
Sommario dei documenti relativi alla proposta di commessa CNR-IntraWeb presentati dai diversi gruppi (in parentesi).

- 1 - "1_Presentazione_v2.pdf" (ISTC-CNR): Slides introduttive della prima presentazione della proposta IntraWeb - i documenti successivi dettagliano gli accenni forniti nella presentazione;
- 2 - "2_Modello_IntraWeb_v2.pdf" (ISTC-CNR): Descrizione dei requisiti per lo sviluppo dell'IntraWeb Semantico del CNR come sistema di applicazioni a medio termine (allegato 2b_UML-IWS.pdf: modello UML completo);
- 3 - "3_Presentazione della commessa CNR (ISTC-CNR/ILC-CNR) "IntraWeb Semantico";
- 4 - "4_Caratteristiche_motore_di_ricerca_v02.pdf" (SRT-CNR): Descrizione dello stato dell'arte del motore di ricerca intranet del CNR, principali caratteristiche, tecnologie e prodotti utilizzati;
- 5 - "5_SearchEngine_LOA.pdf" (ISTC-CNR): Descrizione della proposta di evoluzione, ricerca e sviluppo, dell'architettura del motore di ricerca CNR;
- 6 - "6_ilc_spec.pdf" (ILC-CNR): Proposte dell'Istituto di Linguistica Computazionale del CNR riguardo la "Gestione e accesso intelligenti del contenuto di repertori documentali mediante l'uso di strumenti avanzati di analisi linguistica automatica";
- 7 - "7_Descrizione_Modulo_LABDOC.pdf" (Universita' della Calabria): Descrizione del modulo Labdoc per la "Gestione documentale e sistemi di classificazione pre coordinati. Costruzione di lessici specialistici strutturati. Indicizzazione.";
- 8 "8_XtermDESCRIZIONE.pdf" (SSLMIT): Descrizione della piattaforma XTerm di content management system sulla base di schede terminografiche;
- 9 - "9_KEExWP3.2.1.pdf" (Dthink): Descrizione del sistema di knowledge management distribuito KEEx della societa' DThink;
- 10 - "10_Mini-progetti_v6.pdf" (ISTC-CNR): Alcune proposte di progetti su task specifici da svolgere in periodi di tempo brevi (2-6 mesi).

Massimiliano Ciaramita, Aldo Gangemi, Domenico Pisanelli, 28/12/2005

IntraWeb Semantico del CNR

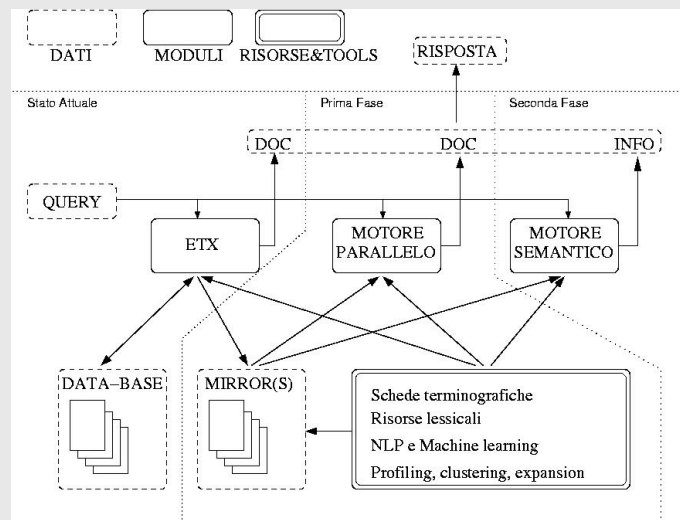
Massimiliano Ciaramita, Aldo Gangemi, Domenico
Pisanelli
ISTC-CNR, Roma

Modello dei requisiti per IWS

- Estrazione e formalizzazione dei requisiti per il *corporate knowledge management* del CNR, secondo lo schema
 - **dominio-task-metodo-applicazione**
- Domini principali: *pagine, documenti, termini, concetti, entità, servizi*
- Vedi diagramma UML



Potenziamento motore di ricerca CNR: Information Retrieval Base: Excalibur + modulo parallelo custom



Componenti

- **Estensione:**
 - Immagini mirror: motore parallelo
 - Estensione query con risorse lessicali aggiuntive
 - Reranking dei risultati (rel. feedback, local clustering, profiling)
 - Generazione dinamica di link (schede terminografiche)
 - Interfaccia semplice (tipo-Google)



Interfaccia semplice

CNR

Consiglio Nazionale delle Ricerche

Directory Ricerca avanzata Navigazione grafica News

LIS



IWS: Requisiti, architettura

5

Ricerca "LIS" in interfaccia semplice

Directory Ricerca avanzata Navigazione grafica News

CNR

LIS

1 Gestualita', oralita' e lingua scritta nello sviluppo e nella lingua dei segni (ISTC)... -> COMMESSA

2 Osservatorio neologico della lingua italiana (Onli) www.cnr.it/istituti/Focus1 ... -> FOCUS

3 UN CD-ROM DECLAMA LE POESIE DEI SORDI. Il CNR ha realizzato il primo CD ... -> NEWS

4 ISTC - Istituto di Scienze e Tecnologie della Cognizione - CNR www.istc.cnr.it/ -> ISTITUTO

5 ILIESI - Istituto per il Lessico Intellettuale Europeo e Storia ... www.iliesi.cnr.it/ -> ISTITUTO

...



IWS: Requisiti, architettura

6

Potenziamento motore di ricerca CNR: Information Extraction

- **Costruzione dell'Ontologia caricata dal motore:**
 - Schede terminografiche: semantica, info lessicale, contesti
- **Information Extraction:**
 - Training di taggers: annotazione di un'immagine apposita con occorrenze di classi
 - IR con informazione semantica: query complesse



Es. da documenti a informazione strutturata

Curriculum Vitae et Studiorum – C.S.

STUDI:

Titolo di dottore in ricerca in "Biologia evoluzionistica: protisti, animali uomo, ecologia marina", conseguito presso l'Università di Pisa.
L'attività di ricerca, svolta dal Gennaio 2001 al Gennaio 2004, [...] ha avuto come obiettivo lo studio della distribuzione del carbonio organico disciolto e della sostanza organica disciolta cromoforica nella acque del Mar mediterraneo. 16/01/2004

↓ Tagging

TITOLI:

Tipo: Dottorato
Data conseguimento: 16 Gennaio 2004
Istituto: Università di Pisa
Disciplina: Biologia
Specifico: ecologia ed evoluzione
Titolo tesi:
Advisor:

Curriculum-1

B.E.R. – CV

Education:

Ph.D. Brown University, Cognitive and Linguistic Sciences 2001
Thesis: Robust Probabilistic Predictive Syntactic Processing
Committee: Mark Johnson (supervisor), Eugene Charniak, Julie Sedivy, Frederick Jelinek

↓ Tagging

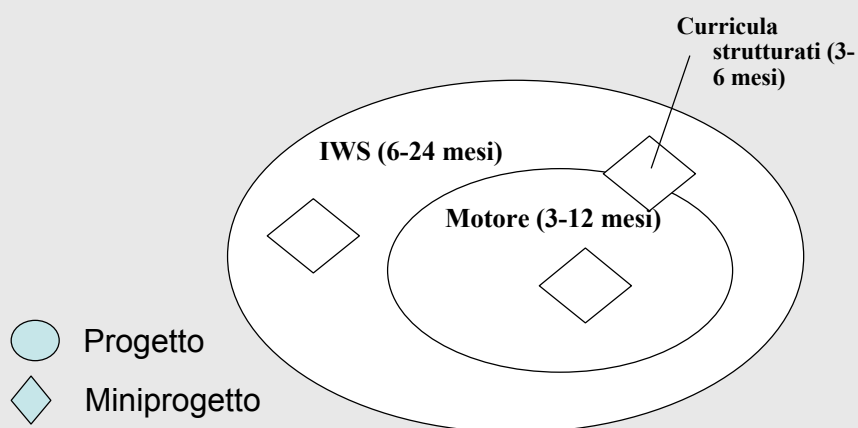
TITOLI:

Tipo: Dottorato
Data conseguimento: 2001
Istituto: Brown University, USA
Disciplina: Scienze Cognitive / Linguistica
Specifico: Linguistica computazionale
Titolo tesi: "Robust Probabilistic ..."
Advisor: Mark Johnson

Curriculum-2



Sviluppo stratificato



Funzionalità del motore e/o di IWS

- Content management
- Motore parallelo: re-ranking documenti, profiling e meta-query, indirizzamento utente multi-lingua
- Motore semantico: mappa documenti (ristrutturazione), validazione documenti, estrazione di relazioni e fatti
- Descrizione semantica CNR: aree tematiche, organizzazione, persone, processi, norme, tecniche. Creazione basi di conoscenza



Miniprogetti possibili

- Strutturazione curricula
- Supporto alla decisione per domande di autorizzazione
- Base di conoscenza commesse (anche con triplete) e/o anagrafica
- Modellazione e linee-guida per contratti specifici
- Workflow manager per attività amministrative
- CNR Wiki
- Integrazione servizi (ex. Bilancio e Commesse o Anagrafica)
- Sviluppo di risorse e strumenti linguistico/semantici



Partner

- **Nucleo di partner coinvolti finora**
 - CNR-SRT, CNR-ISTC, CNR-ILC, UniBologna-SSLMIT, Università della Calabria-LabDoc
- **Partner italiani e stranieri coinvolgibili a breve**
 - DThink, Firenze Tecnologia, OU-KMI, UniKarlsruhe-AIFB, UniTor Vergata, DSI Sapienza
- **Altri partner in una prospettiva di ricerca**
 - UniSheffield, EML-SAP, UniTrento, DSI Sapienza, Centro Studi IBM, ELSAG, ...
- **Possibili sinergie con progetti europei**
 - NeOn, XMedia, SmartWeb, KnowledgeWeb



Possibile distribuzione competenze per motore

- 1. CNR-SRT: sviluppo del motore di ricerca.
- 2. CNR-ISTC: coordinamento scientifico, assistenza al SRT per il trasferimento tecnologico, definizione dell'ontologia con tecniche varie (learning, modelli di expertise -in associazione con esperti CNR, pattern-based), tecniche avanzate di IR e IE per il potenziamento del motore di ricerca.
- 3. CNR-ILC: estrazione di risorse terminologiche dal database del CNR e altre fonti testuali e loro strutturazione e proiezione su classi dell'ontologia attraverso tecniche di machine learning; strumenti di annotazione linguistico-semantica dei documenti basati su tecniche di NLP; analisi delle query utente in linguaggio naturale.
- 4. Laboratorio di Documentazione dell'Università della Calabria: gestione documentale e sistemi di classificazione pre coordinati; costruzione di lessici specialistici strutturati; indicizzazione.
- 5. Gruppo di lavoro SSLMIT dell'Università di Bologna (Forlì): produzione di schede terminografiche per l'annotazione delle occorrenze degli elementi dell'ontologia nel corpus; specifiche terminologiche e contesti d'uso delle classi.



Modello formale preliminare dei requisiti per lo sviluppo dell'IntraWeb Semantico del CNR.

Aldo Gangemi, Massimiliano Ciaramita, Domenico Pisanelli
Laboratorio di Ontologia Applicata, CNR-ISTC, Roma
Versione 2, 20 Dicembre 2005

Cos'è l'IntraWeb Semantico (IWS) del CNR

Un intraweb è un web che comprende tutti i nodi HTTP di una intranet. In molte organizzazioni gli intraweb sono diventate un mezzo privilegiato per la gestione della conoscenza (*corporate knowledge management*).

Le tecnologie semantiche di recente introduzione (ingegneria ontologica, motori inferenziali su logiche descrittive e programmazione logica, linguaggi di marcatura evoluti), usate ancora allo stato prototipale sulle applicazioni del web in generale, possono essere già sfruttate sugli intraweb per le loro dimensioni contenute e la possibilità di accedere a modelli d'uso accessibili e comunità di riferimento abbastanza definite. L'IntraWeb Semantico (IWS) del CNR servirà alla gestione evoluta della *corporate knowledge* del CNR. Sarà un sistema basato su ingegneria ontologica, tecniche di elaborazione del linguaggio naturale e linguaggi progettati per il *semantic web*, e sarà integrato con il sistema di *information retrieval* della ricerca.

Questo documento introduce il modello di requisiti - in base a domini, obiettivi e soluzioni emersi finora, l'architettura di una sua parte (il Motore di Ricerca) e qualche nota sul piano di ricerca e sviluppo.

Il modello dei requisiti

Presentiamo il modello mediante un diagramma di classi UML, qui riportato in frammenti per ragioni di spazio.

Una classe (rettangolo) indica un tipo di elementi. Un'associazione fra classi (freccia semplice) indica una relazione fra quegli elementi, secondo una certa cardinalità (0..*, 1..*). Quando non indicata, è 0..*.

Una freccia con il triangolo pieno indica una generalizzazione, cioè una relazione di sottoclasse.

Una freccia con la losanga indica un'aggregazione, cioè una relazione di parte.

Il package “guida” (Fig. 1) introduce i tipi di classi presentate nel diagramma.

I domini (sfondo blu) sono gli insiemi di elementi a cui un'applicazione si rivolge.

I task (sfondo arancio) sono gli obiettivi dell'applicazione rispetto a un certo dominio. Un task è quindi “svolto_su” un dominio.

I metodi (sfondo celeste) sono le soluzioni fornite per soddisfare un task rispetto a un dominio. Un metodo è quindi “applicato_a” un task.

I tipi di sistema o risorsa (sfondo grigio) sono i tipi di implementazione per un certo metodo. Un tipo di sistema quindi “implementa” un metodo.

Le istanze di sistemi o risorse (sfondo giallo-verde) sono le specifiche implementazioni fornite da un provider.

Queste distinzioni sono basate su un paradigma di modellazione chiamato PSM (Problem-Solving Methods), introdotto negli anni 80 con la metodologia KADS e successivamente raffinato in base ai "Models of Expertise" di Luc Steels e alle ricerche della comunità KA (Knowledge Acquisition).

Negli ultimi anni il paradigma è usato nelle applicazioni più avanzate dell'ingegneria ontologica e del Semantic Web.



Fig. 1. Le classi guida nel paradigma PSM.

Il package “CNR IntraWeb” introduce le sottoclassi delle classi introdotte nella guida, che rappresentano parte dei domini emersi nell’analisi dei requisiti per il sistema di *corporate knowledge management* del CNR. Le associazioni fra queste classi, dove non indicate, sono le stesse definite fra le classi introdotte nella guida.

I domini considerati dai requisiti sono: Pagine web, Documenti, Termini, Concetti, Entità, Servizi. Per ognuno di questi domini sono emersi alcuni task, con relativi metodi, tipi di sistemi o risorse e istanze di quei sistemi o risorse.

Le *pagine web* (Fig. 2) sono documenti che “codificano” altri documenti, nel senso che, una volta richiamate mediante uno URI, ne permettono la realizzazione fisica su un supporto elettronico.

I task svolti (svolgibili) sulle pagine web includono almeno: Presentazione, Versioning, Generazione. I *Content Management System* (CMS) implementano di solito metodi applicati a quei task. L’uso di tecnologie semantiche per i CMS è attivamente studiato in questo periodo e ci sono già delle soluzioni interessanti (vedi sezione sui partner di progetto).

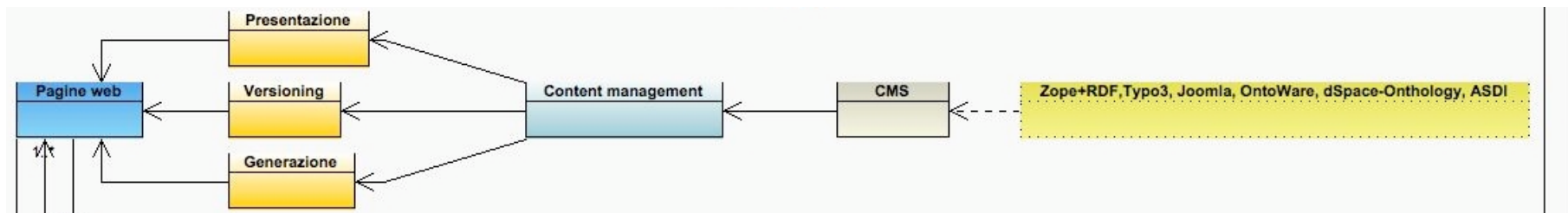


Fig. 2. Le classi nel PSM relativo alle pagine web.

I *documenti* (Fig. 3) sono oggetti informativi come testi, immagini, multimedia, etc. In questo modello ci si riferisce prevalentemente a testi (strutturati, semi- o non-strutturati). Sotto-classi rilevanti di documenti sono i *curricula*, i testi delle *commesse* e le *domande di autorizzazione*. I documenti vengono realizzati da un supporto fisico, per esempio da pagine web. I testi sono costituiti da termini (tra l'altro).

I documenti sono il domio di riferimento di molti task: Ricerca filtrata, Marcatura, Generazione di mappe, Validazione in base a strutture pre-definite o emergenti, Generazione viste, Estrazione d'informazione.

Per esempio, alcuni tipi di documenti (e relativi task) notevoli - probabilmente adatti per dei miniprogetti a breve scadenza (vedi sezione sui tempi di sviluppo) - sono:

I *curricula*, che possono essere strutturati in un formato unico a partire da strutture non-predefinite dopo una fase di training per creare mappature fra le strutture originali e una struttura predefinita (o emergente).

I testi delle *commesse*, che possono essere indicizzati secondo l'ontologia generica del CNR. Possono inoltre essere analizzati e modellati secondo strutture concettuali come le triplete "Area-Processo-Tecnica".

I testi delle *domande di autorizzazione*, che possono essere oggetto di analisi e *clustering* dopo una fase di *training*. Il *clustering* può essere poi utilizzato per creare un servizio di assistenza alla decisione sulle autorizzazioni.

I *log accessi*, che sono liste di registrazione di accessi con i dati dell'agente che accede a una pagina web. Se l'agente si registra come utente, allora è facile classificarlo in un profilo nonché costruire i profili incrementalmente. Altrimenti la profilazione può essere generata statisticamente.

I *log query*, che sono liste di *query* con i dati dell'agente che le ha proposte. Le *query* vengono usate per il riconoscimento della lingua-utente e possono essere analizzate per costruire repertori di “meta-query”, che a loro volta possono essere usate per generare configurazioni ottimali di un sito e dell'informazione contenuta (*community-based design*).

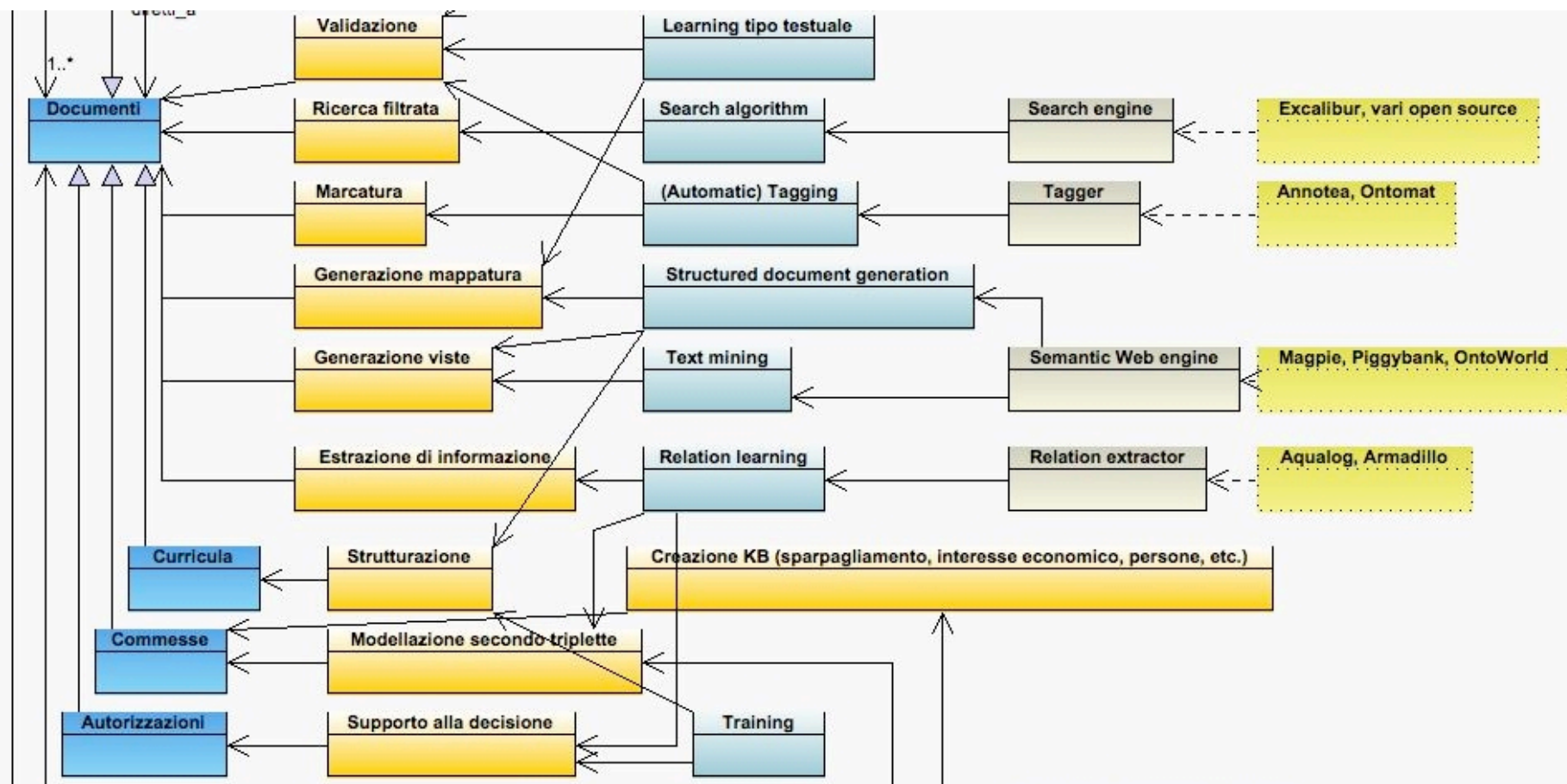


Fig. 3. Le classi del PSM relativo ai documenti.

I *termini* (Fig. 4) sono tra i costituenti dei documenti testuali. Tipici task svolti sui termini sono: Estrazione, Gestione, Uso in IR (*Information Retrieval*). Lessici e schede terminografiche sono metodi per la loro gestione in IWS.

I termini possono esprimere concetti e servono quindi anche a “lessicalizzare” i concetti rappresentati in un'ontologia.

I *concetti* (Fig. 4) sono costrutti cognitivo-sociali che si riferiscono a entità qualsiasi (oggetti fisici, sociali, eventi, astrazioni, informazioni, etc.). Sono espressi da termini e formalizzati da ontologie.

Tipici task svolti sui concetti sono la Gestione mediante ontologie e il loro Uso in sistemi di gestione della conoscenza, come l'IWS del CNR.

Le *entità* (Fig. 4) sono per esempio oggetti fisici, sociali, eventi, astrazioni, informazioni, etc. Il tipico task da svolgere su un dominio di entità qualsiasi è la Descrizione.

Le ontologie permettono descrizioni di tipo diverso a seconda del punto di vista o del tipo di entità prese in considerazione: descrizione *tematica* (per aree di attività, interesse, etc.), *statica* (per tipi di oggetto), *dinamica* (per tipo di processi o eventi), *regolativa* (per piani, norme, tecniche), etc.

Un'ontologia del CNR sarà costruita a partire dall'analisi dell'organizzazione attuale, degli esempi esistenti di ontologie per le organizzazioni di ricerca e soprattutto della necessità di strutturare la conoscenza in funzione di domande specifiche (*competency questions*) che si immagina di porre all'IWS.

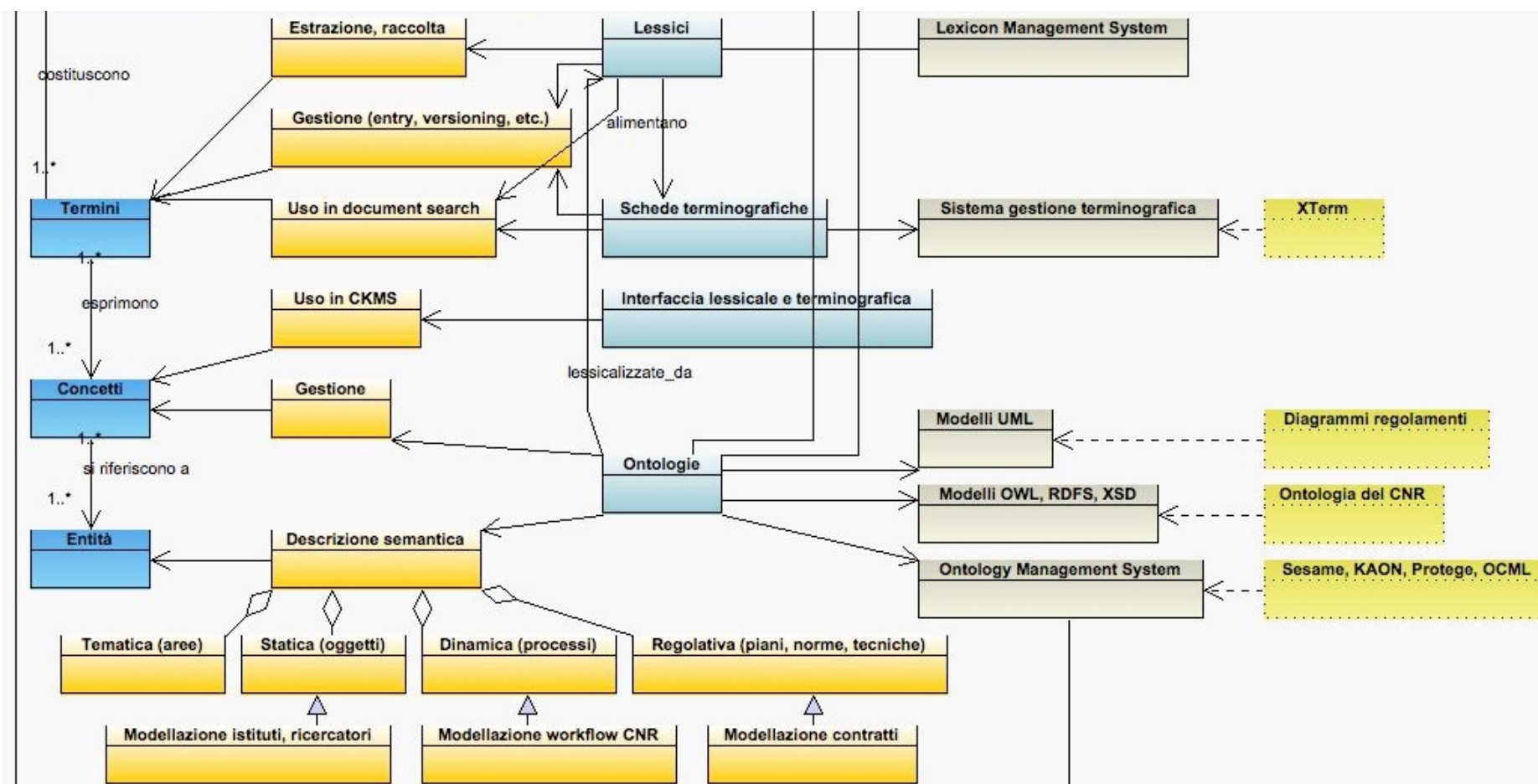


Fig. 4. Le classi dei PSM relativi a termini, concetti ed entità descritte dalle ontologie.

Un'applicazione (Fig. 5) è un sistema informativo che svolge un *servizio* specifico. Il task di Integrazione dei servizi richiede una descrizione semantica del servizio. La descrizione semantica fornita da ontologie permette poi la scoperta, il matching, e la composizione dei servizi.

Tipici applicazioni/servizi integrabili mediante un framework semantico sono: motori di ricerca, gestori di workflow, applicazioni di bilancio e gestione commesse o anagrafica, servizi comunitari come wiki e blog, etc. La gestione integrata dei servizi mediante tecnologie semantiche può anche avere un effetto di “moltiplicazione” sulla *corporate knowledge* estratta e strutturata. Per esempio, una Wikipedia personalizzata (e gestita semanticamente) per le attività di ricerca e gestione del CNR potrebbe portare alla creazione di conoscenza di alta qualità all’interno di IWS, senza fatica e con un semplice coinvolgimento da parte dei ricercatori e degli amministrativi. Anche le ricadute educative sul Web in generale sarebbero rilevanti.

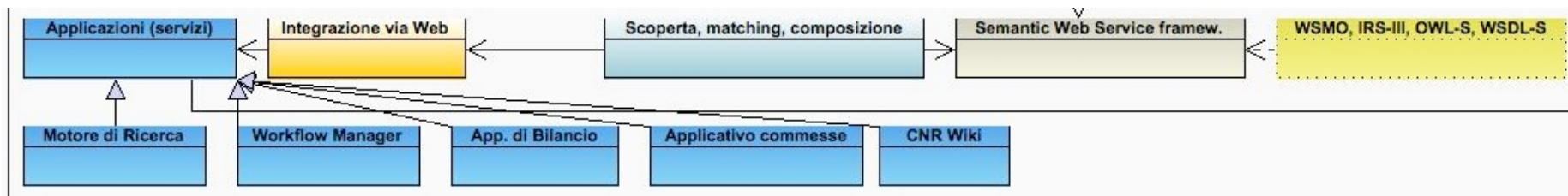


Fig. 5. Le classi del PSM relativo ai servizi e alla loro integrazione semantica.

Il motore di ricerca evoluto

Una buona parte del sistemi e delle risorse presentate nel modello dei requisiti sono coinvolte nel progetto preliminare di *motore di ricerca semantico*. L'allegato 1 presenta in un dettaglio maggiore un'architettura, degli esempi e una proposta di distribuzione di lavoro fra i partner finora coinvolti.

I partner di progetto

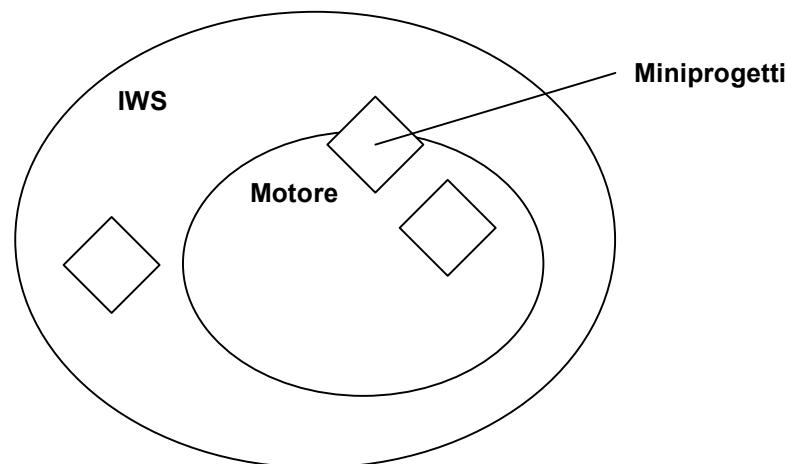
Una volta avviata la commessa, oltre agli attori finora coinvolti (CNR-SRT, CNR-ISTC, CNR-ILC, UniBologna-SSLMIT, Università della Calabria-LabDoc), è possibile coinvolgere altri partner italiani e stranieri, costruendo per esempio sinergie a breve termine con progetti europei come NeOn (*Networked Ontologies*, CNR-ISTC è partner), XMedia (knowledge management ed estrazione della conoscenza, CNR-ISTC ha accordo di collaborazione con Università di Sheffield, coordinatore), SmartWeb (applicazioni avanzate e distribuite di semantic web, partner tedeschi, accordo di collaborazione CNR-ISTC con AIFB e EML), etc. In particolare, il Knowledge Media Institute della Open University e l'AIFB di Karlsruhe hanno già mostrato interesse per collaborare con tecnologie di semantic web a un progetto su IWS. In Italia, un certo

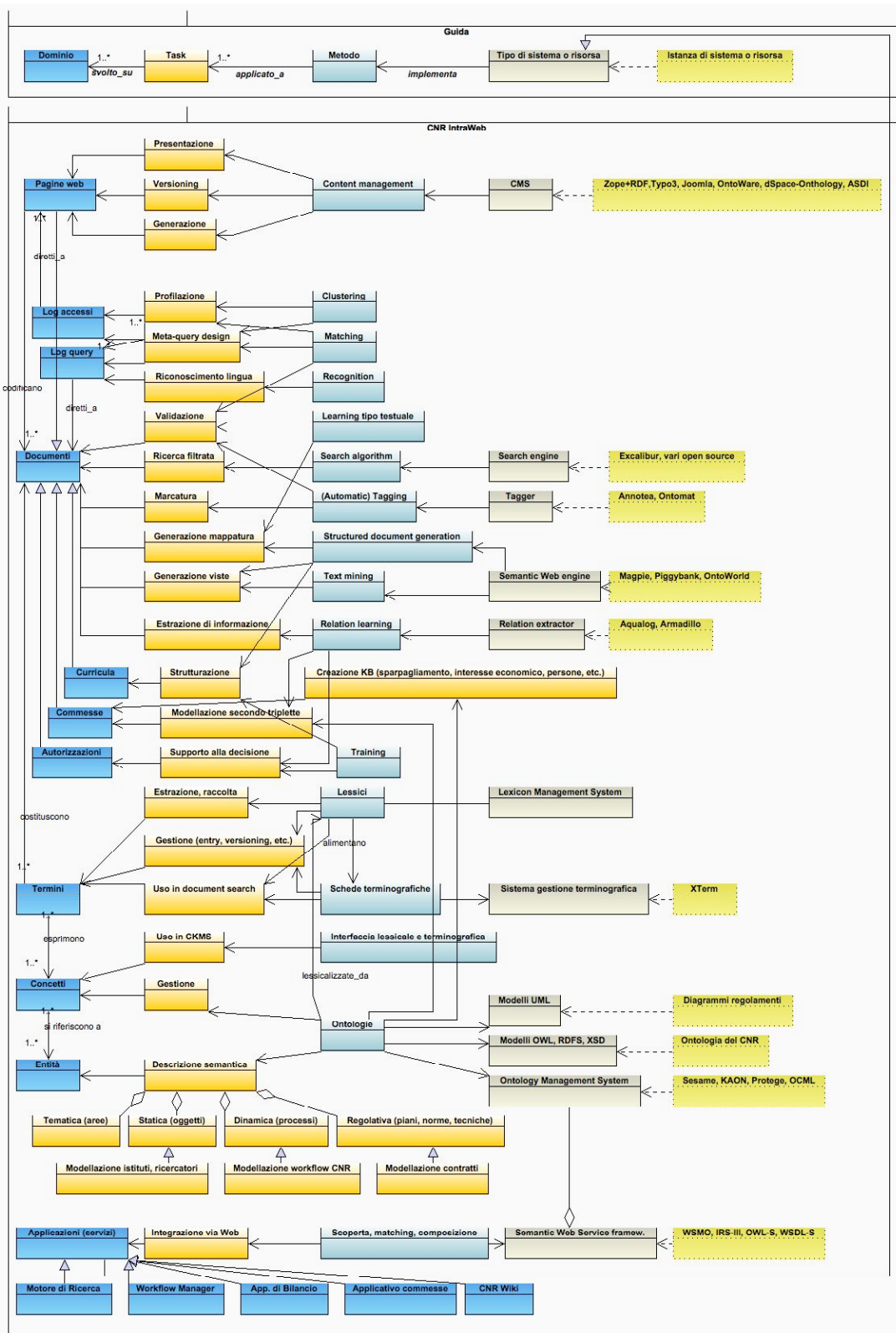
numero di partner accademici (DIS alla Sapienza, Università di Tor Vergata, Università di Trento, etc.) e industriali (Distributed Thinking, Firenze Tecnologia, IBM Centro Studi, etc.), che possiedono tecnologie e risorse interessanti per IWS, è coinvolgibile sia a breve termine, per esempio riusando prodotti esistenti, sia a medio termine mediante sinergie con progetti di ricerca industriale e/o accademica.

I tempi di sviluppo

Non è ovviamente possibile formulare un Gantt senza la conoscenza di chi fa cosa con quali risorse, che andrà dettagliato nella fase di avvio della commessa dal gennaio 2006. Possiamo tuttavia stimare i tempi di sviluppo se stratifichiamo il progetto in tre cicli di vita:

- a) la visione di IWS: è il progetto completo di IntraWeb semantico, la cui realizzazione è da stimare in 24-36 mesi. In questo tempo si intende anche il raggiungimento di accordi di collaborazione con organizzazioni e progetti italiani ed esteri;
- b) il motore di ricerca evoluto: è un sottinsieme notevole dei componenti di IWS, le cui funzionalità possono essere realizzate incrementalmente in un periodo fra 3 e 12 mesi; I partner del nucleo coinvolto finora hanno le risorse e i metodi per la sua realizzazione;
- c) i miniprogetti: sono “progettini” (ex. KB commesse, supporto alle autorizzazioni, normalizzazione curricula, modellazione contratti specifici, KB CNR - anagrafico, delle linee-guida, etc - , CNR Wiki, etc.) per i quali una *task force* specializzata può ottenere risultati interessanti in tempi fra 3 e 6 mesi, senza attendere il completamento del motore o di IWS; una *task force* può includere anche partner esterni al nucleo o utilizzare componenti sub-ottimali ma prontamente reperibili e riusabili.







Previsione 2006-2008

**SCIENZE E TECNOLOGIE DELLA
COGNIZIONE**Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:04:26**Commessa:** (ICT.P04.019) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse**Moduli:** ICT.P04.019.001 ICT.P04.019.002**Anagrafica della Proposta di Commessa**

Dipartimento: ICT (ICT)
Progetto: ICT.P04 / Tecnologia della conoscenza e servizi avanzati
Commessa: ICT.P04.019 / IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse
Stato: Nuova proposta
Istituto esecutore: Istituto di scienze e tecnologie della cognizione (ISTC)
Primo anno di attività: 2006
Anno chiusura previsto: 2008
Tipologia di ricerca Progetti relativi a linee tematiche a carattere strategico
Responsabile della commessa
Codice terzo SIGLA: 795
Cognome: GANGEMI Nome: ALDO
Email: aldo.gangemi@istc.cnr.it Telefono: +390644161535
Sede principale svolgimento attività: Sede principale Istituto
Parole chiave: web semantico trattamento automatico del linguaggio naturale gestione della conoscenza
Descrittori sintetici: Campo non attivo
Abstract
Un intraWeb è un web che comprende tutti i nodi HTTP di una intranet. In molte organizzazioni gli intraWeb sono un mezzo privilegiato per la gestione della conoscenza (corporate knowledge management). Le tecnologie semantiche possono essere sfruttate sui documenti presenti in un intraWeb grazie alle dimensioni contenute del corpus, la disponibilità di modelli d'uso e la presenza di comunità di riferimento definite. Come caso di studio, si intende costruire l'IntraWeb Semantico (IWS) per la gestione evoluta della corporate knowledge del CNR. IWS è un sistema basato su information retrieval avanzato, tecniche di elaborazione del linguaggio naturale, machine learning, ingegneria ontologica e linguaggi progettati per il semantic web. Le funzionalità di un intraWeb semantico comprendono: il potenziamento del motore di ricerca terminologico con componenti morfologici, multi-lingua e modellazione dei log, l'integrazione di servizi, il supporto alla decisione su documenti digitalizzati, la mappatura di documenti su un modello prototipico, la creazione di basi di conoscenza, la modellazione di linee-guida per contratti e workflow, la creazione di know-how comunitario (ex. semantic wiki).

[Indietro](#)



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:05:01

Commessa: (ICT.P04.019) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Moduli: ICT.P04.019.001 ICT.P04.019.002

Descrizione Analitica della Commessa

Tematiche di ricerca

Ricerca di informazione su basi di dati testuali non strutturate tramite analisi semantica, automatica e semi-automatica, dei testi. Ricerca di documenti attraverso tecniche avanzate di interazione utente-motore, per es. tramite profiling, espansione semantica delle query, relevance feedback, local clustering e reranking dei documenti. Valutazione della performance del motore e sua ottimizzazione attraverso l'analisi delle transazioni precedenti (backlog). Costruzione di un'ontologia dell'organizzazione che guidi il processo di ricerca di informazione tramite la definizione di classi, relazioni, e strutture informative rilevanti. Estrazione di informazione, identificazione di contenuti informativi specifici all'interno di uno o più documenti; trattamento automatico del linguaggio per l'identificazione di occorrenze di informazione strutturata nei documenti. Costruzione di strumenti di annotazione automatica di documenti, classificatori in domini di classi strutturate (es. sequenze, tassonomie), e risorse di supporto (terminologie, lessici, tesauri). Modelli di navigazione e ricerca semantica (semantic web, logiche descrittive, sistemi basati su regole, etc.).

Stato dell'arte

La commessa integra modelli stato dell'arte in information retrieval, text mining e ontology engineering, tramite un approccio misto knowledge-based e machine learning, concettuale e automatico-statistico, per l'accesso all'informazione. La base di conoscenza del sistema è un modello concettuale riusabile del dominio di ricerca, di cui alcuni campioni vengono proiettati manualmente su documenti del database. Tecniche di machine learning vengono poi utilizzate per indicizzare l'informazione semantica contenuta nel database su larga scala, cioè sulla totalità dei documenti. L'approccio combina la precisione tipica del lavoro di esperti di dominio e ontologi con la flessibilità e la copertura delle tecniche automatiche e semi automatiche di annotazione e classificazione di testi. L'indicizzazione su larga scala permette la costruzione di query complesse e strutturate sul dominio di ricerca. Il paradigma inoltre integra la valutazione quantitativa dell'efficacia dei metodi sviluppati come elemento fondamentale di ricerca e sviluppo. Altri componenti del sistema servono alla gestione di contratti e workflow e strumenti per la creazione di know-how comunitario.

Competenze, tecnologie e tecniche di indagine

Sono coinvolti principalmente due gruppi di ricerca: uno dall'ISTC-CNR e uno dall'ILC-CNR. Il Laboratorio di Ontologia Applicata (LOA) ha un ruolo di riferimento internazionale nell'ingegneria ontologica e nelle sue applicazioni, oltre a competenze specifiche nell'applicazione di metodi di machine learning e natural language processing combinati con ontologie. L'Istituto di Linguistica Computazionale del CNR, leader in Italia, e con una consolidata visibilità internazionale, nella linguistica computazionale, ha vaste e approfondite conoscenze nell'ambito del trattamento automatico del linguaggio e nella costruzione di risorse lessicali.

Collaborazioni (partner e committenti)

Partner: ISTC-CNR, ILC-CNR. Altri partner mediante convenzioni: LabDoc Università della Calabria (classificazione documentale), SSLMIT di Forlì (Università di Bologna) (risorse terminologiche). Collaborazioni italiane: ITTIG-CNR Firenze (ontologie e terminologie legali), ICIB-CNR Napoli (CMS semantici), Università di Roma 1 (DI, DSI) (NLP, ragionamento).

automatico), Università di Roma 2 (estrazione dell'informazione), Università di Trento (web semantico, ragionamento contestuale), Università di Bologna, Dip. informatica (web semantico). Collaborazioni internazionali con istituzioni: Toyota Technological Institute Chicago (machine learning), European Media Lab Heidelberg (interfacce intelligenti), KMI Open University Milton Keynes (servizi semantici, ingegneria ontologia), AIFB Università di Karlsruhe (CMS semantici), Università di Sheffield (NLP), Università di Madrid (ingegneria ontologica). Collaborazioni con progetti: NeOn (EU-FP6), XMedia (EU-FP6), KnowledgeWeb (EU-FP6), SmartWeb (Germania). Collaborazioni con aziende: Centro Studi IBM (tecnologie semantiche), Elsas (sistemi logistica, sicurezza), ISOCO (tecnologie semantiche), Firenze Tecnologia (motori di ricerca).

Potenziale impiego

- per processi produttivi

Le tecnologie sviluppate dalla commessa hanno una immediata applicabilità alle necessità di gestione informativa del CNR e di altri enti/organizzazioni con simili risorse intranet. La prospettiva è l'integrazione e la fruibilità sull'(intra)web di servizi esistenti e nuovi, mediante la loro ricerca, confronto e composizione automatica o semi-automatica. La creazione di nuovi servizi si concentra sulle funzioni avanzate del motore di ricerca, la migrazione di conoscenza (sia statica sia procedurale) sparsa e semi-strutturata in basi di conoscenza armonizzate con l'ontologia delle organizzazioni, e la reingegnerizzazione di formati documentali.

- per risposte a bisogni individuali e collettivi

La commessa riguarda la ricerca e sviluppo di tecnologie di accesso all'informazione che hanno una rilevanza sia scientifica sia sociale, individuale e collettiva, tramite lo sviluppo di interfacce e strumenti che facilitino l'accesso a basi di dati (strutturate e non), ampie ma circoscritte come le intranet, in una prospettiva a medio termine, anche sul Web. Ulteriori componenti integrati, per esempio una wiki semantica, rispondono a bisogni comunicativi individuali o di piccoli gruppi e alla raccolta di know-how locale.

Obiettivi

L'obiettivo di questa commessa è sviluppare una sofisticata piattaforma di gestione semantica dell'informazione contenuta all'interno di una intranet, e in prospettiva nel web. Componenti di questa architettura sono moduli individuali di immediata applicazione che progressivamente arricchiscono un motore di ricerca e un sistema di gestione del contenuto di nuova generazione. Gli obiettivi a breve-medio termine saranno scelti fra i seguenti: backlog, generalizzazione di query utente, riformulazione di query (query expansion) e ricerca basata su tecniche di elaborazione del linguaggio naturale (NLP), reingegnerizzazione di documenti semi-strutturati, supporto alla decisione su tipi documentali specifici (ex. richieste di autorizzazione), sviluppo di basi di conoscenza sensibile per l'organizzazione (ex. anagrafica, temi, progetti, regolamenti), modellazione di contratti e flussi di lavoro (workflow), acquisizione di know-how mediante una wiki semantica dell'organizzazione.

[Indietro](#)



Previsione 2006-2008

**SCIENZE E TECNOLOGIE DELLA
COGNIZIONE**Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:05:50**Commessa:** (ICT.P04.019) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse**Moduli:** ICT.P04.019.001 ICT.P04.019.002**Previsione attività della Commessa****Attività da svolgere**

1. Analisi e definizione di alcuni task relativi a tipologie specifiche di informazione strutturata, es. analisi ed estrazione di informazione da curricula, descrizioni di commesse etc. 2. Costruzione/implementazione delle risorse necessarie all'estrazione di informazione: risorse terminologiche, modelli ontologici, annotatori automatici, interfacce di query, etc. 3. Sviluppo del motore di ricerca su testi di base con funzionalità aggiuntive e integrazione con i componenti di estrazione di informazione.

Punti critici e azioni da svolgere

1. Implementazione di un'interfaccia di ricerca e sviluppo; content management delle risorse create mediante le nuove tecnologie riguardanti il motore di ricerca 2. Specificazione dell'agenda di lavoro coordinata fra i partner e implementazione della pipeline di sviluppo 3. Modellazione di organizzazioni e flussi di lavoro 4. Valutazione dei risultati

Risultati attesi nell'anno

1. Realizzazione di alcuni task di estrazione/trasformazione di informazione strutturata comprendenti analisi qualitative/quantitative delle performance ottenute nei vari task. 2. Upgrade del motore di ricerca di base a stato dell'arte con tecniche di machine learning e trattamento automatico del linguaggio. 3. Definizione dell'agenda di ricerca e sviluppo del motore di ricerca semantico per task generici. 4. Definizione di ontologie organizzative per il CNR.

Iniziative per l'acquisizione di ulteriori entrate

Un'iniziativa per l'acquisizione di fondi di supporto alle richieste di personale, macchine, viaggi, etc. necessari allo svolgimento della commessa è in fase di negoziazione con il Ministero dell'Innovazione Tecnologica (MIT) nell'ambito di un accordo con la Confartigianato. Altre iniziative sono previste nell'ambito dell'ultima call del Sesto Programma Quadro della Comunità Europea e degli altri bandi nazionali e internazionali per la ricerca e l'innovazione industriale.

[Indietro](#)



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:06:51

Modulo: (ICT.P04.019.001) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Commessa: ICT.P04.019

Anagrafica della Proposta di Modulo

Dipartimento: ICT/ICT
Progetto: ICT.P04 / Tecnologia della conoscenza e servizi avanzati
Commessa: ICT.P04.019 / IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse
Modulo: ICT.P04.019.001 / IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse
Istituto esecutore della commessa: Istituto di scienze e tecnologie della cognizione (ISTC)
Istituto esecutore del modulo: Istituto di scienze e tecnologie della cognizione (ISTC)
Primo anno attività: 2006
Anno di chiusura previsto: 2008
Stato: Nuova proposta
Tipologia di ricerca: Progetti relativi a linee tematiche a carattere strategico
Responsabile della commessa
Codice terzo SIGLA: 795
Cognome: GANGEMI Nome: ALDO
Email: aldo.gangemi@istc.cnr.it Telefono: +390644161535
Sede principale svolgimento attività: Sede principale Istituto
Sedi partecipanti:
Parole chiave: web semantico e gestione della conoscenza trattamento automatico del linguaggio naturale machine learning
Descrittori sintetici: Campo non attivo
Abstract
Il lavoro dell'ISTC sullo sviluppo di intraWeb semantici si concentra su web semantico, gestione della conoscenza, costruzione di ontologie e modelli di organizzazioni, trattamento automatico del linguaggio naturale e tecniche di machine learning. Indichiamo tutto questo come "tecnologie semantiche". Le tecnologie semantiche possono essere sfruttate sui documenti presenti in un intraWeb grazie alle dimensioni contenute del corpus, la disponibilità di modelli d'uso e la presenza di comunità di riferimento definite. Nell'ambito del caso di studio (IWS),

l'ISTC coordinerà la commessa e parteciperà ai progetti focalizzati sulle tematiche di sua competenza. In particolare: l'architettura del motore di ricerca semantico, l'integrazione di servizi, il supporto alla decisione su documenti digitalizzati, la mappatura di documenti su un modello prototipico, la creazione di ontologie per il CNR e le organizzazioni in generale, la modellazione di linee-guida per contratti e workflow, la creazione di know-how comunitario (ex. semantic wiki).

[Indietro](#)



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:08:18

Modulo: (ICT.P04.019.002) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Istituto esecutore: Istituto di linguistica computazionale (ILC)

Commessa: [ICT.P04.019](#)

Dati Generali

Anagrafica

Previsione

Descrizione

Anagrafica della Proposta di Modulo

Dipartimento: ICT/ICT
Progetto: ICT.P04 / Tecnologia della conoscenza e servizi avanzati
Commessa: ICT.P04.019 / IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse
Modulo: ICT.P04.019.002 / IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse
Istituto esecutore della commessa: Istituto di scienze e tecnologie della cognizione (ISTC)
Istituto esecutore del modulo: Istituto di linguistica computazionale (ILC)
Primo anno attività: 2006
Anno di chiusura previsto: 2008
Stato: Nuova proposta
Tipologia di ricerca: Progetti relativi a linee tematiche a carattere strategico
Responsabile della commessa
Codice terzo SIGLA: 169
Cognome: PIRRELLI Nome: VITO
Email: vito.pirrelli@ilc.cnr.it Telefono: 0503152851
Sede principale svolgimento attività: Sede principale Istituto
Sedi partecipanti: Sede principale Istituto
Parole chiave: Web semantico Trattamento automatico della lingua Gestione della conoscenza ontology learning
Descrittori sintetici: Campo non attivo
Abstract
La necessità quotidiana di accedere a grandi quantità di conoscenza digitale non strutturata

contenuta in basi documentali di dominio, disponibili sul Web o su Intranet aziendali, ha dato crescente impulso allo sviluppo di tecnologie per l'acquisizione e gestione automatiche dell'informazione testuale, molte delle quali hanno ormai superato la fase di prototipazione per giungere a una piena commercializzazione in prodotti diretti verso settori di mercato quali il technology watch, i motori di ricerca sul Web e i sistemi di supporto decisionale. Nonostante gli evidenti successi conseguiti, tuttavia, le tecnologie esistenti sono relativamente rigide e poco scalabili a causa di due fattori fondamentali: 1) la rudimentale modellizzazione formale della conoscenza (dichiarativa e procedurale) condivisa, 2) l'insufficiente granularità dell'analisi linguistica dei testi. Obiettivo del presente modulo è quello di consentire un accesso avanzato all'informazione documentale attraverso l'integrazione dinamica di tecnologie per l'intelligenza linguistica del contenuto, tecniche di machine learning e strumenti evoluti di rappresentazione ontologica della conoscenza di dominio.



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:07:21

Modulo: (ICT.P04.019.001) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Commessa: ICT.P04.019

Descrizione Analitica del Modulo

Tematiche di ricerca

Le tematiche dell'ISTC nella commessa includono: costruzione di un'ontologia delle organizzazioni che guidi il processo di ricerca di informazione tramite la definizione di classi, relazioni, e strutture informative rilevanti. Estrazione di informazione, identificazione di contenuti informativi specifici all'interno di uno o più documenti; trattamento automatico del linguaggio per l'identificazione di occorrenze di informazione strutturata nei documenti. Modelli di navigazione e ricerca semantica (semantic web, logiche descrittive, sistemi basati su regole, etc.). Ricerca di informazione su basi di dati testuali non strutturate tramite analisi semantica, automatica e semi-automatica, dei testi.

Stato dell'arte

Il modulo, seguendo la commessa, integra modelli stato dell'arte in information retrieval, text mining e ontology engineering, tramite un approccio misto knowledge-based e machine learning, concettuale e automatico-statistico, per l'accesso all'informazione. La base di conoscenza del sistema è un modello concettuale riusabile del dominio di ricerca, di cui alcuni campioni vengono proiettati manualmente su documenti del database. Tecniche di machine learning vengono poi utilizzate per indicizzare l'informazione semantica contenuta nel database su larga scala, cioè sulla totalità dei documenti. L'approccio combina la precisione tipica del lavoro di esperti di dominio e ontologi con la flessibilità e la copertura delle tecniche (semi) automatiche di annotazione e classificazione di testi. L'indicizzazione su larga scala permette la costruzione di query complesse e strutturate sul dominio di ricerca. Il paradigma inoltre integra la valutazione quantitativa dell'efficacia dei metodi sviluppati come elemento fondamentale di ricerca e sviluppo. Altri componenti del sistema servono alla gestione di contratti e workflow e strumenti per la creazione di know-how comunitario.

Competenze, tecnologie e tecniche di indagine

L'ISTC-CNR ha competenze e tecnologie in intelligenza artificiale, modelli cognitivi, ontologia formale, linguistica, trattamento automatico del linguaggio, gestione della conoscenza, etc. Il Laboratorio di Ontologia Applicata (LOA) in seno all'ISTC-CNR ha un ruolo di riferimento internazionale nell'ingegneria ontologica e nelle sue applicazioni sul web semantico, la modellazione concettuale, i sistemi informativi di organizzazioni e comunità, oltre a competenze specifiche nell'applicazione di metodi di machine learning e natural language processing combinati con ontologie. Tali competenze saranno integrate per l'analisi della conoscenza organizzativa e procedurale del CNR (nel caso di studio) e in generale delle organizzazioni caratterizzate da ricche basi di conoscenza (dati statici, know-how, programmi di ricerca e sviluppo). L'ibridazione di tecniche di modellazione della conoscenza (ontologie, design patterns) con tecniche di apprendimento automatico di strutture da dati distribuiti è uno degli approcci originali dei ricercatori coinvolti nel modulo.

Collaborazioni (partner e committenti)

Partner: ISTC-CNR, ILC-CNR. Altri partner mediante convenzioni: LabDoc Università della Calabria (classificazione documentale), SSLMIT di Forlì (Università di Bologna) (risorse terminologiche). Collaborazioni italiane: ITTIG-CNR Firenze (ontologie e terminologie legali),

ICIB-CNR Napoli (CMS semantici), Università di Roma 1 (DI, DSI) (NLP, ragionamento automatico), Università di Roma 2 (estrazione dell'informazione), Università di Trento (web semantico, ragionamento contestuale), Università di Bologna, Dip. informatica (web semantico). Collaborazioni internazionali con istituzioni: Toyota Technological Institute Chicago (machine learning), European Media Lab Heidelberg (interfacce intelligenti), KMI Open University Milton Keynes (servizi semantici, ingegneria ontologia), AIFB Università di Karlsruhe (CMS semantici), Università di Sheffield (NLP), Università di Madrid (ingegneria ontologica). Collaborazioni con progetti: NeOn (EU-FP6), XMedia (EU-FP6), KnowledgeWeb (EU-FP6), SmartWeb (Germania). Collaborazioni con aziende: Centro Studi IBM (tecnologie semantiche), Elsas (sistemi logistica, sicurezza), ISOCO (tecnologie semantiche), Firenze Tecnologia (motori di ricerca).

Potenziale impiego

- per processi produttivi

Le tecnologie sviluppate in questo modulo per la commessa hanno una immediata applicabilità alle necessità di gestione informativa del CNR e di altri enti/organizzazioni con simili risorse intranet. La prospettiva è l'integrazione e la fruibilità sull'intranet di servizi esistenti e nuovi, mediante la loro ricerca, confronto e composizione automatica o semi-automatica. La creazione di nuovi servizi si concentra sulle funzioni avanzate del motore di ricerca, la migrazione di conoscenza (sia statica sia procedurale) sparsa e semi-strutturata in basi di conoscenza armonizzate con l'ontologia delle organizzazioni, e la reingegnerizzazione di formati documentali.

- per risposte a bisogni individuali e collettivi

Il modulo si identifica nei potenziali impieghi della commessa, in quanto questa riguarda la ricerca e sviluppo di tecnologie di accesso all'informazione che hanno una rilevanza sia scientifica sia sociale, individuale e collettiva, tramite lo sviluppo di interfacce e strumenti che facilitino l'accesso a basi di dati (strutturate e non), ampie ma circoscritte come le intranet, in una prospettiva a medio termine, anche sul Web. Ulteriori componenti integrati dove il modulo è coinvolto, per esempio una wiki semantica, rispondono a bisogni comunicativi individuali o di piccoli gruppi e alla raccolta di know-how locale.

Obiettivi (Max 1200 caratteri)

L'obiettivo di questo modulo, coerentemente a quello della commessa, è di sviluppare una sofisticata piattaforma di gestione semantica dell'informazione contenuta all'interno di una intranet. Gli obiettivi a breve-medio termine saranno scelti fra i seguenti: backlog, generalizzazione di query utente, riformulazione di query (query expansion) e ricerca basata su tecniche di elaborazione del linguaggio naturale (NLP), reingegnerizzazione di documenti semi-strutturati, supporto alla decisione su tipi documentali specifici (ex. richieste di autorizzazione), sviluppo di basi di conoscenza sensibile per l'organizzazione (ex. anagrafica, temi, progetti, regolamenti), modellazione di contratti e flussi di lavoro (workflow), acquisizione di know-how mediante una wiki semantica dell'organizzazione.

[Indietro](#)



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
16:23:36

Modulo: (ICT.P04.019.002) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Istituto esecutore: Istituto di linguistica computazionale (ILC)

Commessa: [ICT.P04.019](#)

Dati Generali

[Anagrafica](#)

Previsione

Descrizione

Descrizione Analitica del Modulo

Tematiche di ricerca

Gestione avanzata della conoscenza condivisa di organizzazioni aziendali, basata sull'analisi e l'accesso selettivo del contenuto testuale di basi documentali di dominio e sulla modellizzazione ontologica di strutture organizzative e processi produttivi tramite la definizione di classi, relazioni, e strutture informative rilevanti. Tecniche e metodi di annotazione semantica di documenti per la rappresentazione esplicita del loro contenuto testuale. Classificatori automatici in domini di classi strutturate. Reperimento intelligente di informazioni a partire da basi di dati testuali annotate semanticamente. Costruzione automatica di risorse lessicali di supporto (repertori terminologici strutturati e lessici specialistici di dominio) e loro personalizzazione in relazione a specifici domini di applicazione. Costruzione automatica di risorse concettuali (ontologie, mappe concettuali ecc.) per la navigazione documentale. Interfacce avanzate per l'interrogazione in linguaggio naturale di basi documentali strutturate, attraverso l'integrazione dinamica di tecnologie di analisi automatica del linguaggio, tecniche di apprendimento automatico e strutture ontologiche dedicate.

Stato dell'arte

Il modulo integra modelli stato dell'arte in information retrieval, text mining e ontology engineering, tramite un approccio misto knowledge-based e machine learning, concettuale e automatico-statistico, per l'accesso all'informazione. La base di conoscenza del sistema è un modello concettuale riusabile del dominio di ricerca, di cui alcuni campioni vengono proiettati manualmente su documenti del database. Una combinazione di tecniche di machine learning e tecnologie di analisi automatica del linguaggio è utilizzata per indicizzare l'informazione semantica contenuta nel database documentale su larga scala. L'approccio combina la precisione tipica del lavoro di esperti di dominio e ontologi con la flessibilità e la copertura delle tecniche automatiche e semi automatiche di annotazione e classificazione di testi. L'indicizzazione su larga scala permette la costruzione di interrogazioni complesse sul dominio di ricerca. Il paradigma inoltre integra la valutazione quantitativa dell'efficacia dei metodi sviluppati come elemento fondamentale di ricerca e sviluppo.

Competenze, tecnologie e tecniche di indagine

L'Istituto di Linguistica Computazionale del CNR, leader in Italia, e con una consolidata visibilità internazionale, nella linguistica computazionale, ha vaste e approfondite conoscenze nell'ambito del trattamento automatico del linguaggio, delle tecnologie di apprendimento automatico del linguaggio, dell'annotazione automatica robusta di vasti repertori testuali e nel disegno e creazione di risorse lessico-concettuali. In particolare l'ILC metterà a disposizione del progetto la sua passata esperienza nello sviluppo di sistemi "ibridi" per l'analisi e l'accesso intelligente di basi documentali annotate, basati sull'integrazione dinamica di compilatori di automi a stati finiti, tecniche stocastiche e utilizzo di indici entropici per l'apprendimento automatico e strutture ontologiche dedicate.

Collaborazioni (partner e committenti)

Partner: ISTC CNR Roma. Altri partner mediante convenzioni: SSLMIT di Forlì (Università di Bologna). Altre collaborazioni italiane: ITTIG-CNR Firenze, Università di Pisa, Università di Pavia, Istituto Trentino di Cultura (ITC-irst), Università della Calabria, Università di Salerno, Università di Trento, Università degli Studi di Bari. Collaborazioni internazionali con istituzioni: Fraunhofer Gesellschaft zur Förderung der angewandten Forschung, University of Sheffield, Université Toulouse II, Ecole Nationale Supérieure des Télécommunications. Coinvolgimento in progetti nazionali e internazionali rilevanti: Vikef (EU-FP6), BootStrep (EU-FP6), Pekita (MIUR), FuLL (Ministero Attività Produttive PIA), Pubblicamente (Formez). Collaborazioni con aziende: Xerox Research Centre Grenoble, Siemens Italdata, BC srl, Meta srl.

Potenziale impiego

- per processi produttivi

Le tecnologie sviluppate dalla commessa hanno una immediata applicabilità alle necessità di gestione informativa del CNR e di altri enti/organizzazioni con simili risorse intranet. La prospettiva è l'integrazione e la fruibilità sull'(intra)web di servizi esistenti e nuovi, mediante la loro ricerca, confronto e composizione automatica o semi-automatica. La creazione di nuovi servizi si concentra sulle funzioni avanzate del motore di ricerca, la migrazione di conoscenza (sia statica sia procedurale) sparsa e semi-strutturata in basi di conoscenza armonizzate con l'ontologia delle organizzazioni, e la reingegnerizzazione di formati documentali.

- per risposte a bisogni individuali e collettivi

La commessa intende integrare ricerca e sviluppo di tecnologie di accesso all'informazione che hanno una rilevanza sia scientifica sia sociale, individuale e collettiva, tramite lo sviluppo di interfacce e strumenti che facilitino l'accesso a basi di dati (strutturate e non), ampie ma circoscritte come le intranet, in una prospettiva a medio termine, anche sul Web. Ulteriori componenti integrati, per esempio una wiki semantica, rispondono a bisogni comunicativi individuali o di piccoli gruppi e alla raccolta di know-how locale.

Obiettivi (Max 1200 caratteri)

Obiettivo di questa commessa è sviluppare una sofisticata piattaforma di gestione semantica dell'informazione contenuta all'interno di un intranet, e in prospettiva nel web. Componenti di questa architettura sono moduli individuali di immediata applicazione che progressivamente arricchiscono un motore di ricerca e un sistema di gestione del contenuto di nuova generazione. Gli obiettivi a breve-medio termine saranno scelti fra i seguenti: backlog, generalizzazione di query utente, riformulazione di query (query expansion) e ricerca basata su tecniche di elaborazione del linguaggio naturale (NLP), reingegnerizzazione di documenti semi-strutturati, supporto alla decisione su tipi documentali specifici (ex. richieste di autorizzazione), sviluppo di basi di conoscenza sensibile per l'organizzazione (ex. anagrafica, temi, progetti, regolamenti), modellazione di contratti e flussi di lavoro (workflow), acquisizione di know-how mediante una wiki semantica dell'organizzazione.



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
17:07:54**Modulo:** (ICT.P04.019.001) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse**Commessa:** ICT.P04.019

Previsione attività del Modulo

Attività da svolgere
1. Analisi e definizione di alcuni task relativi a tipologie specifiche di informazione strutturata, es. analisi ed estrazione di informazione da curricula, descrizioni di commesse etc. 2. Costruzione/implementazione delle risorse necessarie all'estrazione di informazione: risorse terminologiche, modelli ontologici, annotatori automatici, interfacce di query, etc. 3. Sviluppo del motore di ricerca su testi di base con funzionalità aggiuntive e integrazione con i componenti di estrazione di informazione.
Punti critici e azioni da svolgere
1. Implementazione di un'interfaccia di ricerca e sviluppo; content management delle risorse create mediante le nuove tecnologie riguardanti il motore di ricerca 2. Specificazione dell'agenda di lavoro coordinata fra i partner e implementazione della pipeline di sviluppo 3. Modellazione di organizzazioni e flussi di lavoro 4. Valutazione dei risultati
Risultati attesi nell'anno
1. Realizzazione di alcuni task di estrazione/trasformazione di informazione strutturata comprendenti analisi qualitative/quantitative delle performance ottenute nei vari task. 2. Upgrade del motore di ricerca di base a stato dell'arte con tecniche di machine learning e trattamento automatico del linguaggio. 3. Definizione dell'agenda di ricerca e sviluppo del motore di ricerca semantico per task generici. 4. Definizione di ontologie organizzative per il CNR.
Iniziative per l'acquisizione di ulteriori entrate
Un'iniziativa per l'acquisizione di fondi di supporto alle richieste di personale, macchine, viaggi, etc. necessari allo svolgimento della commessa è in fase di negoziazione con il Ministero dell'Innovazione Tecnologica (MIT) nell'ambito di un accordo con la Confartigianato. Altre iniziative sono previste nell'ambito dell'ultima call del Sesto Programma Quadro della Comunità Europea e degli altri bandi nazionali e internazionali per la ricerca e l'innovazione industriale. Un contesto idoneo è costituito dalle azioni di supporto a progetti esistenti del FP6, in particolare il progetto integrato NeOn (http://www.neon-project.org) di cui l'ISTC è partner.

[Indietro](#)



Previsione 2006-2008

SCIENZE E TECNOLOGIE DELLA COGNIZIONE

Stato: **In fase di compilazione**Data corrente: 06-12-2005
16:23:52

Modulo: (ICT.P04.019.002) IntraWeb semantico: gestione avanzata dell'informazione in organizzazioni complesse

Istituto esecutore: Istituto di linguistica computazionale (ILC)

Commessa: [ICT.P04.019](#)

Dati Generali

Previsione

Previsione attività [Ric. pers t. ind.](#) [Ric. pers. t. det](#) [Ric. pers. non di ruolo](#) [Entrate/Spese](#) **Personale CNR**

Previsione attività del Modulo

Attività da svolgere

1. Analisi e definizione di alcuni task relativi a tipologie specifiche di informazione strutturata, es. analisi ed estrazione di informazione da curricula, descrizioni di commesse ecc. 2. Costruzione/implementazione delle risorse necessarie all'estrazione di informazione: risorse terminologiche, modelli ontologici, annotatori automatici, interfacce di query, ecc. 3. Sviluppo del motore di ricerca su testi di base con funzionalità aggiuntive e integrazione con i componenti di estrazione di informazione. 4. Integrazione dell'uso di ItalWordNet con strutture informative specifiche del dominio.

Punti critici e azioni da svolgere

1. Implementazione di un'interfaccia di ricerca e sviluppo; content management delle risorse create mediante le nuove tecnologie riguardanti il motore di ricerca 2. Specificazione dell'agenda di lavoro coordinata fra i partner e implementazione della pipeline di sviluppo 3. Valutazione dei risultati

Risultati attesi nell'anno

1. Realizzazione di alcuni task di estrazione/trasformazione di informazione strutturata comprendenti analisi qualitative/quantitative delle performance ottenute nei vari task. 2. Upgrade del motore di ricerca di base a stato dell'arte con tecniche di machine learning e trattamento automatico del linguaggio. 3. Definizione dell'agenda di ricerca e sviluppo del motore di ricerca semantico per task generici.

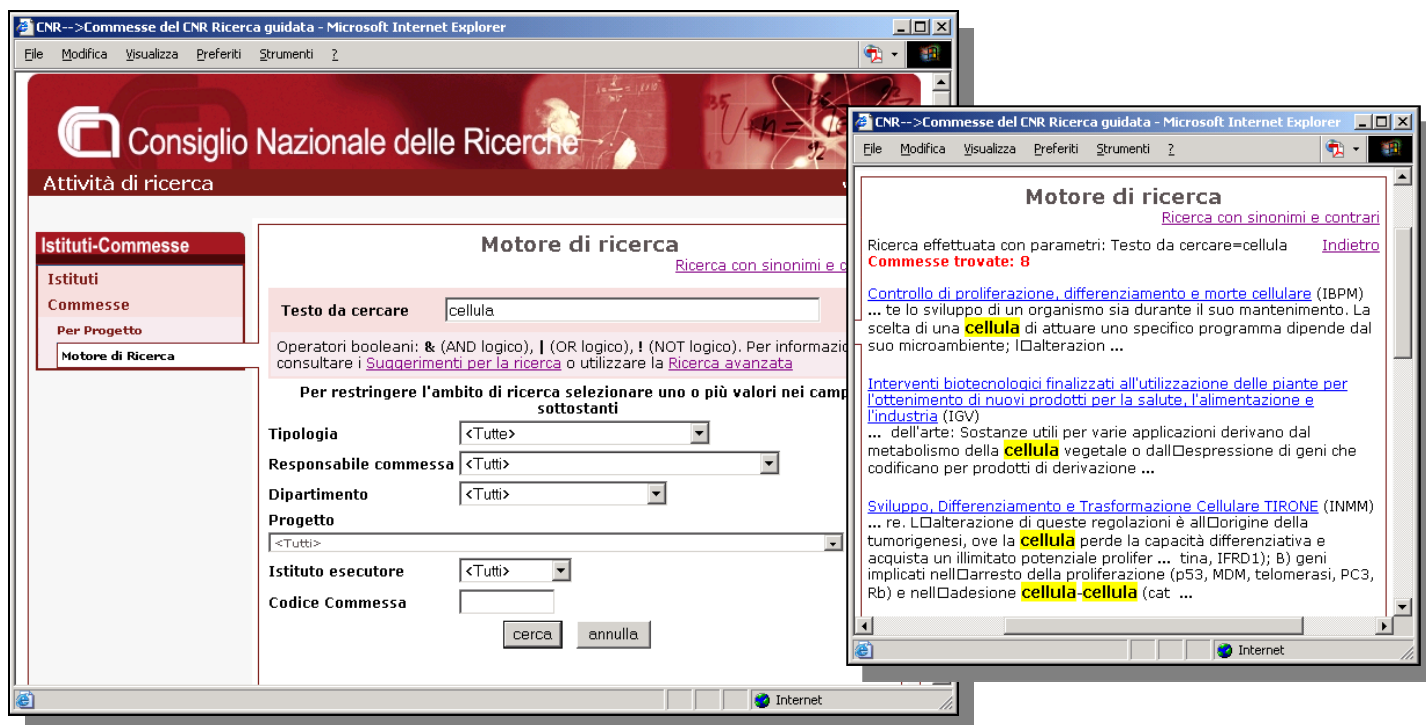
Iniziative per l'acquisizione di ulteriori entrate

Un'iniziativa per l'acquisizione di fondi di supporto alle richieste di personale, macchine, viaggi, etc. necessari allo svolgimento della commessa è in fase di negoziazione con il Ministero dell'Innovazione Tecnologica (MIT) nell'ambito di un accordo con la Confartigianato. Altre iniziative sono previste nell'ambito dell'ultima call del Sesto Programma Quadro della Comunità Europea e degli altri bandi nazionali e internazionali per la ricerca e l'innovazione industriale.



Motore di ricerca

Principali caratteristiche, tecnologie e prodotti utilizzati



INDICE :

FUNZIONALITÀ DEL MOTORE DI RICERCA	2
<i>Ricerche di parole</i>	2
Carattere jolly	2
Esclusione di parole da una ricerca	2
Ricerche booleane	2
Lettere maiuscole/minuscole	2
Sostituzione carattere singolo	3
Distanza massima tra parole	3
Supporto <i>Stopword</i>	3
Gestione del set di caratteri	3
Limitazione del numero di risultati	3
<i>Ricerca di frasi</i>	3
<i>Rappresentazione del risultato</i>	3
Evidenziazione dei termini trovati (Highlight)	3
Visualizzazione del contesto di ricerca	3
RICERCA CON SINONIMI E CONTRARI	4
RICERCA CON THESAURUS	5
<i>Funzionalità del thesaurus</i>	8
<i>Elenco delle relazioni del thesaurus</i>	8
TECNOLOGIE E PRODOTTI UTILIZZATI	9

Funzionalità del motore di ricerca

Ricerche di parole

La ricerca base è per parola e di tipo "AND" automatico, nel senso che vengono cercate le pagine che contengono ogni termine specificato. Per ottimizzare o restringere le ricerche è sufficiente aggiungere altre parole a quelle già inserite, il motore restituirà un sottoinsieme delle pagine visualizzate dopo la prima ricerca.

Carattere jolly

Per una ricerca più generica è possibile utilizzare il carattere jolly, **simbolo ***, in sostituzione di uno o più caratteri. Per esempio indicando **geo*** si cercheranno i termini **geologi**, **geologia**, **geosfera**, **georgia**; oppure indicando ***gia** si cercheranno i termini **biologia**, **geologia**, **regia**. Se si effettua una ricerca per frase, e non per parola, il carattere jolly perde la sua funzionalità e viene trattato come un qualsiasi altro carattere.

Esclusione di parole da una ricerca

Per escludere una parola dalla ricerca, è sufficiente inserire il simbolo di negazione ("!") davanti al termine da escludere. Per ulteriori informazioni sull'operatore NOT si veda quanto riportato nella seguente sezione "Ricerche booleane".

Ricerche booleane

Il motore permette di effettuare ricerche booleane utilizzando gli operatori di seguito indicati:

Simbolo operatore	Funzione	Esempio
&	Trova le pagine contenenti tutti i termini di ricerca (operatore AND)	Se l'utente inserisce acustica&materiali&fenomeni il motore cercherà le pagine che contengono la parola acustica e la parola materiali e la parola fenomeni . L'operatore & può essere omesso: ricercare acustica materiali equivale a ricercare acustica&materiali .
 	Trova le pagine contenenti uno dei termini (operatore OR)	Se l'utente inserisce acustica materiali fenomeni il motore cercherà le pagine che contengono la parola acustica o la parola materiali o la parola fenomeni .
!	Esclude le pagine che contengono un termine (operatore NOT)	Se l'utente inserisce telefonia &!wireless il motore cercherà le pagine che contengono la parola telefonia e che non contengono la parola wireless . L'uso dell'operatore ! è consentito solo congiuntamente all'operatore & .
()	Le parentesi permettono di costruire espressioni complesse composte dagli operatori booleani , &, !.	(dispositivo&mobile) (dispositivo&wireless)

Lettere maiuscole/minuscole

Il motore di ricerca non fa distinzione tra lettere minuscole e maiuscole. Ad esempio, digitando "attivo", "ATTIVO" e "AttIVO" si ottengono sempre gli stessi risultati.

Sostituzione carattere singolo

Abilitando la sostituzione carattere singolo il motore cercherà i termini indicati ed i termini che differiscono per un solo carattere da quelli indicati. Ad esempio, se l'utente inserisce **cassa** ed abilita la sostituzione carattere singolo, il motore cercherà le pagine che contengono la parola **cassa**, ma anche quelle che contengono le parole **casse**, **casta**, **tassa**, ecc..

Distanza massima tra parole

Abilitando la distanza massima tra parole il motore cercherà le pagine in cui la distanza massima tra tutti i termini indicati non supera il valore specificato.

Se l'utente, ad esempio, inserisce "analisi materiali" ed abilita a 2 la distanza massima tra le parole, il motore cercherà le pagine che contengono le parole "analisi" e "materiali" e restituirà solo le pagine in cui tali termini sono ad una distanza tra loro non superiore a 2 (una eventuale pagina contenente "analisi quantitativa senza materiali" sarà inclusa tra i risultati della ricerca perché i termini cercati sono a distanza due, mentre una pagina contenente "analisi quantitativa senza utilizzare materiali" non sarà inclusa tra i risultati della ricerca perché i termini cercati sono a distanza tre).

Supporto *Stopword*

Il sistema offre la possibilità di ignorare parole (sequenze di caratteri) troppo frequenti cioè le cosiddette *stopword*: gli articoli, le congiunzioni, ma anche i segni di interpunzione e i caratteri non alfabetici come "\$", "/", ecc..

Gestione del set di caratteri

Il sistema consente di indicizzare i documenti facendo riferimento a diversi insiemi di caratteri. All'interno di un insieme di caratteri possono essere definite delle equivalenze tra caratteri (ad esempio i tre caratteri **A**, **a**, **à** possono essere considerati equivalenti). Il sistema gestisce tre set di caratteri (ASCII, ISO, OVERLAP_ISO) ma, oltre a quelli già presenti, permette anche di definirne di nuovi.

Limitazione del numero di risultati

Il sistema permette di definire il massimo numero di risultati prodotti da una ricerca agendo su due parametri di tuning (**MAX_MATCHES** e **WORD_SCORE**).

Ricerca di frasi

Per ricercare una frase è sufficiente racchiuderla tra virgolette. Ad esempio se si inseriscono le parole "*tecnica di sintesi*" il motore cercherà la frase "*tecnica di sintesi*" esattamente così come è scritta.

Rappresentazione del risultato

Evidenziazione dei termini trovati (*Highlight*)

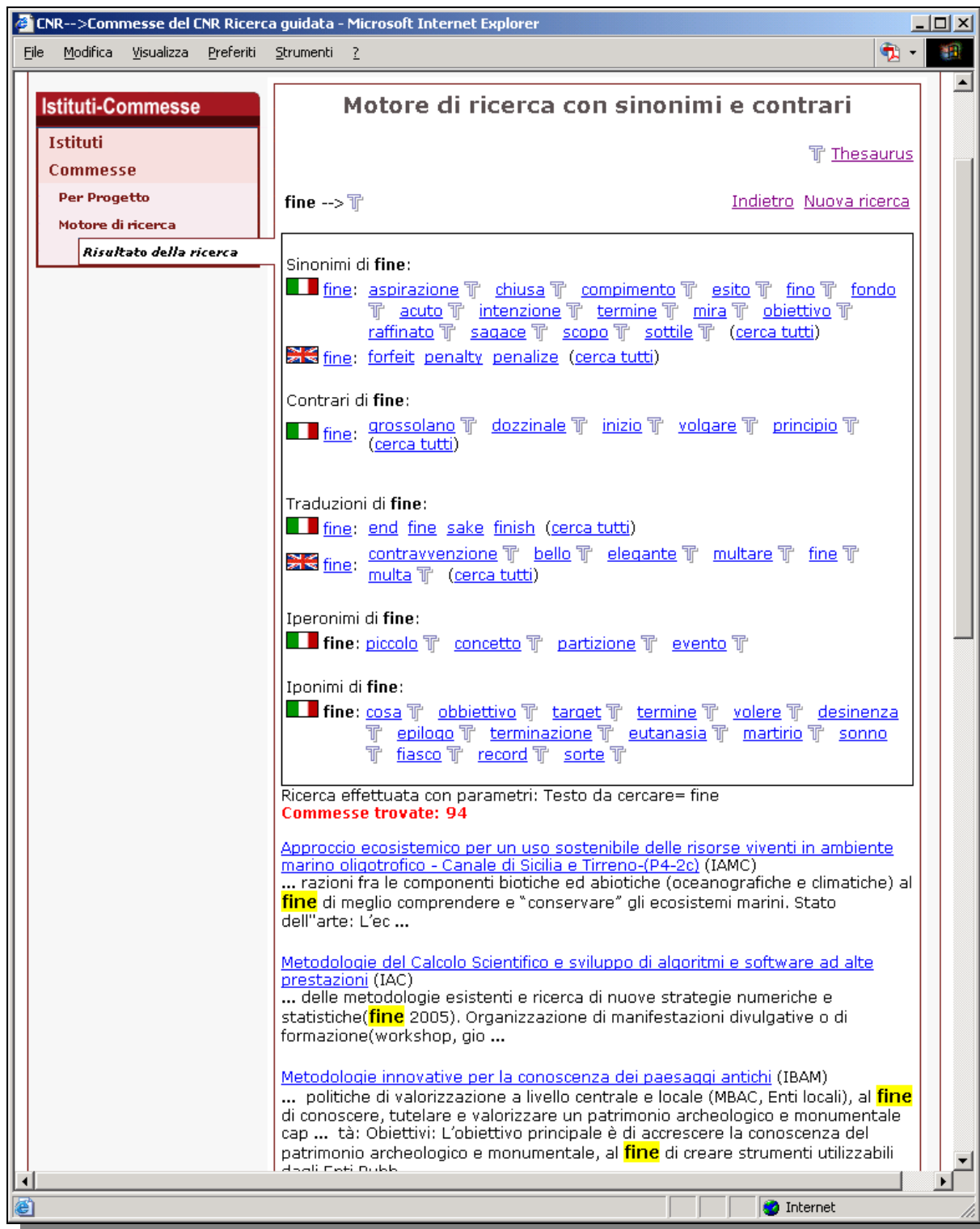
La funzione *Highlight* evidenzia, con colore giallo, tutti i termini trovati all'interno dei risultati della ricerca. La funzione è attivata automaticamente ad ogni ricerca effettuata.

Visualizzazione del contesto di ricerca

Tutti i risultati trovati contengono una o più sintesi che mostrano il contesto in cui vengono utilizzati i termini ricercati.

Ricerca con sinonimi e contrari

Il motore di ricerca per sinonimi contrari offre delle funzionalità avanzate di supporto alla consultazione delle informazioni. Effettuando la ricerca di un termine il sistema restituisce, oltre ai risultati, una serie di suggerimenti relazionati con il termine (o i termini se la query è complessa): sinonimi, contrari, generalizzazioni (iperonimie), specializzazioni (iponimie), nonché la traduzione dall'italiano all'inglese e dall'inglese all'italiano.



The screenshot shows a web browser window titled "CNR-->Commesse del CNR Ricerca guidata - Microsoft Internet Explorer". The browser's address bar shows the URL "http://www.cnr.it/...". The page content is divided into a left sidebar and a main content area.

Left Sidebar:

- Istituti-Commesse** (highlighted)
- Istituti
- Commesse
- Per Progetto
- Motore di ricerca
- Risultato della ricerca

Main Content Area:

Motore di ricerca con sinonimi e contrari

[Thesaurus](#)

fine --> [Indietro](#) [Nuova ricerca](#)

Sinonimi di fine:

- [fine](#): [aspirazione](#) [chiusa](#) [compimento](#) [esito](#) [fino](#) [fondo](#) [acuto](#) [intenzione](#) [termine](#) [mira](#) [obiettivo](#) [raffinato](#) [sagace](#) [scopo](#) [sottile](#) (cerca tutti)
- [fine](#): [forfeit](#) [penalty](#) [penalize](#) (cerca tutti)

Contrari di fine:

- [fine](#): [grossolano](#) [dozzinale](#) [inizio](#) [volgare](#) [principio](#) (cerca tutti)

Traduzioni di fine:

- [fine](#): [end](#) [fine](#) [sake](#) [finish](#) (cerca tutti)
- [fine](#): [contravvenzione](#) [bello](#) [elegante](#) [multare](#) [fine](#) [multa](#) (cerca tutti)

Iperonimi di fine:

- [fine](#): [piccolo](#) [concetto](#) [partizione](#) [evento](#)

Iponimi di fine:

- [fine](#): [cosa](#) [obbiettivo](#) [target](#) [termine](#) [volere](#) [desinenza](#) [epilogo](#) [terminazione](#) [eutanasia](#) [martirio](#) [sonno](#) [fiasco](#) [record](#) [sorte](#)

Ricerca effettuata con parametri: Testo da cercare= fine
Commesse trovate: 94

[Approccio ecosistemico per un uso sostenibile delle risorse viventi in ambiente marino oligotrofico - Canale di Sicilia e Tirreno-\(P4-2c\)](#) (IAMC)
... ragioni fra le componenti biotiche ed abiotiche (oceanografiche e climatiche) al **fine** di meglio comprendere e "conservare" gli ecosistemi marini. Stato dell'arte: L'ec ...

[Metodologie del Calcolo Scientifico e sviluppo di algoritmi e software ad alte prestazioni](#) (IAC)
... delle metodologie esistenti e ricerca di nuove strategie numeriche e statistiche(**fine** 2005). Organizzazione di manifestazioni divulgative o di formazione(workshop, gio ...

[Metodologie innovative per la conoscenza dei paesaggi antichi](#) (IBAM)
... politiche di valorizzazione a livello centrale e locale (MBAC, Enti locali), al **fine** di conoscere, tutelare e valorizzare un patrimonio archeologico e monumentale cap ... tà: Obiettivi: L'obiettivo principale è di accrescere la conoscenza del patrimonio archeologico e monumentale, al **fine** di creare strumenti utilizzabili dagli Enti Pubb...

Figura 1 - Esempio di ricerca con sinonimi, contrari, iperonimi e iponimi

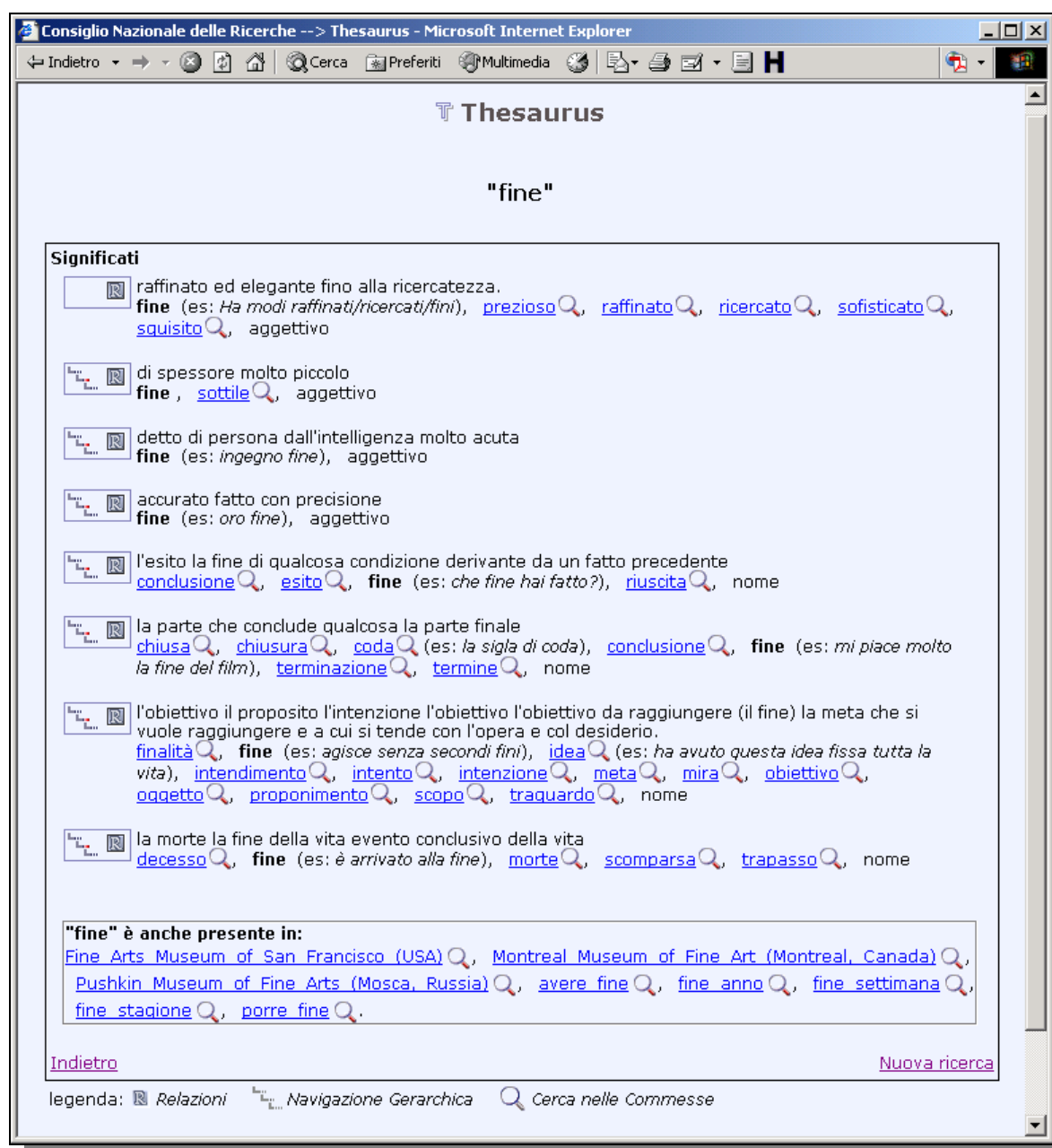
Il sistema, attualmente, utilizza cinque dizionari (il cui numero può essere agevolmente aumentato):

- ❑ dizionario italiano dei sinonimi
- ❑ dizionario italiano dei contrari
- ❑ dizionario inglese dei sinonimi
- ❑ dizionario dall'italiano all'inglese
- ❑ dizionario dall'inglese all'italiano

Ricerca con thesaurus

Il thesaurus è stato sviluppato partendo dai dati dell'ItalWordNet messi a disposizione dall'Istituto di Linguistica Computazionale del CNR e comprende allo stato attuale 50366 set di significati per 68628 lemmi e 133095 relazioni.









La consultazione del thesaurus può essere effettuata navigando per significati, per relazioni e per gerarchie di significati (limitatamente alle relazioni iperonimo/iponimo e classe/istanza).



Thesaurus

"fine"

Significati

-  raffinato ed elegante fino alla ricercatezza.
fine (es: *Ha modi raffinati/ricercati/fini*), [prezioso](#), [raffinato](#), [ricercato](#), [s sofisticato](#), [squisito](#), aggettivo
-  di spessore molto piccolo
fine, [sottile](#), aggettivo
-  detto di persona dall'intelligenza molto acuta
fine (es: *ingegno fine*), aggettivo
-  accurato fatto con precisione
fine (es: *oro fine*), aggettivo
-  l'esito la fine di qualcosa condizione derivante da un fatto precedente
[conclusione](#), [esito](#), **fine** (es: *che fine hai fatto?*), [riuscita](#), nome
-  la parte che conclude qualcosa la parte finale
[chiusa](#), [chiusura](#), [coda](#) (es: *la sigla di coda*), [conclusione](#), **fine** (es: *mi piace molto la fine del film*), [terminazione](#), [termine](#), nome
-  l'obiettivo il proposito l'intenzione l'obiettivo da raggiungere (il fine) la meta che si vuole raggiungere e a cui si tende con l'opera e col desiderio.
[finalità](#), **fine** (es: *agisce senza secondi fini*), [idea](#) (es: *ha avuto questa idea fissa tutta la vita*), [intendimento](#), [intento](#), [intenzione](#), [meta](#), [mira](#), [obiettivo](#), [oggetto](#), [proponimento](#), [scopo](#), [traguardo](#), nome
-  la morte la fine della vita evento conclusivo della vita
[decesso](#), **fine** (es: *è arrivato alla fine*), [morte](#), [scomparsa](#), [trapasso](#), nome

"fine" è anche presente in:
[Fine Arts Museum of San Francisco \(USA\)](#), [Montreal Museum of Fine Art \(Montreal, Canada\)](#),
[Pushkin Museum of Fine Arts \(Mosca, Russia\)](#), [avere fine](#), [fine anno](#), [fine settimana](#),
[fine stagione](#), [porre fine](#).

[Indietro](#) [Nuova ricerca](#)




legenda:  Relazioni  Navigazione Gerarchica  Cerca nelle Commesse

Figura 2 – Consultazione del thesaurus – Significati della parola “fine”

T Thesaurus

la parte che conclude qualcosa, la parte finale

[chiusa](#), [chiusura](#), [coda](#), [conclusione](#), [fine](#), [terminazione](#), [termine](#), [nome](#).

[Indietro](#)[Nuova ricerca](#)

Relazioni

antonym



nascita, origine, il periodo iniziale (fig), primo periodo, momento iniziale, periodo iniziale

[natale](#), [alba](#), [albore](#), [albori](#), [aurora](#), [barlume](#), [esordio](#), [fonte](#), [inizio](#), (es: *all'inizio c'erano dubbi*) [nascita](#), [origine](#), [prima fase](#), [primo periodo](#), [primordio](#), [principio](#), [nome](#).



principio, inizio di qualcosa, parte iniziale di qualcosa, la prima parte di qualcosa, l'inizio, la parte iniziale

[germoglio](#), [attacco](#), [inizio](#), (es: *l'inizio del capitolo non è chiaro*) [principio](#), [nome](#).

has_hyperonym



Navigazione Gerarchica salendo verso:



ciascuna delle porzioni o degli elementi in cui è diviso un tutto

[partizione](#), [parte](#), [nome](#).

has_hyponym



descrizione non disponibile

[desinenza](#), [nome](#).



terminazione di un vocabolo

[terminazione](#), [uscita](#), [nome](#).



descrizione non disponibile

[epilogo](#), [nome](#).

has_pertained



relativo all'ultima parte di qualcosa

[terminale](#), [conclusivo](#), [finale](#), [aggettivo](#).

near_synonym



cosa che sta al termine di qualcosa

[appendice](#), [nome](#).

Figura 3 – Relazioni del termine “fine” nel significato “la parte che conclude qualcosa, la parte finale”

Thesaurus






edificio o parte di esso in cui si abita
[abitazione](#), [casa](#), [dimora](#), [magione](#), [ostello](#), [tetto](#), [nome](#).

[Nuova ricerca](#)

[Indietro](#)

Navigazione gerarchica

fino a:

-  **ciò che esiste**
[ente](#), [entità](#), [essere](#) (es: ogni essere vivente ha diritto a rispetto e considerazione), [nome](#).
-  **oggetto materiale concreto**
[cosa](#), [oggetto](#), [nome](#).
-  **cosa costruita**
[costruzione](#), [struttura](#), [nome](#).
-  **costruzione architettonica.**
[edificazione](#), [edificio](#), [fabbricato](#), [nome](#).
-  **edificio o parte di esso in cui si abita**
[abitazione](#), [casa](#), [dimora](#), [magione](#), [ostello](#), [tetto](#) (es: Non possono più stare sotto lo stesso tetto, non c'è più niente fra loro.), [nome](#).

al livello sottostante:












-  **piccola casa fatta di sassi o di legno e usata come ricovero in alta montagna.**
[[baita](#), [nome](#)]
-  **piccola casa, per lo più in legno, per il ricovero provvisorio di uomini, animali o merci.**
[[baracca](#), [bicocca](#), [nome](#)]
-  **casa signorile rustica adibita a luogo di raccolta per i cacciatori**
[[casino](#), [nome](#)]
-  **piccola casa fatta di rami e frasche**
[[capanna](#), [tucul](#), [nome](#)]
-  **casa molto piccola e povera.**
[[capanna](#), [casipola](#), [casupola](#), [nome](#)]
-  **casa di campagna rustica e isolata**
[[casale](#), [casolare](#), [nome](#)]
-  **tipica abitazione rurale della Russia, fatta di legno e costituita da un'unica grande stanza con stufa.**
[[isba](#), [izba](#), [nome](#)]
-  **abitazione misera, squallida, sporca.**
[[abituro](#), [buco](#), [canile](#), [catapecchia](#), [spelonca](#), [stamberga](#), [tana](#), [topaia](#), [tugurio](#), [nome](#)]
-  **casa grande e lussuosa.**
[[reggia](#), [nome](#)]

Figura 4 – Navigazione gerarchica per il termine “casa” nel significato “edificio o parte di esso in cui si abita”

Funzionalità del thesaurus

Il thesaurus attualmente supporta le seguenti funzionalità:

- ❑ Integrazione completa con il motore di ricerca: l'icona  , a fianco di ogni termine del thesaurus, consente di effettuare la ricerca di quel termine all'interno delle commesse del CNR; viceversa, l'icona  permette di consultare il thesaurus direttamente sul termine corrispondente.
- ❑ Ricerca nei significati e nei lemmi: se il termine è presente nel thesaurus vengono restituiti tutti i significati del termine e in aggiunta tutti i lemmi all'interno dei quali esso è presente (es: cercando "casa" otterremo anche: "Vedi anche: casa di cura, casa del popolo, etc..."); se il termine non è presente nel thesaurus il sistema suggerisce gli eventuali lemmi all'interno dei quali è presente.
- ❑ Gestione di stringhe complesse: il sistema è in grado di elaborare la stringa di ricerca escludendo i caratteri non alfanumerici (es. operatori booleani, spazi o parentesi) separando la digitazione in singole parole e cercando ciascuna di esse.
- ❑ Gestione dei prefissi: se l'utente inserisce 'bio' comparirà tra i risultati in prima battuta anche 'bio- '.
- ❑ Navigazione gerarchica e per relazioni tra termini.
- ❑ Forme verbali: a breve sarà implementata anche la gestione delle forme verbali (se il termine è un verbo flesso il sistema restituisce la forma verbale all'infinito).

Elenco delle relazioni del thesaurus

Relation	Examples	Relation	Examples
Synonymy	bicycle/bike to analyse/to examine	Involved / Role	to hammer/hammer pedestrian/to walk
Near_Synonym	implement/utensil cerebration/opinion	Involved_Agent / Role_Agent	to teach/teacher runner/to run
Xpos_Near_Synonym	arrival/to arrive	Involved_Patient / Role_Patient	to teach/student student/to teach
Has_Hyperonym / Has_Hyponym	dog/animal to move/to travel	Involved_Instrument / Role_Instrument	to paint/paint-brush gun/to shoot
Has_Xpos_Hyperonym / Has_Xpos_Hyponym	arrival/to go to hit/knock	Involved_Location / Role_Location	to swim/water school/to teach
Antonym	to arrive/to leave	Involved_Direction / Role_Direction	to lead/place arrival/to arrive
Compl_Antonym	alive/dead	Involved_Source_Direction / Role_Source_Direction	to disembark/ship outside/to enter
Grad_Antonym	cold/hot	Involved_Target_Direction / Role_Target_Direction	to exit/outside inside/to enter
Xpos_Antonym	arrival/departure	Involved_Result / Role_Result	to freeze/ice ice/to ice
Has_Holonym / Has_Meronym	arm/body hand/finger	Co_Role	piano player/piano



Has_Mero_Part / Has_Holo_Part	foot/toe hip/body	Co_Agent_Patient / Co_Patient_Agent	teacher/student student/teacher
Has_Mero_Member / Has_Holo_Member	team/player student/school	Co_Agent_Instrument / Co_Instrument_Agent	guitar player/guitar guitar/guitar player
Has_Mero_Madeof / Has_Holo_Madeof	jam/fruit	Co_Agent_Result / Co_Result_Agent	painter/painting painting/painter
Has_Mero_Portion / Has_Holo_Portion	bread/slice slice/cake	Co_Patient_Instrument / Co_Instrument_Patient	wood/axe axe/wood
Has_Mero_Location / Has_Holo_Location	city/city-centre oasis/desert	Co_Patient_Result / Co_Result_Patient	skin/scarification scarification/skin
Causes / Is_Caused_By	to kill/to die execute/sentence	Co_Instrument_Result / Co_Result_Instrument	camera/photo photo/camera
Results_In / Is_Result_Of	to kill/to die sick/to fall ill	Be_In_State / State_Of	poor/poorness oldness/old
For_Purpose_Of / Is_Purpose_Of	to search/to find to win/to compete	In_Manner / Is_Manner_For	to whisper/ in a low voice
Is_Means_For / Has_Means	heat/distillation to evaporate/boiling	Derivation	water/water-carrier
Has_Subevent / Is_Subevent_Of	to buy/to pay to snore/to sleep	Liable_To / Has_Liability	judgeable/to judge
Is_A_Value_Of / Has_Value	tall/stature	Fuzzynym	collaborationist/enemy
Pertains_To / Has_Pertained	presidential/president Italian/Italy	Xpos_Fuzzynym	to govern/ government-in-exile
Has_Instance / Belongs_To_Class	river/Po Rome/city		

Tecnologie e prodotti utilizzati

Sistema operativo	Linux/Unix
DBMS	Informix
Web Server / Application server	iPlanet/Netscape/Apache IBM-Informix Web Data Blade
Motore di ricerca	Excalibur Text Data Blade

Evoluzione del Motore di Ricerca CNR: Architettura, Piano di Ricerca e Sviluppo

Massimiliano Ciaramita Aldo Gangemi Domenico Pisanelli
Laboratorio di Ontologia Applicata (LOA) ISTC-CNR*

27 dicembre 2005

Indice

1	Introduzione	2
1.1	Organizzazione del documento	2
2	Architettura generale	2
2.1	Motore di ricerca base	2
2.2	Ottimizzazione della ricerca di documenti	2
2.3	Sviluppo parallelo su Immagini del database	3
3	Potenziamento del motore di ricerca	4
3.1	Information retrieval avanzato	4
3.2	Il modello di re-ranking	5
3.3	Dynamic linking e interfaccia leggera	6
3.4	Componenti	7
4	Evoluzione del motore di ricerca	7
4.1	Costruzione dell'ontologia e Information Extraction	7
4.2	Analisi semantica di query e documenti	8
4.3	Componenti	9
5	Estensioni ulteriori del motore di ricerca	9
6	Gruppi e competenze	9

*Questo documento rielabora e sviluppa anche le informazioni e le proposte fornite dal Servizio Reti e Telecomunicazioni del CNR, dall'Istituto di Linguistica Computazionale del CNR, e dal Gruppo di lavoro SSLMIT dell'Università di Bologna (Forlì).

1 Introduzione

Questo documento presenta un modello di ricerca e sviluppo di un motore di ricerca su base di dati testuale non strutturata. Il caso di studio e applicazione primaria riguarda la rete intranet del CNR. Il modello si basa in primo luogo su un motore di ricerca pre-esistente standard. L'evoluzione del motore di ricerca riguarda principalmente i seguenti aspetti: il potenziamento delle funzionalità di ricerca di documenti e di interazione con l'utente, l'evoluzione a motore di ricerca semantico, dove l'unità informativa è costituita da "pezzi" di informazione specifici e strutturati, eventualmente contenuti in uno o più documenti, anche espressi in modo implicito. Il modello prevede un'evoluzione basata su tecnologie di ontology engineering, tecniche lessicografiche, trattamento automatico del linguaggio e machine learning.

1.1 Organizzazione del documento

Nella Sezione 2 è descritta l'architettura generale del motore di ricerca, nella Sezione 3 gli elementi centrali della prima fase di ricerca e sviluppo che riguarda il potenziamento del motore di base, mentre la Sezione 4 descrive gli aspetti rilevanti della seconda fase, di evoluzione del paradigma di ricerca. Il motore di ricerca così inteso costituisce il cuore di un sistema informativo studiato per supportare applicazioni sofisticate di information management, illustrate nella Sezione 5. Infine, la sezione 6 illustra uno schema delle competenze relative ai diversi aspetti.

2 Architettura generale

2.1 Motore di ricerca base

Il modello presuppone l'esistenza di un motore di ricerca su testi con caratteristiche standard. Nel nostro caso il motore di ricerca è Excalibur Text DataBlade (ETX). Il motore garantisce una serie di funzionalità primarie, in particolare:

1. Creazione di indici testuali da un database esistente, (**etx_access()**);
2. Ricerche su parole e stringhe di testo, con operatori booleani, prossimità, liste di stopword, risorse lessicali tipo sinonimi, etc. (**etx_contains()**);
3. Restituzione delle coordinate dei termini trovati: l'indice del documento e la posizione nel testo dei termini trovati (**etx_GetHilite()**);
4. Output dei documenti indicizzati dopo la formattazione di ETX, quindi un'immagine "mirror" dei testi su cui viene operata la ricerca (**etx_ViewHilite()**);

2.2 Ottimizzazione della ricerca di documenti

La prima fase di ricerca e sviluppo riguarda il fine-tuning del motore con tecniche state-of-the-art di recupero di documenti, come: query expansion supportata da risorse lessicali aggiuntive, local clustering dei documenti selezionati, relevance feedback, e profiling. Lo sviluppo degli strumenti di supporto: l'ontologia del motore di ricerca, le basi di conoscenza, le tecniche di learning e trattamento

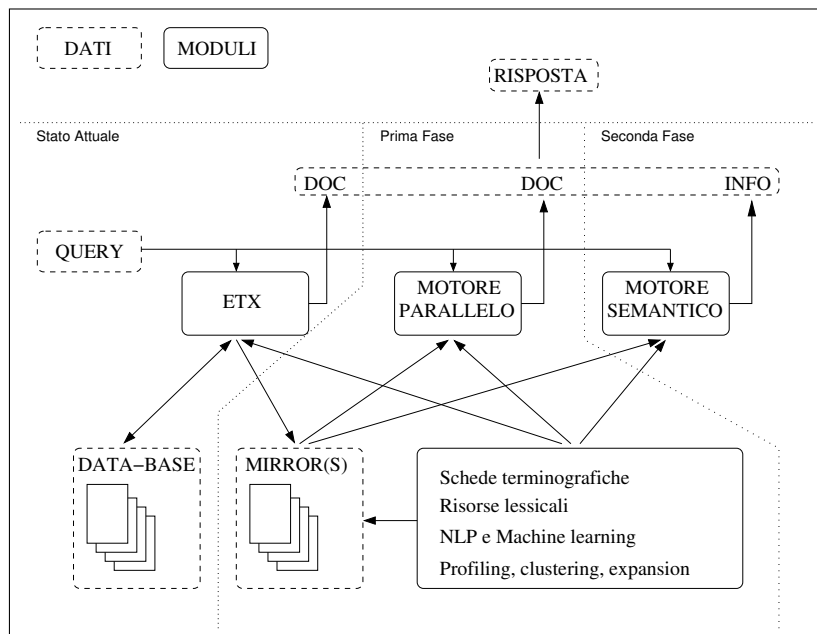


Figura 1: Schema dell'architettura generale del sistema e del suo sviluppo da semplice motore di ricerca su testo a motore di ricerca semantico.

automatico del linguaggio aggiuntive, costituiscono i fondamenti del passaggio al motore semantico in cui vengono ricercate informazioni strutturate e complesse e non documenti. La Figura 1 presenta una visione schematica del motore di ricerca attuale e del suo sviluppo ed evoluzione.

2.3 Sviluppo parallelo su Immagini del database

Lo sviluppo del motore si basa sull'architettura esistente e la estende tramite la creazione di *immagini mirror* del database testuale "copiate" da quella di ETX, cf. (4) sopra. Questa copia dei documenti indicizzati da ETX, ha diversi scopi: costruzione di un motore di ricerca parallelo, supporto al tagging per annotazione di entità, occorrenze, annotate con metadati da usare per la creazione di link ad altri documenti intranet, internet. Nell'immagine, i documenti sono indicizzati nello stesso formato ASCII (o alternativo utilizzato dal modulo ETX). Nell'immagine l'indicativo del documento (per esempio, il nomefile) e l'offset di ogni termine corrispondono ai valori dell'immagine di ETX prodotti da (`etx_ViewHilite()`). In questo modo il lavoro di formattazione, indicizzazione e individuazione dei termini è svolto da ETX, mentre le proprietà associate con le stringhe reperite sono disponibili in tempo costante tramite consultazione dell'immagine. Le immagini sono dunque copie del database soggette a trattamento automatico, per esempio tagging con etichette di classi semantiche, usate per costruire indici contenenti meta-informazione.



Figura 2: Schema di interfaccia leggera.

3 Potenziamento del motore di ricerca

La prima fase riguarda il potenziamento del motore attuale. L'obiettivo è monitorare e migliorare la performance in Document Retrieval e l'interazione utente-motore. Gli interventi riguardano in particolare l'implementazione di concetti affermati in Information Retrieval (IR) come *relevance feedback*, *query expansion* basata su glossari di dominio generale e specifico, *local clustering*, e *profiling*. Inoltre a livello di interfaccia si sviluppano funzionalità aggiuntive come la generazione dinamica di link, e un'interfaccia ergonomica leggera.

3.1 Information retrieval avanzato

Il potenziamento della qualità del motore in Document Retrieval, in particolare la *rilevanza* dei risultati, è centrato su tre concetti base di IR:

- **Relevance feedback:** l'utente immette una query, il motore restituisce una serie di documenti,¹ l'utente seleziona un documento, o altra informazione, oppure riformula un'altra query, oppure ancora abbandona il motore. Queste azioni definiscono un *ciclo* in cui ogni azione condiziona la seguente. In particolare le azioni dell'utente costituiscono la fonte principale di feedback per il motore, da analizzare online per produrre la successiva lista di documenti (attraverso appropriate query expansion o selezione di cluster di documenti appropriati, vedi sotto), oppure off-line come analisi dei baklog.
- **Query expansion:** consiste nella appropriata espansione, operata implicitamente dal motore, delle query formulate dall'utente, principalmente attraverso l'uso di risorse lessicali di dominio specifico, o generale, rilevanti. In prospettiva attraverso il reweighting dei pesi del modello in un'implementazione parallela del motore basata sull'immagine.

¹Insieme, eventualmente, ad altra informazione, come link a dizionari etc.

- **Local clustering:** le selezioni dell'utente, insieme alle query, fanno sì che alcuni insiemi di documenti siano più vicini semanticamente all'informazione cercata dall'utente. Utilizzando le risorse lessicali, tecniche di clustering e di profiling, il motore analizza lo stato corrente della query e opera un re-ranking adattivo dei documenti da restituire, in modo che siano più "vicini" alla query dell'utente. Lo score associato dal motore ad ogni hit produce un ranking iniziale dei documenti, questo ranking viene manipolato utilizzando la knowledge base del motore, cioè le risorse lessicali e la conoscenza relativa al profilo dell'utente e al profilo dei documenti "attivi". L'analisi delle transazioni avvenute viene utilizzata per costruire un modello di re-ranking che viene progressivamente affiancato al motore ETX.

L'implementazione di questi concetti avviene attraverso la classificazione, o clustering, dell'utente corrente in base a proprietà come l'origine (IP), la query formulata, proprietà della query formulata, documenti selezionati etc. La risposta del motore a queste informazioni è la produzione di una query espansa attraverso termini rilevanti estratti da risorse lessicali di dominio. Di conseguenza il motore opera dinamicamente un re-ranking delle hits più probabili o vicine al prototipo di "comunicazione" in corso con l'utente, e la selezione, o clustering locale, di probabili documenti rilevanti. La definizione di un prototipo di sessione, utente, e cluster di documenti sono basati su modelli vettoriali standard. L'analisi dei log supporta la creazione di prototipi vettoriali di utenti, query e cluster di documenti.

3.2 Il modello di re-ranking

Qui definiamo brevemente il paradigma di re-ranking. Supponiamo che, data una query q_i , ETX restituisca una lista di n documenti, a ogni documento $x_{i,j}$ è associato uno score $L(x_{i,j}, q_i)$ che determina il ranking iniziale. Supponiamo inoltre che il documento ricercato dall'utente sia uno degli n documenti, che indichiamo con $x_{i,1}$. Il nostro obiettivo è sviluppare un modello di interazione utente-motore in cui caratteristiche salienti della query e dei documenti e tutte le informazioni ritenute utili al fine del reperimento della corretta informazione sono rappresentate in un modello vettoriale. Questa codificazione avviene attraverso una funzione che estrae "features", caratteristiche salienti, dagli n documenti candidati; per es. una feature potrebbe identificare l'utente in base al suo indirizzo IP, e dunque essere parte di un insieme di caratteristiche di "profiling", nel modo seguente:

$$h_s(x_{i,j}, q_i) = \begin{cases} 1, & \text{se l'indirizzo IP dell'utente comincia con "150.146"} \\ 0, & \text{altrimenti.} \end{cases}$$

In altre parole questa feature, se attiva, identifica l'istituto di provenienza della query. Supponiamo che nel modello siano definite m di queste features, il nostro obiettivo è assegnare un nuovo score agli n documenti uguale a:

$$F(x_{i,j}, \vec{\alpha}, q_i) = \alpha_0 L(x_{i,j}, q_i) + \sum_{s=1}^m \alpha_s h_s(x_{i,j}, q_i) \quad (1)$$

Il problema da risolvere è quello di trovare un vettore di parametri per $\vec{\alpha}$ che produca dei buoni risultati in termini di rilevanza. Il vettore di parametri

Directory	Ricerca avanzata	Navigazione grafica	News
-----------	------------------	---------------------	------

CNR

1 Gestualita', oralita' e lingua scritta nello sviluppo e nella lingua dei segni (ISTC)... -> COMMESSA

2 Osservatorio neologico della lingua italiana (Onli) www.cnr.it/istituti/Focus1 ... -> FOCUS

3 UN CD-ROM DECLAMA LE POESIE DEI SORDI. Il CNR ha realizzato il primo CD ... -> NEWS

4 ISTC - Istituto di Scienze e Tecnologie della Cognizione - CNR www.istc.cnr.it/ -> ISTITUTO

5 ILIESI - Istituto per il Lessico Intellettuale Europeo e Storia ... www.iliesi.cnr.it/ -> ISTITUTO

...

Figura 3: Esempio di ricerca con interfaccia leggera.

può essere “stimato” in diversi modi, per esempio tramite metodi di *boosting*. Il feedback necessario a settare i parametri può essere ottenuto direttamente tramite un periodo di training o, indirettamente, tramite analisi del relevance feedback. La costruzione di questo modello parallelo si basa sull’esistenza delle immagini-copia del database e utilizza le funzionalità base del motore ETX.

3.3 Dynamic linking e interfaccia leggera

Due funzionalità aggiuntive sono considerate primarie a breve termine. Primo, un’interfaccia leggera, tipo-Google, in cui l’utente pu scegliere di fare come in qualsiasi altro motore di ricerca e semplicemente immettere una query iniziale non accompagnata da sintassi ulteriore. L’opzione strutturata, ovviamente, è mantenuta; come opzione per utenti che già abbiamo familiarità con essa e sappiano già come formulare query strutturata e conoscano l’organizzazione del motore CNR. Per gli utenti naive o meglio, abituati a Google (cioè tutti gli altri), dopo la query la risposta del motore è la presentazione dei risultati in modo standard, con l’aggiunta, accanto ad ogni documento, della sua etichetta: commessa, news, focus, istituto, etc. Appositi link al glossario interno spiegano il significato di questi termini. La selezione dell’utente decide di cosa l’utente sia in cerca, e.g., se seleziona un istituto il local clustering si limita agli istituti, etc. Non solamente una questione grafica ma di strategia all’approccio dialogico con l’utente proiettata verso l’interazione in linguaggio naturale.

Le interfacce leggere costituiscono l’approccio dominante in IR e sono volte a schermare l’utente dal processo di (ri)formulazione della query. L’utente arriva sul motore di ricerca con una sua idea di cosa cerca. Utente e motore, insieme, costruiscono la giusta query, Nascondere i dettagli all’utente riduce la ricerca a una serie di passi semplici e intuitivi, locali. Inoltre, determinano una maggiore interazione con l’utente e quindi un maggior numero di opportunità di feedback per il motore, e perciò un maggior numero di dati su come “ragiona” l’utente, riutilizzabili per studiare trend e schemi di queries (implicitamente supporta il profiling). La Figura 2 illustra una pagina tipo-Google per il motore di ricerca, mentre la Figura 3 illustra i possibili risultati di ricerca per una query come

“LIS” (acronimo di Linguaggio Italiano dei Segni) dove sono evidenziati i domini di ogni hit (commessa, istituto, etc.).

La presentazione dei risultati comprende la generazione dinamica di un numero arbitrario di *link* contenuti nel documento selezionato dall’utente. I link riguardano tutte le occorrenze di oggetti per cui esiste un link, quelli correntemente indicizzati dal motore in indici addizionali. In pratica, si definisce un’insieme di oggetti da linkare, per esempio tutte le pubblicazioni, si costruisce un indice dei link attivi nel documento eseguendo un tagging sulla base dell’immagine. Nel documento restituito all’utente sono attivati i link identificati, che possono essere relativi non solo a eventuali pubblicazioni, ma anche a pagine-web di istituti, personale, news, Web, etc.

3.4 Componenti

La prima fase consiste nell’implementazione dei seguenti componenti:

1. Generazione dell’immagine;
2. Costruzione delle risorse lessicali per l’espansione delle query e per l’indicizzazione del materiale documentale;
3. Costruzione delle schede terminografiche per le entità da linkare
4. Modello vettoriale per la rappresentazione di utenti, query e cluster di documenti;
5. Implementazione delle strategie di document retrieval avanzato, re-ranking e profiling.

4 Evoluzione del motore di ricerca

L’evoluzione del motore si basa sulla definizione di un’ontologia del dominio di ricerca che viene “caricata” dal motore. Le occorrenze nei testi delle classi definite dall’ontologia sono identificate, mediante tecniche di estrazione di informazione: in particolare tramite tecniche di supervised sequence learning che utilizzano per il training e il decoding un insieme di contesti e l’insieme delle risorse lessicali. L’unità informativa non è più, solo, il documento, ma pezzi di informazione specifica contenuti al suo interno. L’interfaccia si focalizza sull’analisi della query in linguaggio naturale.

4.1 Costruzione dell’ontologia e Information Extraction

Mano a mano che classi di oggetti informativi rilevanti vengono isolati; per es, “titoli di studio”, queste classi vengono definite e gli viene associata una semantica che servirà alla loro utilizzazione in fase di pattern-matching. Alla definizione di ogni classe dell’ontologia del dominio segue la produzione delle risorse necessarie alla sua identificazione tramite machine learning. Le classi vengono definite in base alla creazione di strutture dati adeguate che ne specificano la semantica:

- formalizzazione, attributi, etc.;
- contesti d’uso campionati dal database di testi;

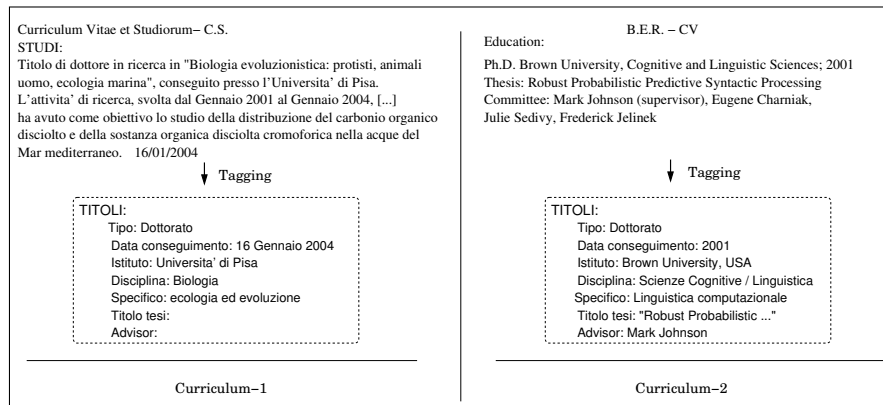


Figura 4: Esempio di estrazione di informazione riguardante titoli di studio, contenuti in curriculum caratterizzati da diversa struttura e in lingue diverse.

- informazione lessicale associata:
 - variazioni terminologiche
 - sinonimi, iperonimi, metonimi, etc.
 - liste di nomi: luoghi, istituti, docenti, titoli, etc.
 - liste di termini associati: (“libro”, “stampante”, “editor”,

La classe così codificata si aggiunge all'ontologia. Tramite tecniche di *information extraction* (IE) un'immagine corrente del database viene periodicamente annotata con le occorrenze di tutte le classi definite. Questa metainformazione viene utilizzata dal motore di ricerca.

Per esempio, informazione strutturata relativa a titoli di studio, contenuta in curricula scritti in lingue diverse e con diversi stili di formattazione, può essere mappata in una nuova rappresentazione strutturata condivisa. La Figura 4 illustra graficamente due casi di curriculum diversi che dopo essere stati annotati semanticamente (IE), sono mappati in una nuova rappresentazione strutturata condivisa. Più concretamente, la fase di estrazione di informazione consiste nell'identificazione nel testo delle stringhe di parole che corrispondono ad occorrenze di classi specifiche definite nell'ontologia, come “advisor” o “Ph.D.” negli esempi sotto:

- (2) “STUDI: [Titolo di dottore in ricerca]_{Dottorato} in “Biologia evolutivistica: protisti, animali, uomo ...”
- (3) “Education: [Ph.D.]_{Dottorato} [Brown University]_{Istituto} ... Committee: [Mark Johnson]_{Advisor} (supervisor), Eugene Charniak, ...”.

4.2 Analisi semantica di query e documenti

L'utilizzazione dell'informazione semantica definita dall'ontologia entra in gioco a livelli di diversa complessità. In primo luogo pu essere usata come metainformazione direttamente nel re-ranking dei documenti, preferendo quelli dove

oltre a un matching di parole c'è un matching semantico. In secondo luogo l'informazione può essere usata per evidenziare e isolare porzioni di testo ritenute rilevanti alla query stessa. Queste porzioni di testo dove applicabile possono anche essere linkate ad altre risorse corrispondenti (e.g. i file delle pubblicazioni). Infine, nella sua implementazione piena, il motore semantico opera un'analisi di query complesse e strutturate come "libri di A.B. ultimi cinque anni", la cui risposta potrebbe dover essere estratta da numerosi documenti in cui l'informazione non neanche espressa esplicitamente, per esempio in una lista di pubblicazioni di A.B. ci potrebbe essere un'occorrenza come "A.B. "Introduction to Algorithms"" (2001), MIT Press", in cui il fatto che si tratti di un libro dipende dall'informazione implicita nella referenza e non nella presenza del termine "libro".

4.3 Componenti

La seconda fase è caratterizzata dai seguenti elementi:

1. Formalizzazione e formattazione machine-readable in strutture adeguate dell'ontologia, contenenti la semantica dei concetti definiti, informazione lessicale associate, contesti d'uso campionati dal database;
2. Tagging dei documenti per l'identificazione di occorrenze di classi;
3. Implementazione di strategie di pattern matching semantico, dalla semplice query expansion/local clustering alla analisi semantica di query strutturate.

5 Estensioni ulteriori del motore di ricerca

Le seguenti sono solo alcune possibili estensioni che il modello definito precedentemente può supportare:

1. Adozione di tecnologie di Semantic Web; e.g. Portale semantico, etc;
2. Navigazione grafica: attraverso la struttura dell'ontologia
3. Interfaccia basata su linguaggio naturale di tipo automatic Q/A";

Inoltre, possibili direzioni di ulteriore ricerca e sviluppo potrebbero riguardare tra l'altro temi come l'implementazione di un *parser* semantico da associare all'ontologia nell'analisi dei testi, oppure l'implementazione di tecniche di indicizzazione e ricerca su supporti multimediali: primariamente immagini (anche audio, video).

6 Gruppi e competenze

Il progetto implementa una visione di ricerca e sviluppo a ciclo rapido, dove i diversi moduli supportano l'architettura corrente e costituiscono gli elementi portanti delle evoluzioni successive. Per esempio, lo sviluppo di un'ontologia complessa e proiettata sul database costituirà non solo il supporto del motore

di ricerca ma anche la navigazione grafica concettuale. Le risorse lessicali supportano il local clustering, profiling e relevance feedback del motore di ricerca su documenti, ma sono anche componenti fondamentali del motore di estrazione di informazioni. La seguente lista identifica i componenti fondamentali e le competenze necessarie:

1. ILC: estrazione di risorse terminologiche dal database del CNR e altre fonti testuali e loro strutturazione e proiezione su classi dell'ontologia attraverso tecniche di machine learning; strumenti di annotazione linguistico-semantica dei documenti basati su tecniche di NLP; analisi delle query utente in linguaggio naturale.
2. Gruppo di lavoro SSLMIT dell'Università di Bologna (Forlì): produzione di schede terminografiche per l'annotazione delle occorrenze degli elementi dell'ontologia nel corpus; specifiche terminologiche e contesti d'uso delle classi;
3. Gruppo LABDOC (Università della Calabria): implementazione di una procedura operativa di aggiornamento semiautomatico del "thesauro" del sistema di classificazione dei documenti.
4. SRT: supporto allo sviluppo del motore di ricerca.
5. LOA: coordinamento scientifico, assistenza al SRT per il trasferimento tecnologico, definizione dell'ontologia, tecniche avanzate di IR e IE;

Gestione e accesso intelligenti del contenuto di repertori documentali mediante l'uso di strumenti avanzati di analisi linguistica automatica

Nicoletta Calzolari, Simonetta Montemagni, Vito Pirrelli
Istituto di Linguistica Computazionale ILC-CNR
Alessandro Lenci
Università di Pisa Dipartimento di Linguistica

Obiettivo generale

Consentire l'indicizzazione e l'interrogazione automatiche di testi documentali (possibilmente corredati di metadati) attraverso l'uso di repertori terminologici di dominio, concetti strutturati, mappe concettuali e domande libere in linguaggio naturale.

Esempi di risultati attesi

Indicizzazione e accesso ai documenti per

- ✓ termini specifici del dominio
 - estratti (semi)automaticamente a partire da testi di dominio
 - lemmatizzati (ricondotti a un unico esponente lessicale)
esempio: *approvare, approvati, approveranno* > APPROVARE
 - sia semplici che strutturalmente complessi (unità mono e multi-lessicali)
esempio: *investimento, aggiornamento della previsione di spesa*
 - normalizzati per classi di equivalenza
esempio:
{*servizio per le imprese, servizio alle imprese, servizio all'impresa*}
{*inserimento al lavoro, inserimento lavorativo*}
- ✓ nomi propri
 - normalizzati (possibili varianti sono ricondotte a un'unica norma)
esempio: *v. A. Saffi, via Aurelio Saffi, via Saffi* > VIA AURELIO SAFFI
 - etichettati per semplici classi semantiche (ad es. toponimo, nome proprio di persona, istituzione ecc.)
- ✓ famiglie di termini concettualmente affini
 - attraverso relazioni gerarchizzate di iponimia/iperonimia

esempio: SERVIZI > SERVIZI ALLE IMPRESE > SERVIZI ALLE IMPRESE PRIVATE > SERVIZI INTEGRATI ALLE IMPRESE PRIVATE

- attraverso relazioni di “quasi-sinominia” (clustering concettuale)
esempio:
{LEGGE, DISEGNO, REGOLAMENTO, DISPOSIZIONE, PIANO ...}
{AZIONE, INIZIATIVA, ATTIVITÀ, SERVIZIO, PROCEDURA, PROGETTO, INTERVENTO, PROGRAMMA ...}
- attraverso relazioni di tipo derivazionale
esempio: {industria, industriale, industrializzare, industrializzazione ecc.}

✓ relazioni concettuali tra più termini (mappa concettuale)

esempio:

{FORNIRE, OFFRIRE}	– oggetto diretto → {AIUTO, PRESTAZIONE, SERVIZIO}
	– soggetto → {AMMINISTRAZIONE PUBBLICA}

Interrogazione di basi di dati documentali mediante domande in linguaggio naturale e l'appoggio di ontologie e metadati

esempio: *quali commesse presentano nel 2005 un finanziamento esterno superiore a 100.000 euro?*

Programmazione attività

Obiettivi a breve termine (6-10 mesi)

- Applicazione estesa del software esistente a vasti repertori documentali omogenei per dominio/area tematica. Analisi dei risultati ottenuti. Personalizzazione del software. (mese 6)
- Applicazione algoritmo di clustering terminologico a repertori terminologici estesi. Analisi dei risultati ottenuti. Personalizzazione del software. (mese 8)
- Implementazione e verifica di strategie di filtraggio della terminologia di dominio attraverso un'analisi comparativa dei repertori estratti per ciascuna area tematica. (mese 10)

Obiettivi a medio termine (16-24 mesi)

- Classificazione semantico-lessicale dei nomi propri (nomi propri di persona, toponimi, nomi di istituzione) (mese 16)

- Classificazione automatica dei fallimenti dell'analisi lessicale per categorie lessicali (forestierismo, tecnicismo, nome proprio, refuso ecc.) e grammaticali (nome, aggettivo, verbo ...) mese (18)
- Sviluppo di mappe concettuali per la navigazione intelligente del contenuto documentale (mese 20)
- Strutturazione automatica dei repertori terminologici per famiglie lessicali di tipo derivazionale (ad es. *industria, industriale, industrializzare, industrializzazione* ecc.) (mese 22)
- Interrogazione della base di dati documentale tramite domande in linguaggio naturale (ad es.; *quali commesse presentano nel 2005 un finanziamento esterno superiore a 100.000 euro?*) (mese 24)

Risorse umane previste

2 assegni di ricerca di fascia media per il potenziamento e la personalizzazione degli strumenti esistenti

2 ricercatori per il disegno e lo sviluppo di nuove funzionalità (classificazione di fallimenti lessicali; strutturazione lessicale per famiglie di derivati; sviluppo mappe concettuali; interrogazione in linguaggio naturale di database documentali)

1 contratto d'opera/assegno per lavoro di integrazione/ingegnerizzazione del software prodotto

Descrizione Analitica del Modulo Labdoc

Tematiche di ricerca

Gestione documentale e sistemi di classificazione pre coordinati. Costruzione di lessici specialistici strutturati. Indicizzazione.

Stato dell'arte

Allo stato i sistemi di classificazione normalmente utilizzati sono, sostanzialmente, divisibili in due macro categorie: pre-coordinati e post-coordinati. Il vantaggio dei primi è la loro esistenza indipendentemente dal contesto documentale di riferimento pur pagando il prezzo di una necessaria genericità rispetto allo specifico contesto di applicazione. Per contro, i secondi, risultano estremamente aderenti al contesto in quanto costruiti sulla base di relazioni specifiche ma presentano il loro maggior limite nella impossibilità di riutilizzo senza adattamenti in ambiti diversi da quello di generazione.

Competenze, tecnologie e tecniche di indagine

Le competenze di tipo documentale (analisi concettuale, estrazione di voci indice e costruzione di relazioni tra termini) e di tipo linguistico-computazionale (costruzione delle classi, definizione di reti semantiche e strutturazione di lessici specialistici) costituiscono la base per la progettazione di sistemi di gestione documentale e di formazione a distanza con particolare attenzione alle problematiche di profiling degli utenti per la creazione di percorsi formativi personalizzati, basati sull'estrazione semantica dai documenti e relativa costruzione semi-automatica di ontologie di dominio.

Collaborazioni (partner e committenti)

MIUR, Ministero Comunicazioni, CNR, Formez, Cnipa, Italdat Siemens, FileNet Italy, Prisma Engineering, ItConsult, Regione Calabria.

Potenziale impiego

- per processi produttivi

Snellimento dei processi di recupero dell'informazione, spesso troppo pesanti a causa dello scarso adattamento del lessico di dominio ai documenti aggiunti in itinere.

- per risposte a bisogni individuali e collettivi.

Riduzione del rumore informativo in risposta alle richieste degli utenti finali, e maggiore efficacia nella gestione dei processi informativi propri del sistema.

Obiettivi (Max 1200 caratteri)

Implementazione di una procedura operativa di aggiornamento semiautomatico di sistemi di classificazione pre coordinati con le relazioni semantiche estratte da corpora testuali strutturati e non strutturati all'interno dello stesso dominio di riferimento. Le metodologie proposte hanno lo scopo di analizzare i testi oggetto di analisi, estraendone i termini rilevanti e strutturandoli sulla base delle relazioni gerarchiche di un thesauro documentale.

Il thesauro ottenuto verrà, quindi, posto in matching con quello contenuto nel glossario terminologico - derivante dal machine learning - ottenendo così un aggiornamento ricorsivo dei termini contenuti e delle relazioni esistenti tra loro, ampliandone e precisandone il "vocabolario" in modo da renderlo maggiormente aderente all'ambito concettuale di riferimento.

XTerm online

XTerm online è un Content Management System (CMS) progettato per la gestione di knowledge bases di tipo terminologico. Il sistema è interamente basato su database e integra a) una banca dati di schede terminologiche, b) un motore di interrogazione di corpora e c) una base dati documentale.

Al momento in cui si scrive il presente documento, XTerm Online è un prototipo funzionante in fase di alpha testing.

Architettura

L'applicazione desktop gira in ambiente Windows NT/2K/XP, l'applicazione web del sistema è basata su un architettura di tipo LAMP (Linux, Apache, MySQL, PHP) e si serve di un normale browser web lato client.

Il Content Management System

A. BANCA DATI TERMINOLOGICA

La banca dati terminologica di XTerm consiste in un database sul quale è possibile effettuare varie viste a seconda delle esigenze di consultazione.

Tipicamente si utilizzano quattro tipi di vista:

1. **glossario monolingue**, che permette di visualizzare tutti i termini nella lingua scelta presenti all'interno della banca dati;
2. **glossario completo multilingue**, una tabella sinottica che permette di visualizzare simultaneamente termini ed equivalenti in tutte le lingue;
3. **per dominio**, visualizzazione monolingue dei termini categorizzati in base al dominio di appartenenza;
4. **singola scheda**, completa di tutte le informazioni terminologiche relative ad un termine (definizione, contesto, fonti ecc).

L'inserimento dei dati nel database può avvenire in due modi:

1. per mezzo di un'applicazione desktop in ambiente Windows che lavora su database Access. L'applicazione è in grado di gestire utenti multipli che lavorano contemporaneamente sullo stesso database, l'unica limitazione di questo approccio è che gli utenti devono essere in grado di accedere allo stesso file, il che normalmente presuppone che tutti gli utenti lavorino nella stessa rete locale.

Una volta creato, il database può essere:

- esportato come ipertesto statico, vengono generate semplici pagine HTML, una per ogni termine, più una pagina generale contenente l'indice sinottico multilingue delle schede;
 - convertito in formato MySQL in modo da poter essere direttamente visualizzato su Internet per mezzo dell'applicazione web XTerm online;
2. con un normale browser web, direttamente su Internet/Intranet servendosi di XTerm online; in questa modalità è naturalmente possibile anche intervenire su database creati utilizzando il metodo 1.

B. GESTIONE DEI CORPORA

Il motore di ricerca sui corpora è un'applicazione interamente web-based che sfrutta il Corpus Workbench sviluppato presso l'Institut für Maschinelle Sprachverarbeitung di Stoccarda (<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>).

Il motore prevede la possibilità di effettuare due tipi di query sui corpora:

1. la **ricerca semplice** permette la ricerca per parola/frase esatta oppure per lemma;
2. la **ricerca avanzata** permette di utilizzare il sofisticato linguaggio di interrogazione previsto dal Corpus Query Processor.

Il sistema è in grado di gestire corpora di grandi dimensioni, attualmente viene impiegato per l'interrogazione (tra gli altri) di "Repubblica", un corpus di italiano giornalistico di 380 milioni di tokens sviluppato dal gruppo di ricerca in Linguistica dei Corpora della SSLMIT di Forlì (Baroni et al. 2004).

C. BASE DATI DOCUMENTALE

La base dati documentale raccoglie i testi presenti nei corpora, permettendone:

- la consultazione diretta da un indice generato dinamicamente;
 - la consultazione a partire dalle schede terminologiche.
3. I documenti vengono formattati in XML e possono essere integrati con metadati quali autore, data di pubblicazione, status amministrativo ecc.

Bibliografia

M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, M. Mazzoleni. 2004. *Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*. Proceedings of LREC 2004.

KEEx

Knowledge Enhancement and Exchange

Versione 3.2.1

Soluzione tecnologica per il knowledge management distribuito

White Paper

*Distributed Thinking S.p.A
Partita IVA 01836390227.
Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)
Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943
Sede Operativa: Via F. Zeni 8 – 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233*

KEEx

Knowledge Enhancement and Exchange

Contenuti

1.	Introduzione	3
2.	Peer-to-peer	3
3.	Gestione della conoscenza	4
3.1.	Plug-in	5
3.1.1.	Plug-in di Microsoft Outlook	5
3.1.2.	Plug-in di Microsoft Internet Explorer	5
3.1.3.	Consistenza dati	5
3.2.	Comunità	6
3.3.	Sicurezza	6
4.	Ricerca	7
4.1.	Ricerca semantica	7
4.2.	Ricerca lessicale	7
4.3.	Ricerca concettuale	7
4.4.	Combinazione delle ricerche	8
5.	Sistema KEEx	8
5.1.	Personal Knowledge Manager peer (PKM peer)	9
5.2.	Source peer	9
5.3.	Normalization peer	9
5.4.	Super peer	9
5.4.1.	Zone	9
5.4.2.	Comunità imposte	10
5.4.3.	Sicurezza con zone e comunità imposte	11
5.4.4.	Configurazione dei PKM peer	11
5.4.5.	Monitoraggio della rete P2P di KEEx	12
5.5.	Rendez-Vous peer	12
5.6.	Rete peer-to-peer di KEEx	13
6.	Tecnologia di base di KEEx	13
6.1.	Comunicazione peer-to-peer	13
6.1.1.	Utilizzo di JXTA in KEEx	14
6.2.	Contesto	15
6.3.	Plug-in	15
6.4.	Sicurezza	15
6.5.	Ricerca semantica	16
6.5.1.	Normalizzazione	16
6.5.2.	Matching semantico	17
6.6.	Ricerca lessicale	18
6.6.1.	dtSearch®	18
6.6.2.	Jakarta Lucene	19
6.7.	Ricerca concettuale	20
6.8.	Topologia rete di KEEx	20
6.8.1.	Utilizzo del Rendez-Vous peer	20

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

1. Introduzione

La soluzione KEEEx (**K**nowledge **E**nhancement and **E**xchange) è un insieme di funzionalità a supporto della gestione della conoscenza distribuita all'interno dell'organizzazione. L'ipotesi su cui si fonda la soluzione è che l'esistenza nelle organizzazioni complesse di molteplicità di gruppi, pratiche, valori e prospettive diverse ed eterogenee, se opportunamente gestita, può trasformarsi da ostacolo alla comunicazione (com'è tradizionalmente intesa tale eterogeneità) a fonte di innovazione e apprendimento per l'organizzazione stessa.

Secondo questo approccio un sistema di KM che voglia valorizzare il patrimonio di conoscenze di un'organizzazione deve fornire gli strumenti necessari ad esplicitare **l'intero sistema conoscenza** che significa non solo **l'informazione** (intesa ad esempio come il documento) ma anche il **contesto interpretativo e relazionale** in cui essa ha un senso: sapere che uno specifico documento genericamente categorizzabile come "tecnologico" è stato creato nell'ambito di un progetto di messa in piedi di una infrastruttura di rete, invece che di Content Management, e sapere inoltre che lo stesso documento è stato raccomandato dal proprio capo, invece che dall'ultimo arrivato in azienda, arricchisce il documento di una serie di conoscenze estremamente rilevanti anche se non intrinseche al mero livello del suo contenuto.

In un'organizzazione esistono molti sistemi di conoscenze almeno, uno per ogni "entità" che ha una sua conoscenza locale in termini di contenuti e di contesti. Tali entità possono essere le singole persone, ma anche gruppi di persone che lavorano su temi simili (comunità di interesse) o obiettivi comuni (team di progetto), e anche repository tecnologici di documenti che contengono sia contenuti che le logiche organizzative necessari a classificarli (contesti). L'organizzazione è quindi vista come una rete di nodi di conoscenza che possono dinamicamente aggregarsi per formare gruppi, come team o comunità, che sono lo spazio dove i singoli nodi di conoscenza condividono le informazioni.

In KEEEx ogni attore (singolo, gruppo o repository) dell'organizzazione può essere identificato da un peer di una rete peer-to-peer (vedi 2). Grazie agli strumenti forniti dall'applicazione, ogni attore può costruire in modo autonomo il proprio "spazio della conoscenza" (vedi 3) in cui organizzare tematicamente diversi elementi del proprio patrimonio conoscitivo, quali i propri documenti, i documenti di altri peer, ma anche i riferimenti ad esperti e gruppi di esperti di dominio presenti nell'organizzazione, e i loro contesti interpretativi e relazionali. Oltre agli strumenti necessari ad organizzare tali conoscenze, l'applicazione fornisce ad ogni attore anche gli strumenti necessari a condividerle con il resto dell'organizzazione in modo selettivo (ad esempio: per ogni documento può essere stabilito chi vi può accedere) e tematico (ad esempio: ogni documento può essere connesso a certi tipi di ricerca tematica). Infine l'applicazione fornisce gli strumenti necessari a ricercare tali informazioni presso gli altri peer della rete. Coerentemente alla filosofia sottostante, per la quale il contesto interpretativo e relazionale in cui è inserita un'informazione ne determina il significato, la ricerca viene effettuata sfruttando diverse logiche in grado di reperire non solo il mero documento, ma anche il suo contesto (vedi 4).

2. Peer-to-peer

Una rete peer-to-peer (P2P) è un particolare tipo di rete nella quale, a differenza del paradigma client-server nel quale solo alcuni nodi server sono dedicati a fornire particolari funzioni, ogni nodo (peer) ha equivalenti capacità e responsabilità, essendo in grado sia di fornire che di accedere a determinati servizi, e ogni nodo può condividere risorse con gli altri nodi senza dipendere da un server centrale.

Nel caso di KEEEx, utilizzare una rete peer-to-peer significa poter abilitare le singole macchine degli utenti (che sono i nodi della rete aziendale) a offrire e fruire di servizi rilevanti per lo

Distributed Thinking S.p.A

Partita IVA 01836390227.





Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

scambio del patrimonio informativo di ognuno, quali il servizio di condivisione dei propri documenti ad altri peer, o il servizio di ricerca di documenti presso altri peer.

Le principali funzionalità che permettono l'interazione tra i peer sono:

-  **Discovery:** permette ad ogni peer di scoprire dinamicamente la presenza di altri peer (vedi 6.1.1), di comunità e dei servizi da essi offerti, senza ricorrere ad un servizio centralizzato di intermediazione (tipo yellow pages). Da personalizzare in base ai risultati dell'analisi organizzativa (in termini ad esempio di numero di peer e di gruppi).
-  **Ricerca:** permette ad ogni peer di indirizzare una ricerca (vedi 4) a uno o più peer o comunità (vedi 3.2).
-  **Download:** permette lo scambio di documenti tra peer.
-  **Chat:** permette a due utenti di comunicare direttamente.

3. Gestione della conoscenza

Ogni peer permette ad un attore dell'organizzazione di dare significato alle proprie risorse inserendole all'interno di una particolare struttura gerarchica denominata *contesto* (vedi 6.2). È possibile creare più contesti rispetto a cui organizzare le informazioni.

Nello specifico, un contesto è una classificazione gerarchica, composta da un certo insieme di nodi, denominati concetti, con associate delle etichette in linguaggio naturale che siano descrittive del loro contenuto (che può essere costituito da documenti o altre informazioni associate) e del significato di quel nodo all'interno del contesto stesso. La categoria è invece l'insieme dei concetti a partire dal concetto radice, che compongono un percorso sul contesto e che identificano univocamente il concetto.

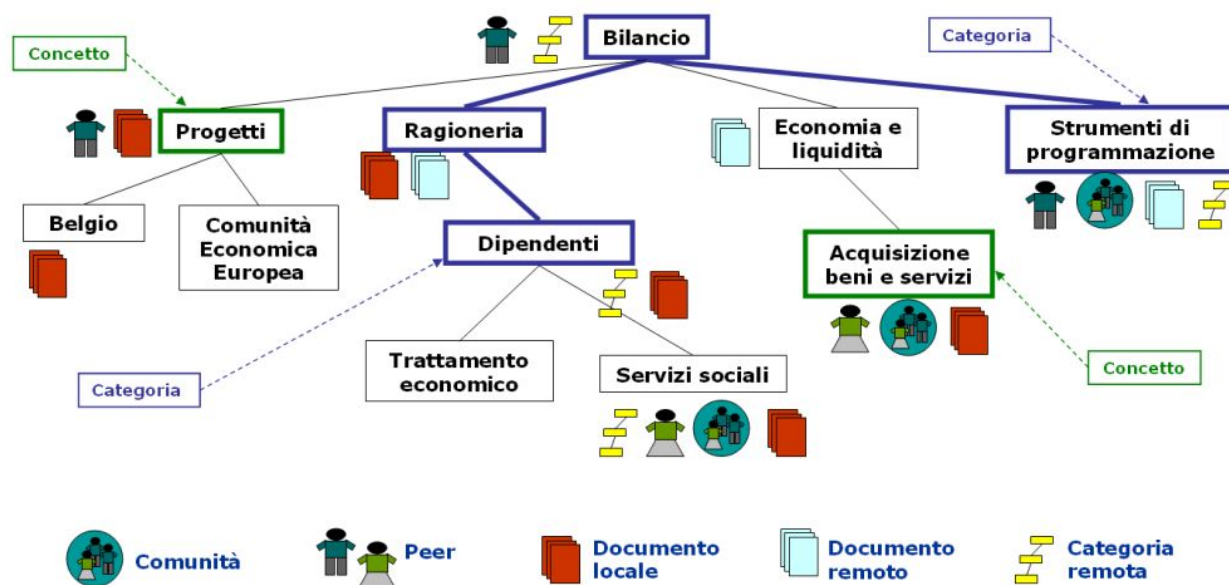


Figura 1: Contesto

A differenza del file system, tali strutture permettono di organizzare "concettualmente" non solo i documenti locali (documenti sul file system, mail) ma anche altre informazioni come collegamenti a indirizzi di pagine web (URL), documenti remoti (cioè documenti che sono fisicamente presso altri peer), singoli esperti (peer) o gruppi di esperti (comunità), e relazioni con contesti di altri peer (categorie remote).

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

Nella soluzione KEEEx i contesti hanno una duplice valenza: vengono utilizzati sia come strumento per organizzare la propria conoscenza, che come strumento per ricercare informazioni sulla conoscenza di altri peer attraverso la ricerca semantica (vedi 4).

Le stesse informazioni possono essere simultaneamente classificate rispetto a più categorie (vedi 6.2) e rispetto a più contesti, realizzando così viste tematiche personalizzate sullo stesso insieme di informazioni.

Per costruire un contesto si può partire da zero utilizzando un ambiente di editing oppure è possibile importare strutture gerarchiche già esistenti nel pc dell'utente, come ad esempio porzioni di file system.

3.1. Plug-in

L'insieme dei plug-in proposti nella soluzione KEEEx ha lo scopo di abilitare i peer ad interagire con applicazioni locali di uso comune rendendone disponibili i contenuti ad altri peer. Tali contenuti, pur essendo creati da applicazioni diverse, possono essere gestiti in maniera unitaria nelle varie strutture categoriali (i contesti) che l'utente avrà creato in KEEEx, alla pari dei contenuti presenti nel file system.

3.1.1. Plug-in di Microsoft Outlook

Il plug-in di Microsoft Outlook aggiunge due bottoni all'interno di Microsoft Outlook che permettono all'utente rispettivamente di condividere una o più mail in un PKM peer (vedi 5.1) e vedere quali mail sono state condivise.

Nel momento in cui l'utente condivide una o più mail, queste vengono estratte da Outlook e passate al PKM peer che ne indicizza il testo, considerando anche eventuali documenti allegati. In questo modo una mail può essere trovata via ricerca lessicale (vedi 6.6) sia se contiene la parola chiave nel testo della mail sia se la contiene in qualcuno degli allegati.

3.1.2. Plug-in di Microsoft Internet Explorer

Come per il precedente, anche il plug-in di Microsoft Internet Explorer aggiunge due bottoni all'interno di Explorer che permettono all'utente rispettivamente di condividere l'indirizzo della pagina web (URL) correntemente visualizzata nel browser con il PKM peer (vedi 5.1), e vedere quali URL sono stati condivisi.

Un utente che sta navigando nel web usando Microsoft Internet Explorer può decidere di inserire nel proprio PKM peer (e quindi condividerlo con gli altri peer) l'indirizzo della pagina che sta visualizzando. Nel momento in cui l'utente condivide l'indirizzo tramite il bottone di condivisione, il link a tale pagina viene passato al PKM. In questo caso non viene indicizzata la pagina HTML, perché la maggior parte delle pagine web sono dinamiche e quindi il contenuto cambia molto rapidamente. Quello che viene indicizzato è il titolo e l'URL.

3.1.3. Consistenza dati

Poiché il sistema KEEEx tratta documenti che sono gestiti da altre applicazioni, (File System, Microsoft Outlook) è necessario controllare se tali documenti condivisi con un PKM peer (vedi 5.1) o con un Source peer (vedi 5.2) sono stati rimossi, rinominati o spostati nell'ambito delle applicazioni native. Il sistema KEEEx segnala l'eventuale inconsistenza dei dati all'utente.

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

3.2. Comunità

Una comunità in KEEEx è vista come un insieme di peer che condividono l'interesse a fornire un servizio come gruppo invece che come singoli (ad esempio fornire informazioni su un determinato argomento). L'elemento aggregante della comunità di interesse può essere rappresentato da ogni peer con uno o più contesti.

Tutti i peer delle rete possono proporre comunità aperte a cui gli altri peer possono liberamente decidere di aderire. Quando una ricerca (vedi 4) è rivolta ad una comunità, tutti i peer attivi appartenenti alla comunità ricevono la richiesta e ognuno risponde se può alla richiesta.

Nella configurazione di base una comunità è di tipo "aperto", il che significa che tutti i PKM peer sono liberi di creare comunità e di aderirvi, senza nessun tipo di vincolo; alternativamente si possono avere le zone (vedi 5.4.1) o le comunità imposte dall'organizzazione (vedi 5.4.2). Questi ultimi due tipi di comunità, rappresentando in KEEEx gruppi di utenti decisi per qualche ragione dall'organizzazione, possono essere utilizzati anche per creare le ACL (Access Control List) associate ai documenti (vedi 3.3 e 5.4.3).

3.3. Sicurezza

Questa funzionalità permette di gestire la sicurezza della rete P2P a due livelli differenti:

- accesso degli utenti ai documenti condivisi: ogni utente crea localmente ad ogni peer delle ACL (Access Control List) associate ad ogni documento che decide di rendere disponibile tramite il sistema KEEEx. Tali ACL possono contenere utenti singoli (rappresentati nel sistema KEEEx da peer) o gruppi di utenti (zone, comunità imposte, o altri gruppi decisi dall'organizzazione e presenti nel suo sistema di gestione della sicurezza).
- scambio di messaggi: è possibile criptare i messaggi che i peer si scambiano per rendere questo scambio sicuro sulla rete P2P.

Per fare questo vengono utilizzati meccanismi basati su creazione ed uso di certificati, utilizzando l'infrastruttura a chiave pubblica e privata (PKI), firma digitale e crittografia dei messaggi. Il certificato rappresenta univocamente una entità della rete P2P, la firma digitale permette di verificare che un messaggio è stato inviato da una determinata entità e non è stato manomesso, ed infine la crittografia rende un messaggio leggibile solo ad una specifica entità destinataria (vedi 6.4).

Ci sono vari modi in cui ogni peer può ottenere la lista di utenti e gruppi dell'organizzazione: può ottenere la lista degli utenti in modo dinamico, basandosi sul meccanismo di discovery dei peer (vedi 2), ricevere la lista di zone e comunità imposte tramite configurazione inviata dal peer di amministrazione (vedi 5.4.3), oppure integrandosi direttamente con il sistema per la gestione della sicurezza utilizzato dall'organizzazione stessa. Quest'ultima soluzione permette anche di verificare l'identità di un utente oltre ad accedere alla lista di tutti gli utenti e gruppi presenti in una rete. La soluzione KEEEx è in grado di integrarsi con sistemi di autenticazione quali Active Directory di Microsoft e sistemi LDAP (Lightweight Directory Access Protocol).

Il livello di sicurezza può essere ulteriormente raffinato utilizzando un ente certificatore (CA, Certificate Authority) se all'interno di una stessa organizzazione, o utilizzando sistemi di cross certificazione fra domini differenti nel caso di organizzazioni diverse o distribuite.

4. Ricerca

Il principale metodo che un peer utilizza per interagire con gli altri peer della rete è la ricerca di documenti. Ogni peer può ricercare documenti presso altri peer in modi diversi che possono essere combinati tra loro.

4.1. Ricerca semantica

I contesti utilizzati per rappresentare la conoscenza locale vengono anche utilizzati per ricercare documenti presso altri peer; il parametro della ricerca in questo caso è una categoria (vedi 6.2) scelta in un contesto. L'algoritmo di matching semantico (vedi 6.5.2) che sta alla base di questa funzionalità è in grado di comparare la categoria ricevuta come input con le categorie appartenenti ai contesti dei peer cui è rivolta la ricerca, e individuare tra queste quelle con cui la categoria input ha una relazione semantica (vedi 6.5.2). Il risultato della ricerca sarà quindi l'insieme dei documenti classificati nelle categorie che hanno una relazione semantica con la categoria input della ricerca.

Al fine di poter individuare relazioni semantiche tra nodi di contesti diversi, questi ultimi devono essere sottoposti ad una prima fase detta *normalizzazione* (vedi 6.5.1.) che ha lo scopo di arricchire di informazioni grammaticali di varia natura (semantiche, morfologiche, lessicali) le etichette che lo compongono, al fine di renderne più esplicito il significato e favorire l'individuazione di risultati da parte dell'algoritmo di matching.

La ricerca semantica utilizza come input una categoria di un contesto che è stato normalizzato e tramite l'algoritmo di matching cerca le relazioni semantiche tra la categoria in input e le categorie dei contesti a disposizione presso i peer interrogati. Come output restituisce una lista ordinata di quelle categorie che hanno almeno un documento classificato. L'ordinamento è definito dalle relazioni semantiche trovate. In particolare verranno presentate prima le categorie che hanno una relazione semantica di "equivalenza" con la categoria di input, poi quelle che risultano essere "meno generali" di questa, quindi quelle "più generali" e infine quelle categorie per le quali è stata trovata una percentuale di compatibilità con la categoria di input (vedi 6.5).

4.2. Ricerca lessicale

La ricerca lessicale permette di cercare documenti tramite una o più parole chiave legate da operatori logici (*AND* e *OR*). Le parole chiave vengono ricercate all'interno dei documenti o nel nome del file. Per questo tipo di ricerca KEEEx si avvale di un indicizzatore lessicale (vedi 6.6).

L'algoritmo per la ricerca lessicale restituisce una lista ordinata di documenti. L'ordinamento è basato su un punteggio assegnato ad ogni documento (*confidence*), calcolato dall'indicizzatore lessicale, che si basa sull'occorrenza delle parole chiave, usate come parametro della ricerca, all'interno del documento stesso. I documenti trovati solo perché le parole chiave si trovano all'interno del nome del file hanno *confidence* 0 quindi compariranno in fondo alla lista dei documenti della risposta. I documenti vengono presentati all'utente con la categoria cui appartengono nel contesto di origine (cioè uno dei contesti del peer che sta rispondendo alla ricerca in questione).

4.3. Ricerca concettuale

Come per la ricerca lessicale anche questo tipo di ricerca è basata su una o più parole chiave legate da operatori logici (*AND* e *OR*), ma le parole chiave vengono ricercate nelle etichette che descrivono i concetti dei contesti. Il risultato della ricerca sarà quindi l'insieme dei

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

documenti che sono stati classificati in un concetto che soddisfa i parametri della ricerca (vedi 6.7).

L'algoritmo per la ricerca concettuale restituisce una lista ordinata di concetti (nodi dell'albero) nei quali è stato classificato almeno un documento (vedi 6.2). L'ordinamento è fatto in base al numero di volte in cui il/i termine/i ricercato appare in un concetto. All'inizio dell'elenco compariranno i concetti identici alla stringa ricercata, di seguito si avranno i concetti che contengono parti della stringa.

4.4. Combinazione delle ricerche

Le ricerche sopra descritte possono essere combinate e quindi anche il risultato sarà una combinazione dei risultati delle singole ricerche. L'ordinamento globale è il seguente:

1. *Categoria e Parole Chiave.* Contiene i documenti trovati con ricerche:
 - Semantica, concettuale e lessicale. I documenti appartengono a delle categorie che sono in relazione semantica con quelle utilizzate per effettuare la ricerca semantica, e sono classificati in concetti che rispondono alla ricerca concettuale. Inoltre hanno nel testo o nel nome del file le parole chiave usate per la ricerca lessicale.
 - Semantica e lessicale. I documenti appartengono a delle categorie che sono in relazione semantica con quelle utilizzate per effettuare la ricerca semantica. Inoltre hanno nel testo o nel nome del file le parole chiave usate per la ricerca lessicale.
 - Concettuale e lessicale. I documenti appartengono a dei concetti che rispondono alla ricerca concettuale. Inoltre hanno nel testo o nel nome del file le parole chiave usate per la ricerca lessicale.
2. *Categoria.* Contiene i documenti trovati con ricerche:
 - Semantica e concettuale. I documenti restituiti appartengono a delle categorie che sono in relazione semantica con quelle utilizzate per effettuare la ricerca semantica; inoltre tali documenti sono classificati in concetti che rispondono alla ricerca concettuale.
 - Semantica. I documenti restituiti appartengono a delle categorie che sono in relazione semantica con quelle utilizzate per effettuare la ricerca semantica.
 - Concettuale. I documenti restituiti sono classificati in concetti che rispondono alla ricerca concettuale.
3. *Parole Chiave.* Contiene i documenti trovati solo con ricerca lessicale cioè che hanno nel testo o nel nome del file le parole chiave usate come parametro della ricerca lessicale.

5. Sistema KEEx

Il sistema KEEx permette di rappresentare un'organizzazione come un insieme di peer in grado di raggrupparsi o essere raggruppati dinamicamente in comunità, e scoprire e interagire con altri peer e comunità della rete. L'insieme di tutte le conoscenze gestite localmente ai peer compongono la conoscenza dell'organizzazione nella sua totalità.

La soluzione KEEx prevede cinque tipi di peer diversi: PKM peer, Source peer, Normalization peer, Super peer e Rendez-Vous peer.

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

5.1. Personal Knowledge Manager peer (PKM peer)

Il Personal Knowledge Manager peer (PKM peer) permette ad un utente della rete di gestire la propria conoscenza rendendola allo stesso tempo disponibile alla rete P2P di KEEEx secondo le logiche decise dall'utente stesso (vedi 3). L'utente, sfruttando i vari meccanismi della rete P2P di KEEEx (vedi 2) e le funzionalità di ricerca (vedi 4), può interagire con gli altri peer e con le comunità.

5.2. Source peer

Il Source peer è un peer in grado di integrare sorgenti di dati documentali presenti in azienda eventualmente classificati con tassonomie. A differenza del PKM peer che è gestito dal singolo utente, il quale mette a disposizione degli altri peer della rete la sua conoscenza di livello locale, il Source peer viene gestito a livello di "organizzazione" (attraverso ad esempio un utente amministratore) e permette a quest'ultima di rendere disponibili agli utenti della rete P2P di KEEEx (PKM peer) il livello più generale della "conoscenza istituzionale".

Le sorgenti dati che un source peer può integrare sono:

- *Siti Intranet.* Il source peer è in grado di rendere reperibili i contenuti statici di un sito Intranet (pagine html, documenti in vari formati) attraverso le funzionalità di ricerca semantica, lessicale e concettuale sfruttando la struttura a directory del sito stesso.
- *Strumenti Content Management.* Il source peer è in grado di rendere reperibili i contenuti di un sistema di Content Management attraverso le funzionalità di ricerca semantica, lessicale e concettuale.
- *Search Engine.* Il source peer è in grado di trasformare la struttura categoriale di un Search Engine in un contesto. Tale contesto può essere utilizzato per effettuarvi ricerche di tipo semantico e concettuale. Inoltre consente di sfruttare le potenzialità dei search engine in termini di ricerca lessicale e concettuale, rendendola disponibile alle rete P2P di KEEEx.
- *Sorgenti Data Base.* Il source peer è in grado di rendere reperibili i contenuti di una base di dati integrando opportunamente le funzionalità di ricerca del peer con la base di dati.

5.3. Normalization peer

Il Normalization peer offre il servizio di normalizzazione dei contesti (vedi 4.1) ai vari peer della rete (PKM peer e Source peer). Tale servizio viene pubblicato nella rete e trovato in modo dinamico dai peer. Per la natura del servizio, tale peer dovrebbe essere installato su una macchina sempre attiva.

5.4. Super peer

Il Super peer permette di configurare, amministrare e monitorare la rete P2P di KEEEx. Combinando opportunamente la configurazione dei singoli peer (vedi 5.4.4) e la suddivisione in zone e comunità imposte (vedi 5.4.1 e 5.4.2), si ottiene la rete P2P che meglio rappresenta sia la rete fisica che funzionale dell'organizzazione.

5.4.1. Zone

PKM peer e Source peer sono istanziati dal Super peer all'interno di quelle che in KEEEx sono chiamate "zone" che partizionano la rete da un punto di vista logico per rispecchiare al meglio l'organizzazione. Se a un peer non viene assegnata una zona, o non vengono create zone i peer appartengono ad un unico gruppo di default chiamato "main group".

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

Una zona può ad esempio corrispondere ad un'unità organizzativa (un ufficio, una divisione dell'organizzazione) mentre il main group rappresenta tutta l'organizzazione. Un PKM peer o un Source peer hanno un unico gruppo di appartenenza (main group o una zona) e a seconda della configurazione possono scoprire e interagire con altri peer, zone, comunità imposte o aperte (vedi 5.4.4).

Su grandi numeri di peer, la scelta di non partizionare il main group con zone è sconsigliata. Avere un elevato numero di peer nel proprio gruppo di interazione non permette l'utilizzo ottimale dell'applicazione, perché l'utente accede a troppa informazione senza alcun filtro. Sapendo ad esempio che un certo peer appartiene ad una certa unità organizzativa (zona), o fa parte di un progetto (comunità imposta vedi 5.4.2) l'utente potrebbe orientare meglio la sua ricerca. Nella figura (Figura 2) supponendo di avere 36 peer (che non è un numero elevato per la soluzione KEEEx, ma serve solo per l'esempio), senza una partizione in zone e abilitando il discovery di peer ad un PKM peer (vedi 5.4.4), l'utente vedrebbe una lista di 36 peer da cui scegliere per fare la ricerca. Invece, configurando opportunamente i PKM peer e partizionando i peer in zone e comunità, l'utente vedrebbe prima i peer della propria zona, eventualmente altre zone o comunità e quindi i peer membri della altre zone o comunità.

Un altro aspetto per cui è conveniente partizionare il main group della rete P2P di KEEEx, avendo un elevato numero di peer, è legato alle performance di rete: avere molti peer che possono potenzialmente rispondere tutti insieme a fronte ad esempio di una richiesta di discovery da parte di un PKM peer, potrebbe aumentare sensibilmente il traffico di rete (vedi 6.1.1).

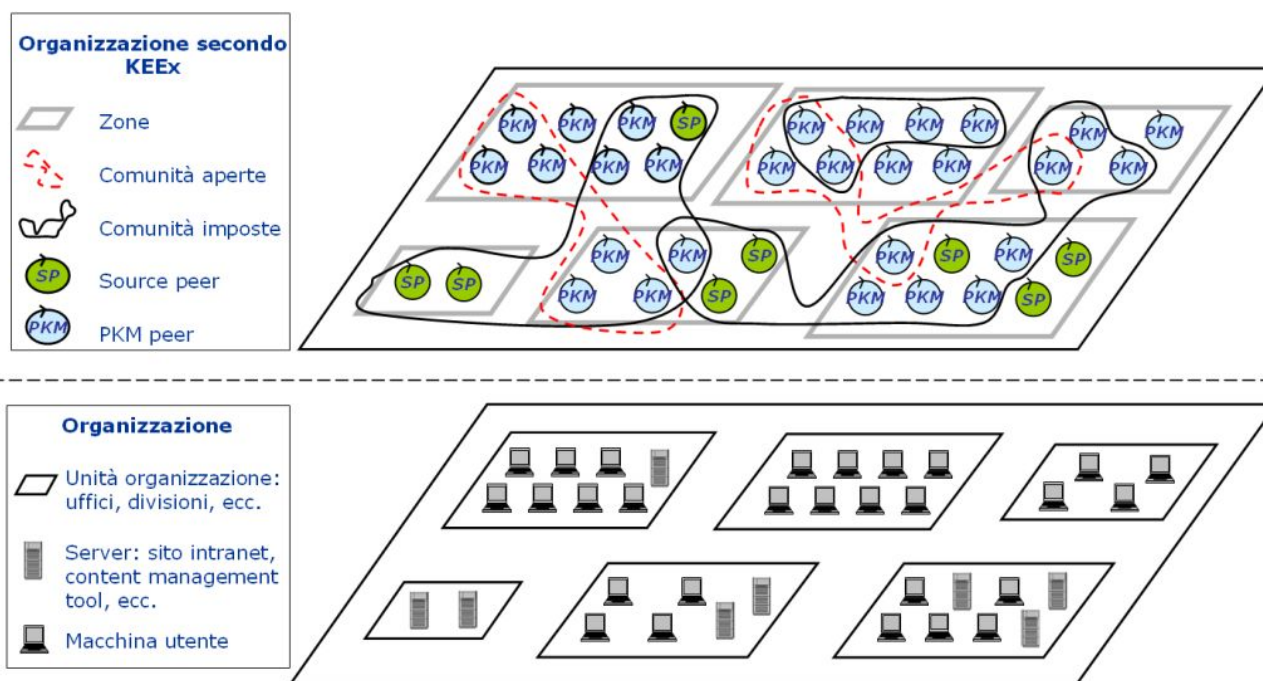


Figura 2: Soluzione KEEEx in un'organizzazione

5.4.2. Comunità imposte

Un'ulteriore partizione della rete può essere fatta dalle comunità che in KEEEx possono essere aperte (vedi 3.2) o imposte. Questo ultimo tipo, simile come filosofia a quella aperta, viene però creata dal Super peer e forza alcuni peer a parteciparvi. Può essere trasversale rispetto alle zone, cioè può comprendere peer di zone diverse. Un PKM peer o un Source peer inoltre possono appartenere a più comunità imposte. Le comunità imposte permettono

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

all'organizzazione di raggruppare utenti (PKM peer) e rendere disponibile, tramite Source peer, la conoscenza istituzionale (vedi 5.2) contestualmente a un gruppo. In questo modo si possono rispecchiare nella rete P2P di KEEEx i gruppi di lavoro dinamici e eventualmente temporanei (ad esempio i team di progetto) dell'organizzazione stessa.

Le comunità imposte possono in un certo senso derivare da comunità aperte, fondate e partecipate in modo spontaneo dai singoli PKM peer. Infatti se l'organizzazione monitorando la rete vede che una certa comunità aperta è utile a livello di organizzazione, potrebbe pensare di istituzionalizzarla creandone una imposta che la sostituisca.

5.4.3. Sicurezza con zone e comunità imposte

Zone e comunità imposte in KEEEx, come già detto, sono usate dall'organizzazione per rappresentare gruppi di utenti secondo esigenze di vario genere (vedi 5.4.1 e 5.4.2). Questi gruppi possono essere utilizzati anche dagli utenti per creare le ACL (vedi 3.3) associate ai singoli documenti. In questo modo se un utente aggiunge ad un ACL associata ad un documento una zona o una comunità imposta, tutti gli utenti (PKM peer) che sono membri della zona o comunità imposta potranno accedere al documento.

La lista delle zone e comunità imposte viene passata al PKM peer dal Super peer in fase di configurazione iniziale del PKM peer (vedi 5.4.4), e viene aggiornata ogni volta che l'amministratore crea o distrugge comunità imposte e zone.

5.4.4. Configurazione dei PKM peer

I singoli PKM peer possono essere configurati dal Super peer per abilitare o meno l'interazione con i vari elementi di KEEEx (peer, zone, comunità imposte e aperte). In particolare per ogni PKM peer è possibile configurarne le funzionalità di discovery, ricerca presso altri peer o comunità e creazione e adesione a comunità aperte.

Abilitazione funzionalità discovery

- *Discovery peer all'interno della zona di appartenenza.* Un PKM peer che appartiene ad una zona può scoprire i peer che appartengono alla stessa zona.
- *Discovery peer all'esterno della zona di appartenenza.* Un PKM peer può scoprire altri peer che sono nel main group e non appartenenti quindi a nessuna zona. Abilitare questa funzionalità potrebbe essere utile nella situazione in cui esistono tanti PKM peer divisi in zone e pochi Source peer che non appartengono a nessuna zona ma al main group. In questo modo i PKM peer possono scoprire i Source peer e tutti i PKM peer che stanno nella stessa zona (se abilitato la funzionalità del punto precedente).
- *Discovery di zone.* Un PKM peer può scoprire le altre zone oltre a quella di appartenenza.
- *Discovery membri per ogni zona.* Per ogni zona creata è possibile permettere ad un PKM peer di scoprire i peer membri di tale zona.
- *Discovery di comunità imposte.* Un PKM peer può scoprire le comunità imposte.
- *Discovery membri per ogni comunità imposta.* Per ogni comunità imposta creata si può permettere ad un PKM peer di scoprirne i membri.
- *Discovery di comunità aperte.* Dare la possibilità ad un PKM peer di scoprire le comunità aperte create dagli altri PKM peer.
- *Discovery di membri di comunità aperte.* Per tutte le comunità aperte, dare il permesso ad un PKM peer di scoprire i membri delle comunità aperte. Non è filtrato sulla singola comunità aperta come per le comunità imposte o per le zone, perché non è il Super peer che gestisce direttamente le comunità aperte.

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

Abilitazione funzionalità di ricerca presso altri peer o comunità

- *Ricerca broadcast a tutti i peer.* Dare la possibilità a un PKM peer di indirizzare una ricerca a tutti i peer che appartengono alla stessa zona (o main group) di appartenenza.
- *Ricerca indirizzata ai peer.* Permessso ad ogni PKM peer di indirizzare una ricerca ad altri peer che può scoprire. Un PKM peer potrebbe non avere la possibilità di fare ricerche presso altri peer direttamente, anche se scoperti via discovery membri di un qualche gruppo (comunità o zona), ma potrebbe avere la possibilità di fare query solo alle comunità o alle zone.
- *Ricerca indirizzata alle zone.* Per ogni zona creata si può decidere se un PKM peer può indirizzare una ricerca a tale zona oppure no.
- *Ricerca indirizzata alle comunità imposte.* Per ogni comunità imposta creata si può dire se un PKM peer può indirizzare una ricerca a tale comunità imposta oppure no.
- *Ricerca indirizzata alle comunità open.* Permessso ad un PKM peer di indirizzare una ricerca a tutte le comunità aperte comunità aperte che il peer è in grado di scoprire. Non è filtrato sulla singola comunità aperta come per le comunità imposte o per le zone, perché non è il Super peer che gestisce direttamente le comunità aperte.

Abilitazione funzionalità di creazione e adesione a comunità aperte

- *Creazione comunità aperte.* Dare la possibilità ad un PKM peer di creare comunità aperte.
- *Adesione a comunità aperte.* Dare la possibilità ad un PKM peer di aderire alle comunità aperte che vengono trovate con un'operazione di discovery. Anche in questo caso il permesso è generale per tutte le comunità aperte.

5.4.5. Monitoraggio della rete P2P di KEEx

Questa funzionalità permette un'analisi quantitativa di come gli utenti utilizzano il PKM peer e di come interagiscono tra loro all'interno della rete P2P di KEEx. Il Super peer in ogni momento può fare richiesta ai PKM peer attivi dei dati che ne descrivono l'utilizzo, permettendo così all'amministratore di fare l'analisi. I dati che ogni PKM peer fornisce sono elencati di seguito.

- *Numero di avvii del PKM peer e date del primo e dell'ultimo avvio.*
- *Numero documenti condivisi distinti in: condivisi con tutti, con nessuno o con qualcuno.*
- *Numero di discovery PKM peer fatti e data dell'ultimo.*
- *Numero di discovery Comunità fatti e data dell'ultimo.*
- *Numero di ricerche fatte e data dell'ultima.*
- *Numero di ricerche ricevute e data dell'ultima.*
- *Numero di download documenti fatti e data dell'ultimo.*
- *Numero di download documenti forniti e data dell'ultimo.*

5.5. Rendez-Vous peer

Nella soluzione KEEx questo tipo di peer non interagisce direttamente con gli altri tipi, ma è di supporto alla comunicazione tra i vari peer nel caso di alcune tipologie di rete, o dove si vuole mettere in comunicazioni due o più reti fisiche distinte, magari protette da firewall (vedi 6.8).

Come per il Normalization peer, anche questo peer (quando necessario) dovrebbe essere installato su una macchina sempre attiva per garantire in ogni momento la comunicazione nella rete P2P di KEEx.

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

5.6. Rete peer-to-peer di KEEEx

I peer descritti precedentemente compongono la soluzione KEEEx. In generale i PC della rete gestiti dagli utenti possono avere installato il PKM peer e, eventualmente, il Super peer se tali utenti hanno mansioni di amministratori della rete P2P di KEEEx, mentre macchine sempre attive (server) avranno installati il Normalization peer e il Source peer.

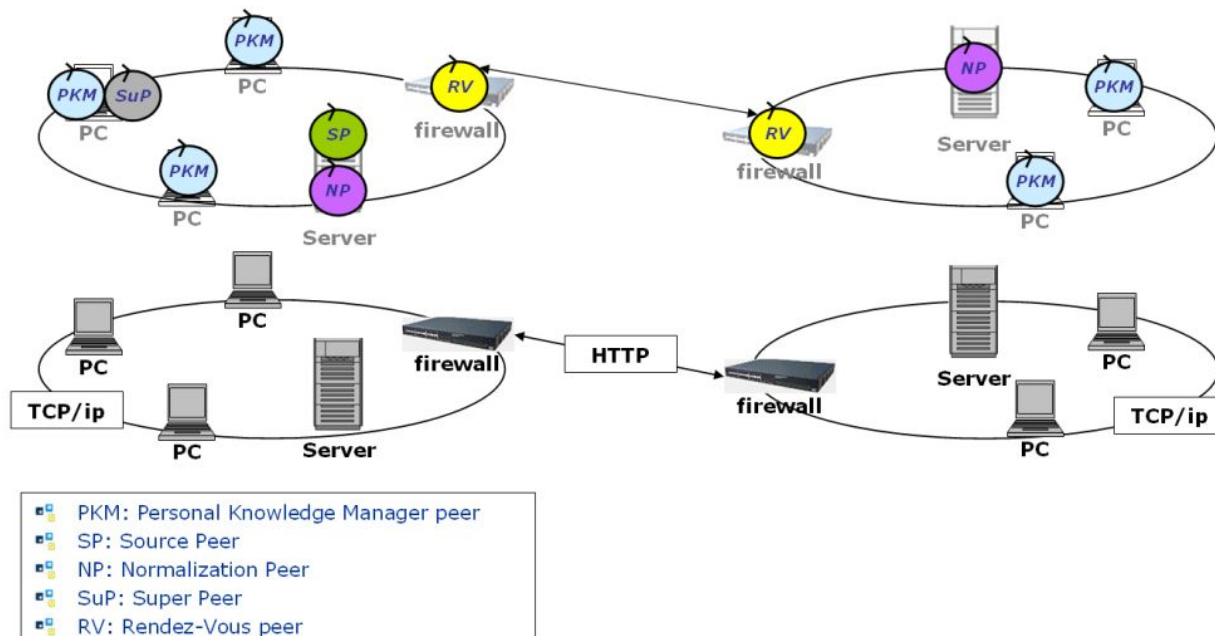


Figura 3: Rete P2P di KEEEx

Nella figura è mostrato un esempio di come possono essere installati in due reti distinte i vari peer (Figura 3). Da notare che i Rendez-Vous peer, che nella figura per semplicità sono posizionati sui firewall, in realtà vengono installati su macchine accessibili via HTTP o TCP a ogni altro computer su cui sono installati gli altri tipi di peer. Grazie ai Rendez-Vous peer, un PKM peer di una LAN può interagire con un PKM peer di un'altra LAN in modo trasparente. Inoltre all'interno delle due LAN il rispettivo Rendez-Vous peer garantisce la comunicazione nel caso in cui un comune indirizzo di multicast non possa essere utilizzato (vedi 6.8). Da notare che la comunicazione via Rendez-Vous peer è completamente trasparente per l'utente.

6. Tecnologia di base di KEEEx

6.1. Comunicazione peer-to-peer

L'infrastruttura di comunicazione P2P nella soluzione KEEEx si basa su un insieme di protocolli, definiti dal progetto JXTA, progetto open source (www.jxta.org) sponsorizzato da Sun Microsystems. Tali protocolli definiscono le funzionalità di base necessarie ad una rete P2P:

- indipendenza dalla piattaforma: essendo JXTA basato su protocolli e non su API può essere implementato in differenti linguaggi, sistemi operativi, etc;
- focalizzazione sulle funzionalità P2P: JXTA non fornisce nessuna soluzione a problemi/applicazioni specifici, ma è interamente orientata alla tecnologia di base per il P2P.

L'implementazione di tali protocolli introduce cinque livelli di astrazione della rete:

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

- *Peer ID*: un modello di indirizzamento dei peer basato sulla generazione di ID (128-bit random UUIDs). L'id permette di identificare univocamente un peer in modo indipendente dall'indirizzo fisico.
- *Peer Group*: permette ai peer di organizzarsi in domini virtuali.
- *Advertisement*: permette ad ogni peer di pubblicare nella rete le risorse disponibili.
- *Pipe*: canali virtuali di comunicazione che permettono alle applicazioni di comunicare tra loro.
- *Resolver*: meccanismo che permette di risolvere le operazioni di routing dei messaggi dei peer basandosi sui Peer ID.

Per una descrizione dettagliata della tecnologia JXTA rimandiamo alla documentazione di JXTA stesso (www.jxta.org).

6.1.1. Utilizzo di JXTA in KEEEx

Tutti gli elementi della soluzione KEEEx a livello P2P sono stati sviluppati su un'implementazione dei protocolli di JXTA fatta con il linguaggio JAVA.

KEEEx è un sistema peer-to-peer puro, non prevede cioè un indice centralizzato dei peer. Ogni peer è identificato univocamente da un *Peer ID*. I peer scoprono gli altri peer, le comunità e le zone (la cui implementazione è basata sui *Peer Group*), tramite il meccanismo di discovery basato sulla pubblicazione nella rete di *Advertisement* da parte dei vari peer e utilizzando le *Pipe*.

Esiste un advertisement per ogni peer, uno per ogni comunità e uno per ogni zona. Tale advertisement è rappresentato da un documento XML che ogni peer pubblica in locale. Nell'advertisement è descritta la modalità di interazione con la risorsa rappresentata. In particolare:

- *Advertisement del Peer*: pubblicato da ogni peer, descrive la tipologia del peer (PKM peer, Source peer, Normalization peer, Super peer e Rendez-Vous peer), i servizi disponibili (diversi a seconda della tipologia di peer) e come utilizzare tali servizi.
- *Advertisement della comunità o della zona*: pubblicato dal peer che crea la comunità o la zona, ne descrive la tipologia, la modalità di adesione e la modalità di interazione con la comunità o la zona.

I vari elementi della soluzione KEEEx (peer, comunità e zone) mettono a disposizione un insieme di servizi ognuno dei quali possiede una pipe di comunicazione. Tali servizi permettono l'accesso alle funzionalità descritte nelle sezioni precedenti: discovery, ricerca, download, chat, adesione ad una comunità, ecc.

Il servizio più importante a livello P2P è quello del discovery, perché permette di recuperare gli advertisement presenti nella rete che descrivono gli altri servizi. Quando un peer vuole fare discovery manda un messaggio su una pipe comune a tutti i peer e tutti i peer attivi rispondono con i propri advertisement e con gli advertisement delle comunità che hanno memorizzato in locale (perché da loro pubblicati o memorizzati da una operazione di discovery). A questo punto il peer può utilizzare i servizi descritti negli advertisement trovati comunicando direttamente con i peer, le comunità o le zone.

Nell'implementazione attuale, per il servizio di discovery viene utilizzato un indirizzo di multicast comune. Se nella rete in cui viene istanziata l'applicazione la multicast è disabilitata è possibile garantire la comunicazione grazie al Rendez-Vous peer (vedi 5.4.5 e 6.8).

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

6.2. Contesto

Un contesto nasce con l'esplicito scopo di organizzare e classificare un certo insieme di informazioni e viene realizzato come una gerarchia di concetti (albero di concetti) ciascuno dei quali ha associata un'etichetta in linguaggio naturale. I concetti vanno a formare una gerarchia perché sono connessi da un certo insieme di archi che ne individuano le relazioni. Dal punto di vista implementativo in KEEEx i contesti vengono rappresentati utilizzando un linguaggio di specifica denominato *CTXML* (ConTeXt Markup Language) basato su XML e XML-Schema. Ogni contesto si compone di un header e di un contenuto. L'header contiene informazioni generali sul contesto, come ad esempio chi ne è il proprietario, una breve descrizione del suo contenuto, informazioni sulle modalità di accesso, etc.. Il contenuto è invece la rappresentazione della gerarchia dei concetti vera e propria: la scelta di base nella definizione del linguaggio di specifica è stata quella di utilizzare i costrutti di base di XML-Schema (quali il *ComplexType* e il meccanismo di *Extension* che permette di applicare il principio dell'ereditarietà tra i tipi definiti in file XML) per rappresentare i concetti (*ComplexType*) e le loro relazioni gerarchiche (*Extension*).

Un concetto molto importante relativo ad un contesto è quello di *categoria*. Una categoria individua in modo univoco e completo un nodo all'interno di un contesto; dal punto di vista pratico rappresenta l'insieme dei nodi che collegano (in un unico percorso) il concetto in questione con la radice del contesto. La categoria contiene anche tutte le informazioni linguistiche relative al nodo e a tutti i nodi che lo collegano alla radice. Anche per la categoria è stata scelta una rappresentazione in formato XML, che raccoglie i dati descritti.

6.3. Plug-in

I plug-in di Microsoft Outlook e Microsoft Internet Explorer sono componenti sviluppate con il linguaggio Visual Basic 6.0 utilizzando librerie standard dell'SDK di Visual Studio 6.0. Per il plug-in di Microsoft Outlook è stata anche utilizzata la libreria Redemption (<http://www.dimastr.com/redemption/>) per riuscire ad estrarre le mail da un'applicazione esterna a Outlook stesso (PKM peer).

6.4. Sicurezza

La sicurezza, intesa come gestione dell'accesso degli utenti ai documenti condivisi tramite le applicazioni KEEEx può essere integrata con quella di sistema a vari livelli. Ogni utente crea delle ACL (Access Control List) associate ad ogni documento che decide di rendere disponibile tramite il sistema KEEEx.

In generale la sicurezza in KEEEx, è basata sul meccanismo di creazione di coppie chiave pubblica - chiave privata utilizzando algoritmi di crittazione RSA a 1024 bit e generando per ogni peer un certificato conforme allo standard X509. Il sistema determina l'identità di chi sta eseguendo l'applicativo basandosi sul sistema di log-on del sistema operativo. Oltre all'identità dell'utente, il certificato sarà utilizzato per verificare la provenienza e l'integrità dei messaggi scambiati tra i peer. Infine, a richiesta dell'utente, è possibile criptare le informazioni che vengono trasferite tra i peer utilizzando per il download funzioni di cifratura a chiavi asimmetriche (algoritmo DES a 1024 bit).

In un primo livello di sicurezza in KEEEx la lista degli utenti è dinamica, nel senso che ogni presenza di un utente di KEEEx sulla rete P2P viene registrata dagli altri tramite il meccanismo discovery (vedi 6.1.1): ogni peer si crea una propria lista di utenti presenti e attivi sulla rete P2P. Non esiste quindi una lista statica degli utenti KEEEx sulla rete.

Nel secondo livello, la sicurezza di KEEEx si appoggia al sistema di sicurezza dell'organizzazione (basata ad esempio su Active Directory di Microsoft o un Server LDAP). Il sistema KEEEx accede

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

ad un server di autenticazione per ottenere la lista completa degli utenti e dei gruppi definiti nell'organizzazione per il sistema KEEEx. E' necessario quindi avere per ogni installazione di KEEEx l'accesso a questo server d'autenticazione.

6.5. Ricerca semantica

La ricerca semantica consiste nella capacità di individuare relazioni di tipo semantico tra nodi di contesti diversi. L'algoritmo di *matching semantico* (vedi 6.5.2), partendo dal significato di due nodi in due contesti (vedi 6.2) cerca di individuare se i significati di questi nodi hanno una qualche relazione di tipo semantico (vedi 6.5.2).

Il significato dei singoli nodi di un contesto viene definito nella fase di *normalizzazione* (vedi 6.5.1). Tale fase non è parte della ricerca ma deve essere fatta al momento della creazione di un contesto ed è necessaria per definire l'interpretazione semantica delle etichette che compongono il contesto, ovvero rendere esplicito il significato delle etichette associate ai nodi sulla base del linguaggio in cui esse sono scritte.

6.5.1. Normalizzazione

La normalizzazione di un contesto ha lo scopo di definire l'interpretazione semantica delle etichette che lo compongono. Ogni contesto, tramite normalizzazione, viene arricchito di un certo insieme di sensi scelti da *WordNet* che individuano appunto il significato delle etichette. WordNet è un database lessicale utilizzato in tutto il mondo e disponibile per molte lingue diverse. La praticità di WordNet è data dal fatto che esso contiene allo stesso tempo sia le informazioni sul significato delle parole che le informazioni sulle relazioni tra esse esistenti; inoltre le versioni per le diverse lingue sono tra loro allineate e permettono quindi di lavorare contemporaneamente con lingue diverse utilizzando gli stessi strumenti e le stesse modalità di accesso. WordNet è una risorsa linguistica definita di tipo "generale" nel senso che contiene termini appartenenti ai domini più svariati senza essere specifica per nessuno di essi. Potrebbe quindi rendersi necessario effettuare quella che si chiama customizzazione di WordNet, ovvero l'operazione con cui esso viene arricchito con termini specifici di un particolare dominio.

Nell'attuale implementazione di KEEEx sono possibili due diverse modalità di accesso a WordNet, con caratteristiche diverse in termini di prestazioni e occupazione di memoria.

1. WordNet è rappresentato da un insieme di file di testo che vengono caricati in memoria sotto forma di hash. In questo caso il peer di normalizzazione ha una maggiore occupazione di memoria ma tempi di normalizzazione nettamente inferiori, in quanto l'accesso ai dati in memoria è molto più veloce.
2. WordNet è rappresentato da un insieme di tabelle relazioni in un database MySQL. In fase di normalizzazione l'accesso a tale database viene realizzato via JDBC, ovvero quelle API che permettono di accedere ad un qualsiasi tipo di sorgente dati in formato tabellare da un programma Java. Questo tipo di soluzione comporta una minore occupazione di memoria perché i dati rimangono nel database e non in memoria, ma ha tempi di normalizzazione più lunghi dovuti all'accesso al database.

È possibile utilizzare una o l'altra versione provvedendo a seconda dei casi all'installazione del peer di normalizzazione opportuno. La normalizzazione si divide in due fasi: nella prima (analisi linguistica) vengono analizzate le singole etichette dei nodi del contesto, indipendentemente dalla loro posizione al suo interno. Nella seconda fase (*sense refinement*) i risultati così ottenuti vengono raffinati in base alla relazione tra le etichette appartenenti a concetti diversi. In particolare il raffinamento avviene confrontando la rappresentazione della conoscenza "personale" costituita dal contesto con quella "del mondo" rappresentata da WordNet.

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

L'analisi linguistica di un contesto richiede l'uso di un certo insieme di tecnologie nate e sviluppate nell'ambito delle problematiche di processo del linguaggio naturale e si compone di alcune sottofasi:

- Tokenizzazione. Utilizzando un lessico definito a seconda della lingua del contesto, il modulo tokenizzazione estrae dalle etichette i singoli tokens, ovvero le parole che le compongono.
- Analisi morfologica. Utilizzando un radiciario definito in base alla lingua del contesto, l'analisi morfologica individua per ogni token le possibili categorie grammaticali (ovvero se si tratta di un nome, piuttosto che un verbo, aggettivo, etc.).
- Lemmatizzazione. Utilizzando il risultato dell'analisi morfologica dei tokens, il modulo di lemmatizzazione individua il lemma corrispondente ad ognuno di essi, ovvero la forma normale della parola (ad esempio nel caso di parole al plurale, ne individua il singolare; nel caso di verbi coniugati ne individua la forma infinita, etc.).
- PoS-Tagging. Utilizzando un lessico specifico per la lingua del contesto, il Pos tagger sceglie per ogni token la corretta categoria grammaticale (detta **Part of Speech** o **PoS**) tra quelle proposte dall'analizzatore morfologico.
- Analisi semantica (interrogazione di WordNet). Per ogni singolo lemma individuato nelle fasi precedenti viene verificato se esso è presente in WordNet, interrogando il database in modo puntuale con il lemma e la categoria grammaticale individuate. Se il lemma viene trovato, i sensi presenti in WordNet relativamente ad esso vengono associati al concetto. Questa fase può essere paragonata alla ricerca del significato di una parola su di un vocabolario. Per ogni significato individuato vengono anche raccolte, sempre tramite interrogazioni di WordNet, l'insieme delle relazioni strutturali tra sensi (olonimia, antonimia, iperonimia) esistenti, relazioni che verranno utilizzate in fase di *sense refinement*.
- Riconoscimento di multiwords. Per ogni concetto viene verificato se esistono delle multiwords, ovvero più parole in successione per le quali esiste un significato preciso, diverso da quello delle parole isolate (ad esempio collega di lavoro, mano d'opera, forza lavoro, carta di credito, Presidente della Repubblica, ecc.). I sensi corrispondenti alle multiwords vengono aggiunti con priorità alta alla lista dei sensi creata precedentemente.

Tutti i componenti descritti sono integrati all'interno di un modulo denominato JTextPro, che viene distribuito in forma di dll e al quale ci si interfaccia utilizzando la tecnologia JNI (Java Native Interface) dall'interno del modulo matching semantico di KEEx. JTextPro riceve come input un contesto in un opportuno formato e restituisce lo stesso contesto con le informazioni linguistiche individuate sulla base della lingua definita per il contesto stesso.

Successivamente i dati così ottenuti vengono opportunamente inseriti nel contesto e sottoposti alla fase successiva di *sense refinement*, che ha lo scopo di determinare come il significato dell'etichetta di un nodo cambia in funzione della sua posizione all'interno della gerarchia. Alla base di questo processo vi è l'analisi delle relazioni strutturali tra sensi (olonimia, antonimia, iperonimia) presenti in WordNet. La logica di fondo consiste nell'eliminare i sensi presenti nel contesto che non rispecchiano tali relazioni strutturali, a favore di quelli che richiamano la struttura presente in WordNet.

6.5.2. Matching semantico

Dato un contesto normalizzato, la fase di *semantic matching* è in grado di confrontare due categorie e di verificare se tra essi esiste una relazione semantica {"meno generale di", "più generale di", "equivalente a"}, o di generica compatibilità nel caso in cui non sia possibile individuare l'esatta relazione semantica, ma esistano comunque forti correlazioni linguistiche tra le parole delle due categorie.

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

Data una categoria viene costruito il significato semantico, costituito dall'insieme dei sensi associati ai nodi della categoria stessa. Tale significato semantico viene utilizzato per riempire una struttura dati, denominata "matrice di matching", che riporta l'insieme dei sensi così individuato in corrispondenza di righe e colonne. La matrice viene riempita individuando le relazioni esistenti tra questi insiemi di sensi sulla base di quanto esistente tra le relazioni strutturali di WordNet. La matrice così ottenuta diviene l'input della successiva fase di calcolo in cui essa viene "trasformata" in un problema di soddisfacibilità (ovvero di verifica se una certa formula di logica del primo ordine è deducibile da un certo insieme di premesse). Le premesse di tale problema sono date dalle relazioni individuate in matrice, mentre le conseguenze sono date di volta in volta dalle relazioni semantiche tra i concetti che si stanno verificando. Quando non è possibile individuare una relazione semantica i dati della matrice vengono comunque utilizzati per fornire una percentuale di compatibilità (in questo caso puramente linguistica) tra i due concetti, che indica semplicemente quanto i due concetti contengono parole tra loro simili o in relazione dal punto di vista linguistico.

Quello che è stato considerato nella realizzazione e implementazione dell'algoritmo di matching tra contesti è il fatto che lo scopo non è quello di cercare delle relazioni tra strutture gerarchiche con arbitrarie etichette ma bensì trovare relazioni tra strutture le cui etichette sono espressioni appartenenti al linguaggio (naturale) delle persone che formano la comunità, in cui quindi ogni parola ha un proprio significato in quella comunità e un certo insieme di relazioni con altri termini. Nell'attuale implementazione, il *semantic matching* è realizzato interamente in Java con il supporto, per il problema di soddisfacibilità di una libreria (Jsat, anch'essa scritta in Java) e di un modulo (Orbital, sempre Java) per la creazione delle forme normali congiuntive che il modulo di Sat prende come input.

6.6. Ricerca lessicale

La ricerca lessicale è basata sull'indicizzazione di documenti, che permette al peer una veloce ricerca full-text basata su parole chiave in documenti di vari formati. La creazione di indici permette di evitare di aprire ed effettuare la ricerca in ogni documento al momento della richiesta, operazione che risulterebbe troppo dispendiosa in termini di tempo. La soluzione più facile è fare la maggior parte del lavoro in anticipo, ossia estrarre le informazioni sui termini di ogni documento e memorizzarle in modo tale che siano facilmente recuperabili. A fronte di una query il motore di ricerca non analizzerà tutti i documenti, ma interrogherà l'indice creato.

KEEx prevede l'indicizzazione del testo e del nome del documento (nome del file). Saranno quindi possibili ricerche di documenti basate su una o più parole chiave da ricercare all'interno del documento o nel nome del file. È inoltre possibile specificare gli operatori logici AND e OR per la ricerca con più parole chiave.

Il sistema KEEx può utilizzare diversi indicizzatori a seconda delle esigenze.

6.6.1. dtSearch®

È uno dei più potenti motori di ricerca full-text commerciale (www.dtsearch.com). Oltre ad essere molto veloce nella ricerca su grandi quantità di documenti, permette l'indicizzazione efficace di molti formati di documenti:

- Adobe Acrobat (PDF) tutte le versioni fino alla versione 6
- HTML
- Microsoft Excel (fino alla versione Excel 2003)
- Microsoft Outlook Express 5 and 6 (*.dbx) message stores
- Microsoft PowerPoint 97, PowerPoint 2000, PowerPoint XP, PowerPoint 2003

Distributed Thinking S.p.A
Partita IVA 01836390227.

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

- Microsoft Rich Text Format
- Microsoft Word per DOS
- Microsoft Word per Windows (fino alla versione Word 2003)
- Ansi Text
- CSV (Comma-separated values)
- MSG files (e-mail salvate da Outlook)
- XML
- ZIP
- EBCDIC
- Ami Pro
- EML files (e-mail salvate da Outlook Express)
- Eudora MBX message files
- MBOX email archives
- MHT archives (HTML archive salvati da Internet Explorer)
- MIME messages
- Microsoft Works
- Multimate Advantage II
- Multimate version 4
- Treepad HJT files
- Unicode
- WordPerfect (tutte le versioni dalla 5.0 fino a WordPerfect 2002)
- WordStar versions 4, 5, 6
- WordStar 2000
- Write
- XBase (compreso FoxPro, dBase, e fli altri XBase - formati compatibili)

6.6.2. Jakarta Lucene

È un veloce motore di ricerca full-text scritto interamente in Java. Jakarta Lucene è un progetto open source di Jakarta (<http://jakarta.apache.org/lucene/docs/index.html>).

L'estrazione del testo da documenti con formati proprietari è possibile grazie all'utilizzo di filtri.

- Formati Microsoft: filtri basati sulle API POI, del progetto open source Jakarta POI (<http://jakarta.apache.org/poi>), che permettono di manipolare vari formati di file basati sul formato OLE 2 Compound Document di Microsoft.
- Formato PDF: filtro basato sulle librerie java PDFBox che permettono di accedere alle componenti di un documento PDF (<http://www.pdfbox.org>).
- Formato RTF: filtro basato sulle librerie della distribuzione standard di java (jdk 1.4) package javax.swing.text.rtf.

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

I formati dei file che KEEx indicizza utilizzando Jakarta Lucene sono: txt, html, xml, doc (Microsoft Word), xls (Microsoft Excel), ppt (Microsoft Power Point) e pdf.

6.7. Ricerca concettuale

La ricerca concettuale consiste nel ricercare parole chiave nelle stringhe che descrivono i concetti dei contesti (6.2). Le parole chiave vengono ricercate come sotto-stringhe non *case sensitive*. È possibile specificare gli operatori logici AND e OR per la ricerca con più parole chiave. La visita del contesto (albero) viene fatta in modalità Bread-First, ricercando quindi per livelli di profondità del contesto.

6.8. Topologia rete di KEEx

I protocolli di JXTA permettono a KEEx di creare una rete virtuale sopra l'infrastruttura fisica di rete dove ogni peer può interagire con altri peer tramite scambio di messaggi in modo indipendente dalla posizione fisica nella rete. I messaggi vengono instradati in modo trasparente, potenzialmente attraversando firewall o NAT, usando differenti protocolli di trasporto (TCP/IP o HTTP) per raggiungere i peer destinatari dei messaggi.

Dal punto di vista della rete fisica, per poter ovviare al problema che due peer non sono collegati direttamente perché protetti da firewall o NAT, è possibile utilizzare il peer Rendez-Vous per garantire la comunicazione. In questo caso il Rendez-Vous peer, via protocollo TCP/IP o HTTP su due porte stabilite, deve essere raggiungibile da tutti i peer e deve a sua volta poter comunicare con tutti i peer delle reti che sta mettendo in comunicazione. Nel caso di presenza di firewall è necessario che questo permetta almeno la comunicazione HTTP tra le macchine interne al firewall e la macchina su cui è installato il Rendez-Vous peer. Inoltre le macchine fuori dal firewall devono poter raggiungere il Rendez-Vous peer via HTTP o TCP/IP. Dove è possibile è consigliabile utilizzare il protocollo TCP/IP perché le performance sono migliori.

6.8.1. Utilizzo del Rendez-Vous peer

Non è sempre necessario utilizzare il Rendez-Vous peer nella rete P2P di KEEx. In una situazione in cui in una rete ha un comune indirizzo di multicast abilitato e i computer della rete sono direttamente connessi tra loro, il Rendez-vous peer non è necessario (Figura 4).

Per utilizzare un Rendez-Vous peer è necessario che ogni peer sia configurato opportunamente con indirizzo IP e porta su cui viene fornito il servizio.

Il Rendez-Vous peer viene utilizzato per propagare i messaggi tra i peer quando non esiste un indirizzo di multicast comune perché non abilitato o perché le macchine sono su reti LAN distinte e non possono comunicare direttamente tra loro via protocollo TCP/IP. In quest'ultimo caso il Rendez-Vous peer fornisce la funzionalità di instradamento dei messaggi nella rete P2P di KEEx.

*Distributed Thinking S.p.A
Partita IVA 01836390227.*

Sede Legale: Via Fontana, 6 - 38068, Rovereto (TN)

Sede Commerciale: Via Traiano, 7 - 20149, Milano tel. 02-45499020 fax. 02-45499943

Sede Operativa: Via F. Zeni 8 - 38068, Rovereto (TN) tel. 0464-443232 fax. 0464-443233

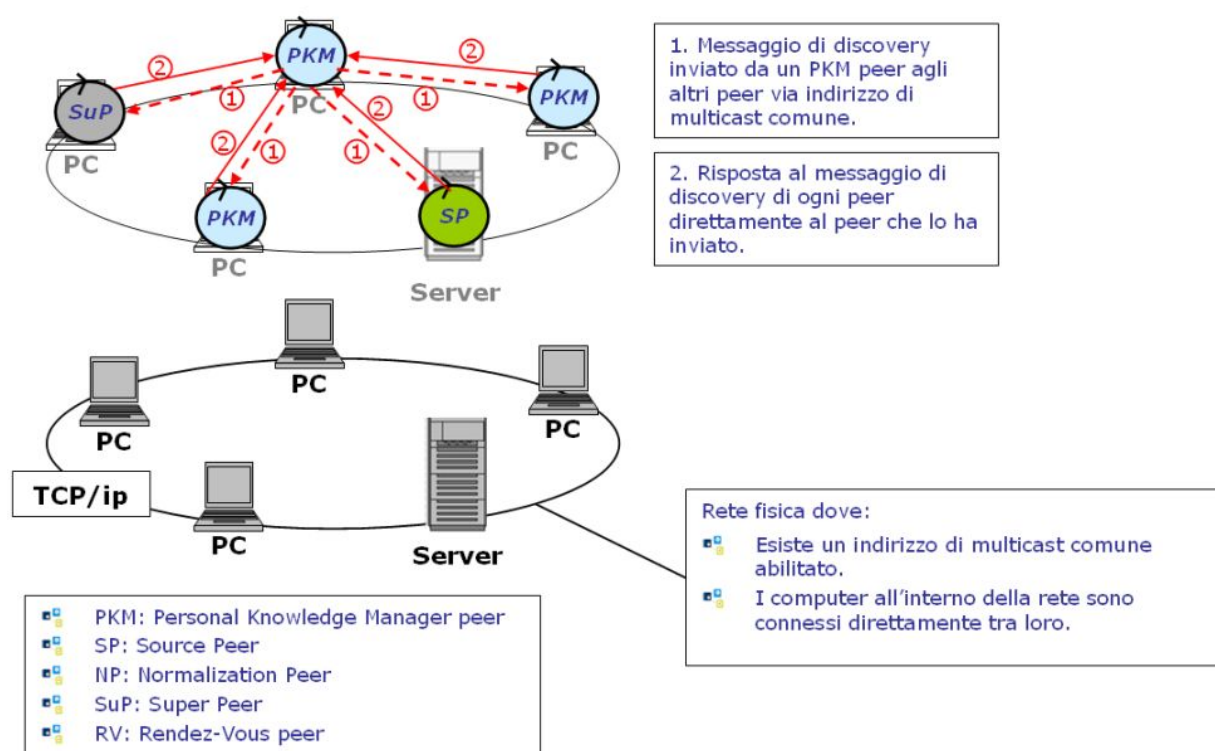


Figura 4: Discovery su rete fisica con indirizzo multicast comune abilitato e computer della rete connessi direttamente tra loro.

Nel primo caso (Figura 5), cioè quando non esiste un comune indirizzo di multicast abilitato ma le macchine possono comunicare direttamente via TCP/IP il Rendez-Vous peer viene utilizzato per la funzionalità di discovery. Quando un peer vuole fare discovery manda il messaggio di discovery al Rendez-Vous peer e questo lo propaga a tutti i peer registrati. I peer registrati risponderanno direttamente al peer che ha inviato il messaggio di discovery.

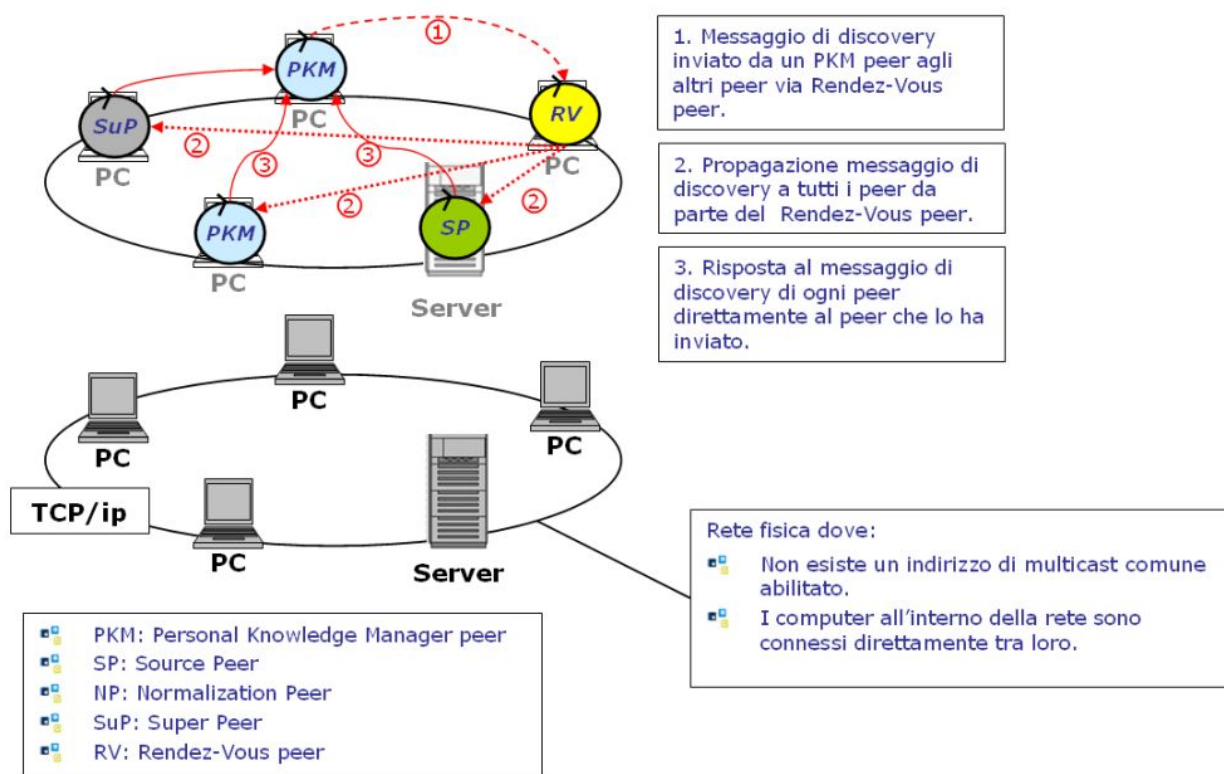


Figura 5: Discovery su rete fisica senza indirizzo multicast comune abilitato e computer della rete connessi direttamente tra loro.

Nel secondo caso (Figura 6), quando oltre a non avere la multicast abilitata le macchine non comunicano direttamente via TCP/IP, il Rendez-Vous propaga e instrada i messaggi opportunamente. Quando un peer deve mandare un qualsiasi messaggio a uno o più peer lo invia sempre al Rendez-Vous peer che provvede a propagarlo a tutti i peer nel caso del discovery e a instradarlo ai peer corretti nel caso di un messaggio inviato direttamente a uno o più peer. Tutti i messaggi, anche eventuali messaggi di risposta, passano per il Rendez-Vous peer.

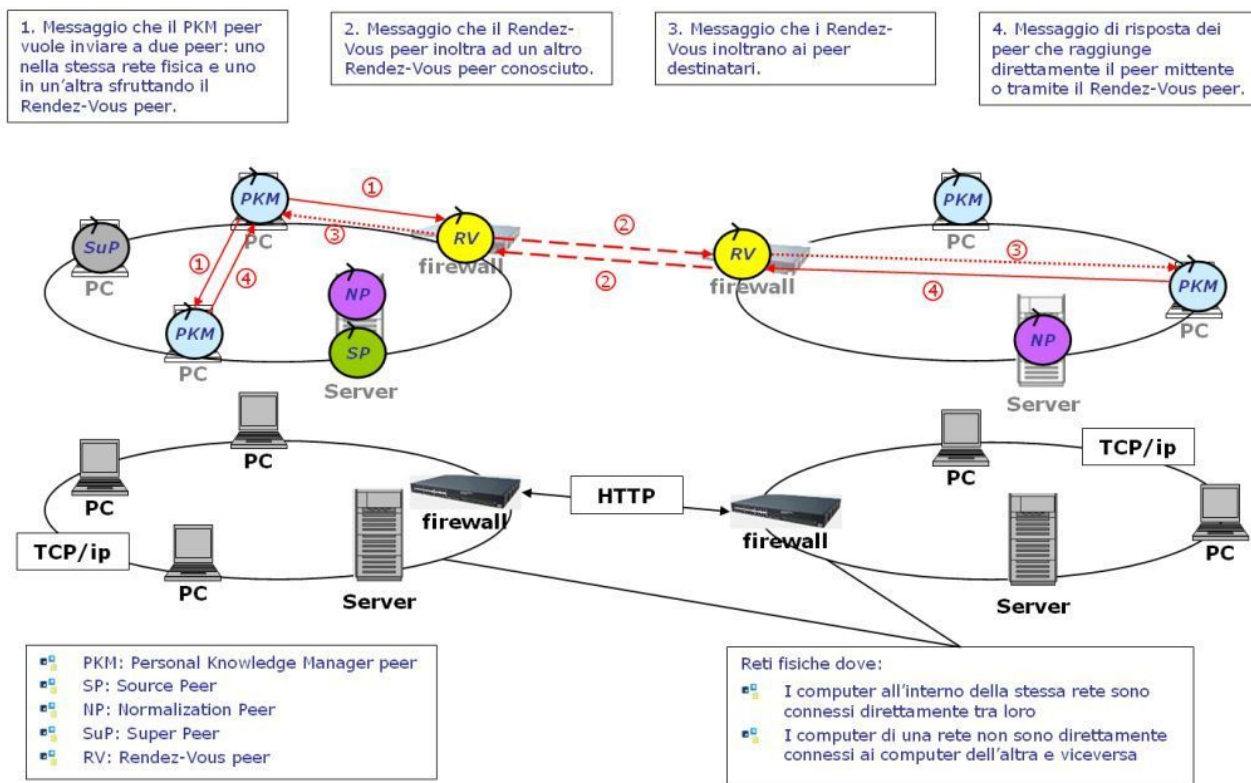


Figura 6: Scambio messaggi tra peer su reti fisiche distinte.

I Rendez-Vous peer possono essere incrociati tra loro per poter garantire l'instradamento dei messaggi in reti distinte ed eventualmente protette da firewall. Quindi un Rendez-Vous peer instrada oltre che agli altri peer anche ad altri Rendez-Vous peer registrati per connettere più reti fisiche distinte.

Alcuni esempi di mini-progetti

Massimiliano Ciaramita, Aldo Gangemi, Domenico Pisanelli

Laboratorio di Ontologia Applicata, CNR-ISTC, Roma

6 Dicembre 2005

Un “mini-progetto” è un progetto finalizzato a una funzionalità molto specifica, da completare a breve termine utilizzando tecnologie che fanno già parte dell'architettura di *information management* generale, oppure che saranno integrate successivamente, o anche che servano da soluzioni sub-ottimali per le finalità contingenti del mini-progetto. Hanno un tempo di sviluppo breve e una *task force* specializzata (che può contenere anche partner esterni, via convenzioni o contratti di collaborazione), può ottenere risultati interessanti in tempi fra 3 e 6 mesi, senza attendere il completamento del motore o di IWS.

I seguenti sono alcuni esempi di mini-progetti che riguardano aspetti di ricerca e analisi semantica su testi non strutturati e reingegnerizzazione di basi di dati e altri formati semi-strutturati. Nella prospettiva dell’IntraWeb Semantico del CNR, tutte le funzionalità descritte saranno comunque sussunte dal sistema finale. Altri possibili mini-progetti, come l’analisi di *log*, la costruzione di *meta-query*, il potenziamento del motore di ricerca terminologico con componenti morfologici e multi-lingua, etc. sono o già descritti nel documento sull’architettura del motore di ricerca, o possono essere specificati in versioni successive.

1. Modulo per il supporto alla decisione sul contenuto di documenti digitalizzati.

Questo modulo svolge una funzione di classificazione automatica dei documenti rispetto a una serie di categorie pre-definite. Per esempio, il compito potrebbe essere il supporto al *decision making* riguardo l'autorizzazione alla finalizzazione di processi, come l'acquisto di hardware, o approvazione di domande etc.¹ Le categorie che definiscono le possibili risposte vanno dal più semplice caso binario (sì/no, accettato/scartato) a categorizzazioni più complesse, cioè definite da un numero di categorie maggiore (accettato/scartato/dubbio, nel caso ternario etc.).

Il supporto alla decisione avviene tramite l'analisi del testo attraverso un classificatore automatico che viene addestrato a replicare le decisioni già prese da un esperto.

Il classificatore è una funzione che dato un documento restituisce una delle risposte possibili. La funzione si basa su un vettore di parametri che specificano proprietà importanti del documento, su cui è basata la categorizzazione, a cui è associato un peso individuale. Il vettore su cui si basa la funzione viene “imparato” empiricamente su una base di dati, per esempio le decisioni precedentemente prese dal responsabile alla decisione stessa. In questo modo il classificatore cerca esplicitamente di riprodurre le decisioni che sono più coerenti con la base di dati sistente. Per valutare e migliorare l'accuratezza del classificatore, il risultato di alcuni campioni di decisioni prese dal classificatore viene confrontato con le decisioni corrette in modo che: 1) viene stimata la percentuale di correttezza, o inversamente di errore, del sistema, e 2) i

¹ Ovviamente, in quanto *supporto* alla decisione, va prevista una verifica manuale di ogni decisione rilevante.

parametri del sistema vengono riaggiustati tenendo conto dell'informazione addizionale. Il sistema è cioè adattivo.

Questo sistema si basa su una serie di passi semplici, la digitalizzazione (spesso già disponibile all'origine) dei documenti, l'estrazione delle caratteristiche del documento (estrazione delle *features*) e il *training* di un classificatore stato dell'arte.

Tempo di realizzazione: 2 mesi.

2. Mappatura di documenti non strutturati ad un modello prototipico

Molti documenti che contengono informazione simile sono strutturati in modo diverso. Per esempio curricula vitae diversi, soprattutto in un ambiente intranet, contengono tipicamente lo stesso tipo di informazione: informazioni personali, titoli di studio, pubblicazioni, collaborazioni, esperienze etc. L'obiettivo di questo mini-progetto è di mappare automaticamente documenti diversi in una rappresentazione condivisa attraverso l'applicazione di tecniche di *information extraction*.

Il task può essere svolto in un ciclo breve di sviluppo se focalizzato su alcuni tipi specifici di documentazione abbastanza uniformemente strutturata, come le pubblicazioni, le commesse, etc. Prima di tutto, si identifica la tipologia di documento e si definisce la struttura in cui devono essere mappati. In seguito, si costruiscono degli esempi di curricula mappati manualmente, campionati dal database stesso. Infine, su questi dati si fa il "training" di un annotatore automatico statistico (*tagger*). Sugli stessi dati, usando un approccio di tipo *cross-validation*, si valuta la performance del sistema. Se la percentuale di errore del sistema è sotto la soglia desiderata, si campiono alcuni esempi di curricula su cui il sistema ha fallito, si annotano a mano e si introducono nei dati di *training*. Questa procedura viene ripetuta finché l'accuratezza necessaria non sia stata raggiunta.²

Tempo di realizzazione: 3 mesi.

3. Basi di conoscenza sensibili

Questo mini-progetto è finalizzato alla reingegnerizzazione di alcune basi di dati "sensibili" del CNR. In particolare: il database semi-strutturato delle commesse, il database strutturato dell'anagrafica istituti, ricercatori, aree tematiche e il database semi-strutturato dei regolamenti. L'obiettivo è di ottenere conoscenza di dettaglio sulle commesse, integrarla con l'anagrafico e i regolamenti.

I task di questo mini-progetto includono l'interpretazione semantica degli schemi dei database sorgente, la loro migrazione in ontologie, l'esportazione dei dati in formati gestibili (ex. RDF) con tecniche di ingegneria della conoscenza, la formulazione di estensioni delle ontologie per classificare i dati in maniera integrata o per gestire una tassonomia di "viste".

Nella mancanza temporanea di un ambiente integrato di gestione di ontologie, motore di ricerca ed estrazione d'informazione, soluzioni sub-ottimali possono essere trovate con l'impiego di esperti, strumenti esistenti open-source (ex. Sesame) e/o commerciali, etc. In particolare, la base di conoscenza delle commesse richiederebbe

² Per esempi, vedi il documento sull'architettura del motore di ricerca.

comunque un'analisi semi-automatica dell'informazione sorgente, che ci si aspetta estremamente variegata quanto a contenuti, modalità di presentazione, accuratezza, etc. Un altro lavoro che richiede competenze intellettuali non automatizzabili è l'integrazione (finale) degli schemi, per esempio per sussumere gli schemi anagrafici e quelli delle commesse in rapporto a uno schema più generale, come quello suggerito: Processo-Tecnica-Area.

Tempo di realizzazione: 4 mesi.

4. Modellazione e linee-guida per contratti o regolamenti specifici

Un'altra sorgente di informazione sensibile per il CNR è il database dei regolamenti. Nell'ambito del mini-progetto 3. è possibile reingegnerizzarlo e aggiungere metadati intorno ai regolamenti, nel tentativo di ottenere un'indicizzazione più accurata della semplice ricerca per stringa. Tuttavia, per la natura dell'informazione regolativa, sarebbe opportuno costruire una base di conoscenza che importi la conoscenza sensibile per l'attuazione e l'esecuzione di un regolamento.

Questo mini-progetto consiste nell'identificazione della conoscenza procedurale e normativa all'interno dei regolamenti e nella sua integrazione con eventuali flussi di lavoro esistenti nell'amministrazione CNR (vedi 5.). Nell'attesa del sistema integrato di IntraWeb Semantico, metodi e strumenti sub-ottimali possono essere adottati, fra cui l'annotazione manuale e semi-automatica dei testi regolativi e la predisposizione di *template* per il *draft* di regolamenti (cf. progetto NormeInRete). I *template* per l'esecuzione dei regolamenti possono anche beneficiare dei risultati del progetto EU Metokis, in cui l'ISTC-CNR è partner.

I risultati principali di questo mini-progetto sono quindi la definizione di linee-guida e di spiegazioni strutturate per la comprensione, l'esecuzione e la gestione dei contratti e dei regolamenti.

Tempo di realizzazione: 6 mesi (a seconda del numero di contratti e regolamenti).

5. Workflow manager per attività amministrative

Alcuni uffici del CNR hanno cominciato a esplicitare il *know-how* relativo alle procedure tipiche da loro svolte. In alcuni casi, mediante modelli UML (diagrammi di classe, attività, stato, sequenza). Il *know-how* è un patrimonio fondamentale per un'organizzazione complessa e va integrato con le conoscenze statiche (*know-that*) nella cosiddetta *corporate knowledge*. L'IntraWeb Semantico sarà la piattaforma ideale per tale integrazione. Preliminarmente tuttavia, è possibile cominciare a sperimentare sulla gestione delle procedure già esplicitate mediante *workflow manager* semantici.

Questo mini-progetto consiste nella reingegnerizzazione semantica di alcuni modelli procedurali finora esplicitati e in una fase di sperimentazione della gestione avanzata di quei flussi di lavoro insieme al personale degli uffici coinvolti in quelle procedure. L'obiettivo è di identificare il tipo di conoscenza e le viste che possono migliorare il lavoro all'interno degli uffici, evitando ridondanze, colli di bottiglia, etc.

I risultati principali del mini-progetto includono l'implementazione di un *workflow manager* open-source, la formalizzazione di alcuni modelli di flusso e l'analisi dei dati relativi alla fase di sperimentazione.

Tempo di realizzazione: 6 mesi.

6. CNR Wiki

Il patrimonio di conoscenze di un'organizzazione non si risolve solo nelle basi di dati ufficiali e nei modelli di flussi di lavoro, ma anche nella comunicazione fra gli agenti che agiscono per l'organizzazione; in questo caso: dirigenti, ricercatori, tecnici, amministrativi. La comunicazione avviene ormai in parte in forma digitale, e formati di comunicazione come i forum, le liste di discussione e i blog sono largamente usati. Tuttavia, per un'organizzazione come il CNR, la conoscenza sensibile derivante dalla comunicazione fra agenti è utilizzabile al meglio se viene strutturata "spontaneamente" da quegli agenti. Un formato in pieno sviluppo attualmente è il Wiki, un magazzino di appunti e informazioni che viene mantenuto, modificato e arricchito comunitariamente, senza speciali mediazioni (a parte forme di inibizione di utenti con intenti aggressivi). Sorprendentemente, questi magazzini raggiungono un'accuratezza notevole, specialmente quando il task della Wiki è l'accordo intersoggettivo su un termine, sulla descrizione di un fatto, su un processo, etc. L'esempio di Wikipedia: <http://www.wikipedia.org> è illuminante al riguardo.

Questo mini-progetto consiste nell'implementazione di una o più Wiki dedicate all'immagazzinamento di informazioni sensibili intorno alle procedure, ai fatti noti, alle soluzioni locali, etc. del CNR. L'obiettivo è quello di creare una base di conoscenza "dal basso", coinvolgendo gli agenti del CNR direttamente. Per esempio, l'accordo sull'interpretazione di un regolamento in termini di procedure locali potrà essere raggiunto per successivi aggiustamenti in una Wiki. Un altro obiettivo è l'integrazione delle Wiki con le basi di conoscenza e a medio-termine con l'IntraWeb Semantico, secondo il modello di "Wiki semantica" che sta emergendo in ambito di ricerca.

Tempo di realizzazione: 3 mesi.

7. Integrazione servizi (ex. Bilancio e Commesse o Anagrafica)

In un'organizzazione complessa è possibile oggi integrare i sistemi informativi esistenti o in sviluppo per ottenere servizi che sarebbero impossibili per le applicazioni locali. Per esempio, nel CNR, i servizi forniti dall'applicazione per la gestione del Bilancio possono essere integrati con i servizi della base di conoscenza anagrafica o della futura base di conoscenza delle commesse, ottenendo la possibilità di associare una *query* sul bilancio a una sulle commesse e comporla in funzione di una richiesta di servizio più complessa. Tale integrazione può essere fatta sui cosiddetti *Web Service*, che sono già stati introdotti in alcuni uffici del CNR.

Questo mini-progetto consiste nell'implementazione di componenti sperimentali per l'integrazione semantica di *Web Services* (per esempio la piattaforma IRS-III, sviluppata dal KMI, partner di ISTC-CNR), la descrizione di altre applicazioni esistenti e il loro interfacciamento sull'IntraWeb, e la costruzione di servizi integrati a partire da viste espresse formalmente.

Tempo di realizzazione: 6 mesi.

8. Sviluppo di risorse e strumenti linguistico/semantici

Quasi tutti i mini-progetti descritti richiedono lo sviluppo di risorse linguistico/semantiche. E' quindi opportuno accorpare le tecniche e gli strumenti per tale sviluppo all'interno di un mini-progetto specifico, consistente nella creazione di risorse linguistico-semantiche, per es. lessico-terminologiche, relative ad aree tematiche rilevanti all'interno del database del CNR.

La costruzione delle risorse si basa su tecniche stato dell'arte di trattamento automatico del linguaggio, attraverso l'analisi del database intranet con strumenti esistenti di parsing sintattico/semantico e clustering su base distribuzionale, come quelli sviluppati da ILC-CNR di Pisa, che ha una vasta esperienza in proposito, per esempio nel progetto PeKITA del Ministero della Ricerca, con il quale sarebbe opportuno attuare una collaborazione.

Tempo di realizzazione: 2 mesi.