D2I

Integrazione, Warehousing e Mining di sorgenti eterogenee

Tema 1: Integrazione di dati provenienti da sorgenti eterogenee

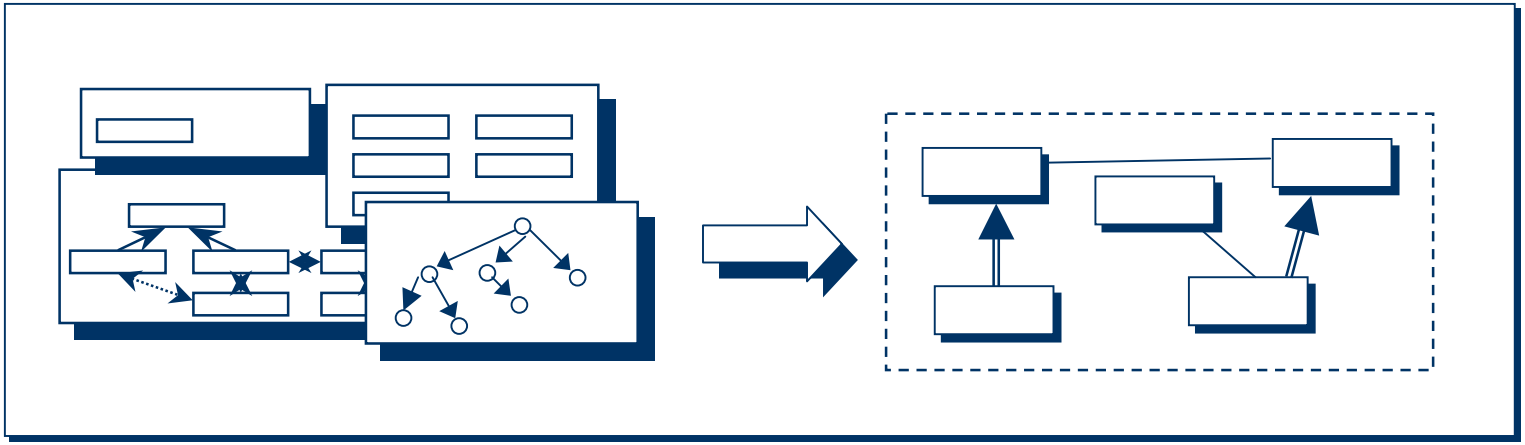
11 Marzo 2003 - Workshop D2I
MILANO – Centro Congressi Le Stelline

Metodi e strumenti per l'integrazione di sorgenti Informative eterogenee

http://www.dis.uniroma1.it/~lembo/D2I/Temi/tema1.html

Sonia Bergamaschi
Università di Modena e Reggio Emilia,
DII - Via Vignolese 905, 41100 Modena
http://www.dbgroup.unimo.it

Integrazione di dati provenienti da sorgenti eterogenee



L'integrazione fornisce uno schema concettuale globale (Vista Virtuale Integrata) sul quale l'utente può porre una interrogazione e ricevere una singola risposta unificata in maniera trasparente rispetto alle sorgenti coinvolte.

D2I

Integrazione di dati provenienti da sorgenti eterogenee

- Unita’ di Ricerca coinvolte:
  - Universita’ di Bologna (BO)
  - Universita’ della Calabria (CS)
  - Universita’ di Milano (MI)
  - Universita’ di Modena e Reggio Emilia (MO)
  - Universita’ di Roma "La Sapienza" (RM)

D2I

Integrazione di dati provenienti da sorgenti eterogenee

- Obiettivi del tema:
  - sviluppo di metodi e strumenti per l'integrazione di dati provenienti da sorgenti fortemente e strutturalmente eterogenee:
    - di tipo strutturato

      (ad es., database relazionali, ad oggetti, file)
    - di tipo semistrutturato

      (ad es., documenti XML)

Integrazione di dati provenienti da sorgenti eterogenee

- Scoperta di proprietà inter-schema, che permettono di caratterizzare le relazioni semantiche tra dati in diverse sorgenti
  - le informazioni semantiche sulle sorgenti sono spesso implicite
- Tecniche di risoluzione di query globali
  - occorre risolvere problemi sia per la suddivisione della query in sottoquery, sia per la ricostruzione della risposta.

D2I

Integrazione di dati provenienti da sorgenti eterogenee

- Principali Risultati attesi
  - definizione di una metodologia di integrazione di sorgenti fortemente eterogenee
  - definizione di tecniche di clustering basate su proprieta' di affinita' e corrispondenze semantiche
  - progetto di algoritmi per la riscrittura di interrogazioni su viste globali in termini di interrogazioni sulle sorgenti
  - definizione di metodi per la gestione di versioni diverse delle sorgenti
  - caratterizzazione di opportuni parametri di descrizione della qualità dei dati
  - studio di tecniche per la riconciliazione di dati provenienti da sorgenti diverse
  - progetto e realizzazione di un ambiente che supporti l'attivita' d'integrazione, basato sulla gestione di meta-dati

D2I

Integrazione di dati provenienti da sorgenti eterogenee

- Le Fasi di lavoro
  - Fase 1: analisi approfondita dello stato dell'arte con particolare riferimento al confronto dei modelli di dati semistrutturati proposti in letteratura e allo studio dei metodi esistenti per il query rewriting (durata 4 mesi - dal 1/12/2000 al 31/3/2001)
  - Fase 2: individuazione di un quadro metodologico per l'integrazione di dati provenienti da sorgenti fortemente eterogenee (durata 8 mesi - dal 1/4/2001 al 30/11/2001)
  - Fase 3: realizzazione di prototipi che implementino le funzioni evidenziate nella fase 2 (durata 8 mesi - dal 1/12/2001 al 31/7/2002)
  - Fase 4: verifica sperimentale dei prototipi sviluppati (durata 4 mesi - dal 1/8/2002 al 30/11/2002)

D2I

Fase 1
(Integrazione dei dati)

- Fase 1
  - studio e analisi di metodi e tecniche di estrazione, rappresentazione ed integrazione di sorgenti strutturate e semistrutturate (tutte le unita').
  - confronto tra i modelli per dati semistrutturati proposti in letteratura, allo scopo di individuarne il potere espressivo (tutte le unita').
  - formalizzazione della fase di estrazione delle relazioni interschema di tipo lessicale estratte attraverso l'interazione con WordNet (MO,MI)
  - individuazione dei requisiti per la scoperta e la rappresentazione di proprietà intra e inter-schema delle sorgenti, sia intensionali sia estensionali (MO,MI).

D2I

Fase 1
(Integrazione dei dati)

- metodi per definire parametri di qualità delle sorgenti (affidabilita', completezza, ridondanza, accuratezza, ecc.) e per la riconciliazione di dati provenienti da sorgenti eterogenee (tutte le unita').
- metodi e tecniche per la traduzione di informazioni da modelli di dati sorgente a modelli di dati target e introduzione del modello concettuale SDR-Network (CS)
- studio dell'impatto della presenza di diverse versioni dello schema di una sorgente: introduzione del modello CVM basato sulla logica descrittiva ALCQIO (BO)
- analisi del ruolo dei meta-dati e delle ontologie in un contesto in cui si integrano sorgenti strutturate e semi-strutturate (MO-MI)
- definizione della struttura del meta-data repository e dei metodi e delle strutture per ottenere la rappresentazione globale integrata e uniforme (tutte le unità)

D2I

Fase 1
(Query sulla Vista Virtuale Integrata)

- Definizione di approcci per specificare i mapping tra schemi delle sorgenti e schema globale (RM)
  - Local-as-view (LAV): le strutture delle sorgenti sono definite come viste sullo schema globale
  - Global-as-view (GAV): ogni concetto globale e' definito in termine di viste delle sorgenti locali
- analisi dei metodi esistenti per il query rewriting e del query answering using views (RM):
  - Query rewriting (LAV): riscrittura della query avviene in termini di viste e poi si valuta tale riscrittura
  - Query answering (LAV): si risponde direttamente alla query che e' basata su estensioni della vista globale
  - Unfolding (GAV) di ogni concetto globale posto nella query con la sua definizione in termini di sorgenti

D2I

Fase 1- Prodotti
(Integrazione e Query)

- D1.R1 Metodi e tecniche di estrazione, rappresentazione ed integrazione di sorgenti strutturate e semistrutturate (tutte le unita')
- D1.R2 Utilizzo di ontologie e proprieta' inter-schema di tipo estensionale (MO,MI)
- D1.R3 Metodi e tecniche per la traduzione di informazioni da modelli di dati sorgente a modelli di dati target (CS)
- D1.R4 Introduzione di un approccio formale per la gestione di versioni di schema in ambiente eterogeneo (BO)
- D1.R5 Rassegna sui metodi per query rewriting e il query answering using views (RM)

D2I

Fase 2
(Integrazione dei dati)

- Fase 2
  - definizione di una metodologia per la costruzione di viste riconciliate di dati semi-strutturati provenienti da sorgenti eterogenee, basata su:
    - tecniche intelligenti di tipo semi-automatico per l'identificazione e riconciliazione di eterogeneità basate su affinita' e clustering
    - estrazione semi-automatica di proprietà interschema
    - creazione di ontologie di dominio
    - in presenza di diverse versioni di schema, le tecniche impiegate sono basate su proprieta' inter-versione dedotte dalle modifiche di schema applicate (MI,BO,MO)
  - definizione dell'architettura funzionale di un prototipo che implementa la metodologia sviluppata (CS,MI,MO)
  - arricchimento della struttura del meta-data repository con riferimento alle interrogazioni globali e al loro mapping in interrogazioni locali alle sorgenti (tutte le unità).

D2I

Fase 2
(Query sulla Vista Virtuale Integrata)

- definizione di metodi e tecniche per il trattamento di interrogazioni formulate sulla vista integrata (CS,MI,MO)
- definizione di linguaggi fuzzy per l'interrogazione di viste riconciliate/sorgenti in cui pesare termini e filtrare le risposte in base alla rilevanza, tenendo conto della esistenza di sorgenti strutturate, semistrutturate e versionate (MI)
- definizione di algoritmi per la riscrittura di interrogazioni rispetto ad un insieme di viste (query rewriting e query answering using views), estendendo, modificando e adattando gli approcci attuali tenendo conto della esistenza di sorgenti semistrutturate (RM)

D2I

Fase 2
(Query sulla Vista Virtuale Integrata)

- algoritmi per la traduzione di informazioni da modelli di dati sorgente a modelli di dati target (CS)
- produzione delle specifiche funzionali di un "Query Manager" che supporti interrogazioni globali rispetto ad una vista virtuale integrata delle sorgenti. La decomposizione di una query globale in sub-query relative alle sorgenti e l'ottimizzazione della esecuzione delle sub-query e` ottenuta sulla base di un metodo unfolding (GAV) nel caso di “viste esatte”(MO).

D2I

Fase 2 – Prodotti
(Integrazione e Query)

- D1.R6 Descrizione della metodologia di integrazione di sorgenti fortemente eterogenee (MI,BO,MO)
- D1.R7 Architettura funzionale di un ambiente di ausilio al progettista per la costruzione di viste riconciliate di sorgenti fortemente eterogenee basato sulle tecniche sviluppate (CS,MI,MO)
- D1.R8 Specifiche funzionali del Query Manager (MO)
- D1.R9 Algoritmi per la traduzione di informazioni da modelli di dati sorgente a modelli di dati target (CS)
- D1.R10 Descrizione del linguaggio fuzzy per l' interrogazione di viste riconciliate (MI)
- D1.R11 Descrizione della metodologia e degli strumenti per la riconciliazione dei dati (RM)

D2I

Fase 3
(Integrazione dei dati)

- Fase 3
  - realizzazione di un prototipo di un ambiente di ausilio al progettista per la costruzione di viste riconciliate di sorgenti fortemente eterogenee basato sulla metodologia sviluppata. Il prototipo ingloberà anche un ambiente di ausilio alla costruzione della vista virtuale globale, con particolare riferimento
    - agli aspetti ontologici,
    - ai risultati di clustering interattivo basato su affinita'
    - alla conoscenza inter-schema di tipo sia intensionale che estensionale (MO).
  - realizzazione di un prototipo per l'estrazione di proprietà inter-sorgente esistenti tra oggetti rappresentati in sorgenti informative aventi formati eterogenei considerando: sinonimie, omonimie e similarità tra sotto-sorgenti (CS).

D2I

Fase 3
(Integrazione e Query)

- realizzazione di un prototipo per gli algoritmi di query rewriting e query answering using views e per la riconciliazione dei dati (RM).
- realizzazione di un prototipo di un query manager per la gestione di query globali. L'utente ha la possibilita' di interrogare una vista globale virtuale di un insieme di sorgenti eterogenee sottoponendo query SQL-like (MO).
- progettazione di un sistema di supporto alla realizzazione di sistemi per la gestione di versioni di schemi relativi a dati provenienti da sorgenti eterogenee (BO).
- realizzazione di prototipo per la traduzione di informazioni da sorgenti informative rappresentate attraverso modelli dei dati eterogenei in un modello dei dati uniforme denominato SDR-Network (CS)

D2I

Fase 3 – Prodotti
(Integrazione e Query)

- D1.P1 Prototipo di ambiente di ausilio al progettista per la costruzione di una vista globale basato su ontologie e assiomi inter-schema [SI-Designer - MOMIS] (MO)
- D1.P2 Prototipo per l'estrazione di proprietà inter-schema [SIPE] (CS)
- D1.P3 Prototipo per gli algoritmi di query rewriting e query answering using views e per la riconciliazione dei dati [IBIS] (RM)
- D1.P4 Prototipo di strumento per la manipolazione di versioni di schema in ambito eterogeneo [SVMgr](BO)
- D1.P5 Prototipo di un query manager per la gestione di query globali (MO) [Query Manager - MOMIS]
- D1.P6 Prototipo per la traduzione di informazioni da modelli di dati sorgente a modelli di dati target [SDR-TRAN] (CS)
- D1.P7 Prototipo di ambiente di ausilio al progettista per la costruzione di viste globali riconciliate basato su valutazione di affinità e clustering interattivo [ARTEMIS] (MI)

D2I

Fase 4

- Fase 4 (tutte le unita')
  - realizzazione e integrazione dei prototipi sviluppati conduzione di esperimenti per verificarne l'efficacia in problemi reali d'integrazione
- Fase 4 – Prodotti (tutte le unita')
  - D1.R12 Risultati della sperimentazione delle metodologie e dei prototipi per l'integrazione. In particolare:
  - La sperimentazione di DIKE (Universita' della Calabria) ha coinvolto tre differenti gruppi di sorgenti di dati, caratterizzati da (a) eterogeneita' nel formato, nella dimensione e nella struttura, (b) alto livello di eterogeneita' nei formati di rappresentazione, (c) dimensione rilevante delle sorgenti
  - L'Universita' di Milano ha sperimentato Artemis considerando quattro casi di integrazione in differenti domini composti da differenti generi di sorgenti di dati: relazionali, oggetti e XML

D2I

Fase 4

- Il prototipo dell' Universita' di Bologna SVMgr -- Schema Versioning Manager e' stato provato su differenti sorgenti
- L'Universita' di Modena e Reggio Emilia ha condotto la sperimentazione di MOMIS su un progetto reale confrontandone i risultati ottenuti con quelli dell'applicazione di una metodologia manuale di integrazione
- L'Universita' di Roma "La Sapienza" ha sperimentato il prototipo IBIS integrando differenti sorgenti di tipo relazionale, file system, sorgenti Web e sistemi legacy.

D2I

Pubblicazioni

http://www.dis.uniroma1.it/~lembo/D2I/Pubblicazioni/index.html

Sono state prodotte **numerose** e **prestigiose** pubblicazioni (85):

- articoli in riviste internazionali (22) e nazionali (2)
- capitoli in libri internazionali (6)
- atti di conferenze (25) e workshop (20) internazionali
- atti di conferenze nazionali (10)

Riviste internazionali:

- IEEE TKDE
- IEEE Intelligent System
- ACM –TODS
- SIGMOD RECORD
- Information Systems
- Data & Knowledge Engineering

Conferenze internazionali:

- VLDB
- DL
- ISWC
- PODS
- ICDE
- CooPIS
- IJCAI
- ER
- KRDB
- ICDT
- SEBD

D2I

Disseminazione dei risultati e riconoscimenti

- Le tecniche proposte per l'individuazione e l'estrazione automatica di proprietà interschema e la generazione automatica di viste riconciliate di dati ottenute dalle Università della Calabria, di Modena e Reggio Emilia e di Milano, sono state oggetto di un'analisi comparativa, che ha coinvolto sia aspetti metodologici che aspetti legati alle prestazioni, condotta da Jayant Madhavan (Università di Washington), Philip A. Bernstein (Microsoft Research), e Erhard Rahm (Università di Leipzig). Tale analisi è riportata in:

  - *Generic schema matching with Cupid* - J. Madhavan, P. A. Bernstein, and E. Rahm *In Proc. of the 27th International Conference on Very Large Databases, VLDB 2001, Roma, Italia, Settembre 2001*.