Tecniche intelligenti di Mining Intensionale per la reingegnerizzazione di sistemi di basi di dati

Luigi Palopoli, Domenico Ursino
DEIS
Univ. della Calabria
87036 Rende (CS), Italy
email:{palopoli,ursino}@si.deis.unical.it
tel: +39-984-494752 +39-984-494691

Molte organizzazioni utilizzano sistemi informativi basati su modelli di dati non più in uso (quali i DBMS gerarchici o, persino, file di dati piatti) che devono essere riorganizzati secondo le metodologie e le tecniche attuali allo scopo di ottenere delle prestazioni ragionevoli o di procedere ad integrazioni multisorgente. Risulta pertanto utile estrarre le informazioni riguardanti la struttura dei sistemi informativi coinvolti al fine di consentirne un'appropriata riorganizzazione (reingeneering).

Ciò potrebbe aver luogo utilizzando tecniche varie. Ad esempio si potrebbe pensare di riprogettare ex-novo la base di dati di interesse. Tuttavia questo risulta spesso troppo dispendioso, ed in pratica improponibile, ad esempio nel caso in cui un numero rilevanti di basi di dati debbano essere riorganizzate in un sistema federato unitario. In particolare proprio il problema dell'integrazione di grandi insiemi di basi di dati preesistenti appare oggigiorno di grande rilevanza. Una tecnica promettente in questo caso consiste nella individuazione e/o derivazione di proprietà esistenti tra gli schemi delle varie basi di dati da utilizzare in un processo di integrazione/astrazione di tipo semiautomatico. In presenza di numerose basi di dati, tale processo di individuazione-derivazione di proprietà può essere vantaggiosamente realizzato applicando tecniche di mining e di inferenza automatica sviluppate ad hoc che indicheremo complessivamente con il termine di IKDD.

Molte organizzazioni possiedono già un numero elevato di basi di dati. Talora queste sono così numerose da risultare incontrollabili. Appare pertanto necessaria una riorganizzazione logica delle basi di dati in modo che le stesse vengano raggruppate in cluster, a ciascuno dei quali afferiranno basi di dati di contenuto omogeneo. A questo punto gli schemi delle basi di dati appartenenti a ciascun gruppo possono essere analizzati per rilevare le proprietà in comune tra elementi di schemi diversi (proprietà interschema); le proprietà interschema costituiscono la base del processo di integrazione che, una volta rilevate tali proprietà, provvede ad ottenere, a partire dai singoli schemi delle basi di dati, un unico schema (schema astratto) che le rappresenta tutte. Tuttavia anche i gruppi risultanti potrebbero essere numerosi e così si potrebbe procedere all'integrazione degli schemi astratti relativi ai cluster ottenendo degli schemi astratti al secondo livello. Procedendo ancora in questo modo si ottiene una struttura a più livelli denominata repository. Una delle organizzazioni che possiede tale strumento è la Pubblica Amministrazione Centrale, dove si è provveduto alla realizzazione del repository avvalendosi di tecniche di integrazione "manuale".

La nostra ricerca si propone di utilizzare l'IKDD per ottenere un sistema intelligente di tipo semi-automatico di supporto al processo di integrazione di basi di dati. Il sistema, dati gli schemi di partenza e interagendo con gli utenti, dovrebbe essere in grado di ricavare gli schemi astratti ai vari livelli del repository. La struttura che si otterrebbe sarebbe molto più flessibile per cui, nel momento in cui dovesse presentarsi un cambiamento, nella struttura o nel numero di basi di dati, la ristrutturazione del repository avverrebbe in maniera agevole.

Un aspetto peculiare della nostra ricerca consiste nell'utilizzo di una variante della Logica Descrittiva (un formalismo nato nell'ambito dell'Intelligenza Artificiale) per rappresentare proprietà interschema. Infatti, essendo questo formalismo adatto a rappresentare conoscenza su domini di struttura complessa, consente la rappresentazione di proprietà di entità e relazioni che, in genere, non sono rappresentabili mediante i modelli concettuali tradizionali. Essendo, inoltre, la Logica Descrittiva, un linguaggio formale caratterizzato da un ben definito sistema di inferenza è possibile effettuare ragionamenti complessi sull'insieme delle proprietá interschema. In particolare, definiamo e proponiamo l'utilizzo di un sottoinsieme della Logica Descrittiva arricchito di una componente probabilistica che sembra essere necessaria per essere in grado di ragionare su proprietá degli schemi di natura non esatta.

Il nostro approccio consiste in una serie di passi: innanzitutto utilizziamo una fase di preprocessing durante la quale vengono ricavate le sinonimie e le omonimie tra oggetti appartenenti ai vari schemi. Successivamente si provvede alla risoluzione dei conflitti di tipo.

Per la ricerca delle omonimie e delle sinonimie viene utilizzato un algoritmo che, dati due oggetti (entità, relazioni, attributi) appartenenti a schemi diversi, esamina gli oggetti ad essi correlati, i vincoli di correlazione e, facendo uso di un dizionario dei sinonimi di base, cerca di verificare se tra gli oggetti in esame vi sia qualche proprietà di sinonimia o di omonimia e di fornire una misura probabilistica della sua attendibilitá.

La risoluzione dei conflitti di tipo avviene considerando alcune proprietà degli oggetti in esame quali il nome, il tipo, il dominio e le dipendenze funzionali che vedono coinvolto l'oggetto in esame.

Dopo la fase di preprocessing si passa alla fase vera e propria di IKDD che prevede due stadi: il primo stadio, elementare, tenta di sfruttare le proprietà di sottoinsieme (così come è espressa nella nostra Logica Descrittiva con probabilitá). Il procedimento utilizza tre dizionari: il dizionario dei sinonimi e il dizionario delle negatività, ricavati durante la fase di ricerca dei sinonimi e degli omonimi nel passo precente, e il dizionario dei sottoinsiemi, fornito dagli esperti dei sistemi informativi coinvolti. In questa fase si tenta di ricavare nuove informazioni utili alla individuazione di proprietà interschema applicando alcuni assiomi della Logica Descrittiva che trattano in maniera opportuna gli indici statistici coinvolti nelle proprietà di partenza.

Un procedimento analogo, basato però su tecniche più sofisticate di data mining e su una maggiore correlazione con la Logica Descrittiva, è alla base del secondo stadio di IKDD, in cui si ricercano pattern complessi di correlazione tra schemi.

Il prodotto delle fasi di estrazione e inferenza subisce, infine, validazioni sia formali (ad esempio consistenza) che di contenuto. attraverso lo scambio di informazioni suppletive con esperti umani.

La nostra ricerca è attualmente in corso e si sta concentrando sullo sviluppo degli algoritmi di mining/inferenza di proprietà interschema. Obiettivo finale della ricerca è quello della progettazione e sviluppo di un tool semiautomatico, flessibile ed intelligente di ausilio all'integrazione di basi di dati federate di grandi dimensioni.