PROGETTO DI UNA UNITA' DI RICERCA - MODELLO B
Anno 2004 - prot. 2004095494_002

1.1 Tipologia del programma di ricerca

Interuniversitario

Aree scientifico disciplinari

Area 09: Ingegneria industriale e dell'informazione (%)

1.2 Durata del Programma di Ricerca

24 Mesi

1.3 Coordinatore Scientifico del Programma di Ricerca

BERGAMASCHI ***SONIA*** ***sonia.bergamaschi@unimo.it***

ING-INF/05 - Sistemi di elaborazione delle informazioni

Università degli Studi di MODENA e REGGIO EMILIA

Facoltà di INGEGNERIA

Dipartimento di INGEGNERIA DELL'INFORMAZIONE

1.4 Responsabile Scientifico dell'Unità di Ricerca

CIACCIA ***PAOLO***

Professore Ordinario ***16/04/1959*** ***CCCPLA59D16G489Q***

ING-INF/05 - Sistemi di elaborazione delle informazioni

Università degli Studi di BOLOGNA

Facoltà di INGEGNERIA

Dipartimento di ELETTRONICA, INFORMATICA E SISTEMISTICA

051/2093070	***051/2093540***	***pciaccia@deis.unibo.it***
(Prefisso e telefono)	*(Numero fax)*	*(Email)*

1.5 Curriculum scientifico del Responsabile Scientifico dell'Unità di Ricerca

Testo italiano
Paolo Ciaccia è nato il 16/04/1959 a Peschiera del Garda, VR, Italia. Nel 1985 si è laureato in Ingegneria Elettronica all'Università degli Studi di Bologna, ricevendo per la sua tesi il "Premio IBM Italia", e nel 1992 ha conseguito il titolo di Dottore di Ricerca in Ingegneria Elettronica ed Informatica. Dal 2000 è Professore Straordinario con afferenza al DEIS e allo IEIIT-CNR, dove coordina il MultiMedia Database Group. I suoi attuali interessi di ricerca includono il trattamento di interrogazioni di similarità e basate su

preferenze, lo studio di tecniche di Data Mining e la gestione di profili d'utente. È uno degli ideatori dell'M-tree, un indice per dati metrici utilizzato da molti gruppi di ricerca in ambito multimediale e di data mining nel mondo.
Ha partecipato a progetti di ricerca nazionali e internazionali, tra cui LTR HERMES (Foundations of High Performance Multimedia Information Management Systems). Nel 1997-98 è stato il coordinatore nazionale del progetto CNR MIDA su modellazione e interrogazione di basi di dati multimediali. Dal 2001 è il responsabile per il DEIS nella Thematic Network IST/FET PANDA (Patterns for Next-Generation Database Systems). È ricercatore principale in un progetto triennale finanziato dall'agenzia messicana CONACYT sull'indicizzazione di spazi metrici.
Ha pubblicato più di 60 lavori nei settori delle basi di dati, reti neurali, ingegneria del software e sistemi autonomi, in importanti riviste (tra cui IEEE TKDE, IEEE TSE, ACM TODS, ACM TOIS, Information Systems e Biological Cybernetics) e conferenze internazionali (tra cui VLDB, ACM-PODS, EDBT e ICDE).
Nel 1999 è stato "programme co-chair" del 1st International Workshop on Similarity Search e nel 2002 coordinatore del Comitato di Programma della conferenza italiana su basi di date evolute SEBD.
Dal 1999 è Associate Editor della rivista IEEE TKDE, per la quale è responsabile delle aree di strutture dati e algoritmi e di dati spaziali e multimediali.

Testo inglese
Paolo Ciaccia was born on April 16th 1959, in Peschiera del Garda, VR, Italy. In 1985 he got the "Laurea" degree in Electronic Engineering from the University of Bologna, Italy. His Laurea thesis was awarded the "IBM-Italy Prize". He received a PhD in Electronic and Computer Engineering (1992) from the same University. Since 2000 he has been a Full Professor at University of Bologna, with DEIS and the IEIIT institute of CNR, where he coordinates the activity of the MultiMedia Database Group. His current research interests include similarity and preference-based query processing, Data Mining techniques and user profile management. He is one of the designers of the M-tree, an index for metric data, which is used by several multimedia and data mining research groups in the world.
He participated several international and national research projects, among which ESPRIT IV LTR HERMES (Foundations of High Performance Multimedia Information Management Systems). In 1997-1998 he was the national coordinator of the Italian MIDA project on modeling and retrieval in multimedia databases. Since 2001 he has been the responsible for DEIS within the IST/FET Thematic Network PANDA (Patterns for Next-Generation Database Systems). He is major researcher in a 3 years project funded by the CONACYT mexican agency on the indexing of metric data.
He has published more than 60 papers in the areas of database systems, neural networks, software engineering, and autonomous systems, in major international journals (including IEEE TKDE, IEEE TSE, ACM TODS, ACM TOIS, Information Systems, and Biological Cybernetics) and international conferences (including VLDB, ACM-PODS, EDBT, and ICDE).
In 1999 he was programme co-chair of the 1st International Workshop on Similarity Search and in 2002 the PC chair of the SEBD italian conference on advanced data bases.
Since 1999 he has been Associate Editor of IEEE TKDE, responsible for the areas of Data Structures and Algorithms and Spatial and Multimedia Data.

1.6 Pubblicazioni scientifiche più significative del Responsabile Scientifico dell'Unità di Ricerca

1. *CIACCIA P.; PENZO W (2003).* ***The Collection Index to Support Complex Approximate Queries on XML Documents*** *1st VLDB XML Symposium (XSym 2003). pp. 164-180*
2. *CIACCIA P.; PATELLA M. (2002).* ***Searching in Metric Spaces with User-Defined and Approximate Distances*** *ACM TRANSACTIONS ON DATABASE SYSTEMS. (vol. 27(4) pp. 398-437)*
3. *CIACCIA P.; MONTESI D.; PENZO W.; TROMBETTA A. (2001).* ***Fuzzy Query Languages for Multimedia Data*** *In SYED M.R. Design and Management of Multimedia Information Systems: Opportunities and Challenges. pp. 201-212 ISBN: 1-930708-00-9 HERSHEY PA: Idea Group Publishing (UNITED STATES)*
4. *BARTOLINI I.; CIACCIA P.; WAAS F. (2001).* ***FeedbackBypass: A New Approach to Interactive Similarity Query Processing*** *27th International Conference on Very Large Data Bases (VLDB 2001). pp. 201-210*
5. *CIACCIA P.; PATELLA M.; ZEZULA P. (1997).* ***M-tree: An Efficient Access Method for Similarity Search in Metric Spaces*** *Proc. 23rd Int. Conf. on Very Large Data Bases (VLDB'97). pp. 426-435*

1.7 Risorse umane impegnabili nel Programma dell'Unità di Ricerca

1.7.1 Personale universitario dell'Università sede dell'Unità di Ricerca

Personale docente

nº	Cognome	Nome	Dipartimento	Qualifica	Settore Disc.	Mesi Uomo	
						1° anno	2° anno
1.	*CIACCIA*	*Paolo*	*Dip. ELETTRONICA, INFORMATICA E*	*Prof. Ordinario*	*ING-INF/05*	*6*	*6*

			SISTEMISTICA				
2.	*RIZZI*	*Stefano*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA*	*Prof. Associato*	*ING-INF/05*	*3*	*2*
3.	*GOLFARELLI*	*Matteo*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA*	*Ricercatore Universitario*	*ING-INF/05*	*2*	*2*
4.	*PATELLA*	*Marco*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA*	*Ricercatore Universitario*	*ING-INF/05*	*4*	*4*
5.	*PENZO*	*Wilma*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA*	*Ricercatore Universitario*	*ING-INF/05*	*4*	*4*
	TOTALE					**19**	**18**

Altro personale

nº	Cognome	Nome	Dipartimento	Qualifica	Mesi Uomo 1° anno	Mesi Uomo 2° anno
1.	*Bartolini*	*Ilaria*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA-DEIS*	*Dottore di Ricerca*	*2*	*2*
	TOTALE				**2**	**2**

1.7.2 Personale universitario di altre Università

Personale docente

Nessuno

Altro personale

Nessuno

1.7.3 Titolari di assegni di ricerca

Nessuno

1.7.4 Titolari di borse

nº	Cognome	Nome	Dipartimento	Anno di inizio borsa	Durata(in anni)	Tipologia	Mesi Uomo 1° anno	Mesi Uomo 2° anno
1.	*Baldacci*	*Lorenzo*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA-DEIS*	*2004*	*3*	*Dottorato*	*2*	*2*
2.	*Linari*	*Alessandro*	*Dip. ELETTRONICA, INFORMATICA E SISTEMISTICA-DEIS*	*2004*	*3*	*Dottorato*	*4*	*4*
	TOTALE						**6**	**6**

1.7.5 Personale a contratto da destinare a questo specifico programma

Qualifica	Costo previsto	Mesi Uomo 1° anno	Mesi Uomo 2° anno	Note
Altre tipologie	*9.000*		*6*	*laureato con contratto a termine*
Altre tipologie	*9.000*		*6*	*laureato con contratto a termine*
TOTALE	**18.000**	**0**	**12**	

1.7.6 Personale extrauniversitario indipendente o dipendente da altri Enti

Nessuno

2.1 Titolo specifico del programma svolto dall'Unità di Ricerca

Testo italiano
Elaborazione di interrogazioni distribuite basata su ontologie di dominio e profili di sorgenti

Testo inglese
Distributed query processing based on domain ontologies and source profiles

2.2 Settori scientifico-disciplinari interessati dal Programma di Ricerca

ING-INF/05 - Sistemi di elaborazione delle informazioni

2.3 Parole chiave

Testo italiano
ONTOLOGIA DI DOMINIO ; ELABORAZIONE DI INTERROGAZIONI ; SELEZIONE DI SORGENTI ; PROFILI DI SORGENTI ; SISTEMI INFORMATIVI DISTRIBUITI

Testo inglese
DOMAIN ONTOLOGY ; QUERY PROCESSING ; SOURCE SELECTION ; SOURCE PROFILING ; DISTRIBUTED INFORMATION SYSTEMS

2.4 Base di partenza scientifica nazionale o internazionale

Testo italiano
Nell'ambito del progetto l'attività dell'Unità di Bologna si concentra su problematiche proprie dei Temi 1 (Creazione ed estensione di un'ontologia di dominio) e 3 (Elaborazione di interrogazioni) e conseguentemente su aspetti inerenti tecniche di descrizione delle sorgenti e di ricerca intelligente basate su ontologie e mapping semantici.

Il problema della selezione delle sorgenti più "rilevanti" per una data interrogazione è fondamentale per la risoluzione efficiente della stessa. Ad esempio, due sorgenti possono fornire lo stesso tipo di informazioni, ma per la prima si può fare affidamento su dei tempi di risposta più brevi. In alternativa, una sorgente può mettere a disposizione informazioni attinenti la richiesta specifica, ma non esattamente coincidenti con i requisiti della stessa. In ogni caso ciò che risulta necessario è una caratterizzazione (strutturale, semantica e statistica) di ciascuna sorgente. Mentre per i primi due aspetti si può fare affidamento, rispettivamente, sulle descrizioni fornite da wrapper e da ontologie di dominio, la caratterizzazione statistica necessita di informazioni riguardanti il contenuto della sorgente in termini di dati (istanze) gestiti.
Ovviamente, dal momento che risulta impossibile caratterizzare in maniera esatta tale contenuto, la rilevanza può solamente essere stimata mediante indicatori statistici preventivamente estratti dalla sorgente. Le soluzioni attualmente esistenti in letteratura (Gravano, 1999; Ipeirotis, 2002; Gravano 2003) si basano essenzialmente sull'estrazione di un insieme di parole chiave con associate frequenze di occorrenza, e quindi non sono in grado di tenere conto delle relazioni semantiche esistenti tra i termini (concetti e valori) presenti nell'interrogazione e quelli propri della sorgente (Ganesan, 2003).
L'utilizzo di ontologie per stabilire la rilevanza delle sorgenti può, inoltre, permettere la gestione di richieste riguardanti concetti che si trovano a diversi livelli della gerarchia di astrazione delle ontologie stesse: ad esempio, un sito che contiene informazioni riguardanti la pallacanestro dovrebbe essere giudicato come rilevante per un'interrogazione che contenga il termine "NBA", in quanto quest'ultimo è un concetto più specifico rispetto al termine "pallacanestro" (Ipeirotis, 2002).

Per quanto riguarda il recupero dei risultati per l'interrogazione formulata dall'utente, per motivi di efficienza non è pensabile un semplice modello basato prima sul recupero dalle sorgenti selezionate di tutti gli oggetti che, sia pur minimamente, possono risultare rilevanti e poi sulla determinazione di un sottoinsieme, tipicamente di dimensione limitata, contenente gli oggetti che meglio soddisfano la richiesta. Si rende pertanto necessario mettere a punto meccanismi di esecuzione che limitino al minimo le risorse necessarie e che, al contempo, garantiscano la correttezza del risultato, tenendo in considerazione sia la rilevanza delle sorgenti rispetto all'interrogazione, che le modalità di accesso alle sorgenti stesse (Fagin, 2001). In particolare, è necessario conoscere quali sorgenti sono in grado di fornire i propri risultati in ordine di rilevanza rispetto all'interrogazione fornita e come il mapping tra la richiesta originale dell'utente e quella effettivamente inviata alla sorgente, e quindi basata su ontologia locale, influenza la rilevanza dei risultati. Lo scenario generale, che include anche aspetti più propriamente legati ai vincoli architetturali (ad es., tempi di risposta delle sorgenti), prevede pertanto che, per produrre il risultato di una interrogazione, si debbano andare a "pesare" opportunamente diversi aspetti che incidono sulla rilevanza dei risultati (aspetti relativi a sorgenti, alle singole istanze, alla completezza rispetto a quanto richiesto, all'incertezza derivante da atttività di "object fusion", ecc.). In tal senso, in letteratura il problema è stato affrontato per il caso di singola sorgente strutturata (Ciaccia, 2000; Bruno, 2002) o di più sorgenti strutturate, accessibili via Web (Bruno, 2004). Il caso di sorgenti eterogenee, come sono quelle considerate da WISDOM, e per le quali l'accesso è mediato da un'ontologia di dominio locale, non è stato tuttavia ancora considerato. In tale scenario non esiste inoltre nessuno studio relativo all'impatto di criteri di integrazione "qualitativi", che generalizzano quelli puramente numerici e che si sono già dimostrati più efficaci in diversi ambiti (Chomicki, 2002; Torlone, 2002; Bartolini, 2004).

Al fine di restituire all'utente un risultato significativo e facilmente fruibile, assume notevole importanza la possibilità di navigare e sintetizzare efficacemente i dati ottenuti dall'interrogazione. A questo proposito, gli approcci presenti in letteratura sono inquadrabili in tre distinti filoni. Da un lato, nell'ambito della business intelligence e dei database multidimensionali sono state studiate tecniche per la fruizione del contenuto informativo a differenti livelli di aggregazione attraverso l'applicazione di operatori OLAP (Gyssens, 1997). In questi casi l'informazione di interesse è prevalentemente numerica, differentemente da quanto richiesto in WISDOM, e il procedimento di sintesi si riduce al calcolo di indicatori statistici finalizzato alla definizione di cruscotti decisionali per il manager aziendale; la possibilità di navigare interattivamente il contenuto informativo è garantita da front-end sofisticati che implementano primitive di roll-up, drill-down, slice-and-dice. Dall'altro lato, nell'ambito data mining e knowledge extraction, sono state studiate tecniche per la modellazione di pattern intesi come rappresentazioni, compatte e semanticamente ricche, di un insieme di dati. Alcuni lavori sono specificamente orientati alla rappresentazione di pattern tipici del data mining, quali regole associative e cluster (Imielinski, 1996); in altri approcci, l'enfasi è invece sulla derivazione di modelli general-purpose, estendibili e riusabili, per la rappresentazione di pattern (Rizzi, 2003). I modelli proposti da questi lavori non sono però applicabili al caso di WISDOM poiché la sintesi dell'informazione è pre-cablata dal progettista invece di essere guidata da ontologie. Infine, nell'ambito delle interfacce avanzate sono stati messi a punto linguaggi visuali che permettono la formulazione di interrogazioni su basi di dati navigando schemi concettuali definiti dal progettista (Benzi, 1999).

Testo inglese

Within the WISDOM project the University of Bologna Unit will work on issues related to Theme 1 (Building and extending a domain ontology) and Theme 3 (Query processing) and consequently on aspects related to techniques for describing the data sources and for intelligent querying based on ontologies and semantic mappings.

Selecting the most "relevant" data sources for a given query is a fundamental prerequiste for efficiently processing the query. For example, two data sources may contain the same information but the first one could guarantee a faster response time. Alternatively, a data source could contain some information relevant to the problem but not exactly coinciding with the query requirements. In all these cases a (structural, semantic, and statistical) characterization is necessary for each data source. While for the first two aspects we can rely, respectively, on the descriptions obtained from the wrappers and on the domain ontologies, statistical characterization requires information concerning the data instances. Obviously, since an exact content description is not feasible, the relevance of a source has to be estimated by means of statistical indicators obtained from the data source. The approaches available in the literature (Gravano, 1999; Ipeirotis, 2002; Gravano 2003) are based on the extraction of a set of keywords and their frequencies, thus they are not capable of keeping into account the semantic relationships between terms (concepts and values) present in a query and those that characterize the data source (Ganesan, 2003). When domain ontologies are used to determine data source relevance it is also possible to handle requests involving concepts at different aggregation levels: For example, a source containing basketball information should be rated as relevant for a query containing the term "NBA" , since the latter is a more specific term than "basketball" (Ipeirotis, 2002).

The actual retrieval of the results of a user query is another major research topic. Indeed, due to efficiency constraints it is not possible to adopt simple models that (1) first retrieve all the objects in the selected data sources that are in some way, even marginally, relevant to the query, and (2) then determine the subset, usually of limited cardinality, of objects that best satisfies the query. Rather, execution mechanisms have to be developed that ensure the result correctness and, at the same time, require the minimum amount of resources. Such execution techniques have to properly take into account both the relevance of data sources and their access modalities (Fagin, 2001). In particular, it is necessary to know which data sources are capable of ordering the results based on their relevance to the user query and how the result relevance is influenced by the mapping between the original query and the one sent to the data source, the latter being determined by the local ontology of the source. The global scenario, that may also include aspects related to architectural constraints (e.g., data sources response times), thus requires to properly "weigh" and combine all the different aspects that impact on the relevance of results (e.g., aspects related to data sources, instances, completeness with respect to the desired answer, uncertainty deriving from "object fusion", etc.). This issue has been addressed in the literature both in the case of a single structured data source (Ciaccia, 2000; Bruno, 2002) and in the case of multiple structured data sources accessible on the Web (Bruno, 2004). However, the case of multiple heterogeneous data sources in which, as in WISDOM, the access is mediated by a local domain ontology has not been considered yet. Further, in this scenario no previous work has focused on the impact of "qualitative" integration criteria. These criteria have already proved to be more effective than those based only on numerical scores in several application domains (Chomicki, 2002; Torlone, 2002; Bartolini, 2004).

Returning the user a significant and easy to use result requires to effectively navigate and summarize data obtained from the query. The approaches proposed in the literature can be classified in three distinct groups. On the one hand, in the business intelligence and in the multidimensional database field many techniques have been studied that allow the data to be analyzed at different aggregation levels using OLAP operators (Gyssens, 1997). In these cases, differently from WISDOM, information is mainly numeric and the summarization process can be seen as the computation of simple statistical indicators oriented to define a set of decision-making dashboards for the managers; the ability of interactively navigating the contents is ensured by advanced front-ends that implement roll-up, drill-down, and slice-and-dice operators.
On the other hand, research on data mining and knowledge extraction has studied many techniques for modeling patterns meant as compact and semantically rich representations of raw data. Some works focus on the representation of typical data mining patterns as association rules and clusters (Imielinski, 1996); other approaches emphasize how to derive general-purpose models characterized by extensibility and reusability (Rizzi, 2003). The models proposed in such works cannot be applied to the WISDOM scenario since the way they summarize information is pre-defined by the designer and cannot be driven by using domain ontologies. Finally, as concern smart front-ends, several visual languages have been defined that allow query definition on pre-defined conceptual schemata (Benzi, 1999).

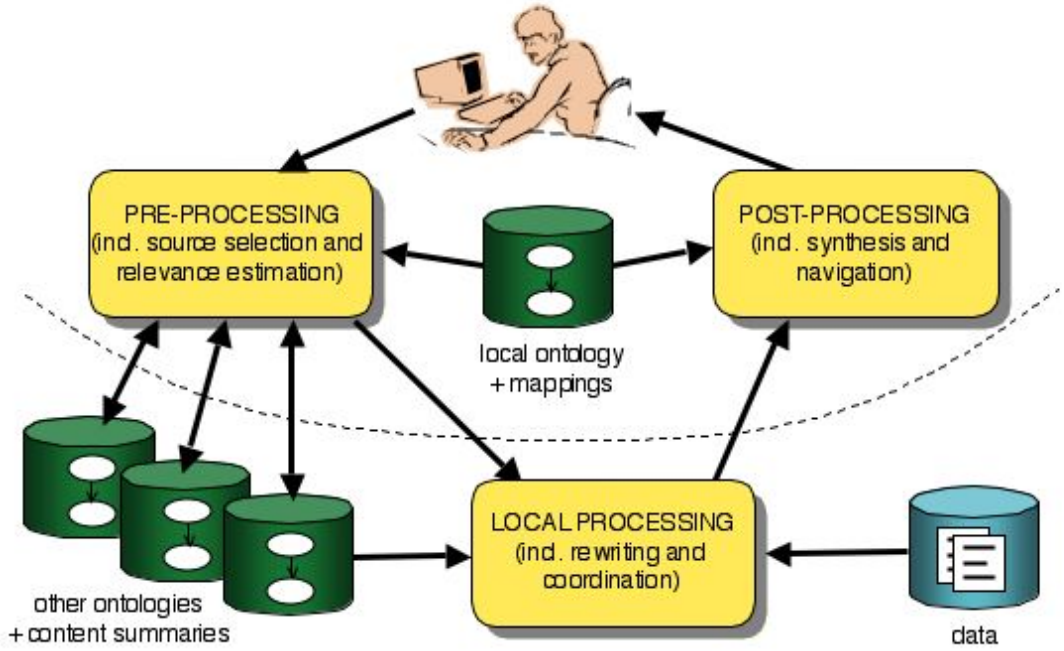
2.4.a Riferimenti bibliografici

(Bartolini, 2004) I. Bartolini, P. Ciaccia, V. Oria: Integrating the Results of Multimedia Sub-Queries Using Qualitative Preferences. Technical Report IEIIT-BO, 2004.
(Benzi, 1999) F. Benzi, D. Maio, S. Rizzi: Visionary: a viewpoint-based visual language for querying relational databases. Jour. Visual Languages and Computing 10(2): 117-145 (1999).
(Bruno, 2002) Nicolas Bruno, Surajit Chaudhuri, Luis Gravano: Top-k selection queries over relational databases: Mapping strategies and performance evaluation. ACM TODS 27(2): 153-187 (2002).
(Bruno, 2004) Nicolas Bruno, Luis Gravano, Amélie Marian: Evaluating Top-k Queries over Web-Accessible Databases. To appear in ACM TODS (2004).
(Callan, 1995) James P. Callan, Zhihong Lu, W. Bruce Croft: Searching Distributed Collections with Inference Networks. SIGIR 1995: 21-28.
(Chomicki, 2002) Jan Chomicki: Querying with Intrinsic Preferences. EDBT 2002: 34-51
(Ciaccia, 2000) P. Ciaccia, D. Montesi, W. Penzo, and A. Trombetta: Imprecision and User Preferences in Multimedia Queries: A Generic Algebraic Approach. FoIKS 2000: 50-71
(Fagin, 2001) Ronald Fagin, Amnon Lotem, Moni Naor: Optimal Aggregation Algorithms for Middleware. PODS 2001: 102-113
(Ganesan, 2003) Prasanna Ganesan, Hector Garcia-Molina, Jennifer Widom: Exploiting hierarchical domain structure to compute similarity. ACM TOIS 21(1): 64-93 (2003).
(Gravano, 1999) Luis Gravano, Hector Garcia-Molina, Anthony Tomasic: GlOSS: Text-Source Discovery over the Internet. ACM TODS 24(2): 229-264 (1999).
(Gravano, 2003) Luis Gravano, Panagiotis G. Ipeirotis, Mehran Sahami: QProber: A system for automatic classification of hidden-Web databases. ACM TOIS 21(1): 1-41 (2003).
(Gyssens, 1997) M. Gyssens, L.V .S. Lakshmanan: A Foundation for Multi-Dimensional Databases. VLDB 1997: 106-115.
(Imielinski, 1996) T. Imielinski, H. Mannila: A Database Perspective on Knowledge Discovery, CACM 39(11):58-64 (1996).
(Ipeirotis, 2002) Panagiotis G. Ipeirotis, Luis Gravano: Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection. VLDB 2002: 394-405.
(Rizzi, 2003) S. Rizzi et al.: Towards a logical model for patterns. ER 2003: 77-90.
(Torlone, 2002) Riccardo Torlone, Paolo Ciaccia: Which Are My Preferred Items? Workshop on Recommendation and Personalization in E-Commerce (RPEC'02).

2.5 Descrizione del programma e dei compiti dell'Unità di Ricerca

Testo italiano
Gli argomenti che l'Unità di Bologna tratterà nell'ambito del progetto sono relativi ai Temi 1 e 3 e così sintetizzabili:
Tema 1:
- creazione di "content summaries" per le sorgenti informative (Content summaries)
Tema 3:
- esecuzione di interrogazioni distribuite in WISDOM (Esecuzione)
- fruizione e navigazione basata su ontologia dei risultati di interrogazioni (Navigazione)

Lo scenario rilevante per le attività di ricerca dell'Unità è concisamente riassunto in figura



In particolare, l'attività relativa ai "Content summaries" ha lo scopo di fornire una caratterizzazione ("profilo") delle sorgenti dal punto di vista statistico che permetta una più precisa valutazione della rilevanza delle sorgenti stesse relativamente a una data interrogazione e, conseguentemente, la selezione delle sorgenti più significative. L'attività di "Esecuzione" include aspetti di esecuzione distribuita dell'interrogazione sulle diverse sorgenti e il coordinamento/sincronizzazione di tali esecuzioni, al fine di determinare, con il minimo dispendio di risorse, i risultati ritenuti più rilevanti. Infine, l'attività di "Navigazione", che opera a valle del processo di elaborazione vero e proprio, ha lo scopo di fornire meccanismi per poter fruire dei risultati in forma compatta e flessibile, considerando i livelli di astrazione offerti dalla specifica ontologia di dominio.

La ricerca su tali argomenti, rispettando le fasi previste dal progetto, si articolerà come segue:

PRIMA FASE (6 MESI)
Per tutti i 3 argomenti la prima fase del progetto sarà innanzitutto dedicata all'analisi critica dello stato dell'arte, allo scopo di definire compiutamente i limiti delle soluzioni attualmente disponibili per i problemi di interesse. Si procederà quindi alla formulazione dei requisiti specifici per i 3 argomenti di ricerca. Nello specifico:
(Content summaries) La fase di analisi verterà principalmente sulla possibilità di estendere le tecniche oggi disponibili per la costruzione di profili al caso di sorgenti descritte da un'ontologia di dominio. In particolare si definiranno compiutamente i requisti che i content summaries devono soddisfare al fine di poter essere efficacemente usati per determinare la rilevanza di una sorgente rispetto a un'interrogazione.
(Esecuzione) Partendo da un'analisi delle principali tipologie di elaborazione in ambiente distribuito ed eterogeneo, si definiranno compiutamente i limiti delle stesse in relazione all'architettura di WISDOM (nella quale, si ricorda, una sorgente è visibile esternamente solo attraverso l'ontologia di dominio che la integra). In particolare, considerando i vari aspetti che possono contribuire a determinare la rilevanza di un risultato, si analizzerà se e come tali aspetti sono influenzati dall'architettura di WISDOM.
(Navigazione) Si analizzerà la possibilità di elaborare i dati restituiti dalle interrogazioni al fine di presentarli all'utente in forma compatta e facilmente fruibile. Si valuterà quindi in che misura sia opportuno abbinare tecniche di navigazione e sintesi proprie della business intelligence a forme di rappresentazione di pattern proprie del data mining con l'obiettivo di garantire massima flessibilità nella specifica del livello di granularità a cui analizzare i dati. Inoltre, si studierà in che modo estendere i paradigmi sviluppati per l'interrogazione visuale di basi di dati al caso di interrogazione di insiemi di documenti, tenendo in considerazione le peculiarità derivanti dall'utilizzo di ontologie.

Infine si lavorerà, congiuntamente a tutte le altre Unità, alla definizione dell'architettura metodologica e funzionale di riferimento per l'intero progetto (prodotto D0.R1).

PRODOTTI DELLA FASE 1
I prodotti attesi in questa fase del progetto sono di tipo rapporto tecnico (sigla R). La sigla dopo la D indica il tema (0 se il prodotto è comune a tutti i temi).

D0.R1 Rapporto sull'architettura metodologica e funzionale di riferimento (in collaborazione con Modena e Reggio Emilia - MO, Roma - RM, Trento - TN)
D1.R1 Analisi critica dei linguaggi e standard emergenti per le ontologie (in collaborazione con MO, RM, TN)
D3.R1 Analisi critica di linguaggi di interrogazione e tecniche di riscrittura basati su ontologie (in collaborazione con MO, TN)
D3.R2 Analisi critica delle tecniche di esecuzione di interrogazioni in ambiente eterogeneo

SECONDA FASE (6 MESI)
Nella seconda fase del progetto verranno messe a punto le soluzioni di base per i 3 argomenti specifici trattati dall'Unità:
(Content summaries) Nella seconda fase verranno definiti i meccanismi attraverso i quali ogni ontologia di dominio può essere corredata da informazioni di natura quantitativa. L'idea di base è estendere le tecniche esistenti di "probing" (interrogazione) delle sorgenti considerando le informazioni di natura ontologica, e i vincoli che da tali informazioni sono desumibili. Le soluzioni che si deriveranno saranno ispirate a principi di economia, nel senso di: 1) richiedere il minor numero di "probes", e 2) fornire le informazioni quantitative più significative, a parità di memoria richiesta per la memorizzazione dei "content summaries".
In linea con gli obiettivi del Tema 1, verranno inoltre dettagliate le modalità di aggiornamento dei "content summaries" a fronte dell'introduzine di una nuova sorgente e all'estensione della relativa ontologia di dominio.
(Esecuzione) Scopo di questa fase è pervenire alla definizione di tecniche per l'esecuzione di interrogazioni distribuite che, considerando i limiti imposti dall'architettura di WISDOM, siano in grado di restituire i risultati ritenuti più rilevanti minimizzando le risorse necessarie. Poiché la rilevanza di un oggetto può dipendere da vari fattori e dal modo come tali fattori sono tra loro combinati, le tecniche che si svilupperanno saranno di tipo generalizzato, ovvero in grado di funzionare correttamente ed efficientemente anche al variare del criterio di combinazione. Per tale criterio, che nel caso base può ridursi ad una somma pesata dei vari fattori, verrà anche considerato il caso più generale di tipo "qualitativo", ovvero definito non necessariamente mediante tecniche numeriche.
(Navigazione) Relativamente alla fruizione dei risultati, verranno messi a punto metodi per permettere all'utente di specificare in maniera puntuale il livello di risoluzione desiderato. In particolare, si definiranno tecniche per rappresentare l'informazione in forma compatta e ricca di semantica a differente livelli di astrazione, e si individueranno operatori per la navigazione interattiva dell'informazione sui vari livelli in accordo con l'ontologia di dominio.
Infine si lavorerà, congiuntamente a tutte le altre Unità, alla definizione delle interfacce dei componenti per il prototipo integrato (prodotto D0.R2).

PRODOTTI DELLA FASE 2
D0.R2 Specifiche delle interfacce dei componenti del prototipo integrato (in collaborazione con MO, RM, TN)
D1.R2 Definizione del linguaggio per la specifica di una ontologia di dominio (in collaborazione con MO, TN)
D1.R3 Definizione di tecniche per la creazione di "content summaries"
D3.R3 Definizione del linguaggio di interrogazione e delle tecniche di riscrittura basate su ontologie (in collaborazione con MO, TN)
D3.R4 Definizione di tecniche per l'esecuzione di interrogazioni in WISDOM

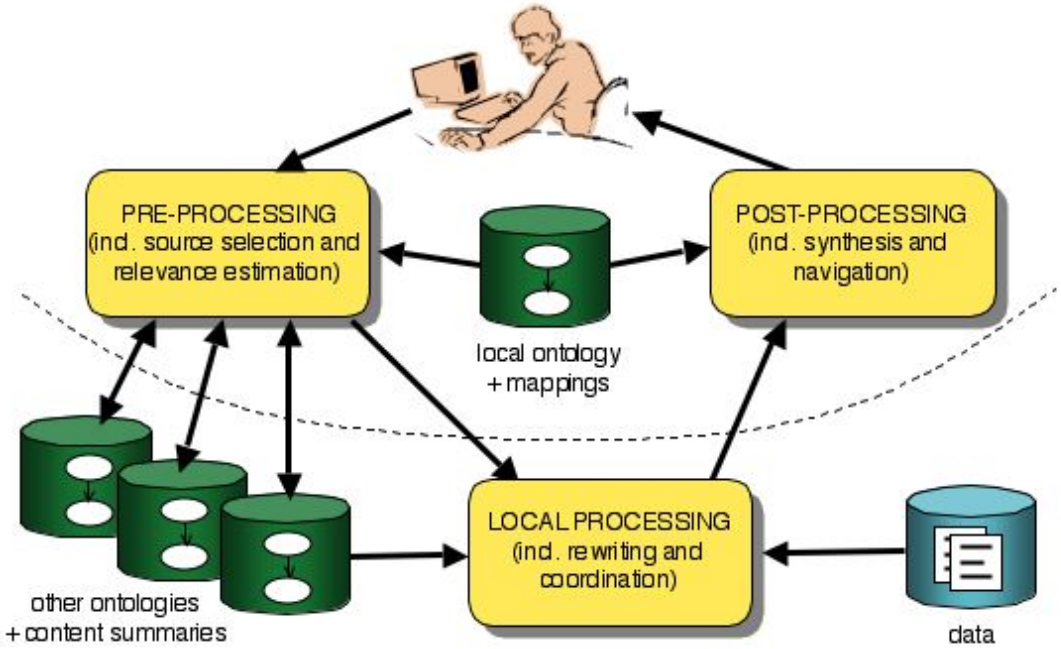
TERZA FASE (12 MESI)
Nella terza fase del progetto verranno sviluppati 3 prototipi, e si collaborerà con le altre Unità all'integrazione dei prototipi realizzati durante il progetto.
Il primo prototipo, a partire da un'ontologia di dominio preesistente, implementerà tecniche di "probing" delle sorgenti relative e di produzione dei relativi "content summaries" a partire dai risultati ottenuti
Il secondo prototipo, realizzato in collaborazione con l'Unità di Modena e Reggio Emilia (MO), si farà carico dell'acquisizione e analisi delle interrogazioni, oltre che della determinazione delle sorgenti rilevanti per l'interrogazione stessa.
Il terzo prototipo implementerà le tecniche di esecuzione di interrogazione messe a punto durante la fase 2, e includerà un'interfaccia per la navigazione interattiva dell'informazione a diversi livelli di astrazione sulla base dell'ontologia di dominio.
Per tutti i prototipi verrà condotta un'intensa attività sperimentale volta a verificare gli aspetti di natura prestazionale.

PRODOTTI DELLA FASE 3
I prodotti attesi in questa fase del progetto sono sia di tipo rapporto tecnico (sigla R) che di tipo prototipo software (sigla P).
D0.P1 Prototipo integrato di sistema (in collaborazione con MO, RM, TN)
D1.P2 Prototipo per la creazione di "content summaries"
D3.P1 Prototipo per la formulazione di interrogazioni (in collaborazione con MO)
D3.P2 Prototipo per l'esecuzione di interrogazioni distribuite in WISDOM

Testo inglese
The topics the Bologna Unit will deal with are part of the resarch Themes 1 and 3, and can be summarized as follows:

Theme 1:
- "Content summaries" creation for information sources (Content summaries)
Theme 3:
- Distributed query execution in WISDOM (Execution)
- Usage and navigation, based on ontologies, of the query results (Navigation)

The scenario for the research activities is summarized in the following figure:



In particular, the task concerning "Content summaries" aims to deliver a characterization ("profile") of the data sources from the statistical point of view in order to accurately evaluate their relevance with respect to a given query and, consequently, to allow a smart selection of the most relevant data sources. The "Execution" task deals with aspects related to the distributed execution of queries on different data sources and their coordination/synchronization, so as to determine, with a minimal amount of resources, the most relevant results. Finally, the "Navigation" task, which takes place after query processing, is aimed at defining mechanisms for exploiting the results in a synthetic and flexible representation by relying on the multiple abstraction levels available with a specific domain ontology.

Research on such topics, given the project structure, will be organized as follows:

FIRST PHASE (6 MONTHS)
The first phase will be devoted to accurately define the requirements for the 3 research topics, then we will analyze and criticize the related literature in order to identify the limits of the available solutions with respect to our current goals. In detail:
(Content summaries) The analysis phase will survey state-of-the-art techniques for building profiles, in order to determine how they can be extended to the case where a data source is described by a domain ontology. In particular, we will define the requirements that must be satisfied by the content summaries so as to ensure that they can be effectively exploited to determine the relevance of a data source in answering a query.
(Execution) A thorough analysis of the different distributed query processing techniques will be carried on, so as to highlight the limits of such techniques with respect to the WISDOM architecture (we remind that in WISDOM a data source is externally perceived only through its domain ontology). In particular, the different aspects that may influence the relevance of a result will be analyzed to see at which extent they are influenced by the WISDOM architecture.
(Navigation) We will analyze how query results can be elaborated in order to be returned to the user in a compact and easy to use form. Then, we will evaluate how navigation and aggregation techniques experienced in business intelligence and data mining can be combined in order to ensure the maximum flexibility in choosing the level of granularity for presenting data. Furthermore, we will study how the paradigms devised for visual querying databases can be extended to queries that involve the use of ontologies.
Finally, we will work, together with the other Units, on the definition of the methodological and functional architecture for the whole project (deliverable D0.R1).

PRODUCTS OF PHASE 1
The expected deliverables in this phase are technical reports (R). The number after the letter D represents the theme (0 for propdotti common to all the themes).
D0.R1 Report on the methodological and functional architecture (in collaboration with Modena e Reggio Emilia - MO, Roma - RM, Trento - TN)
D1.R1 Review of the languages and emerging standards for ontologies (in collaboration with MO, RM, TN)
D3.R1 Review of the query languages and of the rewriting techniques based on ontologies (in collaboration with MO, TN)
D3.R2 Review of query processing techniques in heterogeneous environments

SECOND PHASE (6 MONTHS)
During the second phase we will work on solutions for the 3 topics handled by the Unit:
(Content summaries) During the second phase we will define the mechanisms for adding numeric information to the domain ontologies. The basic idea is to extend the existing techniques for "probing" the data sources by considering ontological information and the derived constraints. The extension will be inspired by economy principles as: 1) require as few "probes" as possible, and 2) return the most significant quantitative information given fixed quantity of memory for storing content summaries. According to the targets of Theme 1, we will specify the update methods for the content summaries when a new data source is added and when the corresponding domain ontology is extended.
(Execution) The aim of this phase is the definition of a set of techniques for the execution of distributed queries that, considering the limits imposed by WISDOM architecture, return the most relevant results while minimizing the used resources. Since the relevance of a given object depends on several factors and on their relationships, the techniques that will be developed will be very general in order to be capable of working properly and efficiently even when the combination criterion is changed. For this criterion, which initially may be implemented as a weighted sum of the different factors, we will also consider the more general and expressive "qualitative" case, that is, based not only on numerical techniques.
(Navigation) As concern the exploitation of the query results, we will identify the techniques necessary to precisely define the desired granularity level. In particular, we will define the compact and rich in semantic representations for information available at different abstraction levels, and we will identify the operators necessary for an interactive navigation on the different levels according to the domain ontology.
Finally, we will work, in collaboration with the other Units, to the definition of the interfaces of the different components of the integrated prototype (deliverable D0.R2).

PRODUCTS OF PHASE 2
D0.R2 Specification of the component interfaces of the integrated prototype (in collaboration with MO, RM, TN)
D1.R2 Definition of the language for the specification of domain ontologies (in collaboration with MO, TN)
D1.R3 Definition of the techniques for the creation of content summaries
D3.R3 Definition of the query language and of the ontology-based query rewriting techniques (in collaboration with MO, TN)
D3.R4 Definition of query execution techniques in the WISDOM environment

THIRD PHASE (12 MONTHS)
During the third phase we will develop 3 prototypes and, jointly with the other Units, we will collaborate to the integration of the prototypes developed in the project.
The first prototype, starting from a pre-existing domain ontology, will implement "probing" techniques for the corresponding data sources and it will define algorithms for building the content summaries starting from the results obtained.
The second prototype (joint work with the MO Unit) will accept and analyze user queries. It will also determine the sets of relevant sources for the query at hand.
The third prototype will implement the query execution techniques defined during phase 2. Further, it will include an interface for an ontology-based interactive navigation at different abstraction levels.
An extensive experimental activity will be carried on in order to asses the performance of prototypes.

PRODUCTS OF PHASE 3
Deliverables expected for this phase are software prototypes (P).
D0.P1 Integrated system prototype (in collaboration with MO, RM, TN)
D1.P2 Prototype for the creation of content summaries
D3.P1 Prototype for query specification (in collaboration with MO)
D3.P2 Prototype of the query execution engine

2.6 Descrizione delle attrezzature già disponibili ed utilizzabili per la ricerca proposta con valore patrimoniale superiore a 25.000 Euro

Testo italiano

Nessuna

Testo inglese

Nessuna

2.7 Descrizione delle Grandi attrezzature da acquisire (GA)

Testo italiano

Nessuna

Testo inglese

Nessuna

2.8 Mesi uomo complessivi dedicati al programma

		Numero	Mesi uomo 1° anno	Mesi uomo 2° anno	Totale mesi uomo
Personale universitario dell'Università sede dell'Unità di Ricerca		*6*	*21*	*20*	*41*
Personale universitario di altre Università		*0*	*0*	*0*	*0*
Titolari di assegni di ricerca		*0*			
Titolari di borse	*Dottorato*	*2*	*6*	*6*	*12*
	Post-dottorato	*0*			
	Scuola di Specializzazione	*0*			
Personale a contratto	*Assegnisti*	*0*			
	Borsisti	*0*			
	Dottorandi	*0*			
	Altre tipologie	*2*	*0*	*12*	*12*
Personale extrauniversitario		*0*			
TOTALE		**10**	**27**	**38**	**65**

3.1 Costo complessivo del Programma dell'Unità di Ricerca

Testo italiano

Voce di spesa	Spesa in Euro	Descrizione
Materiale inventariabile	*20.000*	*Personal computer, workstation, libri*
Grandi Attrezzature		
Materiale di consumo e funzionamento	*5.000*	*software, cancelleria, toner stampante*
Spese per calcolo ed elaborazione dati		
Personale a contratto	*18.000*	*personale per la realizzazione di prototipi*
Servizi esterni		
Missioni	*40.000*	*riunioni di progetto, trasferte per convegni nazionali e internazionali di interesse per la ricerca*
Pubblicazioni		
Partecipazione / Organizzazione convegni	*7.000*	*Iscrizioni a convegni e workshop*
Altro		
TOTALE	**90.000**	

Testo inglese

Voce di spesa	Spesa in Euro	Descrizione
Materiale inventariabile	*20.000*	*Personal computers, workstation, books*
Grandi Attrezzature		
Materiale di consumo e funzionamento	*5.000*	*software, paper, printer toner*
Spese per calcolo ed elaborazione dati		
Personale a contratto	*18.000*	*personnel to implement prototypes*
Servizi esterni		
Missioni	*40.000*	*project meetings, travel expenses for attending national and international conferences relevant to the research activity*
Pubblicazioni		
Partecipazione / Organizzazione convegni	*7.000*	*payment of workshops and conference fees*
Altro		
TOTALE	**90.000**	

3.2 Costo complessivo del Programma di Ricerca

		Descrizione
Costo complessivo del Programma dell'Unità di Ricerca	*90.000*	
Fondi disponibili (RD)	*16.600*	*Fondi progetti RFO, convenzioni*
Fondi acquisibili (RA)	*10.400*	*Budget virtuale di Ateneo per COFIN2004*
Cofinanziamento di altre amministrazioni		
Cofinanziamento richiesto al MIUR	*63.000*	

3.3.1 Certifico la dichiarata disponibilità e l'utilizzabilità dei fondi di Ateneo (RD e RA)

SI

Occorre precisare che la quota di cofinanziamento MIUR più la quota di cofinanziamento di altre amministrazioni cofinanziatrici del Programma di Ricerca non potrà superare il 70% per programmi Interuniversitari e il 50% per programmi Intrauniversitari del costo totale ammissibile del Programma stesso.

(per la copia da depositare presso l'Ateneo e per l'assenso alla diffusione via Internet delle informazioni riguardanti i programmi finanziati e la loro elaborazione necessaria alle valutazioni; legge del 31.12.96 n° 675 sulla "Tutela dei dati personali")

Firma _____ Data 19/03/2004 ore 16:29