.Company
e2 = fibre2fashion.fibre2fashion
e3 = prontocomune.Azienda
e4 = prontocomune.Categoria
e5 = prontocomune.Indirizzo
e6 = prontocomune.Prontocomune

e7 = subfor.Azienda
e8 = subfor.Indirizzo
e9 = subfor.Settore
e10 = subfor.SottoSettore
e11 = subfor.Subfor

	e0	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11
e0		0.05	0.25	0.05	1.00	0.25	0.25	0.05	0.25	1.00	0.08	0.25
e1	0.05		0.05	1.00	0.05	0.05	0.05	1.00	0.05	0.05	0.04	0.05
e2	0.25	0.05		0.05	0.25	0.25	1.00	0.05	0.25	0.25	0.02	1.00
e3	0.05	1.00	0.05		0.05	0.05	0.05	1.00	0.05	0.05	0.04	0.05
e4	1.00	0.05	0.25	0.05		0.25	0.25	0.05	0.25	1.00	0.08	0.25
e5	0.25	0.05	0.25	0.05	0.25		0.25	0.05	1.00	0.25	0.02	0.25
e6	0.25	0.05	1.00	0.05	0.25	0.25		0.05	0.25	0.25	0.02	1.00
e7	0.05	1.00	0.05	1.00	0.05	0.05	0.05		0.05	0.05	0.04	0.05
e8	0.25	0.05	0.25	0.05	0.25	1.00	0.25	0.05		0.25	0.02	0.25
e9	1.00	0.05	0.25	0.05	1.00	0.25	0.25	0.05	0.25		0.08	0.25
e10	0.08	0.04	0.02	0.04	0.08	0.02	0.02	0.04	0.02	0.08		0.02
e11	0.25	0.05	1.00	0.05	0.25	0.25	1.00	0.05	0.25	0.25	0.02	

Tabella 3.7: Matrice di affinità per il coefficiente di naming

Matrice di affinità per il coefficiente strutturale

e0 = fibre2fashion.Category
e1 = fibre2fashion.Company
e2 = prontocomune.Azienda
e3 = prontocomune.Categoria
e4 = prontocomune.Indirizzo
e5 = subfor.Azienda
e6 = subfor.Indirizzo
e7 = subfor.Settore

	e0	e1	e2	e3	e4	e5	e6	e7
e0		0.09		0.05	0.16	0.07	0.11	0.04
e1	0.09		0.31	0.09		0.14		0.08
e2		0.31				0.10		
e3	0.05	0.09			0.16	0.07	0.11	0.04
e4	0.16			0.16			0.54	0.14
e5	0.07	0.14	0.10	0.07				0.06
e6	0.11			0.11	0.54			0.01
e7	0.04	0.08		0.04	0.14	0.06	0.01	

Tabella 3.8: Matrice di affinità per il coefficiente strutturale

Matrice di affinità globale

e0 = fibre2fashion.Category
e1 = fibre2fashion.Company
e2 = fibre2fashion.fibre2fashion
e3 = prontocomune.Azienda
e4 = prontocomune.Categoria
e5 = prontocomune.Indirizzo

e6 = prontocomune.Prontocomune
e7 = subfor.Azienda
e8 = subfor.Indirizzo
e9 = subfor.Settore
e10 = subfor.SottoSettore
e11 = subfor.Subfor

	e0	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11
e0		0.29	0.12	0.25	0,052	0.20	0.12	0.28	0.18	0.07	0.04	0.12
e1	0.29		0.25	0,045	0.29	0.25	0.25	0.57	0.25	0.29	0.02	0.25
e2	0.12	0.25		0.25	0.12	0.12	0.05	0.25	0.12	0.12	0.01	0.05
e3	0.25	0,045	0.25		0.25	0.25	0.25	0.55	0.25	0.25	0.02	0.25
e4	0,052	0.29	0.12	0.25		0.20	0.12	0.28	0.18	0.07	0.04	0.12
e5	0.20	0.25	0.12	0.25	0.20		0.12	0.25	0,053	0.19	0.01	0.12
e6	0.12	0.25	0.05	0.25	0.12	0.12		0.25	0.12	0.12	0.01	0.05
e7	0.28	0.57	0.25	0.55	0.28	0.25	0.25		0.25	0.28	0.02	0.25
e8	0.18	0.25	0.12	0.25	0.18	0,053	0.12	0.25		0.17	0.01	0.12
e9	0.07	0.29	0.12	0.25	0.07	0.19	0.12	0.28	0.17		0.04	0.12
e10	0.04	0.02	0.01	0.02	0.04	0.01	0.01	0.02	0.01	0.04		0.01
e11	0.12	0.25	0.05	0.25	0.12	0.12	0.05	0.25	0.12	0.12	0.01	

Tabella 3.9: Matrice di affinità globale

Queste matrici di affinità, e in particolare quella di affinità globale, sottolineano valori sufficienti ad affermare la formazione dei seguenti cluster. Essi hanno somiglianza con quelli presentati nella procedura di integrazione precedente ma hanno in più gli attributi di *subfor*. La formazione di questi cluster esprime la realizzazione della GVV dalle tre sorgenti separate.

Marchio (rinominata)	fibre2fashion(fibre2fashion)	Prontocomune(prontocomune)	Subfor(subfor)
Company	Company	Azienda	Azienda

Category	Categoria(prontocomune)	Category(fibre2fashion)	Settore(subfor)
SottoSettore			SottoSettore
CategoryCode	CodiceCategoria	CategoryCode	CodiceSettore
Descrizione	Descrizione	Description	DescrizioneSettore

Indirizzo	Indirizzo(subfor)	Indirizzo(prontocomune)
Via	Via	Via
TEL	TEL	
POSTA-ELETTRONICA	POSTA-ELETTRONICA	
CAP	CAP	CAP
FAX	FAX	
Regione	Regione	Regione
Comune	Comune	Comune

Azienda	Company(fibre2fashion)	Azienda(prontocomune)	Azienda(subfor)
ContactPerson	ContactPerson		CONTATTI
Indirizzo	Address	Indirizzo	Indirizzo

AboutUs	AboutUs		DescrizioneAttivita
Fax	Fax	Fax	
E-Mail	E-Mail	E-mail	
Web	Web	Web	SITO-INTERNET
Tel	Tel	Telefono	
Category	Category	Categoria	
Nome	Name	Nome	Nome
Settore	Category	Categoria	Settore
ADDETTI			ADDETTI
CONTROLLI-QUALIT.			CONTROLLI-QUALITA
CAPITALE-SOCIALE			CAPITALE-SOCIALE
FATTURATO			FATTURATO
TECNOLOGIE			TECNOLOGIE
ORDINI			ORDINI

SottoSettore	SottoSettore(subfor)
Garante	Garante
TitoloSottoSettore	TitoloSottoSettore
DataIstituzione	DataIstituzione

Tabella 3.10

Gli obiettivi di integrazione siano stati raggiunti. Le tre sorgenti integrate tra di loro hanno dato luogo a una GVV costituita da cinque classi globali. La prima di esse, *Marchio*, è stata mantenuta per la stessa ragione affermata nella sezione precedente. La classe *Categoria* contiene attributi provenienti da una classe affine per ciascuna sorgente. Si può notare come l'apporto della sorgente *subfor* aggiunga alcuni attributi a quelli mappati nel cluster. Questo evento è in linea con il primo scenario previsto. Lo stesso si può affermare per le classi *Indirizzo* e *Azienda*. La prima delle due è generata dall'apporto delle sole sorgenti *pronotocomune* e *subfor*, la seconda dal contributo di tutte e tre le sorgenti.

La classe *Azienda* è l'espressione più evidente del verificarsi del primo scenario: l'integrazione della classe *subfor.Azienda* è determinata dalla forte affinità nel nome (relazione SYN) tra le classi e non richiede altre affinità. L'aggregazione si è dunque ottenuta anche con un debole coefficiente di affinità strutturale: 0,05 tra *subfor.Azienda* e *prontocomune.Azienda* e 0,07 tra *subfor.Azienda* e *fibre2fashion.Company*. I bassi coefficienti riportati esprimono la scarsa sovrapposizione degli attributi nel cluster azienda. L'esempio si riferisce a una situazione realistica in ciascuna sorgente dispone di una struttura per organizzare dati con lo stesso obiettivo espressivo ma i valori di interesse per ciascuna sorgente (attributi) sono in larga misura differenti. La sinonimia tra i nomi delle classi ne determina ugualmente la giusta unione in un unico cluster.

Il secondo scenario è espresso nell'ultima classe globale: *SottoSettore*. Confrontando la GVV ottenuta dalle tre sorgenti con quella ottenuta dalle due iniziali si nota che questa classe è in più.

Subfor.SottoSettore non ha infatti legami sufficienti con classi di altre sorgenti per determinare alcuna integrazione (si vedano la riga e la colonna 9 della matrice di affinità globale). Per questo motivo la classe è mappata in un cluster separato.

La procedura appena trattata ha ottenuto quella che è stata definita GVV_total seguendo il primo approccio previsto, pertanto l'integrazione di una nuova sorgente è avvenuta senza differenze rispetto al processo descritto nel capitolo 1. Nella prossima sezione sarà trattato il procedimento che prevede un processo di integrazione tra la GVV generata dalle due sorgenti iniziali e *subfor*.

3.2.3 Generazione della GVV da una GVV e una nuova sorgente

Questa procedura utilizza come sorgente la GVV risultante dall'integrazione di *prontocomune e fibre2fashion*. La sorgente da aggiungere al progetto è *subfor*. La seguente tabella riporta le struttura della GVV (GVVold) espressa come sorgente completamente annotata e importata nel sistema:

Sorgente: GVVold

Nome e tipo attributo	Forma e senso attributo

Indirizzo (address, 2)	
CAP (string)	zip_code (1)
Via (string)	street (1)
Regione (string)	region (3)
Comune (string)	town (1)

Azienda (business_organization, 1)	
AboutUs (string)	description (1)
Fax (string)	fax (1)
E-Mail (string)	e-mail (1)
ContactPerson (string)	contact (5)
Web (string)	web (5)
Address (string)	address (2)
Category (Categoria)	sector (2)
Telefono (string)	telephone (1)
Name (string)	name (1)

Marchio (trade_name, 1)	
Azienda (set <Azienda>)	business_organization (1)

Categoria (sector, 2)	
Description (string)	description (1)
CategoryCode (string)	code (2)

Tabella 3.11

La seconda sorgente da integrare, *subfor*, si considera identica a quella già espressa per la rappresentazione del primo approcio. L'annotazione delle due sorgenti è completa, si procede alla creazione del Common Thesaurus.

Generazione del Common Thesaurus

La seguente tabella presenta le relazioni valide tra *GVVold* e *subfor*. Queste relazioni non hanno alcun collegamento con quelle che nel caso precedente si erano instaurate tra *subfor* e le due classi che compongono *GVVold* in quanto al suo interno non esiste traccia delle due classi separate che abbia valore in questa fase. Le relazioni evidenziate dal fondo più scuro sono quelle di origini strutturale.

Elemento 1	Relazione	Elemento 2
GVVold.Azienda	RT	GVVold.Categoria
GVVold.Marchio	RT	GVVold.Azienda
subfor.subfor	RT	subfor.Azienda
subfor.Azienda	RT	subfor.Indirizzo
subfor.Azienda	RT	subfor.Settore
GVVold.Azienda	SYN	GVVold.Marchio.Azienda
GVVold.Azienda	SYN	subfor.Subfor.Azienda
GVVold.Azienda	SYN	subfor.Azienda
GVVold.Azienda.AboutUs	SYN	GVVold.categoria.Description
GVVold.Azienda.AboutUs	SYN	subfor.Settore.DescrizioneSettore
GVVold.Azienda.AboutUs	SYN	subfor.Azienda.DescrizioneAttivita
GVVold.Azienda.Name	SYN	subfor.Azienda.Nome
GVVold.Azienda.Web	SYN	subfor.Azienda.SITO-INTERNET
GVVold.Categoria	BT	subfor.Settore.SottoSettore
GVVold.Categoria	SYN	GVVold.Azienda.Category
GVVold.Categoria	SYN	subfor.Settore
GVVold.Categoria	SYN	subfor.Azienda.Settore
GVVold.Categoria.CategoryCode	SYN	subfor.Settore.CodiceSettore
GVVold.Categoria.Description	SYN	subfor.Settore.DescrizioneSettore
GVVold.Categoria.Description	SYN	subfor.Azienda.DescrizoneAttivita
GVVold.Indirizzo	SYN	GVVold.Azienda.Address
GVVold.Indirizzo	SYN	subfor.Indirizzo
GVVold.Indirizzo	SYN	subfor.Azienda.Indirizzo
GVVold.Indirizzo.CAP	SYN	subfor.Indirizzo.CAP
GVVold.Indirizzo.CAP	NT	GVVold.Categoria.CategoryCode
GVVold.Indirizzo.CAP	NT	subfor.Settore.CodiceSettore

GVVold.Indirizzo.Comune	SYN	subfor.Indirizzo.Comune
GVVold.Indirizzo.Comune	NT	GVVold.Indirizzo.Regione
GVVold.Indirizzo.Comune	NT	subfor.Indirizzo.Regione
GVVold.Indirizzo.Regione	SYN	subfor.Indirizzo.Regione
GVVold.Indirizzo.Via	SYN	subfor.Indirizzo.Via
GVVold.Indirizzo.Marchio	SYN	subfor.Subfor
GVVold.Indirizzo.Marchio	NT	GVVold.Azienda.Name
GVVold.Indirizzo.Marchio	NT	subfor.Azienda.Nome
subfor.Azienda	SYN	GVVold.Marchio.Azienda
subfor.Azienda	SYN	subfor.Subfor.Azienda
subfor.Indirizzo	SYN	GVVold.Azienda.Address
subfor.Indirizzo	SYN	subfor.Azienda.Indirizzo
subfor.Indirizzo.CAP	NT	GVVold.Categoria.CategoryCode
subfor.Indirizzo.CAP	NT	subfor.Settore.CodiceSettore
subfor.Indirizzo.Comune	NT	GVVold.Indirizzo.Regione
subfor.Indirizzo.Comune	NT	subfor.Indirizzo.Regione
subfor.Settore	BT	subfor.Settore.SottoSettore
subfor.Settore	SYN	GVVold.Azienda.Category
subfor.Settore	SYN	subfor.Azienda.Settore
subfor.Settore.DescrizoneSettore	SYN	subfor.Azienda.DescrizioneAttivita
subfor.SottoSettore	SYN	subfor.Settore.SottoSettore
subfor.SottoSettore	NT	GVVold.Azienda.Category
subfor.SottoSettore	NT	GVVold.Categoria
subfor.SottoSettore	NT	subfor.Settore
subfor.SottoSettore	NT	subfor.Azienda.Settore
subfor.Subfor	NT	GVVold.Azienda.Nome
subfor.Subfor	NT	subfor.Azienda.Nome

Tabella 3.12

Balza subito all'occhio il numero delle relazioni generate. La loro quantità è drasticamente inferire a quella considerata nella generazione della GVV dalle tre sorgenti separate. In particolare il numero di relazioni valide realizzate tramite il secondo approccio è 53 mentre quello ottenuto per il primo approccio è 97, quindi il secondo approccio. In altri termini il secondo approccio ha prodotto il 54,6% di relazioni del secondo. Ci si aspettava una situazione di questo genere a favore del secondo approccio.

Un altro particolare da notare riguarda la presenza di relazioni di origine strutturale all'interno di GVVold. Si tratta di *GVVold.Azienda RT GVVold.Categoria* e *GVVold.Marchio RT GVVold.Azienda*. La presenza di relazioni di origine strutturale in una GVV è quantomeno insolita poiché la GVV non conserva una memoria della relazioni alla base del Thesaurus da cui è stata ricavata. In questo caso si sono mantenute relazioni strutturali perché entrambe le sorgenti erano di origine semistrutturata ed erano sorgenti XML con percorsi analoghi. In entrambi era presente una root (Marchio) a cui era collegata Azienda alla quale era collegata Categoria. La struttura si è mantenuta tale nell'integrazione simmetrica delle sorgenti, quindi nella GVV.

Generazione della GVV

In questo caso le sorgenti da integrare sono due, quindi non è necessario riportare le matrici di affinità che legano le classi. Infatti, ponendosi nella sperimentazione dei primi due casi di integrazione, i cluster sono formati da al massimo due classi. Non si verifica l'evento in cui forti relazioni con una classe comune fanno unire due classi che altrimenti sarebbero disgiunte. I cluster generati in questo passaggio sono riportati in seguito.

Indirizzo	Indirizzo(GVVold)	Idirizzo(subfor)
CAP	CAP	CAP
POSTA-ELETTRONICA		POSTA-ELETTRONICA
FAX		FAX
Via	Via	Via
Regione	Regione	Regione
TEL		TEL
Comune	Comune	Comune

NA = 0,5
SA = 0,27

SottoSettore	SottoSettore(subfor)
Garante	Garante
DataIstituzione	DataIstituzione
TitoloSottoSettore	TitoloSottoSettore

Marchio	Sobfor(subfor)	Marchio(GVVold)
Azienda	Azienda	Azienda

NA = 0,5
SA = 0

Azienda	Azienda(GVVold)	Azienda(subfor)
ADDETTI		ADDETTI
CONTROLLI-QUALITA		CONTROLLI-QUALITA
E-Mail	E-Mail	
FATTURATO		FATTURATO
DescrizioneAttivita	AboutUs	DescrizioneAttivita
Nome	Name	Nome
TECNOLOGIE		TECNOLOGIE
SITO-INTERNET	Web	SITO-INTERNET
Fax	Fax	
CONTATTI	ContactPerson	CONTATTI
Address	Address	Indirizzo
ORDINI		ORDINI
Category	Category	Settora
CAPITALE-SOCIALE		CAPITALE-SOCIALE
Telefono	Telefono	

NA = 0,5
SA = 0,07

Categoria	Settore(subfor)	Categoria(GVVold)
SottoSettore	SottoSettore	
DescrizioneSettore	DescrizioneSettore	Description
CodiceSettore	CodiceSettore	CategoryCode

NA = 0,5
SA = 0,2

Tabella 3.13

3.3 Confronto tra le due GVV

Figura 3.1: Confronto tra le due GVV

Le due immagini rappresentano le i cluster ottenuti seguendo i due procedimenti. Purtroppo non è possibile fornire una rappresentazione esaustiva della seconda GVV (a destra) in cui risulti chiaro come le classi mappate nella vecchia GVV (usata come sorgente) siano nuovamente esplose come classi locali nella mappatura, per rendere possibile la sesura di query sulla GVV ottenuta. Il numero dei cluster è lo stesso. Il confronto tra i due approcci non si può quindi basare sul risultato, conseguito senza difficoltà progettuali in entrambi i casi riproducendo gli scenari di integrazione previsti. La scelta tra i due dipenderà dal numero e dalla varietà di operazioni da compiere. Durante lo svolgimento dell'esperimento sono state riscontrate le seguenti osservazioni:

Numero delle operazioni

Il secondo approccio è decisamente meno oneroso nel numero delle operazioni richieste al sistema. Già dalla fase di generazione del common thesaurus pare più agevole impegnare il sistema con meno elementi tra cui calcolare relazioni. Il numero delle relazioni riscontrato, 53 per il secondo approccio contro le 97 del primo è indice di come anche in una situazione in cui sono coinvolte poche classi la differenza tra il numero delle relazioni sia così elevata (il secondo approccio ha prodotto il 54,6% delle relazioni del primo). Il secondo approccio ha dunque un numero di elementi inferiori rispetto al primo (numerosi gruppi di elementi sono già mappati in elementi globali), di conseguenza può risultare più rapido nella realizzazione, in particolare quando le strutture delle sorgenti sono complesse. Questa osservazione ha anche riscontro nella realizzazione dei cluster, dove il sistema deve affrontare matrici di affinità che coinvolgono un minor numero di classi.

Potere espressivo

Salvo alcune eccezioni (come dimostrato nell'esperimento dalle root XML) il secondo approccio ha un potere espressivo più ridotto rispetto al primo. Nella maggior parte dei casi le relazioni di origine strutturale che sono state considerate nel calcolo della prima GVV non sono poi utilizzabili nel processo di aggiunta. Infatti la GVV non mantiene al suo interno una struttura tale da giustificare la creazione di tali relazioni, le quali altrimenti sarebbero desunte da sorgenti semistrutturate (XML) o strutturate (Relazionali). La scelta del primo approccio preserva totalmente la struttura delle sorgenti e sfrutta tutto il potere espressivo delle relazioni trattate.

Automazione e indipendenza progettuale

Gli esperimenti sono stati svolti senza prevedere interventi da parte del progettista (eccetto nel rinominare un cluster per chiarezza). Interventi alla struttura della GVV sono comunque possibili a partire dalla fase di annotazione, dove il progettista sceglie il significato da considerare giusto e la word form da affidare a ogni elemento. Nell'ipotesi di una necessità da parte del progettista di intervenire manualmente in alcune fasi del processo di integrazione il primo approccio diviene più conveniente. Esso infatti non sacrifica nessuna informazione ottenuta dalla struttura delle sorgenti importate (anche se sono state già integrate) e consente al progettista ti operare su parametri che non sarebbero più accessibili utilizzando una GVV ad espressione di alcune sorgenti. Al contrario il secondo approccio offre la possibilità di operare un processo automatico di interazione dedicandosi il meno possibile a dettagli progettuali.

4 Aggiunta di una nuova sorgente ad una GVV: esperimento

Il terzo scenario di integrazione necessita di uno studio più approfondito, differente dalla valutazione del comportamento del sistema nei due approcci di integrazione possibili. Per affrontare lo scenario si consideri una situazione semplificata che prevede solo 3 classi, due delle quali rappresentano due cluster distinti di una stessa GVV e la terza appartiene a una nuova sorgente da integrare. La situazione esemplifica correttamente le occorrenze del terzo scenario di integrazione se l'unione dei due cluster a opera di una nuova classe è determinata prevalentemente da relazioni che i due cluster hanno rispettivamente con tale classe e in minor misura tra di loro. Dato questo obiettivo non è necessario un numero elevato di classi da unire per studiare la situazione. Inoltre il fatto che i due cluster siano il risultato di un processo di integrazione precedente è utilizzato solo come ipotesi con cui confrontare volta per volta la realizzabilità delle casistiche. Di seguito sarà presentato un esperimento di riferimento, che riproduce una delle occorrenze più frequenti del terzo scenario, ma tutte le possibili situazioni di integrazione saranno trattate in modo analitico prescindendo da un esempio con dati reali (nomi e annotazioni) ma basandosi sulla valutazione dei coefficienti di affinità.

4.1 Le sorgenti dell'esempio

Per rappresentare una particolare occorrenza de terzo scenario si utilizzano due sorgenti. Una, *GVVold* contiene i due cluster (denominati A e B nell'analisi successiva) da unire, l'altra, *SorgenteC*, contiene la classe che deve unire i due cluster (denominata C nell'analisi).

Sorgente 1: GVVold

Nome e tipo attributo	Forma e senso attributo

Agenzia (agency, 2) (classe A, cluster)	
Clienti (string)	customer (1)
Fatturato (string)	revenue (1)
Titolare (string)	owner (1)
Nome (string)	name (1)
Descrizione (string)	description (1)

Azienda_manifatturiera (manufacturer, 1) (classe B, cluster)	
Indirizzo (string)	address (2)
Prodotti (string)	product (1)
Categoria (string)	business_sector (1)
Stabilimenti (string)	facility (1)
Descrizione (string)	description (1)

Tabella 4.1

Sorgente 2: SorgenteC

Nome e tipo attributo	Forma e senso attributo

Organizzazione (business_organization, 1) (classe C, nuova)	
Anno_istituzione (string)	year (2)
Controllo_qualita (string)	quality_control (1)
Indirizzo (string)	address (2)
Clienti (string)	customer (1)
Presidente (string)	president (1)
Prodotti (string)	product (1)
Titolare (string)	owner (1)
Categoria (string)	business_sector (1)
Nome (string)	name (1)

Tabella 4.2

L'annotazione del nome delle due classi *GVVold.Agenzia* e *GVVold.Azienda_manifatturiera*, come anche la loro affinità strutturale, non sono sufficienti a unire le due classi senza l'apporto della classe *SorgenteC.Organizzazione*. I due cluster hanno 5 attributi di cui 3 sono presenti anche in *Organizzazione*, per ricreare l'affinità strutturale con la classe che deve unirle, uno comune alle classi: *Descrizione*. Un terzo attributo per ciascuna delle due classi si presenta solo il quella classe: i due attributi sono *GVVold.Agenzia.Fatturato* e *GVVold.Azienda_manifatturiera.Stabilimenti*. Anche la classe *Organizzazione* introduce attributi che non sono presenti in altre classi. Nessuna relazione di origine strutturale è stata utilizzata per questo esempio.

4.2 Svolgimento dell'esperimento

Il primo passo è quello di importare la sorgente GVVold e la sorgente *SorgenteC* come parti di un nuovo progetto di integrazione. La prima è già nella forma consona al trattamento nel sistema (è infatti il risultato di un precedente processo di integrazione), per la seconda si utilizza un wrapper DTD.

Generazione del Common Thesaurus

Considerando completata l'annotazione delle sorgenti nella loro descrizione del paragrafo precedente, si procede direttamente alla realizzazione del Common Thesaurus. Il tentativo di derivare relazioni del Thesaurus dallo schema non porta alcun risultato. Infatti non ci sono relazioni strutturali tra le classi e gli attributi delle sorgenti importate. Le relazioni derivate dal lessico sono rappresentate nella seguente tabella:

Elemento 1	Relazione	Elemento 2
GVVold.Azienda_manifatturiera	NT	sorgenteC.Organizzazione
GVVold.Agenzia	NT	sorgenteC.Organizzazione
GVVold.Azienda_manifatturiera	RT	GVVold.Agenzia
GVVold.Agenzia.Clienti	SYN	sorgenteC.Organizzazione.Clienti
GVVold.Agenzia.Titolare	SYN	sorgenteC.Organizzazione.Titolare
GVVold.Agenzia.Nome	SYN	sorgenteC.Organizzazione.Nome
GVVold.Agenzia.Descrizione	SYN	GVVold.Azienda_m.Descrizione
GVVold.Azienda_m.Prodotti	SYN	sorgenteC.Organizzazione.Prodotti
GVVold.Azienda_m.Indirizzo	SYN	sorgenteC.Organizzazione.Indirizzo
GVVold.Azienda_m.Categoria	SYN	sorgenteC.Organizzazione.Categoria

Tabella 4.3

Nota: *Azienda_manifatturiera* è stato abbreviato in alcuni casi a *Azienda_m*.

Tutte le relazioni sono valide e nessuna relazione è aggiunta per opera del progettista o chiusura transitiva. Nella tabella sono evidenziate dallo sfondo colorato le relazioni tra i nomi di classe.

Generazione della GVV

Una volta costituito il Common Thesaurus si procede con la realizzazione dei cluster. A scopo esplicativo le relazioni rilevanti per la determinazione del Naming Affinity Coefficient sono state evidenziate dallo sfondo grigio. Il valore predefinito delle relazioni NT è 0.8, per le relazioni RT si usa 0.5. Il sistema calcola il coefficiente di affinità legato al nome scegliendo il cammino minore (cioè quello corrispondente a un valore più alto) tra le classi. In questo caso il cammino minore tra *Agenzia* e *Azienda_manifattuirera* passa per *Organizzazione*, infatti il prodotto dei valori delle due relazioni NT che legano le due classi secondo questo cammino è un valore maggiore a quello che le lega lungo il cammino diretto determinato dalla relazione RT (0,8 x 0,8 = 0,64 > 0,5). Calcolato il Naming Affinity Coefficient il sistema lo moltiplicherà per il coefficiente di normalizzazione legato al nome (0,64 x 0,5 = 0,32). Il coefficiente di affinità strutturale è calcolato in base al numero degli attributi in relazione diviso il numero degli

attributi totali, moltiplicando il risultato per il coefficiente di normalizzazione, ottenendo un valore di 0,05 (1/10 x 0,5). Il coefficiente di affinità globale è dunque 0,37. Questo coefficiente è decisamente inferiore alla soglia (0,37 < 0,5), quindi non giustifica l'unione delle classi *Agenzia Azienda_manifatturiera*. L'unione avviene infatti in virtù dei coefficienti di affinità di ciascuna delle due classi con la nuova classe *Organizzazione*. Le relazioni tra ciascuna delle classi di partenza con *Organizzazione* sono simmetriche sia nella quantità che nella qualità, quindi è sufficiente trattare il calcolo del coefficiente per una delle due classi, si consideri dunque *Agenzia*. Il Naming Affinity Coefficient che la lega a *Organizzazione* è più alto di quello visto in precedenza ed è dovuto al peso della relazione RT moltiplicato per il coefficiente predefinito (0,8 x 0,5 = 0,4). Anche il coefficiente di affinità strutturale è più elevato poiché più attributi sono in relazione tra le due classi rispetto al caso precedente (3/14 x 0,500 = 0,107). Il coefficiente di affinità globale si afferma dunque su 0,507, un valore superiore alla soglia (0,400 + 0,107 = 0,507 > 0,5) anche se di poco. In base a questo calcolo in sistema mette nello stesso cluster *Agenzia* e *Organizzazione*. Come detto il coefficiente di affinità globale tra *Organizzazione* e *Azienda_manifatturiera* è lo stesso, quindi il sistema unisce al cluster anche quest'ultima classe. La conseguenza è che, ad opera delle forti relazioni (sia semantiche che strutturali) delle due classi di partenza con quella aggiunta, le tre classi sono agglomerate nello stesso cluster.

Annotazione della GVV

L'annotazione della GVV non presenta problemi in situazioni come questa. Nell'esempio, dal punto di vista dei significati dei nomi, una classe padre unifica due classi figlie, dando loro una struttura più generale; di conseguenza è evidente come la preponderanza di attributi presenti in *Organizzazione* faccia scegliere al sistema un'annotazione del cluster coincidente con quella di questa classe.

4.3 Analisi delle casistiche

4.3.1 Preliminari all'analisi

Il terzo scenario ha caratteristiche che necessitano una analisi più approfondita: alcune occorrenze risultano impossibili, altre possibili con restrizioni, altre ancora improbabili. In questo paragrafo è riportata un'analisi di ciascuna delle casistiche emerse dalle riflessioni sul terzo scenario. I valori di riferimento non sono più le classi specifiche e relazioni tra gli attributi, ma i coefficienti di affinità tra le classi. In questa fase le classi in esame sono tre e l'ipotesi presentata è che due classi A e B siano il risultato di una precedente procedura di integrazione e una classe C appartenga a una nuova sorgente. Nel rispetto degli eventi classificati come terzo scenario l'unione delle classi A e B deve essere determinata dalle relazioni che ciascuna ha con C, pertanto A e B devono essere disgiunte per ipotesi, cioè appartenenti a due cluster separati generati in un processo di integrazione precedente. Nella trattazione delle occorrenze del terzo scenario sono analizzati o coefficienti di affinità di naming associati alle terne di relazioni che legano le classi e i coefficienti di affinità strutturale necessari ad affermare la realizzabilità di ciascun caso. I grafici esplicativi associati alle trattazioni sono da leggersi come esposto nella seguente figura.

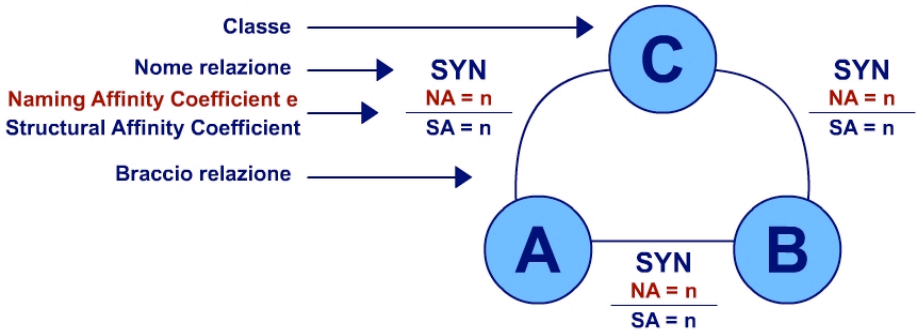

Figura 4.1: Riferimento

La figura rappresenta i collegamenti tra le tre classi. Accanto a ciascun collegamento è riportata la relazione tra le due classi, completo del coefficiente di affinità legato al nome (il più alto, in rosso) e del coefficiente legato alla struttura (più in basso, in blu). Nel corso della trattazione è possibile che sia fatto riferimento a un insieme di valori accettabili per i coefficienti. Si tenga presente che, nella forma in cui sono presentati, essi devono essere compresi tra 0 e 0,5 e che ad esempio un intervallo indicato come $SA < 0,1$ è da intendersi come $0 < SA < 0,1$. I

coefficienti sono presenti nella forma già computata: cioè già moltiplicata per il coefficiente fornito di normalizzazione.

4.3.2 Caso 1: nessuna affinità lessicale con C

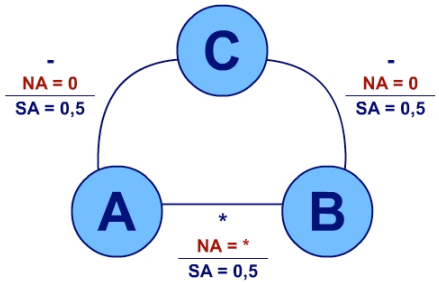

Figura 4.2: Caso 1

La situazione è rappresentata in figura. Si impone che nessuna relazione di origine lessicale esista tra A e C e tra B e C. Perché si verifichi l'integrazione delle tre classi è quindi necessario che il solo coefficiente di affinità strutturale superi la soglia, quindi che le classi rispettivamente A e C e B e C siano strutturalmente identiche. In modo transitivo questo causa l'identità strutturale tra A e B. Se A e B sono strutturalmente identiche non possono presentarsi come due cluster separati, indipendentemente dalla relazione lessicale che lega il loro nomi. La situazione è dunque impossibile.

4.3.3 Caso 2: nessuna affinità strutturale con C

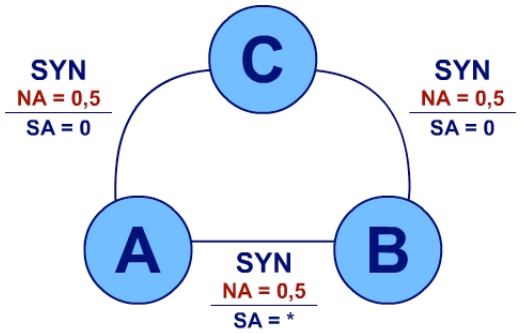

Figura 4.3: Caso 2

Se non esiste alcuna relazione strutturale tra A e C e B e C, necessariamente il coefficiente di affinità legato al nome deve essere tale da consentire l'integrazione. Un coefficiente così alto può essere ottenuto solo tramite relazioni SYN tra in nomi di A e C, B e C. Ancora una volta la

chiusura transitiva delle relazioni impone un vincolo, quello del SYN tra i nomi di A e B non compatibile con la presenza di due cluster separati. Anche questo caso non si può verificare.

4.3.4 Caso 3: relazioni RT di A e B con C

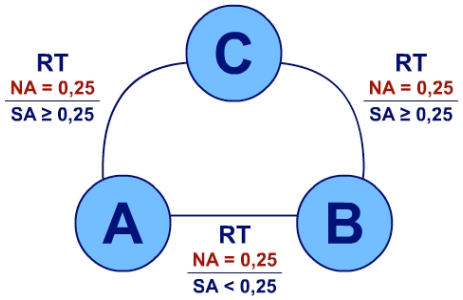

Figura 4.4: Caso 3

In questo caso sono presenti relazioni RT tra A e C e tra B e C. Il debole coefficiente di affinità sul naming fornito dalle relazioni RT necessita che lo Structural Affinity Coefficient sia almeno pari a 0,25 per garantire l'integrazione. Inoltre i nomi delle classi A e B devono avere necessariamente una relazione tra di loro (sono entrambi RT di uno stesso termine). Posto che una relazione di SYN renderebbe il caso irrealizzabile, in figura si è scelta la relazione di RT perché è quella che ha meno peso. Tutto ciò consente di affermare un'affinità strutturale minore di 0,25 tra A e B per garantire che esse siano disgiunte. Se la relazione fosse di NT/BT, sarebbe necessario un SA minore di 0,1. Questa casistica è possibile, per quanto poco realistica. Infatti si propone una situazione di aderenza maggiore 50% di A con C e B con C nella struttura, mentre la stessa aderenza si pone minore del 50% tra A e B, il tutto richiedendo relazioni sui nomi piuttosto deboli. Il caso in cui al posto della RT si presenti una NT/BT è ancora più improbabile ma realizzabile in teoria.

4.3.5 Caso 4: relazioni NT di A e B con C

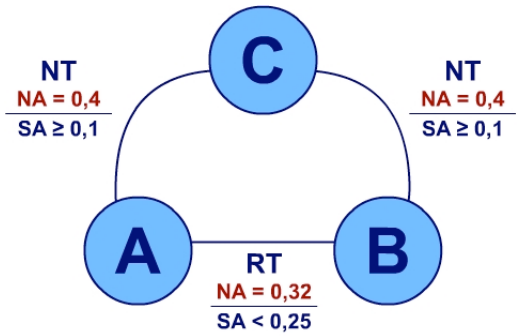

Figura 4.5: Caso 4

La prima espressione di questo caso è quella a cui si riconduce l'esperimento visto precedentemente. Esistono relazioni semantiche NT tra A e B e tra A e C, questo richiede che il coefficiente di affinità strutturale sia almeno pari a 0,1, quindi che un attributo su cinque sia in relazione tra le classi di ciascuna coppia. E' da notare che il coefficiente di affinità legato la nome tra A e B è 0,32, e non 0,25. Questo accade perché il percorso minore tra A e B è quello passando per C (0,8 x 0,8 = 0,64 > 0,5). Ciononostante si continua a richiedere che il SA tra A e B sia inferiore a 0,25 per assicurare che siamo divise. Infatti nel precedente processo di integrazione ipotizzato (senza C) sarebbe stato considerato un NA tra A e B pari a 0,25. In questa rappresentazione per A e B è stata scelta una relazione RT, tuttavia potrebbe esserci anche una relazione NT, questa casistica è rappresentata in figura, dove sono state aggiunte le frecce per simboleggiare la direzione delle relazioni, importante più che negli altri casi.

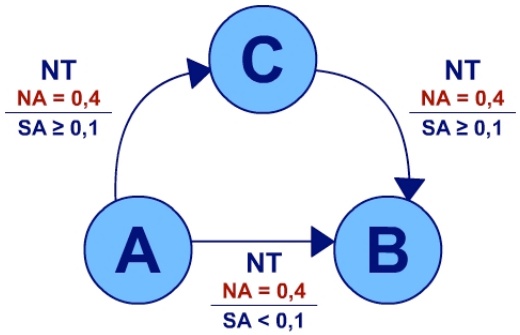

Figura 4.6: Caso 4, seconda rappresentazione

La situazione in esame è realizzabile, ma prevede una affinità strutturale tra A e B molto ridotta (0,1 cioè un attributo su cinque in relazione). La direzione delle frecce indica la chiave di lettura della relazione: esiste una classe A il cui nome ha un significato più ristretto dei nomi

delle lassi B e C, dove però C ha un nome dal significato più ristretto di B. Nei significati B è "nonno" di A e "padre" di C, quindi C è "padre" di A.

4.3.6 Caso 5: SYN tra A e C o tra B e C

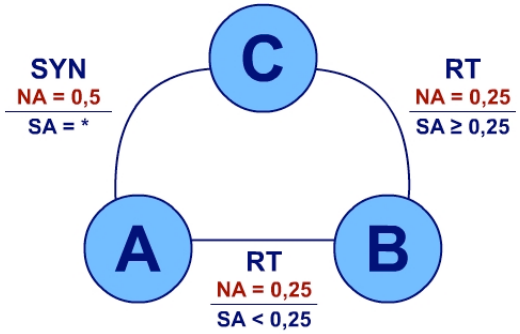

Figura 4.7: Caso 5

Quella in figura è soltanto una delle numerose rappresentazioni possibili che prevedono un SYN tra i nomi di A e C o B e C e altri tipi di relazioni differenti. La situazione rappresentata prevede il caso più generale, in cui B RT C e A RT B. Per la realizzabilità non è rilevante il SA tra A e B in quanto l'aggregazione è assicurata dal SYN, mentre si richiede che quello tra A e B sia minore di 0,25 per evitarne l'aggregazione e che quello tra B e C sia maggiore della stessa cifra. Questo caso non è di facile realizzazione e lo sarebbe ancora di meno ponendo un NT tra A e B. Più probabilmente i dati saranno tali da garantire soltanto l'aggregazione di A e C, cadendo di fatto nel secondo scenario di integrazione presentato. Tra i casi possibili uno di interesse è quello che prevede un NT tra C e B e di conseguenza una NT tra A e B. In questo caso si impone che l'affinità strutturale tra A e C sia inferiore a 0,5, altrimenti l'affinità di A con B e di B con C coinciderebbero a un valore non accettabile per affermare la separazione dei cluster di A e B. Il caso è rappresentato in figura.

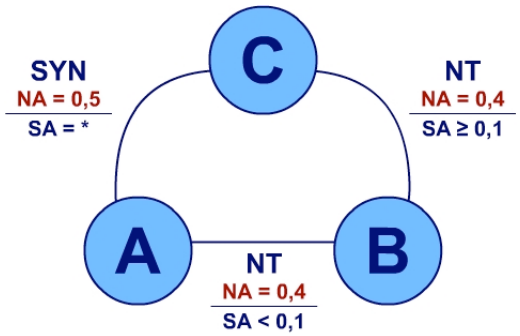

Figura 4.8: Caso 5, seconda rappresentazione

4.3.7 Caso 6: RT on NT tra A e C o tra B e C

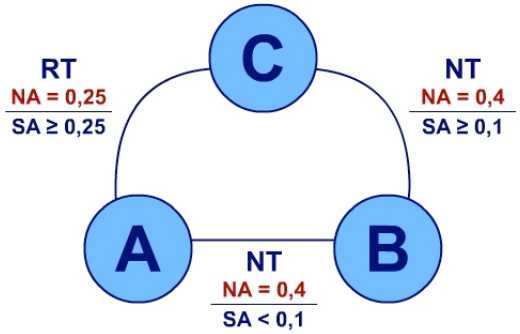

Figura 4.9: Caso 6

Si tratta di un caso analogo al precedente e con molte possibili espressioni. L'espressione trattata in figura può essere letta in modo simmetrico: tra A e C c'è una relazione RT e tra B e C c'è una relazione NT. Il caso specifico è realizzabile qualunque sia la relazione lessicale che lega A e B (eccetto SYN ovviamente), purché i coefficienti di affinità strutturali che legano le due classi con C siano compatibili con quelli di naming posti per ipotesi.

4.3.8 Caso 7: nessuna affinità lessicale tra A e B

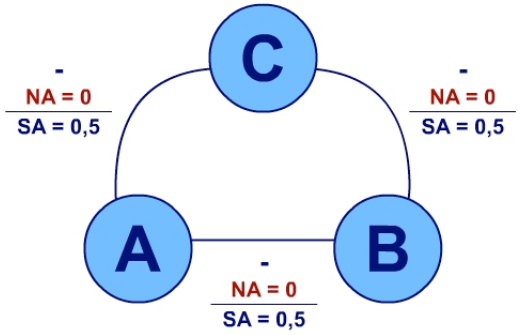

Figura 4.10: Caso 7

Il caso in cui non esista alcuna relazione lessicale tra le classi A e B non è realizzabile in tutte le sue occorrenze. Infatti perché non esista alcuna relazione tra A e B necessariamente non deve esistere alcuna relazioni tra A e C o tra B e C o entrambe. Il caso in cui non siano presenti relazioni semantiche è espresso in figura: come si può vedere per affermare l'integrazione senza l'ausilio di relazioni semantiche è necessario che l'affinità strutturale sia massima sia tra A e C che tra B e C. In altre parole si afferma la coincidenza strutturale. Essa causa la coincidenza strutturale tra A e B per la proprietà transitiva, negando la separazione delle due classi. Il caso in

cui esista affinità lessicale solamente tra due nomi, ad esempio di B e di C è presentato nella prossima figura.

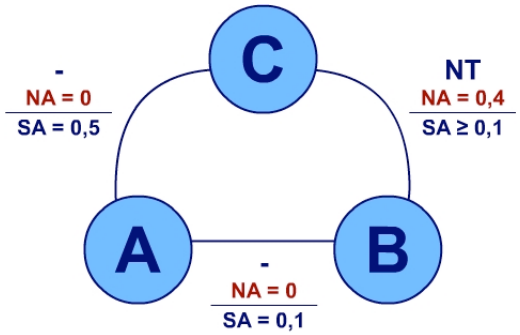

Figura 4.11: Caso 7, seconda rappresentazione

Il caso della relazione NT è solo uno dei tre possibili. La situazione è realizzabile in quanto l'affinità strutturale tra A e B determinata da quella che deve esistere tra A e C e tra B e C non è necessariamente in contrasto con la disgiunzione di A e B, qualunque sia la relazione lessicale che lega B e C.

4.3.9 Caso 8: nessuna affinità strutturale tra A e B

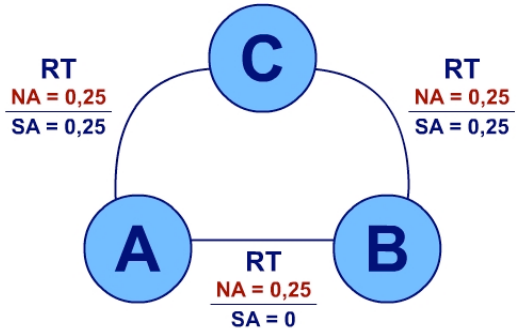

Figura 4.12: Caso 8

Anche questo caso è realizzabile sotto alcune condizioni. Se si impone che non vi sia alcuna affinità strutturale tra A e B, necessariamente si impone anche che l'affinità di A con C e B con C sia inferiore o pari 0,25, altrimenti l'ipotesi sarebbe negata da una sovrapposizione di attributi tra tutte e tre le classi. Un caso possibile (il più generale) prevede solo relazioni RT e coefficienti di affinità strutturale esattamente pari a 0,2.

4.3.10 Riepilogo

La seguente tabella riassume i casi presi in esame. E' possibile notare come soltanto i casi limite si sono verificati completamente irrealizzabili. Tutti gli altri lo sono, per quanto alcuni siano idealmente piuttosto rari.

Caso	Rel A-C	Rel B-C	Rel A-B	NA(A,B)	SA(A,B)	NA(A,C)	SA(A,C)	NA(B,C)	SA(B,C)	Possibile
1	-	-	*	*	0,5	0	0,5	0	0,5	NO
2	SYN	SYN	SYN	0,5	*	0,5	0	0,5	0	NO
3-a	RT	RT	RT	0,25	< 0,25	0,25	≥ 0,25	0,25	≥ 0,25	SI
3-b	RT	RT	NT/BT	0,4	< 0,1	0,25	≥ 0,25	0,25	≥ 0,25	SI
4-a	NT	NT	RT	0,32	< 0,25	0,4	≥ 0,1	0,4	≥ 0,1	SI
4-b	NT	NT	NT	0,4	< 0,1	0,4	≥ 0,1	0,4	≥ 0,1	SI
5-a	SYN	RT	RT	0,25	< 0,25	0,5	*	0,25	< 0,25	SI
5-b	SYN	NT	NT	0,4	< 0,1	0,5	*	0,4	≥ 0,1	SI
6	RT	NT	NT	0,4	< 0,1	0,25	≥ 0,25	0,4	≥ 0,1	SI
7-a	-	-	-	0	0,5	0	0,5	0	0,5	NO
7-b	-	NT	-	0	0,1	0	0,5	0,4	≥ 0,1	SI
8	RT	RT	RT	0,25	0	0,25	0,25	0,25	0,25	SI

Tabella 4.4

5 Conclusioni

5.1 Esperimento 1: la scelta dell'approccio

Gli esperimenti descritti hanno studiato l'occorrenza dei casi di integrazione previsti nell'utilizzo delle due tecniche (o approcci) proposte. Entrambi gli approcci impiegati si prestano a interventi di modificazione della GVV, anche se essi costituiscono pesanti variazioni, come quelle tipiche del terzo scenario di integrazione. Qualsiasi modifica alla struttura può essere operata seguendo indistintamente un approccio o un altro.

Nel corso dell'analisi dei risultati sono state evidenziate alcune differenze tra i due approcci che rendono ciascuno più adeguato a una situazione le cui caratteristiche prescindono dallo scenario di integrazione che deve essere affrontato. La scelta nell'utilizzo di un approccio o l'altro dipende prima battuta dallle preferenze progettista e in secondo luogo dalla varietà e complessità delle strutture da integrare.

E' stato evidenziato che il primo approccio può essere descritto come più dettagliato e preciso ma più oneroso in termini di calcolo. La maggiore quantità di relazioni che costringe il sistema a gestire rispetto al secondo approccio ne estende la complessità, tuttavia il controllo del progettista sul lavoro di integrazione è maggiore. Egli infatti può adattare l'annotazione di alcune sorgenti vecchie alla presenza di sorgenti nuove, pilotando in questo modo il sistema nell'integrazione. Se utilizzasse il secondo approccio potrebbe personalizzare soltanto le caratteristiche della vecchia GVV, senza scendere nel dettaglio delle sorgenti che la costituiscono. A questa maggiore libertà segue una maggiore pesantezza del procedimento sia per il sistema sia per il progettista.

Sfortunatamente l'unico modo per valutare l'efficienza effettiva di un approccio rispetto a un altro, basandosi sul numero di operazioni da eseguire, sarebbe eseguire il processo di integrazione nei due modi confrontando quantitativamente la complessità, come è stato fatto nel primo esperimento.

5.2 Esperimento 2: conclusioni sul terzo scenario

Il terzo scenario di integrazione, come anticipato nella sua descrizione, introduce una possibile inefficacia della GVV dall'occorrenza tutt'altro che rara. Ricreare una situazione realistica in questo ambito operativo è stato relativamente facile, a testimonianza di quanto asserito. Il sistema ha reagito secondo le aspettative a una situazione di integrazione generale come quella proposta, sconvolgendo come necessario la struttura della precedente GVV in un modo in cui eventuali applicativi basati su di essa avrebbero difficoltà a gestire. Questa reazione è tuttavia inevitabile, volendo mantenere un'annotazione più corretta possibile e lasciando gestire al sistema il processo di integrazione.

Lo studio generale delle occorrenze del terzo scenario ha evidenziato come solamente casi limite di difficile realizzazione siano effettivamente impossibili, mentre altri casi, rari e non, siano realizzabili in teoria. Questo fa pensare che, il una possibile diffusione, il sistema potrebbe incontrare numerose casistiche di questo tipo e ottenere i risultati riscontrati nella loro gestione.

Dall'analisi della situazione corrente emerge un'osservazione: è possibile che alcune delle casistiche trattate siano affrontate con maggiore efficienza come estensione del primo scenario.

Il progettista potrebbe intervenire sull'annotazione delle sorgenti o (in modo meno intuitivo) sulle relazioni del Thesaurus per far si che non vi sia effettivamente l'unione delle classi A e B, ma che esse siano mappate in cluster differenti, in uno dei quali è mappata anche la classe C. Questo tuttavia impone che sia il progettista ad applicare un criterio secondo il quale stabilire l'accorpamento. E' possibile suggerire una linea di riferimento per la scelta da operare. Un caso che può essere affrontato con questo tipo di collasso verso l'alto è il primo esempio del numero 5. Si tratta di una situazione in cui l'accorpamento di A e C è assicurato dalla relazione di SYN che lega i nomi delle classi, mentre le relazioni tra C e B sono decisamente più deboli. Il progettitsta, vedendo un caso di questo tipo, potrebbe scegliere di favorire gli accorpamenti causati da relazioni SYN ed eliminare gli altri mantenendo comunque un buon grado di precisione.

Un'altra ipotesi è quella di gestire il terzo scenario di integrazione come un'occorrenza del secondo. Rispetto al caso precedente, non potendo stabilire con esattezza (o convenienza progettuale) quale accorpamento sia più adeguato (se A+C o B+C), il progettista potrebbe scegliere di non accorpare le classi tra di loro e lasciare che esse siano divise in tre cluster. In termini di linee guida si può affermare che un'aderenza strutturale minore o uguale al 50% tra le tre classi, unita a una relazione RT tra i loro nomi può essere affrontata con un collasso verso il basso separando le tre classi e negando qualsiasi inclusione. Il caso più significativo di questo

evento è il numero 3 dove l'applicazione di questo metodo è a discrezione del progettista. La scelta può risultare conveniente nelle situazioni in cui le classi da affrontare sono molte e hanno un numero di attributi non troppo elevato.

Entrambe le ipotesi sacrificano parte della precisione dell'integrazione che altrimenti il sistema garantirebbe in modo automatico ma introducono modifiche alla GVV che non ne compromettono l'utilizzo. Gli interventi del progettista dovrebbero essere operati con cautela e seguendo il primo approccio di integrazione, più adeguato alla personalizzazione.

5.3 Osservazioni sul procedimento

Durante la sperimentazione sono emerse alcune caratteristiche del sistema che meritano ulteriori approfondimenti.

Nel paragrafo 3.2.1 è stata evidenziato il comportamento del sistema nel gestire le relazioni tra elementi appartenenti a domini di dato non compatibili. La reazione riscontrata è la non validazione di queste relazioni, che vengono ugualmente calcolate ma trascurate nella realizzazione dei cluster. Si osserva inoltre che la presenza di queste relazioni è in parte legata alla struttura delle sorgenti. E' infatti comune che in presenza di sorgenti semistrutturate (come quelle presentate nell'esempio) un attributo non abbia un tipo di dato semplice (o elementare, come *String*) ma sia associato a un tipo di dato complesso. La situazione è frequente e rappresenta casi in cui un attributo referenzia un'altra classe della stessa sorgente. Nel caso due attributi di questo tipo siano appartenenti a due classi analoghe di due differenti sorgenti la relazione di origine lessicale instaurata tra i due attributi verrebbe trascurata. Infatti ciascun attributo avrebbe un tipo di dato complesso e differente dall'altro, quindi apparterrebbe a un dominio incompatibile.

Malgrado esista una incompatibilità nei domini, è opportuno sottolineare che la relazione di origine lessicale tra i due attributi non perde valore espressivo a causa di questa incompatibilità, pertanto sarebbe opportuno considerarla nell'algoritmo di clustering.

Come soluzione al problema si propone di consentire al progettista di interagire sul risultato della validazione, in questo modo gli sarebbe possibile ricontrollare le relazioni che sono state scartate, rimediando all'errore commesso in questi casi).

Nello stesso paragrafo 3.2.1 è stata avanzata un'altra osservazione che riguarda le relazioni instaurate tra il nome di una classe e il nome di un attributo. L'apporto di queste relazioni deve essere chiarito con osservazioni approfondite.

Anzitutto, esse non contribuiscono al calcolo del coefficiente di affinità strutturale, dove sono considerate soltanto relazioni tra attributi. Tuttavia, nel calcolo del coefficiente di affinità legato al nome possono avere importanza. Nel calcolo dell'affinità tra i nomi di due classi viene scelto il valore affidato al percorso semantico più breve tra i due termini. Si considerino per esempio due classi *Class1* e *Class2*. La relazione tra i nomi delle due classi è di tipo RT, tuttavia la classe *Class1* ha un attributo *Att1.1* e la relazione tra il nome della classe e quello dell'attributo è NT. Si supponga che sia NT anche la relazione tra i nomi di *Att1.1* e *Class2*. Nel computo del Coefficiente di Affinità legato al nome il sistema confronterà il peso del percorso semantico diretto (RT = 0,5) con quello del percorso mediato da *Att1.1* (NT x NT = 0,8 x 0,8 = 0,64). E' immediato notare come le due relazioni classe-attributo di tipo NT aumentino l'affinità tra due nomi di classe. Si osserva che questo percorso potrebbe acquisire valore soltanto in situazioni in cui non esiste una relazione più forte tra i nomi delle due classi (RT in questo caso), pertanto il suo utilizzo è considerato piuttosto raro.

5.4 Prospettive di ricerca future

L'attività sperimentale è stata svolta su una versione non definitiva del tool di sviluppo Momis Ontology Builder, versione che comunque soddisfa tutti i requisiti necessari per lo volgimento delle attività descritte. Nel proseguo dello sviluppo del sistema si ritiene utile un approfondimento dell'attività sperimentale orientata all'analisi della dinamica dell'Ontologia.

Gli esperimenti svolti offrono infatti una rappresentazione semplificata delle casistiche espresse a livello teorico. Le conclusioni tratte sono conformi alle aspettative formulate in precedenza, tuttavia mancano di una formalizzazione matematica e logica. Un possibile approfondimento dell'attività potrebbe riguardare la ricerca di relazioni rigorose per avvalorare o negare le conclusioni di questo documento.

Sulla via sperimentale si ritiene utile un'attività di benchmarking del sistema e della sua attività di integrazione di nuove sorgenti senza cercare di ricreare casi specifici ma verificandone il comportamento in presenza di molte sorgenti con schemi di elevata complessità.

Parallelamente è opportuno verificare con precisione il peso dell'interazione del progettista nel corso della realizzazione dello schema integrato. Nel corso degli esperimenti si è considerata

una situazione di interazione minima e inferenza nulla. Nelle conclusioni si è invece sottolineata l'importanza dell'interazione per affrontare determinate situazioni in cui le reazioni automatiche del sistema potrebbero non essere completamente soddisfacenti. Si ritiene dunque utile approfondire lo studio dell'apporto del progettista per ottenere linee guida alla progettazione come è stato accennato nel paragrafo 5.2. Quesot studio può essere condotto parallelamente all'attività di benchmarking suggerita.

Bibliografia

[1] http://dbgroup.unimo.it/Momis/

[2] D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: *"Synthesizing an Integrated Ontology "*, IEEE Internet Computing Magazine, September-October 2003,42-51.

[3] F.Guerra: *"Dai dati all'informazione: il sistema MOMIS"*, 2002/2003.

[4] D. Calvanese, S. Castano, F. Guerra, D. Lembo, M. Melchiori, G. Terracina, D. Ursino, M. Vincini: *"Towards a comprehensive methodological framework for integration"*, 8th International Workshop on Knowledge Representation meets Databases.