Università degli Studi di Modena e Reggio Emilia

Facoltà di Ingegneria – Sede di Modena

Corso di Laurea in Ingegneria Informatica – *Nuovo Ordinamento*

Semantic Web: valutazione sistema NetWatcher

Relatore:
Prof. Sonia Bergamaschi

Candidato:
Luca Tassi

Correlatore:
Mirko Orsini

Anno Accademico 2004-2005

Parole chiave:

Semantic_Web
Web_Search
NetWatcher
Ontologie
Query

Ringraziamenti

Desidero ringraziare la prof. Sonia Bergamaschi, l'ing. Mirko Orsini e l'ing Alberto Corni per l'aiuto fornito e per la costante disponibilità dimostrata.

Ringrazio tutti i coloro che mi hanno affiancato e supportato durante la redazione di questo elaborato, fra cui: Marco Tosi, Ermes Spadoni, Annarosa Tavernari, Andrea Varone.

Un ringraziamento speciale va poi alla mia famiglia, alla mia ragazza, Valentina, e ai miei amici e compagni Pen e Giova, che mi sono stati vicini per tutto il percorso universitario.

Introduzione

Lo sviluppo delle tecnologie telematiche, tanto per i sistemi di elaborazione, quanto per le reti di calcolatori, ha portato ad una sempre maggiore presenza di sorgenti informative determinando un vera e propria esplosione nella quantità e varietà di dati accessibili. Poter gestire in modo efficace questa mole di dati è diventato quindi un fattore cruciale per il successo aziendale ma, paradossalmente, l'aumento nell'offerta di informazione fatica a tradursi in un effettivo vantaggio per l'utente. Questo perché i metodi di ricerca che si utilizzano generalmente non soddisfano in modo pieno le esigenze dell'utente, il quale, quando si appresta a fare una ricerca con essi, rischia di trovare un enorme quantità di risultati non significativi.

Il lavoro svolto, nell'ambito del tirocinio formativo effettuato da me e Marco Tosi alla CNA servizi di Modena, si inserisce perfettamente nel settore dell'informatica chiamato information retrieval, ovvero la disciplina che studia l'insieme delle operazioni che permettono il recupero di informazioni archiviate in formato elettronico. Il nostro studio non verte però nei sistemi di ricerca tradizionali, ma va ad analizzare una nuova frontiera di essi: i motori di ricerca semantici.

I motori di ricerca semantici sono una nuova tipologia di sistemi che si basa sulle nuove tecnologie che si stanno sviluppando nell'abito della ricerca sul Semantic Web. Il Semantic Web è un iniziativa del consorzio di World Wide Web (W3C) il cui obiettivo è di generare un mezzo universale per lo scambio di dati. Secondo questa prospettiva verranno collegati e integrati uniformemente i processi di gestione delle informazioni. Questo farà si che il Web diventi uno strumento più completo, efficace, utile, in cui le informazioni possano essere reperite in modo più pertinente e congruo.

La CNA servizi, nel periodo in cui è stato svolto il mio tirocinio, ha acquistato una licenza di prova dell'applicativo NetWatcher, di proprietà di Expert System s.p.a. Durante questo periodo siamo andati a effettuare tutta una serie di prove per valutare le potenzialità, i limiti e l'utilità di questo sistema. Per poter analizzare al meglio il sistema ci è stato però necessario effettuare anche una valutazione comparativa con un altro motore di ricerca.

Il sistema di confronto da noi trattato è il SEWASIE, il quale nasce da un progetto europeo coordinato dall'Università degli Studi di Modena e Reggio Emilia, che ha realizzato un prototipo di ricerca semantica che accede a sorgenti di dati eterogenee nel Web. Questo sistema ha però caratteristiche e peculiarità molto differenti da NetWatcher, e per questo abbiamo dovuto trovare un campo nel quale poterli analizzare entrambi: le news.

La CNA servizi di Modena, mediante un suo organo chiamato assomeccanica, gestisce un portale web del settore dell'industria meccanica e plastica; esso tra i vari servizi che offre mette anche a disposizione la spedizione settimanale di newsletter nelle quali vengono inserite tutte le notizie dei settori trattati. Con l'obiettivo di velocizzare i tempi di reperimento delle informazioni, durante il periodo di stage abbiamo cercato, mediante l'utilizzo del sistema NetWatcher di automatizzare il più possibile questo procedimento.

Grazie a questa attività di monitoring è quindi stato creato un grande archivio di notizie, sul quale abbiamo successivamente sviluppato una serie di query.
Mediante l'analisi dei risultati di esse, comparate con quelle ottenute dal sistema SEWASIE, siamo riusciti a effettuare una valutazione dei due sistemi.

Nel primo capitolo di questa tesi viene presentato e spiegato il contesto nel quale sono creati entrambi i sistemi da analizzare: il Semantic Web. Di esso vengono mostrate le principali caratteristiche innovative, l'architettura e alcuni dei linguaggi utilizzati, inoltre vengono esposti gli obiettivi e i metodi che utilizzano le applicazioni di web search semantiche.

Il secondo capitolo consiste invece in una descrizione sintetica del sistema SEWASIE che comprende una breve spiegazione dell'architettura, l'introduzione al sistema MOMIS, alcuni possibili scenari d'utilizzo e l'interfaccia grafica SQoogle.

Nel terzo capitolo viene presentato l'applicativo NetWatcher di Expert System s.p.a. partendo dalla piattaforma semantica sul quale si basa, passando all'architettura e, in fine, all'interfaccia utente.

Nel quarto capitolo ho illustrato i metodi utilizzati, per impostare l'attività di monitoring di news, svolta da NetWatcher, ai fini di creare la newsletter del portale di settore tuttostampi.com

Infine l'ultimo capitolo è il riassunto dell'esperimento di confronto. Si renderanno note tutte le specifiche e i parametri di confronto, le query formulate su entrambi i sistemi e i dati raccolti come risposta alle interrogazioni.

Indice:

1 Contesto del web semantico

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

[Tim Berners-Lee, James Hendler, Ora Lassila, “The Semantic Web”, *Scientific American*, Maggio 2001]

1.1 Evoluzione del web:dall’ HTML all’XML

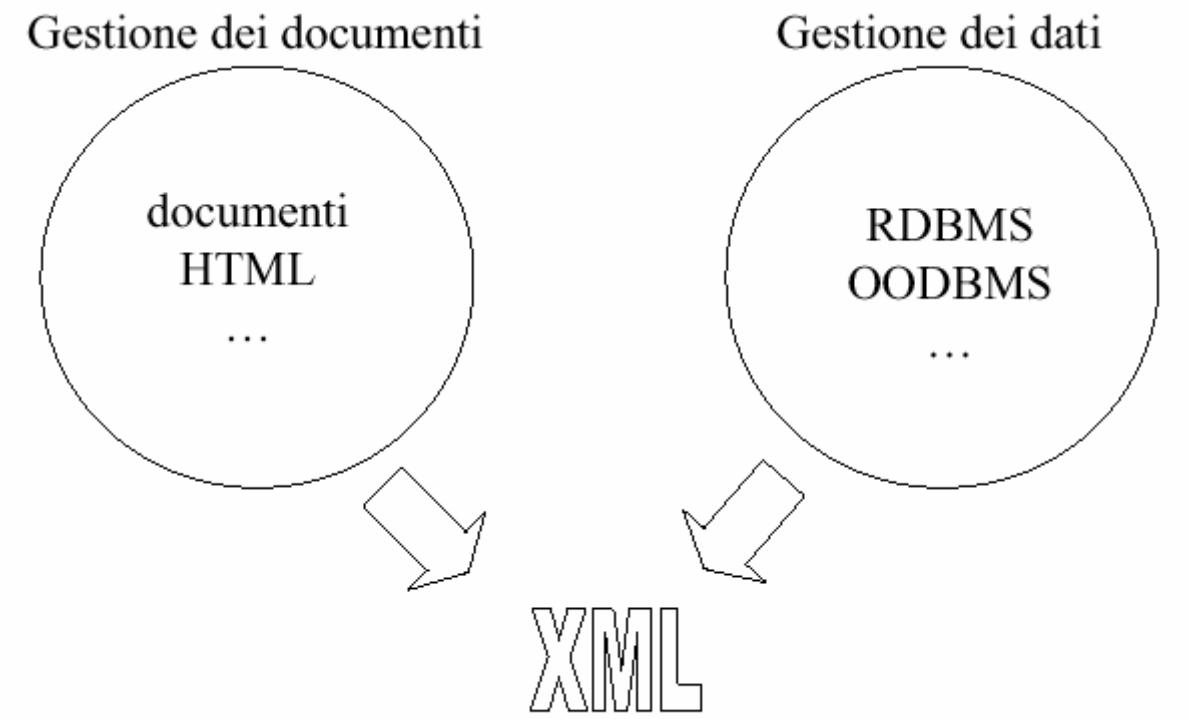
La rapidità della diffusione di Internet è dovuta certamente all'utilità dello strumento, alle sue implicazioni rivoluzionarie, ma non si sarebbe mai realizzata senza la presenza di un substrato tecnico facilmente diffondibile. HTML ha rappresentato lo strumento ideale per fungere da substrato. Innanzitutto si tratta di un linguaggio fruibile da qualsiasi sistema operativo, e in secondo luogo il suo utilizzo è semplice, sufficientemente intuitivo e aperto alla comprensione anche di chi non possegga particolari competenze informatiche.

Queste due caratteristiche derivano dal fatto che HTML non è altro che un linguaggio di formattazione e visualizzazione di testo che, partendo da un testo qualsiasi, definisce per ogni elemento la sua dimensione, forma, colore, posizione ecc… HTML (*Hyper Text Markup Language*)

L'informazione viene espressa ad un unico livello e questo rende l'utilizzo di HTML semplice, tuttavia ne compromette fin dall'inizio la potenza. Definendo a uno stesso livello link, formattazione e altre informazioni (quali ad esempio i metadati), il documento presenterà tutti questi elementi come unità rendendo difficoltosa la trasmissione, l'interscambio differenziato di questo tipo di informazioni. Se esprimo dati, collegamenti e formattazione sullo stesso documento e poi volessi posizionare quel documento diversamente nella mia Rete ipertestuale, non potrò utilizzare solo le informazioni che mi servono per costituire la nuova relazione ma dovrò "trascinarmi" tutto il file. In HTML i collegamenti fra documenti risultano molto spesso rozzi: si presentano o incapaci di comunicare la reale relazione semantica che attivano o incapaci di offrire un testo che la realizzi.

HTML è diventato il linguaggio universale per lo scambio di documenti nel web, ma non permette la descrizione della struttura di questi documenti. Un altro grosso limite dell’HTML è la scarsa interoperabilità con fonti di dati strutturate come i database, siano essi relazionali o ad oggetti o di altro tipo; in questo senso è utile un linguaggio che possa funzionare

come ponte tra la gestione grafica dei testi e la gestione delle strutture e dell'organizzazione dei dati.



Oggi per la comunità informatica si offre la possibilità di rendere più gestibile l'informazione in internet, come anche nelle reti locali. Questa possibilità è offerta dal linguaggio di marcatura XML (Extensible Markup Language), un linguaggio che ha come sue caratteristiche fondamentali quelle di rendere libera e personale la marcatura del testo e di suddividere l'informazione finale prodotta attraverso più documenti.
XML si adatta a tutti gli ambienti, può essere trasformato nei documenti più svariati e non costringe quindi ad adottare una nuova piattaforma di elaborazione. Proprio per la sua adattabilità offre dati facilmente attingibili, organizzati in una struttura permeabile, riconoscibile, ma non vincolante per la visualizzazione.

1.2 Cos'è il Semantic Web

Il web è stato creato come un grande spazio di informazioni con l'obiettivo di facilitare gli scambi di informazioni tra persone, per fare ciò è però utile un intervento partecipativo delle macchine. Un grosso ostacolo è dato dal fatto che, le informazioni sul web, sono generalmente create ad hoc per essere visualizzate e interpretate dalle persone, ma non dalle macchine. L'informazione contenuta all'interno della risorsa è quindi strutturata in modo da essere machine-readable (leggibile da una macchina) ma non machine-understandable (comprensibile da una macchina). In questo modo è molto difficile analizzare, in maniera automatica, le numerose informazioni presenti sul Web proprio perchè queste mancano di una

caratterizzazione semantica che possa essere interpretata e compresa automaticamente da un programma.

L'approccio del web semantico, tende a creare linguaggi per esprimere le informazioni in una forma facilmente interpretabile dalle macchine. Un ruolo molto importante in questo panorama è ricoperto dai Metadati, ovvero le informazioni, comprensibili dalla macchina, relative a una risorsa web, attraverso le quali è possibile ricavare delle informazioni sulla risorsa a cui sono associate.

Nel world wide web le informazioni vengono create da tanti utenti diversi e ad esse si può accedere mediante un semplice URI (unfied resource identifier), questo metodo di accesso ha reso il web molto popolare ma, questa grande semplicità di utilizzo, ha un prezzo: è molto facile perdersi o trovare informazioni inutili o irrilevanti per i nostri fini.

L'idea di fondo del web semantico è quella di far diventare la Rete in grado di capire le nostre richieste. Non in senso proprio, ovviamente: i documenti non dovrebbero più risultare come delle "isole di dati", ma piuttosto come dei database aperti nei quali un "applicativo" possa di distinguere le informazioni contenute, ricavandone solo quelle richieste.

Il web semantico si propone d'inserire nell'architettura della Rete elementi in grado di consentire ad agenti informatici una certa capacità d'azione. Tim Berners Lee nell'articolo sul Semantic Web apparso sul Scientific Amercan [2] immaginava che un motore di ricerca, scorrendo le pagine alla ricerca di una prenotazione aerea, fosse in grado di capire quali link portano alle pagine relative alla destinazione richiesta, quali siano i costi e gli orari dei biglietti, di confrontare tra loro le offerte e di coordinare la partenza con l'agenda dell'utente o con le limitazioni sui costi prima impostate.

Tutto questo non in virtù di sistemi di intelligenza artificiale, ma molto più semplicemente in virtù di una marcatura dei documenti, di un linguaggio gestibile da tutte le applicazioni e dell'introduzione di vocabolari specifici, cioè di collezioni di frasi alle quali possano associarsi relazioni stabilite fra gli elementi marcati. In pratica, il web semantico per funzionare deve poter disporre di informazioni strutturate e di regole di deduzione per gestirle.

Il web semantico si pone l'obiettivo di creare degli standard e delle tecnologie designate per aiutare le macchine a capire molte informazioni sui dati. Grazie a queste tecnologie, le macchine, potranno integrare e automatizzare diversi compiti, e inoltre sarà possibile arricchire il potere informativo dei dati che trattano. Con il web semantico, oltre ad avere risultati di ricerche di informazioni più precise, sarà possibile integrare questi risultati anche se essi provengono da risorse differenti.

Per migliorare l'automazione di servizi il web semantico si propone di aggiungere informazioni specifiche ai dati che si trovano nel web quindi non viene assegnato un URI soltanto ai documenti, come avviene nel web tradizionale, ma anche ai concetti, alle relazioni, e alle persone legate a quel documento. Così ad esempio, aggiungendo metadati sul creatore di

un documento, si potrebbero effettuare ricerche specificando “Marco Rossi” come editore.

XML costituisce la base ideale per la realizzazione di quello che è stato battezzato da Tim Berners Lee “Semantic Web”.

Per far sì che il web semantico rappresenti una svolta è necessario garantire due livelli di interoperabilità:

- Sintattica
- Semantica

L’interoperabilità sintattica è la capacità di leggere i dati ed ottenere una rappresentazione utilizzabile da un’applicazione.

L’interoperabilità semantica è invece la capacità di comprendere i dati da parte di un applicazione

Per stabilire questi due livelli di interoperabilità è necessario definire un linguaggio che permetta di formalizzare una semantica e di fornire un supporto al ragionamento (individuazione di relazioni d’inclusione, equivalenze implicite ed inconsistenze)

Il primo livello d’interoperabilità, quella sintattica, è già da tempo raggiunta grazie all’affermarsi dell’ XML che permette di definire la struttura dei propri documenti ma non dice nulla sul significato della struttura; mentre il secondo livello d’interoperabilità, quella semantica, ha ricevuto negli ultimi tempi un notevole aiuto dai linguaggi RDF, RDF – Schema e OWL.

1.3 Principi del Web Semantico

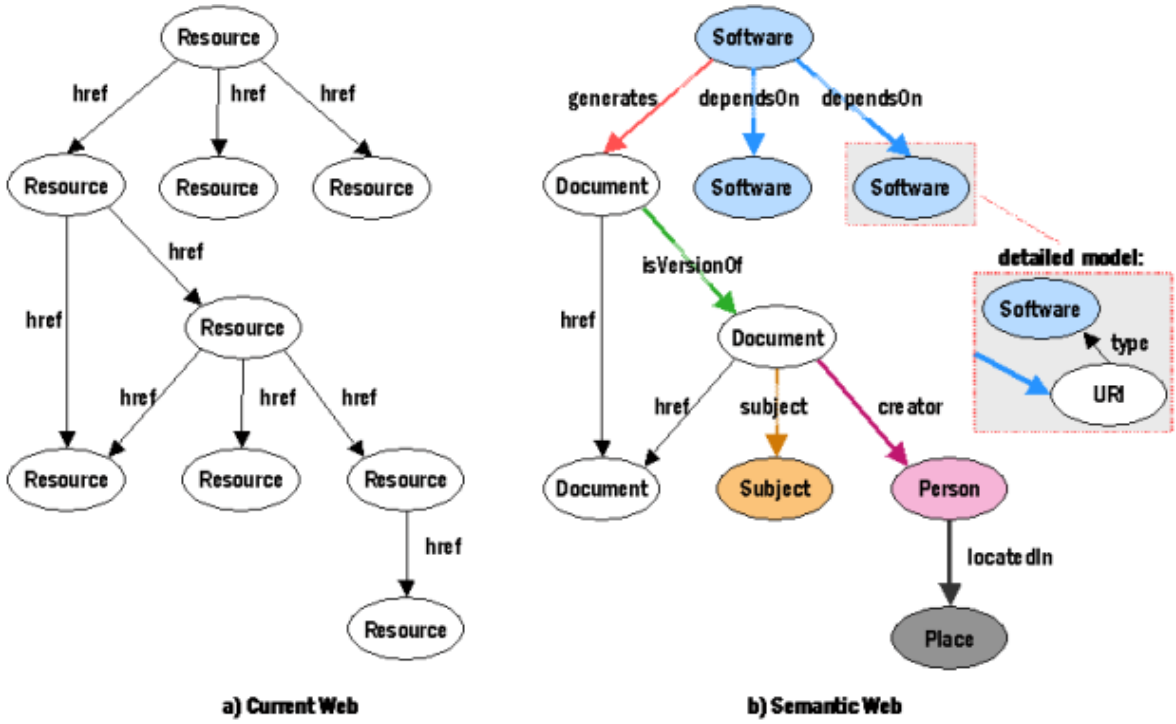
1) Tutto può esser identificato con un URI

Persone, luoghi, cose nel mondo fisico possono essere identificate nel web semantico utilizzando diversi tipi di identificatori. Chiunque abbia il controllo su una parte dello spazio dei nomi del web può creare un URI e, sostenere che identifica qualcosa nel mondo fisico. Si può inoltre riferirsi a un entità fisica indirettamente, ad esempio, per riferirsi alla città di Helsinki, ci si può riferire all’URI della pagina contenente le informazioni sulla città di Helsinki tenuta negli uffici cittadini.

2)Risorse e collegamenti possono avere dei tipi

Il web odierno consiste in risorse e collegamenti, le risorse generalmente non contengono metadati, che spiegano per cosa sono usate, e quali sono le relazioni con altri documenti web. Mentre, per le persone che leggono un documento web, è facile capire che tipo di collegamento è, un determinato link che porta a un'altra pagina, ciò è sicuramente più difficile

per una macchina, la quale non distingue un href da un altro. Per consentire alla macchina di capire il significato dei collegamenti esistenti in un documento, il web semantico, approva la definizione di tipi di link. Ad esempio un link può comunicare alla macchina che una risorsa "dipende da" , "è una versione di" o "ha come oggetto" la risorsa bersaglio del collegamento. I tipi di collegamento sono generalmente mappati in un nodo e hanno quindi anch'essi un URI a cui riferirsi.



3)Sono tollerate informazioni parziali

Nel web odierno non esistono limitazioni e l'integrità dei collegamenti è spesso sacrificata a favore della scalabilità. Un autore può facilmente collegarsi a risorse di altri e chiunque può accedere alle sue, tutto ciò senza tener conto di eventuali spostamenti o cancellazioni dei dati collegati. Questo politica permissiva ha però come rovescio della medaglia il fatto che è possibile arrivare a link irraggiungibili, che causeranno quindi errori 404 http.

Allo stesso modo il web semantico può essere considerato illimitato, infatti chiunque può dichiarare qualcosa su qualsiasi oggetto e creare tipi differenti di collegamenti tra risorse. Inoltre devono essere tollerate eventuali perdite di dati, causate da link che hanno cessato di esistere o che sono stati modificati.

4)Non è necessaria la verità assoluta delle informazioni

Non tutto quello che si trova nel web è attendibile, e il web semantico non cambierà questa situazione. L'affidabilità di un informazione è valutata dalle applicazioni che processano i dati nel web. Le applicazioni decidono

quali risorse ritenere affidabili usando il contesto delle dichiarazioni effettuate.

5)E' supportata l'evoluzione

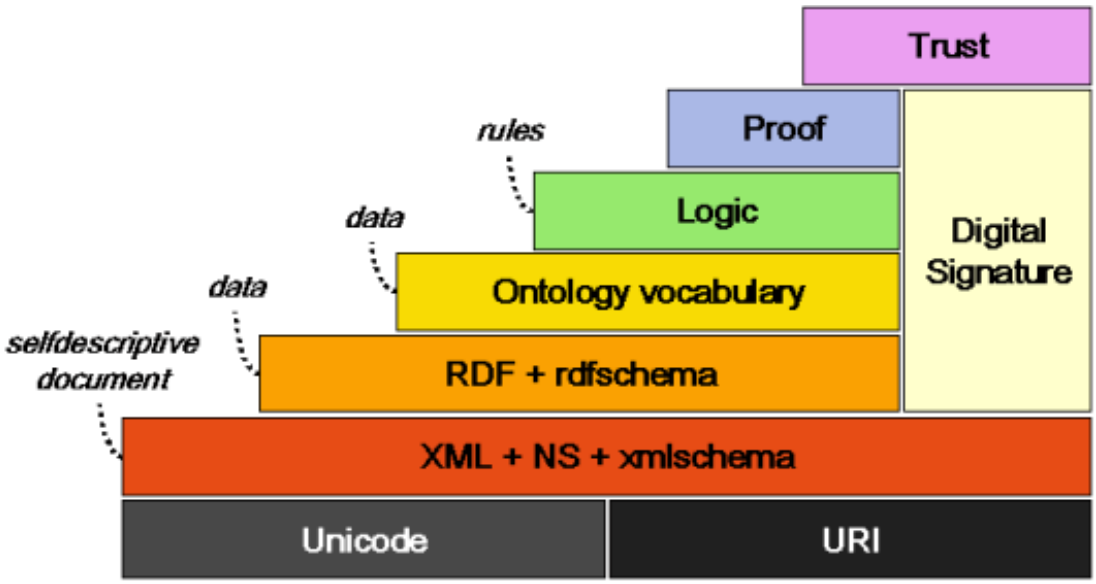
Il Web Semantico combina le informazioni disponibili che utilizzano gli stessi concetti, anche se questi sono definiti da persone diverse in momenti diversi. Per fare questo vengono utilizzate convenzioni che possono espandersi allo stesso modo con cui si espandono le conoscenze umane. Le convenzioni permettono l'integrazione di lavori indipendenti svolti da comunità diverse anche quando queste utilizzano vocabolari differenti, inoltre possono essere aggiunte informazioni senza modificare quelle esistenti.

6)Design minimalista

Questo principio mira a standardizzare soltanto le basi della tecnologia del Web Semantico. Una volta posta un infrastruttura accurata di standard sarà poi possibile lavorare per sviluppare nuove applicazioni basandosi su di essi.

1.4 Architettura del Web Semantico

L'architettura di riferimento per il Semantic Web è quella proposta dal W3C:



Tra gli strati sopra presentati ve ne sono alcuni già sviluppati:

- *UNICODE, URI:* si occupa del controllo dell'utilizzo di set di caratteri internazionali e dell'identificazione degli oggetti presenti.

- *XML, XMLSCHEMA, NAMESPACE:* garantisce l'integrabilità di nuove risorse definite seguendo gli standard basati sull'XML.
- *RDF, RDFSCHEMA:* consentono la possibilità di fare dichiarazione sugli oggetti e di creare vocabolari ai quali accedere mediante URI appropriati.
- *ONTOLOGIA, VOCABOLARIO:* consentono di definire relazioni e significati tra i diversi termini e concetti.
- *DIGITAL SIGNATURE:* Sono blocchi di dati criptati che gli agenti software possono utilizzare per verificare che le informazioni a cui sono allegate siano state reperite da una certa risorsa fidata. Con queste viene inoltre controllato se i documenti hanno subito alterazioni o corruzioni.
- *LOGICO:* consente la definizione di norme di utilizzo.

Mentre questi sono ancora in via di sviluppo:

- *PROVA:* . La prove devono essere una catena di asserzioni e regole di ragionamento con puntatori a tutto il materiale di supporto disponibile. Per provare e garantire i documenti che vengono spediti lungo la rete è necessario un linguaggio di validazione.
- *FIDUCIA:* valuta se le prove effettuate sono attendibili o meno.

Questa suddivisione è molto dettagliata, in alcuni lavori vengono proposti quattro macrolivelli fondamentali:

- Al primo livello abbiamo i *dati*
- A un secondo le informazioni sui dati e le relazioni che intercorrono tra i dati: i *metadati* e lo *Schema*.
- A un terzo livello vocabolari, o *ontologie*, che definiscono il ruolo semantico dei metadati.
- Il quarto è invece il livello logico nel quale sono presenti le logiche descrittive.

Il primo livello, come è già stato più volte ripetuto, è affidato a XML. Le caratteristiche fondamentali di questo linguaggio permettono una marcatura dei dati, e quindi una loro individuazione dentro una struttura, nonché la possibilità di rendere trattabili i dati da più applicazioni, con più visualizzazioni.

Il secondo livello, quello dei metadati, sarà affidato, secondo le intenzioni del W3C, al linguaggio RDF (Resource Description Framework). RDF è un linguaggio per la descrizione di metadati pensato per affiancare XML. I suoi elementi costitutivi e la sua sintassi sono XML, e tramite l'utilizzo di alcuni "namespace" sono specificate le funzioni dei suoi elementi.

RDF è il linguaggio utilizzato per introdurre il cosidetto *Modello basic assertion* il quale si basa su due concetti fondamentali: asserzione e quotazione.
Un asserzione è una parte di una descrizione di una risorsa e si presenta come una sua proprietà. Una quotazione è invece un asserzione fatta su un'altra asserzione. RDF consente di costruire delle asserzioni relativamente ai contenuti di una pagina web. Esse si creano in base a "dichiarazioni triple" costituite da soggetto, predicato e complemento oggetto. Tali asserzioni individuano relazioni fra i dati di cui trattano ma non esplicitano ancora il loro significato.

Per definire il significato delle relazioni serve un ulteriore livello, quello delle così dette ontologie. Le ontologie sono dei vocabolari nei quali collezioni di frasi sono associate a concetti. Le ontologie di dominio, cioè relative ad una collezione di documenti, possono essere scritte in vari linguaggi: il più adatto a XML è RDF Schema, ma ne esistono anche di più evoluti come OWL.

Nel quarto livello si scrivono espressioni logiche all'interno dei documenti che andranno a permettere determinate operazioni su di esso. Si possono definire ad esempio regole di deduzione del tipo di un documento partendo da un documento di un altro tipo, oppure il controllo di un documento basato su determinate regole di consistenza. Abbiamo quindi a che fare con un linguaggio che consente di svolgere calcoli e espressioni logiche sulle quotazioni.

1.5 Linguaggi del Web Semantico

Nei successivi paragrafi verrà presentata una semplice e non esaustiva descrizione di alcuni tra i più importanti e sviluppati linguaggi del web semantico.
Per uno studio più approfondito di questi argomenti è possibile trovare molte altre informazioni nelle sezione relative ai linguaggi del sito www.w3.org [7] e nel sito www.dbgroup.unimo.it [5]all'interno del materiale didattico del corso di *Rappresentazione della conoscenza* per quanto riguarda RDF e OWL, e *Tecnologia database* per quanto riguarda XML.

1.5.1 XML

La definizione di XML, in un primo momento, aveva come obiettivo originario il superamento dei limiti di HTML relativamente al mark-up per il

Web, successivamente nasce poi l'idea che XML possa servire per qualcosa più, come un linguaggio di markup per trasferire dati, non pensato per la visione umana, ma per essere prodotto ed usato da programmi, un meccanismo per convertire dati dal formato interno dell'applicazione ad un formato di trasporto, facile da convertire in altri formati interni. Attraverso l'uso di XSLT (Extensible Stylesheet Language Trasformation) infatti un documento XML può essere trasformato abbastanza facilmente in un qualsiasi altro linguaggio.

XML non consente solo di descrivere dati liberamente, ma anche di interpretarli in maniera automatica tramite tool standard. Esso è infatti utilizzato per creare una struttura dei documenti, cioè i documenti sono rappresentati in maniera da essere interpretati da un computer, affinché le informazioni contenute possano essere elaborate, memorizzate, ricercate o stampate.

Prima di tutto, si deve definire una struttura logica del documento che è comune a tutti i documenti di un certo tipo e individuare i suoi componenti. Ognuno di questi è un elemento XML ed è legato agli altri da vincoli di dipendenza.

La libertà di invenzione di nuovi tag permessa dall'XML non crea situazioni anarchiche, in quanto essa è regolata da norme ben precise. Un documento XML, infatti, è valido solo se fa riferimento ad una DTD (Document Type Definition), o ad uno Schema, e ne segue le regole grammaticali.

La funzione delle DTD è quella di definire formalmente la struttura e la sintassi di ogni tipo di documento XML, specificandone gli elementi, gli attributi e le entità, ma non il loro contenuto. Se il documento non è valido, potrebbe essere ben formato (cioè contenente un mark-up comprensibile), ma non conforme ad una DTD. Un documento XML deve dichiarare a quale DTD si conforma per verificarne la validità, ma tale dichiarazione non è obbligatoria e, se è assente, il documento non è considerato né valido, né non valido.
Le DTD hanno una sintassi particolare, diversa dall'XML, perciò sono richiesti strumenti appositi per la validazione. Inoltre, esse non distinguono tra nome del tag e tipo del tag, ed hanno solo due tipi: complesso (strutturato) e semplice (CDATA o #PCDATA).

Uno Schema XML è una collezione di definizioni di tipi e dichiarazioni di elementi utilizzati nel documento XML, che ne costituisce un'istanza. Gli schemi sono molto più potenti e precisi rispetto alle DTD, infatti sono stati pensati per fornire quel supporto di validazione che le DTD permettono solo parzialmente, in particolare sul contenuto degli elementi e degli attributi dei documenti XML.

Già da queste semplici osservazioni si può capire come un linguaggio di questo tipo offra un'alta flessibilità, sia nella definizione della struttura dei dati, sia nella rappresentazione.
In pratica, i dati sono collocati a livello diverso rispetto a struttura e visualizzazione, inseriti in strutture diverse e visualizzati in più modi. Un unico gruppo di dati può essere usato secondo necessità diverse.

Tutti gli elementi che compongono un documento hanno una struttura ad albero, dove l'elemento principale è la radice, gli altri sono i suoi sottoelementi, che sono a loro volta rami o foglie a seconda che abbiano o no dei discendenti.

Ogni elemento può avere degli attributi, un attributo e' una coppia (nome, valore) che permette di caratterizzare l'elemento. Si noti che la strutturazione di un documento in questi termini non è immediata, alcuni aspetti possono essere rappresentati sia come elementi, sia come attributi e da alcuni tipi di documento non è possibile ricavare una struttura perfettamente ad albero. Anche alla presenza di collegamenti fra un elemento dell'albero ed un altro, XML non dà problemi.
La struttura fisica di un documento XML è formata da entità. Un'entità è un costrutto che rappresenta una parte del testo, ha una dimensione variabile (da un carattere a tutto il documento) ed ha un nome. Per fare i riferimenti, come gli elementi indicano i componenti della struttura logica del documento, le entità specificano la posizione di insiemi di byte. Si deve fare in modo che questi due aspetti siano correlati tra loro.
L'insieme di comandi che stabiliscono questo legame è il mark-up. Una dichiarazione di markup e' una dichiarazione di un elemento, oppure una dichiarazione di una lista di attributi, oppure una dichiarazione di una entità la quale sarà poi compresa dagli elaboratori. Questi comandi sono distinti dal testo in quanto sono racchiusi fra i simboli di minore ("<") e maggiore (">") i quali vengono chiamati delimitatori.
La forma è analoga a quella dei tag HTML, ma, se in questo caso danno le indicazioni per la formattazione delle informazioni, in un documento XML non ci dovrebbero essere istruzioni per la loro visualizzazione. Quello che non è racchiuso fra i tag, è interpretato come testo e non è generalmente analizzato.

1.5.2 RDF

RDF (Resource Description Framework) è lo strumento proposto dal W3C per descrivere i metadati relativi ad una risorsa, mettendo a disposizione un linguaggio per esprimere la semantica di una risorsa.

La specifica di RDF è costituita da sei documenti i quali si trovano nella sezione RDF del sito del consorzio del world wide web [3]. RDF ha due componenti: il primo riguarda la definizione del data model RDF (modello

dei dati), tramite il quale descrivere le risorse, e della sintassi XML utilizzata per specificare questo modello.
RDF Schema[4] invece permette di definire il significato e le caratteristiche delle proprietà e delle relazioni che esistono tra queste e le risorse descritte nel data model RDF.

Una risorsa, identificata univocamente da un URI, viene descritta utilizzando il data model RDF.

Questo modello è basato su tre oggetti:

Resource (risorsa): indica ciò che viene descritto mediante RDF e può essere una risorsa Web (ad esempio una pagina HTML, un documento XML o parti di esso) o anche una risorsa esterna al Web (ad esempio un libro, un quadro, etc.);
Property (proprietà): indica una proprietà, un attributo o una relazione utilizzata per descrivere una risorsa. Il significato e le caratteristiche di questa componente vengono definite tramite RDF Schema;
Statement (espressione/asserzione): è l'elemento che descrive la risorsa ed è costituito da un soggetto (che rappresenta la Resource), un predicato (che esprime la Property) e da un oggetto (chiamato Value) che indica il valore della proprietà.

La struttura delle espressioni RDF è una collezione di triple nelle quali vi sono un soggetto, un predicato e un oggetto. Un insieme di queste triple è un grafo e viene illustrato mediante nodi e archi. Il primo nodo è il soggetto, l'arco orientato è il predicato e il secondo nodo è l'oggetto.

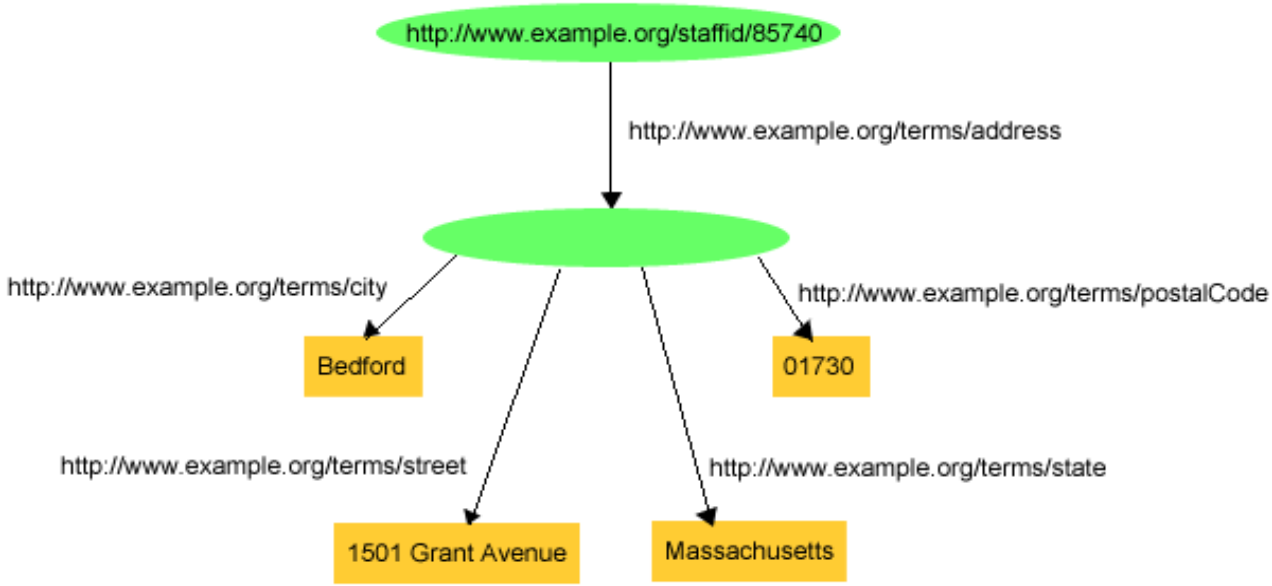
Le asserzioni di una tripla RDF dichiarano l'esistenza di una relazione (il predicato) tra il soggetto e l'oggetto. Il significato del grafo RDF è quindi la congiunzione (AND logico) di tutte le triple dichiarate.
Lo statement RDF che descrive la risorsa è quindi del tipo:

<soggetto> HAS <predicato> <oggetto>.

In accordo col primo principio del Semantic Web ogni tipo di oggetto è referenziato da un URI, cioè un identificatore in grado di localizzare esattamente e senza ambiguità una ed una sola risorsa. Un URI, usato come nodo, identifica cosa rappresenta quel nodo. Se invece la referenza URI è utilizzata come predicato identifica una relazione esistente tra i nodi che connette. Inoltre, si può impiegare una referenza URI di un predicato, come nodo di un grafo. Nella sintassi RDF un blank node è invece un nodo

che può esser usato in una o più espressioni ma che non ha un nome intrinseco.

Un'altra semplice rappresentazione della stessa espressione può essere quella della tabella appartenente a un database relazionale. In questa rappresentazione si ha una tabella con due colonne, che corrispondono a soggetto e oggetto della tripla RDF, mentre il nome della tabella corrisponde al predicato. All'interno di un database relazionale le tabelle possono contenere un numero arbitrario di colonne. Per poter convertire questa rappresentazione in una RDF sarà quindi necessaria una decomposizione, in quanto le triple RDF hanno solo un oggetto. Questa decomposizione può avvenire inserendo un blank node, corrispondente a una riga, e una tripla per ogni cella della riga. Il soggetto di ogni tripla è il blank node inserito, mentre i predicati corrispondono ai nomi delle colonne e gli oggetti sono il valore che vi era all'interno della cella.



RDF Schema, sempre in sintassi XML e attraverso l'uso di namespace, definisce relazioni per gli elementi RDF usati nella descrizione dei metadati.
Mentre RDF è un linguaggio general-purpose per la rappresentazione di informazioni sul web RDF Schema è la sua evoluzione per la creazione di vocabolari web condivisi.
In RDF le proprietà possono essere pensate sia come attributi di una risorsa, sia come relazioni tra di esse, ma non esistono meccanismi che consentono di descriver queste proprietà o le relazioni tra queste proprietà e altre risorse. Per ovviare a questa mancanza RDF Schema definisce classi e proprietà che possono esser utilizzate per descrivere altre proprietà, classi e altre risorse.
RDF Schema è dunque l'estensione semantica di RDF e consente di creare meccanismi per la descrizione di gruppi di informazioni e di relazioni esistenti tra queste risorse.

L'approccio utilizzato da RDF Schema per la descrizione di classi e proprietà è abbastanza simile a quello utilizzato dai linguaggi di programmazione orientati agli oggetti. Questo approccio consente a chiunque di aggiungere proprietà a un dominio degli attributi di un documento, senza il bisogno di ridefinire la descrizione di quella classe, concordando col principio architetturale del web, che consente l'estensione delle informazione sulle risorse a chiunque.
Il vocabolario di RDF Schema è informalmente chiamato rdfs ed è richiamato tramite il riferimento all'URI http://www.w3.org/TR/rdf-schema [4].
Nel documento RDF Schema si dovranno inoltre specificare le gerarchie tra gli elementi e in questo modo fornire ad un agente informatico la capacità di applicare ad essi regole di deduzione.
I costrutti linguistici più importanti introdotti dalla sintassi di RDF Schema sono le classi, con relative superclassi e sottoclassi, il concetto di sottoproprietà e superproprietà e in fine i *Container* e le *Collection*.

Le risorse possono essere divise in gruppi chiamati classi, i membri di esse sono le istanze della classe. Le classi possono essere referenziate con un URI, e inoltre possono essere descritte utilizzando proprietà RDF.
RDF Schema distingue le classi e il set delle sue istanze. Associata ad ogni classe vi è quindi un set di istanze della stessa, che viene chiamata estensione della classe. In questo modo due classi differenti possono avere la stessa estensione della classe; una classe può inoltre essere membro della sua stessa estensione, o può anche essere un istanza di se stessa.
RDF Schema introduce i concetti di sottoclasse per il quale se C è una sottoclasse di C' allora tutte le istanze di C sono anche istanze di C', in modo duale viene definito il concetto di superclasse. Sono considerate classi anche i tipi di dati e, le loro istanze, sono gli spazi dei valori.

Una proprietà è una relazione, dichiarata da una tripla RDF, tra una risorsa soggetto e una risorsa oggetto. RDF Schema introduce anche il concetto di subproprietà per il quale se una proprietà C è subproprietà di C' allora tutte le coppie di risorse che mette in relazione C sono relazionate anche da C'. In modo duale viene poi definito il concetto di superproprietà.

I *container* RDF sono risorse utilizzate per rappresentare insiemi di informazioni. Alcune risorse possono apparire più di una volta nello stesso container e inoltre, a differenza del mondo reale, i container possono essere contenuti in se stessi. I container RDF sono considerabili "aperti", nel senso che non esiste una sintassi RDF per creare meccanismi che dichiarano che non ci sono più membri.
Le RDF *collection* sono invece "chiuse" quindi in esse si può determinare se non hanno più elementi.

1.5.3 OWL

Il linguaggio per la creazione di ontologie OWL [6] è designato per l'utilizzo da parte di applicazioni che devono processare il contenuto delle informazioni invece di presentarlo semplicemente alle persone. Tramite OWL è possibile avere una maggiore comprensibilità della semantica di concetti da parte di agenti software rispetto a XML, RDF e RDF-Schema.
OWL può essere utilizzato per rappresentare il significato e le relazioni tra termini in un vocabolario. Questa rappresentazione di termini e di relazioni è chiamata ontologia. In uno scenario di sviluppo del web semantico nel quale, alle risorse viene assegnato un significato esplicito interpretabile dalle macchine, il linguaggio OWL si pone a un livello superiore a RDF e RDF Schema. A questo livello, il linguaggio, deve descrivere formalmente il significato delle terminologie utilizzate nei documenti web.
La differenza principale con RDF Schema è dunque il fatto che, in quest'ultimo viene creato un meccanismo per descrivere le risorse web, ma non si specificano i metodi con cui le applicazioni utilizzeranno queste meta informazioni, invece OWL è appositamente predisposto per creare vocabolari di informazioni sulle risorse interpretabili da un ragionatore automatico.

OWL ha tre sottolinguaggi i quali consentono un grado di espressività crescente.

- OWL Lite supporta semplici vincoli utilizzabili per la creazione di gerarchie. Non consente l'utilizzazione di cardinalità diverse da 0 e 1 ed è meno complesso di OWL DL.

- OWL DL è utilizzabile da coloro che vogliono massimizzare l'espressività garantendo comunque, che le espressioni siano calcolabili e che il sistema di ragionamento prenda decisioni in un tempo finito. Questo sottolinguaggio include tutti i costrutti di OWL, ma consente di utilizzarli rispettando determinate restrizioni. Il nome DL deriva da description logics che è un ramo della ricerca che ha studiato i fondamenti logici di tutto il linguaggio.

- OWL Full è utilizzabile da coloro che vogliono massimizzare l'espressività e avere tutte le libertà sintattiche consentite da RDF, tutto ciò però ha scapito della calcolabilità delle espressioni. E' possibile ad esempio trattare una classe simultaneamente sia come un insieme di individual, sia come un unico individual.

1.6 Web Search nel contesto del Web Semantico

Un motore di ricerca è uno strumento per mezzo del quale è possibile ricercare alcuni termini (parole) all'interno di una grande quantità di siti web. In seguito ad una ricerca, il motore di ricerca riporta una lista di siti che contengono i termini cercati.

Partendo da un esempio, i motori di ricerca attuali non capiscono la differenza fra il signor "Giuseppe Verdi" che abita al piano di sotto e il famoso compositore italiano di melodrammi dell'ottocento. Se potessimo associare alla ricerca la connotazione, potremmo in un sol colpo eliminare il rumore costituito da tutti quei documenti non pertinenti, implicando che i documenti devono fare riferimento ad una ontologia specifica sulla musica, sulle opere e sui compositori e che i termini significativi vadano "marcati" all'interno del testo, così:

```
"Quando <compositore>Giuseppe Verdi</compositore>
scrisse <opera>l'Aida"</opera> non immaginava di vederla
rappresentata nel 1871 al <capitale_egitto>Cairo</capitale_egitto>
per l'inaugurazione del canale di Suez.."
```

Grazie ai metadati, le pagine indicizzate dai motori di ricerca diverranno parte integrante di una immensa struttura semantica e saranno anche in grado di prevenire ragionevolmente le esigenze degli utenti e di assisterli nel portare a termine un compito.

Un passo importante verso la semanticità della ricerca è l'LSI (alias Latent Semantic Analysis),si tratta del tentativo di organizzare dei documenti (raccogliendoli appunto) con un'ottica semantica e con il minimo apporto 'umano', in sostanza facendo in modo che la macchina, o più precisamente l'agente software, una volta avute le "istruzioni" da un essere umano (o più probabilmente tanti esseri umani) possa autonomamente "capire" di che cosa trattano i documenti raccolti di "capirlo" con una "mentalità" umana e di fare sì che quando un uomo cerchi determinati concetti/significati possa trovare dei documenti che vi siano attinenti, a prescindere dagli specifici termini utilizzati nella domanda.

La tecnologia del web semantico applicata alle pagine HTML tende a trasformare esse direttamente in indici di oggetti RDF. Questo renderà una ricerca sul web sempre più simile a una ricerca all'interno di un grande database strutturato, mentre senza di essa una ricerca in internet è paragonabile a una ricerca all'interno di un libro.

Un motore di ricerca che indicizza pagine HTML trova una grande quantità di risultati ma molti sono generalmente inappropriati e non corretti; ad esempio possono esser trovati molti "falsi positivi" (pagine che contengono la parola chiave da noi inserita, ma non parlano di quello che

cercavamo, per esempio per problemi di polisemia) oppure “falsi negativi” (nei risultati della ricerca non compaiono pagine che sarebbero state di nostro interesse, ma che non contenevano esattamente la parola chiave da noi immessa, magari contenevano un suo sinonimo). Un motore logico, che lavora applicando determinate regole dei inferenza e deduzione, invece riesce a mandare in output solo risultati corretti ma non riesce a rovistare all’interno della grande massa di dati intrecciati tra loro in modo da costruire una risposta valida.

Un buon sistema dovrà quindi integrare insieme tra loro le capacità di ragionamento logico e di ricerca mediante indici, in modo da raccogliere i risultati migliori di entrambe le tecnologie; deve dunque essere in grado di cercare tutte le occorrenze di un determinato termine e, utilizzare le regole logiche per estrapolare solo e soltanto quelle utili alla soluzione della richiesta.

1.6.1 Semantic Search

Il Semantic Search è un applicazione del Web Semantico che tende a superare i limiti degli attuali motori di ricerca [8].
Il Semanc Search mira a aumentare e migliorare i risultati delle ricerche, effettuate mediante i motori di ricerca tradizionali, utilizzando dati presi dal Semantic Web. Per dati del Semantic Web intendiamo dati tradizionali con l’aggiunta di metadati che stabiliscono relazioni ben definite tra di essi.
Le tecnologie utilizzate dai sistemi di information retrieval tradizionale sono tutte fondate su teorie di calcolo dell’occorrenza di una determinata parola in un documento, mentre, quelle utilizzate dal Semantic Search, avendo a che fare con dati commentati e spiegati machine understandable, sono in grado di far eseguire ragionamenti logici alla macchina.

Nel mondo del Web Search possono essere distinte due diverse tipologie di ricerche:

- La prima tipologia prevede una ricerca di una semplice combinazione di parole chiave, le quali dovranno esser presenti all’interno dei documenti cercati. Per queste ricerche non è dunque necessaria un interpretazione del significato della frase cercata da parte della macchina.

- Nella seconda tipologia di ricerca, invece la combinazione delle parole denota un determinato oggetto, con un determinato significato nel mondo reale, sul quale l’utente ricerca informazioni. In questa tipologia l’utente non conosce documenti precisi sui quali sa di trovare il concetto specificato e nemmeno sa altri termini collegati a questo.

Per avere un buon motore di ricerca semantico è necessaria un infrastruttura di applicazioni che consenta di:

- *Rendere Semantiche le pagine web*. Questa operazione può essere effettuata mediante l'annotazione delle pagine effettuata con linguaggi basati sulle asserzioni (RDF). Con il linguaggio RDF si possono organizzare le informazioni a grafo; nel quale i nodi dono informazioni e gli archi sono collegamenti semantici.

- *Interrogare le risorse del Semantic Web creato*. Seguendo l'infrastruttura basata sui grafi prima introdotta, sarebbe utile un interfaccia in grado di, navigare i grafi corrispondenti alle informazioni cercate e mandare in output quelle che ci interessano.

Il principali miglioramenti auspicabili, utilizzando il Semantic Search, saranno utili per le ricerche della seconda tipologia indicata prima, e sono:

- Aggiungere alla lista di risultati, trovati mediante le tecniche di information retrieval classico, altri documenti del Semantic Web, trovati navigando i vari collegamenti semantici relativi al concetto ricercato.

- Visto che, una stringa di ricerca generalmente denota un concetto del mondo reale, sarebbe utile far capire al motore di ricerca quale sia questo concetto. Grazie a questa informazione, il motore sarà capace di capire il contesto nel quale viene effettuata la ricerca, quindi escluderà dai risultati quelle che non appartengono a quel contesto.

La prima prospettiva di miglioramento ottenibile mediante le ricerche semantiche, cioè l'incremento dei risultati trovati dai motori di ricerca tradizionali con l'aggiunta di risultati del Semantic Web, fa insorgere alcune problematiche.

1) Quali risultati visualizzare?

2) Come presentarli?

Una possibile soluzione è:

Seguendo il modello introdotto prima, creare un grafo dei risultati. In questo grafo, un nodo (chiamato ancora) è quello in cui viene trovata l'esatto concetto cercato e, a questa ancora, viene attaccato un sottografo che comprenderà i nodi collegati al nodo di partenza nel grafo sul quale è stata effettuata la ricerca.

Per quanto riguarda il secondo obiettivo, visto che, spesso una parola ha diversi significati, per risolvere la query al meglio sarà necessario capire quale concetto denota la richiesta effettuata.
Per risolvere questo problema il motore può valutare questi fattori:

- La frequenza dell'occorrenza dei vari significati.

- La presenza o meno di un profilo dell'utente, nel quale si possono mantenere i suoi campi di interesse. Sapendo i campi d'interesse dell'utente l sistema può dunque escludere alcuni significati possibili.

- Il contesto nel quale viene effettuata la ricerca. Come per il punto precedente un contesto della ricerca può escludere alcuni significati possibili.

Per avere una selezione precisa del significato della parola da cercare si potrebbe anche proporre all'utente un interfaccia grafica mediante la quale egli decide il significato più appropriato, tra quelli proposti dal sistema.

Un'altra soluzione possibile è invece quella di aggiungere, per ogni risultato trovato, un link mediante il quale l'utente può indicare al sistema che quello è il significato che cercava, dopodiché il sistema visualizzerà nuovamente i risultati togliendo quelli non coerenti al significato selezionato.

2 SEWASIE

2.1 Introduzione

SEWASIE, acronimo di Semantic Webs and AgentS in Integrated Economies [9], è un progetto di ricerca finanziato dalla comunità europea nell'azione IST (Information Society and Technology), coordinato dall'Università degli Studi di Modena e Reggio Emilia e si avvale della collaborazione di: Confederazione Nazionale dell'Artigianato e della Piccola e Media Impresa, CNA Servizi Modena s.c.a.r.l. (Italia), Università degli Studi di Roma "La Sapienza" (Italia), Rhenisch Westfaelische Technische Hochschule Aachen (Germania), The Victoria University of Manchester (Regno Unito), Thinking Networks AG (Germania), IBM Italia S.p.A (Italia), Fraunhofer-gesellschaft zur Forderung der angewandten Forschung eingetragener Verein (Germania), Free University of Bozen - Bolzano.

Lo scopo del progetto SEWASIE è l'implementazione di un motore di ricerca avanzato, che consenta un accesso personalizzato alle informazioni provenienti da sorgenti eterogenee, tramite una semantica di dati processabili direttamente dalle macchine, fornendo le basi per una sicura struttura di comunicazione basata sul web.

Il target di utilizzatori previsti per questo sistema sono in prevalenza le piccole e medie imprese, le quali mediante questo strumento potrebbero risolvere i problemi principali che riscontrano al momento della ricerca di informazioni in internet. Oggi le attività di business dipendono fortemente dall'avere le necessarie informazioni nei tempi prescritti; a volte però la ricerca è difficoltosa a causa di fonti distribuite in un gran numero di risorse differenti che portano una conoscenza frammentaria.

Le ricerche effettuate coi motori di ricerca tradizionali presentano infatti diversi problemi. Innanzitutto è molto difficile, in alcuni casi, esprimere ciò che si vuole ricercare mediante una semplice stringa inoltre, vengono presentati un numero elevato di risultati, molti dei quali completamente irrilevanti.
SEWASIE invece, mediante l'utilizzo di un interfaccia guidata dai termini di un ontologia di settore, permette di effettuare ricerche molto dettagliate. Questo alto grado di dettaglio viene raggiunto grazie alla presenza di informazioni strutturate.

Un altro obiettivo importante da raggiungere per le piccole e medie imprese è la capacità di comunicare con altri soggetti economici, ed avere una interazioni più completa e costante col mercato, che permetta una crescita professionale e economica. SEWASIE, mediante il tool di negoziazione, fornisce alle piccole e medie imprese uno strumento capace

di rendere la comunicazione, con altre imprese o con clienti, semplice e precisa.

In sintesi il sistema SEWASIE per le piccole e medie imprese mira a:

- Sviluppare uno strumento per trovare, in un ambiente multinazionale, le informazioni strategiche al momento giusto.
- Fornire dei servizi di monitoraggio e collegamento delle diverse informazioni, in un contesto di analisi dei rischi o dei competitor.
- Creare un meccanismo di comunicazione, basato sulle ontologie, per la negoziazione in un ambiente multilinguistico.
- Creare un linguaggio col quale integrare tra loro informazioni di tipo differente e poterle poi interrogare in modo uniforme, semplice e intuitivo.

2.2 Architettura

Tutti i tool e i metodi sono sviluppati per mantenere o creare un ontologia multilingua, e pongono le loro fondamenta sulle raccomandazioni degli standard del W3C (XML, RDF, RDF Schema, OWL), che sono le basi per i meccanismi di ricerca avanzati e per lo scambio di informazioni strutturate.

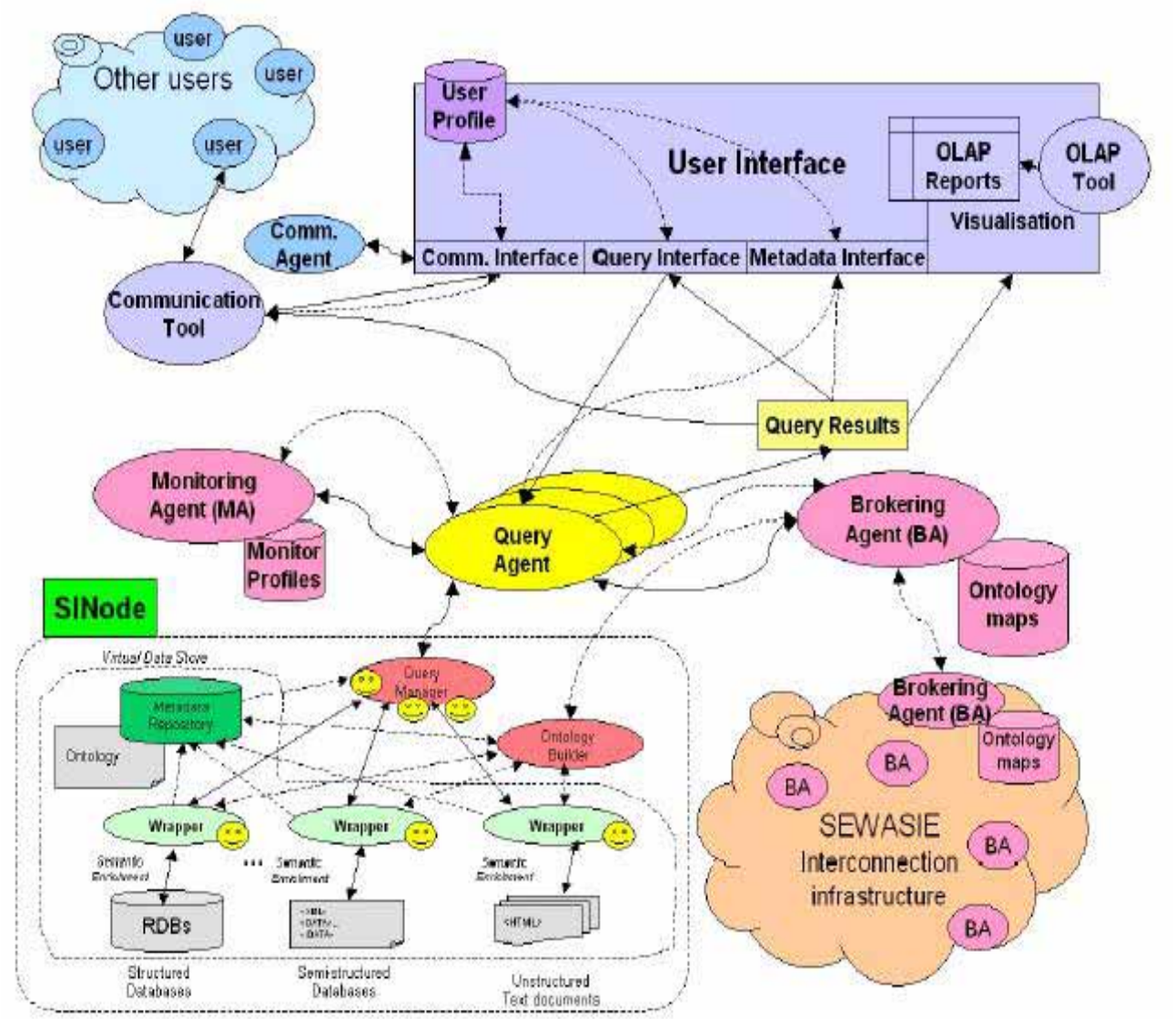
SEWASIE mira a creare un architettura aperta e distribuita, basata su agenti intelligenti (broker, mediator, wrappers) , capace di cambiare e adattarsi alle nuove situazioni ambientali e di interoperare con altri sistemi, offrendo comunque un punto di accesso sicuro all'utente [10].

L'architettura SEWASIE può essere divisa in tre differenti stati:

- Il *front-end*, che offre all'utente supporto per le ricerche di informazioni, la visualizzazione e l'interpretazione dei risultati anche mediante strumenti che utilizzano la tecnica OLAP e la negoziazione elettronica delle informazioni.
- Il livello di definizione degli information node e di supporto nella ricerca di query. A questo livello abbiamo la gestione delle query ricevute dal front-end e un infrastruttura per l'arricchimento semantico e il wrapping dei nodi informativi.

- Il livello degli agenti intermedi che effettuano le interconnessioni tra i nodi informativi e li ageti delle ricerche.

Essendo un architettura distribuita si basa su una rete chiamata SEWASIE Virtual Network (SVN).



Nella figura sono rappresentati i principali componenti della rete:

- Interfaccia utente(con i Comunication e Monitoring Agent)
- I Query Agent
- I Brokering Agent
- I SEWASIE Information node

Un componente architetturale fondamentale di SEWASIE sono I *SEWASIE Information Node* (*SINode*), cioè i nodi virtuali che sintetizzano l'ontologia di un dominio sulla base di sorgenti di dati eterogenee residenti su sorgenti web integrate. I SINode raggruppano i moduli, basati sui mediatori, che lavorano per definire, e mantenere, un nodo di informazioni pubblicato in rete.
In ogni SINode abbiamo un Ontology Builder, un Query Manager e un Virtual Data Store che comprende tutte le sorgenti di informazione, i wrapper e un metadata repository.

Le sorgenti di informazione sono di tipo eterogeneo infatti possono essere strutturate, semistrutturate, o non strutturate. Per ogni risorsa informativa esiste un Wrapper che opera da mediatore tra essa e il SINode.

Ogni Wrapper deve tradurre la struttura della sorgente informativa locale nel linguaggio del SINode, viceversa deve supportare la traduzione di query dal query language del SINode a quello della sorgente locale sottostante.

L'Ontology Builder è un insieme di funzionalità che supporta la creazione e il mantenimento della Vista Globale Virtuale (GVV) del SINode. Per creare la GVV l'Ontology Builder interagisce coi Wrapper delle sorgenti dati e coi designer del SINode, esegue l'integrazione poi salva l'ontologia risultante all'interno del Metadata Repository.

Nei Metadata Repository vi sono dunque le ontologie relative al SINode e tutte le conoscenze che permettono di creare relazioni semantiche tra esso e i suoi vicini.

Il Query Manager è il set di funzionalità necessarie per risolvere le query ricevute. L'approccio utilizzato dal Query Manager è il Global As View (GAV). Mediante questo approccio una query effettuata sulla GVV del SINode viene decomposta in tante piccole query effettuate dai wrapper sulle sorgenti di dati collegate.

L'architettura dei SINode è distribuita quindi, questi componenti, per poter interagire tra loro, devono avere dei protocolli di comunicazione; i protocolli selezionati sono quelli della famiglia TCP/IP in quanto sono universalmente supportati a tutti i livelli.

I *Query Agent* sono gli agenti che trasportano le richieste degli utenti, dalle interfacce utente ai vari SINode, e hanno il compito di risolvere le query interagendo con i vari Brokering Agent. Partendo da un determinato SINode possono accedere ad altri SINode, mediante richieste ai Brokering Agent, raccogliere risposte parziali e integrarle poi tra loro.

I *Brokering Agent* sono responsabili di mantenere la vista d'insieme delle informazioni impiegate dalla rete di SEWASIE (SVN), devono quindi contenere le relazioni semantiche tra i vari SINode. Sono inoltre responsabili della gestione della creazione di un nuovo SINode e conseguentemente dell'aggiornamento della SVN.
Un'altra funzione fondamentale dei Brokering agent è quella di rispondere alle richieste ricevute dagli altri componenti. Un Brokering Agent, quando riceve una query dal Query Agent, controlla che le informazioni ricercate siano riconosciute dalla propria ontologia e quindi siano presenti nelle sorgenti da lui gestite. In questo caso ritorna al Query Agent le informazione necessarie per poter riscrivere la query da porre al Query Manager del SINode interessato. Se inoltre il Brokering Agent riscontra che la query è riconosciuta anche dall'ontologia di un nodo circostante, esso dirige il Query Agent anche al Brokering Agent che gestisce questo nodo. Una volta trovati tutti i risultati dai diversi SINode il Query Agent procede alla fusione di essi e alla presentazione all'utente.

L'interfaccia utente è il gruppo di moduli che lavorano insieme per offrire un'interazione, tra l'utente e il sistema SEWASIE, deve esser semplice e immediata da utilizzare. L'interfaccia utente per la creazione delle query istanzia un nuovo Query Agent il quale provvederà a effettuare la ricerca.

I *Monitoring Agent* si occupano del monitoraggio di risorse informative compatibili con gli interessi degli utenti, definiti nei profili di monitoring. Per fare quest'attività devono impostare, ad intervalli regolari, i Query Agent per trovare le informazioni d'interesse dell'utente.

I *Comunication Agent* si occupano di trovare e contattare potenziali partner di business iniziando quindi la negoziazione elettronica delle informazioni.

2.3 Scenari d'utilizzo

Lo scenario più elementare di utilizzo di questo sistema è quello che vede un utente in una workstation che cerca informazioni su un argomento specifico. L'utente fa la richiesta di informazioni al sistema e l'interfaccia utente traduce questa in una query e, tenendo conto del contesto di utilizzo e delle caratteristiche dell'utente, imposta il Query Agent per la ricerca.

Il Query Agent si connette alla rete dei Brokering Agent di SEWASIE e gli chiede dove trovare le informazioni di interesse. A questo punto il Brokering Agent, o instrada il Query Agent verso uno o più SINode che contengono le informazioni ricercate, oppure gli indica un altro Brokering Agent a cui chiedere informazioni. Una volta arrivato nel SINode ricerca le informazioni al suo interno e, dopo averle trovate, le integra tra loro e le visualizza formattate all'utente.

E’ inoltre possibile effettuare una richiesta di monitoraggio a lungo termine. Mentre la richiesta precedente ritornava gli eventuali risultati subito, quest’ ultima tipologia non ha precisi limiti di tempo, infatti viene effettuata ripetutamente per rintracciare eventuali cambiamenti e aggiornamenti delle informazioni all’interno del dominio di competenza. Per tenere monitorato un determinato dominio di competenza il Monitoring Agent pianifica delle ricerche da far svolgere al query agent, a intervalli di tempo prestabiliti.

Un'altra famiglia di funzioni d’interesse del SEWASIE sono quelle relative alla creazione, modifica e cancellazione di un nodo informativo.

La creazione di un nuovo nodo è l’acquisizione di nuove risorse di informazioni e l’organizzazione di esse in un unità di informazione SINode. Questo processo viene svolto semiautomaticamente e ha i seguenti obiettivi.

- Configurare i Wrappper appropriati, per accedere ai dati e alle loro strutture.
- Creare un ontologia, che sia una Global Virtual View (GVV), e le descrizioni dei mapping tra essa e le altre risorse.
- Configurare il Query Manager, per gestire in modo ottimale le query, all’interno di questo nodo.
- Notificare alla rete dei Brokering Agent il nuovo nodo (o inizializzare un nuovo Brokering Agent)

L’aggiornamento di un nodo esistente, effettuato mediante modifiche strutturali, ha i seguenti obiettivi.

- Cambiare l’ontologia.
- Cambiare la struttura delle risorse, che implica adattamenti a livello del nodo.
- Aggiunta/cancellazione di una risorsa, che implica un cambio dell’ontologia e adattamenti a livello dei Brokering Agent.

La cancellazione di un nodo causa la cancellazione di tutti i riferimenti a quel nodo in tutti i Brokering Agent della rete che vi erano collegati, e la seguente terminazione delle attività del nodo.

2.4 Interfaccia grafica SQoogle

Alla pagina http://sewasie.ing.unimo.it:8080/sewasie/home.jsp

Il sistema SEWASIE mette a disposizione dell'utente un interfaccia grafica mediante la quale lo supporta nell'attività di ricerca di informazioni.

Questa interfaccia è basata sulle ontologie create e le richieste sono strutturate come un albero di classi e proprietà.

Selezionando il pulsante ricerca viene visualizzata la schermata dei domini

con questa si va a selezionare, tra le varie ontologie presenti, quella relativa al settore di interesse della ricerca.

Una volta selezionata l'ontologia (in questo caso quella tessile con Brokering agent) è necessario selezionare che tipo di informazione cercare tra quelle presenti mediante la schermata seguente.

In questa vogliamo cercare un venditore quindi selezioniamo supplier. Successivamente si passa alla schermata di composizione della query.


SEWASIE Query Tool - Mozilla Firefox
Interfaccia SEWASIE
Inizio
Ricerca
Monitoraggio
Rapporto OLAP
Visualizzazione
Negoziazione
Conclusione
SEWASIE
Documentazione
SQoogle
Domini Informativi
Inizio Ricerca
Composizione
Risultati
Configura
Scegli il punto di partenza della richiesta.
Nome
Descrizione
category
Type of companies
enterprise
Companies offering services and products in the "textile" sector
goods
Textile-related products
supplier
Suppliers of textile-related goods and services
Completato


SEWASIE Query Tool - Mozilla Firefox
Interfaccia SEWASIE
Inizio
Ricerca
Monitoraggio
Rapporto OLAP
Visualizzazione
Negoziazione
Conclusione
SEWASIE
Documentazione
SQoogle
Domini Informativi
Inizio Ricerca
Composizione
Risultati
Configura
supplier
XML
Sostituisci con generalizzazione o specializzazione
aggiungi una proprietà
Cerca con SQoogle
Mi sento fortunato
Completato

Nella schermata di composizione della query è possibile formulare passo-passo la richiesta di informazioni iniziando dal punto di partenza selezionato precedentemente in “Inizio Ricerca”. E’ possibile sia aggiungere proprietà che specializzare o generalizzare una classe inserita, i quali andranno a comporre l’albero della ricerca. Una volta terminata la composizione è possibile premere entrambi i bottoni sottostanti:

Con cerca “Cerca con Sqoogle” si dovranno poi selezionare i campi da visualizzare in output mentre “Mi sento fortunato” imposta la pagina di output con tutte le proprietà selezionate.

Nella nostra richiesta andiamo a inserire nome e telefono di un supplier , e quindi la finestra dei risultati apparsa è la seguente:

In essa abbiamo nella parte alta l’albero della richiesta effettuata e sotto la tabella dei risultati con i campi selezionati.

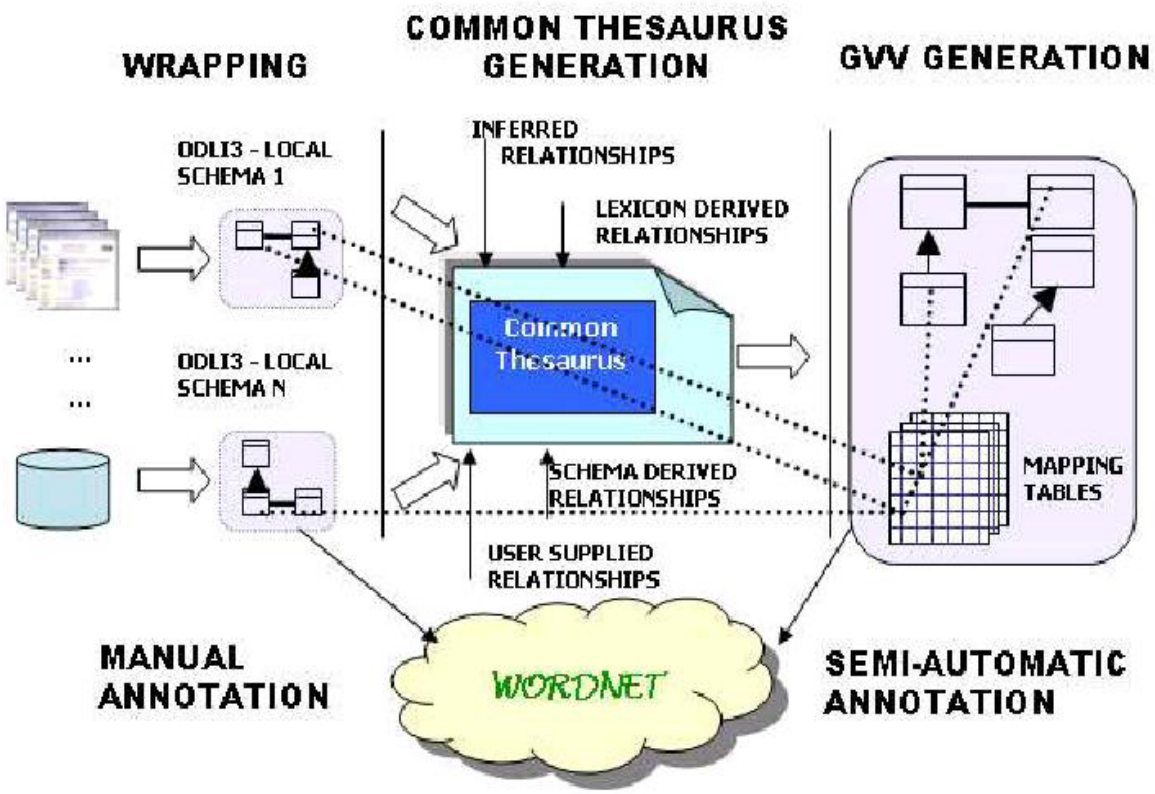
2.5 MOMIS

MOMIS è un sistema basato su un’architettura a mediatore che sfrutta l’estensione ODLi3 dello standard Object Definition Language per generare una GVV (vista virtuale globale) che esprime un’integrazione dello schema di sorgenti di dati eterogenee[11]. Esso provvede a tradurre in una base

comune la struttura di varie sorgenti indipendentemente dalla loro localizzazione ed espressione. MOMIS è stato progettato per fornire un accesso integrato ad informazioni eterogenee memorizzate sia in database di tipo tradizionale (e.g. relazionali, object-oriented) o file system, sia in sorgenti di tipo semistrutturato(XML).

Successivamente accompagna il progettista nella procedura di integrazione di queste sorgenti 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.

L'altro compito fondamentale di MOMIS è quello di rendere trasparente l'accesso ai dati, permettendo all'utente di interagire direttamente con lo schema globale virtuale. Mediante la GVV infatti, l'utente che ricerca informazioni non si deve preoccupare di come è formata la struttura complessa dei dati sottostanti in quanto effettua delle interrogazioni considerandola come un insieme omogeneo. Quando un utente invia una query a MOMIS il modulo mediatore Query Manager (QM) produce un insieme di sotto-query da inviare alle sorgenti coinvolte dalla query. Ciò avviene attraverso le fasi di ottimizzazione semantica e formulazione del piano di interrogazione.



Il procedimento per la creazione della GVV prevede diverse fasi [12]:

- *Estrazione delle sorgenti locali:* wrapper dedicati per ciascuna struttura delle sorgenti ne estraggono lo schema e lo traducono in ODLi3.

- *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. Le relazioni possono essere di diverso tipo: *schema-derived* derivate automaticamente dalle sorgenti, *lexicon-derived* derivate dalle annotazioni effettuate sul lessico inter-schema, *designer-supplied* create dal progettista su conoscenze specifiche di un determinato dominio, *inferred* derivate da un motore inferenziale basato sulle Description Logics.

- *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 assegnato agli elementi globali che compongono la GVV. Il procedimento è analogo a quello utilizzato per l'annotazione delle sorgenti locali.

MOMIS fornisce un effettivo supporto al progettista dell'integrazione grazie allo strumento SI-Designer, un'interfaccia grafica che guida il progettista nel percorso di integrazione.

3 NetWatcher®

In questo capitolo verrà illustrato il sistema NetWatcher, il quale è un software proprietario di Expert System SPA. Il software è stato testato da me e Marco Tosi nell'ambito del tirocinio formativo svolto alla CNA servizi Modena.

3.1 La piattaforma COGITO®

La semantica consente una completa e rapida organizzazione delle informazioni non strutturate e rappresenta oggi la risposta più innovativa ed efficace per supportare l'OSINT (Open Source Intelligence) e risolvere qualsiasi problema di gestione della conoscenza: ricerca, filtraggio, classificazione, correlazione, mining e discovery.

Cogito è la piattaforma linguistica, sulla quale si basano molti strumenti software distribuiti da Expert System[13], tra i quali vi è il NetWatcher.

Cogito è in grado di elaborare in modo intelligente la conoscenza contenuta nei testi scritti nel linguaggio di tutti i giorni, per l'estrazione di significati, la comprensione del linguaggio naturale, la traduzione, la condivisione e la circolazione della conoscenza.

Attraverso le tecnologie linguistiche cattura tutti gli aspetti

- Strutturali
- sintattiche
- lessicali

di un testo, attraverso un approfondita analisi e disambiguazione di tutti gli elementi che incontra.
Il risultato dell'elaborazione di Cogito è una mappa cognitiva e concettuale, vale a dire una rappresentazione strutturata degli aspetti qualificanti del flusso di dati non strutturati in ingresso. La strutturazione dell'output consente ogni tipo di trattamento automatico degli elementi più significativi dei documenti.

Cogito si compone di diversi moduli, ognuno dei quali compie un attività necessaria alla comprensione automatica dei testi:

- una rete semantica
- un parser
- un motore linguistico

- un sistema di disambiguazione

La rete semantica di Expert System è chiamata Sensigrafo® ed è un grafo che contiene la rappresentazione concettuale della nostra lingua. In essa sono presenti informazioni relative a relazioni fra oggetti, specifiche sull'appartenenza ad un certo dominio lessicale e informazioni sulla frequenza d'utilizzo. Nel Sensigrafo ogni synset (significato) rappresenta un nodo della rete semantica ed è collegato agli altri da relazioni semantiche in una struttura gerarchica.

Il parser esegue una completa analisi morfologica, grammaticale e sintattica della frase, esso è costruito ad hoc per l'interrogazione della rete semantica.

Il motore linguistico serve per effettuare l'interrogazione della rete semantica quindi per istituire un confronto tra essa e gli elementi trovati nel testo.

Il disambiguatore è il programma che analizza singole frasi, o interi documenti e distingue il giusto significato per ogni elemento che incontra , eliminando ogni possibile ambiguità.

3.1.1 Procedimento di analisi linguistica del testo

Per poter effettuare un procedimento di analisi automatica del testo è necessario creare delle precise regole di disambiguazione. Queste regole di disambiguazione vengono ricavate da un corpus, cioè una serie di documenti rappresentativi della totalità degli argomenti che si possono incontrare in un testo. Il corpus utilizzato da Expert System comprende testi tratti da enciclopedie, quotidiani, romanzi, riviste ecc.

La procedura di analisi del testo comincia con la lettura di esso e con l'estrazione degli elementi dalle frasi. Questa estrazione degli elementi viene effettuata con il tagging, cioè la marcatura di alcune porzioni di testo alle quali si assegna una data proprietà o caratteristica.

Una volta estratti gli elementi vanno confrontati con i significati presenti nella rete semantica e va quindi cercato quello appropriato da assegnargli.

A questo punto entra in gioco il disambiguatore che, per capire il significato da dare all'elemento, utilizza prima un analisi morfologico/grammaticale della frase, poi interagendo col Sensigrafo stabilisce la priorità da dare alle parole o a gruppi di parole all'interno della frase. Viene poi effettuata un analisi logica e sintattica della frase, con la quale si va a capire quale sia il ruolo logico della parola da disambiguare. In fine data la coppia parola/ruolo il sistema andrà a decidere quale sia il significato da assegnargli analizzando i seguenti aspetti:

- frequenza d'uso
- dominio
- attributi sugli aggettivi, sostantivi e verbi
- informazioni contestuali.

Il disambiguatore si presta quindi perfettamente a applicazioni di ricerca di informazioni; grazie ad esso infatti si potrebbero eliminare tutte le occorrenze di una data parola che non corrispondono al determinato significato.

3.2 Cos'è NetWatcher®

Nell'ambito dei sistemi di comunicazione, internet riveste nelle sue varie forme (notizie on line, forum, siti settoriali ecc.) un'importanza sempre maggiore come strumento privilegiato di informazione. La rete,grazie alla dinamicità dei contenuti, alla sua capillarità ed alla interattività che offre agli utenti, è divenuta, infatti, la fonte aperta per eccellenza.

I giornali on line sono aggiornati in tempo reale e, in modo ormai costante, anticipano i contenuti che saranno ripresi dalle edizioni cartacee del giorno dopo, nei forum si commentano fatti e personaggi, nei siti web aziendali si annunciano in anteprima prodotti e strategie, nei portali di settore sono presenti articoli, editoriali e recensioni.

In alcuni casi, la quantità di dati disponibili è però talmente elevata da scoraggiare le attività di monitoraggio e di rielaborazione delle informazioni.A questo scopo, Expert System ha implementato NetWatcher, un'applicazione basata sulla piattaforma di analisi semantica COGITO in

grado di fornire un concreto supporto all'attività di analisi e reperimento dei dati da internet.
Gli obiettivi principali che si pone il sistema sono:

- innanzitutto la riduzione dei costi e soprattutto dei tempi di reperimento delle informazioni da internet.

- Permettere un attività di monitoraggio personalizzata delle fonti di interesse, come ad esempio l'analisi della concorrenza, che permette di analizzare il comportamento dei principali competitors.

- Invio tempestivo all'utente del resoconto dei risultati riscontrati dalle ricerche.

- Sfruttando l'analisi semantica e concettuale delle informazioni limitare l'overload di risultati.

- Personalizzazione delle ricerche, mediante la creazione di profili, che consentiranno di ricevere soltanto notizie pertinenti agli argomenti dichiarati.

- Archiviazione su disco dei risultati trovati per ovviare alla volatilità del web.

NetWatcher è un sistema multi-utente di news feeling, il quale tiene aggiornati i suoi utenti quando su Web, in siti indicati dagli utenti stessi, appaiono documenti interessanti. Per definire se un documento è interessanti o meno vengono utilizzati dei filtri che verranno presentati successivamente.
Un interfaccia intuitiva e flessibile permette all'utente di creare profili d'interesse, i quali sono utilizzati dal sistema per impostare le ricerche dei contenuti in rete.
La creazione di un profilo consente all'utente di impostare i seguenti parametri:

- nome del profilo, che funge da titolo;

- modalità di notifica: consultazione tramite interfaccia web e/o ricezione dei risultati via mail;

- fonti, ovvero le URL internet dalle quali far partire lo spider per la ricerca;

- parametri che regolano l'attività di crawl dei siti stessi: periodicità con cui i siti indicati devono essere visitati e l'estensione del crawl (quali link seguire, fino a quale profondità, il numero massimo di documenti da leggere, ecc.).

- filtri che NetWatcher applica al contenuto dei documenti letti da Web per determinare se il singolo documento è da ritenersi interessante e se, quindi, ne deve essere notificata l'esistenza all'utente.

Ogni filtro viene sviluppato combinando a piacere questi parametri di ricerca:

- argomenti
- concetti
- keywords

Sfruttando la piattaforma semantica COGITO i risultati vengono categorizzati automaticamente; Il primo parametro quindi consente di scegliere tra una tassonomia di oltre 600 categorie uno o più argomenti di interesse.

L'utente ha a disposizione un wizard col quale può selezionare l'esatto concetto che cerca, dove per concetto si intende l'esatto significato di una data parola. Questa selezione viene effettuata basandosi sui differenti significati presenti sulla rete semantica.

L'ultimo parametro è la classica parola chiave, che viene comunemente utilizzata dai motori di ricerca classici, a differenza di essi è però possibile considerare essa come lemma cioè includere nei risultati della ricerca tutte le forme flesse di essa.

Per comodità è possibile distribuire su più filtri, combinati tra loro mediante operatori booleani, criteri di selezione particolarmente complessi.

NetWatcher visita le fonti scelte con periodicità impostata dall'utente e designa un documento come nuovo solo quando esso non risulta visitato in precedenza oppure quando è cambiato in modo rilevante, vale a dire quando il nocciolo testuale del documento stesso ha subito variazioni.

Gli esiti delle ricerche vengono comunicati, in tempo reale, all'utente mediante l'invio di una mail (push) e saranno poi archiviati per successive consultazioni. NetWatcher permette cioè di impostare un sistema di alert che segnala il rilevamento di contenuti corrispondenti al proprio profilo di interesse.NetWatcher mantiene tutte le versioni dei documenti letti fino a quando non raggiungono l'età stabilita per l'operazione di garbage collection.L'utente può effettuare in qualsiasi momento ricerche nell'archivio storico, usando criteri inseriti al momento eventualmente combinati con quelli memorizzati e associati alle operazioni di aggiornamento automatico (filtri) impostati nel profilo in cui viene effettuata la ricerca.

3.3 Architettura

NetWatcher è costituito da componenti che possono essere dislocati su server diversi:

Back-end

- Searchbox linguistico
- Server GSL di analisi linguistica con Dispatcher

Front-end

- Applicazione Web
- Pannello di controllo searchbox

3.3.1 Searchbox

Searchbox è un motore di indicizzazione e ricerca di documenti Web (crawler). Effettua il crawl dei siti indicati nei profili definiti dagli utenti, indicizzando e archiviando i documenti man mano letti.

La funzione fondamentale di searchbox è rispondere alle richieste provenienti dal front-end, vale a dire:

- query sugli indici dei documenti;
- richieste di apertura di versioni archiviate di un documento;
- chiamate per la gestione dei profili;

searchbox è accessibile al front-end come servizio Web: questo significa che è possibile interagire con esso tramite un semplice client SOAP.

3.3.2 Server GSL analisi linguaggio con Dispatcher

Server GSL è un servizio TCP/IP, che fornisce a searchbox informazioni di carattere tematico (argomenti trattati) e linguistico (concetti espressi, lemmi) frutto dell'analisi linguistica dell'eventuale nocciolo testuale di ciascun documento.

Searchbox e Server GSL sono entrambi client di Dispatcher. Questo è un servizio TCP/IP che mette in comunicazione le due parti:

- ricevendo le richieste da searchbox (documenti da analizzare);
- inoltrandole a Server GSL;
- ricevendo da questo i risultati (flusso XML contenente le informazioni tematiche e linguistiche) e trasmettendoli a Searchbox.

Tutte le parti coinvolte si scambiano messaggi TCP/IP in formato prestabilito attraverso socket. Queste permette di distribuirle su elaboratori diversi collegati in rete.

3.3.3 Applicazione Web

L'applicazione Web rappresenta lo strumento attraverso cui gli utenti interagiscono con NetWatcher. Si tratta di pagine dinamiche ASP integrate da componenti COM lato server che permettono all'utente di scegliere dati linguistici dalla rete semantica Sensigrafo per usarli all'interno di criteri di selezione.

Attraverso l'applicazione Web, l'amministratore definisce gli account, mentre gli utenti definiscono e configurano i propri profili oltre a richiederne il push e a poter effettuare ricerche dirette in archivio. I dati degli account e quelli dei profili non strettamente attinenti l'attività di crawl e l'indicizzazione sono memorizzati in un database gestito direttamente dall'applicazione Web.

3.3.4 Pannello di controllo Searchbox

Si tratta di uno strumento interattivo che permette di monitorare searchbox.
Consente di visualizzare lo stato dei profili, di attivarne o sospenderne l'attività di crawl e indicizzazione, nonché di ottenere log delle diverse attività, tra qui le query sottoposte dagli utenti in modo automatico o interattivo.

Tutti i componenti del front-end (applicazione Web e Pannello di controllo searchbox) comunicano col componente searchbox del back-end tramite SOAP. Le due parti si scambiano buste SOAP, cioè normali pacchetti HTTP contenenti una forma concordata di XML.

3.4 Interfaccia grafica

Per prima cosa l'utente deve effettuare il login al sistema inserendo lo username e la password che gli sono stati assegnati. Una volta autenticato l'accesso si hanno a disposizione due sezioni dell'applicativo: la principale e quella dei profili.

3.4.1 Sezione profili

Nella sezione profili vengono creati, gestiti e modificati i vari profili di indicizzazione che implementerà NetWatcher.

Ogni profilo di NetWatcher crea in locale un archivio all'interno del quale vengono memorizzate tutte le pagine indicizzate. All'interno di ogni archivio è poi possibile effettuare ricerche mediante l'interfaccia presentata nella sezione principale dell'applicazione. E' inoltre possibile impostare una ricerca nel momento stesso in cui viene effettuato un crawl, in modo da creare un archivio già filtrato coi criteri desiderati.

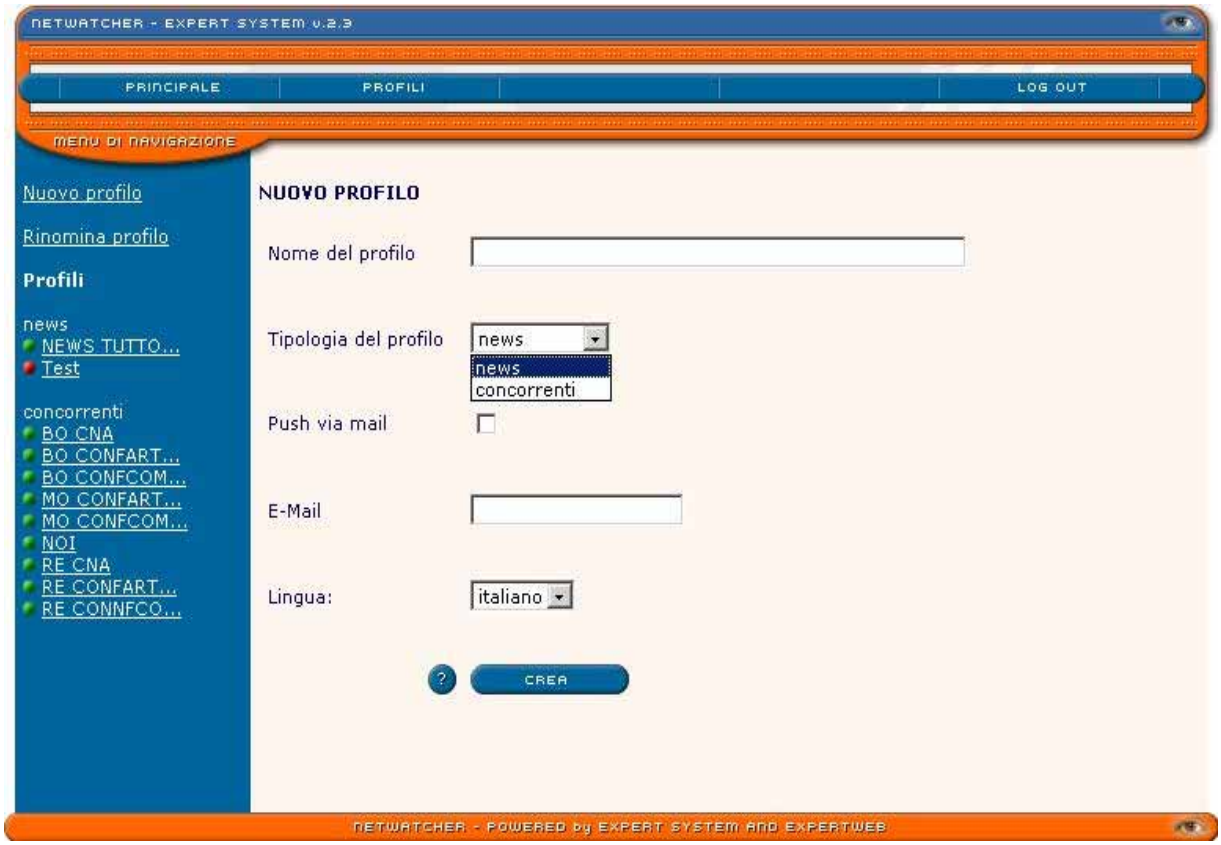
3.4.1.1 Creazione nuovo profilo

L'interfaccia della sezione profili, alla sua apertura, visualizza, al centro, la schermata per l'inizializzazione di un nuovo profilo e, nella colonna di sinistra, vi sono il collegamento alla schermata con cui si rinominano i profili e un elenco cliccabile di tutti i profili creati dall'utente suddivisi per tipo. Di fianco ad ogni profilo creato vi è un semaforo che indica se il profilo è stato attivato; cliccando su uno di essi si passa alla schermata con la quale si possono modificare i vari parametri di ricerca ed indicizzazione del profilo stesso.

In questa implementazione sono stati previsti due tipi di profilo, News e Concorrenti, la cui selezione viene effettuata mediante una combobox; la differenza tra un tipo e l'altro sta nella profondità a cui si spinge il sistema nell'esame dei link contenuti nelle pagine.

Per i profili di tipo News, adatti a siti che cambiano spesso contenuto, il sistema si ferma alle pagine di partenza indicate nel profilo e a quelle raggiungibili coi link contenuti in esse. La scansione del sito si ferma in superficie, cioè al primo livello di link. Profili di questo tipo possono avere frequenze di refresh elevate (anche più volte al giorno).

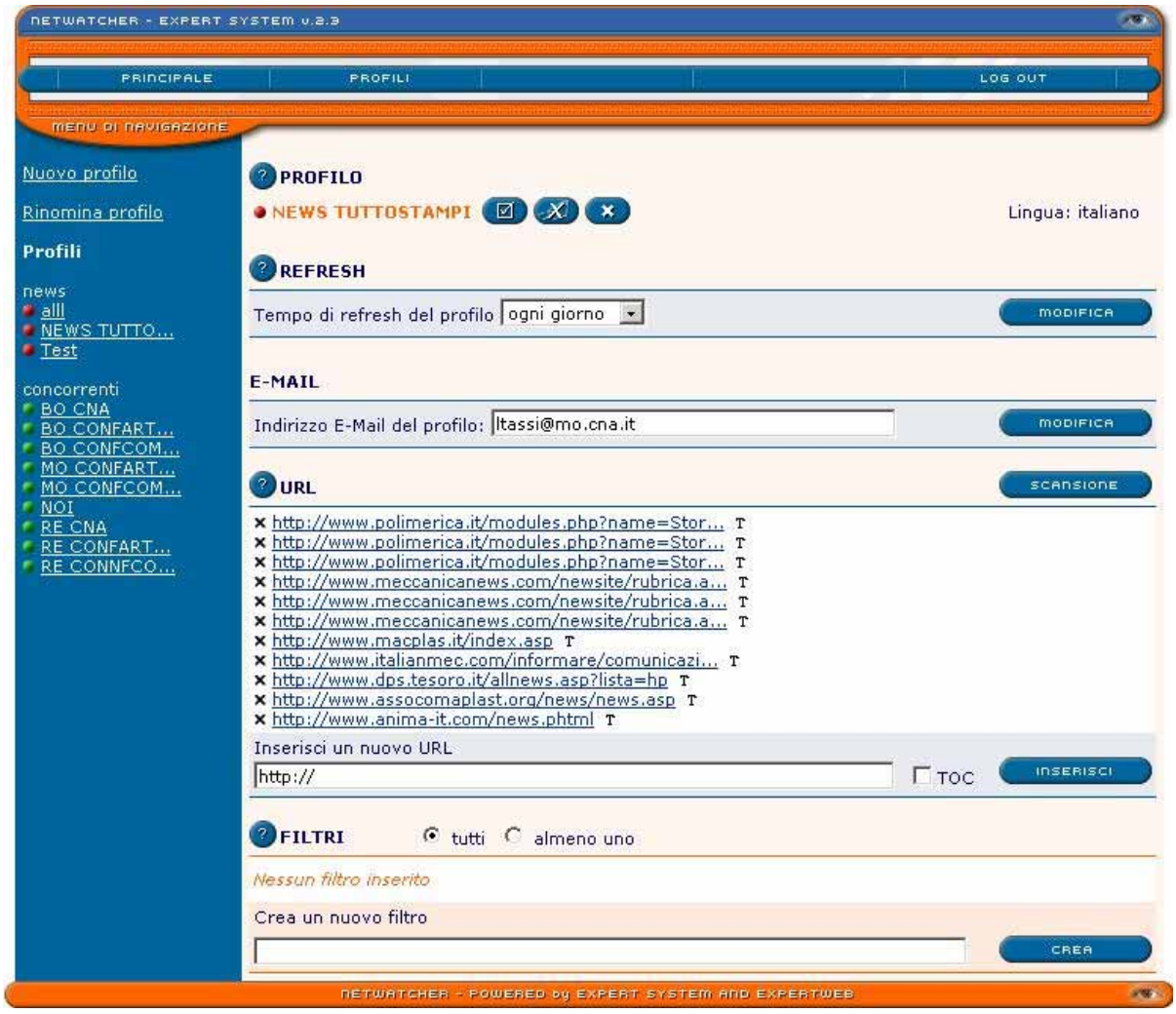
I profili di tipo Concorrenti, invece, sono adatti a siti in cui la maggior parte del contenuto cambia raramente. Per profili di questo tipo il sistema legge le pagine di partenza, ne segue i link contenuti, esamina le pagine collegate, ne segue i link e così via fino all'esame completo del sito. La scansione raggiunge la massima profondità possibile. Profili di questo tipo si prestano a frequenze di refresh basse (una volta al giorno o più raramente).



L'opzione "Push via mail" consente di ricevere i risultati dei crawl alla propria casella e-mail con tempestività e completezza. All'indirizzo di posta, indicato nella text box, ogni qual volta verranno trovati aggiornamenti o nuove pagine che soddisfano le condizioni di un determinato profilo si riceverà un messaggio con un elenco di pagine strutturato analogamente al "push interattivo" che si trova nella sezione principale dell'applicazione.

La combobox della lingua consente di decidere se far svolgere al NetWatcher l'analisi semantica dei contenuti della pagina in lingua inglese o italiana. Sono infatti disponibili due Sensigrafi per l'analisi semantica dei contenuti. Questa selezione è utile per un eventuale ricerca per concetti, categorie o lemmi, in quanto per la ricerca di keyword non è necessaria l'analisi del significato semantico del contenuto delle pagine, ma basta una ricerca full text.

Una volta premuto crea dalla finestra di inizializzazione profilo si passa a un'altra schermata di modellazione del profilo. Questa è anche la pagina che appare quando si va a selezionare un profilo tra quelli presenti sulla colonna di sinistra



Nella parte superiore della pagina, al fianco del nome del profilo creato e del piccolo semaforo che indica lo stato del profilo, vi sono tre pulsanti che servono rispettivamente per:

- attivare/disattivare il profilo. L'attivazione inizializza e programma il crawler sottostante, impostando automaticamente i vari parametri del pannello di controllo Searchbox;
- attivare/disattivare il push via mail;
- eliminare il profilo.

Il tempo di refresh, che viene selezionata tra diversi intervalli con una combobox (ogni modifica alla selezione all'interno deve inoltre essere attivata mediante il pulsante modifica), è la periodicità con cui il sistema scandisce le risorse indicate dal profilo, per verificare se ci sono novità da comunicare all'utente.

Allo scadere del tempo indicato, il sistema visita tutti gli URL definiti. Se trova link ad altre pagine visita anche queste, proseguendo in questo modo fino ad una profondità determinata dal tipo di profilo ("superficiale"

per profili di tipo News, "fino al fondo" per profili di tipo Concorrenti). E' inoltre possibile impostare la profondità del crawl programmando direttamente il crawler dal pannello di controllo Searchbox.

La durata della scansione dipende quindi da diversi fattori tra cui:

- il tipo di profilo
- il numero di fonti
- la velocità di accesso alle pagine dei singoli URL("velocità" del sito)
- il numero di link contenuti nelle pagine (complessità del sito)

Se per il profilo è stato abilitato il push via e-mail, allora il tempo di refresh rappresenta anche la frequenza con cui arriveranno le e-mail coi risultati.

Un URL (Uniform Resource Locator) è l'indirizzo di una pagina Web.
In questa area si dovrà quindi elencare le pagine, che nel loro insieme, costituiscono una fonte di potenziali informazioni e che NetWatcher deve tenere sotto controllo, filtrandone i contenuti in modo da comunicare all'utente solo le novità veramente interessanti e non il "rumore".

È quindi possibile indicare:

- l'home page di un sito: http://www.sito.it
- la sezione di un sito: http://www.sito.it/sezione
- una pagina qualunque: http://www.sito.it/sezione/pagina.html

URL particolarmente lunghi possono apparire "troncati" sullo schermo, ma posizionando il cursore su di essi appare un tooltip che li mostra per esteso. Facendo clic su un URL la pagina corrispondente si apre in nuova finestra del browser Internet.

La casella di controllo TOC ("Table Of Contents", sommario) serve ad indicare gli URL che corrispondono a pagine-sommario. Si tratta di pagine interessanti per i collegamenti ad altre pagine, ma non per il testo in esse contenuto. Selezionando la casella TOC la pagina corrispondente all'URL non verrà quindi visualizzata tra i risultati di ricerche in archivio e push.

3.4.1.1.1 Scansione

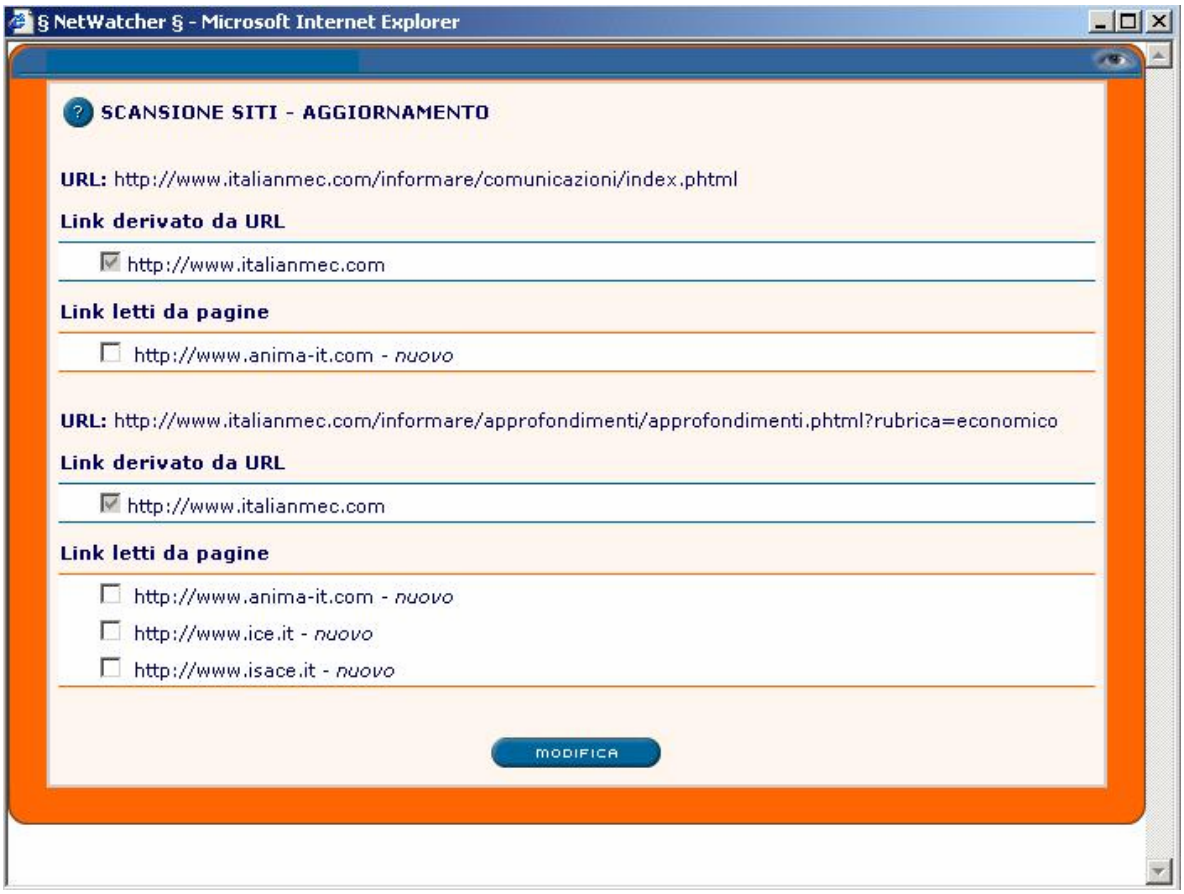
Il processo di scansione è l'attività fondamentale con cui NetWatcher visita i siti corrispondenti agli URL. Partendo dai documenti Web direttamente associati agli URL, NetWatcher segue i link che eventualmente trovati in tali documenti per raggiungerne altri. Tra i link trovati nei documenti

vengono seguiti solo quelli che iniziano in un certo modo: in particolare vengono seguiti solo i link che iniziano come i link-modello elencati in questa pagina.

I link-modello colorati di grigio sono derivati automaticamente dagli URL di partenza e quindi non possono essere modificati o rimossi. Questi rappresentano la parte di URL che tutte le pagine, per essere visitate, dovranno avere. Quindi dato un URL di partenza, tutti i documenti raggiungibili da esso e che hanno, come prima parte dell'indirizzo il link modello indicato, vengono sempre raggiunti (quindi letti e indicizzati).

È possibile tuttavia fare in modo che il processo di scansione segua anche i link che hanno un indirizzo non conforme al modello indicato prima. La seconda visualizzazione di questa pagina, chiamata con la pressione del pulsante aggiorna, mostra, se definiti, tutti i link presenti nelle pagine di partenza, non conformi ai link-modello in modo che l'utente possa scegliere, spuntando le appropriate opzioni, quali debbano essere seguiti in fase di scansione del profilo.

Con queste selezioni si cerca di modellare il tragitto da far percorrere al crawler, anche se è consentito impostare solo i link da seguire all' esterno del sito di partenza; per effettuare una modellazione precisa del tragitto è necessario impostare direttamente il crawler dal pannello di controllo Searchbox.



3.4.1.1.2 Filtri

Se il ruolo degli URL nel profilo è indicare "dove" cercare informazioni, i filtri sono fondamentali per indicare "cosa" rappresenta informazione.

Un filtro è un criterio per valutare il contenuto informativo delle pagine corrispondenti agli URL: NetWatcher considera interessanti (e quindi "da comunicare" all'utente) solo le pagine che soddisfano tutti i criteri impostati. Un filtro serve, ad esempio, ad esprimere criteri di ricerca di informazioni come: "solo le pagine che trattano certi argomenti" oppure: "solo le pagine in cui sono espressi certi concetti ed in cui si trovano certe keyword".

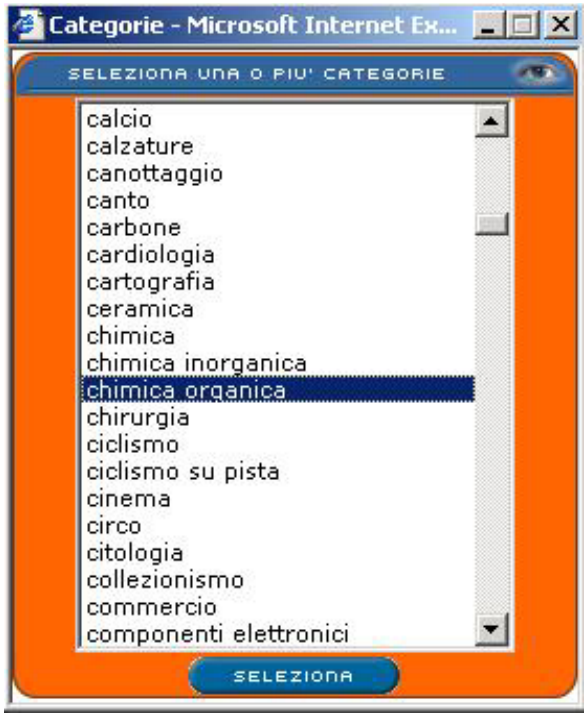
Un filtro è costituito da una combinazione di criteri su keyword, categorie e concetti e, in un profilo, è possibile anche impostarne più di uno. Date le pagine lette attraverso gli URL di un profilo, il filtro serve a NetWatcher per determinare se una determinata pagina deve essere "comunicata" all'utente come novità riguardante il profilo stesso. Lo scopo del filtro, infatti, è stabilire le caratteristiche (in termini di contenuti) che una pagina letta mediante il profilo deve avere per essere considerata interessante e, quindi, "da comunicare".

Una keyword è una parola che appare nel testo della pagina. E' possibile cercare una parola esattamente come viene scritta, oppure cercare anche tutte le forme flesse di essa: singolari, plurali, maschili e femminili per i sostantivi e i vari tempi e modi per le forme verbali. Quest ultimo metodo di ricerca è detto per lemmi. Nel caso pratico della ricerca della parola *bambino* come semplice keyword , il sistema troverà le sole pagine che contengono la parola scritta esattamente, mentre cercando come lemma il sistema troverà anche le pagine contenenti *bambina*, *bambine*, *bambini*. Per distinguere una ricerca per keyword da una per lemma è necessario spuntare l'opzione lemma di fianco alla text box di inserimento prima di premere inserisci.

Quando vengono inseriti più di un lemma o una keyword è necessario scegliere se considerarli collegati dall'operatore logico "and" o dall' "or" , per questo si utilizza un radio button con "tutti" per applicare l'operatore "and" e "almeno una" per l' "or".

Una categoria è un argomento o tema. Quando NetWatcher esamina una pagina determina e memorizza gli argomenti trattati nella parte più significativa del testo mediante il categorizzatore. Premendo sul pulsante inserisci si accede all'elenco completo di tutte le categorie memorizzate da NetWatcher e se ne può scegliere una o più.

Le categorie non sono rappresentate da un albero, ma da un elenco puro , non sono quindi trattabili gerarchie di categorie con padri e figlie. Esistono però alcuni ambiti, per esempio *chimica* e *industria*, nei quali sono presenti categorie con relative sottocategorie di specializzazione come, per quanto riguarda la chimica, chimica organica e chimica inorganica. In questo modo si può modulare a piacere il grado di specializzazione delle pagine ricercate.



Una volta inserito un argomento è possibile eliminarlo dal filtro premendo la x al fianco dell'argomento stesso e, se vi sono più categorie è necessario selezionare con quale operatore logico collegarle tra loro mediante un radio button analogo a quello delle keywords.

Un concetto è un determinato significato espresso da una parola, una stessa parola può dunque essere collegata a diversi concetti. Quando NetWatcher esamina una pagina cerca di determinare il significato delle parole contenute nel testo. Interpreta il testo, risalendo dai termini scritti ai significati che essi rappresentano e distinguendo, grazie all'analisi del contesto, tra possibili significati alternativi di forme scritte nello stesso modo (disambiguazione dei significati). Se incontra la parola *pesca*, ad esempio, cerca di scoprire se si tratta del frutto oppure dell'attività compiuta dai pescatori oppure di una lotteria a scopo benefico.

Le informazioni sui significati rilevati vengono poi memorizzate e diventano "ricercabili". Premendo inserisci si apre il pannello che permette di scegliere un concetto. Si deve inserire un parola del dizionario all'interno della text box e, una volta premuto esegui, è necessario selezionare il significato della parola che si intende cercare.

Se un filtro viene impostato con più di una tipologia di parametri (keywords, argomenti, concetti) è necessario, mediante i radio button presenti sulla destra, selezionare come combinare tra loro essi (and/or).

Una volta creato un filtro può essere:

- abilitato o disabilitato (i filtri disabilitati non sono presi in considerazione)
- clonato (viene creato un nuovo filtro identico che può essere poi modificato)
- eliminato

3.4.2 Sezione principale

Al centro vi è la schermata degli aggiornamenti nella quale vengono presentati tutti i profili creati dall'utente e i collegamenti alle relative pagine di push interattivo.
Nel dettaglio sotto al nome del profilo abbiamo la data dell'ultima indicizzazione e uno o due link, il primo apre la pagina degli aggiornamenti riscontrati dall'ultima indicizzazione che ha prodotto risultati, mentre il secondo apre quella di tutti i crawl effettuati negli ultimi sette giorni.

Nella colonna di sinistra vi è il menu di navigazione delle varie schermate della sezione: ricerca e gestione dei dati personali. In alto abbiamo invece il pulsante per disconnettersi dal sistema e il collegamento alla sezione profili dell'applicativo.

3.4.2.1 Push interattivi

I push interattivi vengono aperti in una nuova finestra non ingrandibile. Nell'intestazione in alto sono indicati la data dalla quale sono stati riscontrati aggiornamenti e la data dell'ultima indicizzazione. Sotto vengono invece visualizzati i risultati trovati. Per ognuno vi è, sulla prima riga il titolo della pagina, che è cliccabile e collega all'ultima versione pubblicata sul sito di provenienza, sulla seconda abbiamo l'URL di provenienza, sulla terza abbiamo il collegamento alla pagina archiviata nella cache del sistema. Sotto vi è una piccola sintesi del testo contenuto all'interno della pagina e in fine abbiamo le categorie di appartenenza delle pagine trovate, riconosciute dal categorizzatore, e la data e l'ora della loro indicizzazione.

3.4.2.2 Ricerca in archivio

La ricerca in archivio è uno degli strumenti più utili e importanti del sistema, consente infatti di impostare ricerche, complesse a piacere, tra tutte le pagine che sono state indicizzate durante i crawl sfruttando le tecnologie semantiche.

Tutte le pagine lette da NetWatcher durante la sua attività sono memorizzate in diversi archivi, uno per ogni profilo creato, perciò quando si vuole eseguire una ricerca bisogna prima di tutto impostare l'ambito nel quale andrà ad operare.

L'ambito si definisce mediante la scelta di un profilo, attraverso i cui URL le pagine sono state lette, e mediante l'indicazione del periodo nel quale le pagine sono state archiviate. Sia la scelta del profilo dal quale andare a cercare le informazioni, sia quella del periodo vengono effettuate mediante un menu a scelta multipla combobox. Nella prima combobox vi sono tutti i profili creati dall'utente mentre nella seconda vi sono una serie di periodi che vanno dagli ultimi tre giorni fino agli ultimi tre mesi.

Le ricerche all'interno dell'archivio possono essere effettuate mediante diversi parametri: keyword, lemmi, concetti e categorie. La procedura per

l’impostazione della ricerca è analoga in tutto e per tutto alla creazione di un filtro che è stata trattata in precedenza.
Sotto al menu a tendina dei profili c’è un opzione spuntabile che se attiva inserisce automaticamente nella ricerca corrente i filtri impostati nel profilo selezionato.

Se una ricerca viene impostata selezionando più di una tipologia di parametri di ricerca (keywords, argomenti, concetti) o, se viene selezionata l’opzione di utilizzo dei filtri del profilo nel quale si va a cercare, nella parte destra dello schermo, sotto alla selezione dell’ambito, appaiono alcuni radio button coi quali si definisce un riassunto dei criteri di ricerca impostati. Per definire questo criterio complessivo è quindi necessario selezionare le relazioni logiche (and o or) tra i differenti tipi di parametri precedentemente definiti.

4 Utilizzo di NetWatcher® come supporto alla creazione di newsletter

Nell'ambito del tirocinio svolto alla CNA servizi di Modena, per poter effettuare una valutazione completa del sistema NetWatcher, sono stati effettuati da me e Marco Tosi diversi test di vario tipo. Tra questi i più importanti e degni di nota sono stati:

- quelli di monitoring di fonti di dati giornalistici;
- quelli di confronto tra i risultati di query effettuate con diversi sistemi di information retrieval.

In questo capitolo verranno presentati i test della prima tipologia, mentre nel capitolo successivo saranno trattati quelli di confronto.

4.1 Area di utilizzo del sistema

L'obiettivo principale della CNA servizi di Modena è quello di offrire servizi efficienti per le piccole e medie imprese in tutti i vari settori di interesse. Per ottenere questo obiettivo la CNA si avvale di 57 sedi nel territorio provinciale e di altre 20 società strategiche correlate. Nel settore della meccanica la CNA ha creato la struttura Assomeccanica la quale, grazie alla partecipazione di esperti del settore, aggiorna e aiuta le piccole e medie imprese di questo settore favorendone la competitività e aiutandole nella ricerca di Business. A CNA Assomeccanica Modena aderiscono le imprese Artigiane e della piccola e Media Industria della meccanica di produzione della committenza e della subfornitura.

Un valido strumento informativo gestito da CNA Assomeccanica è il portale www.tuttostampi.com. Tuttostampi è il portale dello stampaggio industriale rivolto a stampisti, stampatori e indotto di materie plastiche e metalliche. Si tratta di un portale orientato a una logica di servizio, commercio elettronico e business to business, che mira a creare un punto di incontro per l'intero panorama stampistico in internet. L'obiettivo di Tuttostampi è di supportare le aziende dello stampaggio industriale ad acquisire progressivamente gli strumenti per poter entrare nell'area delle nuove forme di economia, garantendo a tale scopo supporto formativo, informativo e di servizio e fornendo uno spazio per le nuove forme di transazione commerciale.

All’interno di questo portale è inoltre presente un servizio di newsletter a cui tutti gli utenti possono aderire. Questo servizio permette agli utenti registrati di ricevere periodicamente informazioni sulle novità, gli eventi, le fiere e gli annunci di compravendita del settore.

La creazione di questa newsletter prevede una ricerca di informazioni standard. Colui che gestisce l’invio della newsletter infatti visita regolarmente una serie di siti del settore e quando trova delle news che reputa importanti le inserisce nella newsletter del portale indicando anche la fonte da cui provengono.

Essendo questo procedimento molto meccanico e ripetitivo, per ridurre i tempi di ricerca delle informazioni, si è pensato di utilizzare un sistema automatico di news monitoring come NetWatcher.

4.2 Attività di news monitoring

Come introdotto poco sopra, si vuole automatizzare la ricerca di news relative al settore meccanico e dello stampaggio.

Per effettuare questa attività mediante il sistema NetWatcher, è necessario impostare dei profili di ricerca nei quali inserire tutti gli URL delle fonti richieste ed eventuali filtri per selezionare le informazioni di interesse.

Per prima cosa quindi, mi sono stati dati, dall’utente che generalmente compie la ricerca delle news, l’elenco dei siti nei quali generalmente vengono trovate esse:

http://www.italianmec.com/

http://www.meccanicanews.com/home.asp

http://www.opifici.com/

http://www.anima-it.com/

http://www.ipi.it/

http://www.aluplanet.it/ita/home.asp

http://www.assocomaplast.org/index_ita.asp

http://www.assomet.it/

http://www.ilb2b.it/

http://www.istat.it/

http://www.ilsole24ore.com

http://www.siderweb.com/main/

http://www.subfor.camcom.it/show.jsp

http://emarketservices.it/

http://www.c-s-m.it/

http://www.cheminitaly.it/

http://www.dps.tesoro.it/

http://www.gommanews.it/

http://www.marketpress.info/

http://www.ovest.it/mpe/

http://www.polimerica.it/

http://www.meccanicautile.com/

http://www.plastica.it/default.php

http://www.macplas.it/index.asp

http://www.deformazione.it/

http://www.apri-rapid.it/

http://www.eplast.it/chi.html

http://www.apogeonline.com/

http://praxa.it/news.php/main?uid=/kgBSTvMTmvAqAAF

In questo elenco sono presenti siti molto eterogenei tra loro, infatti si va dal sito specializzato in un determinato settore a quello di una testata giornalistica economica.

Il primo approccio da me utilizzato è stato quello di creare un unico profilo nel quale inserire tutti gli URL di partenza senza filtri precisi ma questo ovviamente ha dato, nei vari giorni successivi un numero esagerato di risultati. Questi risultati, oltre ad essere troppi, erano anche inadeguati, infatti in essi non erano presenti soltanto news, ma anche qualsiasi altro tipo di pagina, da quelle pubblicitarie a quelle dei contatti del sito.

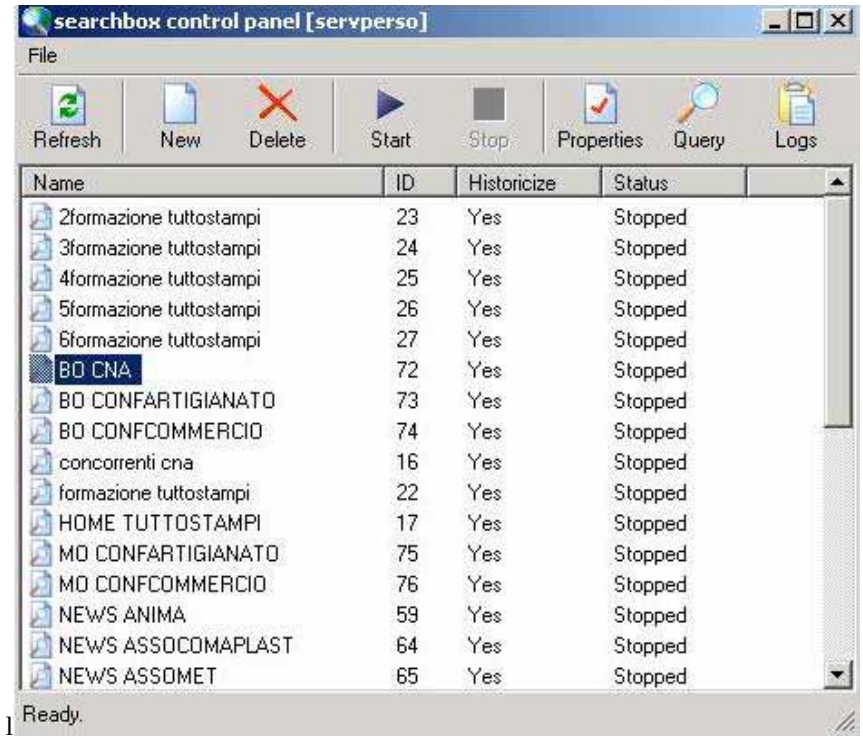
Nei giorni successivi quindi ho analizzato, insieme all'esperto di settore che effettua generalmente le ricerche, i vari siti in modo da poter creare dei profili più dettagliati, nei quali inserire soltanto le sezioni di essi in cui si possono trovare notizie adeguate. Inoltre per poter analizzare al meglio i risultati ho diviso le ricerche in tanti profili con un solo sito per ognuno.

Netwatcher non consente di indicare quali sono i link da seguire e quali non lo sono. Così partendo da una pagina di un sito il crawler scandisce tutti gli URL che trova in essa e va a indicizzare pagine inutili per gli obiettivi della ricerca. In questo caso sarebbe necessario, partendo, ad esempio, da una pagina in cui vi è un elenco di news scaricare solo e soltanto esse, e non tutti gli altri link presenti. Per questo, utilizzando il pannello di controllo searchbox è possibile una semplice modulazione del tragitto da far percorrere al crawler.

Nella schermata principale del pannello di controllo searchbox vi è l'elenco di tutti i profili creati da tutti gli utenti di NetWatcher e alcune informazioni su di essi:

- nome

- numero identificatore
- presenza dei dati nella cache
- stato del profilo, cioè se è in corso o meno un crawl



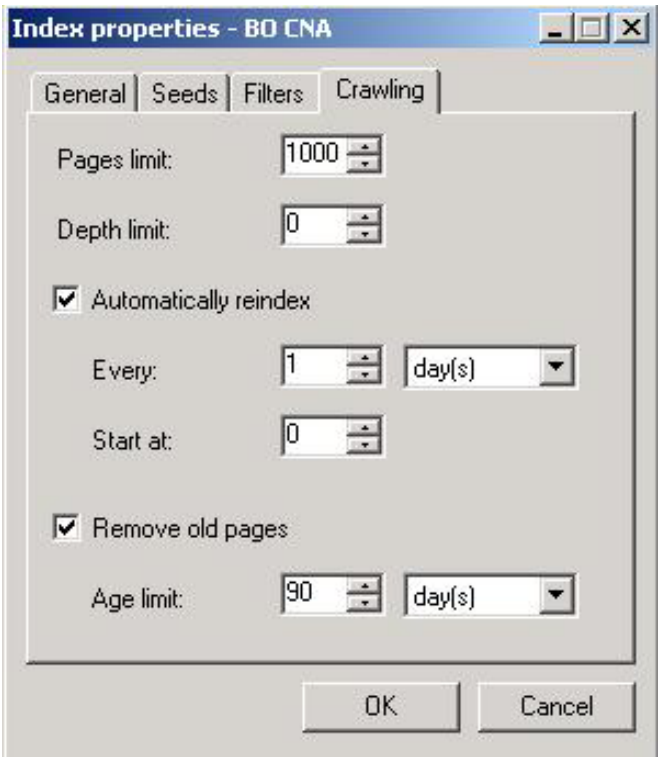
Da questa interfaccia è possibile creare o modificare profili di scansione dei siti. Selezionando col mouse un profilo e cliccando il pulsante proprietà è possibile andar a cambiare alcuni parametri di indicizzazione che non sono modificabili nell'interfaccia web di NetWatcher. NetWatcher infatti al momento della creazione di un profilo dall'interfaccia web imposta tutti questi automaticamente.
La finestra delle proprietà ha quattro diverse sezioni e ognuna di esse ha una scheda associata:

- General
- Seeds
- Crawling
- Filters

Nella prima scheda viene indicato il nome del profilo e se l'indicizzazione di esso va storicizzata o meno.

La scheda dei seed, ovvero i semi, visualizza le pagine di partenza delle scansioni e consente quindi di modificare gli URL creati da NetWatcher.

Questo tipo di modifica non ha però una grande utilità essendo possibile effettuarla direttamente dall'interfaccia grafica di NetWatcher.



Nella sezione crawling vi sono alcuni parametri di indicizzazione.

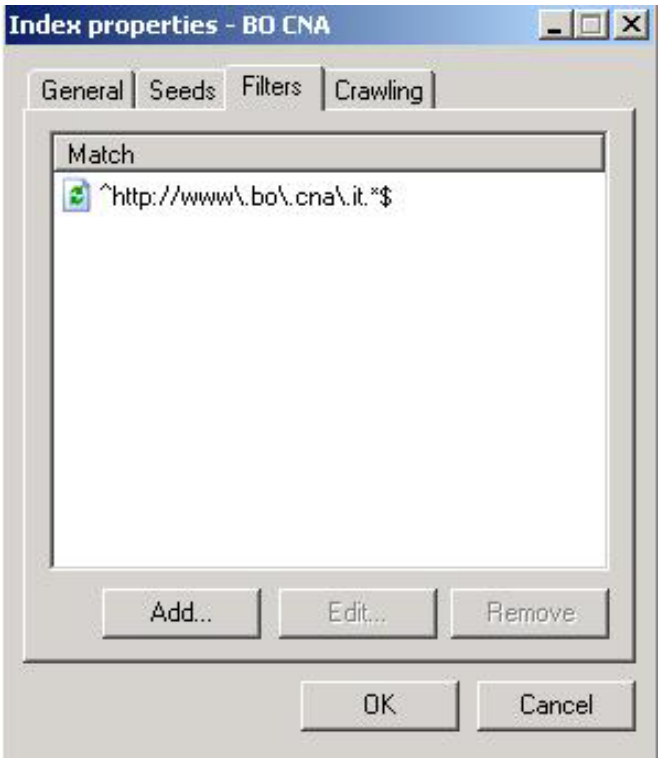
E' possibile infatti decidere la frequenza e l'ora in cui effettuare l'indicizzazione dei siti, oppure si può andar a modificare il numero massimo di pagine da indicizzare e decidere se rimuoverle o meno dalla cache del sistema dopo un determinato periodo.

Il parametro più significativo di questa scheda è però il depth limit, ovvero il limite di profondità. Nell'interfaccia web di NetWatcher la decisione della profondità della ricerca in un sito è limitata a due sole possibilità: profilo di tipo news, il quale ha profondità 1, oppure profilo di tipo concorrente che non ha limiti di profondità, cioè va a rovistare in tutte le pagine del sito di partenza.

Quando NetWatcher imposta un profilo concorrente il limite di profondità in questa scheda viene settato a 0 che indica nessuna limitazione, mentre per quelli news a 1. In alcuni casi è però necessario avere una profondità intermedia, ad esempio 2 o 3 livelli, quindi è necessario andar a modificare questo parametro manualmente all'interno di questa scheda.

Un'altra sezione molto utile per l'utente è la sezione filtri.

In questo strumento un filtro è inteso in modo completamente diverso da quelli che abbiamo introdotto per l'interfaccia grafica di NetWatcher. In essa infatti i filtri erano considerati una combinazione di parametri di ricerca nel testo; essi erano quindi necessari per valutare il contenuto informativo di una pagina.

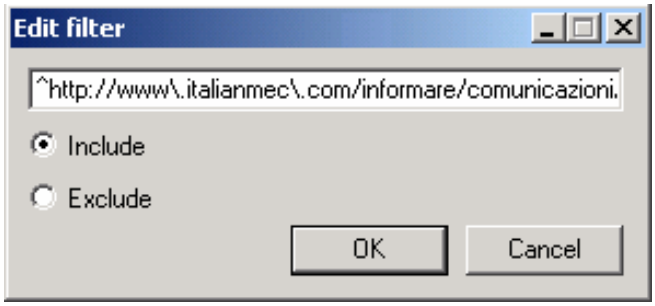


In questo contesto invece il filtro serve per indicare quali pagine indicizzare non valutando ciò che è presente nel testo ma l'indirizzo della pagina stessa.

Per effettuare questa valutazione dell'indirizzo vengono utilizzate delle stringhe chiamate regular expression.

Le espressioni regolari (Regular Expression in inglese, abbreviato regexp, regex o RE) sono una sintassi definita per descrivere un insieme di stringhe con un modello comune. La costruzione delle espressioni regolari è analoga a quella delle espressioni aritmetiche. Mediante un insieme di metacaratteri e operatori vengono combinate espressioni di piccole dimensioni in modo da creare espressioni estese. I componenti di un'espressione regolare possono essere costituiti da caratteri singoli, set di caratteri, intervalli di caratteri, alternative tra caratteri o qualsiasi combinazione di questi componenti.
Nel caso più semplice un espressione regolare rappresenta una stringa nella quale ogni carattere corrisponde a se stesso, a meno che non siano presenti caratteri speciali che hanno significati specifici come:

^ (inizio stringa)
$ (fine stringa)
. (qualsiasi carattere)
* (ripetizione carattere precedente)



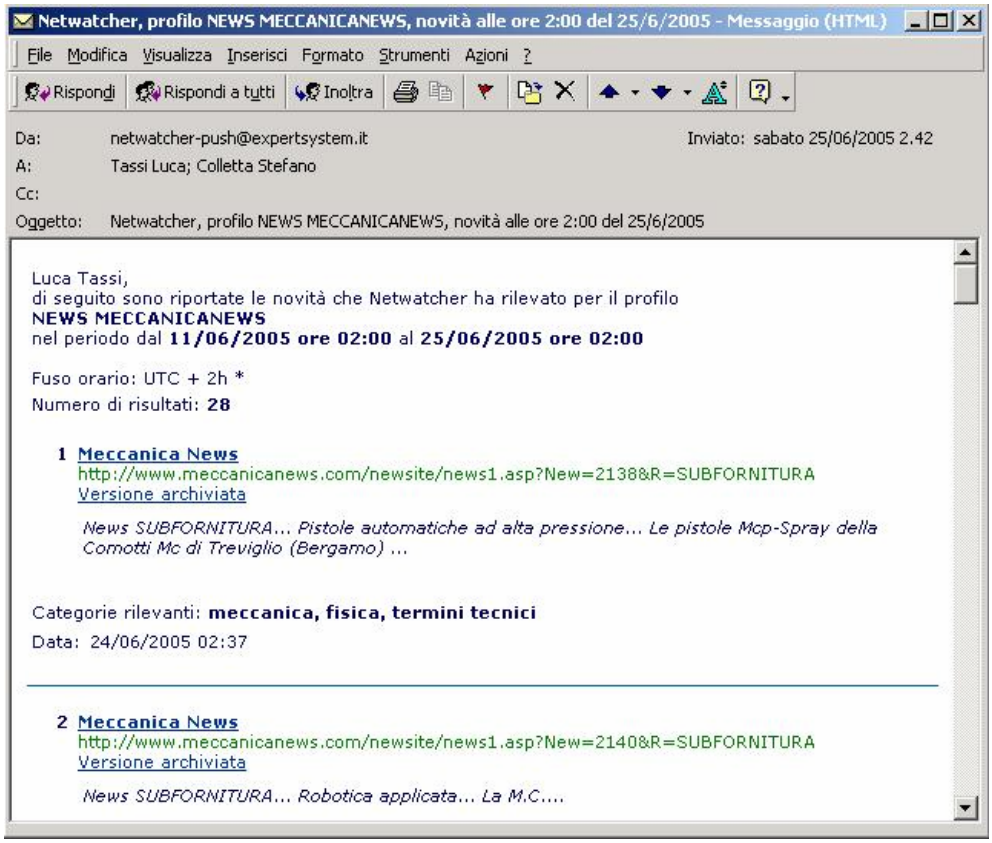
Un filtro di searchbox può essere inclusivo o esclusivo, così è possibile escludere determinate pagine non significative e includere quelle più importanti nella scansione.

4.3 Analisi dei risultati

L'attività di creazione di profili da me svolta aveva come obiettivo principale la creazione di un archivio di news completo ed aggiornabile basato soltanto su news realmente interessanti. Mediante questo archivio colui che gestisce l'invio delle newsletter del portale tuttostampi ha a disposizione informazioni aggiornate e precise sempre consultabili al momento del bisogno.

Per trovare le news da inserire nella newsletter è possibile utilizzare sia il push interattivo, sia il push via mail sia la ricerca in archivio.
I push via mail sono le mail che genera automaticamente il sistema dopo ogni scansione dei siti. In ogni e-mail ricevuta, vengono inseriti gli stessi risultati presenti nei push interattivi della sezione principale del sistema.

Nelle mail abbiamo una prima intestazione nella quale sono indicati il nome dell'utente, il profilo a cui si riferisce il push e il periodo nel quale sono stati riscontrati quei risultati; nella parte seguente del messaggio abbiamo invece un elenco dei risultati trovati. Per ogni risultato vi è un link alla pagina presente sul sito di provenienza, un link alla pagina salvata nella cache del sistema, un breve sommario del testo della pagina trovata e le categorie di appartenenza delle pagine trovate, riconosciute dal categorizzatore, e la data e l'ora della loro indicizzazione.



Questo strumento è risultato molto utile, infatti grazie alle precise espressioni regolari designate l'utente riceve soltanto news e non pagine inutili (come ad esempio quelle relative ai contatti del sito).

Un altro aspetto positivo è la tempestività della ricezione dei risultati. Ogni mattina quindi colui che compone la newsletter ha a disposizione direttamente nella propria casella mail le nuove notizie, senza quindi dover navigare tutti i siti alla ricerca di esse. Inoltre le notizie vengono considerate nuove se cambia sostanzialmente il contenuto della pagina e non soltanto se cambiano parti marginali.

Nei casi da me trattati ogni giorno i push contenevano un numero molto elevato di risultati, grazie all'interfaccia di ricerca in archivio è però stato possibile filtrare essi con criteri a piacere.

Mediante l'analisi giornaliera dei risultati ricevuti ho potuto riscontrare alcuni problemi non trascurabili.

Uno di essi riguarda l'analisi delle pagine: in certi risultati vengono riportate alcune pagine considerate nuove dal sistema, ma che in realtà nuove non sono. Questo avviene probabilmente perché il sistema non riesce a interpretare bene la struttura della pagina e quindi considera alcune parti di essa importanti anche se non lo sono. L'algoritmo che stabilisce l'importanza di ciascuna parte del testo a volte prende decisioni che non corrispondono a quelle che prenderebbe un essere umano.
Un esempio lampante di questo errore lo si ha quando in una pagina con una news molto vecchia vengono inseriti o cambiati banner, oppure quando in una pagina viene visualizzata la data, infatti l'algoritmo non riesce a riconoscer il concetto di data. Una altro caso in cui il sistema ha qualche difficoltà ad individuare le parti importanti della pagina è quando esse sono strutturate su più frame.

Il secondo problema principale del sistema riguarda l'indicizzazione delle pagine. Il crawler infatti non riesce ad indicizzare correttamente le pagine nelle quali sono presenti javascript. Perciò i profili in cui vi sono pagine di questo tipo restituiscono spesso pagine errate o mal visualizzate.

In conclusione gli esperimenti di monitoraggio svolti dimostrano che il sistema NetWatcher è molto utile nel reperimento di notizie da inserire nelle newsletter in quanto velocizza i tempi delle ricerche, ma non può rendere completamente automatico il reperimento di esse in quanto è sempre e comunque necessario un intervento partecipativo dell'utente al quale spetta anche la decisione di quali news pubblicare nelle newsletter.

5 Test di confronto NetWatcher® / SEWASIE

In questo capitolo verranno presentati alcuni test, effettuati da me e Marco Tosi, sui due sistemi di web search presentati nei capitoli precedenti.

L'obiettivo di questa attività è quello di formulare, utilizzando le interfacce grafiche dei due sistemi, delle ricerche di informazioni simili all'interno di una fonte di dati comune. La realizzazione di questo test può essere suddivisa in diverse fasi, ognuna delle quali ha impegnato risorse e tempi differenti:

1) creazione di una fonte comune di dati;

2) identificazione parametri di valutazione e esecuzione delle query;

3) analisi dei risultati.

5.1 Creazione fonte di dati comune

Per poter effettuare un confronto attendibile, abbiamo analizzato i vari campi di applicazione dei sistemi. Come presentato nel capitolo tre, il punto di forza del sistema NetWatcher è l'analisi del testo, in questo contesto infatti riesce a estrapolare al meglio concetti e argomenti dalle parole presenti nel testo. Il sistema SEWASIE invece si pone come prerogativa la creazione di dati strutturati mediante le varie fasi presentate in precedenza, in modo da rendere poi possibile una ricerca di informazioni guidata dall'ontologia creata. Data questa enorme differenza ci siamo orientati verso una tipologia di informazioni che consentisse di sfruttare le potenzialità di entrambi gli strumenti, ovvero le news. Esse infatti consentono sia un analisi testuale, sia una strutturazione dei dati.

Tra i vari siti, che sono stati monitorati da NetWatcher per l'attività di rassegna stampa presentata nel capitolo precedente, ne sono stati selezionati due da utilizzare anche in questa serie di test. La fonte comune dei dati è infatti composta da una serie di news presenti in due portali del settore meccanico/plastico.

All'interno del portale Polimerica.it sono state prese in considerazione tutte le news relative alle categorie Economia e mercati , Osservatorio prezzi e Materie prime per un totale di 613 articoli. Le suddette categorie si trovano all'interno della sezione Notizie.


Polimerica - il portale delle materie plastiche - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti Multimedia
Indirizzo http://www.polimerica.it/modules.php?name=Stories_Archive&sa=show_topic&year=5
Vai Collegamenti
Polimerica
Attualità e notizie dal mondo della plastica
ABCS Formazione Tecnica Ricerca e Sviluppo
Home Notizie Approfondimenti Utilità Info Contatti
LOGIN
Nickname
Password
Login
Non hai ancora un tuo account? Registrati! Come utente registrato potrai utilizzare tutti i servizi.
I NOSTRI SPONSOR
R&D Materie Plastiche Formazione Scientifica Laboratorio di Analisi ABCS
Archivio articoli
Polimerica - il portale delle materie plastiche: Economia e mercati
ARTICOLI LETTURE VOTO DATA AZIONI
L'e-commerce torna di moda 134 0 2005-11-25
Segnali discordanti per ordini e fatturato 106 0 2005-11-22
Crescono le plastiche in Spagna 198 0 2005-11-16
Vetrina: Macchine per imballaggio in ripresa 158 0 2005-11-16
La produzione torna giù 168 0 2005-11-15
Ottobre positivo per l'auto 120 0 2005-11-07
Vetrina: Macchine: export fermo nel I semestre 178 0 2005-11-02
Estrusione di tubi in Europa 222 4 2005-10-31
Fatturato in ripresa anche ad agosto 223 0 2005-10-24
Cavi: Italia leader in Europa 483 3 2005-10-20
Trasformatori europei sotto pressione 553 3.66 2005-10-20
Menu ready for use
Internet


ItalianMec - Comunicazioni - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti Multimedia
Indirizzo http://www.italianmec.com/informare/comunicazioni/
Vai Collegamenti
newsletter contattaci link utili mappa
ITALIAN mec
STANTE & ECOTRANS IL MODO più semplice per spedire in:
Federazione ANIMA
cerca
Opportunità Mercati Formare Informare
Comunicazioni
Comunicazioni
Canale Tecnico
Canale Ambiente e Sicurezza
Canale Economico
Canale Giuridico
PRE-REGISTRATI!
24.11.2005
TAU INTERNATIONAL: l'inaugurazione sarà dedicata ai rapporti tra energia e ambiente
24.11.2005
Siglato l'accordo di collaborazione del Programma Costruction Equipment Lab (C.E.Lab)
24.11.2005
Semplificazione, Confindustria: Un buon segnale, ma serve di più Morandini: "I mercati non ci aspettano, le Regioni siano coerenti"
27.10.2005
Progetto "Linea di credito per i Balcani"
26.10.2005
Finitura & Oltre: una grande prima!
Vero punto d'incontro tra offerta e domanda: tanto business generato a Bergamo Fiera.
20.10.2005
Internet

Per quanto riguarda Italianmec.it le informazioni a cui si è fatto riferimento sono state le 369 notizie appartenenti alla sezione Informare che si suddivide in varie categorie: Comunicazioni , Canale Tecnico, Canale Ambiente e Sicurezza, Canale economico e Canale giuridico.

Nel sistema SEWASIE questa fonte di dati è stata creata grazie all'utilizzo di due wrapper i quali, navigando i sito mediante i link presenti nella pagina cercano di estrarre le informazioni importanti di esse e di salvarle in una base di dati strutturata. Questa procedura va quindi a popolare due database. Mediante le varie fase di integrazione dati del MOMIS essi vengono combinati in una Global Virtual View sulla quale si potranno eseguire tutte le query del test in modo omogeneo grazie all'interfaccia SQoogle, che verrà presentata nei paragrafi successivi.

I due database sono comunque stati salvati e mantenuti in locale per poter effettuare direttamente query SQL [14] su di essi.

Nel sistema NetWatcher invece è stato programmato, mediante l'interfaccia web integrata con il pannello di controllo Searchbox, un profilo nel quale sono stati indicizzati i dati provenienti da entrambi i portali. Per modulare al meglio questo profilo sono stati inseriti 10 URL di partenza (seed), uno per ogni sezione dei due portali da trattare e, vista la similitudine degli URL delle news sono stati inseriti come filtri inclusivi di Searchbox queste espressioni regolari:

^http://www\.italianmec\.com/informare/comunicazioni/pagina\.phtml.ID=[0-9][0-9][0-9][0-9]$

^http://www\.italianmec\.com/informare/canaleambsic/pagina\.phtml.id=[0-9][0-9][0-9][0-9]$

^http://www\.polimerica\.it/modules\.php.name=News&file=article&sid=.*$

Dato che nel portale Italianmec.it le news vengono presentate su diverse pagine, per esso è stato necessario inserire altri sette filtri inclusivi searchbox (uno per ogni seed) per rendere possibile l'indicizzazione delle news presentate nelle pagine successive alla prima:

^http://www\.italianmec\.com/informare/comunicazioni/index\.phtml.nrecs=20.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Newsamb.nrecs=121.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Finanziamenti.nrecs=27.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Pilloleamb.nrecs=32.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Approfondimenti.nrecs=24.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Pillolesic.nrecs=31.from=[1-9].*$

^http://www\.italianmec\.com/informare/canaleambsic/index\.phtml.rubrica=Newssic.nrecs=104.from=[1-9].*$

Le operazione di estrazione dati sono state effettuate, in entrambi gli ambienti, in un periodo di tempo ravvicinato, per impedire di avere dati differenti a causa di aggiornamenti o inserimenti di news all'interno dei siti, in modo da mantenere così una base di dati comune consistente. Inoltre sempre per questo motivo, data l'impossibilità di effettuare il wrapping automaticamente ad intervalli regolari di SEWASIE, nel NetWatcher non è stata impostata una frequenza con la quale rieffettuare i crawl.

In conclusione, entrambi i sistemi, hanno composto la propria fonte di informazioni con le stesse 982 news: questo ha reso quindi possibile un confronto accurato dei risultati ottenuti.

5.2 Query

L'obiettivo in questa fase è stato rivolto alla realizzazione di query il più simili possibile tra i due sistemi da analizzare. In questo tentativo però sono state riscontrate molte difficoltà dovute alle differenze sostanziali tra i due sistemi.

Essi infatti, come presentato nei capitoli precedenti, sono strutturati in maniera molto differente, sia dal punto di vista dell'architettura di base, sia dal punto di vista dell'interfaccia grafica.

La prima scelta effettuata per riuscire a trovare un punto d'incontro tra i due sistemi è quella della fonte di dati, infatti come introdotto precedentemente le news sono una delle poche (o forse l'unica) tipologia di dati che consente di sfruttare le potenzialità di entrambi i sistemi.

Nel SEWASIE le news vengono categorizzate in base alla loro posizione nel sito di provenienza. In essi infatti abbiamo diverse sezioni e ogni news appartiene a una di esse. Utilizzando quindi l'interfaccia grafica SQoogle è possibile effettuare una ricerca per categoria. Altrimenti è possibile andar a cercare l'occorrenza di una parola all'interno del testo o del titolo della notizia.

Nel NetWatcher invece le news vengono categorizzate mediante l'analisi del testo. Il disambiguatore infatti va a esaminare ogni frase per capirne il senso e, fatto questo decide di che argomento si sta trattando.

Pur essendo metodi di classificazione molto differenti è possibile effettuare un confronto tra essi, e per questo la maggior parte delle query realizzate mette in evidenza questo aspetto.

5.2.1 Parametri di valutazione

Tutte le query che verranno esposte in seguito sono state effettuate con tre strumenti differenti:

- SEWASIE
- NetWatcher
- MySql [15]

I primi due strumenti rappresentano l'oggetto del test di confronto, mentre il terzo è stato utilizzato come base sulla quale andare a valutare i risultati dei primi due.

Più precisamente, dopo che i wrappers di SEWASIE hanno creato e popolato i due database di news (test per italianmec.it, test2 per polimerica.it) abbiamo espresso mediante la sintassi SQL le query create negli altri due sistemi.

Tramite queste query effettuate in SQL abbiamo definito due parametri di valutazione dei test:

- Risultati DB
- Risultati attesi

I risultati DB sono l'esatto numero di risultati ricevuti dalla query SQL effettuata su test e test2. Dato che il linguaggio SQL non permette un'analisi semantica delle informazioni presenti in un testo, è possibile avere tra i risultati alcuni record non appropriati al criterio che si vuole impostare.

Ad esempio, se si ricercano tutte le news che parlano dell'ambiente e della tutela ambientale, mediante una query SQL che analizza la presenza o meno della parola ambiente nel testo, si possono ottenere record nei quali la parola ambiente non ha il significato da noi ricercato. Questi record, pur essendo stati trovati mediante un metodo rigoroso come SQL, non combaciano col criterio di ricerca che volevamo impostare.

I risultati attesi sono quel sottoinsieme dei risultati DB nei quali il criterio di ricerca formulato è pienamente soddisfatto.

Questi sono stati ottenuti mediante una procedura di scrematura effettuata manualmente. Con questa attività abbiamo analizzato una ad una le tuple per valutare se esse fossero realmente coerenti coi criteri di ricerca che ci eravamo preposti inizialmente. Nei casi in cui il numero di risultati DB da esaminare superava le 50 unità si è proceduto tramite l'analisi casuale di campioni costituiti dal 50% degli articoli trovati:

*Risultati attesi = risultati DB * (risultati campione DB giusti / campione)*

Il numero di queste news rappresenta il dato al quale i due motori di ricerca valutati devono avvicinarsi maggiormente applicando le rispettive logiche semantiche. Questo valore è alla base della definizione della recall

Con il termine *recall* si intende il rapporto tra il numero di informazioni pertinenti estratte dai vari sistemi e la totalità delle informazioni pertinenti da estrarre, cioè i risultati attesi.

La *precision* di un sistema serve invece per indicare il rapporto tra il numero di informazioni pertinenti estratte e il numero totale di informazioni estratte.

Per valutare la pertinenza o meno dei risultati, in entrambi i sistemi, è stato usato la stessa procedura di scrematura utilizzata con i risultati DB.

Riassumendo, Dati:

Risultati pertinenti NW = *RPNW*

Risultati totali Netwatcher = *RTNW*

Risultati pertinenti SEWASIE = *RPSW*

Risultati totali SEWASIE = *RTSW*

Risultati attesi = *RA*

	NetWatcher	SEWASIE
Precision	*RPNW* / *RTNW*	*RPSW* / *RTSW*
Recall	*RPNW* / *RA*	*RPSW* / *RA*

5.2.2 Metodologia di esecuzione

La *Query1* effettuata aveva il seguente obiettivo:

Ricerca tutte le notizie riguardanti le normative sulla sicurezza del lavoro

Nell'interfaccia di ricerca in archivio di NetWatcher questa query è stata formulata utilizzando una combinazione di concetti e categorie

Categorie	Lavoro	*and*	Legislazione
Collegamento		*And*	
Concetti	Sicurezza	*or*	Sicurezza

I due concetti di "Sicurezza" espressi in or tra loro hanno i seguenti significati:

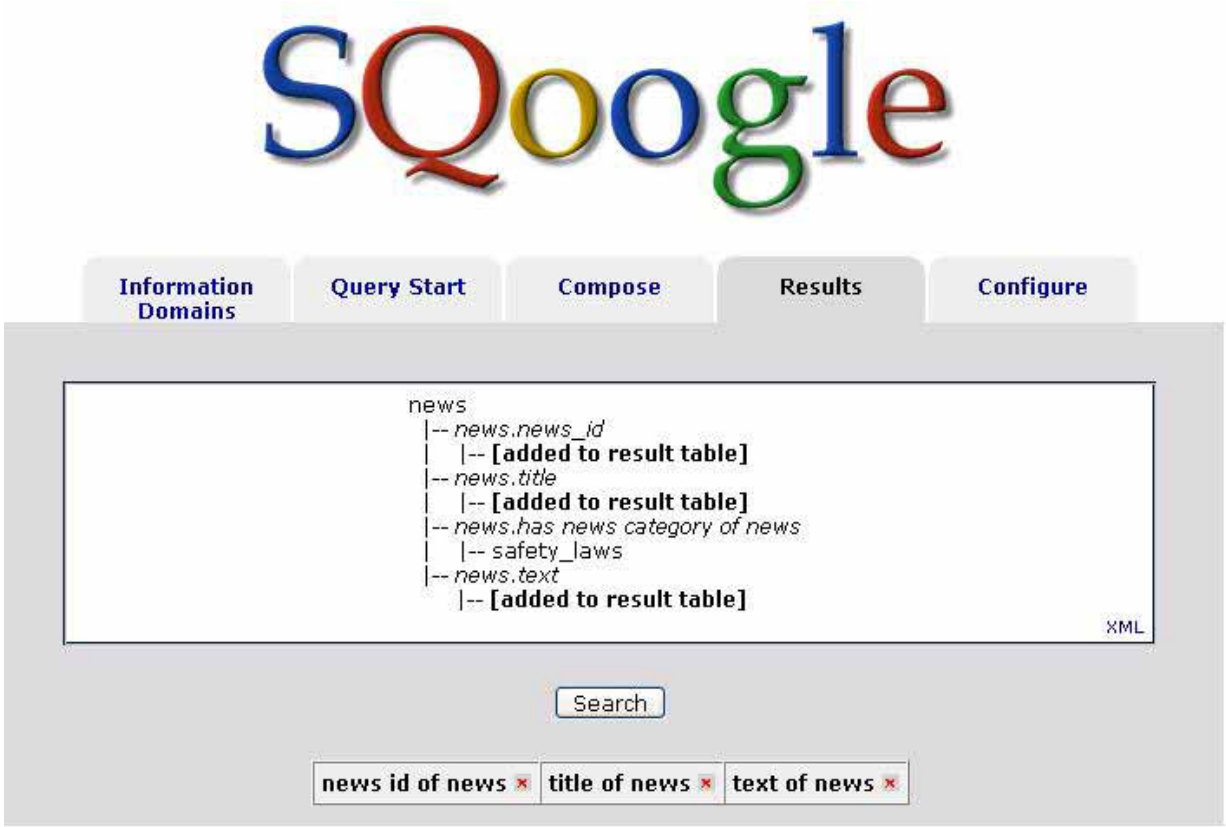
- Insieme degli interventi statali a tutela del benessere del cittadino.
- Prevenzione, eliminazione parziale o totale di danni o pericoli.

In pratica una news viene mandata in output solo e soltanto se:

E' considerata appartenente a entrambe le categorie Lavoro e Legislazione, e contiene almeno uno dei significati espressi sopra della parola sicurezza.

Nell'interfaccia SQoogle di SEWASIE invece è stata composta in questo modo:

- Prima di tutto è stata selezionata l'ontologia della meccanica (*MechanicBA*),
- poi è stata scelta la classe *news*
- sono state selezionate le tre proprietà da visualizzare:
  - *News_Id*
  - *Title*
  - *Text*
- è stata aggiunta la proprietà di join *Has news category of news*
- sono state aggiunte le specializzazioni *environment_safety*
  - *environment_safety*
  - *safety_laws*



Infine in MySql è stata formulata mediante questa espressione SQL:

```
select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT
from test.news n, test.category c
where n.CATEGORY_ID=12
and c.CATEGORY_ID=n.CATEGORY_ID  union

select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT
from test.news n, test.category c
where
(      n.TITLE LIKE "%norma%"
or     n.TITLE LIKE "%legge%"
or     n.TITLE LIKE "%decreto%"
or     n.TITLE LIKE "%normativa%"
or     n.TITLE LIKE "%direttiva%"
or     n.TITLE LIKE "%legislazione%"
or     n.TITLE LIKE "%regolamentazione%"
or     n.TITLE LIKE "%diritto%")
AND
(      n.TITLE LIKE "%lavoro%"
or     n.TITLE LIKE "%impiego%"
or     n.TITLE LIKE "%servizio%"
or     n.TITLE LIKE "%job%"
or     n.TITLE LIKE "%lavorator%"
or     n.TITLE LIKE "%occupazione%")
and
(      n.TITLE LIKE "%sicurezza%"
or     n.title LIKE "%prevenzione%"
or     n.TITLE LIKE "%tutela%"
or     n.TITLE LIKE "%salvaguardia%"
or     n.TITLE LIKE "%incolumit%")

union
select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT
from test.news n, test.category c
where
(      n.text LIKE "%norma%"
or     n.text LIKE "%legge%"
or     n.text LIKE "%decreto%"
or     n.text LIKE "%normativa%"
or     n.text LIKE "%direttiva%"
or     n.text LIKE "%legislazione%"
or     n.text LIKE "%regolamentazione%"
or     n.text LIKE "%diritto%")
AND
(      n.text LIKE "%lavoro%"
or     n.text LIKE "%impiego%"
or     n.text LIKE "%servizio%"
or     n.text LIKE "%job%"
or     n.text LIKE "%lavorator%"
or     n.text LIKE "%occupazione%")
and
(      n.text LIKE "%sicurezza%"
or     n.text LIKE "%prevenzione%"
```

```
or     n.text LIKE "%tutela%"
or     n.text LIKE "%salvaguardia%"
or     n.text LIKE "%incolumit%")

union
select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT
from test2.news n2, test2.category c2
where
(      n2.TITLE LIKE "%norma%"
or     n2.TITLE LIKE "%legge%"
or     n2.TITLE LIKE "%decreto%"
or     n2.TITLE LIKE "%normativa%"
or     n2.TITLE LIKE "%direttiva%"
or     n2.TITLE LIKE "%legislazione%"
or     n2.TITLE LIKE "%regolamentazione%"
or     n2.TITLE LIKE "%diritto%")
AND
(      n2.TITLE LIKE "%lavoro%"
or     n2.TITLE LIKE "%impiego%"
or     n2.TITLE LIKE "%servizio%"
or     n2.TITLE LIKE "%job%"
or     n2.TITLE LIKE "%lavorator%"
or     n2.TITLE LIKE "%occupazione%")
and
(      n2.TITLE LIKE "%sicurezza%"
or     n2.title LIKE "%prevenzione%"
or     n2.TITLE LIKE "%tutela%"
or     n2.TITLE LIKE "%salvaguardia%"
or     n2.TITLE LIKE "%incolumit%")

union
select  n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT
from test2.news n2, test2.category c2
where
(      n2.text LIKE "%norma%"
or     n2.text LIKE "%legge%"
or     n2.text LIKE "%decreto%"
or     n2.text LIKE "%normativa%"
or     n2.text LIKE "%direttiva%"
or     n2.text LIKE "%legislazione%"
or     n2.text LIKE "%regolamentazione%"
or     n2.text LIKE "%diritto%")
AND
(      n2.text LIKE "%lavoro%"
or     n2.text LIKE "%impiego%"
or     n2.text LIKE "%servizio%"
or     n2.text LIKE "%job%"
or     n2.text LIKE "%lavorator%"
or     n2.text LIKE "%occupazione%")
and
(      n2.text LIKE "%sicurezza%"
or     n2.text LIKE "%prevenzione%"
or     n2.text LIKE "%tutela%"
or     n2.text LIKE "%salvaguardia%"
or     n2.text LIKE "%incolumit%")
```

Questo costrutto SQL è composto da cinque query unite da delle union:

- La prima va a selezionare tutte le news che appartengono alla categoria “Norme sicurezza” di italianmec.

- La seconda va a selezionare tra le news di italianmec quelle che nel titolo hanno
  - sia un termine che contrassegna il concetto di legislazione
  - sia uno che contrassegna il concetto di sicurezza
  - sia uno che contrassegna il concetto di lavoro.

- La terza va a selezionare tra le news di italianmec quelle che nel testo hanno
  - sia un termine che contrassegna il concetto di legislazione
  - sia uno che contrassegna il concetto di sicurezza
  - sia uno che contrassegna il concetto di lavoro.

- La quarta va a selezionare tra le news di polimerica quelle che nel titolo hanno
  - sia un termine che contrassegna il concetto di legislazione
  - sia uno che contrassegna il concetto di sicurezza
  - sia uno che contrassegna il concetto di lavoro.

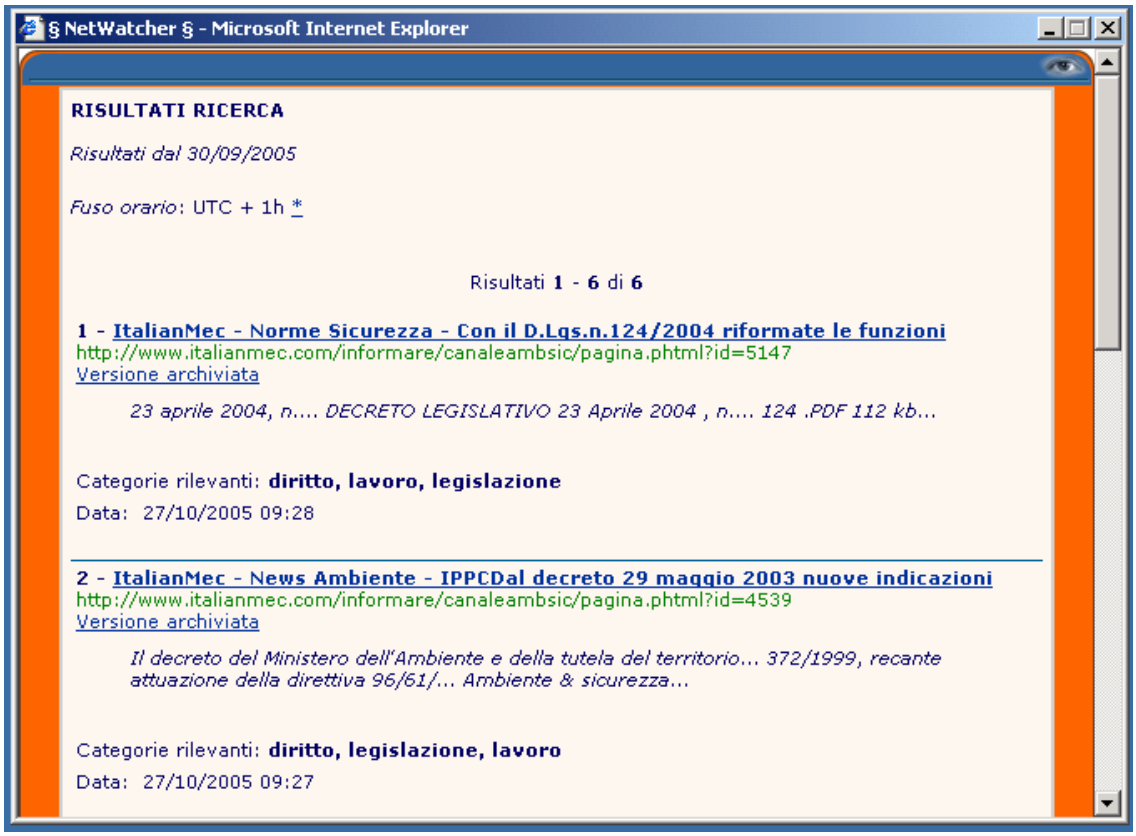
- La quinta va a selezionare tra le news di polimerica quelle che nel testo hanno
  - sia un termine che contrassegna il concetto di legislazione
  - sia uno che contrassegna il concetto di sicurezza
  - sia uno che contrassegna il concetto di lavoro.

I risultati ottenuti da questa query SQL sono 95, dopo un analisi di essi abbiamo però ottenuto che solo 54 di essi erano realmente appropriati (risultati attesi).

Nella pagina successiva sono visualizzati invece i risultati ottenuti con NetWatvher e SEWASIE, ovvero:

- NetWatcher: 6 risultati dei quali 5 appropriati

- SEWASIE: 31 risultati dei quali 25 appropriati



Results **1**-**10** of 31

news id of news	title of news	text of news
5645	Albo installatori: lattività degli iscritti è operativa solo per gli edifici scolastici	La circolare del Ministero delle Attività produttive 14 giugno 2005, n. 3584/C chiarisce che l'Albo degli installatori istituito dall'art. 109, comma 2, D.P.R. n. 380/2001, è operativo unicamente per le attività di installazione degli impianti negli edifici scolastici. Infatti, l'ultima proroga dell'entrata in vigore del Capo V, parte II, D.P.R. n. 380/2001 (cosiddetto TU edilizia), escludeva dal rinvio gli edifici scolastici di ogni ordine e grado, portando il Ministero alla conclusione che i disposti dell'art. 109, comma 2, TU edilizia, per questi edifici sono ormai entrati in vigore. Leggi la Circolare .PDF 222 kb Convenzioni ANIMA Riviste Ambiente & sicurezza Costo abbonamento annuale: ¬ 139 Costo abbonamento associati ANIMA: ¬ 125 Richiesta abbonamento soci ANIMA Banche dati Codice di Ambiente e Sicurezza
5535	Amianto intenzionalmente aggiunto	Il decreto del Ministero della Salute 14 dicembre 2004 modifica l'allegato al D.P.R. 10 settembre 1982, n. 904, recante "Attuazione della direttiva CEE relativa all'immissione sul mercato e all'uso di talune sostanze e preparati pericolosi" , già oggetto di numerose modifiche. Con il provvedimento risulta vietato l'uso delle fibre di Crocidolite, Crisotilo, Amosite, Antofillite, Actinolite e Tremolite, con i rispettivi numeri CAS, anche sotto forma di prodotti contenenti queste fibre, intenzionalmente aggiunte. È, tuttavia, consentito, fino alla data della loro eliminazione o fine della vita utile, l'uso dei prodotti contenenti le medesime fibre di amianto e già installati o in servizio prima della data di entrata in vigore del decreto in questione. (in Gazzetta Ufficiale dell'8 febbraio 2005, n. 31). Leggi il Decreto .PDF 10 kb Convenzioni ANIMA Riviste Ambiente & sicurezza Costo abbonamento annuale: ¬ 139 Costo abbonamento associati ANIMA: ¬ 125 Richiesta abbonamento soci ANIMA Banche dati Codice di Ambiente e

Ora che è stato chiarito il metodo con cui si vanno a comporre le query nei diversi sistemi inserirò soltanto due tabelle riassuntive:

- in una vi andranno le formulazioni delle query successive in linguaggio naturale e in SQL

- nell'altra vi andranno i risultati ottenuti da tutte le query e i vari parametri di analisi di essi.

Tabella query:

Query2	*News riguardanti l'università e la ricerca scientifica*	`select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name`
		`from test2.news n2, test2.category c2`
		`where n2.CATEGORY_ID=8`
		`and c2.CATEGORY_ID=n2.CATEGORY_ID`
		`union`
		`select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name`
		`from test2.news n2, test2.category c2`
		`where (n2.title like "%universit%"`
		`or n2.TITLE LIKE "%ricerca scientifica%"`
		`or n2.TITLE like "%ateneo%"`
		`or n2.TITLE like "%politecnico%")`
		`and c2.CATEGORY_ID=n2.CATEGORY_ID`
		`union`
		`select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name`
		`from test2.news n2, test2.category c2`
		`where (n2.text like "%universit%"`
		`or n2.text LIKE "%ricerca scientifica%"`
		`or n2.text like "%ateneo%"`
		`or n2.text like "%politecnico%")`
		`and c2.CATEGORY_ID=n2.CATEGORY_ID`
		`union`
		`select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.name`
		`from test.news n, test.category c`
		`where (n.title like "%universit%"`
		`or n.TITLE LIKE "%ricerca scientifica%"`
		`or n.TITLE like "%ateneo%"`
		`or n.TITLE like "%politecnico%")`
		`and c.CATEGORY_ID=n.CATEGORY_ID`
		`union`
		`select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.name`
		`from test.news n, test.category c`
		`where (n.text like "%universit%"`
		`or n.text LIKE "%ricerca scientifica%"`
		`or n.text like "%ateneo%"`
		`or n.text like "%politecnico%")`
		`and c.CATEGORY_ID=n.CATEGORY_ID`

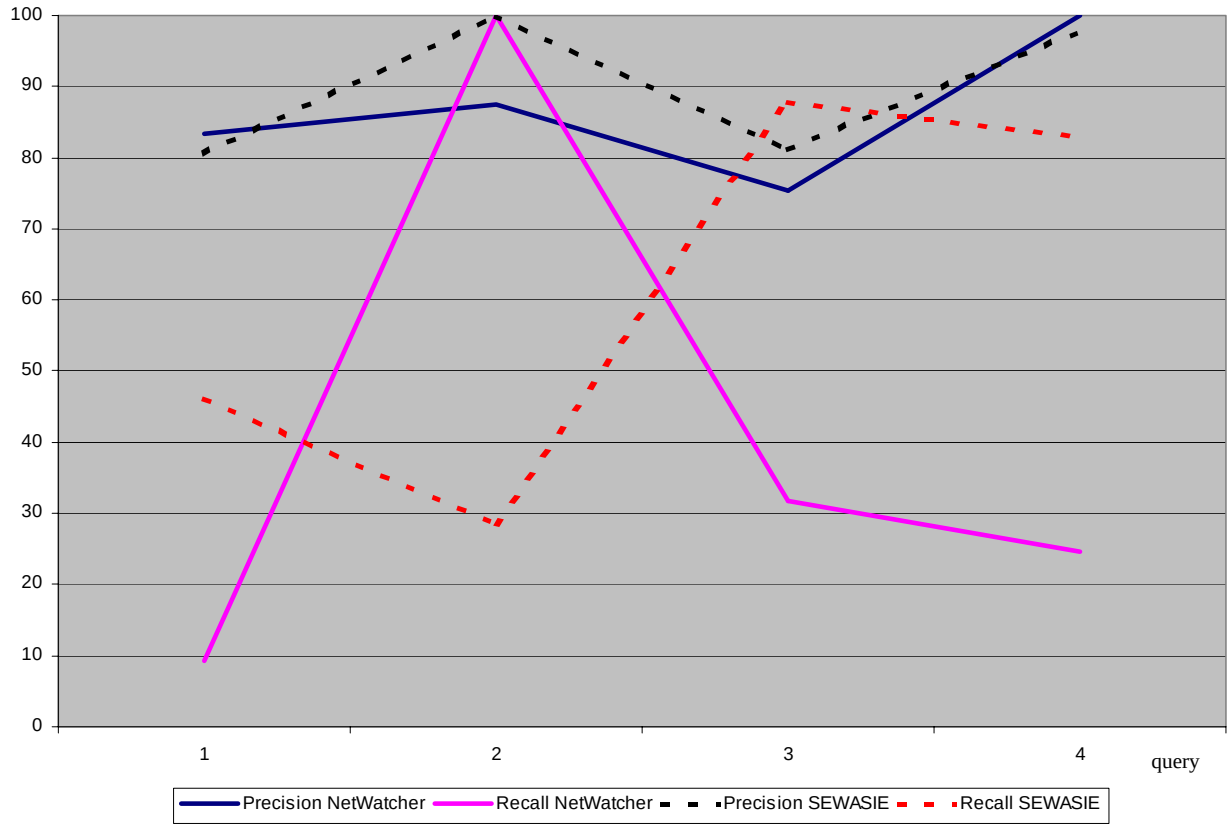
Query3	*News riguardanti economia e mercati*	select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name from test2.news n2, test2.category c2 where n2.CATEGORY_ID=1 and c2.CATEGORY_ID=n2.CATEGORY_ID union select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name from test2.news n2, test2.category c2 where n2.TITLE LIKE "%economia%" or n2.TITLE LIKE "%mercat%" and c2.CATEGORY_ID=n2.CATEGORY_ID union select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name from test2.news n2, test2.category c2 where n2.TITLE LIKE "%economia%" or n2.TITLE LIKE "%mercat%" and c2.CATEGORY_ID=n2.CATEGORY_ID union select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.name from test.news n, test.category c where n.TITLE LIKE "%economia%" or n.TITLE LIKE "%mercat%" and c.CATEGORY_ID=n.CATEGORY_ID union select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.name from test.news n, test.category c where n.TITLE LIKE "%economia%" or n.TITLE LIKE "%mercat%" and c.CATEGORY_ID=n.CATEGORY_ID
Query4	*News su materie prime*	select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.name from test2.news n2, test2.category c2 where n2.CATEGORY_ID=4 and c2.CATEGORY_ID=n2.CATEGORY_ID union select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.NAME from test.news n, test.category c where (n.TEXT REGEXP ".+materi. prim.+" or n.TEXT REGEXP ".+material. grezz.+" or n.TEXT REGEXP ".+row material.+" or n.TEXT REGEXP ".+material. grezzo.+" or n.TEXT REGEXP ".+material. d'opera.+") and c.CATEGORY_ID=n.CATEGORY_ID UNION select n.NEWS_ID, n.PARSING_URL, n.TITLE, n.TEXT, c.NAME from test.news n, test.category c where (n.title REGEXP ".+materi. prim.+"

		or n.title REGEXP ".+material. grezz.+" or n.title REGEXP ".+row material.+" or n.title REGEXP ".+material. grezzo.+" or n.title REGEXP ".+material. d'opera.+") and c.CATEGORY_ID=n.CATEGORY_ID union select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.NAME from test2.news n2, test2.category c2 where (n2.TEXT REGEXP ".+materi. prim.+" or n2.TEXT REGEXP ".+material. grezz.+" or n2.TEXT REGEXP ".+row material.+" or n2.TEXT REGEXP ".+material. grezzo.+" or n2.TEXT REGEXP ".+material. d'opera.+") and c2.CATEGORY_ID=n2.CATEGORY_ID UNION select n2.NEWS_ID, n2.PARSING_URL, n2.TITLE, n2.TEXT, c2.NAME from test2.news n2, test2.category c2 where (n2.title REGEXP ".+materi. prim.+" or n2.title REGEXP ".+material. grezz.+" or n2.title REGEXP ".+row material.+" or n2.title REGEXP ".+material. grezzo.+" or n2.title REGEXP ".+material. d'opera.+") and c2.CATEGORY_ID=n2.CATEGORY_ID

5.3 Analisi dei risultati

N° Query	Risultati DB	Risultati Attesi	Risultati NetWatcher (totali/corretti)	Risultati SEWASIE (totali/corretti)	Precision NetWatcher %	Precision SEWASIE %	Recall NetWatcher %	Recall SEWASIE %
1	95	54	6/5	31/25	83,3	80,6	9,2	46,3
2	14	7		2/2		100		28,6
2BIS	14	7	8/7		87,5		100	
3	209	183	77/58	198/161	75,3	81,3	31,7	88
4	433	428	105/105	363/355	100	97,8	24,5	82,9

Grafico risultati:



Come si può notare dalla tabella dei risultati la query 2, ha due forme:

Query2 e Query2BIS; questo perché nel NetWatcher la stessa versione fatta nel SEWASIE non dava risultati perciò si è passati a un più alto grado di generalizzazione nell'espressione di essa. Da notare inoltre, sempre nella query 2, la bassa percentuale di recall del sistema SEWASIE; questo risultato è causato da un errata politica di integrazione dello schema della GVV di SEWASIE. Nella creazione dello schema infatti abbiamo riscontrato che, le notizie che appartengono alla sezione "news ambiente" del portale italianmec.it sono state erroneamente inserite nella categoria "Università e ricerca" di polimerica.

Dal grafico dei risultati appare che, per quanto riguarda la precision, entrambi i sistemi hanno un comportamento piuttosto buono, essa infatti non va mai sotto al 75% per tutti e due.

Per quanto riguarda la recall abbiamo invece un comportamento piuttosto differente dei due sistemi. Per questo parametro infatti, fatta eccezione per la Query 2, il SEWASIE ha un comportamento decisamente migliore.

Nell'analisi dei risultati va però tenuto conto del fatto che, nell'intenzione di trovare query esprimibili, e che sfruttassero le potenzialità di entrambi i sistemi, si sono andati a selezionare argomenti e criteri di ricerca affini alla struttura dell'ontologia del SEWASIE. Questa metodologia ha quindi influenzato positivamente le prestazioni del SEWASIE, sia dal punto di vista della precision sia della recall. Queste prestazioni infatti, non sarebbero probabilmente raggiungibili se fossero stati utilizzati nei test argomenti non affini alle categorizzazioni presenti nella vista globale di SEWASIE.

Il sistema NetWatcher, per quanto riguarda la recall, ottiene risultati piuttosto bassi, fatta eccezione della query 2. Questo perché, il motore semantico spesso non riesce a categorizzare in modo preciso le informazioni presenti nelle notizie. Questo limite sarebbe stato probabilmente meno significativo se, nelle query fossero stati impostati criteri più generalistici; in tal modo, probabilmente si sarebbero potute sfruttare meglio le capacità di analisi semantica delle informazioni del motore linguistico.

Come si può facilmente intuire dalle considerazioni sopra presentate, la valutazione delle prestazioni di due sistemi è fortemente influenzata dalle metodologie e dai parametri che vengono utilizzati, quindi la scelta di essi è piuttosto importante. In questo senso nei nostri test abbiamo cercato di utilizzare una metodologia che permettesse allo stesso tempo, sia di sfruttare i punti di forza di entrambi i sistemi, sia di valutarli in modo obiettivo.

Conclusioni e lavoro futuro

Questa tesi, partendo da un analisi del contesto del Semantic Web, del quale sono stati esposti i principali concetti e linguaggi, espone le caratteristiche principali di due applicazioni sviluppate nell'ambito di esso: NetWatcher e SEWASIE.

L'obiettivo principale di questa tesi era, dunque, effettuare un analisi dettagliata delle potenzialità del sistema NetWatcher, utilizzando come sistema di confronto il SEWASIE. Queste due applicazioni fanno entrambe parte del ramo di ricerca dell'information retrieval, ovvero la disciplina che studia l'insieme delle operazioni che permettono il recupero di informazioni archiviate in formato elettronico.

Il sistema NetWatcher, utilizzando uno spider, va a indicizzare e archiviare in memoria di massa le pagine Web delle risorse di cui si è interessati. Una volta creata una base consistente di informazioni, è possibile andare ad effettuare delle ricerche all'interno di essa sfruttando la tecnologia semantiche della piattaforma proprietaria COGITO.

SEWASIE, invece, non si limita ad archiviare le pagine web, come vengono presentate nei siti di provenienza, ma, mediante l'utilizzo di wrappers, va ad analizzare la struttura dei dati che sono presenti in esse e, mediante una procedura guidata, mira a integrare questi dati eterogenei tra loro. Una volta creata un vista globale di tutti i dati, è possibile andarli a interrogare in modo omogeneo utilizzando l'interfaccia grafica SQoogle, la quale accompagna l'utente nella composizione di query basandosi sull'ontologia di concetti sottostanti.

Pur essendo presenti differenze sostanziali tra i due sistemi, la serie di test effettuati è stata molto significativa in quanto ha messo in evidenza sia le peculiarità che gli aspetti migliorabili di essi.

Nell'ambito delle news del sistema SEWASIE abbiamo notato che non è presente un alto grado di dettaglio dell'ontologia; i grafi creati dalle ricerche infatti non arrivano a profondità elevate come quelle impostabili in altre sezioni, come quelle riguardanti le aziende o i prodotti meccanici. Questa mancanza è però probabilmente attribuibile al fatto che, l'ambito delle news, al momento dello sviluppo del progetto, non era tra gli obiettivi principali di esso. Un altro aspetto migliorabile del sistema è l'interfaccia grafica, in essa infatti non sono esprimibili richieste con le quali andare a effettuare ricerche con diversi parametri in or tra loro. Inoltre, per poter rendere questo sistema un prodotto di larga scala, sarebbe auspicabile una semplificazione della procedura di composizione delle query per tutti quegli utenti che non hanno dimestichezza coi concetti di classe, attributi e proprietà.

Il sistema NetWatcher non ha come obiettivo la creazione di una struttura precisa di dati interrogabile come un database, ma invece concentra le sue tecnologie nell'analisi semantica dei testi, dai quali andare a estrapolare i concetti e le informazioni importanti. Essendo i dati trattati nei nostri test notizie, quindi testi, possiamo affermare che in linea di principio questo sistema è più adatto a questo utilizzo che SEWASIE. I risultati conseguiti da questo sistema in termini di recall evidenziano però una difficoltà del sistema nell'interpretazione dei concetti quando si ha a che fare con informazioni specifiche di un dato settore, nel nostro caso quello meccanico/plastico. In questo senso sarebbe augurabile un integrazione della rete semantica, sulla quale si basa la capacità di analisi testuale, con una serie di termini di settore. Una volta effettuata l'integrazione della rete semantica sarebbe stato necessario sviluppare tutta una serie di regole, da integrare al motore linguistico, necessarie per consentire ad esso di categorizzare e interpretare, in modo preciso, i testi con linguaggio specifico di settore. Queste attività, che inizialmente rientravano negli obiettivi del tirocinio formativo da me svolto alla CNA servizi di Modena, non sono state realizzabili in quanto avrebbero necessitato di una quantità molto maggiore di tempo e soprattutto della presenza costante di una persona esperta del settore, che mi supportasse nell'integrazione della rete semantica e nella creazione di regole appropriate.

Bibliografia

[1] Tim Berners-Lee. *Semantic web roadmap.* Internal note, 1998. World Wide Web Consortium.
http://www.w3.org/DesignIssues/Semantic.html

[2] Tim Berners-Lee, James A. Hendler, Ora Lassila. *The Semantic We*b. Scientific American, May 2001.
http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

[3] http://www.w3.org/RDF

[4] Dan Brickley ,R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema.* World Wide Web Consortium
http://www.w3.org/TR/rdf-schema

[5] http://www.dbgroup.unimo.it

[6] Deborah L. McGuinness , Frank van Harmelen. OWL Web Ontology Language Overview
http://www.w3.org/TR/2004/REC-owl-features-20040210

[7] World Wide Web Consortium
www.w3.org

[8] R. Guha, Rob McCool, Eric Miller. *Semantic Search*
http://tap.stanford.edu/ess.pdf

[9] http://www.sewasie.org

[10] http://www.dbgroup.unimo.it/sewasie

[11] http://www.dbgroup.unimo.it/momis

[12] S.Bruschi, S.Bergamaschi, F.Guerra. *Dinamica delle ontologie: Inserimento di una nuova sorgente nel sistema Momis*

[13] http://www.expertsystem.it

[14] D. Beneventano, S. Bergamaschi, M. Vincini, *Progetto di Basi di Dati Relazionali: lezioni ed esercizi*

[15] http://dev.mysql.com/doc/refman/5.0/en/