Information Systems Group @ DII-Unimo: Shaping Tomorrow Information Management, Today

Riccardo Martoglia
riccardo.martoglia@unimo.it

ISGroup - Information Systems Group
homepage: www.isgroup.unimo.it

Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
tel. 059 205 6142
fax 059 205 6129

The recent developments in computing power and telecommunications, and, in general, the advanced ICT (Information and Communication Technology) of the 20th century, accelerated the use and value of **Information** in our society. Indeed, Information is the main value of Information Society. In this respect, the World Wide Web, Peer-to-Peer networks, mobile devices and ubiquitous computing systems and sensors give us more and more interesting possibilities today; however, current research on the relevant technologies, structures and services is still not enough mature.

Research at the Information Systems Group (ISGroup), inside the Information Engineering Department (DII) of the Modena and Reggio Emilia University, is focused on the design and development of new systems, algorithms and data structures for the **access and management of Information**. The group constantly devises and puts into practice, also by means of national and international research projects and collaborations, innovative solutions able to answer, both effectively and efficiently, increasingly complex Information needs in several 21st century applications.

Information is everywhere and comes in many flavours: textual information, multi-lingual information, structural (XML) and multimedia information, multi-version information. Think, for instance, to product descriptions, data sheets, notes, web information, real-time data coming from sensors, etc. What follows is a short presentation of the past and present research activities of the group; thanks to this overview, the reader will have a glimpse of many of the practical applications that benefited from the obtained results, also by possibly investigating them further through the provided references. Most importantly, the key message is that information management, in its many forms, is crucial at every level and in every application scenario; indeed, the following is only an example of what can be achieved. ISGroup will be pleased to be contacted and to take up any type of proposed information management challenge, even through new collaborations or projects.

Text retrieval & information extraction

Textual information is the most popular form of electronic information representation. **Text** is everywhere in everyday work, and for this exact reason finding what is needed fast and in large amounts of data often proves to be a very demanding task. ISGroup has worked on many aspects of Text Retrieval and developed several algorithms and data structures allowing users to search for the needed information not only exactly, but also approximately. For instance, in

the field of Example-Based Machine Translation (EBMT), **approximate search** techniques are applied to sentences, also taking care of additional issues such as multilingualism. Specifically, past translations are stored in a database called Translation Memory (TM); when the translator submits a new document to be translated to the system, which is called **EXTRA** [1], it automatically proposes several translation suggestions, by performing a search among the similar sentences in the TM. In this way, the translator can finish the work in a more consistent and efficient way. Experiments have shown that assisted translation of a complete document takes almost 50% less time than the corresponding manual translation [8].

Similar techniques have also been exploited in very different applications, such as assisted label creation: design and print phases are greatly sped-up thanks to the useful information extracted from past works through **advanced information search techniques**, such as graphical browsing for whole labels and their parts (logos, texts, etc.), filtering on different fields or full-text search for specific words and sentences.

Structured & multimedia information management

"Pure" text is not always sufficient for representing the information one has to deal with. It is often necessary to represent and search for specific relationships between concepts (structure) or make use of images, audio and video (multimedia). Concerning the first aspect, ISGroup has studied **XML**, that is today considered the de-facto standard for data representation and interchange, and has proposed structural query processing techniques, algorithms and data structures allowing an efficient handling of very conspicuous quantities of data [9]. Further, the research has focused on managing and querying the more and more widespread **multimedia audio/visual objects**, expressed in the MPEG-7 standard, in multimedia databases (**X-Siter** system). Such techniques are able to support the most varied searching needs, for instance in video-surveillance scenarios ("retrieve all the suspicious behaviours") or in sport repositories ("retrieve all the sequences in which a pilot overtakes immediately after a pitstop"). As for all the ISGroup systems, X-Siter guarantees a very satisfying efficiency level: for instance, a complex search on a 6 million nodes database is performed in less than 0.2 seconds.

ISGroup has also researched advanced aspects such as **personalization** and management of **multi-version** documents. Even if it might not seem so, multi-versioning is a familiar concept in everyday life: for example, in a normative environment, each law changes in time due to successive modifications, while maintaining its own identity. In these cases, it is useful to reconstruct the consolidated version of a norm, but also past versions can be important. ISGroup has created a system for reconstructing and presenting the useful documents on the basis of the relevant retrieved portions; in this way, each citizen can easily access the portions of the laws which are relevant to its condition [4]. A possible query, which is handled in less than 2 seconds on more than 20000 documents, is "retrieve all the norms containing paragraphs dealing with health-care which were valid and in effect between 2002 and 2004 and which are applicable to self-employed citizens".

Intranet & P2P searching

An even more popular scenario is the one in which the information to be searched is not available in a single place (for instance, the computer which we are working on), but is distributed in different locations, such as in an intranet or in a global network (internet and PeerToPeer systems). Specifically, ISGroup has developed **advanced search engines** working on large quantities of **distributed and heterogeneous information**, i.e. information describing the same reality in different and, thus, incompatible ways. In the **SUNRISE** system [2], flexible querying mechanisms are made available in order to allow users to easily express

their search needs in every situation [6]. Further, queries are automatically adapted (query rewriting) and executed in the network with respect to each useful available document. In order to achieve this, schema matching and semantical multi-lingual annotation techniques are exploited for the **automatic meaning discovery** of the available data (**STRIDER** Word Sense Disambiguation module [3, 7]). For instance, the query "retrieve the plot of the movie Indiana Jones IV" is handled by automatically understanding, on the basis of the context, the correct meanings of the terms "plot" and "movie", and by properly adapting them to the way data is expressed in the different nodes of the network (such as "retrieve the *story* of the *film* Indiana Jones IV"). Moreover, in order to ensure a very high efficiency level, specific query routing techniques constantly select the best network directions to propagate the query to [5].

Real-time sensor information management

A very recent research area is the one of **real-time** sensor information management. In several application scenarios it is possible to exploit **sensors** which are able to interact with their environment by measuring and monitoring physical parameters and by wirelessly communicating the acquired information. This represents a new class of networks, known as **Wireless Sensor Networks** (WSNs). WSN nodes are required to be simple and easily programmable, to have reduced dimensions and low cost; in this way, they can be fit and exploited in different situations. Regarding to this topic, ISGroup research focuses on managing the very large amount of information inside WSNs by means of ad-hoc databases, so to allow an easier information extraction thanks to data-mining derived analysis techniques. In particular, a new WSN laboratory has been created and a new project has been started with the aim of building WSNs with proprietary and flexible hardware and software; in the future we also plan to extend the project by considering also RFID (Radio Frequency IDentifier) devices. The proposed solutions will be tested in real WSN application scenarios, such us monitoring biomedical parameters, industrial manufacturing, environmental control, logistics, and more.

The ISGroup research experience at your disposal

The Information Systems Group was established by Paolo Tiberio, Federica Mandreoli and Riccardo Martoglia. The current formation also includes Simona Sassatelli, Giorgio Villani and Fabio Bertarelli, together with several external collaborators. In the past years, the group obtained many national and international scientific achievements, as can be seen by the many successfully accomplished projects, collaborations and publications. The group is active inside the Department of Information Engineering (DII) of the Modena and Reggio Emilia University, directed by Gianni Immovilli; recently performed official research evaluations have put DII inside the **area of excellence** of the said University, thanks to a CIVR rating well above national means.

Effectively and efficiently managing information and knowledge is essential to every application scenario. For more details on ISGroup research activities, publications, projects and software, please refer to the group website: www.isgroup.unimo.it. Most importantly, the group is open to requests and proposals for projects, collaborations and professional advice; for any question, specific need or information request, please contact riccardo.martoglia@unimo.it.

Information Systems Group @ DII-Unimo: Ricercare l'Information Management di Domani, Oggi

Riccardo Martoglia
riccardo.martoglia@unimo.it

ISGroup - Information Systems Group
homepage: www.isgroup.unimo.it

Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
tel. 059 205 6142
fax 059 205 6129

I recenti sviluppi nella potenza di calcolo e nelle telecomunicazioni e, più in generale, le forme avanzate di ICT (Information and Communication Technology) del 20° secolo, hanno incrementato l'uso e il valore dell'**Informazione** nella nostra società. Le Informazioni sono il bene primario della cosiddetta Information Society. A questo proposito, il World Wide Web, le reti Peer-to-Peer, i dispositivi mobili, i sensori e i sistemi di calcolo presenti ormai in ogni luogo offrono oggi possibilità di sempre maggiore interesse. Purtroppo, però, la ricerca sulle relative tecnologie, strutture e servizi non è ancora sufficientemente matura.

L'obiettivo della ricerca all'Information Systems Group (ISGroup) del Dipartimento di Ingegneria dell'Informazione (DII), Università di Modena e Reggio Emilia, è la progettazione e lo sviluppo di nuovi sistemi, algoritmi e strutture dati per **l'accesso e la gestione delle Informazioni**. Il gruppo idea e mette costantemente in pratica, anche nell'ambito di progetti di ricerca e collaborazioni nazionali ed internazionali, soluzioni innovative in grado di soddisfare efficacemente ed efficientemente i fabbisogni informativi sempre più complessi in numerose applicazioni all'avanguardia.

Le informazioni sono ovunque e si presentano in tante forme: informazioni testuali, informazioni multi-linguistiche, informazioni strutturate (XML) e multimediali, informazioni multi-versione. Si pensi, ad esempio, a descrizioni di prodotti, schede tecniche, note, informazioni provenienti da web, informazioni in tempo reale provenienti da sensori, ecc. Quello che segue rappresenta una breve presentazione delle attività di ricerca presenti e passate del gruppo; grazie a questa panoramica, il lettore potrà toccare con mano molte delle applicazioni pratiche che hanno beneficiato dei risultati ottenuti, eventualmente approfondendole tramite i riferimenti che verranno forniti. Soprattutto, quello che si cercherà di rendere chiaro è che la gestione delle informazioni, nelle sue tante sfaccettature, è indispensabile ad ogni livello e in ogni campo applicativo; proprio per questo, tutto quello di cui si parlerà rappresenta soltanto un esempio di ciò che è possibile ottenere. L'ISGroup sarà felice di essere contattato e di risolvere ogni genere di sfida proposta nel proprio settore, anche nell'ambito di possibili nuovi collaborazioni o progetti.

Gestione ed estrazione di informazioni testuali

Le informazioni testuali rappresentano la più diffusa forma di rappresentazione elettronica delle informazioni. Il **testo** è presente ovunque nel lavoro di tutti i giorni, proprio per questo spesso risulta complesso trovare velocemente ciò di cui si ha bisogno all'interno di grandi moli di dati. L'ISGroup si è occupato di vari ambiti di Text Retrieval; sono stati sviluppati

numerosi algoritmi e strutture dati che permettono di recuperare le informazioni richieste effettuando la ricerca non solo in maniera esatta, ma anche approssimata. Ad esempio, nell'ambito della traduzione assistita basata su esempi o EBMT (Example Based Machine Translation), le tecniche di **ricerca approssimata** vengono applicate alle frasi, tenendo conto anche di tutta una serie di problematiche aggiuntive quali il multi-linguismo. In particolare, le traduzioni passate vengono memorizzate in una base di dati detta Translation Memory (TM); quando il traduttore sottopone al sistema, denominato **EXTRA** [1], un nuovo documento da tradurre, questo automaticamente propone una serie di suggerimenti di traduzione, effettuando la ricerca di frasi simili nel database. In questo modo, si è in grado di aumentare la qualità e la velocità del lavoro. Ad esempio, da test effettuati, la traduzione assistita di un testo completo richiede quasi il 50% di tempo in meno della corrispondente traduzione tradizionale [8].

Tecniche simili sono state applicate anche ad ambiti del tutto diversi, quali ad esempio la realizzazione assistita di etichette: il lavoro del progettista e dello stampatore è velocizzato dalla ricerca di informazioni utili estratte da lavori passati, effettuata tramite **tecniche avanzate di information search**, come browsing grafico per etichette e relative parti (loghi, testi, ecc.), filtraggio su diversi campi o ricerche full-text di frasi e parole specifiche.

Interrogazione di informazioni strutturate (XML), multimediali e multi versione

Il testo "puro" non sempre è sufficiente a rappresentare le informazioni con cui si ha a che fare. Spesso è necessario rappresentare e ricercare specifiche relazioni tra concetti (struttura) o fare uso di immagini, audio e video (multimedia). Per quanto riguarda il primo aspetto, si è gestito lo standard **XML**, che sta conoscendo una popolarità sempre maggiore per la rappresentazione dei dati e si sono studiate tecniche per l'elaborazione di interrogazioni strutturali (twig query processing), unitamente ad algoritmi e strutture dati che ne permettono un'esecuzione efficiente anche su notevoli quantità di dati [9]. E' stata poi resa possibile la gestione e l'interrogazione dei sempre più diffusi **oggetti multimediali di tipo audio/video**, descritti attraverso lo standard MPEG-7, in basi di dati multimediali (sistema **X-Siter**). Tali tecniche permettono di venire incontro alle più svariate esigenze di ricerca, ad esempio in ambito di sistemi di video-sorveglianza ("individuare tutti i comportamenti sospetti") o di repository sportivi ("individuare tutte le sequenze in cui un pilota compie un sorpasso subito dopo essere uscito dai box"). Come tutti i sistemi ISGroup, X-Siter garantisce un'eccellente efficienza: ad esempio, una ricerca complessa in una base di dati di più di 6 milioni di elementi è gestita in meno di 0,2 secondi.

Il gruppo ha approfondito anche aspetti avanzati quali la **personalizzazione** e la gestione di documenti **multi-versione**. Quello della multi-versione è in realtà un concetto molto più familiare di quello che non si creda: ad esempio, in ambito normativo, ogni legge cambia nel tempo a causa di modifiche successive, ma mantiene la sua identità. E' utile in questi casi ricostruire la versione consolidata di un documento, ma anche le versioni passate possono essere importanti. La ricerca effettuata dal gruppo ha permesso la creazione di un sistema per la ricostruzione e presentazione dei documenti ricercati a partire dalle porzioni risultate rilevanti, in grado di fornire a ciascun cittadino solo la porzione di legge di interesse per la sua condizione [4]. Una possibile interrogazione, gestita in meno di 2 secondi su oltre 20000 documenti, è ad esempio "recuperare tutte le norme che contengono paragrafi riguardanti la salute che erano validi tra il 2002 e il 2004, e che sono applicabili ai lavoratori autonomi".

Intranet & P2P searching

Sempre più diffuso è poi lo scenario in cui le informazioni da ricercare non sono presenti in un unico luogo (ad esempio il computer che si sta utilizzando), ma sono distribuite in più punti,

ad esempio in una rete locale (intranet) o in una rete globale (internet e sistemi PeerToPeer). In particolare, ISGroup ha affrontato lo sviluppo di **search engine avanzati** su grandi quantità di **informazioni distribuite ed eterogenee**, che descrivono cioè una stessa realtà ma in modo diverso e, quindi, incompatibile tra di loro. All'interno del sistema **SUNRISE** [2], sono state realizzati meccanismi di interrogazione flessibili che permettono in ogni situazione agli utenti di esprimere con facilità le loro richieste [6]. Queste vengono automaticamente adattate (query rewriting) ed eseguite nella rete, rispetto ad ogni documento utile a soddisfarle. Per ottenere questo, si sono utilizzati, tra gli altri, algoritmi di schema matching e tecniche di annotazione semantica multi-linguistica, per la **comprensione automatica del significato** dei dati (modulo di Word Sense Disambiguation **STRIDER** [3, 7]). Ad esempio, l'interrogazione "recuperare la trama del film Indiana Jones IV" è gestita individuando automaticamente, in base al contesto, il significato dei concetti "trama" e "film", ed adattandola opportunamente a come i dati sono espressi nei vari nodi della rete (ad esempio, "il *riassunto* del *lungometraggio* Indiana Jones IV"). Per garantire la massima efficienza, tecniche di query routing consentono inoltre di selezionare nella rete le direzioni migliori per la propagazione dell'interrogazione [5].

Gestione di informazione real-time da sensori

Una recentissima area di ricerca è quella della **gestione in tempo reale** delle informazioni. In molti scenari applicativi è possibile utilizzare **sensori** in grado di interagire con il loro ambiente misurando o controllando parametri fisici e comunicare le informazioni acquisite via wireless. In questo caso si parla di nuove classi di reti, note come **Wireless Sensor Network** (WSN). I nodi hanno come requisito di base la semplicità e devono avere dimensioni ridotte, basso costo ed essere di facile programmazione per adattarli a scopi diversi. In questo ambito, la ricerca dell'ISGroup si concentra sulla gestione di questa grossa mole di dati tramite database ad-hoc, in modo da permettere una più facile estrazione delle informazioni che l'analisi trasversale dei valori può offrire. In particolare, è stato creato uno specifico laboratorio di WSN e si sta sviluppando un progetto che vede la realizzazione di WSN con hardware e software proprietario flessibile, che verrà poi ampliato anche con dispositivi **RFID** (Radio Frequency IDentifier) tipicamente passivi. Le soluzioni proposte verranno provate nell'ambito di WSN e scenari applicativi reali, come la rilevazione di parametri biomedicali, di lavorazioni industriali, di controllo ambientale, di logistica ed altro ancora.

L'esperienza di ricerca dell'ISGroup a vostra disposizione

L'Information Systems Group è stato creato da Paolo Tiberio, Federica Mandreoli e Riccardo Martoglia. La formazione attuale include anche Simona Sassatelli, Giorgio Villani e Fabio Bertarelli, e si avvale ulteriormente di numerosi collaboratori esterni. Negli ultimi anni sono stati ottenuti risultati scientifici di grande rilevanza nazionale ed internazionale, come testimoniato dai molteplici progetti e collaborazioni conclusi con successo e dalle svariate pubblicazioni realizzate. Il gruppo opera all'interno del Dipartimento di Ingegneria dell'Informazione (DII) dell'Università di Modena e Reggio Emilia, diretto da Gianni Immovilli; il DII è stato collocato dalle recenti valutazioni della ricerca nell'**area di eccellenza** dell'Ateneo, grazie ad un rating CIVR ben al di sopra della media nazionale del settore.

La gestione efficace ed efficiente delle informazioni e della conoscenza è cruciale in qualunque scenario applicativo. Per maggiori dettagli e approfondimenti sulle attività di ricerca, sulle pubblicazioni, sui progetti e sui software dell'ISGroup, è possibile consultare il sito del gruppo: www.isgroup.unimo.it. Inoltre, il gruppo è del tutto aperto a proposte e richieste per progetti, collaborazioni e consulenze; per qualunque curiosità, esigenza specifica o richiesta di informazioni, contattare riccardo.martoglia@unimo.it.

Bibliography / Bibliografia

[1] http://www.isgroup.unimo.it/extra.asp

[2] http://www.isgroup.unimo.it/sunrise.asp

[3] http://www.strider.unimo.it/strider/introduction

[4] F. Grandi, F. Mandreoli, R. Martoglia. Issues in Personalized Access to Multi-Version XML Documents. In Open and Novel Issues in XML Database Applications: Future Directions and Advanced Technologies, Eric Pardede (Ed.), IGI Global, 2008.

[5] F. Mandreoli, R. Martoglia, W. Penzo, S. Sassatelli, G. Villani. Paving the Way to an Effective and Efficient Retrieval of Data over Semantic Overlay Networks. In The Semantic Web for Knowledge and Data Management: Technologies and Practices, Zhongmin Ma (Ed.), IGI Global, 2008.

[6] F. Mandreoli, R. Martoglia, W. Penzo, G. Villani. Flexible Query Answering on Graph-modeled Data. In Proc. of the 12th International Conference on Extending Database Technology (EDBT), St. Petersburg, Russia, 2009.

[7] F. Mandreoli, R. Martoglia, E. Ronchetti. Versatile Structural Disambiguation for Semantic-aware Applications. In Proc. of the 14th ACM International Conference on Information Knowledge and Management (ACM CIKM 2005), Bremen, Germany, 2005.

[8] F. Mandreoli, R. Martoglia, P. Tiberio. EXTRA: a system for example-based translation assistance. Machine Translation, Volume 20, Number 3, 2006.

[9] P. Zezula, F. Mandreoli, R. Martoglia. Tree Signatures and Unordered XML Pattern Matching. Invited talk. In Proc. of 30th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2004), Merin, Czech Republic, 2004.