



Deliverable D2.4

Definizione di tecniche semantiche per l'IR e la realizzazione di uno strumento Integrato per il Supporto su basi dati culturali semanticamente annotate e portali WEB

Responsabile:

Di Martino Beniamino

Afferenza

Seconda Università di Napoli

Autori

Prof. Rocco Aversa, Prof. Ida Caracciolo, Prof.
Beniamino Di Martino, Prof. Pasquale Femia,
Ing. Angelo Martone, Ing. Francesco Moscato,
Prof. Francesco Palmieri, Ing. Massimiliano Rak,
Ing. Gianmarco Romano, Ing. Salvatore
Venticinque, Prof. Rosanna Verde

Afferenza

Seconda Università di Napoli

PROGETTO LC3	Revisione n*	0	Del	----
--------------	--------------	---	-----	------

INDICE

- [TR2.4.1 Stato dell'arte ed analisi delle metodologie e delle tecniche per il Natural Language Processing](#)
- [TR2.4.2 Stato dell'arte ed analisi delle tecnologie e degli strumenti per il Natural Language Processing](#)
- [TR2.4.3 Stato dell'arte ed analisi dei modelli e delle metodologie per l'Information Retrieval](#)
- [TR2.4.4 Stato dell'arte ed analisi delle tecnologie e degli strumenti per la Information Retrieval](#)
- [TR2.4.5 Stato dell' arte ed analisi delle metodologie, dei linguaggi standard e delle tecniche per la Rappresentazione della Conoscenza ed Inferenza in ambito Ontologico](#)
- [TR2.4.6 Stato dell'arte ed analisi delle metodologie, dei linguaggi standard e degli engines per il querying semantico](#)
- [TR2.4.7 Stato dell'arte ed analisi delle metodologie e delle tecniche per il matching di ontologie](#)
- [TR2.4.8 Realizzazione di uno strumento prototipale per il Natural Language Processing](#)
- [TR2.4.9 Definizione di una architettura per il Semantic Information Retrieval](#)
- [TR2.4.10 Realizzazione di uno strumento prototipale per il Semantic Information Retrieval](#)
- [TR2.4.11 Definizione di una tecnica per il confronto \(matching\) di ontologie](#)
- [TR2.4.12 Realizzazione di uno strumento prototipale per il confronto \(matching\) di ontologie](#)

Scopo di questo task è la definizione di un insieme di metodologie e tecniche relative all'analisi di natura semantica, sia della query di ricerca, che della base documentale oggetto della ricerca, che della conoscenza disponibile riguardo al dominio di applicazione.

Nel seguito si descrivono nel dettaglio le attività eseguite ed i risultati conseguiti dalle Unità di ricerca della Seconda Università di Napoli (DII e JM).

Sono state esaminate le principali tecniche per l'analisi lessicale del testo, con particolare riferimento alle procedure di Stopwords Removal, Stemming, di estrazione di indici e di pesatura dei termini. Sono state inoltre esaminate le principali metodologie e tecniche per l'

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------

analisi sintattica e semantica, in particolare il Part of Speech (POS), il Parsing e la Word Sense Disambiguation, con particolare riferimento alle metodologie basate su Parsing grammaticale di tipo Context Free e Context Sensitive, ed alle metodologie stocastiche.

Sono stati analizzati e confrontati i principali strumenti Open Source utilizzati in ambito NLP, tra i quali GATE, OpenNLP e UIMA. Di ognuno di questi viene fornita una descrizione, se ne delinea l'architettura e le tecnologie utilizzate, e si conclude con una tabella comparativa, che enfatizzi le funzionalità offerte. Vedi Technical Reports **TR2.4.1 Stato dell'arte ed analisi delle metodologie e delle tecniche per il Natural Language Processing** e **TR2.4.2 Stato dell'arte ed analisi delle tecnologie e degli strumenti per il Natural Language Processing**.

Sono stati esaminati i principali modelli e le tecniche per l'Information Retrieval, con particolare riferimento ai modelli per la rappresentazione dei documenti e delle queries, delle metriche di similarità e delle metriche di valutazione dei risultati.

Sono stati analizzati i principali strumenti e le piattaforme tecnologiche, commerciali ed Open Source, per l'Information Retrieval. Tra questi si analizza Lucene, una libreria open source estremamente potente e versatile per l'Information Retrieval. Si analizzano i sistemi per la ricerca semantica, in particolare gli approcci per la ricerca sul web basati sulle ontologie. Infine si analizzano i motori di ricerca basati sul *document clustering*, ossia sulla classificazione dei documenti, i quali si allontanano dalla ricerca per indirizzi e parole chiave, per andare nella direzione di una ricerca semantica basata su concetti e categorie. Vedi Technical reports **TR2.4.3 Stato dell'arte ed analisi dei modelli e delle metodologie per l'Information Retrieval** e **TR2.4.4 Stato dell'arte ed analisi delle tecnologie e degli strumenti per la Information Retrieval**

Sono state analizzate le tecnologie dei Sistemi Esperti e della programmazione logica, in particolare è stata effettuata una rassegna dei linguaggi simbolici per la rappresentazione della conoscenza ed inferenza in ambito documentale ed ontologico.

Sono stati analizzati e confrontati i principali linguaggi simbolici per la rappresentazione della conoscenza ed inferenza in ambito documentale ed ontologico, tra i quali FLogic, Prolog, RuleML. Vedi Technical Report **TR2.4.5 Stato dell'arte ed analisi delle metodologie, dei linguaggi standard e delle tecniche per la Rappresentazione della Conoscenza ed Inferenza in ambito Ontologico**

Sono stati analizzati e confrontati i principali linguaggi di query del Semantic Web: RDQL, SPARQL e OWL-QL. E' stato analizzato il framework Jena, utilizzato per realizzare applicazioni Java ontology-based. Di questo framework si sono analizzati in dettaglio i moduli che implementano le funzionalità di reasoning ed il modulo ARQ, il quale è un componente che supporta differenti linguaggi di query e utilizza vari motori di esecuzione.

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------

**Vedi Technical Report TR2.4.6 Stato dell'arte ed analisi delle metodologie, dei linguaggi standard e degli engines per il querying semantico**

Sono state analizzate le principali metodologie e tecniche per il matching di ontologie, tra le quali tecniche di graph matching, di schema matching, matching semantico e sintattico. Vedi Technical Report **TR2.4.7 Stato dell'arte ed analisi delle metodologie e delle tecniche per il matching di ontologie**

Si sta progettando uno strumento prototipale che permetta di utilizzare all'interno di un unico framework, le tecniche e gli strumenti messi a disposizione dai principali strumenti Open Source utilizzati in ambito NLP, tra i quali GATE, OpenNLP e UIMA. Si veda il Technical Report (draft) **TR2.4.8 Realizzazione di uno strumento prototipale per il Natural Language Processing** per ulteriori dettagli.

Si sta definendo una architettura per il Semantic Information Retrieval. Il sistema che si sta definendo si basa su tecniche di Information Retrieval, Natural Language Processing e sulle ontologie, L'architettura che si sta realizzando interpreta le query utente in linguaggio naturale e crea, mediante un' analisi grammaticale e logica realizzata in Prolog, una *Query Ontology* che viene poi modificata ed arricchita dall' utente. Tale ontologia viene poi tradotta in un linguaggio di query basato sulle DL (ad esempio nRQL, SPARQL DL, DIG) e poi data in ingresso a dei reasoner (Pellet, Racer, etc.) oppure essere tradotta in un linguaggio di query RDF (SPARQL, SeRQL, RDQL) e data in ingresso ai relativi motori basati su RDF. L'esecuzione di tale query permette di ottenere le istanze che verificano i vincoli posti dalla query ontology. Infine viene utilizzata l'architettura definita nel [TR2.3.3](#), per permettere di associare le istanze trovate alle porzioni dei documenti che costituiscono le istanze stesse mediante annotazioni semantiche realizzate, e che rappresentano il vero obiettivo della ricerca utente. Vedi Technical report (Draft) **TR2.4.9 Definizione di una architettura per il Semantic Information Retrieval** per ulteriori dettagli.

Si sta realizzando uno strumento prototipale (denominato preliminarmente **Semantic Searcher**) che implementa l' architettura e permetta il retrieval semantico di documenti utilizzando la tecnica riassunta sopra. Vedi Technical report (Draft) **TR2.4.10 Realizzazione di uno strumento prototipale per il Semantic Information Retrieval** per ulteriori dettagli.

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------



Si sta realizzando uno strumento prototipale (denominato preliminarmente **Schema Matcher**) che permetta il matching tra differenti tipologie di *schema* ed in particolare permetta di definire un mapping semantico tra due ontologie, ai fini della “riconciliazione” tra ontologie o della ricerca di contenuti e servizi basati sulla similarità di ontologie. Vedi Technical report (Draft) **TR2.4.12 Realizzazione di uno strumento prototipale per il confronto (matching) di ontologie** e l’ articolo: B. Di Martino, “Semantic Web Services Discovery based on Structural Ontology Matching”, to be published in: International Journal of Web and Grid Services (IJWGS), Inderscience Press per ulteriori dettagli.

PUBBLICAZIONI PRODOTTE

F.Moscato, B.Di Martino: “Semantic Web and Semantic Information Management”, Int. J. Web and Grid Services, Vol. 4, No. 3, 2008.

B. Di Martino, “Semantic Web Services Discovery based on Structural Ontology Matching”, to be published in: International Journal of Web and Grid Services (IJWGS), Inderscience Press.

PROGETTO LC3	Revisione n*	0	Del	-----
--------------	--------------	---	-----	-------