UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

tesi di laurea

Sperimentazione di un processo a supporto dell'interoperabilità semantica nel web

Anno Accademico 2009/2010

relatore
Ch.mo prof. Porfirio Tramontana

candidato
Paola Pignata
Matr. 534/1560

Una panoramica generale

Web Semantico

- *Rappresentazione della conoscenza*
- *Ontologie*
- *Agenti semantici*

Interoperabilità semantica

l'abilità di programmi progettati indipendentemente e implementati su piattaforme diverse di cooperare tra di loro sulla base di una "comprensione" del significato dei dati oggetto di scambio

In questo lavoro ci siamo proposti di

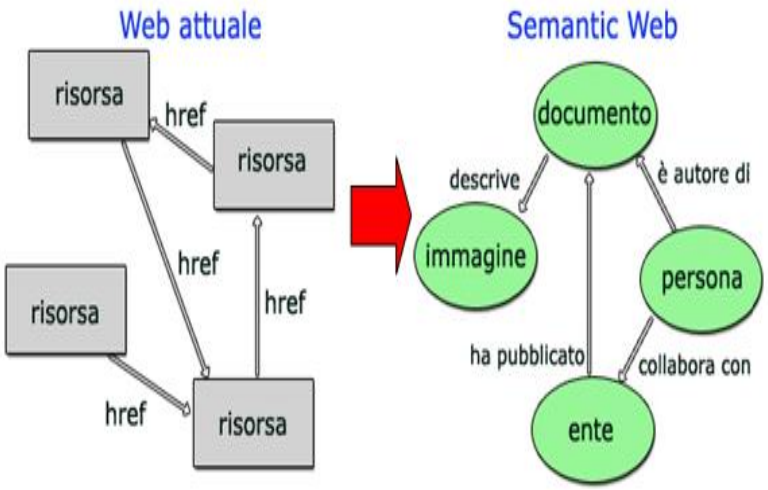
Valutare quantitavivamente i punti di forza e i punti deboli di uno dei modelli di comunicazione esistenti, proposti come approccio all'interoperabilità semantica

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

Web semantico ed Interoperabilità Semantico

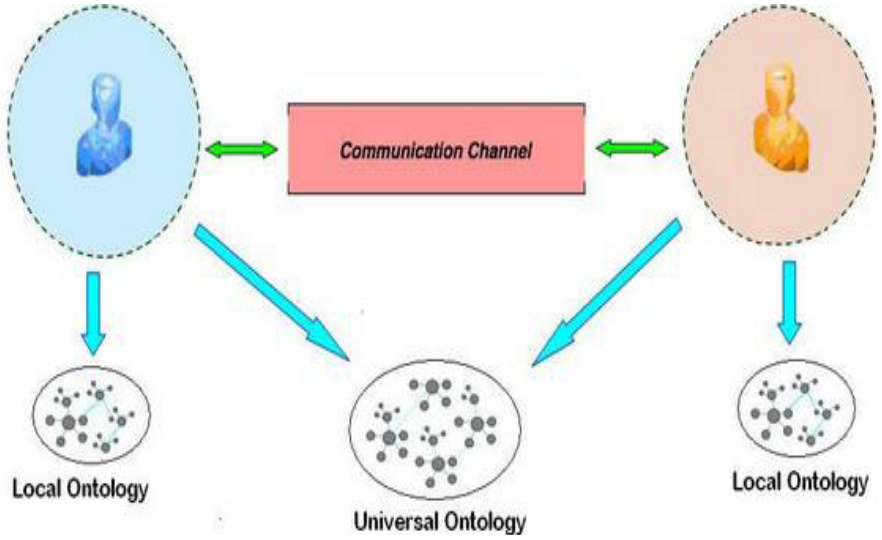
Sperimentazione di un processo a supporto dell'interoperabilità semantica nel web

Il Web Semantico



Il Modello di comunicazione Ibrido

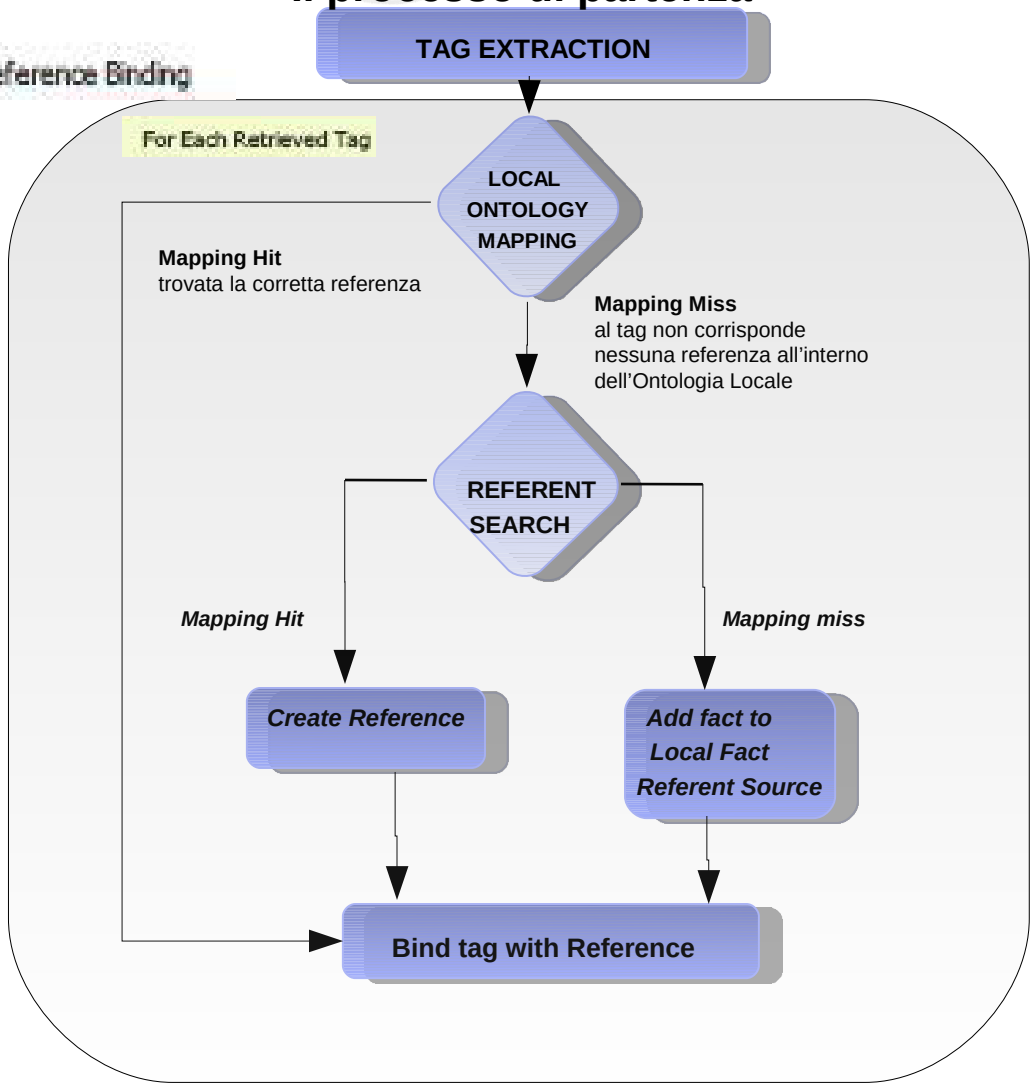
*ogni sender agent possiede la sua **personale conoscenza soggettiva** che può essere sia mappata in **sorgenti di conoscenza oggettiva condivisa**, come le ontologie, o direttamente inclusa nel messaggio codificato con l'obiettivo di preservare la personale interpretazione del concetto trasmesso*



UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

Il Processo di Comunicazione

Il processo di partenza



Il processo ottimizzato

- Ottimizzazione dell'ordine di consultazione delle fonti
- Consultazione parallela delle fonti

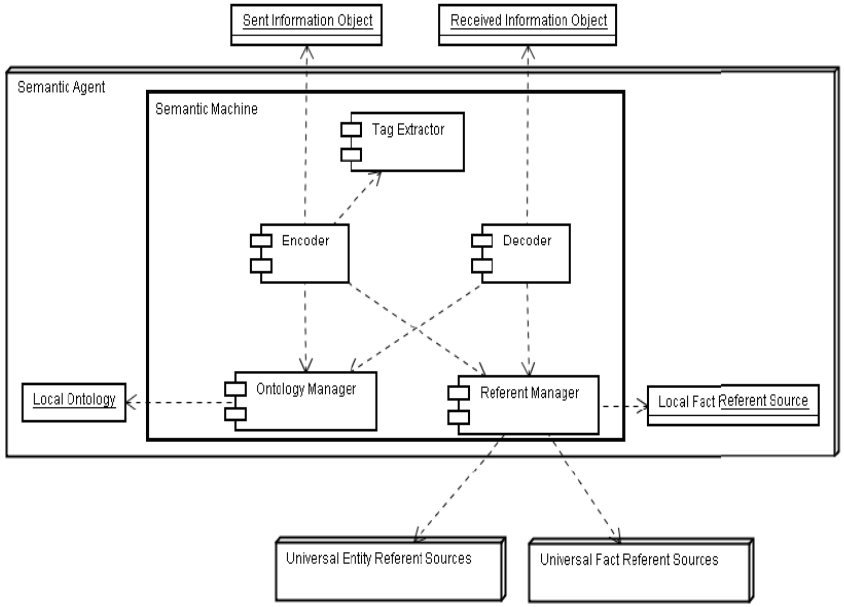
UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

La Macchina Semantica

Sperimentazione di un processo a supporto dell'interoperabilità semantica nel web

All'interno della macchina semantica è possibile distinguere i seguenti componenti:

- **ENCODER**: gestisce le operazioni di codifica;
- **DECODER**: gestisce le operazioni di decodifica.
- **SENDER**: è responsabile della trasmissione dell' Information Object codificato;
- **RECEIVER**: è responsabile della ricezione dell'Information Object trasmesso dal sender;
- **TAG EXTRACTOR**:estrarae e filtra l'insieme dei tag dai digital asset (pagine HTML) che devono essere codificati;
- **ONTOLOGY MANAGER** fornisce le funzionalità per interagire con la Local Ontology
- **REFERENCE MANAGER**: astrae le funzionalità di ricerca, inserimento, modifica e eliminazione delle referenze dall'ontologia locale;
- **LOCAL FACT REFERENT SOURCE**: rappresenta la conoscenza personale del sender.



- ***REFERENT MANAGER*** recupera i referenti all'interno della *Local Fact Referent Source* ed eventualmente li aggiunge
- ***UNIVERSAL ENTITY REFERENT SOURCE***: l'unica sorgente per le entità considerate è Wordnet.
- ***UNIVERSAL FACT REFERENT SOURCE***: come sorgenti di referenti universali abbiamoconsiderato: ***Wikipedia*** *e* ***Imdb***

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

La sperimentazione

Sperimentazione di un processo a supporto dell'interoperabilità semantica nel web

COMPLETEZZA E CORRETTEZZA

Domande	Metriche
Qual è la capacità di codifica del processo di associare a ciascun tag il significato corretto (cioè, il messaggio include un reference ad un referent valido di t)?	Encoding Correctness frazione di tag di D per i quali l'oracolo O giudica se sono corretti i significati associati al tag
Qual è la capacità di codifica del processo di associare ogni tag a un referente (cioè, il messaggio contiene almeno un reference relativo ad un referent)?	Encoding Completeness frazione di tag D per i quali il processo di codifica associa il tag con almeno un referent.

COSTO

Domande	Metriche
Qual è il costo per l'esecuzione delle attività di validazione di reference e referent ?	Validation Complexity (VC) numero totale di diversi significati esaminati per la validazione del tag. Average Validation Complexity (AVC) $AVC(D,\Omega)= \Sigma_D VC(\tau, \Omega) / \|D\|$ Effort $Effort = \alpha_{medio} * n_{accessi} + \beta_{medio} * n_{creazioni}$
Qual è il costo per l'esecuzione dell'attività Add Fact to Local Fact Referent Source?	Local Insertion Ratio (LIR) $LIR(D,\Omega)=\gamma(D,\Omega)/ \|D\|$

Con $\alpha = tval / num_{accessi}$,
$\beta = tcre / num_{creazioni}$
D è un determinato insieme di documenti;
Ω è un elenco ordinato di fonti
τ è un determinato tag

Ipotesi e obbiettivi della sperimentazione

La Teoria della Verifica delle Ipotesi

processo mediante il quale si stabilisce sulla base delle osservazioni campionarie se l'ipotesi formulata può considerarsi esatta oppure no

HO: indica ***l'ipotesi nulla,cioè*** l'ipotesi da verificare ;

H1: è ***l'ipotesi alternativa;***

quando accettiamo un'ipotesi, non stiamo affermando che sia necessariamente vera, ma solo che i dati raccolti non la escludono.

α *:* è il livello di significatività con cui si rifiuta o si accetta l'ipotesi H0; rappresenta la probabilità di commettere l'errore di rifiutare l'ipotesi anche se essa è "vera".

Ipotesi sulla dipendenza dal riempimento iniziale dell'Ontologia Locale

Esempio:

H0: La Validation Complexity (VC) non varia al variare del grado di riempimento iniziale del database

H1: La Validation Complexity (VC) varia al variare del grado di riempimento iniziale del database

Ipotesi sulla dipendenza dall'argomento

Esempio:

H0: Il tempo di validazione (TVal) non varia al variare dell'argomento

H1: Il tempo di validazione (Tval) varia al variare dell'argomento

Ipotesi sulla dipendenza dallo sperimentatore

Esempio:

H0: L'Effort non varia al variare dello sperimentatore

H1: L'Effort varia al variare dello sperimentatore

Procedura sperimentale

Turno	Argomento	Numero di documenti	Numero di tag	Utente
1	Arte	24	238	Sperimentatore
2	Biologia	24	253	Sperimentatore
3	Chimica	24	257	Sperimentatore
4	Film	24	221	Sperimentatore
5	Filosofia	24	236	Sperimentatore
6	Fisica	24	257	Sperimentatore
7	Gastronomia	24	251	Sperimentatore
8	Geologia	24	244	Sperimentatore
9	Musica	24	240	Sperimentatore
10	Storia	24	252	Sperimentatore
11	Teatro	24	253	Sperimentatore
12	Vario	20	400	Utente1/Utente2

*abbiamo eseguito l'esperimento in **12 turni diversi**, che differiscono tra loro per quanto riguarda il **dominio di appartenenza** dei lemmi da codificare e le **persone** che effettuano la codifica:in ognuno dei primi 11 turni abbiamo sottoposto gli stessi dati al processo di codifica due volte,la prima con **Ontologia Locale vuota**,la seconda con **Ontologia Locale preliminarmente riempita** con 1000 lemmi di dominio vario.*

Nell'ultimo turno sono stati selezionati 400 lemmi appartenenti a domini vari e sottoposti al processo di codifica da due utenti differenti.

Esempio

verifica delle ipotesi sulla dipendenza della AVC dal riempimento iniziale del database

	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	i=11
Xi_{AVC}	1,706	1,281	1,68	1,490	1,568	3,590	1,638	1,473	2,694	1,362	1,806
Yi_{AVC}	1,203	1,058	1,291	1,047	1,08	2,471	1,389	1,158	1,795	1,142	1,383
Wi_{AVC}	0,503	0,243	0,389	0,433	0,488	1,119	0,249	0,315	0,899	0,220	0,423

dove **Wi= Xi – Yi**, per **i=1,2,...,11** rappresentano le variazioni nelle variabili misurate a database inizialmente vuoto e inizialmente pieno

i=1,2…,11 indica ognuno degli 11 diversi argomenti

Se non vi fosse nessuna dipendenza,le Wi avrebbero media nulla

H0: $\mu_w = 0$ contro ***H1:*** $\mu_w \neq 0$

Se indichiamo con $\overline{W}$ la media del campione di n dati, con S_w la deviazione standard campionaria di $\overline{W}$ e con **t la variabile causale di Student**

Avvalendoci del **test t di Student** e fissato il livello di significatività α ***=0,05***:

si accetta H0 se $-t_{\alpha 2,n-1} \leq \sqrt{n} \cdot \frac{\overline{W}}{S_W} \leq t_{\alpha/2,n-1}$ *in particolare nel nostro caso* $-2,228 \leq \sqrt{n} \cdot \frac{\overline{W}}{S_W} \leq 2,228$

si rifiuta H0 negli altri casi

E nel caso di AVC $\sqrt{n} \cdot \frac{\overline{W}}{S_W} = 5,62 > 2,228$

Rifiutiamo dunque H0 e riteniamo accettabile

H1: La **Average Validation Complexity** (AVC) non varia al variare del grado di riempimento iniziale del database

Risultati

Tabella di valutazione della VC al variare dell'argomento (problema di Behrens-Fisher)

Arg	ARTE	BIO	CHIM	GAST	GEO	FILM	FILOS	FIS	MUS	STO	TEA
ARTE		-1,47	-0,08	-0,68	-0,34	4,67	-0,02	-0,77	2,42	-1,19	0,27
BIOL	1,47		1,89	1,38	1,05	7,87	1,91	1,55	4,7	0,99	2,28
CHIM	0,08	-1,89		-0,96	-0,36	5,95	-0,18	1,16	3,1	-2,07	2,48
GASTR	0,68	-1,38	0,96		0,26	6,55	0,65	-0,09	3,68	-0,83	1,2
GEO	0,34	-1,05	0,36	-0,26		5,11	0,21	-0,33	2,8	-0,74	0,67
FILM	-4,67	-7,87	-5,95	-6,55	-5,11		-5,76	-6,87	-2,17	-7,57	-4,9
FILOS	0,02	-1,91	0,18	-0,65	-0,21	5,76		-0,79	3,06	-1,46	0,58
FIS	0,77	-1,55	-1,16	0,09	0,33	6,87	0,79		3,87	-0,88	-2,34
MUS	-2,42	-4,7	-3,1	-3,68	-2,8	2,17	-3,06	-3,87		-4,42	2,4
STO	1,19	-0,99	2,07	0,83	0,74	7,57	1,46	0,88	4,42		1,98
TEA	-0,27	-2,28	-2,48	-1,2	-0,67	4,9	-0,58	-2,34	-2,4	-1,98	

Tabella di valutazione del LIR al variare dell'argomento (p-dei-dati, test di Fisher-Irwin)

Arg	BIO	CHIM	GAST	GEO	FILM	FILOS	FIS	MUS	STO	TEA
ARTE	0,0002	0,00001	0,007	0,00001	0,878	0,256	0,212	0,89000	0,00041	0,15200
BIO		0,66200	0,370	0,611	0,00074	0,014	0,657	0,00023	0,99000	0,02700
CHIM			0,130	1,102	0,00007	0,0022	0,578	0,00001	0,52000	0,00470
GAST				0,116	0,01800	0,15600	0,173	0,00530	0,48900	0,24900
GEO					0,00007	0,00199	0,241	0,00001	0,48100	0,00420
FILM						0,41110	0,002	0,81300	0,00144	0,26100
FILO							0,014	0,22600	0,02400	0,88100
FIS								0,00065	0,26000	0,00449
MUS									0,00030	0,12700
STO										0,04480

Conclusioni

I risultati che abbiamo collezionato nei vari turni dell'esperimento ci portano ad accettare le seguenti ipotesi:

- *AVC, Tval, LIR, Effort* **variano** *al variare del grado di riempimento iniziale dell'Ontologia Locale;*
- *AVC, Tval, LIR, Effort* **variano** *al variare dell'argomento*
- *Tval, Effort* **variano** *al variare dello sperimentatore,*

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

Conclusioni e Sviluppi Futuri

Validità interna dell'esperimento

effetto di selezione:

dovuto alle differenze nella capacità tra i soggetti coinvolti nell'esperimento;

SOLUZIONE

agenti selezionati a caso, ma con conoscenze analoghe del dominio di applicazione e circa il processo di codifica.

effetto di strumentazione:

dovuto alla dipendenza dall'ordine delle fonti che potrebbe causare differenze di prestazioni.

SOLUZIONE

l'ordinamento ottimo delle fonti creato con il parallelismo delle interrogazioni

effetto di maturità:

dovuto al progressivo riempimento dell'Ontologia Locale.

SOLUZIONE

si potrebbe limitare tale effetto eseguendo la sperimentazione con il database in parte riempito

Sviluppi Futuri

- *sperimentazione su ordine delle fonti*
- *l'automatizzazione dell'estrazione dei Tag*
- *aggiunta di nuove fonti*
- *l'automatizzazione del processo di REFERENCE VALIDATION*