A. D. MCCXXIV

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

Scuola di Dottorato in Ingegneria dell'Informazione
Dottorato di Ricerca in Ingegneria Informatica ed Automatica

Comunità Europea
Fondo Sociale Europeo

TECNICHE E MODELLI
PER LA RICERCA SEMANTICA SUL WEB
UN APPROCCIO BASATO SU ONTOLOGIE

ANTONIO MARIA RINALDI

Tesi di Dottorato di Ricerca

Novembre 2005

via Claudio, 21- I-80125 Napoli - ☎ *[#39]* (0)81 768 3813 - *[#39]* (0)81 768 3816

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

Scuola di Dottorato in Ingegneria dell'Informazione

Dottorato di Ricerca in Ingegneria Informatica ed Automatica

TECNICHE E MODELLI
PER LA RICERCA SEMANTICA SUL WEB
UN APPROCCIO BASATO SU ONTOLOGIE

ANTONIO MARIA RINALDI

Tesi di Dottorato di Ricerca

(XVIII ciclo)

Novembre 2005

Il Tutore

Prof. Antonio Picariello

Il Coordinatore del Dottorato

Prof. Luigi Pietro Cordella

Dipartimento di Informatica e Sistemistica

via Claudio, 21- I-80125 Napoli - ☎ *[#39]* (0)81 768 3813 - *[#39]* (0)81 768 3816

INTRODUZIONE

L'avanzamento delle tecnologie dell'informazione ha permesso la creazione di amplissime collezioni di documenti in formato elettronico riguardanti gli argomenti più disparati; di queste collezioni possiamo dire che l'esempio più rappresentativo è il World Wide Web.

Non sempre, però, la grande disponibilità d'informazione è da ritenersi un fatto positivo, anzi, in certi casi, può diventare addirittura deleteria.

Studi recenti hanno messo in evidenza come l'immensa quantità d'informazione offerta da internet possa provocare conseguenze psicologiche di vario tipo negli utilizzatori: da problemi nella sfera relazionale, allo scarso rendimento sul lavoro a causa di "navigazione compulsiva" sul web [Greenfield2002].

Ben più grave appare, secondo Francis Heylighen, l'effetto che l'Information Overload (sovraccarico d'informazione) sta avendo sulla nostra società. Sembrerebbe, infatti, che l'eccessiva informatizzazione stia incrementando esponenzialmente la velocità dei processi evolutivi della nostra società e ne stia anche aumentando la complessità [Heylighen2002].

Il fenomeno che sembra caratterizzare i nostri tempi è l'aumento progressivo di produttività in qualunque settore, intendendo con il termine "produttività" il rapporto tra il risultato ottenuto da un processo e le risorse impiegate. Questo fenomeno nel mondo dell'informazione è perfettamente rappresentato da internet: oggi è possibile pubblicare qualunque documento ad un costo praticamente nullo, senza nessun tipo di filtro.

Questo se da un lato ha provocato un incremento della quantità d'informazione disponibile, da un altro lato ne ha, in media, ridotto la qualità.

A causa della complessità nell'organizzazione dei dati e della quantità di materiale presente, la ricerca sul Web di informazioni davvero utili è diventata decisamente complessa. Lo sforzo fatto dalla comunità scientifica e dalle aziende che si occupano di information retrieval ha fornito agli utenti potenti mezzi, come ad esempio i motori di ricerca, per assisterli nella scoperta di risorse. Le tecniche di ricerca sono le più disparate ma i risultati sono lontani dal soddisfare le richieste di una ricerca mirata. Trovare informazioni usando i tradizionali motori si rivela fruttuoso solo in presenza di argomenti di una certa notorietà e importanza e di query molto precise; negli altri casi questo lavoro può implicare

una considerevole perdita di tempo dato che un utente deve raffinare manualmente la ricerca visitando una ad una le pagine restituite.

Questo avviene perché i motori di ricerca tradizionali effettuano ricerche di tipo sintattico: essi restituiscono le pagine che contengono le keywords presenti nelle query degli utenti, indipendentemente dal contesto in cui esse sono utilizzate oppure restituiscono pagine secondo algoritmi differenti, ad esempio basati sulla popolarità.

Se ciò da un lato è conveniente in termini di velocità di reperimento delle pagine e restituzione dei risultati, dall'altro lato porta spesso a risultati errati o imprecisi, dato che vengono restituite molte pagine non attinenti al contesto della query dell'utente. Ad esempio, un utente che voglia cercare pagine relative al jazz, inserendo in un motore come keyword la parola "Davis", oltre a trovare pagine attinenti al dominio musicale troverà sicuramente pagine inerenti al dominio sportivo.

Inoltre, sempre più sono i motori di ricerca che ordinano i documenti web secondo delle politiche commerciali: la registrazione a pagamento di un sito in un motore di ricerca offre una serie di vantaggi, primo fra tutti la certezza che il sito sia presente in testa all'elenco o nelle prime posizioni, anche se tratta marginalmente un determinato argomento.

In un contesto tale ha acquisito sempre più importanza nelle scienze informatiche, ed in particolare nel settore dell'information retrieval, il concetto di "rilevanza" delle informazioni. Questo concetto, che per l'uomo è del tutto intuitivo e nella maggior parte dei casi inconscio, è definito da Schutz come l'inerenza di un informazione ad un tema, cioè al particolare aspetto o oggetto della nostra concentrazione, avendo come base un orizzonte, ossia l'insieme delle conoscenze da noi possedute [Schutz1970].

Sarebbe conveniente avere a disposizione un sistema in grado di "capire" di cosa parla una pagina, valutando la sua attinenza con i domini di interesse per l'utente. Una ricerca di tale tipo è detta semantica in quanto non restituisce semplicemente pagine che contengono le keywords, ma pagine che hanno anche un contenuto semantico aderente al dominio desiderato dall'utente.

I ricercatori stanno cercando di dare risposte a questi problemi e una delle soluzioni più accredidate sembra essere il Semantic Web [BernersLee2001].

E' opinione di chi scrive che, anche se questo modo ci concepire il Web è affascinante e promettente, siamo ancora lontani da un suo utilizzo a larga scala

dato che il metodo proposto implica necessariamente uno stravolgimento dell'attuale struttura del Web.

Lo scopo di questo lavoro è quello progettare e realizzare un meta-motore di ricerca semantico, partendo dalla teorizzazione di tecniche e modelli fino ad arrivare all'implementazione e al testing finale.

Nel primo capitolo saranno introdotti i concetti più importanti attorno ai quali si sviluppa l'Information Retrieval e Representation (IRR); nel secondo capitolo verranno descritti i modelli più importanti per l'IR; nel terzo capitolo parleremo di sistemi noti in letteratura per la ricerca semantica; nel quarto capitolo varrà descritta una tecnica per la rappresentazione della conoscenza, l'ontologia; nel quinto capitolo si parlerà delle metriche per la misura della similarità tra concetti; nel sesto capitolo verrà presentato un modello proposto per l'information retrieval e sarà descritto un sistema che si basa su questo modello; nel settimo capitolo verrà descritta la metodologia per la valutazione del sistema e sarà presentata una sperimentazione; nell'ottavo e ultimo capitolo verranno discussi i risultati ottenuti e presentate le conclusioni.

CAPITOLO 1 IL PROCESSO DI INFORMATION RETRIEVAL E REPRESENTATION E LE NECESSITÀ INFORMATIVE DEGLI UTENTI

Il termine "Società dell'Informazione" ricorre già nel Libro Bianco della Comunità Europea di Jacques Delors pubblicato nel 1993 [Delors1993]. In questo rapporto l'enfasi veniva data non solo alle così dette "autostrade dell'informazione" -termine utilizzato per descrivere le infrastrutture di comunicazione utilizzate per la connettività- ma soprattutto ad una più complessa organizzazione della conoscenza, e quindi dell'informazione, necessaria per il corretto avanzamento sociale ed economico.

Il concetto di *informazione* è stato nel corso del tempo assimilato, comparato e contrapposto a quello di *dato*, *conoscenza*, *sapienza* [Meadow1992] e d'altro canto le parole *informazione, testo, documento* sono spesso usate in maniera intercambiabile. L'avanzamento tecnologico e l'affinamento di tecniche e algoritmi per la gestione, la rappresentazione ed il retrieval dei documenti hanno portato negli ultimi anni ad un passaggio dal *document retreval* al *passage retrieval* [SparkJones2000]. Il passage retrieval, anche detto *information retrieval*, sta ad indicare le tecniche e le metodologie per il recupero di informazioni "realmente utili" per l'utente che non devono coincidere necessariamente con tutto il documento; mentre il document retrieval implica la presentazione all'utente di interi documenti anche se piccole parti di essi riguardano l'argomento di interesse.

1.1 INFORMATION REPRESENTATION AND RETRIEVAL (IRR)

Qualsiasi tipo di informazione ha bisogno di essere rappresentata prima di poter essere recuperata. Da qui l'*Information Representation* va a raccogliere le tecniche per l'estrazione dai documenti di alcuni termini caratteristici (keywords o frasi), o l'assegnazione di termini al documento stesso (descrittori o argomenti). In genere l'Information Representation può essere ottenuta tramite una qualsiasi combinazione di abstracting, indicizzazione, categorizzazione, summarization ed estrazione. Inoltre sia l'*Information Processing* che l'*Information Management* sono spesso usati come sinonimi di Information Representation anche se il primo si riferisce a come l'informazione viene trattata

in funzione del processo di retrieval, mentre il secondo tiene conto di quelle attività che vanno dalla selezione delle informazioni al suo storage.

L'*Information Retrieval* è stata trattata, in linea di massima, come un campo di ricerca che ricopriva sia la rappresentazione che il recupero dell'informazione [SparkJones1997]. In particolare, però, la dimensione legata al recupero informativo è principalmente riferita ad azioni come l'*Information Access*, l'*Information Seeking* e l'*Information Searching*. Questi termini possono essere usati come sinonimi di *retrieval* ma ognuno di loro ha differenti sfaccettature in funzione dei risvolti applicativi e funzionali.

Con information access si enfatizzano gli aspetti legati all'accesso all'informazione; con il termine information seeking il focus è concentrato sull'utente che partecipa attivamente al processo di retrieval e con information searching si intende il "come" cercare l'informazione.

Un altro livello che si può dare all'information retrieval va sotto il nome di *information storage*. Le azioni principali che si trovano a questo livello sono quelle di registrazione ed immagazzinamento dell'informazione.

LA DIMENSIONE UTENTE

L'utente rappresenta un fattore cruciale che deve essere preso in considerazione in tutte le attività dell'IRR che hanno, come obiettivo finale, proprio quello di soddisfare i bisogni informativi degli utenti.

Gli utenti sono individui, ognuno con le loro caratteristiche distintive. Naturalmente è impraticabile lo studio di ogni singolo utente per scopi legati all'IRR. Per questo si cerca di raggruppare gli utenti utilizzando criteri comuni come sesso, età, occupazione, livello economico, cultura, educazione. Queste caratteristiche generali sono alla base dei processi che portano ad una maggiore comprensione delle necessità degli utenti ed inoltre partecipano alla definizione di componenti aggiuntive che permettono una migliore personalizzazione dei bisogni informativi.

Ad esempio prendendo in considerazione l'età, utenti di differenti età hanno differenti necessità informative (maggiore o minore interesse per le nuove tecnologie, musica, itinerari turistici ecc…). Gli utenti in età scolastica o i lavoratori hanno normalmente interessi specifici legati all'educazione o al lavoro mentre altri gruppi tendono a volere informazioni più generali su questi argomenti. L'occupazione è un altro criterio molto importante per determinare i

bisogni dell'utente. Gli scienziati tendono a formare "college invisibili" per scambiare informazioni [Crane1972] [Price1963], mentre gli ingegneri nei laboratori di ricerca e sviluppo fanno affidamento a "gatekeepers" per comunicare con il resto del mondo [Allen1970].

Chowdhury [Chowdhury1999] enumera i bisogni informativi degli utenti per differenti aree di attività.

Il cercare di comprendere gli utenti ed i loro bisogni facilita enormemente il processo di IRR. Molti importanti lavori [Borgman1989] [Fenichel1981] [Marchionini,1993.] [Zhang2001] sono stati portati avanti per esplorare l'impatto degli attributi dell'utente sull'IRR. Comunque, anche dopo decenni di studi, il bisogno informativo rimane un concetto vago dato che non è ben strutturato, suscettibile di diverse interprestazioni e non organizzato. Per queste ragioni in [Belkin1982] questo termine è stato definito come ASK (Anomalous State of Knowledge). In aggiunta le necessità informative hanno altre caratteristiche come il veloce cambiamento, la soggettività, la dipendenza dall'ambiente, restando quindi poco o per nulla espresse [Chowdhury1999]; d'altro canto gli utenti, la maggior parte delle volte, hanno difficoltà nell'esprimere questi bisogni.

Oltre alle loro caratteristiche, Paisley [Paisley1968] identificò altri quattro fattori che influiscono sui bisogni informativi degli utenti:

1. **Information sources**: l'insieme delle sorgenti informative disponibili agli utenti hanno alcuni effetti sui loro bisogni; questo influisce in maniera forte sulle loro aspettative;
2. **Scopo dell'informazione** (What for): a che scopo gli utenti hanno bisogno di informazioni?
3. **Fattori esterni**: fattori sociali, politici ed economici, ad esempio, possono avere forti impatti sugli utenti e sulle loro esigenze;
4. **Risultati**: quali risultati hanno gli utenti in funzioni delle informazioni trovate?

In sintesi gli utenti stessi insieme con le sorgenti informative, gli scopi, i fattori esterni ed i risultati costituiscono i cinque fattori che influenzano i loro *information needs*.

Oltre ai fattori che le influenzano, le necessità informative possono essere divise in differenti tipi. Una classificazione comune è quella di dividerle in due categorie: *know-item need* e *subject need* [Lancaster1993] o *concrete information need* e *problem-oriented information need* [Frants1988].

Anche se con nomi differenti *know-item need* e *concrete information need* rappresentano concetti simili e in maniera analoga sono sinonimi *subject need* e *problem-oriented information need*.

GLI ASPETTI DEL PROCESSO DI IRR

Per dare realmente atto alle esigenze di retrieval di informazioni degli utenti è necessario avere una visione ad alto livello del processo di IRR e successivamente disegnare un modello logico di un sistema adatto al recupero di informazioni realmente utili. Per tali ragioni parleremo di *user task* e di *rappresentazione logica dei documenti*.

L'utente di un sistema di retrieval deve tradurre le sue necessità informative in una query espressa tramite un linguaggio messo a disposizione dal sistema stesso. In un sistema di information retrieval questo implica di solito la specificazione di un set di parole che contengono la semantica del bisogno informativo. In un sistema per il retrieval di dati si utilizza una query che contiene i vincoli che devono essere soddisfatti dagli oggetti facenti parte dell'insieme dei risultati. In entrambi i casi si dice che l'utente cerca "informazioni utili" eseguendo un *retrieval task*.

Consideriamo un utente il cui interesse è scarsamente specificato o che è molto generale, ad esempio vorrebbe avere informazioni sulla città di Roma. Naturalmente gli saranno presentati un numero molto elevato di documenti che vanno dalle notizie storico-culturali, a quelle socio-politiche ad informazioni sportive. In questa situazione l'utente comincia a sfogliare i documenti presentati per cercare quelli che soddisfano le sue richieste. Quindi l'utente è interessato da diversi task mentre utilizza un sistema di information retrieval, in particolare possiamo parlare di *retrieval* e *browsing*. In una distinzione classica esistono sistemi di information e data retrieval che tendono a soddisfare il primo task, mentre il secondo si ottiene tramite sistemi ipertestuali.

Nel paradigma organizzativo del World Wide Web devono essere considerate entrambe queste azioni (pulling actions) portando l'utente a richiedere informazioni in maniera interattiva.

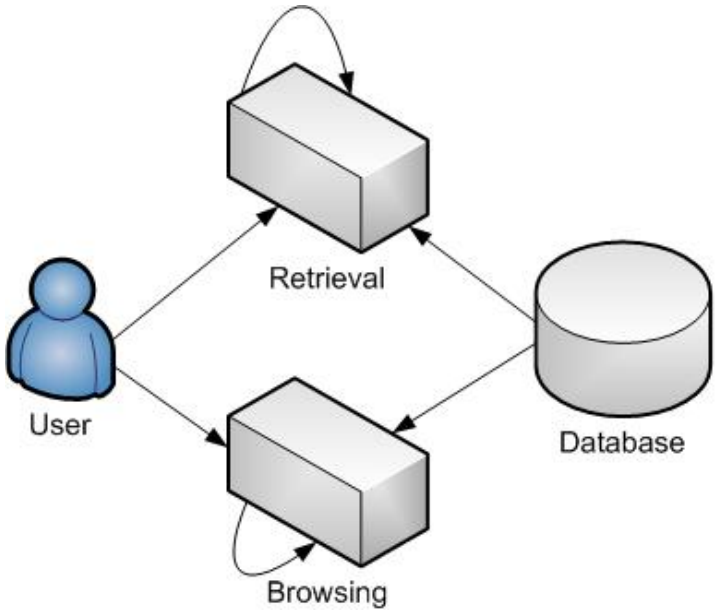


Figura 1.1: Interazione Utente-Sistema

Il contenuto informativo dei documenti dovrebbe essere ben "organizzato" per permettere delle efficienti ed efficaci azioni di retrieval. I documenti devono quindi essere rappresentati ad esempio tramite un insieme di keyword estratte automaticamente dal testo oppure annotate manualmente. Questo permette una *rappresentazione logica dei documenti*. Esistono differenti tipi di rappresentazioni logiche, ad esempio alcuni sistemi possono utilizzare tutte le parole del testo per rappresentare un documento ed in questo caso si parla di *full text representation*. D'altro canto però, in collezioni di documenti molto grandi, bisogna ridurre il numero di termini considerati. Questo si ottiene mediante l'eliminazione delle *stopwords* (come articoli e congiunzioni), l'uso dello *stemming* (che riduce le parole alla loro comune radice grammaticale) e l'identificazione di *noun groups* (eliminando aggettivi, verbi e avverbi). Queste operazioni sono chiamate *text trasformations*.

Chiaramente il full text è la rappresentazione logica più completa, ma il suo uso implica un elevato sforzo computazionale. Un set limitato di termini fornisce una rappresentazione più concisa, ma il corrispondente retrieval è di bassa qualità.

Esistono molte rappresentazioni logiche intermedie che possono essere adottate in un sistema di IRR. Inoltre, adottando queste rappresentazioni il sistema dovrebbe anche essere in grado di stabilire la struttura interna di un documento (capitoli, sezioni, sottosezioni,…). I livelli di rappresentazione intermedia sono mostrati nella figura seguente dove si può notare che la scala di

rappresentazione va da un approccio full text a uno di più alto livello definito da un soggetto umano.

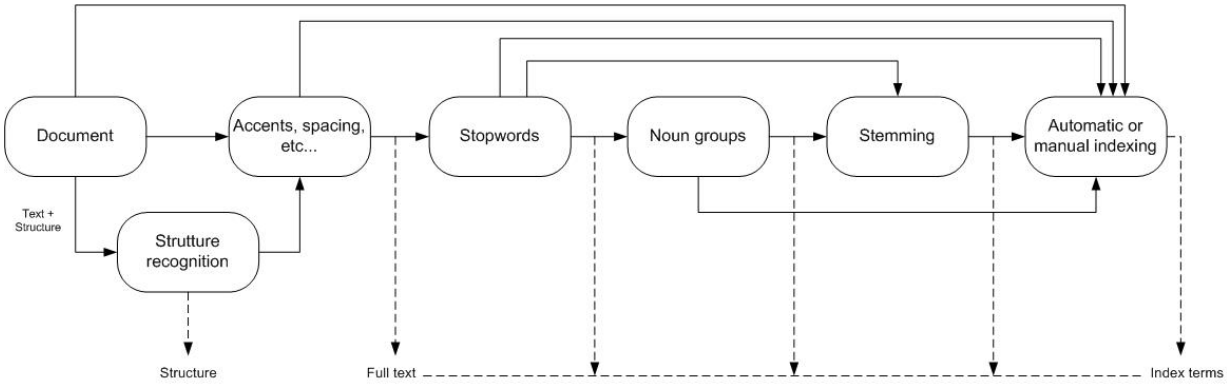


Figura 1.2: Rappresentazione logica di un documento

1.1.1 Le componenti principali

Il processo di IRR è costituito da alcune componenti principali rappresentate dai database, dai meccanismi di ricerca, dai linguaggi e dalle interfacce:

- **Database**: è un deposito di informazioni organizzato in maniera strutturata per permettere un ordinamento, una ricerca ed un retrieval di informazioni il più efficiente possibile. Un database è formato da due parti: file sequenziali e indici. I primi contengono informazioni ordinate in accordo con la struttura dei campi e dei record del database, gli altri permettono l'accesso alle informazioni contenute nei file sequenziali in funzione delle query sottomesse. In senso non tradizionale, un database (ad esempio contenuto in un sistema di Internet retrieval), è sempre costituito da file sequenziali ed indici ma il contenuto informativo è formato da informazioni strutturate in un formato prose-like come ad esempio il contenuto di una pagina Web. Naturalmente il contenuto e la strutturazione del database determinano quello che può essere recuperato da un sistema di IR;
- **Search Mechanism**: l'informazione rappresentata ed organizzata in un database può essere cercata e recuperata solo se si utilizza un meccanismo di ricerca. Tale meccanismo può avere differenti gradi di sofisticazione che sono definiti principalmente dagli algoritmi di ricerca a dalle procedure implementate nel sistema di IR. La capacità di un meccanismo di ricerca determina quali tecniche di ricerca sono a disposizione dell'utente e come le informazioni immagazzinate nei database possono essere recuperate;

- **Linguaggio**: per poter processare, trasferire o comunicare un'informazione è necessario un linguaggio che permetta di esprimerla. Nell'ottica dell'IRR possiamo suddividere il linguaggio in *natural language* e *controlled vocabulary*. Il primo è quello utilizzato naturalmente dalle persone per rappresentare un'informazione o per fare domande. Il secondo è definito come un linguaggio artificiale stabilendo una limitazione nel vocabolario, nella sintassi e nella semantica [Wellish1996]. Il linguaggio può determinare un certo grado di flessibilità nei sistemi di IRR.
- **Interfaccia**: in accordo con [Shaw1991] un'interfaccia è quello che l'utente vede, sente e tocca mentre interagisce con un sistema informatico. In particolare, nei problemi di IRR, l'interfaccia sopperisce alla naturale difficoltà che l'uomo ha nell'utilizzo delle tecnologie dell'informazione. Da qui è necessario effettuare un accurato studio delle tecniche e dei modelli che rendono il più user-friendly possibile un sistema di IRR e per avere una corretta presentazione delle informazioni che soddisfano le necessità degli utenti.

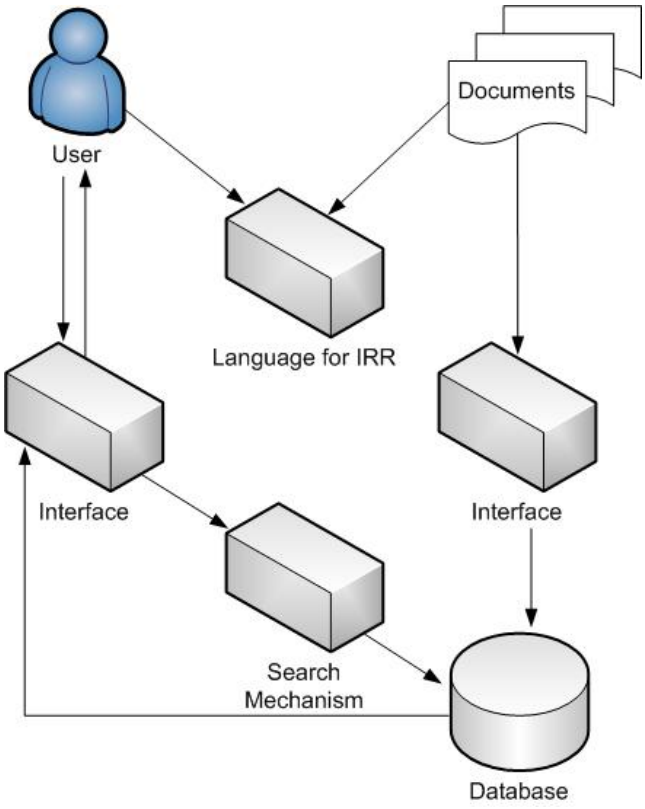


Figura 1.3: Il processo di IRR

Il concetto che più di altri riassume il fine ultimo del processo di IRR è quello di *rilevanza* di un documento.

La rilevanza è definita stabilendo un insieme di criteri che tentano di descrivere, seguendo differenti punti di vista, l'attinenza di un documento ad una determinata ricerca in funzione della soddisfazione dell'utente. In [Barry1995] sono riportati una serie di criteri che possono essere presi come guida nello studio della rilevanza di un documento. I criteri individuati possono essere divisi in categorie che vanno dalle caratteristiche del testo alla conoscenza dell'utente, alle sue aspettative e alle sue preferenze; criteri molto importanti possono inoltre essere individuati nella percezione dell'accuratezza dell'informazione, nell'affidabilità, nella qualità e così via.

Da qui la rilevanza indica una relazione e differenti tipi di manifestazioni della rilevanza includono differenti relazioni.

Nell'ambito dell'information retrieval e delle scienze dell'informazione la rilevanza può avere vari significati in funzione di quali relazioni usiamo e di cosa vogliamo misurare [Saracevic1996]; pertanto possiamo considerarne diverse manifestazioni:

- **system relevance**: è la relazione tra la query dell'utente e la rappresentazione di un documento nel sistema, in questo caso il grado di rilevanza dipende da come il sistema acquisisce, rappresenta, organizza e confronta i documenti e le query;
- **topical relevance**: è le relazione tra il soggetto di una data query e l'argomento di un documento;
- **cognitive relevance**: è la relazione tra lo stato di conoscenza e le necessità informative dell'utente e il documento restituito;
- **situational relevance**: è la relazione tra l'utilità che l'utente ha dal retrieval di un documento;
- **motivational relevance**: è la relazione tra gli obiettivi di un utente e i documenti analizzati.

Queste descrizioni si accordano con i framework di molti sistemi di IR e una loro classificazione è necessaria per diminuire le confusioni semantiche legate al concetto stesso di rilevanza.

CAPITOLO 2 MODELLI PER L'INFORMATION RETRIEVAL

Nella definizione di un sistema di information retrieval, si deve cercare di considerare tutte le possibili componenti in modo tale da avere uno sviluppo, il più possibile formale, del processo di IR.

Per questo motivo dobbiamo parlare di *modelli* di retrieval che stanno alla base dei sistemi di IR, l'architettura sulla quale si dovrà basare il sistema e le funzionalità che dovranno essere implementate nel sistema.

2.1 MODELLI PER L'INFORMATION RETRIEVAL

I sistemi di IR di solito utilizzano termini indicizzati per individuare e restituire documenti. Da un punto di vista specifico, un termine indice è una keyword o un gruppo di parole correlate, che hanno determinati significati nel contesto del documento. Nel caso generale, un termine indice è semplicemente una parola che appare nel testo. Il retrieval basato sul concetto di termini indice è semplice ma si devono fare alcune considerazioni sul processo di IR.

Il considerare termini indice parte dal presupposto che la semantica di un documento e le necessità informative dell'utente possono essere espresse attraverso l'insieme di tali termini. Naturalmente questo porta ad una notevole semplificazione del problema perché una considerevole parte della semantica e delle richieste dell'utente vengono perse quando si sostituisce il testo con un insieme di parole. Il matching tra le richieste utente e i documenti risulta quindi molto impreciso.

In particolare il matching risulta essere l'attività fondamentale per il processo di IR. Esistono differenti approcci per il matching che si basano sulla corrispondenza tra termini, su misure di similarità, sulla frequenza di occorrenza delle parole, ecc…

Il matching tra termini può essere di diversi tipi:

- **exact match:** la rappresentazione della query corrisponde esattamente alla rappresentazione del documento nel sistema;
- **partial match:** solo una parte dei termini usati nella formulazione della query corrisponde con la rappresentazione del documento nel sistema;
- **positional match:** prende in considerazione la posizione dei termini usati nella query nella rappresentazione del documento;

- **range match:** si può utilizzare in espressioni numeriche.

Tutti questi tipi di matching vengono utilizzati senza nessuna trasformazione tra le query e i documenti.

In alcuni modelli il matching tra la rappresentazione della query e la rappresentazione del documento non è esatto. Si deve quindi provvedere ad una trasformazione definendo una misura di similarità tra le rappresentazioni iniziali e finali.

Un problema centrale nel sistemi di IR è quello di identificare quali documenti sono rilevanti e quali no in un determinato contesto. Questa decisione è frutto degli algoritmi di ranking che cercano di stabilire un ordine tra i documenti restituiti. Questo tipo di algoritmo si basa sul concetto di rilevanza di un documento. Differenti premesse nella definizione di rilevanza portano a differenti tipi di modelli per l'IR.

Se in un sistema di IR i documenti nella collezione rimangono relativamente statici mentre nuove query vengono sottomesse, il retrieval è detto *ad hoc*. Se invece le query rimangono abbastanza statiche mentre nuovi documenti sono inseriti o cancellati nel sistema, la tecnica di retrieval è detta *filtering*.

Nel corso del tempo i ricercatori hanno proposto differenti tipi di modelli per l'information retrieval dei quali si può costruire una tassonomia mostrata nella figura seguente.

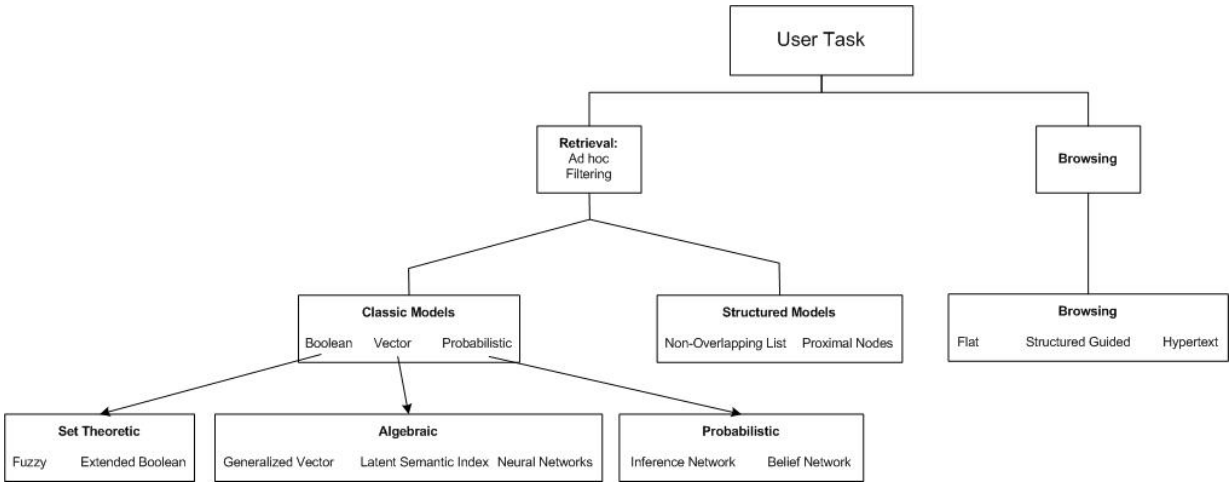


Figura 2.1: Una tassonomia dei modelli per l'IR

Prima di descrivere questi modelli, cercheremo di darne una formalizzazione generale.

Un modello per l'information retrieval è una quadrupla [**D**, **Q**, **F**, **R**(q_i, d_j)] dove:

1. **D** è un insieme composto dalle rappresentazioni logiche dei documenti nella collezione;
2. **Q** è un insieme composto dalle rappresentazioni logiche delle necessità informative dell'utente. Queste rappresentazioni sono chiamate query;
3. **F** è un framework per rappresentare i documenti, le query e le loro relazioni;
4. **R**(q_i, d_j) è una funzione di ranking che associa un numero reale e una rappresentazione di un documento $d_j \in D$ ad una query $q_i \in Q$. Questo ranking definisce un ordinamento tra i documenti in funzione della query q_i.

Per costruire un modello per l'IR si deve prima definire una rappresentazione dei documenti e delle necessità informative; successivamente dobbiamo pianificare il framework nel quale verranno modellate. Il framework fornisce anche una base per la determinazione della funzione di ranking.

2.1.1 Boolean Model

Il Modello Booleano è basato sulla teoria degli insiemi e sull'algebre di Boole. In questo modello le query sono espresse usando espressioni booleane con una semantica ben precisa.

Grazie alla sua semplicità questo modello è stato oggetto di grande attenzione nel passato e ha rappresentato la base di molti sistemi. Sfortunatamente però il modello booleano ha alcuni punti deboli:

- la strategia di retrieval è basata su un criterio di decisione binaria (un documento è rilevante o non-rilevante) senza alcuna nozione di scala ordinale;
- non è semplice trasformare una necessità informativa in un'espressione booleana.

Il modello presuppone che i termini indice siano presenti o assenti in un documento. Da questo i pesi di questi termini sono binari: $w_{i,j} \in \{0,1\}$.

Rispetto al modello una query q è composta da termini indice relazionati tra loro attraverso tre operatori: *not*, *and*, *or*. Pertanto una query è essenzialmente un'espressione booleana convenzionale che può essere espressa come una disgiunzione di vettori congiunti; ogni componente è formata da un vettore binario pesato.

Definiamo ora in maniera formale il Boolean Model e la sua metrica di similarità:

Per il modello booleano i pesi del termine indice sono tutti binari; una query q è un'espressione booleana convenzionale. Sia $\vec{q}_{dnf}$ *la forma normale disgiuntiva per la query q, inoltre sia* $\vec{q}_{cc}$ *una delle componenti congiuntive di* $\vec{q}_{dnf}$*. La similarità di un documento* d_j *alla query q è definita come:* $sim(d_j, q) = \begin{cases} 1 & se \quad \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge \forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc}) \\ 0 & otherwise \end{cases}$*. Se* $sim(d_j,q)=1$*, allora il modello ci dice che il documento* d_j *è rilevante per la query q (potrebbe anche non esserlo); altrimenti il documento non è rilevante.*

Il modello booleano impone che un documento sia rilevante o no; non contempla il concetto di matching parziale.

Possiamo concludere dicendo che questo modello ha un chiaro formalismo ed è semplice da implementare; di contro il matching esatto richiesto porta ad avere come risultato o pochi documenti o un numero eccessivo.

2.1.2 Vector Model

Il vector model [Salton1971] riconosce che l'uso di pesi binari è troppo limitato e propone un framework nel quale è possible un matching parziale. Per questo introduce dei pesi non binari da assegnare ai termini indice; questi pesi sono usati per calcolare un grado di similarità tra i documenti memorizzati nel sistema e le query utente. I documenti restituiti sono ordinati in maniera decrescente in funzione di questo grado di similitudine. Questo modello quindi prende in considerazione anche documenti che non corrispondono perfettamente con la query utente. Il risultato più importante è che l'insieme dei documenti restituiti risulta più preciso, dal punto di vista delle necessità informative, di quello restituito dal modello booleano.

Definiamo in maniera formale il vector model:

Per il vector model, il peso $w_{i,j}$ *associato ad una coppia* (k_i, d_j)*, dove* k_i *è l'i-esimo termine indice e* d_j *è il j-esimo documento, è positivo e non binario. Inoltre i termini indice nella query sono pesati. Sia* $w_{i,q}$ *il peso associato alla coppia* $[k_i, q]$*, dove* $w_{i,q} \geq 0$*. Il vettore della query* $\vec{q}$ *è definito come* $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q},)$ *dove t è il numero totale di termini indici nel sistema. Il documento* d_j*, è rappresentato dal vettore* $\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$*.*

Quindi un documento d_j e una query q sono rappresentati come vettori t-dimensionali. Questo modello cerca di valutare il grado di similarità del documento d_j rispetto alla query q come la correlazione tra i vettori $\vec{d}_j$ *e* $\vec{q}$*.* Questa correlazione può essere calcolata, ad esempio, utilizzando il coseno dell'angolo formato da $\vec{d}_j$ e $\vec{q}$:

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

dove $\left|\vec{d}_j\right|$ e $\left|\vec{q}\right|$ sono le norme del vettore del documento e della query. Il fattore $\left|\vec{q}\right|$ non ha effetto sull'ordinamento perché è lo stesso per tutti i documenti; tramite il fattore $\left|\vec{d}_j\right|$ si effettua una normalizzazione nello spazio dei documenti.

Fino a che $w_{i,j} \geq 0$ *e* $w_{i,q} \geq 0$, $sim(q, d_j)$ varia tra 0 e +1. Quindi, invece di cercare di stabilire se un documento è rilevante o no, il vector model ordina i documenti in funzione del loro grado di similitudine con la query. Un documento può essere restituito anche se corrisponde solo parzialmente con la query. Prima di calcolare il ranking, dobbiamo capire come ottenere i pesi dei termini indice. Questi pesi possono essere calcolati in vari modi. In [Salton1983a] sono presentate varie tecniche per il calcolo dei pesi. Verrà presentato un approccio generale per l'assegnazione di pesi a dei termini i cui principi base sono legati al clustering.

Data una collezione *C* di oggetti e una *vaga* descrizione di un insieme *A*, l'obiettivo di un algoritmo di clustering è quello di dividere *C* in due insiemi: il primo formato da oggetti *legati* ad *A*, il secondo contenente quelli che non sono relazionati ad *A*. Con il termine *vago* vogliamo dire che non abbiamo un'informazione precisa su quali oggetti appartengono o no ad *A*.

In un problema di clustering bisogna determinare quali sono le caratteristiche che meglio descrivono gli oggetti in *A* e devono essere individuate le proprietà che caratterizzano gli elementi di *A* rispetto ai rimanenti in *C*. Il primo set di caratteristiche fornisce una quantificazione della *similarità intra-cluster* mentre il secondo definisce la *dissomiglianza inter-cluster*. Un algoritmo di clustering efficiente cerca di bilanciare questi due effetti.

Un problema di clustering può essere visto come un problema di information retrieval dove i documenti sono gli oggetti nella collezione *C* e la query utente è una vaga specificazione dell'insieme *A*.

Nel vector model, la similarità intra-cluster è calcolata misurando semplicemente la frequenza di un termine k_i in un documento d_j. Questa frequenza è chiamata *tf factor* e dà una misura di quanto bene un termine descrive il contenuto di un documento. La dissomiglianza inter-cluster è calcolata tramite l'inverso della frequenza di un termine k_i nei documenti della collezione. Questo fattore è chiamato *inverse document frequency* o *idf fator*.

L'importanza dell'idf factor risiede nel fatto che i termini che risiedono in molti documenti non sono utili per discriminare tra documenti rilevanti e non.

Come per gli algoritmi di clustering, un algoritmo efficiente per l'assegnazione dei pesi per l'IR cerca di bilanciare questi due fattori.

Formalizziamo ora le considerazioni fatte precedentemente:

Sia N il numero di documenti nel sistema e sia n_i il numero di documenti nei quali il termine indice k_i appare. Sia $freq_{i,j}$ la frequenza del termine k_i nel documento d_j. La frequenza normalizzata $f_{i,j}$ del termine k_i nel documento d_j è data da: $f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$ dove il massimo è calcolato su tutti i termini contenuti nel testo del documento d_j. Se il termine k_i non appare nel documento d_j allora $f_{i,j}=0$. Inoltre sia idf_i l'inverse document frequency di k_i, data da: $idf_i = \log \frac{N}{n_i}$. L'assegnazione dei pesi ai termini viene data da: $w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$ o tramite variazioni di questa formula. Questa tecnica è chiamata strategia tf-idf.

I maggiori vantaggi del vector model sono:

1. la modalità di assegnazione dei pesi utilizzata aumenta le performance nel processo di retrieval;
2. la strategia usata per il matching parziale permette anche un retrieval dei documenti che approssimano le condizioni di query;
3. la formula di ranking basata sul coseno ordina i documenti in base al loro grado di similarità con la query.

A livello teorico lo svantaggio è che questo modello ipotizza che i termini siano mutuamente indipendenti.

2.1.3 Probabilistic Model

Descriveremo ora il classico modello probabilistico [Robertson1976] conosciuto anche come *binary independence retrieval* (BIR) *model*.

Questo modello cerca di definire il problema dell'information retrieval in un framework basato sul concetto di probabilità.

L'idea di base è quella che, data una query utente, esista un insieme di documenti che contiene solo i documenti rilevanti. Ci riferiremo a questo insieme come all'*insieme di risposta ideale*. Data una descrizione di questo insieme ideale, non abbiamo problemi nel retrieval dei suoi documenti. Per questo possiamo pensare che il processo di query sia definito specificando le proprietà di un insieme di risposta ideale. Il problema è che noi non conosciamo esattamente quali siano queste proprietà. Tutto quello che sappiamo è che esistono termini indice la cui semantica può essere usata per caratterizzare queste proprietà. Siccome queste proprietà sono sconosciute al tempo in cui la query è sottomessa, deve essere fatto un tentativo all'inizio ipotizzando quali potrebbero essere. Questa supposizione iniziale ci permette di generare una potenziale descrizione dell'insieme ideale usata per trovare un primo insieme di documenti. Un'interazione con l'utente permette di migliorare la descrizione iniziale dell'insieme ideale. Questa interazione potrebbe partire da un esame dell'utente dei primi documenti restituiti; in questa fase i documenti vengono etichettati come rilevanti o non. Queste informazioni servono al sistema per raffinare la descrizione; ripetendo questa azione si arriva ad una descrizione precisa dell'insieme ideale.

Il modello probabilistico si basa su una premessa fondamentale:

Principio probabilistico: *data una query utente q ed un documento d_j il modello probabilistico cerca di calcolare la probabilità che l'utente trovi il documento rilevante. Il modello assume che questa probabilità di rilevanza dipenda solo dalla rappresentazione della query e del documento. Inoltre il modello suppone che esista un sottoinsieme dei documenti che l'utente potrebbe indicare come insieme di risposta per la query q. Questo insieme ideale viene chiamato R e dovrebbe massimizzare la probabilità di rilevanza per l'utente. Il sistema ipotizza che i documenti nell'insieme R siano rilevanti per la query; i documenti che non appartengono a questo insieme si ipotizzano non rilevanti.*

Questo principio non crea molti problemi perché in esso non viene esplicitato come viene calcolata la probabilità di rilevanza. Infatti non è specificato neanche lo spazio dei campioni usato per definire questa probabilità.

Data una query q, il modello probabilistico assegna ad ogni documento d_j, come misura della sua similarità con la query, il rapporto $P(d_j$ *rilevante per*

q)/P(d_j non rilevante per q) che calcola la probabilità che il documento d_j diventi rilevante per la query *q*, guardando la probabilità di rilevanza come il valore che minimizza la probabilità di errore nel giudizio [Fuhr1979], [vanRijsbergen1979].

Diamo ora alcune definizioni.

Nel modello probabilistico, i valori assumibili dai pesi dei termini indice sono binari: $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. *Una query q è un sottoinsieme di termini indice. Sia R l'insieme di documenti rilevanti conosciuti (o inizialmente ipotizzati). Sia* $\overline{R}$ *il complemento di R, ad esempio i documenti non rilevanti. Sia* $P(R \mid \vec{d}_j)$ *la probabilità che il documento* d_j *sia rilevante per la query q e sia* $P(\overline{R} \mid \vec{d}_j)$ *la probabilità che* d_j *non sia rilevante per q. La similarità del documento* d_j *per la query q è definita come il rapporto:* $sim(d_j, q) = \frac{P(R \mid \vec{d}_j)}{P(\overline{R} \mid \vec{d}_j)}$.

Usando la regola di Bayes si ha che: $sim(d_j, q) = \frac{P(\vec{d}_j \mid R) \times P(R)}{P(\vec{d}_j \mid \overline{R}) \times P(\overline{R})}$ dove $P(\vec{d}_j \mid R)$ è la probabilità che venga selezionato casualmente il documento d_j dall'insieme *R*. P(R) identifica invece la probabilità che un documento selezionato casualmente dall'intera collezione sia rilevante; il significato di $P(\vec{d}_j \mid \overline{R})$ e $P(\overline{R})$ è analogo e complementare ai precedenti. Dato che $P(R)$ e $P(\overline{R})$ sono uguali per tutti i documenti, possiamo scrivere: $sim(d_j, q) \approx \frac{P(\vec{d}_j \mid R)}{P(\vec{d}_j \mid \overline{R})}$ e assumendo l'indipendenza dei termini indice si ha:

$$sim(d_j, q) \approx \frac{(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid R)) \times (\prod_{g_i(\vec{d}_j)=1} P(\overline{k}_i \mid R))}{(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid \overline{R})) \times (\prod_{g_i(\vec{d}_j)=1} P(\overline{k}_i \mid \overline{R}))};$$

$P(k_i \mid R)$ è la probabilità che il termine indice k_i sia presente in un documento selezionato casualmente nell'insieme *R*, $P(\overline{k}_i \mid R)$ indica la probabilità che k_i non sia presente in un documento scelto casualmente in *R*. Le probabilità associate all'insieme $\overline{R}$ hanno significati analoghi a quelle appena descritte. Utilizzando i logaritmi, ponendo $P(k_i \mid R) + P(\overline{k}_i \mid R) = 1$ e non considerando i fattori costanti per tutti i documenti nel contesto della stessa query, possiamo scrivere:

$sim(d_j, q) \approx \sum_i^t w_{i,q} \times w_{i,j} \times (\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})})$ che è l'espressione usata per il ranking dei documenti nel modello probabilistico.

Dato che all'inizio non conosciamo l'insieme *R* è necessario definire un metodo per calcolare inizialmente le probabilità $P(k_i \mid R)$ e $P(k_i \mid \overline{R})$. Per fare questo esistono differenti tecniche [Baeza1999].

Il vantaggio maggiore del modello probabilistico, in teoria, è che i documenti sono ordinati in maniera decrescente in funzione della loro probabilità di essere rilevanti. Gli svantaggi principali risultano essere la necessità di ipotizzare una iniziale separazione tra documenti rilevanti e non; il fatto che il modello non prende in considerazione la frequenza di occorrenza dei termini indice nel documento; l'ipotesi di indipendenza tra i termini indice.

2.1.4 Fuzzy Set Model

Il rappresentare documenti e query con un insieme di keyword che riguardano solo parzialmente il loro reale contenuto implica naturalmente un matching vago e impreciso. Questo può essere modellato considerando che ogni termine di una query definisce un insieme *fuzzy* ed ogni documento ha un grado di appartenenza, di solito minore di uno, in questo insieme.

Questa interpretazione del processo di retrieval si basa su concetti derivanti dalla teoria fuzzy. Molti modelli sono stati proposti utilizzando questa teoria e noi ne presenteremo uno descritto in [Ogawa1991].

Prima di descrivere il modello dobbiamo dare alcune definizioni fondamentali:

Un sottoinsieme fuzzy A nell'universo del dscorso U è caratterizzato da una funzione membro $\mu_A : U \to [0,1]$ *che associa ogni elemto u di U a un numero* $\mu_A(u)$ *nell'intervallo [0,1].*

Le operazioni più comuni su insiemi fuzzy sono il *complemento*, l'*unione* e l'*intersezione* di uno o più insiemi fuzzy. Di seguito è riportata la loro definizione.

Sia U l'universo del discorso, A e B due sottoinsiemi fuzzy di U, e $\overline{A}$ *il complemento di A rispetto a U. Sia inoltre u un elemento di U. Si ha che:*

$\mu_A(u) = 1 - \mu_A(u)$

$\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$

$\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

Il modello fuzzy qui descritto utilizza un thesaurus. L'idea base è quella di espandere l'insieme dei termini indice nella query con termini a loro relazionati presenti nel thesaurus.

Il thesaurus può anche essere usato per modellare un problema di IR in termini di insiemi fuzzy.

Come definito in [Ogawa1991] un thesaurus può essere costruito definendo una matrice di connessione delle keyword $\vec{c}$ in cui le righe e le colonne sono associate ai termini indice nella collezione dei documenti. In questa matrice un fattore di correlazione tra due termini k_i e k_l può essere definito come: $c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$ dove n_i è il numero di documenti che contiene il termine k_i, n_l è il numero di documenti che contiene il ermine k_l e $n_{i,l}$ è il numero di documenti che li contengono entrambi. Possiamo usare $\vec{c}$ per definire un insieme fuzzy associato ad ogni termine k_i. In questo insieme fuzzy un documento d_j ha un grado di membership definito come:

$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$ che calcola una somma algebrica (qui implementata come il negato di un prodotto algebrico) che viene calcolato su tutti i termini nel documento d_j.

Un documento d_j appartiene all'insieme fuzzy associato al termine k_i se i suoi termini sono legati a k_i.

L'utente rappresenta i sui bisogni informativi utilizzando una query Boolean-like nella che può riguardare congiunzioni o disgiunzioni di insiemi, nel nostro caso insiemi fuzzy.

Per questo il grado di membership in un insieme fuzzy disgiuntivo è calcolato usando una somma algebrica invece della più comune funzione *massimo*, nel caso di insiemi fuzzy congiuntivi si utilizza un prodotto algebrico invece della funzione *minimo*. Questo porta ad una transizione meno drastica nel calcolo dei valori $\mu_{i,j}$ che si ritiene essere una scelta più adeguata nei sistemi di information retrieval.

I modelli basati su insiemi fuzzy per l'information retrieval sono stati ampiamente discussi nella letteratura dedicata alla teoria fuzzy, ma non sono molto popolari nella comunità dell'information retrieval. Inoltre la maggior parte di questi sistemi considera solo piccole collezioni di documenti che rende difficile una comparazione con gli altri modelli.

2.1.5 Modello Booleano esteso

Il modello Booleano classico, anche se semplice ed elegante, non permette di pesare i termini e di generare un ranking nell'insieme dei documenti risultatnte da una query. Un approccio alternativo estende il modello Booleano combinando la formulazione delle query booleane con il modello vettoriale. Il *modello booleano esteso* introdotto in [Salton1983b] è costruito su criteri che hanno come assunzioni di base quelle della logica Booleana. Consideriamo una query booleana congiuntiva data da $q = k_x \wedge k_y$. In accordo con il modello booleano, un documento che contiene sia il termine k_x che k_y è irrilevante come un documento che non gli contiene entrambi. Questo criterio di decisione binaria di solito non segue il senso comune. Considerazioni analoghe si possono applicare alle query puramente disgiuntive.

Quando vengono considerati solo due termini possiamo disegnare query e documenti in uno spazio bi-dimensionale.

Un documento d_j è posizionato in questo spazio considerando i pesi $w_{x,j}$ e $w_{y,j}$ associati alle coppie [k_x, d_j] e [k_y, d_j] rispettivamente; ipotizziamo che questi pesi siano normalizzati e il loro valore vari in [0,1]. Ad esempio, calcoliamo questi pesi utilizzando i fattori normalizzati tf-idf: $w_{x,j} = f_{x,j} \times \frac{idf_x}{\max_i idf_i}$. Per semplicità ci riferiremo a $w_{x,j}$ come x, a $w_{y,j}$ come y e al vettore $\vec{d} = (w_{x,j}, w_{y,i})$ rappresentativo del documento d come il punto $d_j = (x, y)$. Le distanza tracciate in figura possono essere considerate come indici di similarità tra i documenti e le query rappresentate:

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}} \qquad sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Indicato con t il generico numero di termini indice in una collezione di documenti il modello booleano descritto precedentemente può essere esteso considerando distanze euclidee in uno spazio t-dimensionale. Inoltre una generalizzazione più completa può essere ottenuta considerando la teoria delle norme dei vettori.

Il *p-norm model* generalizza il concetto di distanza includendo non solo le distanze euclidee ma anche le p-distanze dove $1 \le p \le \infty$ e un paramentro il cui valore deve essere specificato al tempo di query. Una query disgiuntiva generalizzata è a questo punto rappresentata come: $q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$,

in maniera analoga una query congiuntiva generalizzata è rappresentata come: $q_{or} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$;

la misura di similarità query-documento è quindi data da:

$$sim(q_{or}, d_j) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$sim(q_{and}, d_j) = 1 - \left(\frac{(1 - x_1)^p + (1 - x_2)^p + \ldots + (1 - x_m)^p}{m} \right)^{\frac{1}{p}}$$

La p-norma a due interessanti proprietà:

1. per *p=1* si ha che: $sim(q_{or}, d_j) = sim(q_{and}, d_j) = \frac{x_1 + x_2 + \ldots + x_m}{m}$
2. per $p = \infty$ si ha che: $sim(q_{or}, d_j) = \max(x_i)$ e $sim(q_{and}, d_j) = \min(x_i)$

Quindi per *p=1* le query congiuntive e disgiuntive sono valutate tramite la somma dei pesi dei termini del documento come nella formula di similarità del modello vettoriale (che calcola il prodotto cartesiano). D'altro canto per $p = \infty$ le query sono valutate son il formalismo della fuzzy logic, che può essere vista come una generalizzazione della logica Booleana.

Variando il paramentro *p* tra uno ed infinito si può far variare il comportamento del ranking della p-norma da un approccio vector-like ranking ad uno più vicino a quello Booleano.

Osserviamo in fine che questo modello rilassa l'algebra booleana interpretando le operazioni booleane in termini di distanze algebriche.

2.1.6 Vector Space Model Generalizzato

Nei modelli studiati precedentemente viene ipotizzata l'indipendenza dei termini indice che, per il modello vettoriale è definita come:

Sia $\vec{k}_i$ il vettore associato al termine indice k_i. L'indipendenza dei termini indice nel modello vettoriale implica che l'insieme dei vettori $\{\vec{k}_1, \vec{k}_2, \ldots, \vec{k}_t\}$ è linearmente indipendente è costituisce una base per il sottospazio di interesse. Questo spazio ha dimensione t come il numero dei termini indice nella collezione.

In [Wong1985] è stata proposta un'interpretazione del vector space model nella quale i vettori dei termini indici erano linearmente indipendenti ma non ortogonali a coppie. Tali vettori non formano quindi una base ortogonale per lo

spazio in esame ma sono composti da componenti "più piccole" derivanti da particolari collezioni. Queste componenti sono definite nel seguente modo:

Dato un insieme $\{k_1, k_2, \ldots, k_t\}$ *di termini indice in una collezione; sia* $w_{i,j}$ *il peso associato alla coppia termine-documento. Se i pesi* $w_{i,j}$ *sono tutti binari, allora tutte le possibili combinazioni di termini che co-occorrono nei documenti possono essere rappresentati da un insieme di* 2^t *mintermini dati da* $m_1(0,0,\ldots,0), m_2(1,0,\ldots,0), \ldots, m_{2^t}(1,1,\ldots,1)$*. Sia* $g_i(m_j)$ *il valore ({0,1})del peso del termine indice* k_i *nel mintermine* m_j*.*

L'idea base del modello vettoriale generalizzato è di introdurre un insieme di vettori ortogonali a coppie $\vec{m}_i$ associato con l'insieme di mintermini e viene usato questo insieme di vettori come base per il sottospazio di interesse.

Definiamo il seguente insieme di vettori $\vec{m}_i$

$$
\begin{aligned}
m_1 &= (1,0,\ldots,0) \\
m_2 &= (0,1,\ldots,0) \\
&\vdots \\
m_{2^t} &= (0,0,\ldots,1)
\end{aligned}
$$

dove ogni vettore $\vec{m}_i$ è associato con il rispettivo mintermine m_i

Tali vettori sono per definizione ortogonali a coppie e l'insieme $\vec{m}_i$ è una base ortonormale per il modello vettoriale generalizzato. L'ortogonalità a coppie non implica l'indipendenza dei termini indice, anzi questi termini sono correlati tramite i vettori $\vec{m}_i$. Riflettendo su queste considerazioni ci rendiamo conto che queso modello considera il fatto che la co-occorrenza dei termini indice nei documenti della collezione implica una loro dipendenza.

L'uso della dipendenza tra i termini indice per l'IR è un idea precedente alla formalizzazione di questo modello e il loro effettivo utilizzo per l'aumento delle performance dei sistemi di retrieval continua ad essere una questione controversa nella comunità scientifica.

Per determinare il vettore $\vec{k}_i$ associato al termine indice k_i possiamo sommare i vettori per tutti i mintermini m_r nei quali il termine k_i è 1 e normalizzare:

$$\vec{k}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}} \qquad c_{i,r} = \sum_{d_j | g_l(\vec{d}_j) = g_l(m_r) for \ \ all \ \ l} w_{i,j}$$

per ogni vettore $\vec{m}_r$ è definito un fattore di correlazione $c_{i,r}$.

Il prodotto scalare $\vec{k}_i \bullet \vec{k}_j = \sum_{\forall | g_i(m_r)=1 \wedge g_j(m_r)=1} c_{i,r} \times c_{j,r}$ può essere usato per misurare il grado di correlazione tra i termini indice k_i e k_j.

Le rappresentazioni vettoriali del documento d_j e della query q usate nel modello vettoriale classico possono essere tradotte nel modello vettoriale generalizzato utilizzando le equazioni precedenti. Il calcolo del ranking tra i vettori $\vec{d}_j$ e $\vec{q}$ può essere calcolato utilizzando una funzione di similarità standard basata sul coseno.

2.1.7 Latent Semantic Indexing Model

L'idea principale sviluppata in questo modello [Furnas1988] è quella di mappare ogni vettore rappresentativo dei documenti e delle query in uno spazio dimensionale ridotto associato a dei concetti. Questo si ottiene mappando il vettore dei termini indice in questo spazio più piccolo.

L'obiettivo è quello di avere un retriveal migliore in questo spazio ridotto piuttosto che in quello dei termini indice. Definiamo ora la terminologia di base per questo modello:

Sia t il numero dei termini indice nella collezione e sia N il numero totale di documenti. Definiamo con $\vec{M} = (M_{i,j})$ *la matrice di associazione termine-documento con t righe e N colonne. Ad ogni elemento* $M_{i,j}$ *di questa matrice viene assegnato un peso* $w_{i,j}$ *associato alla coppia termine-documento* $[k_i, d_j]$.

Questo peso $w_{i,j}$ potrebbe essere generato usando la tecnica tf-idf del modello vettoriale classico.

L'indicizzazione proposta in questo modello utilizza una decomposizione basata su valori singolari $\vec{M} = \vec{K}\vec{S}\vec{D}^t$ con:

$\vec{K}$: matrice degli autovettori derivata dalla matrice di correlazione termine a termine;

$\vec{D}^t$: e la matrice degli autovettori che è la trasposta della matrice documento a documento

$\vec{S}$: è una matrice diagonale $r \times r$ dei valori singolari dove $r = \min(t, N)$ è il rango di $\vec{M}$.

Considerando ora solo gli s più grandi valori singolari di $\vec{S}$ e prendiamo le colonne ad essi associate in $\vec{K}$ e $\vec{D}^t$ (i rimanenti valori singolari in $\vec{S}$ vengono

cancellati). La matrice risultante è: $\vec{M}_S = \vec{K}_S \vec{S}_S \vec{D}_S^t$ dove *s*, con *s*<*r*, è la dimensionalità dello spazio concettuale ridotto. Il valore di *s* deve essere scelto in modo da bilanciare due effetti; il dover rappresentare in maniera adeguata e quindi generale i dati reali e il dover filtrare i dettagli non rilevanti.

La relationship tra qualsiasi coppia di documenti è data da:

$$\vec{M}_S^t \vec{M}_S = (\vec{K}_S \vec{S}_S \vec{D}_S^t)^t \vec{K}_S \vec{S}_S \vec{D}_S^t = \vec{D}_S \vec{S}_S \vec{K}_S^t \vec{K}_S \vec{S}_S \vec{D}_S^t = \vec{D}_S \vec{S}_S \vec{S}_S \vec{D}_S^t = (\vec{D}_S \vec{S}_S)(\vec{D}_S \vec{S}_S)^t$$

Nella matrice precedente l'elemento *(i,j)* quantifica la relationship tra i documenti d_i e d_j. Il ranking dei documenti rispetto ad una query utente si ottiene considerando la query come uno pseudo-documento nella matrice $\vec{M}$. Quindi la prima colonna nella matrice $\vec{M}_S^t \vec{M}_S$ fornisce il rank di tutti i documenti in funzione di questa query.

2.1.8 Nural Network Model

Il modello descritto si basa su [Wilkinson1991]. Osserviamo immediatamente che la rete neurale in figura è composta da tre livelli:

- i termini nella query;
- i termini nei documenti;
- i documenti.

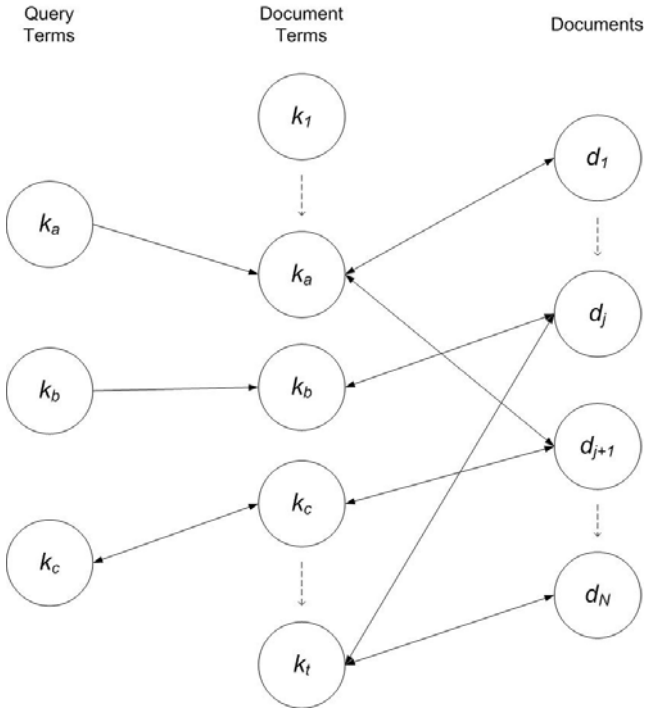


Figura 2.2: Un rete neurale per l'IR

I nodi dei termini della query inviano segnali ai nodi dei termini dei documenti; da qui i nodi dei termini dei documenti generano segnali verso i nodi dei documenti. Questo processo implica dei segnali di ritorno verso i nodi dei termini dei documenti.

In questo scambio l'intensità dei segnali diminuisce fino a che il processo di attivazione non si ferma. Questo processo potrebbe attivare anche un documento che non contiene nessun termine della query.

Ai nodi dei termini della query viene assegnato un livello di attivazione prefissato uguale a 1 (massimo consentito).

Questi nodi inviano segnali al livello intermedio che sono attenuati dai pesi normalizzati dei termini delle query.

Considerando la tecnica di ranking del modello vettoriale i pesi normalizzati sono calcolati tramite la seguente formula: $\overline{w}_{i,q} = \frac{w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$ dove la normalizzazione si effettua utilizzando la norma del vettore della query.

Il segnali che si propagano dal livello intermedio al livello dei documenti vengono attenuati dai pesi normalizzati dei termini dei documenti che, facendo riferimento al modello vettoriale, sono definiti come: $\overline{w}_{i,j} = \frac{w_{i,j}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2}}$ dove la normalizzazione si effettua utilizzando la norma del vettore dei documenti.I segnali che raggiungono il nodo di un documento sono sommati. Pertanto, dopo la prima propagazione, il livello di attivazione del nodo associato al documento d_j è dato da: $\sum_{i=1}^{t} \overline{w}_{i,q}\overline{w}_{i,j} = \frac{\sum_{i=1}^{t} w_{i,q}w_{i,j}}{\sqrt{\sum_{i=1}^{t} w_{i,q}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,j}^2}}$ che è la funzione di ranking del modello vettoriale classico.

Per aumentare le performance del retrieval, la rete continua nel processo di attivazione e propagazione simulando una sorta di user relevance feedback. Per rendere il processo più efficiente viene definita una soglia minima di attivazione, superata la quale il nodo non emette più segnali.

2.1.9 Inference Network Model

Le due scuole di pensiero più affermate riguardo lo studio della probabilità si basano su un punto di vista legato alla *frequenza* e su uno riguardante

l'*epistemologia*. Il primo vede la probabilità in un approccio statistico; il secondo la interpreta come un grado di possibilità (belief) il cui accadimento potrebbe non avere una sperimentazione statistica.

L'inference network model [Turtle1990] [Turtle1991], si basa su un punto di vista epistemologico per affrontare il problema dell'information retrieval. In questo modello vengono associate variabili random ai termini indice, ai documenti e alle query utente. Una variabile random associata al documento d_j rappresenta l'evento di osservazione di quel documento (il modello ipotizza che i documenti siano stati osservati nella ricerca dei documenti rilevanti).

L'osservazione di d_j ipotizza una probabilità sulla variabile random associata ai termini indice. Quindi, l'osservazione di un documento, è la causa per considerare più probabili le variabili associate ai suoi termini indice. I termini indice e le variabili associate ai documenti sono rappresentate come nodi nella rete. Gli archi vanno da un documento ai suoi termini ed indicano che l'osservazione del documento aumenta la possibilità di matching dei sui termini indice con la query.

Le variabili random associate alle query utente modellano l'evento che la richiesta di informazione specificata con la query sia stata soddisfatta. Anche queste variabili random sono rappresentate con dei nodi nella rete. Gli archi di questa rappresentazione vanno dai termini indice alle query.

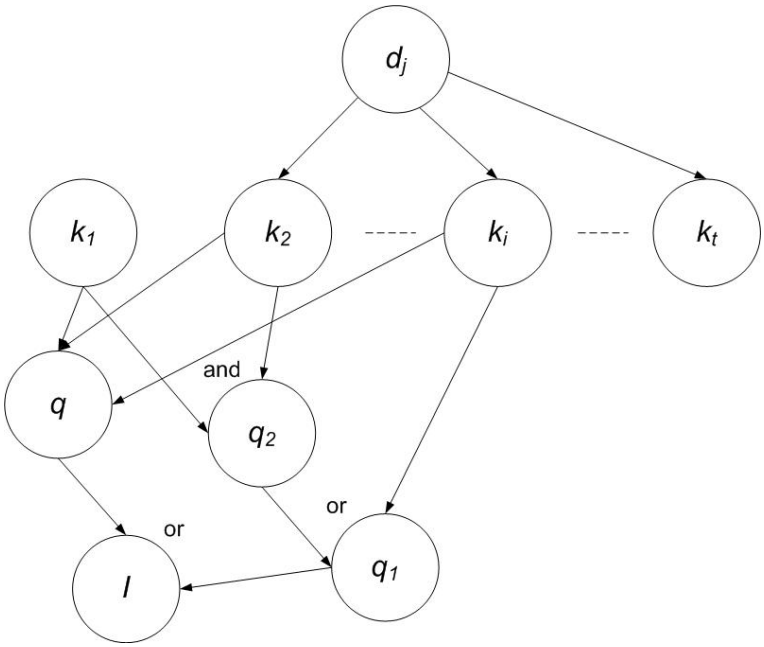


Figura 2.3: Basic inference network model

L'inference network model completo introduce anche nodi testo e nodi associati ai concetti delle query. Viene fatta un'ipotesi aggiuntiva considerando binarie le variabili random nella rete.

Questa ipotesi semplifica il modello ed è abbastanza generale catturando tutte le relationship nel problema di information retrieval.

Diamo ora un definizione:

Sia $\vec{k}$ *un vettore t-dimensionale definito da* $\vec{k} = (k_1, k_2, \ldots, k_t)$ *dove* $k_1, k_2, \ldots, k_t$ *sono variabili binarie random. Queste variabili definiscono* 2^t *possibili stati per* $\vec{k}$*. Inoltre, sia* d_j *una variabile binaria random associata con un documento* d_j *e* q *una variabile binaria random associata con una query* q*.*

Il ranking di un documento d_j in funzione di una query q è una misura di quanto evidentemente la query impone un'osservazione del documento.

In una inference network il ranking di un documento d_j è calcolato come:

$$P(q \wedge d_j) = \sum_{\forall \vec{k}} P(q \wedge d_j \mid \vec{k}) \times P(\vec{k}) = \sum_{\forall \vec{k}} P(q \wedge d_j \wedge \vec{k}) =$$

$$\sum_{\forall \vec{k}} P(q \mid d_j \times \vec{k}) \times P(d_j \times \vec{k}) = \sum_{\forall \vec{k}} P(q \mid \vec{k}) \times P(\vec{k} \mid d_j) \times P(d_j)$$

L'instanziazione del nodo associato al documento d_j (ad esempio l'osservazione di un documento) separa i suoi termini indice figli rendendoli mutuamente indipendenti (si rimanda alla teoria di Bayes per approfondimenti su questa osservazione).

Per questo la probabilità assegnata ad ogni nodo associato al termine indice k_i con l'instanziazione del nodo associato al documento d_j può essere calcolato separatamente.

Questo implica che $P(\vec{k} \mid d_j)$ può essere calcolata in forma di prodotto:

$$P(q \wedge d_j) = \sum_{\forall \vec{k}} P(q \mid \vec{k}) \times \left(\prod_{\forall i \mid g_i(\vec{k})=1} P(k_i \mid d_j) \times \prod_{\forall i \mid g_i(\vec{k})=0} P(\overline{k}_i \mid d_j) \right) \times P(d_j)$$

Attraverso una specificazione appropriata delle probabilità utilizzate nella formula precedente, l'inference network model ricopre un ampio range di strategie per il ranking nei processi di information retrieval.

2.1.10 Belief Network Model

Questo modello, descritto in [Ribeiro1996] si basa su un approccio epistemologico e, partendo dall'inference network, utilizzando uno spazio campione, arriva a formalizzare una differente topologia della rete dividendo la

parte rappresentativa dei documenti da quella delle query. La definizione dello spazio di probabilità utilizzato è stata introdotta per la prima volta in [Wong1995]. Tutti i documenti nella collezione sono indicizzati utilizzando termini indici e l'universo del discorso *U* è l'insieme *K* di tutti i termini indice:

L'insieme $K=(k_1,k_2,\ldots,k_t)$ *è l'universo del discorso e definisce lo spazio campione per il belief network model. Sia* $u \subset K$ *un sottoinsieme di K; ad ogni sottoinsieme u è associato un vettore* $\vec{k} \mid g_i(\vec{k}) = 1 \Leftrightarrow k_i \in u$.

Ogni termine indice è visto come un concetto elementare e *K* è uno spazio concettuale. Un concetto *u* è un sottoinsieme di *K* e potrebbe rappresentare un documento nella collezione o nella query utente. In una belief network le relazioni sono specificate usando variabili random:

Ad ogni termine indice k_i *è associata una variabile random casuale che è anch'essa chiamata* k_i. *Questa variabile è settata ad 1 per indicare che* k_i *è membro di un concetto/insieme rappresentato da* $\vec{k}$.

L'associare i concetti agli insiemi è utile perché ci permette di utilizzare le nozioni logiche di congiunzione, disgiunzione, negazione e implicazione nello stesso modo dei classici operatori dell'algebra degli insiemi come intersezione, unione, complemento e inclusione. I documenti e le query utente possono essere definite come concetti nello spazio campione *K* seguendo la seguente definizione:

Un documento d_j *nella collezione è rappresentato come un concetto composto dai termini che sono usati per indicizzare* d_j. *In maniera analoga una query utente q è rappresentata come un concetto composto dai termini che sono usati per indicizzare q.*

Sia *c* un generico concetto nell spazio *K* rappresentante un documento o una query utente; una distribuzione di probabilità *P* è definita su *K* come: $P(c) = \sum_u P(c \mid u) \times P(u)$ e $P(u) = \left(\frac{1}{2}\right)^t$.

La prima equazione definisce *P(c)* come il grado di copertura dello spazio *K* tramite *c*. Visto che all'inizio del processo il modello non ha informazioni sulla probabilità con la quale un concetto *u* occorre nello spazio *K*, possiamo assumere che ogni *u* è equiprobabile.

Nel belief network model la query utente *q* è modellata come un nodo della rete al quale viene associata una variabile random casuale che chiameremo anch'essa *q*. Questa variabile è settata a 1 ogni volta che *q* ricopre completamente lo spazio concettuale *K*. Quindi quando valutiamo *P(q)* calcoliamo

il grado di copertura dello spazio *K* tramite *q*. In maniera analoga un documento d_j è modellato come il nodo di una rete a cui è associata una variabile binaria casuale che chiameremo anch'essa d_j. Questa variabile vale 1 e sta ad indicare che d_j ricopre completamente lo spazio concettuale *K*. Quando valutiamo $P(d_j)$, calcoliamo il grado di copertura dello spazio *K* tramite d_j.

In accordo con il formalismo descritto, la query utente ed i documenti nella collezione sono modellati come sottoinsiemi dei termini indice. Ognuno di questi sottoinsiemi è interpretato come un concetto racchiuso nello spazio concettuale *K* che lavora come uno spazio di campioni. Le query utente e i documenti sono modellati nello stesso modo. Questa osservazione è importante perché in questo modo viene definita la topologia di una belief network. Nella figura seguente è rappresentata un esempio di belief network.

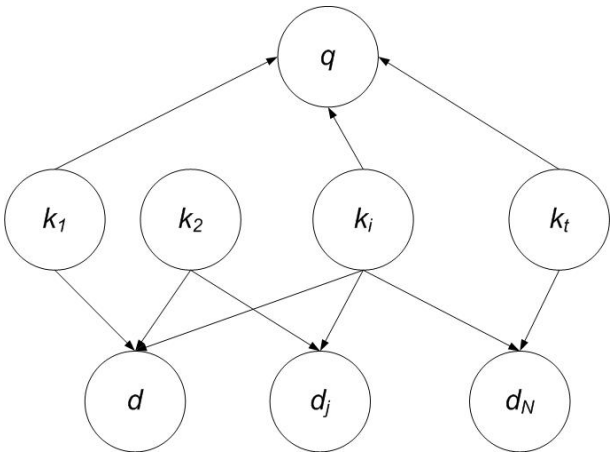


Figura 2.4: Basic belief network model

Una query *q* è modellata come una variabile binaria casuale che è puntata da un termine indice nodo che compone il concetto espresso dalla query.

I documenti sono trattati nello stesso modo (entrambi sono concetti nello spazio *K*). Il nodo che rappresenta il documento è puntato da un termine indice nodo che compone il documento.

Il ranking di un documento d_j relativo ad una data query *q* è interpretato come una relazione tra matching di concetti e riflette il grado di copertura del concetto q su un concetto d_j.

Nel belief network model viene ipotizzato che $P(d_j|q)$ è il rank del documento d_j in funzione della query *q*. Questa probabilità può essere scritta come $P(d_j \mid q) = \sum_{\forall u} P(d_j \wedge q \mid u) \times P(u)$.

In questo modello l'instanziazione delle variabili dei termini indice separa logicamente i nodi *q* e *d* rendendoli mutuamente indipendenti.

Quindi $P(d_j \mid q) = \sum_{\forall u} P(d_j \mid u) \times P(q \mid u) \times P(u)$ che può essere descritta come

$P(d_j \mid q) = \sum_{\forall u} P(d_j \mid \vec{k}) \times P(q \mid \vec{k}) \times P(\vec{k})$.

Le probabilità $P(d_j \mid \vec{k})$ e $P(q \mid \vec{k})$ possono essere definite in funzione delle diverse strategie di ranking.

2.1.11 Metodo basato su le non-overlapping lists

In [Burkowsky1992a] [Burkowsky1992b], viene proposto un metodo, utile per i documenti strutturati, che considera di dividere il testo di ogni documento in regioni non sovrapposte raggruppate in una lista. Esistono molti modi per dividere un testo in regioni non sovrapposte considerano ad esempio capitoli, paragrafi ecc…; le parti così individuate possono essere raggruppate nella *lista dei capitoli*, delle *sessioni* ecc… ed ognuna di queste è una struttura dati distinta e separata.

La ricerca per termini indice e regioni di testo è possibile tramite la costruzione di singoli inverted file dove ogni componente strutturale è un elemento nell'indice.

Ad ogni elemento è associato una lista di regioni di testo rappresentata come lista di occorrenze. Inoltre queste liste possono essere fuse considerando l'inverted file delle parole nel testo.

2.1.12 Metodo basato sui proximal Nodes

Anche questo metodo, presentato in [Baeza1996] [Navarro1995] [Navarro1997] è utilizzato per i documenti strutturati. In esso è descritta una metodologia per definire strutture di indicizzazione gerarchiche indipendenti sul testo contenuto nei documenti. Ognuna di queste strutture indice è una rigida gerarchia composta da capitoli, paragrafi, pagine e linee che sono chiamate nodi. Ad ognuno di questi nodi è associata una regione di testo; due distinte gerarchie potrebbero riferirsi a regioni di testo sovrapposte. Data una query utente che si riferisce a differenti gerarchie, la risposta del modello è formata da nodi contenuti in una sola gerarchia. Considerando la struttura della gerarchia, possono essere ottenute regione di testo innestate nel'insieme di risposta. Una possibile strategia di query è di attraversare la lista inversa della keyword considerata e per ogni elemento nella lista cercare l'indice di gerarchia per le

sessioni, le sottosessioni (e così via) che contengono l'occorrenza del termine. Una strategia più sofisticata è quella di cercare prima l'indice di gerarchia per la keyword considerata. Questo implica lo scorrimento della gerarchia verso il basso fino a che non viene trovato nessun matching o viene raggiunta la fine della gerarchia. L'ultimo componente strutturale trovato è quello che presenta un matching più accurato. Una volta che si conclude questa ricerca non parte nulla sull'elemento seguente nell'inverted list. Invece viene verificato se il componente precedentemente individuato eguaglia anche il secondo elemento nella lista. Se l'esito è positivo, possiamo concludere che il componente strutturale più grande è stato già individuato. Successivamente viene considerato il terzo elemento della lista e così via. Il processo di querying è accellerato perché viene considerato solo il nodo vicino (*proximal node*).

2.1.13 Flat Browsing

Questo tipo di modello si basa sul desiderio che l'utente ha di *sfogliare* (browsing) lo spazio dei documenti cercando riferimenti interessanti. In questo approccio, tale spazio è piatto e i documenti possono essere rappresentati come punti in uno spazio bi-dimensionale o come elementi in un una lista. Ad esempio l'utente potrebbe voler trovare correlazioni tra documenti vicini o parole di suo interesse. Tali parole potrebbero essere aggiunte alla query originale in modo tale da avere una migliore contestualizzazione. L'utente può scegliere un documento e sfogliarlo in maniera *piatta* ad esempio utilizzando un browser per leggere una pagina web.

2.1.14 Structure Guided Browsing

Per facilitare l'azione di browsing i documenti potrebbero essere organizzati in strutture come ad esempio directory. Le directory sono gerarchie in cui gruppi di documenti sono collegati ad un certo argomento. In questo caso l'utente segue una fissata struttura per il browsing tra documenti. La stessa idea può essere applicata ad un singolo documento.

2.1.15 Hypertext Model

Un concetto fondamentale legato alla nozione di scrittura è il concetto di *sequenza*. Il testo scritto è di solito concepito per essere scritto sequenzialmente. Alcune volte, comunque, l'utente cerca informazioni che sono consequenziali al

testo ma che non sono facilmente catturate attraverso una lettura sequenziale. La soluzione è quella di definire una nuova struttura organizzativa del testo oltre a quella già esistente. Un modo per raggiungere questo obiettivo è quello di progettare un *ipertesto*. Un ipertesto è una struttura con un alto livello di interettività che ci permette di sfogliare un testo in maniera non sequenziale. Esso è formato da nodi che sono collegati da link diretti in una struttura a grafo. Ad ogni nodo è associata una regione di testo che potrebbe essere un capitolo in un libro, una sezione in un articolo, una pagina Web. Il processo di navigazione nell'ipertesto può essere interpretato come l'attraversamento di un grafo diretto. Quando l'iperteso è vasto l'utente potrebbe perdere traccia della struttura dell'ipertesto. L'effetto è che potrebbe partire e prendere una direzione di navigazione sbagliata che lo porterebbe lontano dal suo obiettivo specifico. Quando questo succede si dice che l'utente è *perso nell'iperspazio* [Nielsen1990]. Per questo sarebbe utile che l'ipertesto includesse una mappa ipertestuale che mostri all'utente dove si trova. Questa mappa però non deve essere troppo complessa per non disorientare maggiormente l'utente. Per questo l'ipertesto potrebbe essere organizzato gerarchicamente.

Gli ipertesti hanno gettato le basi dell'*hypertext markup language* (HTML) e l'*hypertext transfer protocol* (HTTP) che hanno originato il World Wide Web.

Si intuisce che gli aspetti trattati fin ora: i modelli di IR (booleano, vettoriale, ecc…), la rappresentazione logica dei documenti (full text, termini indice, ecc…), user task (retrieval e browsing), definiscono dimensioni ortogonali dei sistemi di information retrieval. Per questo alcuni modelli sono più appropriati di altri per alcuni task dell'utente e lo stesso modello di IR può essere usato con differenti rappresentazioni logiche in differenti task utente.

		RAPPRESENTAZIONE LOGICA DEI DEOCUMENTI		
		Index Terms	**Full Text**	**Full Text + Structure**
USER TASK	**Retrieval**	Classic Set Theoretic Algebraic Probabilistic	Classic Set Theoretic Algebraic Probabilistic	Structured
	Browsing	Flat	Flat Hypertext	Structure Guided Hypertext

CAPITOLO 3 SISTEMI PER LA RICERCA SEMANTICA SUL WEB

I nuovi sistemi di ricerca sul web che hanno come scopo l'incremento della rilevanza delle informazioni restituite, possono essere classificati in base all'approccio adottato. Una prima categoria è costituita dai sistemi content-based.

Il metodo basato sul contenuto in primo luogo raccoglie le preferenze esplicitate dall'utente e quindi valuta l'attinenza delle pagine in base alle preferenze dell'utente ed al contenuto. I sistemi Syskill & Webert [Ackerman1997], WebWatcher [Armstrong1995], WAWA [Shavlik1998] e WebSail [ChenZ2000] ricadono in questa categoria.

L'approccio collaborativo determina la rilevanza delle informazioni in base alla somiglianza fra gli utenti piuttosto che la somiglianza tra le informazioni in sè. Sistemi di questo tipo sono ad esempio Firefly e Ringo [Maes1994], Phoaks [Terveen1997] e Siteseer [Bollacker2000]. In più, alcuni approcci ibridi comprendono entrambi i metodi per esempio Fab [Balabanovic1997], Lifestyle Finder [Krulwich1997], WebCobra [deVel1998].

La terza categoria è l'approccio basato su domain-knowledge che utilizza sia le preferenze dell'utente sia una base di conoscenza, organizzata in domini, per migliorare la rilevanza dei risultati di ricerca. Yahoo (www.yahoo.com) ad esempio utilizza quest'approccio, presentando un percorso tassonomico predefinito relativo alla ricerca effettuata. In generale classificare automaticamente le pagine in una tassonomia, sia essa predefinita o dinamicamente generata [ChenH2000], è una tecnica tipica di quest'approccio.

NorthernLight (www.northernlight.com) ad esempio è un motore di ricerca che supporta la generazione dinamica della tassonomia. Usando il servizio Custom Search Folder del NorthernLight, gli utenti possono raffinare la loro query di ricerca, specificando un dominio, il che è molto utile quando il motore restituisce troppe informazioni.

Alcuni sistemi cercano di individuare la domain-knowledge dell'utente in modo più esplicito. Per esempio, Aridor ed altri [Aridor2000] rappresentano il dominio di conoscenza dell'utente come un piccolo insieme di pagine fornite dall'utente stesso. Chakrabarti ed altri [Chakrabarti1999] hanno adottato sia una

tassonomia predefinita (ma modificable) che un insieme di pagine esempio fornite dall'utente come base di conoscenza.

L'ultima categoria di approcci per la ricerca sul web è il metodo basato su ontologie. Le ontologie possono essere definite come la descrizione di un insieme di concetti e delle relazioni semantiche che intercorrono tra di essi. Utilizzando un'ontologia come base di conoscenza è possibile fare in modo che un sistema automatico "comprenda" l'argomento trattato all'interno di una pagina web (topic detection) e quindi possa presentare all'utente soltanto le pagine relative al dominio semantico scelto.

Si stanno attualmente sviluppando ontologie per domini specifici a scopo commerciale e pubblico [Clark1999]. Esempi di questo tipo sono: OntoSeek [Guarino1999], On2Broker [Fensel1999], GETESS [Staab1999] e WebKB [Martin2000].

Poiché la trattazione svolta in questo lavoro si basa su approcci utilizzati nell'ultima categoria descritta, presentiamo di seguito alcuni sistemi che utilizzano l'approccio per ontologie.

2.1 WEBSIFTER II

WebSifter II [Kerschberg2002] incorpora uno schema di valutazione di attinenza delle informazioni centrato sull'utente, che integra gli approcci presentati in precedenza. Il sistema mette a disposizione dell'utente degli strumenti per generare una tassonomia che rappresenta la sua intenzione specifica di ricerca. Questa tassonomia fornisce un contesto per la ricerca. La tassonomia è popolata con le pagine trovate da ricerche condotte per mezzo di più motori di ricerca. La valutazione delle pagine può essere vista come un problema di decision-making, dove un decisore (un utente) deve valutare le varie alternative (pagine web) su criteri selezionati (componenti di valutazione) per il suo problema (intenzione di ricerca dell'utente). La valutazione ed il ranking delle pagine è completata usando tecniche analitiche di decisione su cinque componenti di valutazione pesati dall'utente che rappresentano i differenti criteri di valutazione.

Definizione dell'intenzione di ricerca dell'utente

L'utente inquadra il suo fabbisogno informativo in un contesto, specificando un Weighted Semantic Taxonomy Tree (WSTT), ossia un albero tassonomico semantico pesato. Il WSTT consiste di un insieme di nodi che rappresentano un

concetto all'interno dell'intenzione di ricerca dell'utente. Ogni nodo è pesato con un valore da 0 a 10 per rappresentare l'importanza di questo concetto.

Uno degli svantaggi di utilizzare singoli termini come parole chiave per la ricerca è che un termine può avere significati multipli. Questa è una delle ragioni principali per cui i motori di ricerca restituiscono a volte risultati poco rilevanti.

Per ovviare a questo inconveniente, WebSifter II permette che l'utente espanda e raffini il contesto usando WordNet [Miller1995].

L'utente sceglie i significati per rappresentare ogni concetto del WSTT. Si presuppone che i concetti restanti non siano d'interesse, ottenendo così sia gli indicatori positivi che negativi (Positive Concept Terms and Negative Concept Terms) dell'intenzione dell'utente.

Lo schema del WSTT, è tradotto in query booleane espanse dai termini positivi di concetto. I nodi foglia dell'albero denotano i termini di interesse per utente ed i nodi antecedenti per ogni nodo formano il contesto di ricerca. L'intero albero è trasformato in una serie di query separate.

Il numero di query generate dalle combinazioni di termini prevede la copertura dei risultati possibili. I risultati delle query sono immagazzinati per ulteriori elaborazioni.

Meccanismo di Ranking delle pagine web di WebSifter II

Il ranking delle pagine web trovate durante la ricerca effettuata da WebSifter II comprende la valutazione di più attributi, che riflettono le preferenze dell'utente e la loro concezione di fabbisogno informativo. Il ranking è approcciato come se fosse un problema di decisione in presenza di molteplici variabili. I risultati forniti dai motori di ricerca tradizionali (ricordiamo che WebSifter II è un meta-motore) sono ordinati secondo i criteri di decisione scelti dall'utente, usando le tecniche Multi-Attribute Utility Technology (MAUT) [Klein1994], Repertory Grid [Boose1987], Dimensional Analysis [Martin1991] ed il Analytic Hierarchy Process (AHP) [Saaty1980] su cinque componenti. I valori di rilevanza calcolati da ogni componente sono uniti in un'unica misura di rilevanza.

La componente semantica restituisce l'attinenza del contenuto di una pagina web con l'intenzione di ricerca dell'utente. Il valore della rilevanza semantica di una pagina web per ogni query è calcolato contando il numero di volte in cui un termine compare nella pagina rispetto al numero di termini nella query collegata. I concetti negativi compensano la rilevanza semantica aggiustandone il valore in base ai termini irrilevanti della pagina.

Il componente sintattico misura gli aspetti strutturali della pagina in funzione del ruolo di quella pagina all'interno della struttura di un sito Web. Ciò consente una valutazione della pagina indipendentemente dalla ricerca specifica in corso. Il metodo tiene conto della posizione del documento, del suo ruolo, e la forma del relativo URL [Kim2001].

Il componente per il confronto categorico rappresenta la misura della somiglianza fra la struttura della tassonomia creata dall'utente e le informazioni relative alla categoria restuita dal motore di ricerca per ogni pagina web. Molti popolari motori di ricerca, per esempio Yahoo e Google, rispondono alle query dell'utente non soltanto con una lista di URL ma anche con le informazioni relative alla categoria di ogni pagina web. Queste informazioni aiutano gli utenti a filtrare i risultati. Il componente per il confronto di categoria è progettato per fornire i benefici del filtraggio manuale con mezzi automatici.

Il componente Search Engine rappresenta le predilezioni e la fiducia dell'utente nei confronti dei risultati di uno specifico motore di ricerca. Un valore che indica la preferenza dell'utente è assegnato ad ogni motore di ricerca.

Il numero di richieste per una specifica pagina misura la popolarità. Ci sono parecchi popularity services pubblicamente disponibili, che sono raggiunti dal componente di popolarità.

Per concludere, dopo che i risultati sono stati restituiti, l'utente può rivedere le pagine ordinate ed indicare l'attinenza della pagina con il WTTS. Le informazioni sull'attinenza inserite dall'utente alimentano un meccanismo di apprendimento basato su rete neurale che impara di più sulll'intenzione di ricerca dell'utente, rivalutando e riordinando dinamicamente la lista dei risultati [Kim2001].

L'Architettura del Sistema WebSifter II

WebSifter II è in definitiva un sistema che realizza un agente di ricerca personalizzabile, semantico, basato su una tassonomia. La figura seguente mostra l'architettura generale di WebSifter II. L'elicitor del WSTT guida l'utente nella costruzione dell'albero di tassonomia, nell'assegnazione dei pesi su ogni nodo e nella scelta dei significati dei nodi trovati dall'ontology agent. Il WSTT è memorizzato in formato XML. Lo stemming agent trasforma i termini contenuti in una pagina web in termini "stemmati", ossia condotti alla loro forma normale, ad esempio i termini plurali vengono portati al singolare, i verbi all'infinito etc. Il Search Preference Elicitor cattura le preferenze di ricerca dell'utente. L'utente

esprime la sua preferenza di ricerca assegnando un peso a ciascuno dei componenti di preferenza, delle classi sintattiche e dei motori di ricerca.

Il Search Broker interpreta il WSTT e genera le query, interroga i motori di ricerca popolari ed immagazzina i risultati.

Il Page Request Broker ottiene il contenuto di un URL specifico. Il Web Page Rater valuta le pagine e presenta all'utente i risultati ordinati. Lo user profile-learning agent permette all'utente di fornire un feedback sulla rilevanza delle pagine proposte, impara le preferenze di ricerca dell'utente e aggiorna il profilo dell'utente.

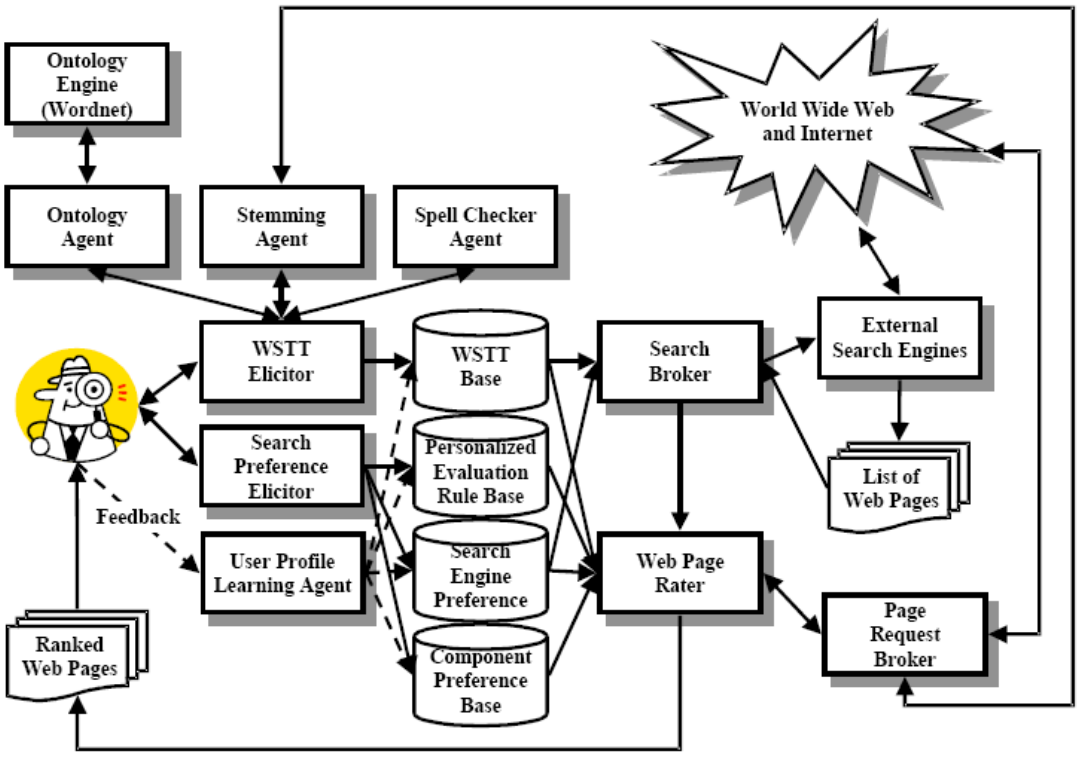


Figura 3.1: Architettura WebSifter II

2.2 IntelliZap

Il sistema IntelliZap [Finkelstein2002] è basato sul paradigma client-server, dove un'applicazione client che funziona sul calcolatore dell'utente cattura il contesto intorno al testo evidenziato dall'utente. Le procedure server-based analizzano il contesto, selezionando le parole più importanti (eliminando così implicitamente eventuali ambiguità sul senso della parola selezionata) e preparano un insieme di query estese per la susseguente ricerca. Il sistema inoltre permette all'utente di modificare l'estensione del contesto che guida una particolare ricerca, variando la quantità di contesto da considerare. Le query che

derivano dall'analisi del contesto sono inviate ad un certo numero di motori di ricerca.

Quando il contesto può essere classificato in modo attendibile all'interno di un insieme predefinito dei domini (quali salute, sport o finanza), vengono inviate delle query supplementari a motori di ricerca specializzati in questi domini. Ciò è fatto per accedere al cosiddetto Web Invisibile, poiché alcuni dei motori a dominio specifico spesso indicizzano dei siti non considerati dai motori di ricerca convenzionali. Un modulo dedicato al re-ranking infine riordina i risultati ricevuti da tutti i motori, in base alla prossimità semantica fra i "riassunti" delle pagine restituiti dai motori ed il contesto originale. A questo scopo viene utilizzata una metrica semantica che, data una coppia di parole o frasi, restituisce un punteggio (normalizzato) che riflette il grado con cui i loro significati sono relazionati.

La significatività di questo nuovo approccio context-based giace nel miglioramento della rilevanza dei risultati di ricerca anche per utenti non esperti di ricerca sul Web. Il sistema ottiene questo risultato, applicando tecniche di elaborazione del linguaggio naturale sul contesto selezionato, al fine di guidare la susseguente ricerca sul testo selezionato dall'utente.

Altri metodi attualmente esistenti prevedono l'analisi dell'intero documento su cui l'utente sta lavorando, o richiedono all'utente di specificare una categoria.

Al contrario IntelliZap analizza automaticamente il contesto nelle vicinanze immediate del testo selezionato. Ciò permette di analizzare appena la giusta quantità di informazioni, senza prendere in considerazione i termini più distanti (e quindi meno relazionati) nel documento di origine. Il sistema raccoglie le informazioni contestuali senza condurre un dialogo esplicito con l'utente.

Di seguito riportiamo una breve descrizione degli elementi fondamentali del sistema IntelliZap, ossia la rete semantica per la misurazione della prossimità semantica, l'algoritmo per l'estrazione delle keyword da utilizzare nelle query ed infine il sistema di re-ranking delle pagine.

La Rete Semantica

Il nucleo della tecnologia di IntelliZap è una rete semantica, che fornisce il supporto per una metrica per la misurazione delle distanze fra coppie di parole. La rete semantica di base è realizzata usando un metodo basato su vettori, in cui ogni parola è rappresentata come un vettore in uno spazio multidimensionale. Per rappresentare ogni parola con un vettore, in primo luogo sono stati identificati 27 domini di conoscenza (ad esempio computer, economia ed

intrattenimento) che approssimativamente fossero partizioni dell'intera varietà di topic possibili.

Per ogni dominio sono stati raccolti circa 10.000 documenti. I vettori relativi ai termini sono stati ottenuti registrando le frequenze di ogni parola in ogni dominio di conoscenza. Ogni dominio può quindi essere osservato come asse nello spazio multidimensionale. La misura della distanza fra i vettori-parola è calcolata usando una metrica basata sulla correlazione:

$$sim(w_1, w_2) = \sum \frac{(\vec{w}_1 - \overline{w}_1)(\vec{w}_2 - \overline{w}_2)}{\sigma_1 \sigma_2}$$

dove $\vec{w}_1$ e $\vec{w}_2$ sono vettori che corrispondono alle parole w_1 e w_2 e $\overline{w}_i$ e σ_i sono stime rispettivamente della loro media e dello scarto quadratico medio. Anche se una metrica siffatta non comprende tutte le proprietà di distanza (si osservi che la diseguaglianza triangolare non è stretta), ha forti motivi intuitivi: se due parole sono usate in domini differenti con un senso simile, queste parole molto probabilmente sono semanticamente correlate.

La rete semantica realizzata su base statistica, descritta precedentemente, viene ulteriormente accresciuta utilizzando le informazioni linguistiche, disponibili attraverso il dizionario elettronico WordNet. Poiché alcune relazioni fra le parole (come iperonimia/iponimia e meronimia/olonimia) non possono essere rilevate usando dati puramente statistici, si utilizza il dizionario di WordNet per correggere la correlazione metrica. È stata sviluppata una metrica basata su WordNet, usando un criterio sul contenuto informativo, e la metrica finale è stata scelta come combinazione lineare fra la correlazione basata sui vettori metrica e la metrica basata su WordNet:

$$sim(w_i, w_2) = \alpha \cdot sim_{VB}(w_i, w_2) + sim_{WN}(w_i, w_2)$$

dove il $sim_{VB}(w_1, w_2)$ e $sim_{WN}(w_1, w_2)$ sono rispettivamente le metriche basate sui vettori e quelle basata su WordNet. I valori ottimali per α e β sono stati ottenuti da training set costituiti da coppie di parole, e sono stati verificati usando una tecnica di cross-validation.

Non essendoci procedure riconosciute per la valutazione delle prestazioni della metrica semantica, il gruppo di sviluppo di IntelliZap ha valutato metriche differenti calcolando la correlazione fra i punteggi delle metriche ed i punteggi assegnati dall'uomo per una lista di coppie di parole. L'intuizione che sta dietro questo metodo è che una buona metrica dovrebbe approssimare bene i giudizi umani.

A questo scopo, è stata redatta una lista di 350 accoppiamenti di sostantivi che rappresentano i vari gradi di somiglianza, ed sono stati impiegati 16 soggetti per stimare la "relatedness" delle parole negli accoppiamenti su una scala da 0 (parole completamente indipendenti) a 10 (parole molto correlate o identiche). La metrica basata sui vettori ha raggiunto il 41% di correlazione con i punteggi assegnati dai soggetti umani, e la metrica basata su WordNet ha raggiunto il 39%di correlazione. La combinazione lineare delle due metriche ha raggiunto il 55% di correlazione con i punteggi umani.

L'Algoritmo per l'Estrazione delle Keyword

La procedura utilizza la rete semantica per estrarre le parole chiavi dal contesto che circonda il testo selezionato dall'utente. Queste parole chiavi sono aggiunte al testo per formare una query estesa, che conduca ad un information retrieval guidato dal contesto.

La procedura per l'estrazione delle parole chiave è stata realizzata attraverso un algoritmo di clustering customizzato per lo scopo che effettua una analisi di clustering ciclica e quindi raffina i risultati statisticamente.

Per una query tipica di 50 parole (da una a tre parole nel testo ed il resto nel contesto), l'algoritmo di estrazione delle parola chiave restituisce solitamente tre o quattro cluster. La ratio del processo di clustering è l'identificazione di cluster di parole che rappresentino i diversi aspetti semantici della query. Le parole chiave nel cluster sono ordinate in base alla loro distanza semantica dal testo, in modo che le parole chiave più importanti compaiono prima nel cluster. Vengono quindi costruite delle query specifiche per ogni cluster combinando le parole del testo con alcune delle parole chiave più importanti di ogni cluster. Rispondendo alle query, i motori di ricerca danno dei risultati che coprono la maggior parte degli aspetti semantici del contesto originale, mentre la procedura di re-ranking elimina i risultati irrilevanti.

L'Algoritmo di Reranking di IntelliZap

L'algoritmo di reranking riordina la lista dei risultati, ottenuta dalla fusione dei risultati ottenuti da tutti i motori utilizzati, paragonandoli semanticamente sia con il testo della query sia con il contesto che la circonda. L'algoritmo computa le distanze semantiche fra le parole del testo e del contesto da una parte, e le parole dei titoli e dei riassunti dei risultati dall'altra parte. Il testo, il contesto, i titoli ed i riassunti sono trattati come insiemi di parole.

La distanza (asimmetrica) fra una coppia di tali insiemi è definita canonicamente come una distanza media dalle parole del primo insieme al contenuto del secondo insieme: $dist(S_1, S_2) = \frac{1}{|S_1|} \sum_{t \in S_1} dist(w, S_2)$

Dove la distanza tra una parola ed un insieme di parole è definita come la distanza minore tra questa parola è l'insieme (cioè la distanza con la più vicina parola dell'insieme): $dist(w, S) = \min_{w' \in S} dist(w, w^i)$

La misura della distanza utilizzata in questi calcoli è esattamente la metrica semantica definita precedentemente.

Il ranking finale è dato dalla distanza pesata fra il testo ed il riassunto, il contesto ed il riassunto, il riassunto ed il testo ed il riassunto ed il contesto. I risultati della ricerca sono disposti in ordine decrescente rispetto ai punteggi e la lista di risultati appena realizzata viene presentata all'utente.

Una caratteristica importante della procedura è che le distanze calcolate fra gli insiemi di parole non sono simmetriche, le distanze tra testo, contesto e riassunti è preso con pesi più grandi rispetto alle distanze reciproche. Si osservi che il testo (e, incidentalmente, il contesto) è selezionato dall'utente, mentre i riassunti sono più arbitrari per loro natura perché dipendono da chi li ha creati al momento dell'inserimento nel motore di ricerca. Secondo le formule presentate, il calcolo della distanza fra il testo e un riassunto considera tutte le parole del testo, ma non necessariamente tutte le parole del riassunto. Quindi, assegnando un peso supplementare alle distanze fra testo e contesto ed i riassunti si da maggiore importanza alle parole del testo e del contesto.

2.3 IL SISTEMA DI MOLDOVAN E MIHALCEA

Il sistema di Moldovan e Mihalcea [Moldovan2000] si caratterizza per l'utilizzo di un interfaccia in linguaggio naturale per incrementare la rilevanza dei risultati di una ricerca sul web. L'analisi semantica della query in linguaggio naturale permette di espandere la query sottoposta ai motori di ricerca. Vediamo in dettaglio il funzionamento del sistema.

Architettura del Sistema

La Figura seguente mostra l'architettura del sistema. La domanda o la frase di input, espressa in inglese, in primo luogo è presentata al modulo d'elaborazione del lessico.

I confini della frase e della parola sono individuati attraverso un processo chiamato tokenizzazione. Le parole sono etichettate con l'indicazione della parte del discorso cui appartengono, usando una versione del tagger di Brill.

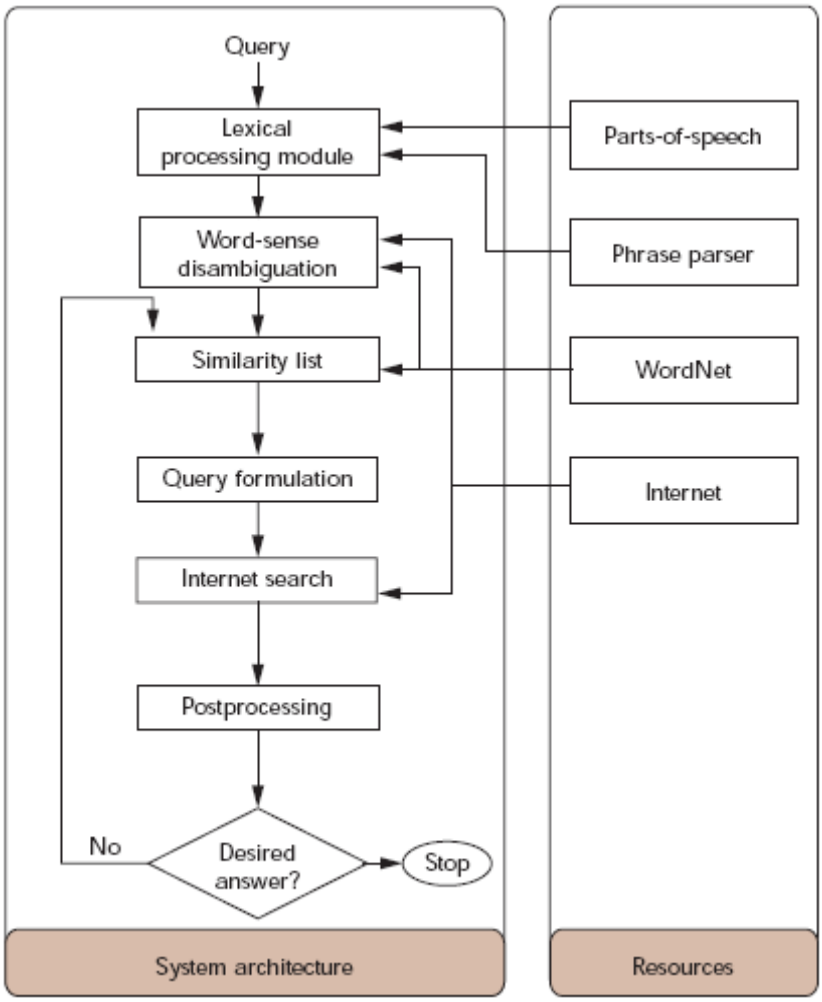


Figura 3.2: Architettura del sistema

Un analizzatore di frase suddivide ogni frase nelle parole e nei verbi costituenti riconoscendone le parole principali. Dopo l'eliminazione delle stopwords (congiunzioni, preposizioni, pronomi e verbi modali), restano alcune parole chiave che rappresentano i concetti importanti della frase in input.

Word-Sense Disambiguation

In questa fase ogni chiave è mappata nella sua corrispondente forma semantica, come definito in WordNet. Questo step consente di abilitare la query expansion basandosi sulla semantica dei concetti anziché le chiavi. Le parole vengono accoppiate, e ogni parola viene disambiguata da una ricerca in Internet con query formulate usando diversi sensi di una parola mantenendo l'altra parola fissata. In questo modo tutte le parole vengono processate e tutti i sensi vengono raccolti. Il passo successivo è quello di raffinare l'ordine dei sensi

ottenuto, con un metodo di densità semantica che tiene conto della distanza semantica tra due o più parole.

Gli algoritmi utilizzati prendono il nome di Contextual Ranking of Word Senses e Conceptual Density Ranking.

Algoritmo Contextual Ranking of Word Senses

Da una coppia di parole semanticamente non taggate (W_1, W_2) ne viene selezionata una, per esempio W_2, e viene creata una lista di similarità per ognuno dei sensi della parola, utilizzando a questo scopo WordNet. Supponendo che W_2 abbia m sensi, questo significa che W_2 compare in m liste di similarità:

$$\left(W_2^1, W_2^{1(1)}, W_2^{1(2)}, ..., W_2^{1(k_1)}\right)$$
$$\left(W_2^2, W_2^{2(1)}, W_2^{2(2)}, ..., W_2^{2(k_2)}\right)$$
....
$$\left(W_2^m, W_2^{m(1)}, W_2^{m(2)}, ..., W_2^{m(k_m)}\right)$$

Dove $W_2^1, W_2^2, ..., W_2^m$ sono i sensi di W_2 e $W_2^{i(s)}$ rappresenta il numero s del senso W_2^i così come definito in WordNet. Si possono quindi formare $W_1 - W_2^{i(s)}$ coppie, e nello specifico:

$$\left(W_1 - W_2^1, W_1 - W_2^{1(1)}, W_1 - W_2^{1(2)}, ..., W_1 - W_2^{1(k_1)}\right)$$
$$\left(W_1 - W_2^2, W_1 - W_2^{2(1)}, W_1 - W_2^{2(2)}, ..., W_1 - W_2^{2(k_2)}\right)$$
....
$$\left(W_1 - W_2^m, W_1 - W_2^{m(1)}, W_1 - W_2^{m(2)}, ..., W_1 - W_2^{m(k_m)}\right)$$

Infine viene effettuata una ricerca su Internet per ogni set di coppie. Le query utilizzano gli operatori forniti da Altavista, che è il motore utilizzato dal sistema per la ricerca sul web, per trovare le occorrenze di W_1 insieme ai sensi di W_2.

Attraverso i risultati delle query, si ottiene il numero di occorrenze per ogni senso i di W_2 e questo fornisce un ranking degli m sensi di W_2 per come si relazionano con W_1.

Algoritmo Conceptual Density Ranking

Questo algoritmo calcola la densità concettuale di una coppia verbo-sostantivo, costruendo un contesto linguistico per ogni senso del verbo e del sostantivo e misurando il numero di sostantivi condivisi tra i contesti del verbo e del sostantivo. In WordNet ogni concetto ha un gloss che funge da microcontesto per quel concetto. Questa fonte, ricca di informazioni linguistiche, è utile nella determinazione della densità concettuale fra le parole, benché si possa applicare soltanto agli accoppiamenti verbo-sostantivo e non agli aggettivi o agli avverbi.

Per ogni coppia verbo-sostantivo v_i-n_j, la densità concettuale C_{ij} è calcolata in questo modo:

1. Si estraggono tutti i gloss dalla sottogerarchia che comprende v_i (la sottogerarchia del verbo v_i è presa dal più alto iperonimo h_i del verbo).
2. Dai gloss si estraggono i sostantivi. Questi costituiscono il contesto del verbo. Ogni sostantivo è immagazzinato insieme ad un peso w che indica il livello, all'interno della sottogerarchia del verbo, nel cui gloss è stato trovato il sostantivo.
3. Si estraggono i sostantivi dalla sottogerarchia in cui è contenuto n_j.
4. Si determina la densità concettuale C_{ij} dei concetti comuni fra i sostantivi ottenuti al punto 2 e quelli ottenuti al punto 3 usando la metrica: $C_{ij} = \frac{\sum_{k}^{|cd_{ij}|} w_k}{\log(discendenti_j)}$

Dove $|cd_{ij}|$ è il numero di concetti comuni tra le gerarchie di v_i e di n_j; w_k rappresenta il livello dei sostantivi nella gerarchia del verbo v_i;e i $discendenti_j$ sono il numero totale di parole nella gerarchia del sostantivo n_j.

Data la densità C_{ij}, l'ultimo step dell'algoritmo è il ranking di tutte le coppie v_i-n_j per ogni i e per ogni j.

Query Expansion

L'operazione di query expansion avviene con l'espansione dell'interrogazione usando termini simili a quelli usati nella query di partenza. Con l'aiuto di WordNet vengono trovate parole semanticamente relazionate ai concetti della query di partenza, in particolare possiamo ottenere la somiglianza semantica tra parole appartenenti allo stesso synset. La query viene espansa usando operatori Booleani.

Postprocessing

L'approccio per filtrare i documenti è diviso in due fasi : si effettua prima una ricerca in Internet usando gli operatori (AND,OR), dopodichè si utilizza un altro operatore per scremare i documenti. Tale operatore lavora come una AND, solo che cattura esclusivamente le parole contenute in n paragrafi consecutivi. Chiaramente l'utilizzo di tale operatore presuppone che i documenti siano, in qualche maniera, segmentati in paragrafi.

2.4 SCORE

Il sistema SCORE (Semantic Content Organization Retrieval Engine) [Sheth2002] fornisce gli strumenti per la definizione delle componenti delle ontologie che gli agenti software del sistema usano per analizzare i documenti. Questi agenti usano regole basate su espressioni congiunte a varie tecniche semantiche di estrazione di metadata da documenti strutturati o semi strutturati.

SCORE supporta alcune fondamentali proprietà che costituiscono il cuore della sua tecnologia: Organizzazione Semantica e uso di Metadata. Una ricerca semantica spesso prevede l'utilizzo di ontologie per organizzare concetti e domini, così come i metadata per annotare ed arricchire il contenuto. I metadata possono essere sintattici o semantici. I primi offrono una descrizione della struttura del documento. I metadata semantici invece offrono informazioni sul dominio specifico a cui appartiene il documento.

- **Normalizzazione Semantica**: la normalizzazione gioca un ruolo molto importante in accordo con l'eterogeneità associata a più sorgenti di dati.
- **Semantic search**: gli attuali motori di ricerca non considerano il contesto della query. Annotazioni semantiche o metadata forniscono un supporto molto importante alla risoluzione di questo problema.
- **Semantic Association**: un generico motore di ricerca fornisce documenti che hanno un qualche grado di familiarità con la keyword specificata nella query. La differenza con la soluzione semantica è che il documento recuperato,certamente appartiene al contesto specificato.

Architettura del Sistema SCORE

La tecnica di classificazione automatica usata aiuta a clusterizzare i documenti in una o più categorie ed estrae (o deduce) i semantic metadata corrispondenti ad uno o più contesti.

L'ontologia è divisa in due componenti relazionati chiamati rispettivamente WordModel (per la definizione dei componenti) e Knowledge-base (che riflette il sottoinsieme di parole con cui un documento composto).

Le operazioni di SCORE coinvolgono tre indipendenti attività. La prima attività definisce WordModel e Knowledge base. Il content processing è la seconda attività che include la classificazione e l'estrazione di metadati dal contenuto. Il risultato è organizzato in accordo alla definizione del WordModel e immagazzinato in un Metabase. Infine troviamo il supporto per applicazioni semantiche.

Per quanto concerne l'estrazione di metadata, SCORE usa i seguenti passi :

- Estrae metadati scorrendo il testo non strutturato servendosi della struttura del contenuto;
- Identifica i metadati sintattici e semantici;
- Migliora l'estrazione di informazioni usando la Knowledge base;
- Aggiorna la Knowledgebase, evitando il problema di dizionari obsoleti.

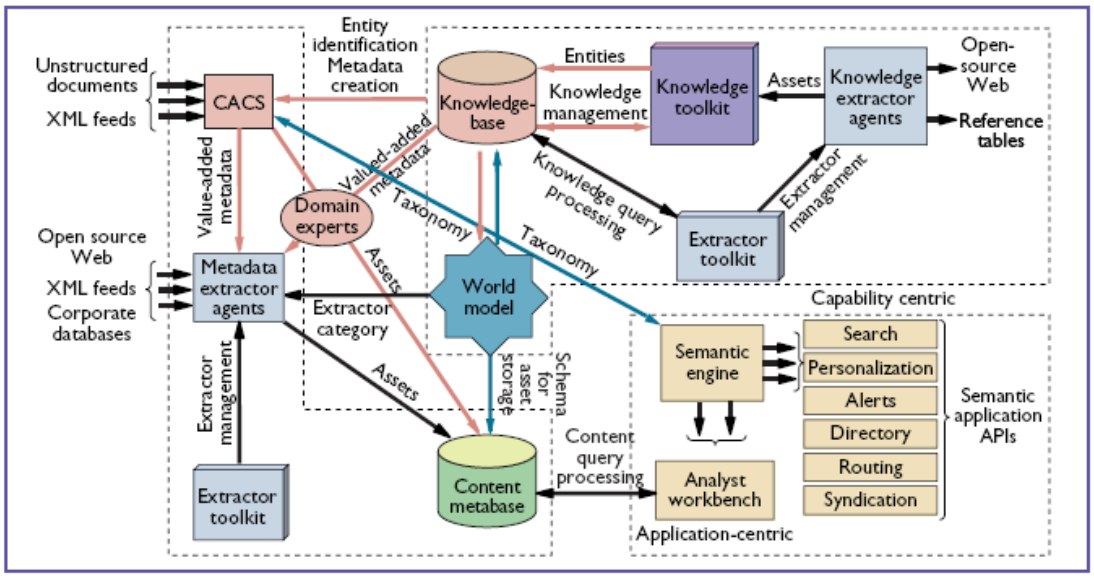


Figura 3.3: Architettura di SCORE

SCORE inoltre utilizza due metodi per risolvere le ambiguità : classificazione e Knowledge base.

Il metodo basato sulla classificazione associa un insieme di entità e classi di entità a categorie di documenti. Mentre il metodo basato sulla Knowledge base risolve le ambiguità facendo leva sul fatto che le entità in un documento sono spesso relazionate ed esse possono anche appartenere alla stessa classe di entità. Combinando questi metodi SCORE assegna un peso a ogni entità e a ogni posizione nel testo nella quale c'è stato un riscontro.

2.5 LaSIE

Diversamente da molti sistemi, che scremano i testi ed usano ampie raccolte di campioni superficiali specifici di un certo campo , il LaSIE (Large Scale Information Extraction) System [Gaizauskas1997] propone un'analisi del testo più completa, dapprima traducendo singole frasi in forma quasi logica, poi costruendo un modello del discorso.

A sostenere il sistema è un modello di mondo generico, rappresentato come una rete semantica, che viene ampliata durante lo svolgimento di un testo aggiungendo le frasi e le classi descritte in quel testo.

Architettura del sistema

LaSIE è stato creato come sistema di ricerca per fini generici, inizialmente adattato soprattutto, ma non esclusivamente, per svolgere i compiti specifici in MUC-6: ricognizione della named entity, risoluzione del coriferimento, riempimento degli elementi del template. Inoltre, il sistema può generare un piccolo sommario relativo allo scenario che ha rilevato nel testo.

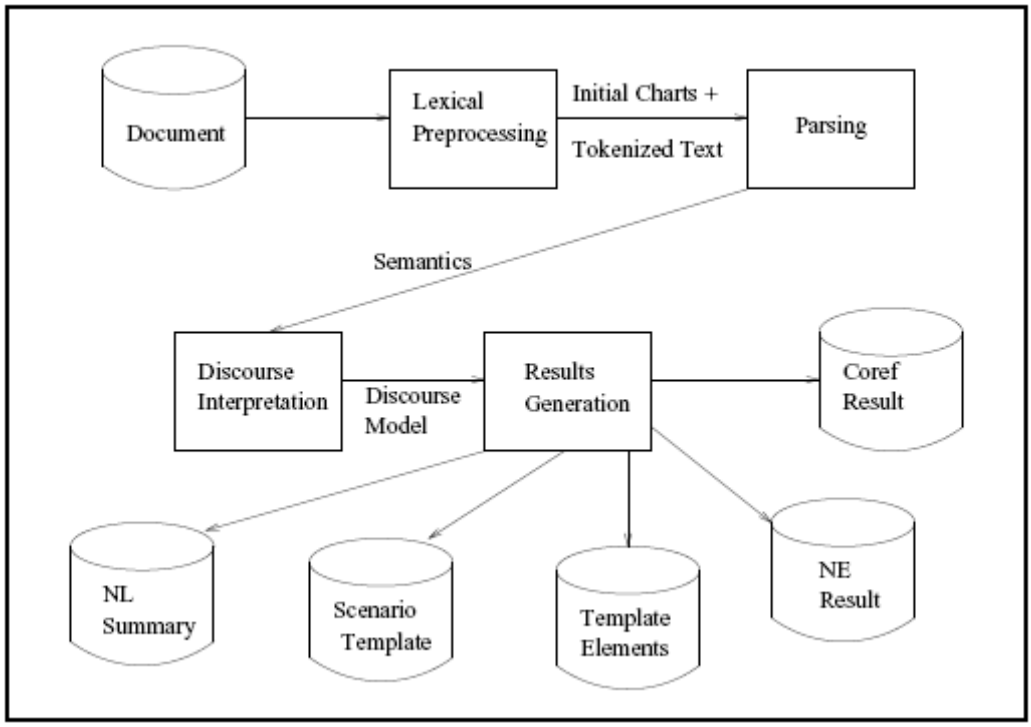


Figura 3.4: Architettura sistema LaSIE

Tutte queste funzioni sono svolte costruendo un modello del discorso da cui vengono letti i diversi risultati. Il sistema ha una architettura reticolare che elabora un testo una frase alla volta e consiste in tre fasi principali di procedura: pre-processamento lessicale, analisi e interpretazione semantica, interpretazione del discorso. Gli apporti generali di queste fasi possono essere descritti in breve come segue:

- il procedimento lessicale legge e trasforma in simboli il testo input abbozzato, etichetta i simboli con parti-del-discorso, esegue un'analisi morfologica, trova delle corrispondenze di espressione tra liste di nomi propri, e costruisce dei limiti lessicali e di espressione in una modalità basata sulla caratteristica che l'analista può tenere a portata di mano;

- il parsing viene fatto in due passaggi, un primo step che usa una speciale grammatica di named entity, e un altro con una grammatica generale e, dopo aver selezionato la "migliore analisi", passa ad una rappresentazione dell'argomento del predicato per la frase in atto;
- l'interpretazione del discorso aggiunge informazioni nel suo predicato in input rappresentato da una rete semantica strutturata. Le interpretazioni semantiche sono assegnate ad ogni frase in un testo durante l'analisi con ciò che è essenzialmente un metodo compositivo classico, ogni regola per la struttura della frase ha una regola semantica corrispondente che specifica il modo in cui deve essere costruita una rappresentazione semantica.

Il World Model

Il Word Model è composto da un'ontologia e da una base di conoscenza degli attributi. L'ontologia è un grafico aciclico diretto con un unico nodo top. I nodi nel grafico sono di classi o di frasi, dove i nodi di frasi sono presenti solo come nodi foglia. Nessun nodo foglia può essere sottoclassificato attraverso la dimensione n, cioè può essere alla radice degli alberi. Ciascuno di questi alberi si divide in rami che si escludono a vicenda , in questo modo, mentre nessun nodo può essere immediatamente dominato da nodi multipli, due di questi nodi non possono essere l'uno alternativo all'altro nella stessa dimensione di classifica.

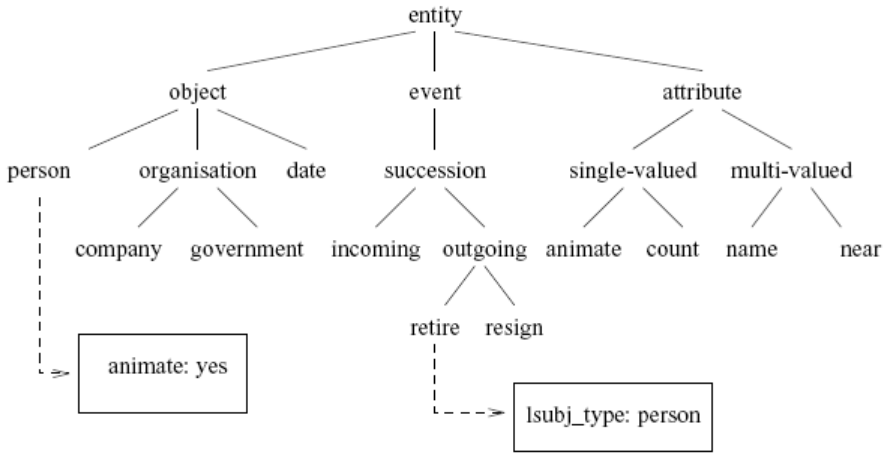


Figura 3.5: Un Frammento dell'Ontologia di LaSIE

Ad esempio, i vini possono essere classificati per colore e nazionalità, in modo che un dato vino possa essere bianco e francese (dominato dai nodi bianco e francese), ma non può essere sia rosso che bianco (dominato da due nodi nella stessa dimensione di classifica). Il Word Model descritto qui può essere considerato una cornice vuota a cui viene aggiunta la rappresentazione

semantica di un particolare testo, occupandola con classi e frasi menzionate nel testo. Il modello del mondo che ne risulta è quindi un modello specializzato per il mondo come descritto nel testo corrente, ci riferiamo a questo modello specializzato come al modello del discorso.

Viene inoltre utilizzato WordNet per tentare di produrre un modello di mondo più generico.

2.6 IL SISTEMA DI ROCHA, SCHWABE E DE ARAGAO

Il sistema di ricerca semantica proposto in [Rocha2004] si fonda sull'idea di arricchire il processo di ricerca per le applicazioni ipermediali con le informazioni estratte da un modello semantico del dominio dell'applicazione. Una delle novità nella ricerca semantica proposta è la combinazione delle tecniche di spread activation con le tecniche tradizionali dei motori di ricerca per ottenere i relativi risultati.

La procedura di Spread Activation funziona principalmente come un esploratore di concetti. Dato un insieme iniziale di concetti attivati e di alcune limitazioni, l'attivazione attraversa la rete raggiungendo altri concetti che sono strettamente collegati ai concetti iniziali. Questa tecnica è molto potente nell'effettuare le ricerche di prossimità, dove, dato un insieme iniziale di concetti, la procedura restituisce altri concetti che sono fortemente collegati ai primi. Una descrizione delle tecniche di spread activation è presentata in [Crestani1997].

Architettura del sistema

L'architettura generale del sistema è indicata nella figura seguente. I primi due punti sono comuni ai metodi di ricerca tradizionali: l'utente esprime la sua query in termini di parole chiavi che sono date in pasto ad un motore di ricerca tradizionale. Questo motore di ricerca ha accesso a tutti i nodi esistenti contenuti nella knowledge base. Per far questo, per ogni nodo nella knowledge base, è generato nel grafico delle istanze un nodo, che è risultato della concatenazione dei valori di tutte le relative proprietà. La ricerca tradizionale avviene sul contenuto dei nodi del grafico delle istanze. La Figura 3.6 mostra il processo di generazione del grafico delle istanze. Dalla figura possiamo vedere che per ogni istanza nell'ontologia, la relativa rappresentazione nel grafico delle istanze contiene tutte le relative proprietà.

Il risultato fornito dal motore di ricerca tradizionale è un insieme di nodi istanza ordinati in base alla loro somiglianza con la query. Questo insieme di nodi

è fornito alla procedura di Spread Activation come l'insieme iniziale dei nodi per la propagazione. In più, vengono anche utilizzate le informazioni d'ordinamento fornite dal motore di ricerca tradizionale. Per ogni nodo, il motore di ricerca tradizionale fornisce un numero reale che misura l'importanza relativa di quel nodo in relazione alla data query. Questo valore numerico è usato come valore iniziale di attivazione per il nodo. Di conseguenza, i nodi che sono allineati meglio degli altri dal motore di ricerca tradizionale avranno priorità nella propagazione poiché l'esplorazione comincierà dai nodi con i più alti valori di attivazione.

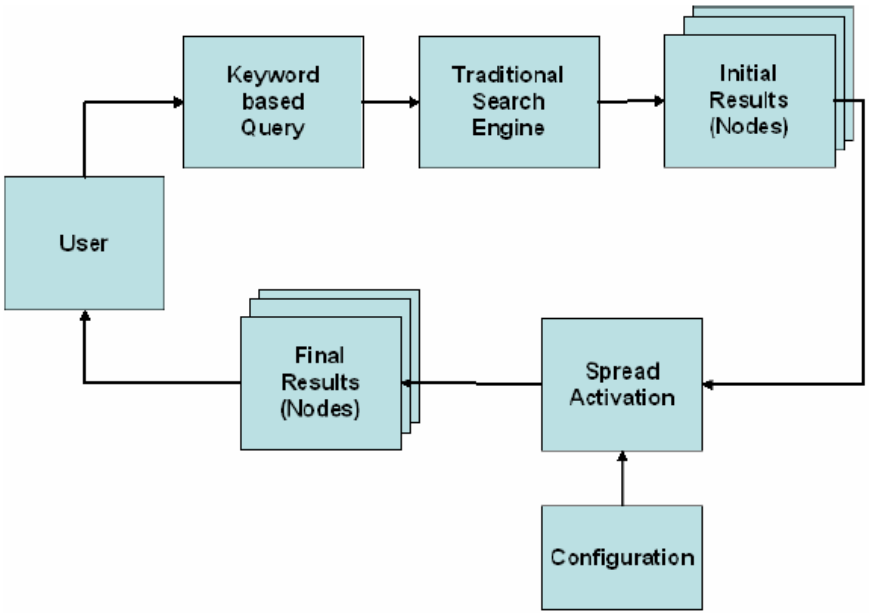


Figura 3.6: Architettura del sistema

La Spread Activation si comporta in conformità con la configurazione specificata usando la semantica di dominio per attraversare i percorsi nel grafico. L'insieme dei nodi ottenuti alla conclusione della propagazione è presentato all'utente come il risultato della ricerca semantica. È interessante notare come la lista dei risultati finali non sarà necessariamente simile all'insieme iniziale dei risultati forniti dalla ricerca tradizionale. I nodi che non sono nella lista iniziale dei risultati potrebbero essere nella lista finale dei risultati e viceversa.

È importante osservare come il transito attraverso un percorso può non sempre essere valido nel modello. Cioè in base alla ricerca che deve essere compiuta dall'utente, potrebbe verificarsi che alcuni percorsi specifici nella rete non debbano essere esplorati. Può anche verificarsi che alcuni percorsi siano più importanti di altri. Questo tipo di conoscenza del dominio semantico deve essere definita nella configurazione della spread activation. Questa configurazione dovrebbe essere fatta da un soggetto specializzato nel dato dominio.

CAPITOLO 4 ONTOLOGIE

L'ontologia, "lo studio dell'essere in quanto essere", diceva Aristotele, è usualmente concepita come una disciplina strettamente filosofica. Eppure, negli ultimi anni grazie all'esplosione delle comunicazioni in rete, gli aspetti ontologici dell'informazione hanno acquistato un valore strategico. Tali aspetti sono intrinsecamente indipendenti dalle forme di codifica dell'informazione stessa, che può essere quindi isolata, recuperata, organizzata, integrata in base a ciò che più conta: il suo contenuto [Turco2002].

La standardizzazione dei contenuti dell'informazione risulta oggi cruciale nelle procedure di knowledge representation e retrieval ed è indispensabile per semplificare i processi di comunicazione. In generale, infatti, la mancanza di un'interpretazione condivisa porta ad una limitatezza di comunicazione tra agenti umani o software. Nel contesto della costruzione di un sistema ICT, tale mancanza di comprensione porta a delle difficoltà nell'identificare i requisiti e nel definire le specifiche dei sistema. Molti tools software, metodi di modellazione, paradigmi e linguaggi limitano l'interoperabilità tra i sistemi, il loro riuso e la loro condivisione.

E' proprio per superare i problemi precedenti che si introduce l'ontologia che cerca di eliminare o, almeno, ridurre le confusioni concettuali o terminologiche, in modo da avere un'interpretazione condivisa; in altre parole un vocabolario comune, con un significato per i vari termini su cui tutti sono d'accordo. Sebbene l'ontologia sia nata nell'ambito filosofico, negli ultimi anni, si è affermata una nuova scuola di pensiero, che propone una caratterizzazione logica rigorosa delle categorie ontologiche fondamentali utilizzate nei sistemi informativi, con lo scopo di aumentarne la trasparenza semantica e l'interoperabilità.

Tale approccio coinvolge attività di modellazione concettuale e d'ingegneria della conoscenza in una prospettiva fortemente interdisciplinare. Una definizione del termine "ontologia" largamente adottata, soprattutto nell'ambito delle "artificial intelligence communities", è quella proposta da Gruber [Gruber1993] , secondo cui un'ontologia è una "specifica esplicita e formale di una concettualizzazione condivisa". La *concettualizzazione* si riferisce ad un modello astratto di un qualche fenomeno, avendone identificato i concetti; *esplicita* significa che i tipi di concetti usati e i vincoli sul loro uso sono esplicitamente

definiti; *formale* si riferisce al fatto che l'ontologia dovrebbe essere "machine-readable"; *condivisa* riflette il fatto che l'ontologia cattura la conoscenza consensuale, cioè quella non propria di un individuo, ma accettata da un gruppo.

4.1 DEFINIZIONI DI ONTOLOGIA

Sono state date diverse definizioni dell'ontologia, oltre a quella prettamente filosofica. Di seguito sono state riportate le varie definizioni che sono state date all'ontologia:

"*Un'ontologia identifica i termini basilari e le relazioni di un determinato dominio, definendone in questo modo il vocabolario, e le regole per combinare tali termini e tali relazioni, andando oltre il vocabolario stesso*" [Neches1991]

Tale definizione indica il modo di procedere per costruire un'ontologia: identificare i termini basilari e le loro relazioni; identificare le regole che li combinano; provvedere a definizioni di tali termini e relazioni. In base a tale definizione, un'ontologia non include solo i termini che sono esplicitamente definiti in essa, ma anche quelli che possono essere derivati usando tali regole.

"*Un'ontologia è un insieme di termini gerarchicamente strutturati per descrivere un dominio che può essere usato come fondamenta per una base di conoscenza*" [Swartout1999]

"*Un'ontologia è un mezzo per descrivere esplicitamente la concettualizzazione presente dietro la conoscenza rappresentata in una base di conoscenza*" [Bernaras1996]

"*Le ontologie sono punti di incontro tra le concettualizzazione condivise. Le concettualizzazioni condivise includono frameworks concettuali per modellare la conoscenza di dominio, protocolli specifici per la comunicazione tra gli agenti e accordi circa la rappresentazione di particolari teorie di dominio. In un contesto di condivisione della conoscenza, le ontologie sono specificate nella forma di un vocabolario di termini*" [Gruber1993]

"L'ontologia è la specificazione esplicita di una concettualizzazione" [Gruber1993]

In altre parole è una caratterizzazione assiomatica di una certa concettualizzazione di un linguaggio, usata per esprimere il significato inteso del vocabolario.

Dal nostro punto di vista l'ontologia può essere vista come "l'insieme dei termini e delle relazioni, che denotano i concetti utilizzati in un dominio". Con

ontologia dunque ci si riferisce a quell'insieme di termini che, in un particolare dominio applicativo, denotano in modo univoco una particolare conoscenza e fra i quali non esiste ambiguità poiché sono condivisi dall'intera comunità d'utenti del dominio applicativo stesso.

Un'ontologia può essere vista come un thesaurus arricchito in cui, oltre alle definizioni e alle relazioni tra i termini di un dato dominio (fornite dal thesaurus), viene rappresentata e fornita una conoscenza più concettuale. L'ontologia viene spesso confusa con una knowledge base; in realtà si tratta di due cose differenti; l'ontologia è una knowledge base particolare, che descrive fatti considerati sempre veri dalla comunità d'utenti, mentre la knowledge base può descrivere fatti e asserzioni relative ad un particolare "state of affaire", che in genere non vale sempre.

Le ontologie hanno una funzione vagamente simile a quella degli schemi di basi di dati, ma con le seguenti, considerevoli differenze:

- un linguaggio per la definizione d'ontologie è sintatticamente e semanticamente più ricco dei comuni approcci alle basi di dati;
- un'ontologia deve essere una terminologia condivisa e consensuale poiché è impiegata allo scopo di condividere e scambiare informazioni;
- un'ontologia fornisce una teoria di dominio e non la struttura di un contenitore di dati.

Sebbene l'ontologia non sia l'unico modo di specificare una concettualizzazione, esso ha l'interessante proprietà di permettere la condivisione della conoscenza tra software di IA.

A partire dalle ontologie, si possono costruire agenti che fanno affidamento su di esse. Le ontologie sono quindi progettate allo scopo di condividere la conoscenza con e tra gli agenti. Esistono molti programmi di IA che prendono in esame frasi in linguaggio naturale tentando di ricavarne il significato. Alcuni hanno suggerito l'idea di utilizzare questi approcci per ricavare in maniera automatica il significato di documenti strutturati o semi-strutturati come ad esempio pagine Web. L'ideale sarebbe che un agente intelligente potesse leggere pagine HTML in modo indipendente, deducendo da esse le informazioni contenute e restituirle all'utente, magari nella forma richiesta.

4.2 FUNZIONI DELLE ONTOLOGIE

In base alla definizione d'ontologia e al significato della nozione di concettualizzazione, possiamo dedurre che un'ontologia consiste di:

1. termini generali che esprimono le categorie principali in cui è organizzato il mondo, come cosa, entità, sostanza, uomo, oggetto fisico, ecc. oppure termini particolari che descrivono un dominio d'applicazione specifico (ontologie di dominio);
2. definizione dei termini;
3. relazioni che li associano o impongono particolari vincoli.

L'ontologia svolge quindi una funzione di:

- lessico comune: la descrizione di un ambiente target necessita di un lessico concordato tra le persone coinvolte. Un notevole contributo è dato dai termini contenuti in un'ontologia;
- spiegazione di ciò che è stato lasciato implicito: in tutte le attività umane, si trovano assunzioni e presupposti impliciti, come la definizione di termini comuni e basilari, le relazioni ed i vincoli tra questi e i punti di vista nell'interpretazione dei fenomeni;
- sistematizzazione della conoscenza: richiede concetti/lessico ben stabiliti, in base ai quali le persone descrivono fenomeni, teorie, ecc. Un'ontologia, perciò, fornisce la backbone della sistematizzazione della conoscenza;
- meta-modello: un modello è generalmente un'astrazione di un oggetto reale. Un'ontologia fornisce concetti e relazioni che sono usati come blocchi costitutivi del modello, cioè specifica il modello da costruire, dando le direttive e i vincoli che dovrebbero essere soddisfatti in questo modello.

Un'ontologia può presentarsi in varie forme, ma include sempre un vocabolario di termini e una descrizione dettagliata del loro significato. Per capire bene che cosa si intende per ontologia, occorre innanzi tutto comprendere che cosa sia la concettualizzazione.

4.3 FORMALIZZAZIONE DELLE NOZIONI SULLE ONTOLOGIE

Come detto precedentemente, è di fondamentale importanza che sia ben chiaro il significato dei vari termini che si utilizzano per definire un'ontologia.

Consideriamo il problema seguente: sia *L* un linguaggio logico usato da un certo insieme *V* di simboli, detto il vocabolario di tale linguaggio. Quando un agente *A* usa *L* per qualche motivo, i modelli di *L* così come vengono intesi da *A*, costituiscono un sottoinsieme più piccolo dell'insieme M(L) di tutti i modelli di *L*; tale sottoinsieme viene chiamato la concettualizzazione di *V* secondo *A*. Consideriamo ora due agenti differenti *A* e *B*, che usano lo stesso linguaggio *L*. Per dare lo stesso significato ai vari termini del vocabolario usato, *A* e *B* devono condividere la stessa concettualizzazione, o mettersi d'accordo nell'adottarne una comune che è l'intersezione delle due concettualizzazioni originarie distinte. L'ontologia aiuta a stabilire tale accordo.

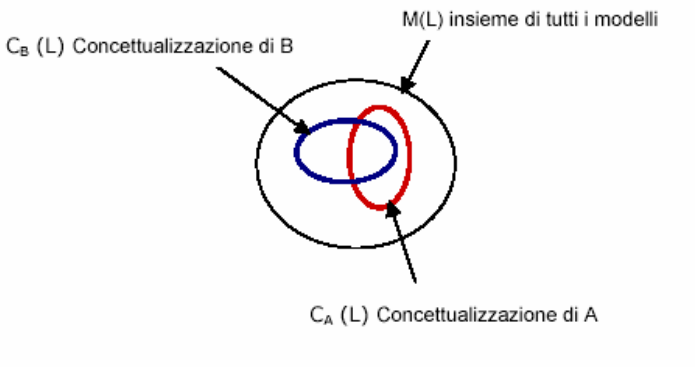


Figura 4.1: Due agenti A e B che usano lo stesso linguaggio L possono comunicare solo se le loro concettualizzazioni CA(L) e CB(L) si sovrappongono

In [Genesereth1987] gli autori hanno introdotto la nozione di concettualizzazione. Per spiegare la loro teorizzazione, viene considerata una situazione in cui due pile di blocchi sono collocate su di un tavolo.

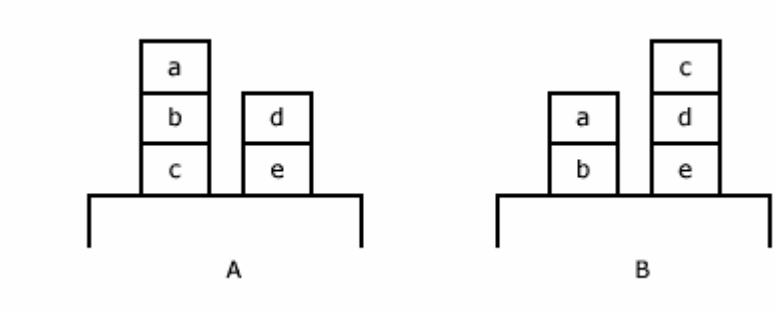


Figura 4.2: Due diverse disposizioni di blocchi su un tavolo

Secondo loro, una possibile concettualizzazione della scena raffigurata sopra è data dalla seguente struttura:

<{a, b, c, d, e }, {su, al di sopra, sgombro, tavolo}>

dove {a, b, c, d, e } è l'universo del discorso, costituito dai cinque blocchi interessati, e {su, al di sopra, sgombro, tavolo} è l'insieme di relazioni

importanti tra i vari blocchi, delle quali le prime due, su e al di sopra, sono binarie, mentre le ultime due, sgombro e tavolo, sono unarie. Genesereth e Nilsson sostengono che (A) e (B) sono due concettualizzazioni diverse; questa visione può andare bene se si è interessati ad una fotografia istantanea del problema; ma nel nostro caso siamo interessati al problema in generale e quindi, poiché il significato dei termini usati per denotare le relazioni rilevanti è sempre lo stesso, come sostiene Guarino [Guarino1995] gli "state of affairs" sono diversi, ma la concettualizzazione è la stessa, poiché è relativa al modo di assegnare una particolare relazione a ciascun simbolo nella nuova situazione.

In altre parole, la concettualizzazione è dunque la struttura formale della realtà così come è percepita e organizzata da un agente, indipendentemente dal vocabolario che viene usato (cioè il linguaggio) e dall'occorrenza attuale di una specifica situazione. Differenti situazioni che riguardano lo stesso oggetto, descritte con diversi vocabolari, possono condividere la stessa concettualizzazione.

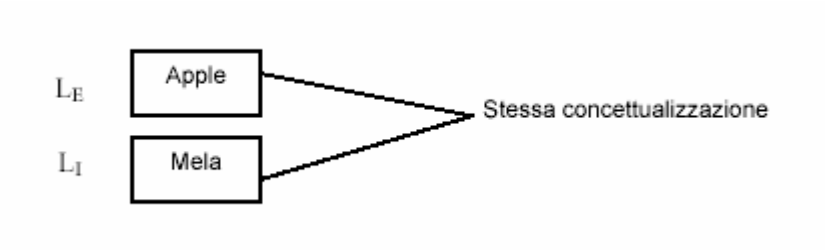


Figura 4.3:Condivisione della stessa concettualizzazione

Possiamo ancora dire che una concettualizzazione è una "vista del mondo", un modo di pensare riguardo un determinato dominio; può essere vista come un insieme di concetti (entità, attributi, processi), le loro definizioni e le loro relazioni. Una concettualizzazione può essere implicita, quando esiste solo nella mente delle persone, o esplicita. Se sorgenti di informazioni differenti sono state progettate e modellate da persone differenti, è molto improbabile che queste persone condividano la stessa concettualizzazione del mondo esterno, cioè non esiste nella realtà una semantica univoca a cui chiunque possa riferirsi. La causa principale delle differenze semantiche si può identificare nelle diverse concettualizzazioni del mondo esterno che persone diverse possono avere. E' possibile classificare le contraddizioni semantiche in tre gruppi principali:

eterogeneità tra classi di oggetti: benché due classi in due differenti sorgenti rappresentino lo stesso concetto nello stesso contesto, possono usare nomi diversi per gli stessi attributi, per i metodi, oppure avere gli stessi attributi

con domini di valori diversi o ancora (dove questo è permesso) avere regole differenti su questi valori;

eterogeneità tra le strutture delle classi: comprendono le differenze nei criteri di specializzazione, nelle strutture per realizzare un'aggregazione, ed anche le discrepanze semantiche;

eterogeneità nelle istanze delle classi: ad esempio, l'uso di diverse unità di misura per i domini di un attributo, o la presenza/assenza di valori nulli.

L'ontologia nasce proprio dal bisogno di superare queste contraddizioni semantiche.

4.4 CLASSIFICAZIONE DELLE ONTOLOGIE

Vi sono diversi tipi di ontologia, che possono variare lungo tre dimensioni chiave: grado di formalità, natura del soggetto e scopo (campi di applicazione delle ontologie) [Uschold1996].

Grado di formalità

Riguarda il grado di formalità in base al quale è creato un vocabolario e vengono specificati i significati dei vari termini; tale grado può essere:

- *altamente informale*: l'ontologia viene espressa in linguaggio naturale;
- *semi-informale*: viene utilizzata una forma ristretta e strutturata di linguaggio naturale, in modo da aumentare la chiarezza e ridurre l'ambiguità.
- *semi-formale*: l'ontologia viene espressa in un linguaggio artificiale, in modo formale;
- *rigorosamente-formale*: i termini vengono definiti in un linguaggio con semantiche formali e teoremi.

Natura del soggetto

Riguarda la natura del soggetto che viene caratterizzato attraverso l'ontologia. Esistono differenti categorie:

- *ontologie di rappresentazione della conoscenza o meta-ontologie*: descrivono le primitive di rappresentazione usate per formalizzare la conoscenza in una base di conoscenza (come concetti, attributi, relazioni, ecc.);
- *ontologie generali/comuni*: includono vocabolari relativi alle cose, eventi, tempo, spazio, causalità, comportamento, funzioni, ecc…;

- *top-level ontologies*: descrivono concetti molto generali (sotto cui sono legati tutti i termini delle ontologie esistenti) come spazio, tempo, materia, oggetto, evento, azione, ecc., che sono indipendenti da un particolare problema o dominio;
- *ontologie di dominio*: descrivono il vocabolario relativo a un generico dominio, le attività presenti in tale dominio, le teorie e i principi elementari che governano tale dominio;
- *ontologie linguistiche*: indipendenti dal dominio;
- *task ontologies*: descrivono il vocabolario relativo ad un generico obiettivo (come la diagnostica e le vendite), dando una specializzazione dei termini introdotti nella top-level ontology;
- *application ontologies*: contengono la conoscenza necessaria per modellare una particolare applicazione.

Questa distinzione tra le varie categorie di ontologia non è mai netta e una stessa ontologia può appartenere a più categorie.

Campi di applicazione delle ontologie

Riguarda i vari campi di applicazione o usi che vengono fatti dell'ontologia. Una lista non esaustiva potrebbe essere:

- Esplicitazione e riutilizzo della conoscenza di dominio
- Comunicazione tra le persone
- Inter-operabilità
- System Engineering
- Supporto al Web Semantico

CAPITOLO 5 MISURE PER LA SEMANTIC RELATEDNESS

In letteratura gli autori usano due termini che sono differenti e qualche volta intercambiabili tra loro per esprimere il concetto di similarità: "semantic relatedness" e "semantic similarity". Resnik [Resnik1995] tenta di dimostrare la distinzione fra i termini mediante un esempio: cars e gasoline, sembrerebbero essere più relazionate rispetto alla coppia cars e bicycles, tuttavia l'ultimo accoppiamento è certamente più simile. La somiglianza rappresenta un caso speciale della relatedness e questa è quello che cercheremo di investigare in questo capitolo. Il concetto di "semantic relatedness" si riferisce in qualche modo ai rapporti esistenti fra le parole ed i concetti. Per misurare la relatedness di due parole, sono state definite varie metriche, che possono essere raggruppate nelle seguenti categorie [Budanitsky1999]:

- Dictionary–based approaches
- Thesaurus–based approaches
- Semantic network–based approaches
- Integrated approaches

Di seguito è riportata una descrizione della metriche basate su questi approcci per la misurazione della semantic similarity basato essenzialmente sul tradizionale utilizzo delle gerarchie lessicali come ad esempio WordNet.

5.1 DICTIONARY-BASED APPROACHES

Nella mentalità più comune, il dizionario è da sempre la risorsa associata ad una conoscenza linguistica. Non deve dunque meravigliare il fatto che ci siano stati tentativi di adattamento di questa risorsa a problematiche relative alla misura della distanza semantica tra parole. Il Longman Dictionary of contemporary English (LDOCE, vedi appendice A) è stato il primo dizionario su supporto magnetico disponibile ai ricercatori. Questo, è certamente il dizionario inglese più usato per il language processing ed in particolar modo è usato per spiegare alcuni concetti che verranno esposti più avanti.

5.1.1 Metodo di Kozima e Furugori: “spreading activation on an English Dictionary”

Dalla definizione data da Kozima e Furugori [Kozima1993], un dizionario ben composto può essere visto come un “closed paraphrasing system of a natural language” : ognuno dei suoi lemmi è definito in funzione di altri lemmi o da esso derivanti o meno. Una strada naturale per trasformare un dizionario in una rete, è quella di creare dei nodi per ogni parola e collegare questi nodi a quelli corrispondenti a tutti i vocaboli incontrati nella sua definizione. Fatto questo, notiamo immediatamente che l’insieme base di lemmi usati dal dizionario, corrisponde alla parte più densa della rete: i rimanenti nodi, che rappresentano i vocaboli al di fuori del set di base, possono essere immaginati come ai margini della rete e sono collegati solo ai vocaboli appartenenti all’insieme di base.

Queste osservazioni, assieme alla struttura del Longman Dictionary of Contemporary English (LDOCE), sono alla base della tecnica di creazione di una rete semantica a partire da un dizionario inglese, ideata da Kozima e Furugori.

Questi ultimi hanno iniziato estraendo da LDOCE solo i vocaboli appartenenti al Longman defining Vocabulary (LDV). Il risultante “sottodizionario” è stato chiamato Glossème. Esso contiene 2851 entries comprendenti in tutto 101861 parole. Ogni entry di Glossème è composta da un lemma , da una word-class (parti di parole) e da una o più unità corrispondenti alle sense-definition entry nel rispettivo LDOCE. Ogni unità consiste in una head-part che corrisponde al genere, ed una o più det-part che corrispondono alla specie. Glossème è stato trasformato in una rete semantica chiamata Paradigme. Quest’ultima conta 2851 nodi, corrispondenti alle entry di Glossème, interconnesse da 295.914 link senza nome. Ogni nodo di Paradigme, include un lemma, una word class ed un activity value ed è collegato ai nodi che rappresentano le parole nella definizione delle entry di Glossème a cui ci si riferisce (approssimativamente abbiamo 104 link per nodo).

I links che partono da un determinato nodo si suddividono in due distinti sets: rèfèrant e rèfèrè. Il primo set non è altro che un collegamento uscente dal nodo; il secondo è un collegamento entrante.

Il rèfèrant contiene un insieme di sotto-set chiamati subrèfèrant, ognuno dei quali corrisponde ad una Glossème unit. Il rèfèrè di un nodo n contiene informazioni relative all’estensione di n.

Una volta costruita la rete , la similarità tra parole di LDV può essere calcolata attraverso le "varie attivazioni" sparse sulla rete. Ogni nodo può tenersi attivo se rappresenta un nodo di passaggio da e verso altri nodi. Il valore di attivazione del nodo n al tempo T+1 è calcolato come segue:

$$v_n(T+1) = \phi * \left(\frac{R_n(T) + R'_n(T)}{2} + e_n(T) \right)$$

dove $R_n(T)$ ed $R'_n(T)$ sono rispettivamente le attività composte , al tempo T, dei rèfèrant e rèfèrè del nodo n, $e_n(T)$ è l'attività distribuita, al tempo T, su n dall'esterno (cioè l'attività assegnata ad n grazie ai nodi esterni), e ϕ è una funzione che limita il valore di uscita nell'intervallo [0,1]. Di quest'ultima funzione Kozima e Furugori non ne forniscono un calcolo preciso. L'attivazione di un nodo n per un periodo di tempo t impone $e_n(T) > 0 \forall t \in [0,t]$ cha causa l'attivazione. Questi risultati in uno schema attivato, arrivano all'equilibrio dopo 10 unità di tempo. Queste considerazioni possono essere usate per stimare il calcolo della similarità tra i nodi-lemma ed altri vocaboli in LDV. L'algoritmo per il calcolo della similarità $sim_{KF}(\omega_k, \omega_t)$ tra due parole ω_k e ω_t consiste nei seguenti passi:

1. Reset degli "Activity-values" di tutti i nodi nella rete.
2. Attivazione del nodo k, corrispondente alla parola ω_k, con strenght $e_k = s(\omega_k)$ per ottenere uno schema attivato $P(\omega_k)$. Il termine $s(\omega_k)$ è il significato di ω_k,definito come "l'informazione normalizzata della parola ω_k" nell'insieme di 5.487.056 parole [West1953].
3. Osservato il valore $a(P(\omega_k), \omega_l)$ -l'activity value del nodo *l* nello schema $P(\omega_k)$ è calcolato come $v_l(10)$ dall'equazione precedente- il valore della similarità cercato è: $sim_{KF}(\omega_k, \omega_t) = s(\omega_t) * a(P(\omega_k), \omega_l)$

La procedura descritta precedentemente definisce una misura della similarità sugli elementi di LDV, $sim_{KF} : LDVxLDV \rightarrow [0,1]^5$ che è solo il 5% di LDOCE. Dunque il prossimo passo della ricerca di Kozima e Furugori è ststa quella di provare ad estendere la misura a $sim_{KF} : LODCExLODCE \rightarrow [0,1]$. Questo è fatto indirettamente estendendo l'equazione che esprime la similarità

precedentemente definita a $sim_{KF} : LDV^n x LDV^m \rightarrow [0,1]$ dove n ed m sono interi positivi arbitrari. Ogni parola nel complemento LODCE di LDV è trattata come una lista $W = \{\omega_1, ..., \omega_r\}$ di parole e la similarità tra le liste W e W' è definita come: $sim_{KF}(W, W'') = \varphi(\sum_{\omega' \in W'} s(\omega').a(P(W), \omega'))$

Qui P(W) è lo schema risultante dall'attivazione di tutti gli $\omega_i \in W$ con strenght $s(\omega_i)^2 / \sum s(\omega_k)$ per 10 passi e φ che limita il valore dell'uscita a $[0,1]$.

5.1.2 Metodo di Kozima ed Ito: "Adaptive Scaling of the Semantic Space"

Attraverso alcune ricerche, Kozima e Ito [Kozima1997] si resero conto che i metodi di Kozima e Furugori [Kozima1993], Morris ed Hirst [Morris1991] ed altri, potevano essere categorizzati come *context-free* o *static,* ossia nella misura della distanza tra parole essi erano indipendenti dal contesto.

Questi allora, basandosi sul lavoro di Kozima e Furugori, idearono una misura di tipo *context-sensitive* o *dynamic*, tenendo conto della *associative direction* per una data coppia di parole.

Nel lavoro svolto da Kozima ed Ito, si rappresenta un contesto tramite un set *C* di parole caratteristiche. Ad esempio C={*car, bus*} impongono una *associative direction* a "*vehicle"* (i set di associazione sono quelli che includono *taxi, railway,airplane,* etc.) mentre C={*car, engine*} impongono una *associative direction* a "*components of car"* (lire seat..etc etc). Denotando con *V* il dato vocabolario, l'obiettivo del metodo può essere espresso come il calcolo della distanza $dist_{KI}(\omega, \omega' | C)$ tra ogni coppia di parole ω, ω' in *V* "sotto il contesto specificato da C". La strategia per il calcolo di $dist_{KI}(\omega, \omega' | C)$ prende il nome di "*adaptive scaling of a semantic space"* nella quale ogni parola in *V* è rappresentata come un vettore multidimensionale. Kozima ed Ito adottarono LDV come loro vocabolario e gli schemi attivati di $P(\omega)$ come vettori.

Attraverso l'attivazione del nodo *s* con ω come lemma, risulta un unico "schema attivo" rappresentabile come un vettore. Dunque si capisce da quanto detto che $P(\omega)$ rappresenta il significato di ω attraverso la sua relazione con il resto di *V*. La distanza geometrica tra $P(\omega)$ e $P(\omega')$ è indicativa della distanza semantica tra ω e ω' ma questa misura è di tipo *static.* Per ottenere una misura

della distanza *context-sensitive*, il vettore P è trasformato in un vettore Q attraverso un'analisi basata sulle componenti principali. Un nuovo sistema di coordinate $X = \{X_1, \ldots\ldots, X_{2851}\}$ è definito in modo tale da fornire al vettore P un sistema di coordinate ortonormali, ordinando gli assi in ordine decrescente in base alla varianza del vettore in questione. Da qui, vengono selezionati i primi m assi $X_1, \ldots\ldots, X_m$, ed ogni $P(\omega_i)$ è proiettato in ognuno di essi. Alla fine i vettori proiettati vengono centrati in modo tale che la loro componente principale sia 0. Poiché la varianza per ogni asse indica la quantità di informazione rappresentata da questi, gli assi in X sono ordinati in ordine discendente del loro significato.

Se tracciassimo la varianza totale in funzione di *m* otterremmo che già una coppia con 100 assi può considerarsi sufficiente per definire quasi metà dell'informazione totale del vettore P.

Il valore esatto m = 281 è ottenuto scegliendo $1 \le m \le 2851$ che rende minima la seguente funzione $noise = \sum_{\omega \in F} |Q(\omega)|$ con F l'insieme di tutte le function words in *V*. Dunque in tal modo riusciamo ad ottenere l'informazione semantica necessaria con un basso valore di rumore, pur riducendo la dimensione dello spazio vettoriale. Poiché, come dimostrato dagli autori, le parole semanticamente relazionate hanno i rispettivi vettori P vicini (o simili) e poiché l'analisi basata sulle componenti principali conserva le loro distanze relative, nel sottospazio semantico dei vettori Q, con una scelta appropriata delle dimension word che sono correlate, si formeranno dei cluster. Questa selezione delle dimensioni appropriate degli assi è realizzata da uno scaling adattativo. Lo spazio semantico è alterato, (aumentato o ridotto) in modo da considerare l'insieme $C = \{\omega_1, \ldots, \omega_n\}$ vicino ad un altro. La distanza $dist_{KI}(\omega, \omega' \mid C)$ tra due parole ω e ω' (con il corrispondente $Q(\omega) = (q_1, \ldots\ldots, q_m)$ e $Q'(\omega) = (q'_1, \ldots\ldots, q'_m)$) è calcolata nella formula seguente:

$dist_{KI}(\omega, \omega' \mid C) = \sqrt{\sum_{i=1}^{m} (f_i(q_i - q'_i))^2}$ con $f_i \in [0,1]$ fattore di scala e definito come:

$$f_i = \begin{cases} 1 - r_i, & r_i \le 1 \\ 0, & r_i > 1 \end{cases} \qquad r_i = SD_i(C) / SD_i(V)$$

Dove $SD_i(C)$ ed $SD_i(V)$ sono rispettivamente le deviazioni standard delle parole in *C* e in *V* proiettate su X_i. Se *C* forma un cluster compatto su X_i, quest'ultimo diventa un asse significativo ($f_i \approx 1$), in caso contrario è insignificante ($f_i \approx 0$).

Dunque il processo dello scaling adattativo , regola la distanza tra il Q-vector ed un dato set *C* , rendendo in tal modo il tutto context-sensitive. Poichè la fase di tuning non è computazionalmente costosa , gli f_i sono gli unici parametri che cambiano da un contesto ad un altro.

5.2 THESAURUS–BASED APPROACHES

Entrambi gli approcci descritti successivamente fanno uso del Roget's Thesaurus (vedi appendice A). Cerchiamo di spendere qualche parola a riguardo. Questa sorgente informativa fu introdotta da Peter Mark Roget [Roget1977] più di 150 anni fa ed è stata sviluppata attraverso un'imponente classificazione di parole e frasi attorno a idee e concetti. I livelli gerarchici utilizzati sono classi, categorie e subcategorie. Una caratteristica molto importante del thesaurus in questione è la presenza di indici, i quali contengono numeri, categorie e labels rappresentative delle categorie per ogni parola. C'è da dire però che come conseguenza della struttura implicita del thesaurus, non possono essere ottenuti valori numerici per il calcolo della semantic distance; potranno piuttosto essere ottenuti solo due valori di tipo booleano, 'close' o 'not close' ossia 'vicino' o 'non vicino'.

5.2.1 Algoritmo di Morris ed Hirst

Lavorando con una versione ridotta del thesaurus di Roget, Morris ed Hirst [Morris1991], identificarono ben cinque tipi di relazioni semantiche tra parole. Nel loro approccio, due vocaboli sono correlati tra loro o semanticamente vicini, se la loro forma base soddisfa una delle seguenti condizioni:

1. se hanno una categoria in comune nel loro insieme di indici
2. se in una sua categoria (individuata da un indice) ci sono puntatori alla categoria dell'altro;
3. se una è in una categoria dell'altra o una è una label nell'indice dell'altra;
4. se sono entrambi contenuti nella stessa sottocategoria;

5. se nelle categorie (individuate dai loro indici) ci sono puntatori a categorie comuni.

Degli esempi tipici, che rientrano in queste categorie, sono moglie e marito, auto ed autista etc. Tra le cinque relazioni enunciate, quelle che trovano maggiori riscontri in esperimenti condotti da ricercatori che utilizzano questo thesaurus come punto di riferimento, sono la 1 e la 2. In aggiunta ai cinque punti esposti sopra, c'è da precisare una cosa, tra l'altro ovvia, e cioè che se due vocaboli sono identici essi sono certamente correlati. Tra l'altro Morrist ed Hirst introducono il seguente concetto di transitività limitata : se la parola *A* è correlata alla parola *B*, la parola *B* con quella *C* e la *C* con un'altra parola *D*, allora *A* è correlata con *C* ma non con *D*.

Morris ed Hirst usano dunque la loro metrica in diversi esperimenti da loro condotti.

5.2.2 Algoritmo di Okumura ed Honda

L'algoritmo di Okumura ed Honda [Okumura1994] è stato derivato dall'algoritmo di Morris ed Hirst. Essenzialmente l'approccio è stato fatto per la lingua giapponese e l'unica differenza è che tra i cinque punti che caratterizzano l'algoritmo precedente, in questo ne viene utilizzato solo il primo, cioè quello secondo il quale due vocaboli sono correlati tra loro o semanticamente vicini se hanno una categoria in comune nel loro insieme di indici.

5.3 SEMANTIC NETWORK-BASED APPROACHES

In accordo con la definizione di Lee [Lee1993], una rete semantica è in generale descritta da una "rappresentazione interconnessa di nodi ed archi, dove i nodi rappresentano concetti ed gli archi rappresentano le varie relazioni tra i concetti". La maggior parte dei metodi discussi più avanti utilizzano il database lessicale Wordnet descritto nell'appendice A.

5.3.1 Metodi basati sulla lunghezza del path

Una strada naturale per valutare la semantic similarity in una tassonomia è quella basata sulla lunghezza del path che unisce due vocaboli. Secondo Resnik [Resnik1995] quanto più è breve il path che congiunge due parole, tanto più queste risulteranno similari. Questo concetto è utilizzato in seguito in alcune metodologie che verranno esposte.

5.3.1.1 Rada et al.'s Simple edge counting

Rada [Rada1989] descrive una metodologia volta a migliorare la ricerca bibliografica in domini specifici, in particolare letteratura biomedica. A differenza di altri approcci che fanno uso del data base lessicale Wordnet, Rada et al. utilizzano come sorgente MeSH (Medical Subject Headings), una rete semantica gerarchica comprendente più di 15000 termini utilizzati in più di cinque milioni di articoli in campo medico. I 15000 termini della rete compongono una gerarchia a 9 livelli che include nodi di livello più alto come anatomia, organismo, malattia ed è basata sulla relazione BROADER-THAN. Quest'ultima è l'esatto opposto di IS-A, ma occasionalmente include dei links opposti a relazioni del tipo PART-OF. Come nel caso di IS-A, le voci BROADER sono poste in alto dell'albero. La principale assunzione fatta da Rada è che il numero di archi tra i termini della gerarchia MeSH è concettualmente una misura di distanza concettuale tra termini. Questa distanza è definita semplicemente come:

$dist_{\text{Re}tal}(l_i, l_j) =$ numero minimo di edges in un path da l_i a l_j

Con questa semplice funzione sono stati ottenuti risultati sorprendenti grazie anche al fatto che il dominio in questione è abbastanza omogeneo ed assicura una altrettanta omogeneità della gerarchia.

5.3.1.2 Hirst and St-Onge's Medium-Strong Relations

Gli studiosi in questione [Hirst1998], affermano che possono esserci tre classi principali di tipologie di relazioni tra i nomi in WordNet. La relazione *extra-strong* si ha tra una parola e la sua ripetizione letterale. Due parole sono *fortemente correlate* se si presenta uno dei seguenti casi:

1. i due vocaboli hanno un synset in comune;
2. i due vocaboli sono associati a due differenti synset che sono connessi da un collegamento orizzontale (ad esempio tramite un link di Antinomia);
3. c'è un tipo di collegamento in cui due synset, sono associati con ogni parola e in aggiunta ogni parola è composta (o è una frase) e include l'altra.

Infine due parole sono correlate in maniera *medium* o regular se esiste un percorso ammissibile che connette i synset associati ad ogni vocabolo. Per percorso ammissibile, si intende un percorso che non ha più di 5 links e si

adegua ad uno degli otto modelli descritti nell'articolo. La giustificazione dei modelli è basata su teorie psicolinguistiche riguardanti l'interazione di generalizzazioni, specializzazioni e coordinazioni. Tutto quello di cui noi abbiamo bisogno è capire se i path ammissibili includono più di un link e le direzioni dei link di uno stesso path possono essere di vario tipo (orizzontali, generalizzazioni -iperonimia e metonimia- e specializzazioni –iponimia e olonimia). Secondo quanto affermato da Hirst e St-Onge's, le relazioni extra strong hanno precedenza su relazioni strong e, quest'ultime, hanno più peso di quelle medium strong. Anche dalla definizione non c'è competizione all'interno delle prime due categorie. Ad ogni path sono assegnati dei pesi dati dalla formula seguente:

$$Weight = C - path\_lenght - k * numero\ di\ cambi\ di\ direzione$$

dove C e k sono costanti, le quali secondo Budanitsky e Hirst [Budanitsky2001] sono settate rispettivamente a 8 ed 1.

Il concetto che sta dietro la formula si basa sul principio secondo il quale più lungo è il path e maggiori sono i cambiamenti di direzione e più piccoli saranno i pesi. L'equazione precedente introduce una funzione di distanze parziali che può essere definita, per esempio, assegnando alle relazioni extra strong il valore di 3C, alle strong il valore di 2C, mentre a quelle medium strong il peso del corrispondente path, e due concetti che non hanno relazioni ricevono un punteggio nullo.

5.3.2 Scaling the network

Malgrado l'apparente semplicità che si riscontra nell'uso delle reti semantiche, un problema ampiamente conosciuto relativo all'approccio "edge counting" è quello per cui in molte metriche esistenti si suppone che i collegamenti tra un nodo ed un altro abbiano lo stesso peso. Questa cosa è certamente non vera anche solo per il fatto che in queste reti ci sono parti più dense di altre e dunque collegamenti in queste zone possono avere un peso di rilevanza diversa rispetto a quelli in parti meno dense.

5.3.2.1 Sussna's Depth-Relative Scaling

Nell'approccio di Sussna [Sussna1993] [Sussna1997] ogni "edge" di Wordnet è composto da due archi rappresentanti relazioni inverse. Ogni relazione r ha un peso o un range di pesi $[\min_r ; \max_r]$:

1. antonimia $\min_r = \max_r = 2.5$.
2. iperonimia , iponimia, olonimia e meronimia $\min_r = 1$ a $\max_r = 2^{14}$.
3. La sinonimia è considerata come relazione intranodo e non ha peso associato.

Dopo aver fatto queste considerazioni, capiamo che la distanza tra due nodi non sarà sempre uguale ad uno ma avrà un certo peso.

I punti dell'intervallo per una relazione *r* dal nodo c_1 *al nodo* c_2 dipende dal numero n_r di archi dello stesso tipo che lasciano c_1:

$$wt(c_1 \rightarrow_r c_2) = \max_r - \frac{\max_r - \min_r}{n_r(c_1)}$$

Questo è il fattore specifico di fanout legato al particolare tipo di relazione *r* che, in accordo con Sussna, riflette la diluizione della potenza di connotazione tra un nodo sorgente ed uno target e tiene in conte della possibile asimmetria tra i due nodi, dove la potenza di connotazione in una direzione è diversa da quella nell'altra. I due pesi inversi per un arco sono mediati e scalati attraverso la profondità *d* dell'arco sull'albero completo della tassonomia. Il motivo fondate per questo scaling deriva dall'osservazione che nodi simili più profondi nell'albero sembrano più correlati di altri che si trovano a livelli più alti. La formula per il calcolo della distanza tra i nodi adiacenti c_1 e c_2 diventa:

$$dist_S(c_1, c_2) = \frac{wt(c_1 \rightarrow_r c_2) + wt(c_2 \rightarrow_{r'} c_1)}{2d}$$

dove *r* è la relazione che lega c_1 a c_2 e *r'* è la sua inversa (la relazione tra e c_2 e c_1).

In conclusione la distanza semantica tra due nodi arbitrari c_i e c_j è calcolata come la somma delle distanze tra la coppia di nodi adiacenti lungo il path più corto che connette c_i a c_j.

5.3.2.2 La similarità concettuale di Wu e Palmer

Wu e Palmer [Wu1994] affrontano il problema della rappresentazione dei verbi nei sistemi computerizzati e il loro impatto sul problema della selezione lessicale in un sistema automatico di traduzione.

L'idea chiave di questa metrica, nella traduzione dei verbi inglesi in mandarino cinese, è "proiettare" verbi semplici e composti di entrambi i linguaggi, in qualcosa che essi chiamano *domini concettuali*. Il primo effetto immediato

dell'operazione di proiezione è la separazione dei differenti significati dei verbi, piazzando questi in differenti domini. Un altro importante aspetto dei domini concettuali è il fatto che il concetto, all'interno di un singolo dominio, può essere organizzato in una stretta struttura gerarchica sulla quale la misura della similarità può essere definita.

Wu e Palmer definiscono la *conceptual similarity* tra una coppia di concetti c_1 e c_2 come : $sim_{WP}(c_1, c_2) = \frac{2 * N3}{N1 + N2 + 2 * N3}$

dove *N1* è la lunghezza (espressa in numero di nodi) del path da c_1 a c_2, dove quest'ultimo è il *lowest common subsumer (lcs)* di c_1 e c_2, *N2* è la lunghezza del path da c_2 a c_3, e *N3* è la lunghezza del path da c_3 alla radice della gerarchia. Notiamo che *N3* rappresenta la profondità globale nella gerarchia. Gli autori infine hanno considerato un fattore di scala per tradurre un linguaggio di similarità in un linguaggio di distanze ottenendo la seguente relazione:

$$dist_{WP}(c_1, c_2) = 1 - sim_{WP}(c_1, c_2) = \frac{N1 + N2}{N1 + N2 + 2 * N3}$$

E' interessante notare che Wu e Palmer descrivono questa metrica relativamente a una tassonomia di verbi, ma essa può essere applicata ugualmente bene ad altre parti del discorso, a patto che i concetti siano organizzati in una gerarchia.

5.3.2.3 La metrica di Leacock e Chodorow

Un metodo intuitivo per la misura della semantic relatedness, incentrata sull'utilizzo di un database lessicale strutturato ad albero, si basa sul conteggio del numero di link tra due gruppi di vocaboli o sinonimi rappresentanti lo stesso concetto. La relazione sarà tanto più grande quanto più piccolo risulterà il percorso che unisce le due parole. La misura proposta da Leacock e Chodorow [Leacock1998] si basa sul considerare un'intera gerarchia di nomi in WordNet nella quale tutti i vocaboli sono collegati ad un unico nodo radice.

Tale affermazione assicura l'esistenza di un percorso che collega ogni coppia di vocaboli e dunque per determinare la semantic relatedness tra due parole ci si riferisce a: $related_{lch}(c_1, c_2) = -\log\left(\frac{schortestpath(c_1, c_2)}{2D}\right)$

dove c_1 e c_2 rappresentano i due concetti, $shortestpath(c_1, c_2)$ rappresenta la lunghezza del percorso più breve tra c_1 e c_2 mentre D è la massima profondità della tassonomia. Questo metodo però assume che la misura dei pesi tra ogni link della tassonomia risulti uguale, assunzione che sicuramente risulta errata in quanto nella gerarchia in questione è possibile trovare link di uno stesso concetto, che risultino lontani a causa di grosse quantità di vocaboli per quel concetto e altri link più vicini ma meno correlati perché appartenenti a concetti diversi.

5.3.2.4 La densità concettuale di Agirre e Rigau

Attraverso alcuni studi, Agirre e Rigau [Agirre1996] [Agirre1997] hanno proposto una misura concettuale della distanza basata su un insieme di parametri:

1. la *lunghezza dello shortest path* che collega i concetti coinvolti;
2. la *profondità della gerarchia*: i concetti appartenenti alla parte più bassa della gerarchia sono considerati più vicini;
3. la *densità dei concetti nella gerarchia*: i concetti in una parte densa della gerarchia sono relativamente più vicini, rispetto ad altri che si trovano in una regione più diradata.

Malgrado questo tipo di idea da loro proposta, non definiscono una formula esplicita della distanza. Essi invece, introducono e sviluppano la nozione di densità concettuale e le cose dette sin'ora potranno essere utilizzate come primo passo utile per determinare la semantic relatedness tra un arbitrario numero di parole.

Vediamo adesso in che modo è stata definita la densità concettuale. Data una sottogerarchia, avente il nodo *c* come nodo radice, che contiene ,tra gli altri, m concetti di interesse, la densità concettuale di *c* rispetto ad m concetti, è definita come: $CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp_c)^i}{\sum_{i=0}^{h-1} (nhyp_c)^i}$

dove $nhyp_c$ è il numero di iponimi per nodo nella sottogerarchia di *c* ed h è l'altezza della sottogerarchia.

Questa formula può essere vista attraverso la figura seguente come un'interpretazione semigeometrica.

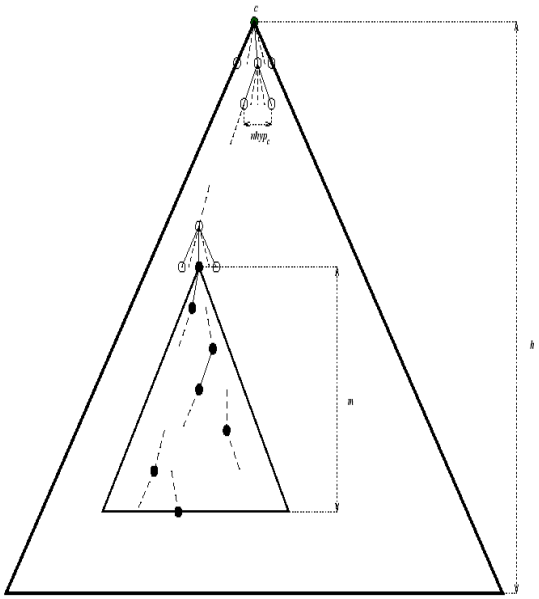


Figura 5.1: Densità concettuale

Se la nostra sottogerarchia fosse un albero $nhyp_c$-ario e l'area della gerarchia fosse calcolata come la quantità di concetti in essa contenuti, allora il denominatore dell'equazione precedente rappresenterebbe l'area della sottogerarchia avente c come radice. In maniera analoga il numeratore andrebbe a rappresentare la più vasta gerarchia minimale per gli *m* concetti.

Da quanto detto capiamo allora che l'equazione precedente vuole esprimere il rapporto tra le aree di due gerarchie: la gerarchia coprente gli m concetti di interesse e l' attuale gerarchia in cui essi si trovano di altezza h.

Tale spiegazione giustifica anche il nome di densità per questa misura. Infatti, la premessa riguardante la perfetta $nhyp_c$-arietà dell'albero non è così irrealistica così come poteva sembrare all'inizio, e il valore di $nhyp_c$ nel metodo in questione viene calcolato per ogni concetto *c* in Wordnet attraverso la seguente equazione: $descendants_c = \sum_{i=0}^{h-1} (nhyp_c)^i$

In questa equazione, $descendants_c$ è il numero di concetti nella sottogerarchia sotto *c* includente *c* stesso. Una volta definita la formula base, Agirre e Rigau, provarono a migliorarne l'efficienza, attraverso l'introduzione di due parametri α e β. In particolare la formula fu modificata come segue:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp_c + \beta)^{i^{\alpha}}}{\sum_{i=0}^{h-1} (nhyp_c)^i}$$

Dopo un certo numero di prove fatte, basate sulla ricerca dei valori di α e β che ottimizzassero questa equazione, si ottenne la migliore performance con α =0.2 e β =0. Di conseguenza la formula finale è la seguente:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp_c)^{i^{0.2}}}{\sum_{i=0}^{h-1} (nhyp_c)^i}$$

Nonostante l'assenza di una specifica formula per il calcolo della semantic relatedness, in base ad un'applicazione fatta da Agirre e Rigau riguardante il problema del Word Sense Disambiguation, è stato dedotta la seguente formula:

$\mathrm{Re}l_{AR}(w_1,...,w_n) = \max_{c \in L} CD(c,m)$ dove $(w_1,...,w_n)$ sono contenute nella finestra *W* ed *L* è l'intero lessico, che nel caso specifico rappresentava la parte nominale di Wordnet.

5.3.3 La misura di Li, Bandar e McLean

Il metodo proposto da Li, Bandar e McLean [Li2000] per il calcolo della semantic similarity si basa essenzialmente sul tradizionale utilizzo delle gerarchie lessicali come ad esempio WordNet. Nel loro articolo mostrano diverse strategie di calcolo in ognuna delle quali è considerato parte o tutto l'insieme di alcune variabili ritenute fondamentali per lo studio di tale problematica. In altre parole essi considerano la seguente dipendenza funzionale: $sim(w_1,w_2) = f(l,h,d)$ dove l , è lo shortest path lenght tra le parole w_1 e w_2, h è la profondità della gerarchia lessicale ed infine d è la densità locale semantica di w_1 e w_2. Ipotizzando di poter riscrivere l'equazione precedente come un contributo di tre funzioni indipendenti, si ha: $sim(w_1,w_2) = f(f_1(l),f_2(h),f_3(d))$ dove $f_1(l),f_2(h),f_3(d)$ sono funzioni dipendenti della lunghezza percorso, dalla profondità e dalla densità locale rispettivamente. Una precisazione da fare è che la funzione $sim(w_1,w_2)$ è non lineare; questa cosa è facilmente intuibile in quanto se la lunghezza del percorso

va a zero si ha $sim(w_1, w_2) = 1$, mentre se se la lunghezza del percorso va all'infinito, $sim(w_1, w_2)$ tende a zero.

Contributo della lunghezza del percorso (path lenght)

Consideriamo una rete semantica organizzata come in figura seguente. In una rete del genere, la lunghezza del percorso tra due parole w_1 e w_2 può essere calcolata in uno dei seguenti casi:

1. w_1 e w_2 appartengono allo stesso concetto;
2. w_1 e w_2 non appartengono allo stesso concetto, ma entrambi i concetti contengono delle parole uguali;
3. w_1 e w_2 non appartengono allo stesso concetto e tali concetti non hanno nemmeno una parola in comune.

Nel caso 1 è evidente che w_1 e w_2 hanno lo stesso significato dunque il path lenght tra w_1 e w_2 sarà uguale a zero.

Nel caso 2 w_1 e w_2 condividono parzialmente il concetto e dunque il path lenght tra w_1 e w_2 viene considerato uguale ad uno.

Quest'ultima osservazione si basa sul fatto che in una gerarchia lessicale, quando i concetti di due parole w_1 e w_2 contengono più parole uguali , essi risultano essere molto simili.

Nel caso 3 viene calcolato l'effettivo path lenght tra w_1 e w_2.

In base alle considerazioni fatte il contributo relativo al path lenght si può calcolare come: $f_1(l) = \exp(-\alpha . * l)$ dove α è una costante. La scelta della funzione esponenziale fa si che questa soddisfi le ipotesi di non linearità fatte precedentemente e che il valore di questa rientri nell'intervallo [0 1].

Contributo della profondità (depth effect)

Prima di esprimere attraverso una funzione il contributo della profondità cerchiamo di capire bene di cosa si tratta. Data una coppia di parole, la profondità in questione è calcolata come quella relativa al primo sovraordinato comune. La funzione che tiene conto del depth effect è stata dedotta in base al seguente principio.

La gerarchia nella rete semantica è organizzata in modo per cui le parole che stanno nella parte alta, appartengono ad un concetto più generale ed hanno poca similarità, mentre le parole che stanno nella parte bassa godono di maggiore similarità ed appartengono a concetti più generali. La funzione in questione è la seguente: $f_2(h) = \frac{e^{(\beta \cdot h)} - e^{(-\beta \cdot h)}}{e^{(\beta \cdot h)} + e^{(-\beta \cdot h)}}$ dove β>0 è un fattore di smorzamento. Al

tendere di $\beta \to \infty$, la profondità delle parole nella rete semantica non è considerata.

Contributo della densità locale (local semantic density)

La *local semantic density* è una questione piuttosto complessa da affrontare in quanto non possiamo calcolarla soltanto attraverso la rete semantica. Comunque è possibile calcolarla ricorrendo all'aiuto di un grande corpus. Il concetto di "contenuto informativo" è stato originariamente sviluppato per la misura della semantic similarity tra parole. Tale nozione è presa in prestito per la rappresentazione della semantic density di un concetto in un corpus.

Detta $P(c)$ la probabilità di incontrare un istanza del concetto *c* in un corpus possiamo definire il "contenuto informativo" del concetto c come: $IC(c) = -\log p(c)$.

Per due concetti c_1 e c_2, la misura di similarità è determinata dal sovraordinato nella rete semantica con il massimo contenuto informativo: $sim(c_1, c_2) = \max_{c \in sub(c_1, c_2)} [-\log p(c)]$ dove $sub(c_1, c_2)$ è il set di concetti che include sia c_1 che c_2. L'informazione condivisa dalle parole w_1 e w_2 è definita come: $wsim(w_1, w_2) = \max_{c_1, c_2} [sim(c_1, c_2)]$ dove c_1 e c_2 comprendono tutti i possibili significati di w_1 e w_2.

Resnik [Resnik1999] suggerisce una modifica dell'equazione precedente dicendo che la misura della similarità può essere migliorata attraverso una somma pesata sui concetti, utilizzando informazioni sul contesto. In accordo con quanto scritto da Resnik il contributo della funzione che riguarda la densità locale sarà:

$$f_3(wsim) = \frac{e^{(\lambda * wsim(w_1, w_2))} - e^{(-\lambda * wsim(w_1, w_2))}}{e^{(\lambda * wsim(w_1, w_2))} + e^{(-\lambda * wsim(w_1, w_2))}}$$

Con $\lambda > 0$. se $\lambda \to \infty$, allora il contenuto informativo delle parole nella rete semantica non è considerato.

Strategie d'utilizzo

Come discusso precedentemente, i contributi utili per la misura della semantic similarity sono quelli relativi alla lunghezza del percorso che unisce due parole, la profondità del primo sovraordinato comune ed il contenuto informativo del concetto. Per avere dunque una buona misura della similarità bisogna fare in

modo di utilizzare una strategia affidabile. Gli autori utilizzano diverse strategie per definire la miglior forma delle equazioni. Dopodiché viene calcolato un coefficiente di correlazione tra il valore di semantic similarity ottenuto dal calcolo e quello ottenuto da un giudizio umano. Quest'ultimo non è altro che relativo ad un set di dati fissato sul quale un insieme di individui hanno espresso una personale misura di similarità. Esso è ormai consolidato da anni ed è un banco di prova per molti studiosi. Tra tutte le strategie la migliore sarà quella col coefficiente di correlazione più alto.

Come accennato precedentemente , la qualità del metodo per il calcolo della similarità è stabilita testando le funzioni su di un set di 65 coppie di parole sulle quali un gruppo di 51 persone ha già espresso la misura di similarità [Rubenstein1965]. La similarità doveva essere espressa attraverso una scala che va da 0 (nessuna similarità) a 4 (perfetta sinonimia). Nel corso di questo esperimento Sono state allora divise le 65 coppie di R-G in due set: 28 coppie denotate con D0, e 37 coppie denotate con D1. Quest'ultimo è usato per progettare il metodo, mentre D0 per testarlo. I set D0 e D1 sono mostrati nella figura seguente.

Word Pair	RG Rating	MC Replica	Resnik Replica	Information Content	Length	Depth
cord-smile	0.02	0.13	0.1	1.1762	12	0
rooster-voyage	0.04	0.08	0	0	30	0
noon-string	0.04	0.08	0	0	30	0
glass-magician	0.44	0.11	0.1	1.0105	8	0
monk-slave	0.57	0.55	0.7	2.9683	4	2
coast-forest	0.85	0.42	0.6	0	6	1
monk-oracle	0.91	1.1	0.8	2.9683	7	2
lad-wizard	0.99	0.42	0.7	2.9683	4	2
forest-graveyard	1.00	0.84	0.6	0	7	1
food-rooster	1.09	0.89	1.1	1.0105	12	0
coast-hill	1.26	0.87	0.7	6.2344	4	3
car-journey	1.55	1.16	0.7	0	30	0
crane-implement	2.37	1.68	0.3	2.9683	4	3
brother-lad	2.41	1.66	1.2	2.9355	4	2
bird-crane	2.63	2.97	2.1	9.3139	3	5
bird-cock	2.63	3.05	2.2	9.3139	1	5
food-fruit	2.69	3.08	2.1	5.0076	4	3
brother-monk	2.74	2.82	2.4	2.9683	1	5
asylum-madhouse	3.04	3.61	3.6	15.666	1	7
furnace-stove	3.11	3.11	2.6	1.7135	2	2
magician-wizard	3.21	3.5	3.5	13.666	0	4
journey-voyage	3.58	3.84	3.5	6.7537	1	5
coast-shore	3.60	3.7	3.5	10.808	1	4
implement-tool-	3.66	2.95	3.4	6.0787	1	4
boy-lad	3.82	3.76	3.5	8.424	1	4
automobile-car	3.92	3.92	3.9	8.0411	0	7
midday-noon	3.94	3.42	3.6	12.393	0	7
gem-jewel	3.94	3.84	3.5	14.929	0	6

Word Pair	RG Rating	MC Replica	Resnik Replica	Information Content	Length	Depth
cord-smile	0.02	0.13	0.1	1.1762	12	0
rooster-voyage	0.04	0.08	0	0	30	0
noon-string	0.04	0.08	0	0	30	0
glass-magician	0.44	0.11	0.1	1.0105	8	0
monk-slave	0.57	0.55	0.7	2.9683	4	2
coast-forest	0.85	0.42	0.6	0	6	1
monk-oracle	0.91	1.1	0.8	2.9683	7	2
lad-wizard	0.99	0.42	0.7	2.9683	4	2
forest-graveyard	1.00	0.84	0.6	0	7	1
food-rooster	1.09	0.89	1.1	1.0105	12	0
coast-hill	1.26	0.87	0.7	6.2344	4	3
car-journey	1.55	1.16	0.7	0	30	0
crane-implement	2.37	1.68	0.3	2.9683	4	3
brother-lad	2.41	1.66	1.2	2.9355	4	2
bird-crane	2.63	2.97	2.1	9.3139	3	5
bird-cock	2.63	3.05	2.2	9.3139	1	5
food-fruit	2.69	3.08	2.1	5.0076	4	3
brother-monk	2.74	2.82	2.4	2.9683	1	5
asylum-madhouse	3.04	3.61	3.6	15.666	1	7
furnace-stove	3.11	3.11	2.6	1.7135	2	2
magician-wizard	3.21	3.5	3.5	13.666	0	4
journey-voyage	3.58	3.84	3.5	6.7537	1	5
coast-shore	3.60	3.7	3.5	10.808	1	4
implement-tool-	3.66	2.95	3.4	6.0787	1	4
boy-lad	3.82	3.76	3.5	8.424	1	4
automobile-car	3.92	3.92	3.9	8.0411	0	7
midday-noon	3.94	3.42	3.6	12.393	0	7
gem-jewel	3.94	3.84	3.5	14.929	0	6

Tabella 5.2: Gli insiemi D0 e D1

Gli autori suggeriscono 10 strategie , a valle dei risultati sperimentale scelgono che la formula per il calcolo della similarity è la seguente:

$$S(w_1,w_2) = e^{(-\alpha * l)} \frac{e^{(\beta * h)} - e^{(-\beta * h)}}{e^{(\beta * h)} + e^{(-\beta * h)}} \qquad \text{con } \alpha = 0.2 \text{ e } \beta = 0.6 .$$

5.4 INTEGRATED APPROACHES

In questo paragrafo si è cercato di presentare un approccio più generale, dato dall'integrazione tra le sorgenti informative utilizzate nei metodi esposti fino a questo momento. In particolare si cerca di considerare il problema relativo alle ontologie generali. Quasi tutte le tecniche esposte di seguito sono incentrate sull'analisi di corpus, con lo scopo di aumentare informazioni presenti sulla rete.

5.4.1 Approccio "information-based" di Resnik

L'idea che sta alla base di questo approccio proposto in [Resnik1995] è quella per cui il contenuto informativo di un concetto è indice della specificità o della generalità del concetto stesso. Il calcolo del contenuto informativo è basato sull'assunzione dell'esistenza di una certa gerarchia. Per essere più chiari, facciamo riferimento a Wordnet e supponiamo che in un determinato testo, il concetto c_1 sia incluso nel concetto c_2. Da questa considerazione è chiaro che ogni volta che incontriamo un'occorrenza di c_1 è come se ne incontrassimo anche una di c_2. Ad esempio se in un dato testo incontriamo il termine *maglione* piochè quest'ultimo si inserisce in un contesto più generale che è quello dell'*abbigliamento*, possiamo dire che il testo parli di abbigliamento. In tal modo, tutti gli oggetti appartenenti allo stesso concetto avranno un determinato grado di specificità. Per trovare l'informazione contenuta in un determinato documento, basta calcolare la frequenza di occorrenza dei vari concetti. In pratica ad ogni concetto verrà attribuito un numero che ne rappresenta la frequenza.

Nel testo da analizzare, ogni volta che occorre un concetto, il numero rappresentante la relativa frequenza verrà incrementato. Tale cosa non è però così semplice come descritto. Il perché di questa affermazione si basa su un'osservazione semplice e cioè, *una stessa parola inserita in due frasi diverse può richiamare due concetti differenti*. Dunque esiste il problema dell'ambiguità. Resnik risolve tale cosa affermando che se viene incontrato uno stesso vocabolo per 100 volte e questo può essere attribuito a 10 concetti distinti, allora verrà fatta una distribuzione equa delle frequenze e cioè come se ogni concetto l'avessimo incontrato 10 volte. Sotto tale ipotesi, il contenuto informativo è

calcolato attraverso la seguente formula: $Ic(c) = -\log\left(\frac{freq(c)}{freq(root)}\right)$ dove $Ic(c)$ è l'informazione contenuta nel concetto c, *root* è il nodo radice della tassonomia mentre *freq(c)* e *freq(root)* sono le frequenze di questi concetti. Un'altra questione importante da affrontare è quella relativa al valore zero della frequenza. In altre parole se *freq(c)*=0 abbiamo un valore indefinito del logaritmo. Questa cosa è possibile affrontarla in due modi diversi. Il primo è permettere questo tipo di valore andando però ad effettuare manipolazioni particolari alla misura.

Il secondo è quello di aggiungere 1 ad ogni valore di *freq(c)* .Così facendo non ci sarà mai un valore nullo; questo vantaggio si paga chiaramente in termini di precisione della misura, la quale verrà leggermente falsata in quanto non avremo mai concetti con frequenza pari a zero.

Resnik, definisce il valore della semantic relatedness di due concetti come quantità di informazione che hanno in comune. Egli calcola la quantità di informazione comune tra due concetti come: $related_{res}(c_1, c_2) = Ic(lcs(c_1, c_2)$ dove *Ic* ha lo stesso significato della definizione precedente, mentre $lcs(c_1, c_2)$ è definito come *lowest common subsumer* ossia il più basso sovraordinato comune ai due concetti considerati nella gerarchia.

La misura in questione dipende allora completamente dall'informazione contenuta dal lowest common subsumer dei due concetti di cui vogliamo misurarne la semantic relatedness.

5.4.2 La misura di Jiang e Conrath

Il lavoro svolto da Jiang e Conrath [Jiang1997] per la misura della semantic similarity si basa su un approccio che combina una tassonomia lessicale con un corpus di informazioni statistiche allo scopo di migliorare la misura della distanza semantica tra nodi dello spazio in questione. Più specificamente, la misura proposta sarà una combinazione di tra un approccio "edge based" arricchito da uno " node based".

Approccio node-based (Information content)

Dato uno spazio multidimensionale nel quale i nodi rappresentano un unico concetto consistente di una certa quantità di informazione e un arco che rappresenta un'associazione diretta tra i due concetti, la similarità può essere

rappresentata come quantità di informazione in comune. L'informazione comune portante può essere identificata come un "nodo" che include entrambi i concetti nella gerarchia. In pratica questa "superclasse" sarà la prima, in alto nella gerarchia, ad includere entrambe le classi. La similarità è definita allora come il valore dell'informazione, contenuta dalla classe in questione ,ottenuta calcolando la probabilità di occorrenza di quest'ultima in un testo.

Utilizzando dunque una notazione data dalla teoria dell'informazione, il contenuto informativo di un concetto può essere scritto nel modo seguente:

$IC(c) = -\log p(c)$ con $p(c)$ probabilità di incontrare un istanza del concetto *c* in un testo. Vale la pena puntualizzare che se un concetto è incluso nella parte più bassa della gerarchia, ci sono più nodi che possano includerlo, ma il contenuto informativo diminuisce.

Data la monotonicità della funzione $IC(c)$, la similarità tra due concetti può essere definita come:

$$sim(c_1, c_2) = \max_{c \in Sup(c_1, c_2)} [IC(c)] = \max_{c \in Sup(c_1, c_2)} [-\log p(c)]$$

dove $Sup(c_1, c_2)$ è il set di concetti che include sia c_1 che c_2. In altre parole ci stiamo riferendo al primo nodo verso l'alto della gerarchia che include entrambi i concetti. Nel caso di eredità multiple e cioè quando due parole possono avere più di un senso ossia più superclassi dirette, la word similarity può essere valutata come la migliore similarity misurata su tutte le coppie di classi appartenenti a quel dato significato. Possiamo riassumere il tutto secondo la seguente formula:

$sim(w_1, w_2) = \max_{c_1 \in sen(w1) c_2 \in sen(w2)} [sim(c_1, c_2)]$ dove $sen(w)$ denota il set di possibili significati per la parola *w*.

Prima di procedere nei dettagli, è opportuno stabilire la differenza tra due sets di concetti: $words(c)$ e $classes(w)$. Il primo è il set di vocaboli che è incluso direttamente o indirettamente dalla classe *c*. Questo può essere visto come un sottoalbero della gerarchia, includente nella propria radice il concetto *c*. $Classes(w)$ è definito come l'insieme delle classi nelle quali la word *w* è contata. Per capire meglio quanto detto, facciamo riferimento alla seguente formula:

$$classes(w) = \{c \mid w \in words(c)\}$$

Si noti che Resnik [Resnik1995] propose la seguente formula per il calcolo della frequenza di concetto o classi: $freq(c) = \sum_{w \in words(c)} freq(w)$ mentre Richardson e

Smeaton [Richardson1995a], [Richardson1995b] proposero una formula leggermente diversa : $freq(c) = \sum_{w \in words(c)} \frac{freq(w)}{|classes(w)|}$.

Infine è possibile calcolare la probabilità dei concetti o classi usando la "maximum likelihood estimation" (MLE): $P(C) = \frac{freq(c)}{N}$

Edge base (distance) approach

Il metodo basato sulla distanza, rappresenta la strada più naturale per valutare la semantic similarity in una tassonomia. Essa calcola la distanza tra due nodi delle cui parole si deve calcolare la similarity.

Dato uno spazio multidimensionale, la variabile in questione può essere misurata come la distanza geometrica tra i nodi rappresentanti i concetti. In uno scenario più realistico non è detto che due nodi adiacenti, aventi la stessa distanza geometrica con un dato vocabolo, abbiano lo stesso valore di similarità con questo. Ciò che viene fatto è essenzialmente un operazione di pesatura delle distanze.

In parole povere, due parole adiacenti equidistanti geometricamente dalla parola sovrastante, non avranno obbligatoriamente lo stesso valore di similarity in quanto i percorsi congiungenti avranno dei pesi diversi. Altri aspetti plausibili da tenere in considerazione sono la densità locale della rete, la profondità del nodo nella gerarchia , tipi di link ed infine l'intensità dei collegamenti.

Sono stati condotti due studi che trattano la problematica appena esposta. Richardson e Smeaton [Richardson1995a] [Richardson1995b] hanno considerato i primi due e l'ultimo degli aspetti , mentre Sussna [Sussna1993], come già detto in precedenza, ha proposto una particolare operazione di pesatura dei collegamenti. Resnik [Resnik1995] propose un modo ancora più semplice per convertire la misura della distanza in una misura di similarità (eliminando la pesatura dei collegamenti), ricavando la seguente relazione:

$sim(w_1, w_2) = 2d_{\max}[\min_{c_1 \in sen(w_1) c_2 \in sen(w_2)} len(c_1, c_2)]$ dove $d_{\max}$ è la massima profondità nella tassonomia e la funzione *len* effettua semplicemente il calcolo dello shortest path lenght considerando unitari tutti i pesi.

Un approccio combinato

Il metodo effettivamente proposto da Jiang e Conrath è una combinaziona tra "l'edge based approach "ed il "node based approach". Essi sostengono che la "forza di un collegamento" ad un nodo figlio è proporzionale alla probabilità

condizionata di incontrate un istanza del nodo figlio c_i data un istanza del suo nodo padre *p*, ossia $P(c_i \mid p)$. Possiamo scrivere quest'ultima nel seguente modo: $P(c_i \mid p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)}$

In base alle considerazioni fatte e utilizzando delle nozioni derivanti dalla teoria dell'informazione, possiamo, definire il "link strength" (forza del collegamento) nel modo riportato di seguito:

$$LS(c_i, p) = -\log(P(c_i \mid p)) = IC(c_i) - IC(p)$$

Dalla relazione precedente si evince che abbiamo a che fare con la differenza tra il contenuto informativo del nodo figlio e il contenuto informativo del nodo padre. Considerando altri fattori come la densità locale, la profondità dei nodi, il tipo di link, possiamo affermare che il peso totale di un link tra un nodo padre e un nodo figlio può essere calcolato con la seguente formula:

$$wt(c,p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}[IC(c) - IC(p)]T(c,p)$$

dove *d(p)* denota la profondità del nodo *p* nella gerarchia, $E(p)$ la densità locale, $\overline{E}$ la densità media in tutta la gerarchia e *T(c,p)* il fattore che denota il tipo o la relazione del collegamento. I parametri α ($\alpha \geq 0$) e β ($0 \leq \beta \leq 1$) sono indice di quanto influiscono il fattore profondità e il fattore densità nella precedente formula. Tali contributi diventano poco significativi se $\alpha \to 0$ e $\beta \to 1$. Dunque la distanza complessiva tra due nodi potrà essere ottenuta come la somma dei vari pesi attraverso lo shortest path:

$$Dist(w_1, w_2) = \sum_{c \in \{path(c_1,c_2) - LSuper(c_1,c_2)\}} wt(c, parent(c))$$

dove $c_1 = sen(w_1)$, $c_2 = sen(w_2)$ e $path(c_1, c_2)$ è l'insieme che contiene tutti i nodi nello shortest path da c_1 a c_2. Uno degli elementi del set è $LSuper(c_1, c_2)$, che denota il più piccolo sovraordinato comune a c_1 e c_2. Nel caso speciale in cui è considerato solo il contributo *LSuper* nella formula dei pesi (cioè ponendo $\alpha = 0$, $\beta = 1$ e *T(c,p)*=1) la funzione distanza può essere semplificata come segue: $Dist(w_1, w_2) = IC(c_1) + IC(c_2) - 2 * IC(LSuper((c_1, c_2))$

ottenendo dunque una relazione sicuramente più facile da manipolare e da interpretare.

5.4.3 La misura di Lin

Date le conoscenze di quel momento sulla misura della semantic similarity, Lin [Lin1997] [Lin1998] intraprese il tentativo di definire una nuova misura della semantic similarity universalmente applicabile e teoricamente giustificata. Per arrivare alla sua definizione, egli usa tre principi base secondo cui:

Principio1: La similarità tra *A* e *B* è proporzionata alla loro "commonality". Quanto più condividono "commonality" tanto più risultano simili.

Principio 2: La similarità tra *A* e *B* è proporzionata alla differenza tra i termini. Quanto più sono differenti, tanto meno saranno simili.

Principio 3: La massima similarità tra *A* e *B* è raggiunta quando questi sono identici e non quando hanno il massimo della commonality.

Lo scopo di tutto il discorso è definire una nuova misura di similarità basata sui principi enunciati sopra.

Inoltre l'autore fa alcune assunzioni:

Assunzione 1: La "commonality" tra *A* e *B* è misurata dalla quantità di informazione contenuta nella proposizione che stabilisce la commonality tra termini. Espletando quanto detto in modo formale possiamo indicare questa quantità come: $IC(common(A,B))$

Per il calcolo della commonality abbiamo anche bisogno di misurare la differenza tra gli oggetti in questione; per questo Lin fa la seguente

Assunzione 2: La differenza tra *A* e *B* è misurata come:

$IC(description(A,B)) - IC(common(A,B))$ dove $description(A,B)$ è una proposizione che descrive cosa sono *A* e *B*.

Assunzione 3: La similarità tra *A* e *B* $Sim(A,B)$ è una funzione della loro commonality e della loro differenza, cioè:

$Sim(A,B) = f(I(common(A,B)) - I(description(A,B)))$

Il dominio della *f* è : $\{(x,y) \mid x \geq 0, y > 0, y \geq x\}$

Il principio 3 porta alla seguente

Assunzione 4: Il valore della similarità tra due oggetti identici è costante ed è uguale ad 1. Da questa considerazione capiamo che la funzione *f* ha un'altra proprietà: $\forall x > 0, f(x,x) = 1$

Quando non c'è commonality tra due oggetti, si assume che la loro similarità sia uguale a 0.

Assunzione 5: $\forall y > 0, f(0, y) = 0$.

Assunzione 6: Supponiamo di avere due oggetti *A* e *B*, i quali possono essere visti in due prospettive distinte. La loro similarità può essere dunque calcolata separatamente sulla base di ogni prospettiva.

Per fare un esempio più concreto, supponiamo di avere due documenti distinti. Si assuma che per calcolarne la similarità totale, andremo a fare una media pesata delle similarità viste dalle due prospettive diverse.

I pesi saranno le quantità di informazione nelle descrizioni ossia:

$$\forall x_1 \le y_1, x_2 \le y_2 : f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1 + y_2} * f(x_1, y_1) + \frac{y_2}{y_1 + y_2} * f(x_2, y_2)$$

In base a quanto detto e grazie alla teoria dell'informazione Lin fu in grado di dimostrare il seguente teorema della similarità:

La similarità tra A e B è calcolata tramite il rapporto tra la quantità di informazione necessaria per definire la loro commonality e l'informazione necessaria per descrivere completamente cosa sono, e cioè:

$$sim_L(A, B) = \frac{\log P(common(A, B))}{\log P(description(A, B))}$$

deducendone perciò la formula operativa per il calcolo della similarità:

$$sim_L(c_1, c_2) = \frac{2 * \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

CAPITOLO 6 IL MODELLO SEMANTICO E IL SISTEMA DYSE: DYNAMIC SEMANTIC ENGINE

La teorizzazione di un modello nel quale vengono formalizzate le tecniche e le metodologie proposte è un passo importante per la definizione di un problema di information retrieval.

Verrà ora affrontato questo aspetto e, nel corso del capitolo, verrà descritto un framework che, basandosi su concetti e relazioni tra simboli, cercherà di proporre una possibile soluzione ai problemi di IRR.

Inoltre nel capitolo verrà presentata l'implementazione di un sistema la ricerca semantica sul web nel quale è stato utilizzato il modello proposto.

Non avendo a disposizione un servizio di indicizzazione si è scelto di costruire un meta-motore che utilizza i motori di ricerca tradizionali per recuperare le pagine; sarà poi compito del nostro sistema analizzarle utilizzando una metrica di similitudine semantica per assegnare un voto di similitudine tra i documenti restituiti e la query utente.

L'architettura ad alto livello prevede la possibilità di sottoporre la query inserita dall'utente ai motori di ricerca selezionati. In questa fase, il sistema dialoga con ciascun motore di ricerca al fine di ottenere un'insieme di link che viene archiviato, insieme alle relative pagine web, in un Web Repository. In una seconda fase, le pagine archiviate vengono preprocessate al fine di eliminare tutto ciò che non deve essere analizzato da un punto di vista semantico e isolarne le parti essenziali: titolo, descrizione, parole chiavi e corpo. Successivamente, il Miner effettua l'analisi semantica, al fine di assegnare un punteggio ad ogni documento; vengono calcolate le varie parti della metrica per ogni tag definito significativo e viene dato un punteggio globale alla pagina. La metrica proposta è formata da varie componenti come in [Li2000].

Le operazioni di recupero delle pagine restituite dai motori, di ripulitura delle pagine dagli elementi non importanti da un punto di vista semantico e l'analisi semantica vera e propria, non sono eseguite dal sistema in modo strettamente sequenziale, ma in parallelo su moduli diversi per garantire una maggiore efficienza.

È stato previsto quindi, come descriveremo meglio in seguito, la distribuzione delle operazioni eventualmente su sistemi hardware diversi e per ogni modulo l'utilizzo del multithreading per parallelizzare l'elaborazione delle diverse pagine.

Le informazioni relative all'intero sistema sono inserite in una base dati che sarà presentata in questo capitolo, a livello progettuale, utilizzando il modello Entità-Relazione.

6.1 IL MODELLO SEMANTICO

Cercare di cogliere gli aspetti semantici dell'informazione è un processo di non facile soluzione. Inoltre, le relazioni tra concetti, espressi in un qualsiasi linguaggio utilizzando differenti simboli, impone una complessa analisi del discorso.

Prima di tutto quindi, andiamo a chiarire alcuni dei termini che sono già emersi in questa primissima introduzione.

Nella semiotica, un segno è in generale definito come "something that stands for something else, to someone in some capacity" [Danesi1999]. Esso potrebbe essere interpretato come un'unità finita di significato. I segni non sono solo parole, ma possono includere immagini, gesti, odori, sapori, suoni; cioè tutti i modi in cui l'informazione può essere processata in una forma codificata e comunicata come un messaggio da una mente senziente e razionale ad un'altra.

Nella nostra trattazione, senza perdere di generalità, il segno sarà una parola, composta da simboli, correlata ad altre parole tramite alcune proprietà.

Diamo ora alcune definizioni

Simbolo: *Si definisce simbolo ogni segno convenzionale che denota un elemento, una funzione o un'operazione entro un linguaggio simbolico* [DeMauro2005].

Vocabolo: *Si definisce vocabolo un insieme di caratteri (simboli) isolato, colto fuori da un contesto e da ogni legame grammaticale o logico, nella sua individualità lessicale* [Garzanti1980].

Lingua: *Si definisce una lingua come l'insieme dei vocaboli utilizzati da un popolo, insieme alle regole che li governano* [Garzanti1980].

Regole sintattiche: *Si definiscono regole sintattiche (o sintassi) l'insieme delle norme che studiano le relazioni che i vocaboli hanno in una frase (gruppo di vocaboli)* [Garzanti1980].

Regole semantiche: *Si definiscono regole semantiche (o semantica) l'insieme delle norme che si occupano del significato dei vocaboli e dei cambiamenti di essi* [Garzanti1980].

Concetto: *Per concetto linguistico si intende una parola, frase, acronimo o nome ricco di significato, che è stato estratto da componenti non strutturati del testo, incluso blocchi isolati di testo, sommari, intestazioni, paragrafi etc. Esso è definito anche come termine linguistico, mentre ogni documento si riferisce a un gruppo di concetti o termini* [Latiri2001].

Universo del discorso: *L'insieme U di tutti i concetti conosciuti*.

L'universo del discorso rappresenta l'orizzonte di conoscenza di un agente umano o software.

Dominio: *Detto U l'universo del discorso si definisce dominio D un sottoinsieme di U, ovvero* $D \subseteq U$.

Dato un dominio D e detti C_i i concetti in esso definiti si ha che l'unione di tutti i concetti è una copertura per il dominio stesso, ovvero $D = \bigcup_{i=1}^{n} C_i$

Ogni concetto è costituito da vocaboli della lingua alla quale si ci riferisce. Possiamo, però considerare che i vocaboli non appartengano ad un solo concetto. Soprattutto nei casi in cui si vuole rappresentare un dominio con un insieme sintetico di concetti, può accadere che alcuni vocaboli facenti parte del dominio, siano inseriti in concetti ai quali non appartengono pienamente.

Peso: *Definiamo peso il grado con cui un vocabolo rappresenta il dominio al quale è associato*.

Parola: *Definiamo parola un vocabolo di una lingua, a cui è associato un concetto C, il dominio D a cui il concetto appartiene e il peso della parola nel dominio; una parola è quindi una quadrupla:*

<vocabolo lessicale, concetto, dominio, peso>

Grado di associazione: *Dato un dominio D costituito da un insieme di concetti* $C_1,...,C_n$ *è possibile definire una grado di associazione tra i concetti* C_i *e* C_j *, con i, j = 1..n*

<C_i*,* C_j*, gradodiassociazione>*

Utilizzando il grado di associazione i concetti, nell'universo del discorso, sono tutti legati tra di loro. Possibili tipi di associazioni possono essere, ad esempio, le proprietà linguistiche.

6.1.1 La rappresentazione delle informazioni nel modello semantico

Dallo studio dei modelli per l'IR, dopo una prima definizione teorica per la modellazione dello spazio di interesse, si devono formalizzare le necessità informative degli utenti e la rappresentazione dei documenti.

Nel nostro modello la rappresentazione delle necessità informative dell'utente è data da una query *q* che utilizza le parole appartenenti all'universo del discorso e quindi all'orizzonte di conoscenza. Intuitivamente questa è una scelta obbligata perché, naturalmente, non è possibile esprimere qualcosa che non si conosce e, in maniera analoga, non è possibile rappresentare un'informazione se non si sa qual è il simbolo che la rappresenta. Le keyword utilizzate vanno a formare il dominio *D* di interesse per l'utente.

D'altro canto deve anche essere definita la rappresentazione dei documenti nel modello. Scegliamo una rappresentazione full-text in modo tale da avere una descrizione più dettagliata del documento.

Notiamo che non è detto che nel documento siano presenti solo termini appartenenti all'universo del discorso; inoltre, analizzando documenti in linguaggio naturale, questi presentano dei morfismi che necessitano di una fase di preprocessing per riportare tutti i termini in forma base e per eliminare eventuale rumore (parole non interessanti per l'analisi) nel testo. Le keywords risultanti vanno a formare la rappresentazione *d* del documento considerato.

Una volta date le definizioni del modello ed individuate le rappresentazioni delle informazioni dobbiamo presentare un framework per l'utilizzo del modello.

Supponiamo che la query utente *q* sia rappresentata da un vettore $\vec{q}$ di lunghezza N composto da tutti i termini riferiti ad un concetto inserito dall'utente che rappresenta il suo dominio di interesse. Chiamiamo inoltre $\vec{d}$ il vettore di lunghezza T contenete i termini del documento *d* presente nella collezione dei documenti.

Partendo dal vettore $\vec{q}$ possiamo costruire la matrice termine-termine nella quale viene dato ad ogni coppia di termini un grado di associazione che rappresenta la proprietà che li lega.

Dall'intersezione dei due vettori $\vec{q}$ e $\vec{d}$ è possibile ricavare un vettore $\vec{l}$, di lunghezza $S \leq N$. Questo vettore è quindi una lexical chain [Halliday1976] che contiene i termini del documento *d* che soddisfano le necessità informative

dell'utente definite mediante la query *q* rispetto all'orizzonte di conoscenza. A tali termini è associato un peso che indica quanto il singolo termine sia rappresentativo del dominio di interesse dell'utente e un grado di similitudine che dovrebbe essere calcolato mediante una metrica composta da vari elementi come il peso del path tra i singoli termini e la loro generalità/specializzazione.

6.2 ARCHITETTURA DEL SISTEMA

La necessità di uno sviluppo rapido per adeguare le tecniche e gli strumenti per l'analisi del Web alle più recenti scoperte scientifiche e per testare velocemente il sistema a seguito di variazioni, anche profonde, dello stesso, ha imposto la scelta di un'architettura modulare e scalabile.

Da una visione di alto livello possiamo inquadrare il sistema in una classico schema three tier come mostrato nella figura seguente.

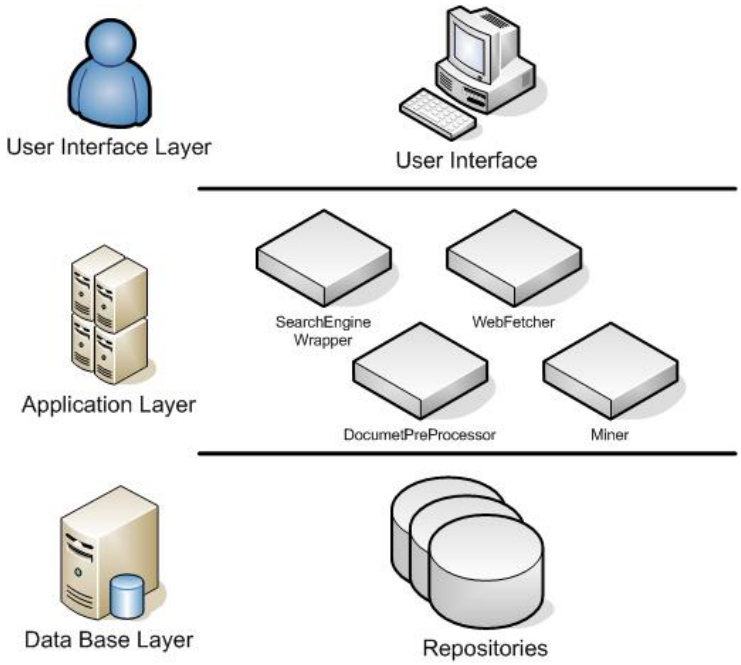


Figura 6.1: Schema three tier del sustema

Da qui si evincono immediatamente le varie fasi che il nostro sistema deve implementare per un corretto retrieval di pagine Web:

- **Querying:** l'utente interroga il sistema mediante un interfaccia nella quale vengono inserite le keyword rappresentative della query utente;
- **Fetching**: consiste nella ricerca di documenti Web che contengono le keyword specificate dall'utente nella sua query. Nel nostro caso questa fase viene svolta da un apposito modulo che si appoggia sui motori di ricerca tradizionali;

- **Preprocessing**: in questa fase vengono eliminate tutte le parti dei documenti Web che non rappresentano informazioni utili per l'analisi semantica (tag HTML, stopwords, ecc…);
- **Mining**: è la fase in cui i documenti sono esaminati da un punto di vista semantico e alla fine della quale viene assegnato un punteggio di similarità con la query utente;
- **Reporting:** i documenti vengono ordinati e sottoposti all'utente mediante un'apposita interfaccia

La figura seguente mostra l'architettura globale del sistema e il simbolismo utilizzato.

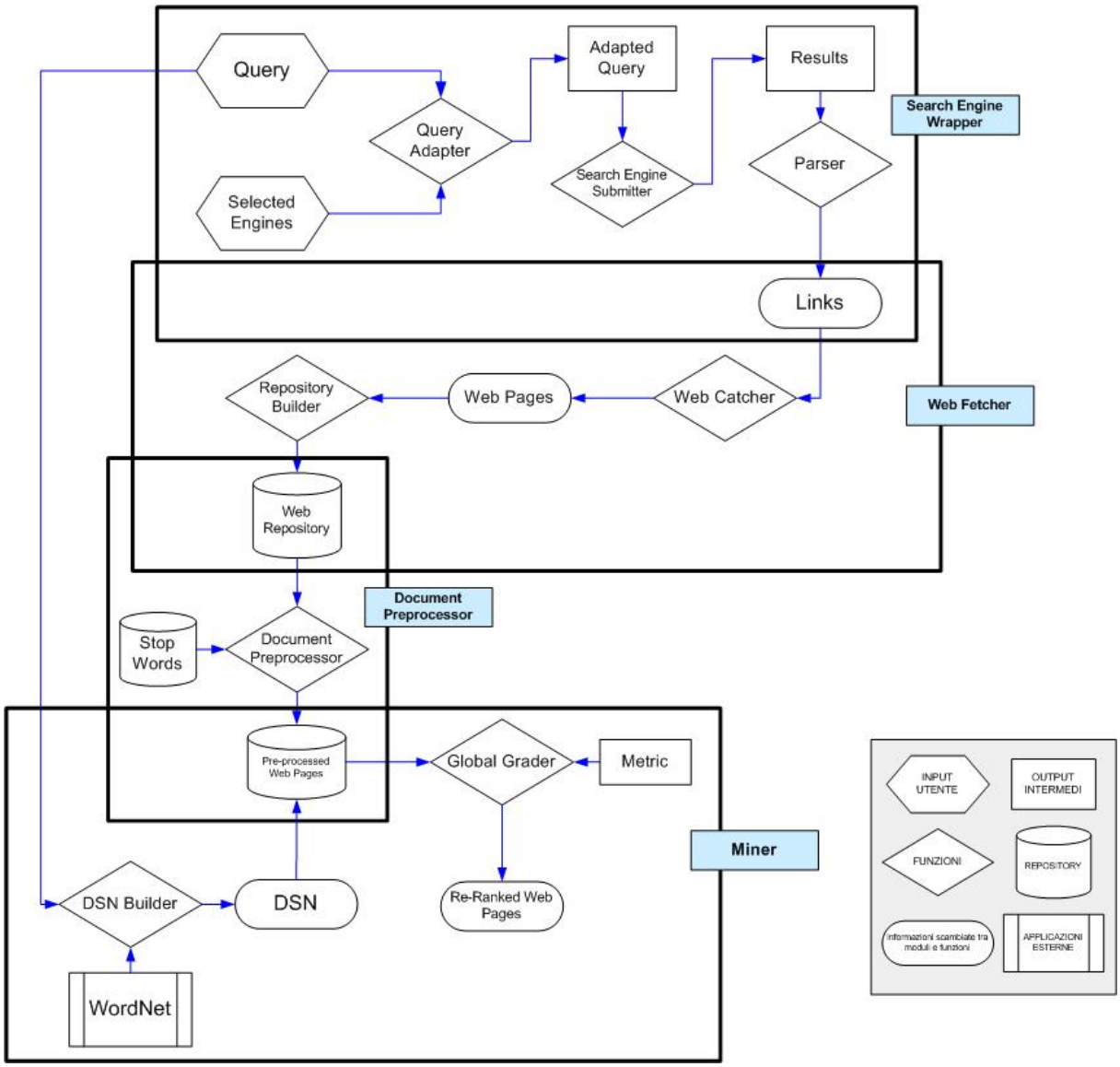


Figura 6.2: Architettura del sistema

Un utente può sottomettere una query al sistema specificando un insieme di keywords (*subject keywords*) come in un tradizionale motore di ricerca e un

dominio al quale è interessato (*domain keyword*). Le subject keywords sono utilizzare per il retrieval delle pagine dai motori di ricerca tradizionali, mentre la domain keyword viene utilizzata per creare, in maniera dinamica, una rete semantica estraendola da una base di conoscenza rappresentata nel nostro caso da WordNet.

Un utente può specificare più subject keywords ma solo un singolo dominio. Comunque possono essere utilizzate parole composte nella specificazione del dominio ma solo se queste sono presenti in WordNet.

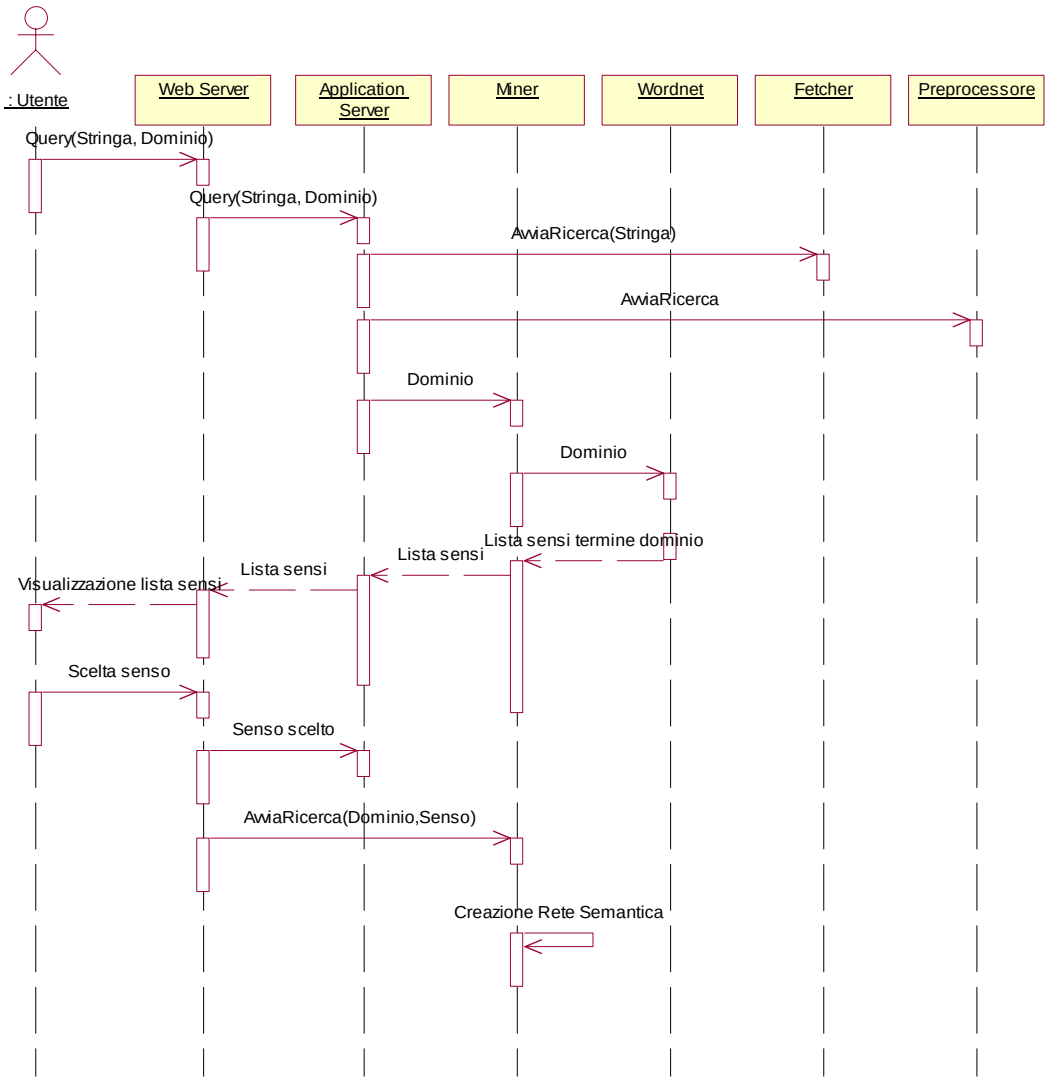


Figura 6.3: Sequence Diagram Interazione Utente-Sistema

Come si può notare l'Application Server, ricevuta la query dell'utente dal web server attiva tutti i moduli del sistema, ossia il Fetcher inizia a scaricare le pagine dai motori di ricerca tradizionali ed il Preprocessore comincia ad elaborarle. Contemporaneamente il Miner riceve dall'application server il termine inserito

dall'utente per rappresentare il dominio semnatico. Il Miner consulta WordNet per avere la lista dei sensi del termine che viene poi restituita al Web Server per la visualizzazione a schermo.

Il Miner a questo punto si dispone in attesa, finché l'utente non sceglie il senso e lo invia attraverso l'Application Server al Miner stesso che genera così la rete semantica.

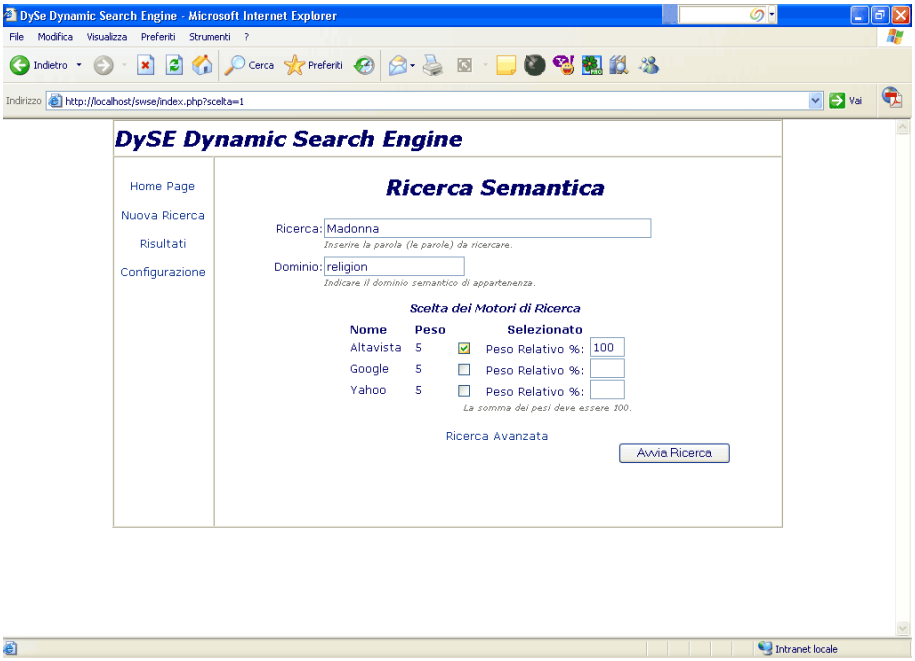


Figura 6.4: Interfaccia inserimento query

6.3 SEARCH ENGINE WRAPPER

Il modulo Search Engine Wrapper riceve le subject keywords inserite mediante l'interfaccia utente e la adatta alle specifiche sintassi dei motori di ricerca, mediante il Query Adapter, creando la stringa di interrogazione per i singoli motori scelti per la ricerca.

Allo scopo di rendere del tutto trasparente per l'utente l'interrogazione dei motori di ricerca, il Serch Engine Wrapper si occupa della sottomissione della parte della query di suo interesse, adattata dal Query Adapter, ai motori di ricerca e di prelevarne i risultati.

Il Search Engine Submitter sottopone la query ad un motore di ricerca ottenendo la pagina con i risultati restituiti ed il Parser analizza tale pagina allo scopo di prelevarne i link in essa contenuti.

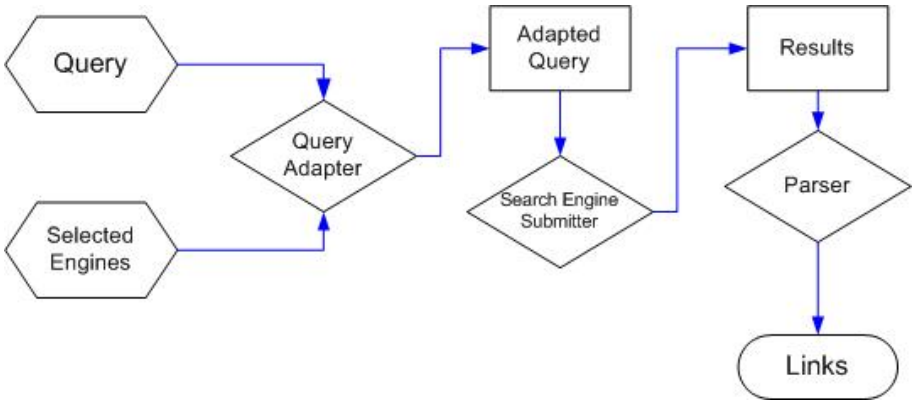


Figura6.5: Search Engine Wrapper

6.4 WEB FETCHER

Il compito del Web Fetcher è quello di individuare le pagine corrispondenti ai link restituiti dai motori di ricerca e di memorizzarle nel repository locale. Le pagine vengono localizzate dal Web Catcher, mentre il Repository Builder provvede ad inserirle nel Database del sistema.

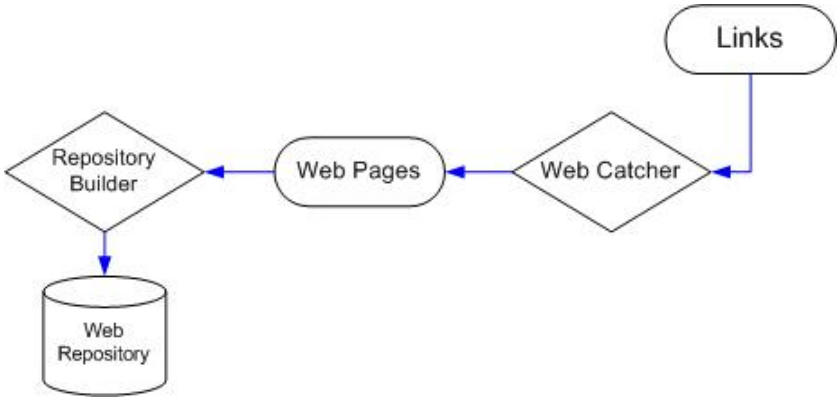


Figura 6.6: Web Fetcher

6.5 DOCUMENT PREPROCESSOR

Questo modulo sottopone le pagine scaricate dal Web Fetcher ad un processo di analisi e segmentazione.

Una pagina web è tipicamente composta da parecchi elementi. Il nostro sistema divide una pagina web in alcune componenti fondamentali:

- titolo;
- descrizione;
- keywords;
- body.

Il document pre-processor analizza le pagine e le segmenta in questi elementi base che saranno sottoposti successivamente all'elaborazione semantica.

Il testo contenuto in questi tag viene sottoposto a varie fase di pre-processing:

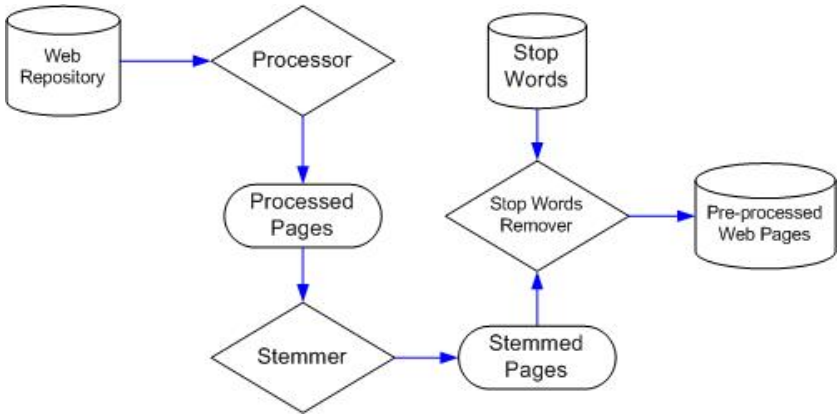


Figura 6.7: PreProcessor

Processing

La fase di processing comporta la eliminazione dei simboli di punteggiatura, caratteri numerici, simboli HTML, ecc. e delle stopwords come articoli e proposizioni. Questa lista è stata ottenuta dall'unione di 5 liste *Oracle 8 ConText*, *SMART*, *Hyperwave*, *University of Kansas* e *Ohio State University* per un totale di 980 stop words.

Parsing

Il parser effettua l'analisi grammaticale della frase: ad ogni parola associa un tag che ne indica il tipo: nome, verbo, aggettivo, avverbio, articolo, preposizione, ecc.. Per effettuare questo tipo di operazione, è stata utilizzata la versione 1.2 del tagger Monty Tagger [Liu2003].

Stemming

Consiste nel riportare le varie parti del discorso nella loro forma originale (ad es. i nomi al singolare e i verbi nella forma all'infinito). Per eseguire la stemmizzazione di una parola è stata utilizzata una funzione apposita messa a disposizione da WordNet chiamata Morph.

Ci sono da fare alcune considerazioni per quanto riguarda l'ordine di esecuzione delle operazioni.

Il parser associa il corrispondente tag ad una parola, tenendo presente anche il contesto in cui questa è inserita, quindi delle altre parole contenute nel testo. La eliminazione delle stopwords prima del parsing fa si che la operazione di etichettatura perda di efficienza, per cui la fase di processing è in realtà divisa in due parti: la prima consiste solo della eliminazione dei vari simboli, come punteggiature, caratteri numerici, ecc.., la seconda, che viene eseguita dopo il parsing, consiste nella eliminazione delle stopwords.

E' ovvio, quindi, che la stemmizzazione delle parole avvenga per ultima, in modo da considerare solo quei termini che effettivamente sono necessari per l'analisi semantica.

Infine il testo in uscita all' extractor viene privato dei tag e quindi memorizzato all'interno del Data Base in una apposita tabella, assieme all'Url, il titolo, la descrizione, il body e le keywords.

6.6 MINER

Una volta che il sistema ha passato la domain keyword al Miner, il DSN Builder costruisce la rete semantica, utilizzando WordNet. Le pagine preprocessate vengono elaborate dal Semantic Grader, che attribuisce ad ogni pagina un voto semantico (SeG) basato sulla reletedness tra i concetti contenuti nella pagina. Il Syntactic-Semantic Grader, invece, attribuisce alla pagina il grado sintattico-semantico che, come vedremo meglio nel paragrafio dedicato alla metrica proposta, si ricava dal peso, calcolato sul livello di polisemia del termine, di ogni parola contenuta nel testo. Entrambi i Grader descritti finora utilizzano le informazioni contenute nella DSN per le loro operazioni.

Il Syntatic Grader attribuisce alle pagine esclusivamente un voto sintattico e per questo motivo non utilizza la DSN, ma calcola il SyG (Syntactic Grade) in base la ranking delle pagine ottenuto dai motori di ricerca tradizionali.

Infine il Global Grader fonde i tre voti (SeG, SSG e SyG) in un unico voto e restituisce all'utente la lista degli URL riordinata.

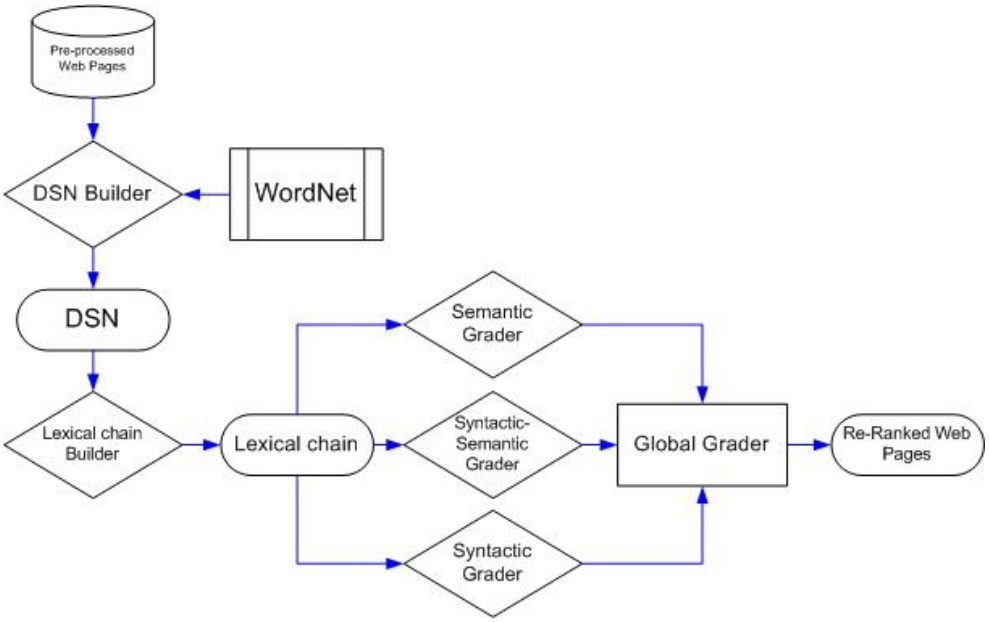


Figura 6.8: Miner

6.6.1 DSN Builder

Il DSN Builder si occupa della costruzione della rete semantica dinamica. Al DSN Builder viene passata la domain keyword; da questa vengono estratti i corrispondenti sensi (polisemia) utilizzando Wordnet. I sensi individuati sono presentati mediante un'apposita interfaccia e viene mostrata anche una descrizione (gloss) estratta sempre da Wordnet; dopo la scelta dell'utente, viene individuato il sysnest corrispondente al senso di interesse.

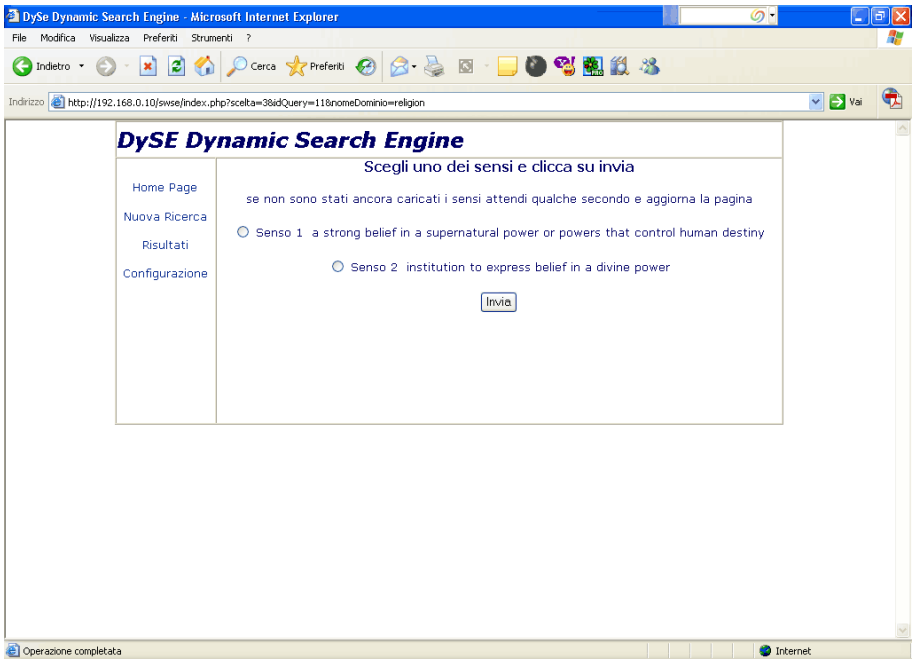


Figura 6.9: Sensi della domain keyword: Religion

6.6.2 La Rete Semantica

La rete semantica costituisce il cuore del mining del sistema. Essa realizza il contesto in cui la ricerca viene operata o, utilizzando la definizione di Schutz [Schutz1970], costituisce l'orizzonte che viene fornito al sistema automatico affinché abbia la base di conoscenza necessaria a riconoscere la rilevanza di un'informazione. E' stato progettato ed implementato un algoritmo ad hoc per l'estrazione della rete semantica utilizzando la struttura di Wordnet.

1. Si parte dal concetto scelto dall'utente e si individuano, attraverso WordNet, il synset corrispondente ed i synset che appartengono allo stesso dominio semantico;
2. Di ogni synset ottenuto si individuano ricorsivamente tutti gli iponimi, cioè per ogni sysnet si individuano gli iponimi di primo livello e da questi

i successivi fino ad arrivare alla base della gerarchia IS-A (gerarchie di generalizzazione/specializzazione) di Wordnet;

3. Da tutti i synset così ottenuti si individuano tutti gli altri concetti (synset) ad essi legati attraverso le altre proprietà linguistiche previste da WordNet (Antinomia, Meronimia/olonimia, implicazione etc.).

In figura viene mostrata una visualizzazione tridimensionale (generata automaticamente dal sistema) della rete semantica.

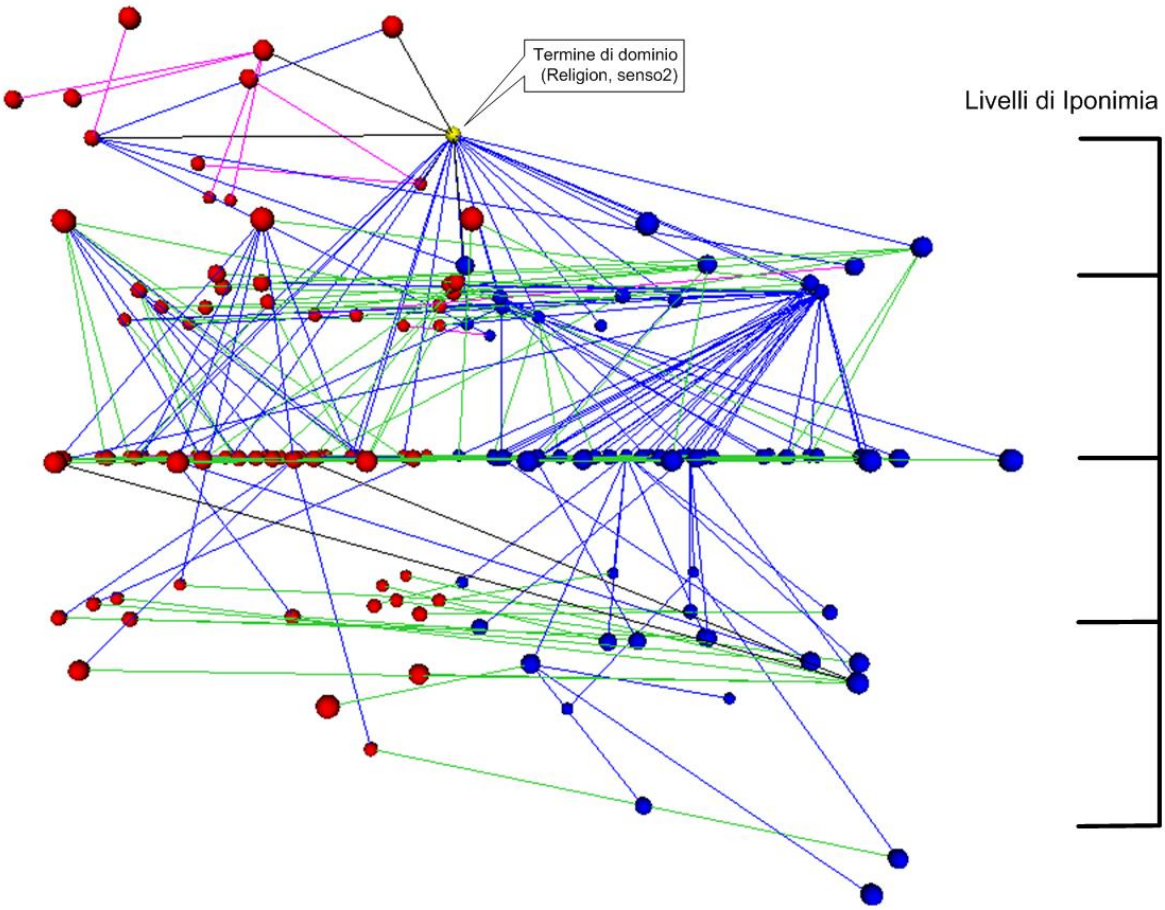


Figura 6.10: Rappresentazione tridimensionale della rete semantica

Il termine di dominio inserito dall'utente è in questo caso "religion" ed il sesno scelto è il numero 2 in WordNet, ossia "institution to express belief in a divine power". Le sfere rappresentano i synset mentre i collegamenti rappresentano le relazioni semantiche. Il synset di colore giallo è il synset principale (individuato dalla *domain_keyword* dell'utente), mentre i synset di colore blu sono quelli che costituiscono l'intelaiatura iperonimica/iponimica della rete. In questo caso la rete si estende su 5 livelli rappresentati nell'immagine da piani paralleli.

Le sfere di colore blu rappresentano invece i synset collegati alla struttura IS-A attraverso tutti gli altri legami semantici previsti da WordNet.

Ad ogni proprietà è collegato un peso , utilizzato per il calcolo della distanza semantica tra due concetti. Il valore dei pesi è stato determinato sperimentalmente.

PROPRIETA'	PESO
ANTONYM	0.8
ATTRIBUTE	0.7
CATEGORY_DOMAIN	1
CAUSE	0.6
DERIVED	0.8
ENTAILED_BY	0.7
ENTAILMENT	0.7
HYPERNYM	0.9
HYPONYM	0.9
MEMBER_HOLONYM	0.5
MEMBER_MERONYM	0.5
MEMBER_OF_CATEGORY_DOMAIN	1
NOMINALIZATION	0.7
PART_HOLONYM	0.7
PART_MERONYM	0.7
PARTICIPLE_OF	0.7
SEE_ALSO	0.6
SIMILAR_TO	0.5
SUBSTANCE_HOLONYM	0.5
SUBSTANCE_MERONYM	0.5
SYNONYMY	1

Tabella 6.1: Proprietà linguistiche e pesi

6.6.3 La metrica per il ranking delle pagine

Per descrivere la metrica utilizzata per il ranking delle pagine introduciamo le definizioni di tre diverse tipologie di voti.

Essi prendono in considerazione tre aspetti diversi dell'analisi semantica ossia: la polisemia dei termini presenti nel testo, il livello di relazione (relatedness) tra i concetti presenti nel testo ed infine il ranking assegnato alle pagine dai motori di ricerca tradizionali.

Definizione del Syntactic-Semantic Grade (SSG)

Data una pagina *v* il Synctatic-Semantic Grade di *v* è definito come:

$$SSG(v) = \sum_{i=0}^{n} \overline{w}(i)$$

Dove *n* è il numero di termini contenuti nella lexical-chain estratta da *v* e $\overline{w}(i) = \frac{1}{poly(i)}$ è l'inverso della polisemia del termine i-esimo.

La giustificazione dell'uso della proprietà di polisemia sta nel fatto che se una parola ha un solo senso, essa esprime in maniera forte un determinato concetto;

quindi, più una parola è associabile a più concetti, minore sarà la sua importanza in un determinato contesto.

Definizione del Semantic Grade (SeG)

Data una pagina *v* il Semantic Grade di v è definito come:

$$SeG(v) = \sum_{(w_i,w_j)} e^{-\alpha \cdot l(w_i,w_j)} \frac{e^{\beta \cdot d(w_i,w_j)} - e^{-\beta \cdot d(w_i,w_j)}}{e^{\beta \cdot d(w_i,w_j)} + e^{-\beta \cdot d(w_i,w_j)}}$$

Dove (w_i, w_j) sono coppie di parole nella lexical chain, α e β sono dei parametri di scala che sono stati definiti attraverso degli esperimenti. La *l* è la distanza semantica tra due concetti all'interno della DSN, definita come il percorso minimo che si deve effettuare all'interno della rete per giungere da un concetto all'altro. Ogni percorso va calcolato sommando i pesi degli archi che collegano i nodi attraversati. Il valore del peso di ogni arco è rappresentato dall'inverso di σ attribuito alla relazione semantica rappresentata dall'arco stesso. Possiamo esprimere sinteticamente la *l* come:

$$l(w_i, w_j) = \min_j \sum_{i=1}^{h_j(w_1,w_2)} \frac{1}{\sigma_j}$$

La *d* invece rappresenta la profondità (depth) relativa a due concetti, definità come la distanza minima, espressa come numero di nodi attraversati, tra l'iperonimo comune ai due concetti considerati ed il nodo radice della gerarchia in Wordnet. Per calcolare la *d* viene utilizzata solo la gerarchia IS-A.

La correlazione tra i termini è calcolata attraverso una funzione non lineare. La scelta di tale funzione deriva da varie considerazioni. I valori della lunghezza del path e della profondità del sovraordinato comune possono variare, secondo la loro definizione, tra zero ed infinito, mentre la similarità semantica tra due termini dovrebbe essere espressa da un numero relae compreso nell'itervallo [0,1] (0=non esiste nessuna correlazione; 1=sono completamente correlati).

In particolare, quando la lunghezza del path va a 0, la similitudine dovrebbe crescere monotonicamente a 1 e dall'altra parte essa dovrebbe andare a zero se la lunghezza del path va ad infinito. Inoltre, dato che le parole presenti nei livelli più alti della gerarchia IS-A esprimono concetti più generali rispetto a quelli nei livelli più bassi, usiamo una funzione non lineare per scalare verso il basso il contributo dei sovraordinati nei livelli alti e viceversa per quelli nei livelli bassi.

Da queste considerazioni viene utilizzata una funzione esponenziale che soddisfi i vincoli discussi precedentemente. Inoltre tale scelta è supportata anche

dal lavoro di Shepard [Shepard1987] che ha dimostrato che le funzioni esponenziali decrescenti sono una legge universale nelle scienze cognitive.

Definizione del Syntactic Grade (SyG)

Data una pagina *v* il Synactic Grade di *v* è definito come:

$$SyG(v) = \frac{1}{\sum_{i=1}^{n_{se}} \varphi_i \cdot p_i}$$

Dove n_{se} è il numero di motori di ricerca, φ_i è il peso assegnato all'i-esimo motore di ricerca e p_i è la posizione della pagina web *v* nel ranking prodotto dall'i-esimo motore di ricerca.

Nei nostri test questo contributo non è utilizzato dato che si voleva conservare una completa indipendenza dai motori utilizzati.

Definizione del Global Grade

Data una pagina *v* il Global Grade di *v* è definito come:

$$GG(v) = K_{SyG} \cdot SyG(v) + K_{SSG} \cdot SSG(v) + K_{SeG} \cdot SeG(v)$$

Dove K_{SyG}, K_{SSG} e K_{SeG} sono rispettivamente i pesi relativi assegnati a SyG, al SSG e al SeG.

E' possibile assegnare anche un peso per ogni tag significativo considerato.

I valori dei pesi della metrica e dei tag è stato calcolato sperimentalmente.

6.7 BASE DATI DEL SISTEMA

La base dati di supporto al sistema è formata essenzialmente dalle seguenti parti:

- dati relativi ai Motori di Ricerca che è possibile utilizzare;
- dati relativi alle caratteristiche delle query richieste al sistema;
- dati relativi allo stato di elaborazione dei server;
- Web Repository;
- Informazioni statiche necessarie al funzionamento del livello applicativo.

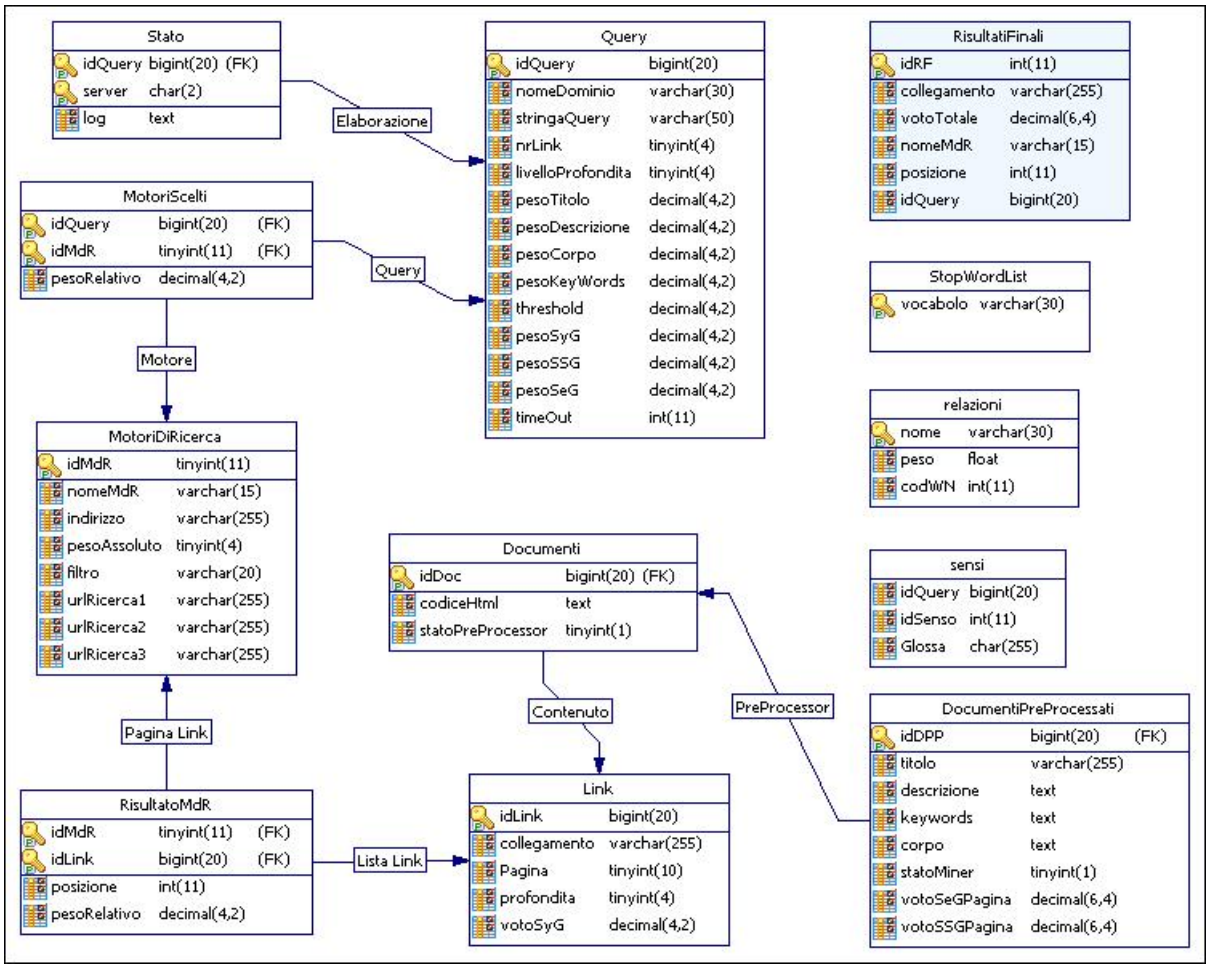


Figura 6.11: Schema del Repository

Nello schema l'entità colorata rappresenta una vista materializzata che riporta informazioni di sintesi a cui accede la pagina di presentazione dei risultati finali.

6.7.1 Repository

Il Repository rappresenta tutti i dati relativi ai motori, alle query e ai documenti da elaborare o elaborati.

Le entità che vi troviamo sono:

1) Query: le sue occorrenze rappresentano le query sottoposte al sistema.

2) MotoriDiRicerca: è l'entità associata ai motori di ricerca.

3) Link: ogni occorrenza rappresenta un link, inteso come universal reference locator relativo a documenti individuati e scaricati nel sistema.

4) Documenti: Rappresenta le pagine web individuate e memorizzate con tutto il codice HTML.

5) DocumentiPreProcessati: Le occorrenze di tale entità si riferiscono a documenti già preprocessati e pronti all'elaborazione semantica.

6) Stato: Indica, per ogni query, lo stato di elaborazione dei server. Ogni occorrenza è relativa ad una query e ad un server. Se è presente indica che quel server, per quella query ha terminato l'elaborazione, altrimenti è ancora in corso.

Inoltre le associazioni permettono di indicare l'insieme dei motori scelti per una data query (MotoriScelti, indicando per ogni associazione il peso relativo assegnato) e di rappresentare l'insieme dei motori che hanno individuato un determinato Link (RisultatiMdR, indicando per ogni associazione la posizione assunta dal link nel motore di ricerca e il peso relativo del motore).

Di seguito indichiamo alcuni dettagli sulle entità del Repository.

6.7.1.2 MotoriDiRicerca

I primi tre attributi rappresentano l'identificativo primario (idMdR), il nome (nomeMdR) e l'indirizzo http (indirizzo) del motore di ricerca.

Vi è poi, il pesoAssoluto, un valore numerico da 1 a 10 che rappresenta l'efficacia che l'amministratore assegna al motore di ricerca. Tale informazione viene utilizzata per consigliare all'utente il peso relativo da assegnare.

Un altro attributo molto importante è il filtro, che consente di filtrare dalla pagina dei risultati i link che non sono attinenti alla ricerca. Di solito esso coincide con il nome del motore di ricerca (per Google sarà google).

Ogni motore di ricerca, tra l'altro, si contraddistingue per il modo in cui rappresenta la congiunzione tra vocaboli e per il modo in cui gestisce l'elenco totale dei link trovati, ovvero il numero di link per pagina.

Per rendere la gestione dei motori di ricerca il più semplice possibile, è necessario per ogni motore inserire tre url-http:

- urlRicerca1;
- urlRicerca2;
- urlRicerca3.

Queste informazioni riguardano le differenti sintassi utilizzate dai motori di ricerca per la gestione del salto pagina e della congiunzione.

6.7.1.3 Query

Tra i vari attributi si evincono:

- idQuery: è l'identificativo della query;
- stringaQuery: rappresenta l'insieme di vocaboli da ricercare;
- nomeDominio: indica il dominio di appartenenza della ricerca;

- nrLink: rappresenta il numero massimo dei link che deve restituire ogni motore di ricerca;
- timeOut: rappresenta il tempo massimo (in msec) di attesa di una pagina Web.

6.7.1.4 Link

Il link rappresenta il vero risultato della ricerca. In tale entità vengono memorizzati tutti i voti calcolati. A tale entità si farà riferimento per avere il set di risultati.

I voti verranno inseriti dai server in fasi diverse, non necessariamente consecutive. Solo quando sarà stata determinata la posizione media e i voti semantici sarà possibile determinare il voto finale.

L'assegnazione dei voti avviene utilizzando la DSN. Solo in questa entità del repository si riflettono le elaborazioni sintattiche e semantiche sulla pagina web: tali informazioni rappresentano l'interfaccia tra la pagina web e la metrica utilizzata.

6.7.1.5 Documenti e DocumentiPreProcessati

In tali entità si evincono due attributi importanti:

- statoPreProcessor: è un valore booleano e indica, se vero, che il documento è stato preprocessato e vi è una sua occorrenza in DocumentiPre-Processati.
- statoMiner: è un valore booleano e indica, se vero, che il documento è stato oggetto di mining.

Tali attributi permettono ai server di comprendere quali documenti bisogna ancora elaborare.

L'attributo server indica il tipo di server: *SS* per Server Spider, *SP* per Server Pre-Processor ed *SM* per Server Miner.

Per ogni query vengono inserite tre occorrenze, ognuna relativa ad un server. Se l'occorrenza è presente significa che il server ha terminato l'elaborazione della query.

Nell'attributo log è possibile inserire informazioni relative ai tempi di elaborazione di una data query, per un dato server. In questo modo è possibile analizzare le prestazioni del sistema e individuare eventuali "colli di bottiglia".

6.8 TECNOLOGIE UTILIZZATE

Le tecnologie utilizzate per la realizzazione del meta-motore sono state:

Linguaggi: Java per la realizzazione dei moduli principali, JSP per l'interfacciamento tra i moduli e le pagine web utilizzate come interfaccia utente, PHP per la realizzazione delle pagine di interfaccia utente.

DBMS: MySQL 5.0

Ambienti di sviluppo: SunOne Studio per Java e JSP, Macromedia Dreamweaver per il PHP

Programmazione distribuita: RMI (Remote Method Invocation) di Java

Per l'interfacciamento con il Database Lessicale Wordnet è stata utilizzata la libreria JWNL (Java Word Net Library).

CAPITOLO 7 RISULTATI SPERIMENTALI E VALUTAZIONE DEL SISTEMA

La necessità di testare in maniera adeguata i sistemi di information retrieval impone l'adozione di metodologie che rendono ripetibili e comparabili i test effettuati [vanRijsbergen1979].

Per mettere in luce il problema della valutazione, occorre rispondere a tre domande principali:

- perché valutare
- cosa valutare
- come valutare

La risposta alla prima domanda è principalmente di carattere sociale ed economico: occorre definire una misura dei benefici o degli svantaggi che si possono ottenere dai sistemi di reperimento delle informazioni, quanto costa usare uno di questi sistemi e soprattutto se ne vale la pena.

Alla seconda domanda diede una risposta Cleverdon [Cleverdon1966], elencando sei principali quantità misurabili:

1. il riempimento della collezione, ossia la quantità di informazioni di base da apprendere sul sistema;
2. il ritardo, cioè l'intervallo medio che intercorre fra il tempo in cui viene fatta la richiesta di ricerca ed il tempo in cui viene data la risposta;
3. la forma della presentazione dei risultati;
4. lo sforzo compiuto da parte dell'utente per ottenere le risposte alle sue richieste di ricerca;
5. la recall del sistema, cioè, la proporzione di materiale rilevante che viene reperito in risposta ad una richiesta di ricerca;
6. la precision del sistema, cioè, la proporzione di materiale reperito che è effettivamente rilevante.

È facile osservare che le prime quattro misure siano prontamente valutabili, mentre la precision e la recall tentano di misurare quella che è conosciuta come l'efficacia del sistema di ricerca dell'informazione. In altre parole è una misura della capacità del sistema di ricercare i documenti rilevanti ed allo stesso tempo tenere "lontani" quelli non rilevanti. Si assume quindi che la precision e la recall siano sufficienti per la misura di efficacia del sistema.

La risposta alla domanda finale, cioè come valutare, tiene in conto che le tecniche di misura dell'efficacia sono in gran parte influenzate dalla particolare strategia di ricupero adottata e dalla forma del suo output.

La Precision e Recall [vanRijsbergen1979] sono due criteri standard di valutazione dei sistemi di information retrieval.

Abbiamo una collezione di oggetti all'interno della quale fare una particolare ricerca e di conseguenza estrarre solo quegli oggetti che rispondono a determinate regole; nel nostro caso gli oggetti a disposizione sono rappresentati dalle pagine web che verranno utilizzate per la sperimentazione e le regole in base alle quali fare la ricerca tra queste pagine sono rappresentate dal contenuto semantico del documento stesso.

Questo è lo scopo della nostra sperimentazione: capire fino a che punto gli algoritmi implementati, con le relative metriche, riconoscono le pagine nelle quali è espresso un certo contenuto semantico (nel nostro caso cercheremo tutte le pagine in un determinato dominio).

Allora, dato un set di pagine effettuiamo la nostra ricerca, come risultato si otterrà un sottoinsieme di pagine trovate (retrieved), che secondo l'algoritmo applicato dovrebbero essere attinenti al dominio di ricerca.

In realtà soltanto una parte di queste pagine sarà realmente attinente al dominio d'interesse: la precisione sarà rappresentata dall' intersezione tra pagine trovate (valutate attinenti) e pagine realmente attinenti (relevant). La rimanente parte sarà composta da pagine non attinenti che in realtà sono state scambiate per pagine attinenti .

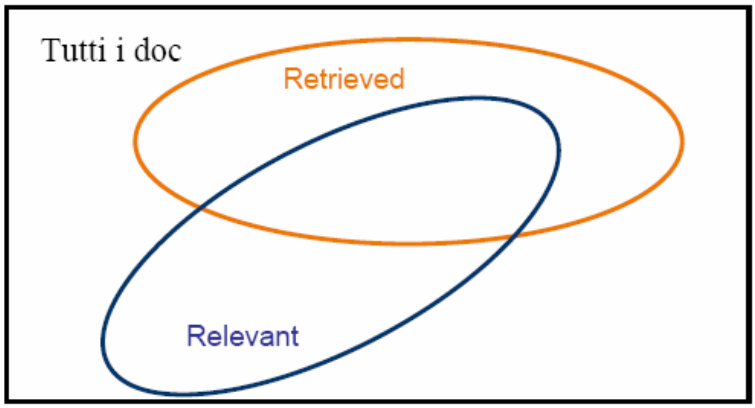


Figura 7.1: relazioni tra le pagine sottoposte ad analisi

Allora, dette:

- Retrieved: le pagine trovate (presumibilmente attinenti);
- Relevant: pagine realmente attinenti nella collezione;

- RelRetrieved: le pagine trovate realmente attinenti (intersezione tra Retrieved e Relevant),

si ha che:

$$precision = \frac{relretrieved}{retrieved} \qquad recall = \frac{relretrieved}{relevant}$$

Quindi il parametro recall viene calcolato sui documenti rilevanti recuperati, e ne misura la percentuale rispetto al totale contenuto nella raccolta. Il parametro precision invece riguarda i documenti restituiti dalla ricerca e rappresenta la percentuale di documenti rilevanti.

7.1 IL TEST SET

Per eseguire le prove è stato utilizzato un test set composto da pagine reperite dal web mediante il motore di ricerca Yahoo, il quale permette la ricerca per directory ritornando le categorie in cui sono raggruppate le pagine contenenti la keyword immessa. In questo modo è possibile associare l'Url alla categoria di appartenenza assegnatagli dal motore, necessaria per il confronto a posteriori con i risultati ottenuti dalla applicazione delle varie metriche.

La collezione comprende quasi 800 pagine, ed è stata definita inserendo nel motore keywords che avessero un elevato grado polisemico, in modo che le relative pagine appartenessero a categorie diverse.

Keyword	Categorie
Angels	Show (18), Sport(10), Religion (18), Movie (11), Music (6), Motorcycles (9)
Apple	Computer (19), Fruit (8)
Cellular	Telephone (65), Science (22), Movies (2)
Davis	Actors (28), Sport (5)
Eagles	Music (6), Sport (9), Bird (2), Car (1), Scout (6), Service (8)
Fan	Actor (14), Animation (4), Movie (12), Music (1), Sport (6)
Fisher	Actor (20), Sport (34), Toys (7)
Hercules	Movie (11), Military (12), Mitology (5)
Jaguar	Animal (4), Car (40), Sport (9), Game (4)
Lincoln	History (4), Car (2)
Lion	Anmial (4), Sport (8), Music (10), Art (10), Service (30)
Mouse	Computer (15), Animal (19), Animation (18)
Simpson	Animation (40), Music (13)
Sun	Astronomy (29), Computer (6), Healt (28)
Table	Sport (4), Science (23), Entertaining (7)

Tabella 7.1: Test Set

Di questo insieme sono state considerate alcune keyword e alcune pagine legate a:

- argomenti generali
- argomenti più specifici

Questo tipo di distinzione è stata fatta per testare l'efficacia del sistema in ambiti diversi; poiché WordNet possiede un dizionario non specializzato, si è voluto considerare le prestazioni con pagine di tipo "generico" e pagine relative a settori più specialistici.

Nella tabella seguente, per ciascuna parola chiave sono riportate le categorie in cui sono raggruppate le corrispondenti pagine e tra parentesi il numero di Url per ciascuna categoria.

Keyword	Categorie
Cellular	Telephone (14), Science (6)
Jaguar	Animal (3), Car (7), Sport (6), Game (4)
Mouse	Computer (7), Animal (6), Animation (7)
Simpson	Animation (14), Music (4)
Sun	Astronomy (13), Computer (4), Healt (3)
Table	Sport (8), Science (9), Entertaining (7)

Tabella 7.2: Sub-Test Set

Per un confronto con sistemi e metriche esistenti similari alla nostra, è stato fatto un confronto con la letteratura e sono state utilizzate le seguenti metriche già descritte nei capitoli introduttivi:

- Rada et. al.;
- Leacock and Chodorow;
- Wu and Palmer.

Dato che queste metriche lavorano con parti del discorso differenti, nella fase di pre-processing la tipologia di testo necessaria è stata archiviata in tabelle apposite.

7.2 ESPERIMENTI

In questo paragrafo verranno riportate le query utilizzate negli esperimenti e i relativi grafici di precision-recall.

Le query sono state progettate tenendo conto di alcuni fattori caratteristici:

- Quantità di testo nelle pagine
- Generalità/Specializzazione del dominio (ampiezza della rete semantica)

Per ogni senso scelto vengono riportati i sinonimi e la descrizione fornita da WordNet.

Query 1

Keyword: **Cellular** Dominio: **DNA**: *deoxyribonucleic acid, desoxyribonucleic acid, DNA -- ((biochemistry) a long linear polymer found in the nucleus of a cell and formed from nucleotides and shaped like a double helix; associated with the transmission of genetic information; "DNA is the king of molecules")*

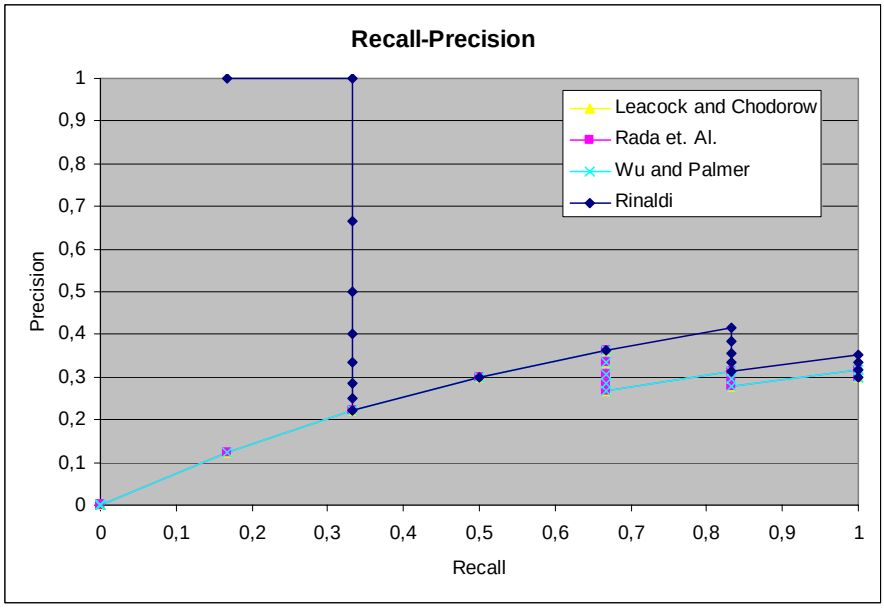


Figura 7.2: Precision-Recall Keyword: Cellular Domain: DNA

Query 2

Keyword: **Cellular** Dominio: **Telephone**: *telephone, phone, telephone set -- (electronic equipment that converts sound into electrical signals that can be transmitted over distances and then converts received signals back into sounds; "I talked to him on the telephone")*

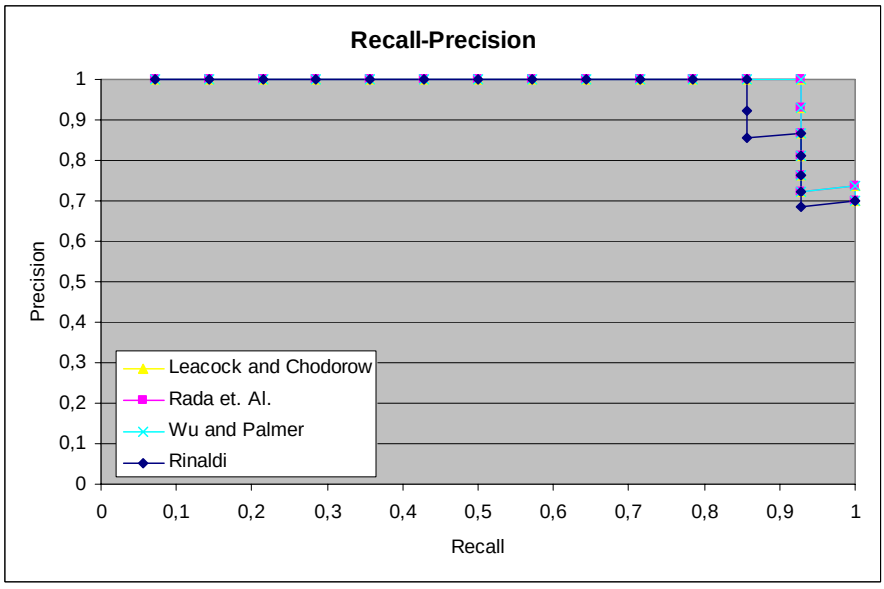


Figura 7.3: Precision-Recall Keyword: Cellular Domain: Telephone

Query 3

Keyword: **Jaguar** Dominio: **Car:** *auto, automobile, machine, motorcar -- (4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work")*

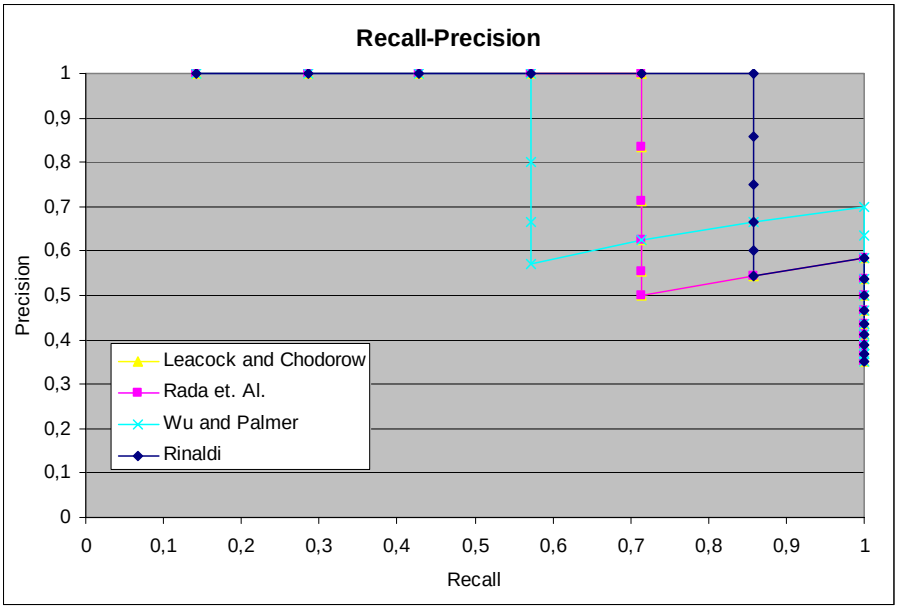


Figura 7.4: Precision-Recall Keyword: Jaguar Domain: Car

Query 4

Keyword: **Jaguar** Dominio: **Mammal***: mammal -- (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)*

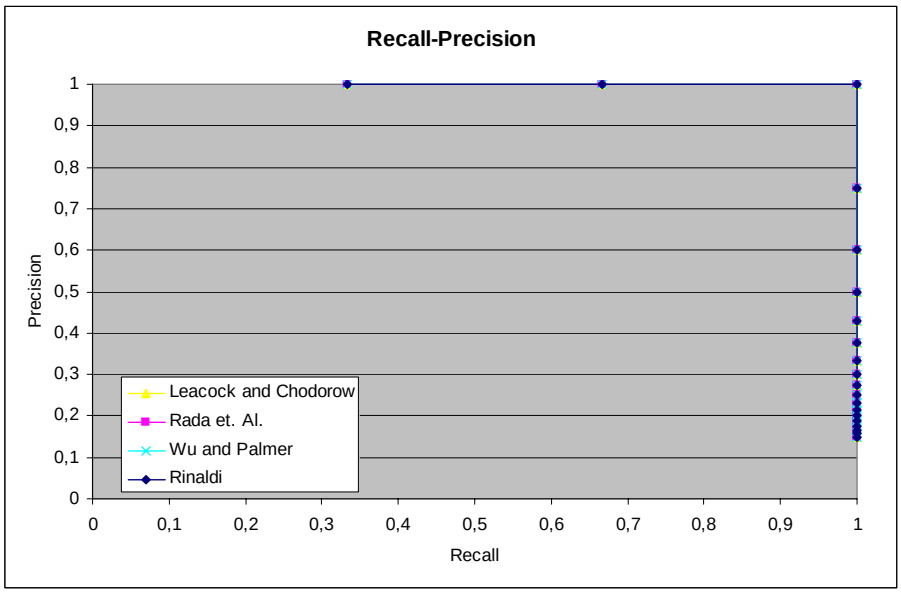


Figura 7.5: Precision-Recall Keyword: Jaguar Domain: Mammal

Query 5

Keyword: **Mouse** Dominio: **Animation:** *animation -- (the making of animated cartoons)*

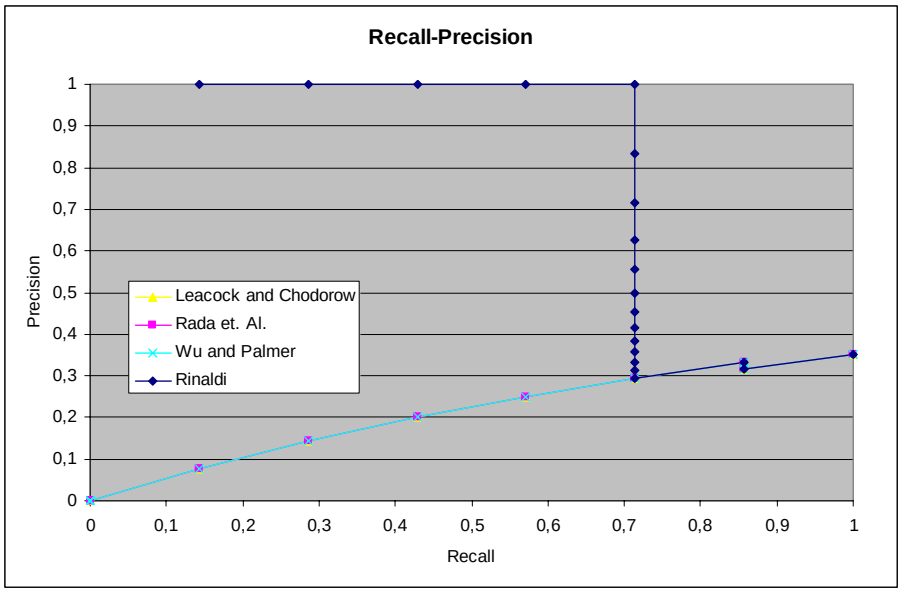


Figura 7.6: Precision-Recall Keyword: Mouse Domain: Animation

Query 6

Keyword: **Mouse** Dominio: **Mammal:** *mammal -- (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)*

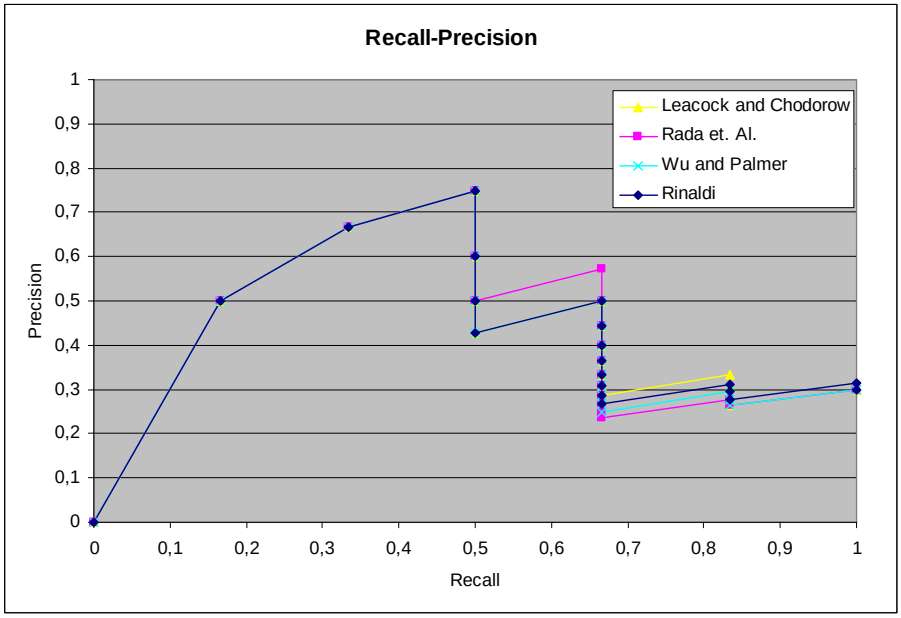


Figura 7.7: Precision-Recall Keyword: Mouse Domain: Mammal

Query 7

Keyword: **Simpson** Dominio: **Animation:** *animation -- (the making of animated cartoons)*

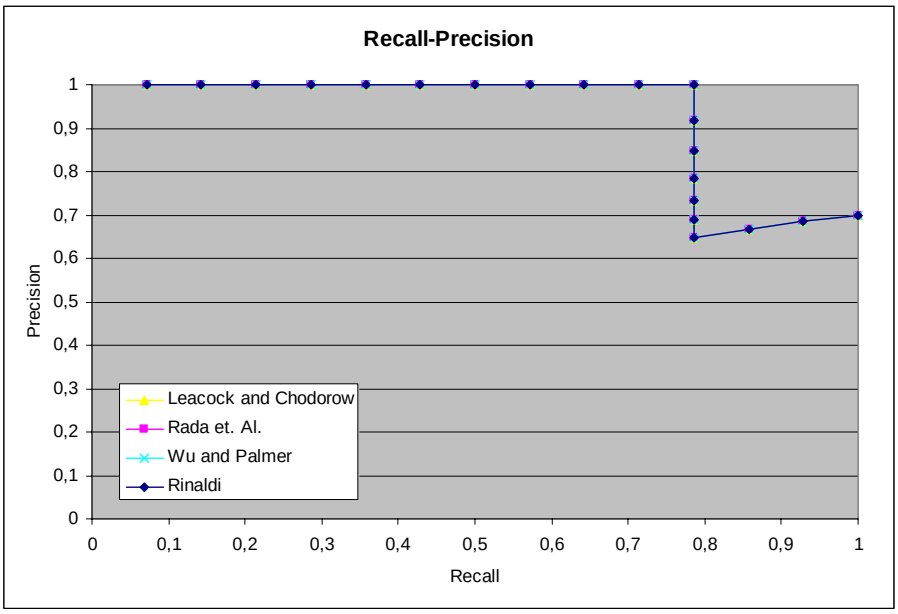


Figura 7.8: Precision-Recall Keyword: Simpson Domain: Animation

Query 8

Keyword: **Simpson** Dominio: **Music**: *music -- ((music) the sounds produced by singers or musical instruments (or reproductions of such sounds))*

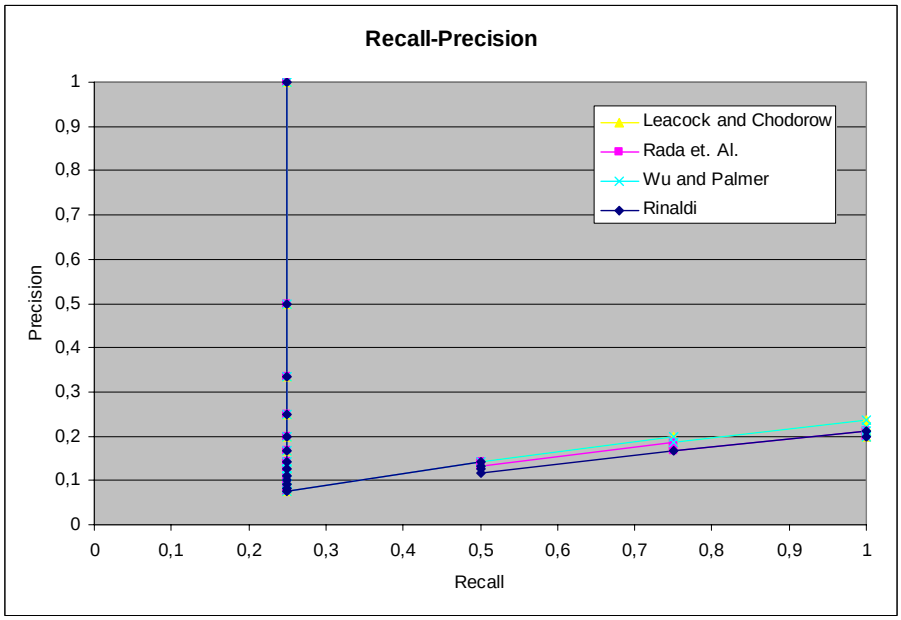


Figura 7.9: Precision-Recall Keyword: Simpson Domain: Music

Query 9

Keyword: **Sun** Dominio: **Astronomy**: *astronomy, uranology -- (the branch of physics that studies celestial bodies and the universe as a whole)*

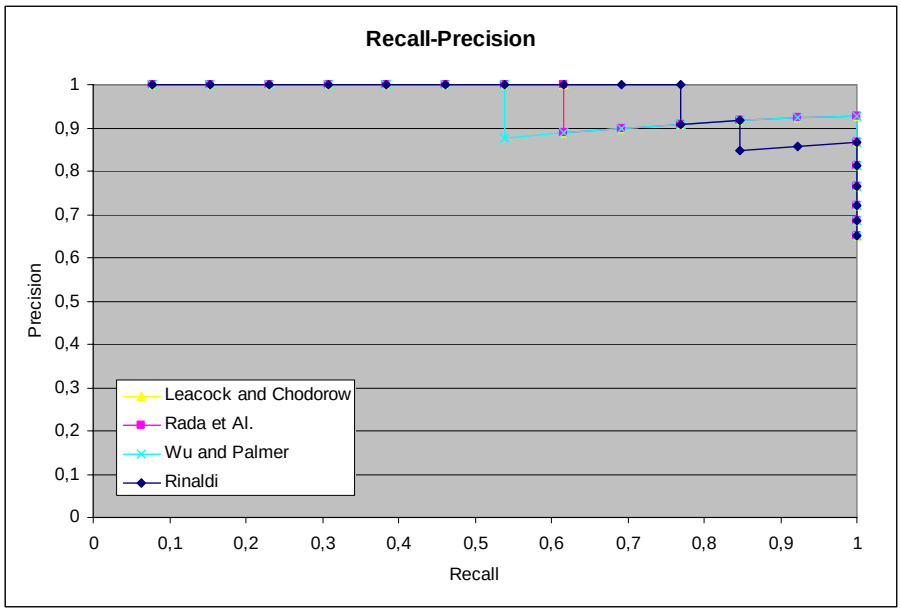


Figura 7.10: Precision-Recall Keyword Sun Domain Astronomy

Query 10

Keyword: **Sun** Dominio: **Computer**: *computer, computing machine, computing device, data processor, electronic computer, information processing system -- (a machine for performing calculations automatically)*

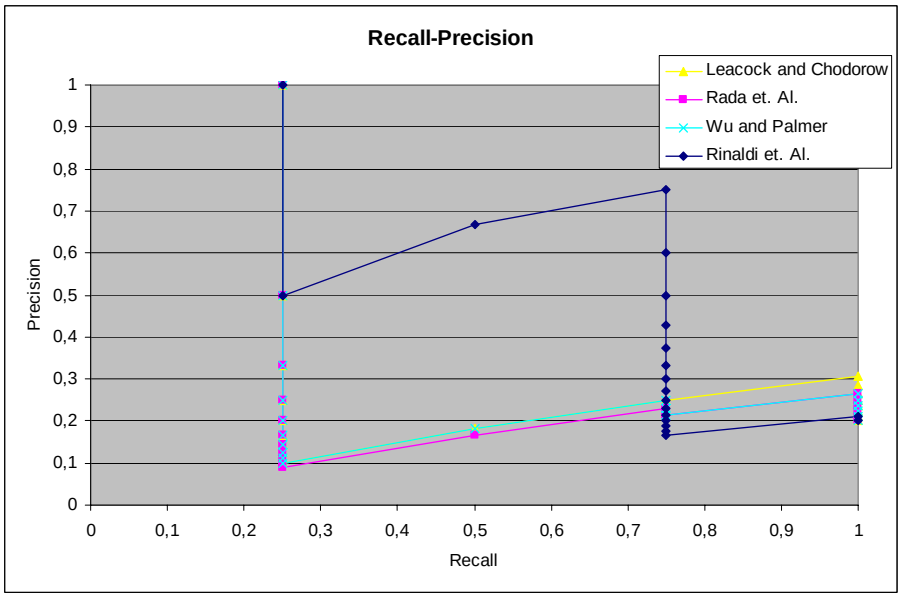


Figura 7.11: Precision-Recall Keyword: Sun Domain: Computer

Query 11

Keyword: **Table** Dominio: **Chemical**: *chemical -- (produced by or used in a reaction involving changes in atoms or molecules)*

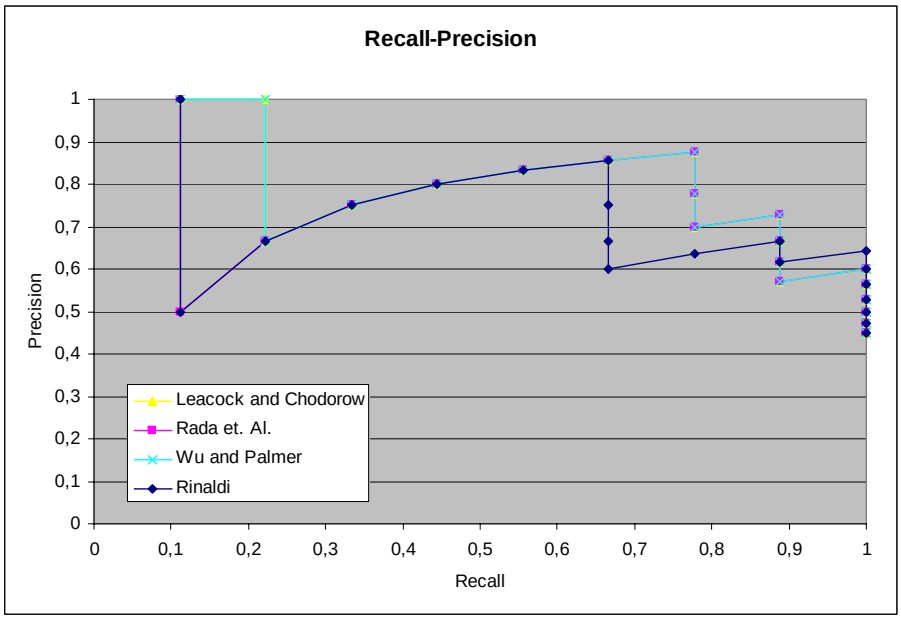


Figura 7.12: Precision-Recall Keyword: Table Domain: Chemical

Query 12

Keyword: **Table** Dominio: **Garden**: *garden -- (a plot of ground where plants are cultivated)*

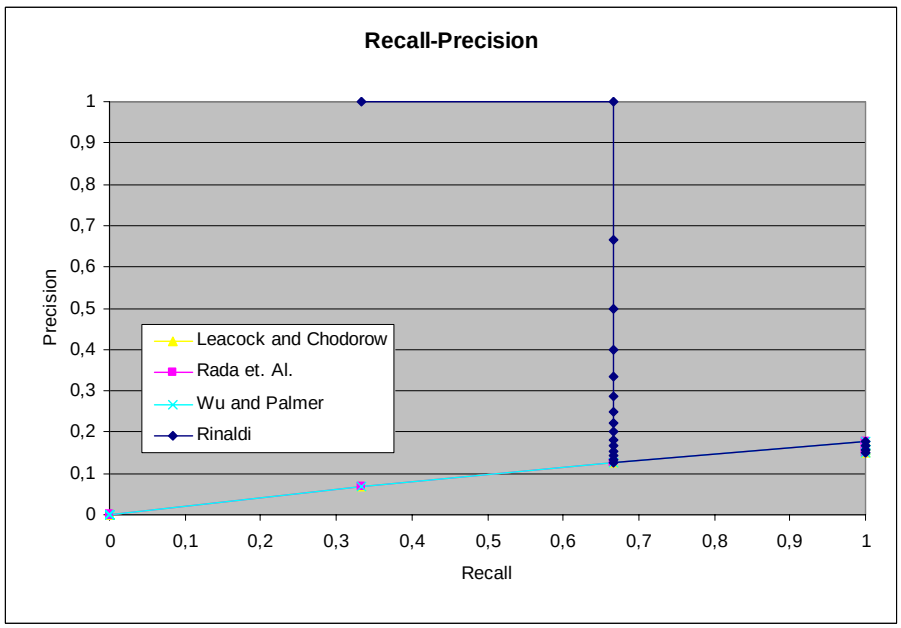


Figura 7.13: Precision-Recall Keyword: Table Domain: Garden

Le performance complete sono state individuate utilizzando una tecnica descritta in [vanRijsbergen1979] che ha portato ad un grafico finale che tiene conto di tutte le query considerate e dei rispettivi risultati.

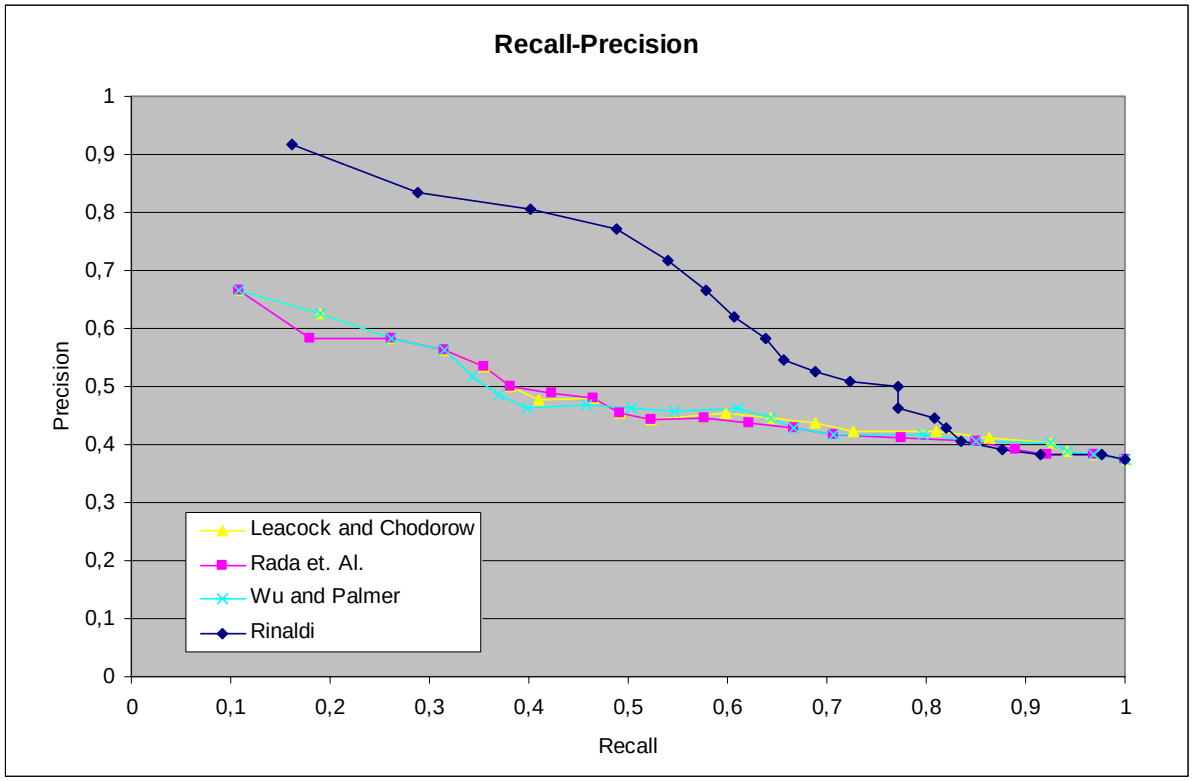


Figura 7.14: Performance complessiva su tutto il test set

7.3 INTERROGAZIONE DIRETTA DI MOTORI DI RICERCA

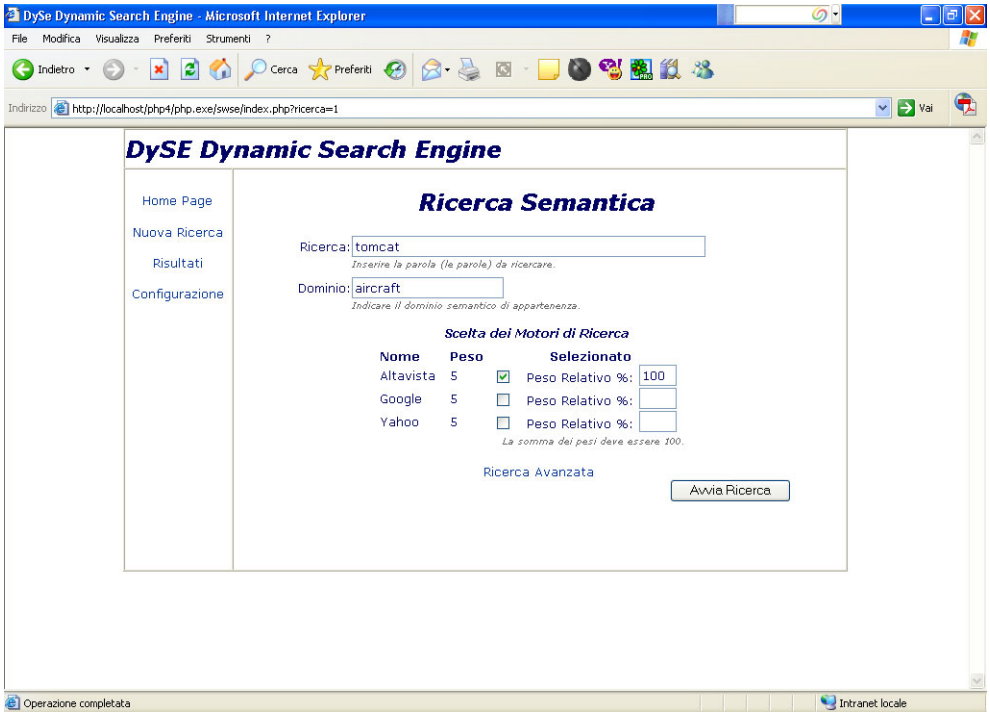
In questo paragrafo verranno mostrati alcuni risultati derivanti dall'interrogazione diretta di motori di ricerca.

Prendiamo come esempio due query: una che utilizza come keyword *tmcat* e come domain_keyword *aircraft*, e l'altra che utilizza *sun* in un contesto astronomico.

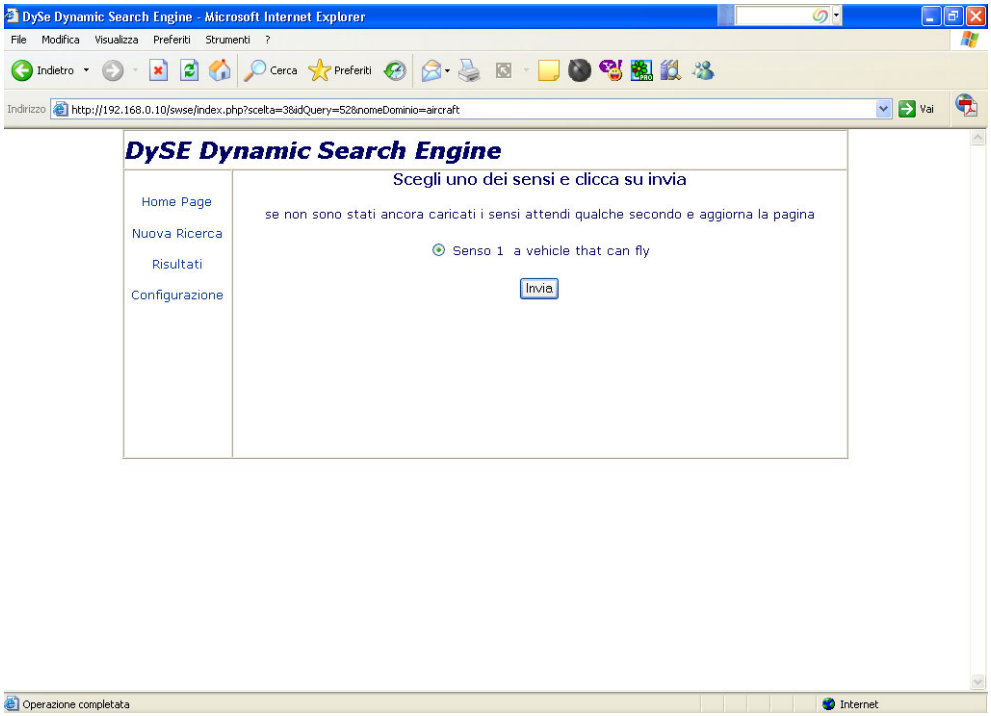
Verrà presentata l'interfaccia di accesso standard, dato che verranno utilizzati i parametri standard mostrati nella schermata per utenti avanzati mostrata nel capitolo precedente.

La prima schermata riguarda l'inserimento delle keyword di ricerca e la scelta del motore.

In questi esempi abbiamo utilizzato Altavista, ma nulla sarebbe cambiato se avessimo altri motori o più di un motore.

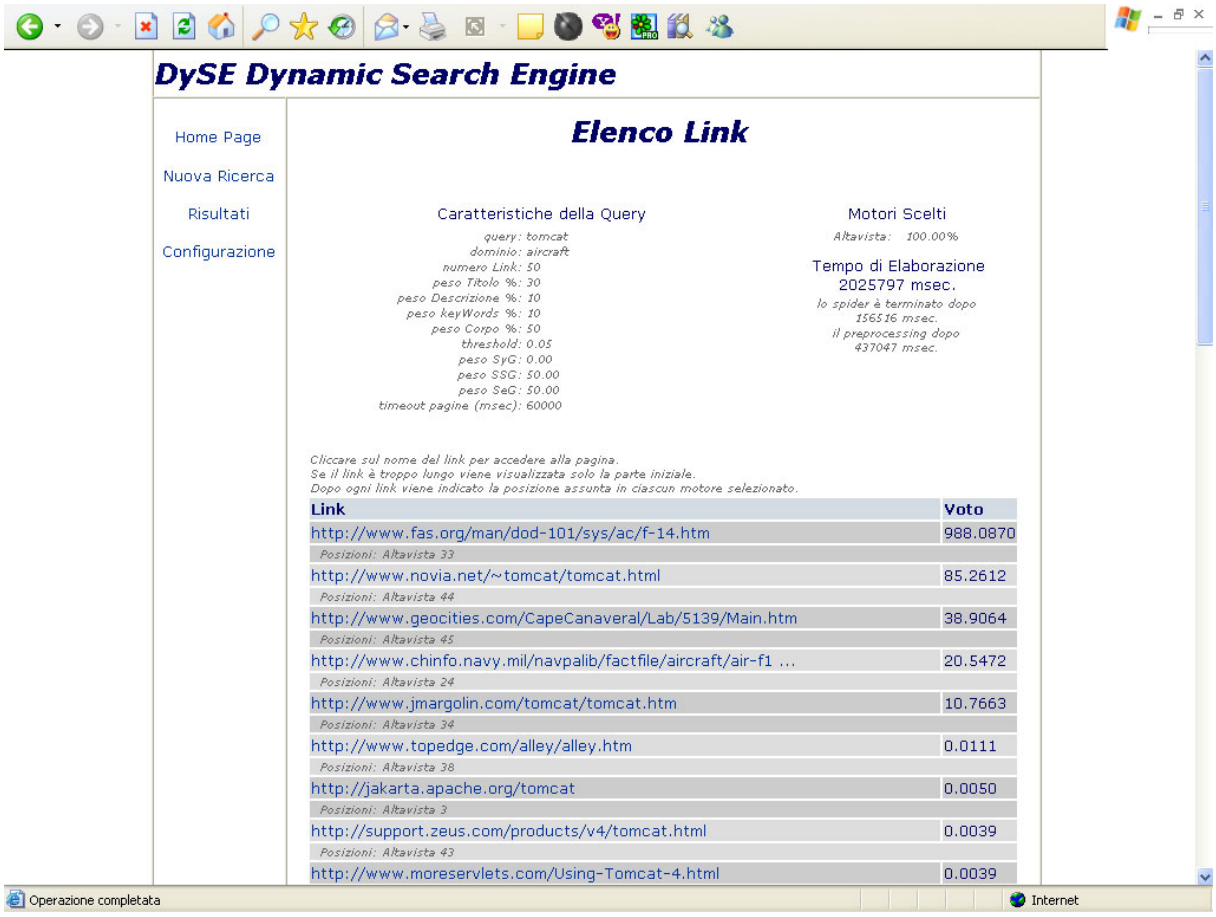


La schermata successiva mostra la scelta del senso per la costruzione della rete semantica.

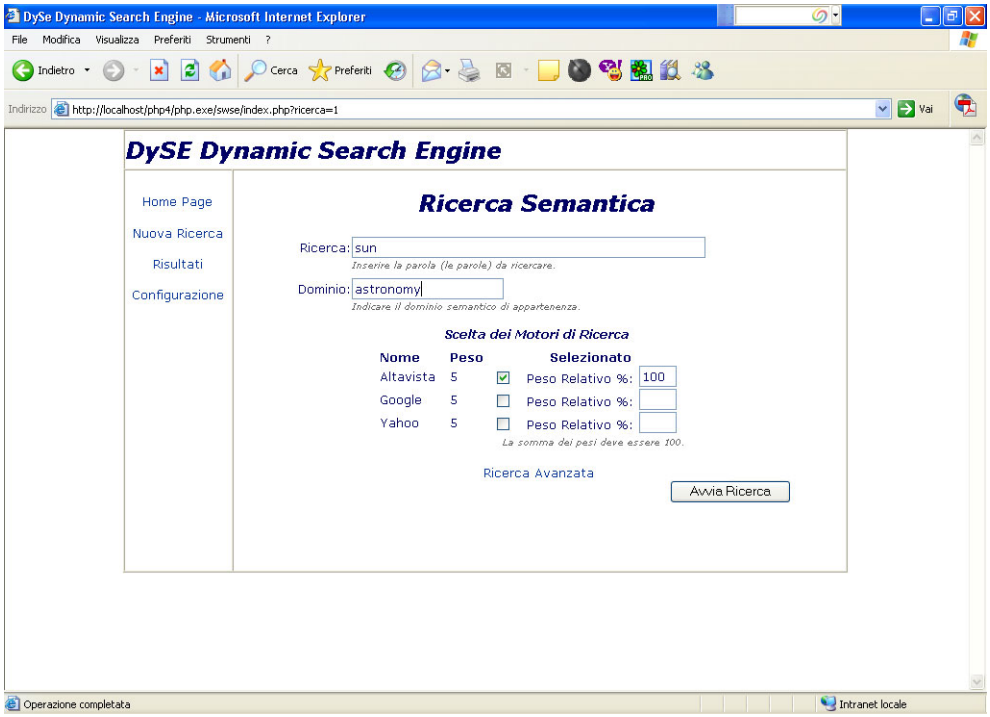


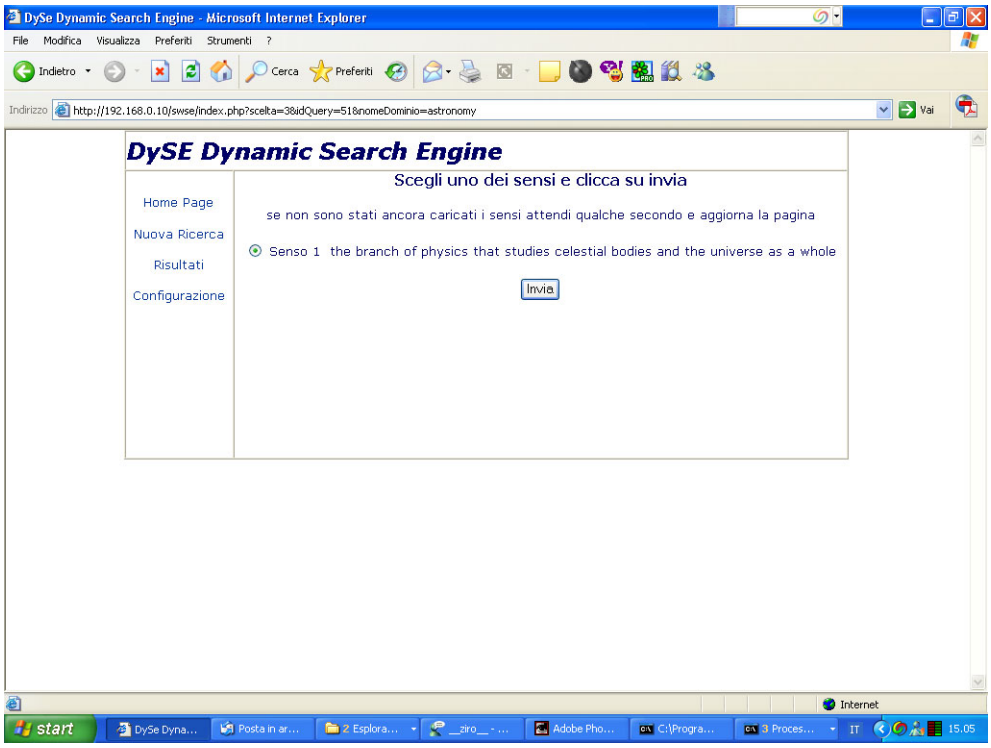
Infine vengono mostrati i risultati insieme alle impostazioni di ricerca. Come si evince dalla figura successiva il nostro motore porta nelle prime posizioni pagine

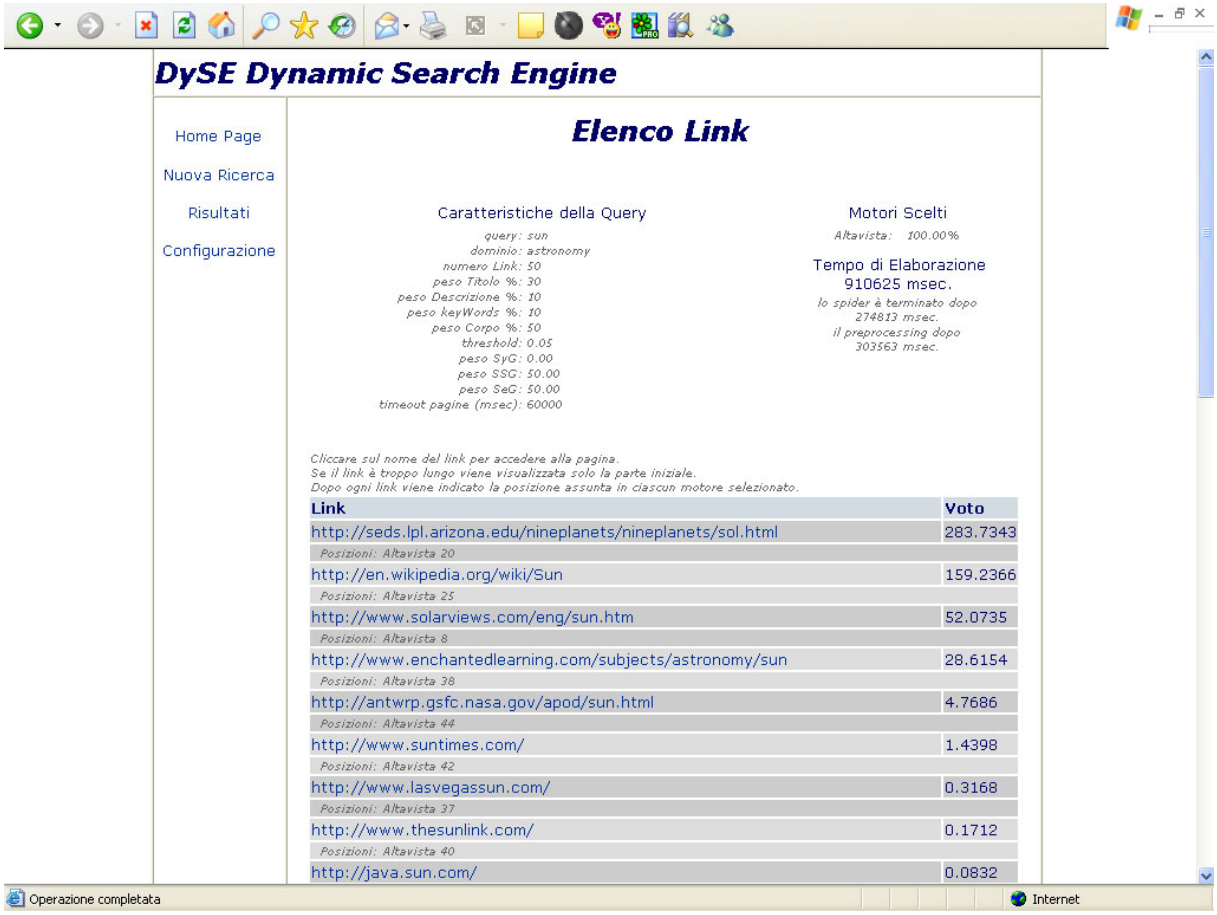
che il motore utilizzato riportava tra la 24 e la 44 posizione. Inoltre il ranking assegna una misura della rilevanza della pagina nel dominio considerato.



In maniera analoga viene effettuata la ricerca per la seconda query e ne vengono mostrati i risultati.




DySe Dynamic Search Engine - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti
Indirizzo http://192.168.0.10/swse/index.php?scelta=3&idQuery=518nomeDominio=astronomy
Vai
DySE Dynamic Search Engine
Home Page
Nuova Ricerca
Risultati
Configurazione
Scegli uno dei sensi e clicca su invia
se non sono stati ancora caricati i sensi attendi qualche secondo e aggiorna la pagina
Senso 1 the branch of physics that studies celestial bodies and the universe as a whole
Invia
Internet
start
DySe Dyna...
Posta in ar...
2 Esplora...
ziro - ...
Adobe Pho...
C:\Progra...
3 Proces...
IT
15.05


DySE Dynamic Search Engine
Home Page
Nuova Ricerca
Risultati
Configurazione
Elenco Link
Caratteristiche della Query
query: sun
dominio: astronomy
numero Link: 50
peso Titolo %: 30
peso Descrizione %: 10
peso keyWords %: 10
peso Corpo %: 50
threshold: 0.05
peso SyG: 0.00
peso SSG: 50.00
peso SeG: 50.00
timeout pagine (msec): 60000
Motori Scelti
Altavista: 100.00%
Tempo di Elaborazione
910625 msec.
lo spider è terminato dopo
274813 msec.
il preprocessing dopo
303563 msec.
Cliccare sul nome del link per accedere alla pagina.
Se il link è troppo lungo viene visualizzata solo la parte iniziale.
Dopo ogni link viene indicato la posizione assunta in ciascun motore selezionato.
Link
Voto
http://seds.lpl.arizona.edu/nineplanets/nineplanets/sol.html
283.7343
Posizioni: Altavista 20
http://en.wikipedia.org/wiki/Sun
159.2366
Posizioni: Altavista 25
http://www.solarviews.com/eng/sun.htm
52.0735
Posizioni: Altavista 8
http://www.enchantedlearning.com/subjects/astronomy/sun
28.6154
Posizioni: Altavista 38
http://antwrp.gsfc.nasa.gov/apod/sun.html
4.7686
Posizioni: Altavista 44
http://www.suntimes.com/
1.4398
Posizioni: Altavista 42
http://www.lasvegassun.com/
0.3168
Posizioni: Altavista 37
http://www.thesunlink.com/
0.1712
Posizioni: Altavista 40
http://java.sun.com/
0.0832
Operazione completata
Internet

CAPITOLO 8 DISCUSSIONE E CONCLUSIONI

In questo lavoro è stato presentato un modello per l'information retrieval e representation basato sull'utilizzo di spazi semantici per la definizione delle necessità informative degli utenti. Il modello può dirsi completo perché, oltre alla definizione delle query utente, comprende sia la rappresentazione dei documenti sia un framework logico per il suo utilizzo.

Viene presentata una metrica innovativa per la misura di similarità tra i documenti presenti in una collezione e una query utente.

Il modello e la metrica sono stati implementati in un meta-motore di ricerca sul Web. In questo sistema, l'orizzonte di conoscenza è dato da una rete semantica estratta dinamicamente da una knowledge-bese mediante un nuovo algoritmo.

I risultati ottenuti sono stati confrontati con altri sistemi similari noti in letteratura, dei quali sono state re-implemetate le metriche.

Il testing è stato condotto costruendo un test set ricavato dal servizio di directory del motore di ricerca Yahoo per avere un riferimento il più possibile certo sull'appartenenza di una pagina Web ad un determinato dominio.

I risultati ottenuti mostrano una aumento considerevole delle prestazioni rispetto a parametri standard come la precision e la recall.

Tali risultati sono dovuti, da un lato, al modello di riferimento utilizzato che sfrutta in maniera estesa le proprietà linguistiche e l'organizzazione cognitiva del linguaggio; dall'altro, ad una metrica più "sensibile" che riesce a misurare in maniera migliore la semantic relatedness.

APPENDICE A: SORGENTI INFORMATIVE

A.1 THE LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH

Come abbiamo visto nei capitoli precedenti, le metriche descritte per il calcolo della semantic relatedness, sono state suddivise in base alla sorgente informativa d'utilizzo. Una di queste è il dizionario, e per molti autori è stato utilizzato un particolare dizionario per la lingua inglese: *The Longman Dictionary of Contemporary English.* Quest'ultimo è ormai da anni che risulta essere un'ottima guida per l'inglese scritto e parlato e, oltre al diffusissimo utilizzo avuto nelle varie scuole del mondo è stato oggetto di utilizzo per molti ricercatori che hanno inventato e realizzato metriche per il calcolo della semantic similarity, utilizzando come sorgente informativa il dizionario. L'ultima versione non solo è stata aggiornata con figure a colori, ma contiene anche molti altri esempi per mantenere aggiornati gli utenti sui cambiamenti della lingua.

A.1.1 Caratteristiche dell'LDOCE

Vediamo in breve le caratteristiche più importanti:

- 155.000 esempi mostrano l'impiego dell'inglese nella lingua reale. E' a tutti ben noto il fatto che c'è moltissima differenza tra l'inglese scritto e quello parlato. A volte, soprattutto gli utenti meno esperti si dilettano a tradurre semplicemente effettuando una traduzione letterale. Questo approccio, certamente sbagliato, può essere ovviato se il vocabolario d'utilizzo mostra degli esempi d'uso dei vocaboli;
- 1.000.000 di frasi in inglese che illustrano vocaboli contestualizzati;
- i 3.000 vocaboli più utilizzati nell'inglese scritto e parlato sono evidenziati in rosso. L'opportunità di avere evidenziati in rosso una molteplicità di vocaboli più in uso, rende le traduzioni meno problematiche,; frasi-specchietto che visualizzano le collocazioni dei vocaboli;
- definizioni chiare e facili da capire;
- tutti i lemmi vengono spiegati utilizzando i 2.000 vocaboli di uso più frequente.

Le caratteristiche inoltre includono:

- chiarificazioni dell'uso di una parola in parlato confrontato all'inglese scritto;
- grafici per evidenziare la differenza nella frequenza dell'uso fra le parole nei discorsi e in scrittura, fra i sinonimi ecc.;
- migliaia di punti culminanti delle frasi e delle collocazioni nel senso possibile più vicino;
- chiave di pronuncia;
- riferimenti ad altre parole, frasi, immagini e note di uso;
- le parole che sono usate spesso insieme sono indicate in grassetto e sono seguite da un esempio o una spiegazione;
- le informazioni grammaticali sono indicate tra parentesi, o in grassetto prima di un esempio.
- esposizioni la differenza fra l'inglese britannico ed americano: differenze di ortografia, di pronuncia, significato delle parole etc etc. l'appendice contiene delle tabelle (numeri, i pesi e misure, verbi irregolari, nomi geografici etc…).

A.2 IL THESAURUS DI ROGET

Nel corso di questo secolo, l'opera di Roget è diventata indispensabile per i produttori di dizionari. La disposizione delle parole e la struttura dell'intero Thesaurus , è diventata così comune che un cambiamento radicale diminuirebbe il valore del libro. Per questo motivo non è stato fatto nessun tentativo di modificare lo schema principale che Roget originalmente ha stabilito, tranne alcune variazioni poco rilevanti lasciando però inaltrata la disposizione. È stato invece arricchito di nuove terminologie, anche in base all'evoluzione della lingua e agli sviluppi scientifici, politici e culturali. Di seguito è riportata una spiegazione di come è interpretato il significato e la sinonimia di una parola.

A.2.1 Significato

La funzione più evidente del significato è denotation, cioè *la cosa significata*, il concetto o l'oggetto citato. Volendo fare subito un esempio, il "denotation" della parola sedia specifica che è un oggetto d'arredamento, che ha un pianale, i piedi, uno schienale, spesso dei bracci e che una persona può sedersi su esso. Sulla base di queste caratteristiche, una sedia può essere identificata come tale,

indipendentemente dal fatto che può essere grande o piccola, fatta di legno, o di ferro; inoltre, la sedia è distinta da tutte le altre parti di una mobilia su cui ci si può sedere. Dunque il "denotation" di una parola include quelle caratteristiche che sono "criterial" per essa, ossia sono fondamentali per distinguere il vocabolo da altri.

Le parole appartenenti ad uno stesso denotation, possono essere organizzati in modo da distinguere i termini più o meno convenzionali, ossia i termini più usati rispetto ad altri, a parità di significato. Oltre che il relativo denotation, una parola può avere una connotazione, ossia le implicazioni indicative o associative di un'espressione, oltre il relativo senso letterale. Due o più parole possono avere lo stesso "denotation" e connotazione, ma differire nella loro gamma di applicabilità, cioè non possono essere usate scambievolmente nello stesso contesto. I termini che hanno una gamma limitata di applicabilità sono identificati con delle etichette legate al settore di utilizzo. Ad esempio ci sono parole il cui utilizzo è limitato solo nel settore della architettura o della musica, e per questo ci saranno delle etichette indicanti tale limitazione.

A.2.2 Sinonimia e sinonimi

Data la complessità dei significati di una generica parola, la definizone del giusto sinonimo, è stata un'operazione non semplice per chi ha lavorato per costruire questo thesaurus. A causa della dettagliata spiegazione che viene data al significato di un vocabolo, c'è sicuramente la possibilità di offrire all'utente una vasta scelta dei sinonimi che si trovano all'interno della gamma "denotative" della parola stessa.

Un sinonimo è una parola con un significato identico o molto simile a quello di un'altra parola. Si dice spesso che, in effetti, non c'è qualcosa come un sinonimo assoluto per una parola, cioè una forma che sia identica in ogni funzione del significato in modo da potere essere applicata scambievolmente tra le due. Secondo questa visione estrema, gli unici sinonimi sono i termini che hanno precisamente la stessa "denotation", connotazione e gamma di applicabilità.

Questa idea della sinonimia è troppo restrittiva, tuttavia nel Roget's thesaurus, i termini sinonimi sono quelli che hanno denotations quasi identici. Il motivo per la scelta dell'uno o dell'altro sinonimo, spesso è un discorso stilistico: si può preferire una parola più semplice o più complessa o si può preferire un termine più convenzionale o meno. Ma il fatto che queste parole condividano un

denotation li rende sinonimi e disponibili per sostituire parole. Il successo rivoluzionario della prima edizione del thesaurus nel 1852 era lo sviluppo di un principio "brand-new": la disposizione delle parole e delle frasi secondo i loro significati. Questo nuovo sistema riunisce in un punto tutti i termini connessi da un singolo pensiero o concetto; permette una vasta indagine della lingua all'interno di un libro dal formato relativamente modesto, senza le ripetizioni spesso "space-consuming" che limitano tantissimo la portata dei thesauri organizzati come dizionari. Questa organizzazione brillante rende il Thesaurus di Roget molto efficiente. Questo strumento viene aggiornato periodicamente per tener conto dei cambiamenti della lingua.

A.3 IL DATA BASE LESSICALE WORDNET

WordNet è un database lessicale basato sulle teorie psicolinguistiche della memoria lessicale umana.

Realizzato manualmente da una équipe di psicologi e linguisti guidata dal prof. George A. Miller presso il laboratorio di Scienze Cognitive all'Università di Princeton, WordNet è disponibile gratuitamente al sito della stessa università, inoltre la licenza d' uso ne permette l'utilizzo gratuito anche a fini commerciali ed al di fuori della ricerca, purché siano citati gli autori ed il sito ufficiale del progetto. Poiché numerosi esperimenti svolti in psicolinguistica hanno dimostrato che molte proprietà del lessico mentale possono essere sfruttate nell'ambito della lessicografia, il gruppo di Princeton ha deciso di combinare le due discipline per costruire un lessico che fosse anche un modello della memoria umana.

I dizionari tradizionali sono organizzati mediante un ordinamento alfabetico per rendere più semplice e rapida l'individuazione da parte del lettore del termine cercato. Mediante questo approccio però, vengono accostati termini che rappresentano concetti totalmente diversi, inoltre per ognuno di essi vengono raccolti tutti i significati insieme, sebbene non abbiano niente in comune. Infine un dizionario tradizionale è ridondante per quanto riguarda i sinonimi, poiché lo stesso concetto viene ripetuto più volte.

Come vedremo nel dettaglio in seguito, WordNet non è organizzato alfabeticamente ma su base concettuale: in esso nomi, verbi, aggettivi ed avverbi sono organizzati in quattro categorie sintattiche, ognuna delle quali è suddivisa in insiemi di sinonimi (*synset*), che a loro volta rappresentano un concetto lessicale condiviso da tutti i termini ad esso associati. Un termine

ovviamente può possedere più di un significato ed essere quindi presente in molti di questi insiemi ed anche in più di una categoria sintattica.

Altro importante fattore che contraddistingue WordNet da un semplice dizionario di vocaboli è che questi insiemi di sinonimi sono collegati tra loro e strutturati in una rete tramite una serie di relazioni che riflettono i principi in base a cui tali concetti sono organizzati nella mente, e che permettono di creare all' interno della categoria sintattica gerarchie di significato.

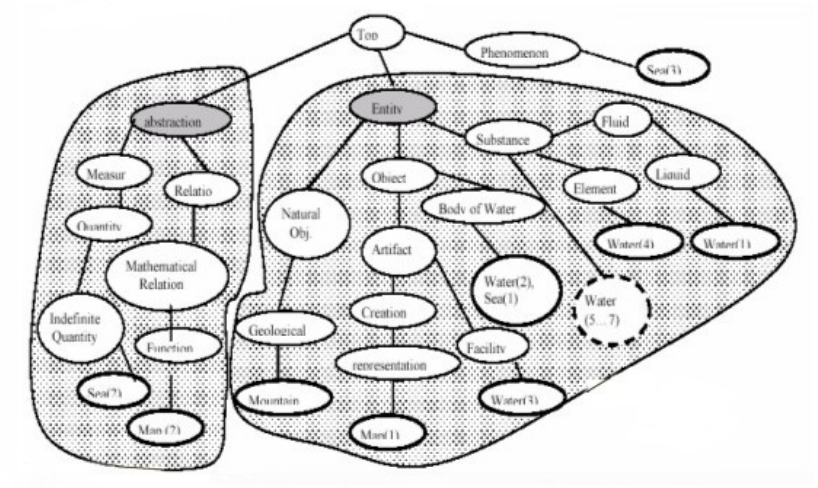


Figura A.1: Spaccato dell'organizzazione gerarchica dei nomi in WordNet

Il termine *map* ha due sensi: 1- carta geografica 2- corrispondenza (intesa dal punto di vista matematico, ad esempio ad un elemento di un insieme corrisponde uno ed un sol elemento di un altro insieme) ognuno dei quali è associato ad un synset diverso che appartiene a una distinta sottogerarchia. Dal punto di vista della ricerca psicolessicologica, WordNet rappresenta un unicum in quanto ha esteso all'intero lessico i risultati di ricerche che solitamente sono svolte su un numero ristretto di parole. È interessante osservare che WordNet si basa su un insieme ristretto di semplici ipotesi psicolinguistiche, tuttavia il fatto che tali ipotesi siano state testate sull'intero lessico della lingua inglese le rende più robuste e plausibili. Attualmente è considerato la più importante risorsa disponibile per i ricercatori nei campi della linguistica computazionale, dell'analisi testuale, e di altre aree associate.

POS	Unique Strings	Synsets Word-Sense	Total Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	3101
Adverb	4669	3664	5808
Totals	152059	115424	203145

Tabella A.1: Numero di lemmi, di synset e totale coppie lemmi-synsets per categoria sintattica in WordNet 2.0

A.3.1 Terminologia di WordNet

In questo paragrafo si introducono un insieme di termini significativi propri della terminologia di WordNet.

- Categoria sintattica: sono le grandi categorie in cui sono suddivisi i termini (ed anche i file in cui sono contenuti) di WordNet. Le categorie sintattiche trattate sono quattro: nomi, verbi, aggettivi ed avverbi.
- Lemma: è la parola, il termine a cui viene associato uno o più significati. A volte un lemma è costituito da due o più parole ed in tal caso i singoli termini sono uniti dal carattere *underscore* (_).
- Synset: un synset rappresenta il significato che viene associato ad un insieme di lemmi appartenenti alla stessa categoria sintattica. In pratica è corretta l'affermazione che ad un synset appartengono un certo numero di lemmi. Un synset, infatti, può essere rappresentato, oltre che tramite il suo *Gloss*, per mezzo dell'insieme dei sui lemmi (di solito racchiusi all'interno di parentesi graffe (e.g. S_j $\{l_1, l_2, l_3...\}$, dove a l_i corrisponde l'i-simo lemma collegato al synset j-simo).
- Gloss: un Gloss è una descrizione a parole di uno specifico significato; ogni synset oltre a contenere un insieme di sinonimi possiede anche un gloss.
- Relazione semantica: si tratta di una relazione presente fra due synset appartenenti alla stessa categoria sintattica.
- Relazione lessicale: è una relazione tra due lemmi appartenenti a due synsets distinti, sempre relativamente alla stessa categoria sintattica.

I diversi tipi di relazioni semantiche e lessicali saranno trattate in seguito.

A.3.2 La Matrice Lessicale

WordNet distingue in modo netto i significati delle parole (concetti lessicali), dalle forme ad essi associati (per forma si intende il modo in cui viene letta e scritta una parola), tramite le quali tali concetti sono espressi.

Per ridurre l'ambiguità derivata dal termine parola, useremo:

- "word meaning" o significato, se ci riferiamo al concetto lessicale o significato;
- *"word form" o lemma*, se ci riferiamo al mondo in cui viene letta o scritta.

La corrispondenza tra *word form* e *word meaning* è, soprattutto per i lemmi più frequenti, in un rapporto molti a molti e da luogo alle seguenti proprietà:

- Sinonimia: proprietà di un significato di avere due o più forme di parole in grado di esprimerlo;
- Polisemia: proprietà di una *forma di parola* di esprimere due o più significati.

Tale corrispondenza è rappresentata tramite una matrice bidimensionale, detta *matrice lessicale* esemplificata nella tabella A.2, dove nelle righe sono rappresentate le *word meaning* (significati) e nelle colonne le *word form* (forma/lemma base).

In pratica, volendo leggere la matrice lessicale tramite la terminologia di WordNet, ad ogni riga è associato un synset e ad ogni colonna un lemma.

Un elemento $E_{i,j}=(M_i, F_j)$ della matrice rappresenta una definizione: il lemma F_j è usato per esprimere il significato dato da M_i.

Word Meaning	**Word Form**			
	F_1	F_2	…	F_n
M_1		$E_{1,2}$	…	
M_2	$\{E_{2,1}$	$E_{2,2}$	…	$E_{2,n}\}$
…	…	…	…	…
M_m		$E_{m,2}$	…	

Tabella A.2: Matrice lessicale

Se ci sono due elementi nella stessa riga significa che le due word form sono sinonimi, mentre se ci sono due elementi nella stessa colonna significa che la word form è polisemica cioè ha più significati.

In WordNet un significato è dunque rappresentato da un insieme di sinonimi. Questa modalità di rappresentazione è motivata dal fatto che WordNet è stato originariamente concepito come uno strumento di consultazione destinato a parlanti inglesi i quali possiedono già il concetto e le parole per identificarlo. In questa ottica qualsiasi descrizione che consenta di distinguere un concetto da un altro, come appunto un insieme di sinonimi, può essere adeguato come rappresentazione del concetto stesso [1].

[1] La teoria che sta alla base di WordNet può essere definita come differenziale in quanto non ha come scopo quello di definire i singoli concetti ma solo di identificarli e differenziarli rispetto agli altri. Al contrario, una teoria costruttiva richiede necessariamente che la rappresentazione del concetto contenga informazioni sufficienti per permettere un'accurata costruzione dello stesso.

Word Meaning	Word Form		
	F_1	F_2	F_3
an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner	music		
any agreeable (pleasing and harmonious) sounds; "he fell asleep to the music of the wind chimes	music,	euphony	
musical activity (singing or whistling etc.); "his music was his central interest".	music		
(music) the sounds produced by singers or musical instruments (or reproductions of such sounds)	music		
punishment for one's actions; "you have to face the music"; "take your medicine"	music		medicine

Tabella A.3: Esempio di matrice lessicale semplificata relativa

I synset possono quindi essere considerati come designatori non ambigui dei significati delle parole: la loro funzione non è di definire i concetti ma semplicemente di indicare che esistono.Tuttavia nel corso del lavoro di sviluppo di WordNet è emerso che in alcuni casi i synset non sono sufficientemente informativi in quanto alcuni concetti sono espressi da una sola parola, la quale non può così essere disambiguata nemmeno da parte di persone che abbiano già acquisito tutti i suoi significati.Per questo motivo, e per rispondere alle richieste di maggior informazione provenienti dall'ambito della linguistica computazionale, a partire dalla versione di WordNet 1.6, ad ogni synset è stata aggiunta un gloss, cioè una breve definizione del concetto accompagnata da un esempio, che non svolge tuttavia nessun ruolo dal punto di vista psicolinguistico. In figura A.2, è riportato un esempio di come WordNet mostra tali informazioni.

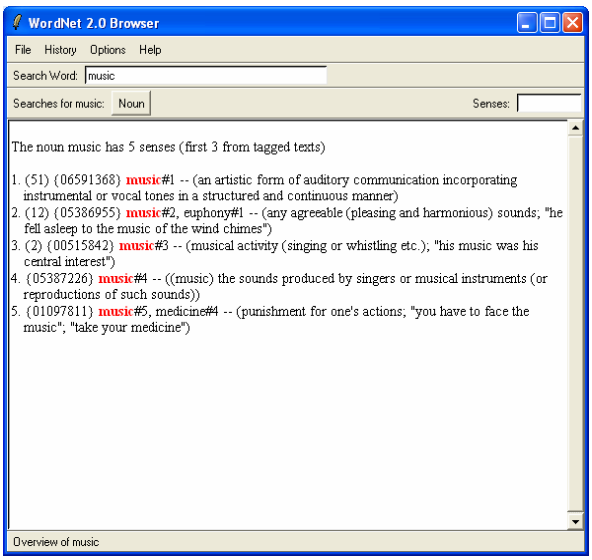


Figura A.2: Esempio di informazioni usanto la word form "music"

Ovviamente i requisiti richiesti da una teoria differenziale sono molto più modesti rispetto a quelli di una teoria costruttiva.

Dunque i modi per rappresentare un significato in WordNet sono:

- Gloss: un commento/spiegazione del significato. E' molto aleatorio oltre che scomodo per la sua lunghezza.
- Elenco dei nomi e dei sinonimi: insieme di tutti i lemmi che singolarmente possono essere utilizzati per rappresentare il significato.
- Chiave surrogata: un identificatore univoco, ma scorrelato dai contenuti del record. In WordNet è usato il synset_offset che indica la posizione, all'interno del file di dati, del synset.
- SenseKey: composta da lemma+numero_senso è una chiave che collega le due entità lemmi e significati.

A.3.3 Relazioni lessicali e semantiche

Il fatto che in WordNet un concetto sia rappresentato da synset rende necessario mantenere separati nomi, verbi, aggettivi a avverbi in quanto parole di categorie sintattiche diverse non possono essere sinonimi. Ciò è coerente con le prove psicolinguistiche che dimostrano che ci sono differenze fondamentali nell'organizzazione semantica di queste categorie.

Nomi, verbi, aggettivi e avverbi devono dunque essere trattati separatamente in quanto esprimono concetti diversi, sono organizzati indipendentemente e strutturati in modo diverso nella memoria lessicale: i nomi sono organizzati come gerarchie ad ereditarietà, gli aggettivi sono organizzati in base all'opposizione semantica e i verbi sono organizzati da una varietà di relazioni di implicazione. Queste strutture diverse riflettono modi diversi di categorizzare l'esperienza e il tentativo di imporre un singolo principio organizzativo su tutte le categorie significherebbe rappresentare erroneamente la complessità psicologica della conoscenza lessicale.

Inoltre in WordNet sono tenuti distinti due tipi di relazioni: semantiche e lessicali, che si differenziano a seconda del tipo di operatore cui sono applicate.

Quindi, le relazioni semantiche valgono tra i concetti ed hanno perciò come operandi i synset, mentre le relazioni lessicali valgono tra le parole contenute nei synset stessi, ossia hanno come operandi i lemmi. Non possono esistere relazioni fra un lemma ed un synset o fra operandi appartenenti a diverse categorie sintattiche, ad esempio fra un nome ed un verbo.

Nei prossimi paragrafi saranno presentate tutte le tipologie di relazioni lessicali e semantiche presenti in WordNet con relativa breve spiegazione.

A.3.4 Relazioni semantiche

Le relazioni di tipo semantico coinvolgono sempre due synsets, non semplicemente due lemmi, e sono per la maggior parte di tipo simmetrico: se esiste una relazione R tra il synset {x1, x2,….} e il synset {y1, y2 ,….}, allora esiste anche una relazione R' tra {y1, y2, …. } e {x1, x2,….}.

A.3.4.1 Iponimia (hyponymy)/ Iperonimia (hyperonymy) e Troponimia (troponymy)

La relazione di iponimia, lega un concetto (nel nostro caso un synset) ad uno più generale, quello che può essere ritenuto una sua generalizzazione, per cui è lecito affermare che:

Un concetto rappresentato dal synset {x, x' …} è un iponimo di un concetto rappresentato dal synset {y, y', …} se un madrelingua inglese valida una sentenza costruita come: x is a (kind of) y.

Per quanto riguarda la relazione opposta, quella di iperonimia, lega un concetto ad uno più particolare, più specializzato. In pratica si può affermare che:

Un concetto rappresentato dal synset {x, x' …} è un iperonimo di un concetto rappresentato dal synset {y, y', …}, se {y, y', …} presenta tute le caratteristiche di {x, x' …} più, almeno, una sua caratteristica particolare e aggiuntiva.

La relazioni di iponimia e iperonimia, chiamate anche relazioni ISA, sono transitive e simmetriche, e generano una struttura gerarchica semantica simile alla gerarchia di specializzazione/generalizzazione presenti nei modelli per database relazionali o dell'ereditarietà per i modelli ad oggetti. La radice di tale gerarchia è occupata dai concetti più generali, mentre a livello delle foglie ci sono i concetti specializzati. Come nei modelli relazionali e nei modelli ad oggetti, ogni iponimo eredita tutte le caratteristiche dei concetti più generici e addiziona almeno uno per distinguerlo dal sovraordinato e dagli altri iponimi del sovraordinato.La relazione di iponimia/iperonimia stabilisce la regola centrale per l'organizzazione dei nomi in WordNet, essa è definita solo per i nomi e i verbi, ed è la relazione più frequente in WordNet.

Nel caso dei verbi la relazione di iponimia è chiamata toponimia (troponymy).

Un verbo che esprime una maniera specifica di operare di un altro verbo: X è un troponimo di Y se fare X è fare Y in qualche maniera.

A.3.4.2 Meronimia (meronymy) / Olonimia (holonomy)

La meronimia è una relazione semantica e corrisponde alla relazione "has a". Dal punto di vista lessicale è definita come:

Un concetto rappresentato dal synset {x,x',…} è un meronimo di un concetto rappresentato dal synset {y,y',…} se un madrelingua inglese convalida la frase costruita come: An y has an x (as a part) or An x is a part of y (x come parte/sostanza/membro di y).

Sono definiti tre tipi di aggregazione:

HAS MEMBER – esempio *school* inteso come istituzione educativa ha due aggregazioni:

- school HAS MEMBER: staff, faculty
- school HAS MEMBER: schoolteacher, school teacher

HAS PART – esempio *school* inteso come luogo dove si riceve l'educazione ha un asola aggregazione:

- school HAS PART: classroom, schoolroom.

HAS SUBSTANCE – esempio *quartz* inteso come minerale ha due aggregazioni:

- quartz HAS SUBSTANCE: silicon, Si, atomic number
- quartz HAS SUBSTANCE: silica, silicon oxide, silicon dioxide.

La meronimia è una relazione transitiva e asimmetrica, la sua relazione duale è l'olonimia; anch'essa può essere usata per costruire gerarchie di concetti meronimi/olonimi sulla categoria sintattica dei nomi, con la differenza che, in questo caso, uno stesso meronimo può avere più olonimi: in altri termini, un concetto può contemporaneamente far parte di differenti concetti composti. I meronimi sono fattori distintivi che gli iponimi possono ereditare. Per esempio, se beak (becco) e wing (ala) sono meronimi di bird (uccello), e canard (canarino) è un iponimo di bird, allora, ereditando, beak e wing deve anche essere meronimo di canary. Quando tutte e tre le relazioni, iponimia, meronimia, e antinomia si incrociano, il risultato è altamente interconnesso in una rete complessa; sapere dove una parola è situata in questa rete è un'importante informazione per la conoscenza del significato della parola stessa.

A.3.4.3 Implicazione (entailment)

Questa relazione può essere utilizzata solo per i verbi.

Un verbo X implica (entail) Y, se X non può essere fatto senza che Y sia, o sia stato, fatto.

La relazione di implicazione è un tipo di relazione unidirezionale, infatti se *x* implica *y,* non è vero il contrario, tranne nei casi in cui i due verbi siano sinonimi. Questo tipo di relazione è simile a quella di meronimia nei nomi. Per esempio, l'azione di comprare un bene si compone prima di una fase di decisionale e di pagamento, così come l'azione di paracadutarsi è composta da lanciarsi, planare e andare giù.

Inoltre si può osservare dalla definizione di implicazione, che la troponimia è una particolare relazione di implicazione, in quanto se X è troponimo di Y, allora X implica (entails) anche Y. (Esempio: se "zoppicare" è troponimo di "camminare" allora non posso zoppicare se non cammino,ovvero zoppicare implica camminare).

A.3.4.4 Relazione causale (cause to)

La relazione causale è simile alla relazione di entailment però senza l'inclusione temporale; la relazione causale è una specie di implicazione:

se V_1 necessariamente causa V_2, allora V_1 implica anche V_2, dove il verbo implicante V_1 denota la causa dello stato o l'attività riferita dal verbo implicato V_2.

Come la relazione di implicazione, anche la relazione causale è unidirezionale: "dar da mangiare causa che una persona mangia, il fatto che persona sta mangiando non implica che qualcuno gli stia dando da mangiare."

A.3.4.5 Raggrupamento di verbi (verb group)

Questa relazione viene utilizzata per produrre raggruppamenti nella categoria sintattica dei verbi. In un gruppo formato in tale maniera i synsets hanno tutti un significato semantico molto simile.

A.3.4.6 Similarità (similar to)

La relazione di similarità è tipica della categoria degli aggettivi, ed è simile alla sinonimia. WordNet divide gli aggettivi in due classi principali: descrittivi e relazionali. Un aggettivo descrittivo è un aggettivo che attribuisce un valore di un attributo ad un nome.

Ad esempio, dire *The package is heavy* presuppone che ci sia un attributo *weight* tale che weight(package) = heavy.

I synset riguardanti gli aggettivi descrittivi sono organizzati in cluster di aggettivi. Al centro di questi cluster c'è un aggettivo a cui gli altri componenti sono legati da similarità.

Gli aggettivi che si trovano al centro dei cluster sono legati mediante la relazione di antonimia ad altri cluster. Si distingue tra antonimi diretti come heavy/light e antonimi indiretti come heavy/weightless. In questo modo gli aggettivi che non hanno antonimi diretti hanno però antonimi indiretti (questi ultimi vengono ereditati attraverso la relazione di similarità).

A.3.4.7 Attributo (attribute)

Questa relazione può essere applicata alle categorie dei nomi e degli aggettivi. Attributo è il nome per cui uno o più aggettivi esprimono un valore. Il nome *weight* è un attributo a cui gli aggettivi light e heavy danno un valore.

A.3.4.8 Coordinazione

Non è un tipo di relazione base, ma derivata. Due synset sono coordinati se condividono uno stesso iperonimo, cioè se sono la specializzazione dello stesso concetto.

A.3.5 Relazioni lessicali

Le relazioni lessicali diversamente da quelle semantiche, coinvolgono sempre due lemmi, no due synsets.

A.3.5.1 Sinonimia (synonymy)

La relazione lessicale più comune e pervasiva è la sinonimia tra forme di parola, che si ritrova all'interno di tutte le categorie lessicali e che serve come criterio costitutivo dei synset. Non a caso tale relazione è alla base della costruzione della matrice lessicale.

Le altre relazioni semantiche e lessicali tendono invece ad essere specifiche delle diverse categorie lessicali.

Pertanto, possiamo affermare che la sinonimia è la relazione più importante del WordNet, infatti, un buon conoscitore della lingua inglese, dalla sola relazione di sinonimia tra i vari lemmi, può risalire al significato..

Consideriamo, ad esempio la parola *board,* essa può assumere i significati di "tavola di legno" e "scheda per computer", se osserviamo i due *synset* che la

contengono*:* {*board,plank*} e {*board, circuit card*}, possiamo risalire all'esatto significato; quindi è bastato introdurre due sinonimi, *plank* e *circuit card*, per eliminare l'ambiguità.

Per distinguere la relazioni di sinonimia dalle altre relazioni, i termini sinonimi sono racchiusi fra parentesi graffe {}, mentre gli altri insiemi prodotti dalle altre relazioni lessicali sono racchiusi fra parentesi quadre [].

La definizione di sinonimia, generalmente attribuita a Leibinitz, è:

"Due espressioni sono sinonime se la sostituzione di una per l'altra non cambia il vero significato della frase nella quale è fatta la sostituzione."

Questa definizione impone una condizione estrema e rende i sinonimi molto rari o inesistenti. Una seconda definizione invece, adottata da WordNet, fa cadere l'ipotesi sul contesto ed è:

"Due espressioni sono sinonime in un contesto linguistico C se la sostituzione di una per l'altra, nel contesto C, non altera il vero valore della frase."

Per meglio comprendere, prendiamo come esempio *board,* che in un contesto di falegnameria, può quasi sempre essere sostituito con *plank* senza alterare il significato, mentre in altri contesti questa sostituzione risulterebbe inappropriata.

La definizione di sinonimia in termini di sostituibilità divide necessariamente il WordNet in nomi, verbi, aggettivi e avverbi, infatti, parole che appartengono a diverse categorie sintattiche non possono essere interscambiate.

A.3.5.2 Antinomia (antynomy)

L'antinomia è una relazione lessicale tra lemmi, la quale indica che uno è il contrario dell'altro.

Solitamente l'antinomia di una parola *"x"* è "non *x"*, ma ciò non è sempre vero, infatti, *ricco* e *povero* sono antinomi, ma dire *non ricco,* non equivale a povero; molta gente può essere né ricca e né povera.

Questo tipo di relazione può essere usato per coppie di termini appartenenti a tutte le categorie sintattiche in cui è suddiviso il WordNet. Pur apparendo come una semplice relazione di simmetria, l'antinomia è una relazione abbastanza complessa.

A.3.5.3 Relazione di pertinenza (pertainym)

La relazione di pertinenza può essere applicata alla categoria degli aggettivi (la relazione inversa "derived from adjective" viene applicata agli avverbi).

Gli aggettivi che rimangono fuori dall'organizzazione per cluster contrapposti sono gli aggettivi relazionali. Tali aggettivi sono definiti da frasi del tipo "di o pertinenti a" e non possiedono antonimi. Un aggettivo di pertinenza può essere in relazione con un nome o con un altro aggettivo di questo tipo

A.3.5.4 Vedi anche (see also)

E' una relazione lessicale che lega singoli lemmi di diversi synset, ed i motivi di tale relazione possono essere molto differenti fra loro.

A.3.5.5 Relazione participiale (participle)

Anche questa relazione è tipica della categoria degli aggettivi e riguarda i cosiddetti aggettivi participiali, cioè che derivano da un verbo. Ad esempio l'aggettivo burned deriva dal verbo burn.

A.3.5.6 Derivato da (derived from)

Alcuni aggettivi relazionali derivano da antichi nomi Greci o Latini. Questa affermazione risulta essere vera sia per la lingua italiana che per quella inglese (idioma su cui è costruito WordNet). L'aggettivo relazionale verbale, deriva dal nome neutro latino verbum, mentre lessicale deriva dal corrispondente nome greco. La relazione "derivato da" lega gli aggettivi ad i nomi stranieri da cui derivano.

A.3.5.7 Relazioni morfologiche

Un'altra importante classe di relazioni lessicali sono le relazioni morfologiche tra word form. Inizialmente questo tipo di relazioni non era stato considerato ma con l'avanzamento del progetto, cresceva la necessità di inserire questo nuovo tipo di relazione; infatti, bastava inserire un termine al plurale, ad esempio trees, per far si che WordNet dava come risultato che il termine non era incluso nel database. La morfologia delle parole nella lingua inglese è abbastanza semplice e si manifesta come declinazioni per i sostantivi, ad esempio distinzione fra singolare e plurale, e coniugazione dei verbi. Per far si che WordNet riconoscesse queste forme è stato inserito un software che non modifica la struttura del database, nel senso che all'interno le parole non sono replicate, infatti, le parole sono memorizzate nella forma canonica. Il software introdotto ha il compito di interfacciarsi fra l'utente e il database lessicale, in modo tale da tradurre ogni

termine in input nella forma canonica e successivamente inviarlo al database. Nonostante la semplicità della morfologia inglese rispetto ad altre lingue, la realizzazione di questo componente non lo è stata a causa della massiccia presenza di verbi irregolari.

A.3.6 Simboli dei puntatori utilizzati in WordNet

Per rappresentare le relazioni tra parole appartenenti a synset diversi si usano i puntatori. I seguenti tipi di puntatore sono usati per indicare relazioni di tipo lessicale: "antonym", "pertainym", "participle", "also see"; i restanti tipi di puntatore sono generalmente usati per rappresentare relazioni di tipo semantico.

Sebbene ci siano molti tipi di puntatori, solo alcuni tipi di relazione sono permessi all'interno di ogni categoria. In tabella A.4, sono indicati i simboli dei puntatori usati da WordNet divisi per categoria sintattica; in tabella A.5, sono riportate le relazioni di tipo simmetrico e le corrispondenti relazioni "inverse".

Nomi	Verbi	Aggettivi	Avverbi
! Antonym	**!** Antonym	**!** Antonym	**!** Antonym
@ Hypernym	@ Hypernym	**&** Similar to	**** Derived from adjective
~ Hyponym	**~** Hyponym	**<** Participle of verb	**;c** Domain of synset - CATEGORY
#m Member holonym	***** Entailment	**** Pertainym (pertains to noun)	**;r** Domain of synset - REGION
#s Substance holonym	**>** Cause	**=** Attribute	**;u** Domain of synset - USAGE
#p Part holonym	^ Also see	^ Also see	
%m Member meronym	**$** Verb Group	**;c** Domain of synset - CATEGORY	
%s Substance meronym	**+** Derivationally related form	**;r** Domain of synset - REGION	
%p Part meronym	**;c** Domain of synset - CATEGORY	**;u** Domain of synset - USAGE	
= Attribute	**;r** Domain of synset - REGION		
+ Derivationally related form	**;u** Domain of synset - USAGE		
;c Domain of synset - CATEGORY			
-c Member of this domain - CATEGORY			
;r Domain of synset - REGION			
-r Member of this domain - REGION			
;u Domain of synset - USAGE			
-u Member of this domain - USAGE			

Tabella A.4: Simboli dei puntatori divisi per categoria sintattica

Pointer	Reflect
Antonym	Antonym
Hypernym	Hyponym
Hyponym	Hypernym
Holonym	Meronym
Meronym	Holonym
Similar to	Similar to
Attribute	Attribute
Verb group	Verb group

Tabella A.5: Relazioni di tipo semantico e loro inverse

All'interno del database di Wordnet, una entry per una parola contiene i puntatori alle entry di altre parole.

A.3.7 Organizzazione delle categorie sintattiche

Come è stato detto in precedenza, WordNet divide le parole in categorie sintattiche. Questo tipo di suddivisione genera un pò di ridondanza in quanto alcuni termini sono inclusi in più categorie, per esempio la parola *cream* è inclusa nella categoria dei sostantivi, in quella dei verbi e in quella degli aggettivi. Si ha però un grande vantaggio: quello di rendere evidenti le differenze di organizzazione semantica. WordNet organizza in modo diverso le categorie:

- i nomi sono organizzati mediante gerarchie di generalizzazione/specializzazione basate sulla relazione semantica di iperonimia/iponimia;
- i verbi sono organizzati tramite varie relazioni di implicazione;
- gli aggettivi e avverbi sono organizzati in iperspazi n-dimensionali.

A.3.7.1 Organizzazione dei nomi

Wordnet nella versione 2.0 contiene più di 110.000 nomi organizzati in circa 75.000 synsets; tali valori sono indicativi in quanto WordNet continua a crescere. In termini di copertura, WordNet raggiunge quella di un buon dizionario tascabile. Ogni vocabolario contiene dei "circoli viziosi", ossia può capitare che per definire la parola X si usi Y e viceversa. I lessicografi che hanno progettato WordNet hanno tentato di dare alla memoria semantica dei nomi una forma ad albero.

Questo albero può essere ricostruito a partire dal cammino effettuato dagli ipernimi. Questa gerarchia, che è limitata in profondità (può comunque raggiungere 16 livelli di profondità), può essere percorsa sia partendo dai concetti più generali andando verso quelli più specializzati sia nella direzione

opposta, poiché in WordNet viene codificata sia la relazione di iperonimia sia la sua relazione inversa, l'iponimia. Questo tipo di struttura ad albero è molto utile perchè le informazioni comuni a più termini vengono memorizzate una volta sola; tale struttura possiede tutte le proprietà di ereditarietà tipiche delle gerarchie di specializzazione.

Oltre ai puntatori per le relazioni di iperonimia e iponimia, ogni synset relativo ad un nome contiene anche i puntatori per gli altri tipi di relazione validi per tale categoria.

La definizione di un significato di un nome viene data da un iperonimo e da altre caratteristiche (es: un meronimo, un omonimo, oppure un antonimo) che servono a distinguerlo da altri termini e/o da altri significati.

Si può dire che la struttura dei nomi di WordNet segue i principi che governano la memoria lessicale umana, infatti quando ci si chiede di dare la definizione di un nome, generalmente si tende a dare o un sinonimo della parola stessa o un concetto più generale, in alcuni casi si tende a dare anche altre caratteristiche che lo possano qualificare; se si tratta di un oggetto tangibile si tende ad indicare il materiale di cui è composto o le parti che lo compongono.

Da quanto è stato detto finora, sembrerebbe che la gerarchia dei nomi abbia un'unica radice; in realtà se così fosse il concetto contenuto nella radice sarebbe così generico da fornire poche informazioni a livello semantico.

L'alternativa, adottata da WordNet, è quella di ripartire i nomi in modo da selezionare un numero limitato di concetti generici che verranno trattati ognuno come il capostipite di una gerarchia a sé. La lista dei concetti capostipiti selezionati da WordNet è riportata in tabella A.6.

Capostipiti per i nomi
{entità, physical thing}
{psycological feature}
{abstraction}
{state}
{event}
{act, human action, human activity}
{group, grouping}
{possesion}
{phenomenon}

Tabella A.6: Lista dei capostipiti per i nomi

A.3.7.2 Organizzazione dei verbi

Sebbene ogni frase di lingua inglese, per essere corretta dal punto di vista grammaticale, debba contenere almeno un verbo, i verbi inglesi sono in numero

decisamente inferiore rispetto ai nomi. In compenso, i verbi hanno una polisemia molto più alta rispetto a quella dei nomi.

Per esempio, il "*Collins English Dictionary*" contiene 43,636 nomi e solo 14,190 verbi. Da ciò si deduce che i verbi sono più polisemici dei nomi, infatti, nel *Collins* i nomi hanno una media di 1.74 significati, mentre i verbi hanno una media di 2.11, ed inoltre sono anche più flessibili poiché cambiano il loro significato in base al contesto, mentre i nomi tendono ad essere più stabili.

Un'altra caratteristica dei verbi è quella di cambiare significato in base al nome che li accompagna; nelle frasi *I have a Ferrari* e *I have a headache* il verbo *have* assume due significati molto diversi.

I verbi di WordNet sono divisi sulla base di criteri semantici in 15 gruppi, ognuno memorizzato in un file. Tutti questi gruppi, tranne uno, corrispondono a ciò che i linguisti chiamano domini semantici e ogni verbo contenuto in questi insiemi descrive eventi o azioni.

Il restante insieme non costituisce un dominio semantico e contiene verbi che descrivono stati. Nella tabella seguente, viene mostrata la lista di questi gruppi e una breve descrizione di ciò che contengono.

Nome del file	Descrizione (di cosa trattano i verbi)
verb.body	vestire e cura del corpo
verb.change	cambiamento di dimensione, temperatura, intensità
verb.cognition	pensare, giudicare, analizzare
verb.communication	parlare, domandare, cantare, ecc.
verb.competition	combattimento, attività atletiche, ecc.
verb.consumption	mangiare e bere
verb.contact	toccare, colpire, tirare, ecc.
verb.creation	cucire, cucinare, dipingere, attività creative manuali
verb.emotion	sentimenti
verb.motion	camminare, nuotare, volare, ecc.
verb.perception	vedere, ascoltare, percepire, ecc.
verb.possession	comprare, vendere, possedere, trasferire
verb.social	eventi di attività sociali e politiche
verb.stative	essere, avere, relazioni spaziali
verb.weather	pioggia, neve, meteorologia in genere

Tabella A.7: Lista dei gruppi in cui sono divisi i verbi di WordNet

Il tipo di struttura usato per i nomi non può essere usato per i verbi. Questi ultimi infatti sono organizzati tramite gerarchie più complesse basate su una serie di relazioni di entailment. Queste strutture sono meno profonde e ramificate rispetto a quelle dei nomi, ed è possibile contare 628 concetti capostipite.

Alcune attività possono essere spezzate in altre attività ordinate in modo sequenziale. Fare questo dal punto di vista lessicale significa suddividere l'azione

descritta da un verbo in più azioni che possono essere ordinate dal punto di vista temporale oppure no.

Come già spiegato precedentemente, per i verbi viene usata la relazione di troponymy al posto dell'iponimia.

Le azioni che si riferiscono ad un troponym sono sempre di tipo simultaneo (totale o parziale). Consideriamo ora l'entailment che non possiede la caratteristica dell'inclusione temporale. Molti verbi con significati opposti condividono una relazione di entailment con lo stesso verbo. La relazione causale è un tipo di entailment senza inclusione temporale, visto che i verbi connessi da questo tipo di legame lo sono dal punto di vista logico, non c'è un vincolo temporale.

A.3.7.3 Aggettivi e avverbi

Gli aggettivi sono strutturati in cluster che contengono un synset principale e dei synset satellite. Ogni cluster è organizzato intorno a coppie di antinomi (occasionalmente triplette). Le coppie di antinomi sono indicate nei synset principali dei cluster e sono dette relazioni di antinomia dirette. I synset principali sono collegati ai synset satellite da una relazione di similarità ed i synset satelliti mediano tramite il synset principale le relazioni di antinomia che diventano derivate. Un modo per pensare l'organizzazione a cluster è visualizzare una ruota, con il synset principale al centro ed i satelliti sui raggi. Due o più ruote sono connesse logicamente con la relazione di antinomia che può essere pensata come un asse tra le due ruote.

Gli aggettivi pertinenti sono aggettivi relazionali che non seguono la struttura a cluster: non hanno antinomi, di solito i synset contengono un solo lemma ed hanno una relazione con un nome. Gli aggettivi partecipiali hanno una relazione lessicale con il verbo dal quale derivano.

Gli avverbi spesso derivano da un aggettivo, e di tanto in tanto hanno un antinomo. Di solito contengono solo la relazione con l'aggettivo da cui derivano.

BIBLIOGRAFIA

[Ackerman1997] Ackerman M., et al., "Learning Probabilistic User Profiles - Applications for Finding Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunities," AI Magazine, vol. 18, issue 2, pp. 47-56, 1997

[Agirre1996] Agirre E., Rigau G., "Word sense disambiguation using conceptual density". In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 16-22, Copenhagen, Denemark, 1996

[Agirre1997] Agirre E., Rigau G. (1997), "A proposal for word sense disambiguation using conceptual distance". In Ruslan Mitkov and Nicolas Nicolov, editors, Recent Advances in Natural Language Processing: Selected Papers from RANLP'95, vol. 136, chapter 2, pp. 161-173. John Benjamins Publishing Company, Amsterdam/Phildadelphia

[Allen1970] Allen T., Communication networks in R&D laboratories, R&D Management, 1, ristampa in Beliver C. Griffith (Ed.), Key papers in information science, pp. 66-73, White Plains, Knoweldge Industry Pubblications, 1970

[Aridor2000] Aridor Y., et al., "Knowledge Agent on the Web," Proceedings of the 4th International Workshop on Cooperative Information Agents IV, pp. 15-26, 2000

[Armstrong1995] Armstrong, R., et al., "WebWatcher: A Learning Apprentice for the World Wide Web," Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995

[Baeza1996] Baeza-Yates R. and Navarro G., "Integrating contents and structure in text retrieval", SIGMOD Rec. vol. 25, issue 1, 1996

[Baeza1999] Baeza-Yates R., Ribeiro-Neto B. (1999), Modern Information Retrieval, Addison Wesley

[Balabanovic1997] Balabanovic M. and Shoham Y., "Content-Based, Collaborative Recommendation," Communications of the ACM, vol. 40, issue 3, pp. 66-72, 1997

[Barry1995] Barry C. B., Schamber L., "User-defined relevance criteria: A comparison of two studies", Proceedings of the 58th Annual Meeting of the American Society for Information Science, vol. 32, pp. 103-111, 1995

[Belkin1982] Belkin N., Oddy R. N., Brooks H. M., "ASK for information retrieval. Part I, Background and Theory", Journal of documentation, vol. 38, issue 2, pp. 61-71, 1982

[Bernaras1996] Bernaras A., Laresgoiti I., and Corera J., "Building and Reusing Ontologies for Electrical Network Applications", In Proceedings of the European Conference on Artificial Intelligence ECAI-96, 1996

[BernersLee2001] Berners-Lee T., Hendler J. and Lassila O., "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", Scientific American, vol. 284, issue 5, pp. 28-37, 2001

[Bies1995] Bies A., et al., "Bracketing Guidelines for Treebank II Style Penn Treebank Project", University of Pennsylavnia, 1995

[Bollacker2000] Bollacker K. D., et al., "Discovering Relevant Scientific Literature on the Web," IEEE Intelligent Systems, vol. 15, issue 2, pp. 42-47, 2000

[Boose1987] Boose J. H. and Bradshaw J. M., "Expertise Transfer and Complex Problems: Using AQUINAS as a Knowledge-acquisition Workbench for Knowledge-Based Systems," Int. J. Man-Machine Studies, vol. 26, pp. 3-28, 1987

[Borgman1989] Borgman C. L., "All users of information retrieval systems are not created equal: An exploration into individual differences", Inf. Process. Manage., vol. 25, issue 3, pp. 237-251, 1989

[Budanitsky1999] Budanitsky A., "Lexical semantic relatedness and its application in natural language processing", Technical report CSRG-390, Department of Computer Science, University of Toronto, 1999

[Budanitsky2001] Budanitsky A., Hirst G., "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", In Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, pp. 29-34, Pittsburgh, 2001

[Burkowski1992a] Burkowski F. J., "An algebra for hierarchically organized text-dominated databases", Inf. Process. Manage., vol. 28, issue 3, pp. 333-348, 1992

[Burkowski1992b] Burkowski, F. J. (1992), "Retrieval activities in a database consisting of heterogeneous collections of structured text", In Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Copenhagen, Denmark, June 21 - 24, 1992). N. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. SIGIR '92, pp. 112-125 ACM Press, New York.

[Chakrabarti1999] Chakrabarti S., et al., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," Proceedings of the Eighth International WWW Conference,pp. 545-562, 1999

[ChenH2000] Chen H. and Dumais S., "Bringing Order to the Web: Automatically Categorizing Search Results," Proceedings of the CHI 2000 conference on Human factors in computing systems, The Hague Netherlands,pp. 145-152, 2000

[ChenZ2000] Chen Z., et al., "WebSail: from On-line Learning to Web Search," Proceedings of the First International Conference on Web Information Systems Engineering, vol. 1, pp. 206-213, 2000

[Chowdhury1999] Chowdhury G. (1999), Introduction to modern information retrieval, Library Association Publishing, London

[Chu2003] Chu H.(2003), Information Representation and Retrieval in the Digital Age, ASIST Monograph Series, Information Today Inc

[Clark1999] Clark D., "Mad Cows, Metathesauri, and Meaning," IEEE Intelligent Systems, vol. 14, issue 1, pp. 75-77, 1999

[Cleverdon1966] Cleverdon C.W., Mills J., Keen E.M., "Factors determining the performance of indexing systems, vol. 2: Test results", Technical report, Aslib Cranfield Research Project, Cranfield, England, 1966

[Crane1972] Crane D., (1972), Invisibile colleges: Diffusion of knowledge in scientific communities, University of Chicago Press, Chicago

[Crestani1997] Crestani F., "Application of Spreading Activation Techniques in Information Retrieval", Artificial Intelligence Review, vol. 11, issue 6, pp. 453-482, 1997

[Danesi1999] Danesi M., Perron P. (1999), Analyzing Cultures. An Introduction and Handbook, Bloomington, Indiana University Press

[Delors1993] Delors J., "White Paper on Growth, Competitiveness and Employment", CCE, Brussel, 1993

[DeMauro2005] De Mauro (2005), Dizionario della lingua italiana, Paravia

[deVel1998] de Vel O. and Nesbitt S., "A Collaborative Filtering Agent System for Dynamic Virtual Communities on the Web," Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98, Carnegie Mellon University, Pittsburgh, 1998

[ESSIR2005] ESSIR 2005 – European Summer School in Information Retrieval, Dublin, Ireland, Sep. 5-9, 2005

[Fenichel1981] Fenichel C. H., "Online Searching: Measures that Discriminate among Users with Different Types of Experiences", Journal of the American Society for Information Science, vol. 32, pp. 23-32, 1981

[Fensel1999] Fensel D., et al., "On2broker: Semantic-Based Access to Information Sources at the WWW," Proceedings of the World Conference on the WWW and Internet (WebNet 99), pp. 25-30, Honolulu, Hawaii, USA, 1999

[Finkelstein2002] Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z.,Wolfman G., Ruppin E., "Placing Search in Context. The Concept Revisited", ACM Transaction on Information Systems, vol. 20, issue 1, pp. 116-131, 2002

[Frants1988] Frants V. I., Brush C. B., "The need for information and some aspects of information retrieval systems construction", Journal of the American Society for Information Science, vol. 39, issue 2, pp. 86-91, 1988

[Fuhr1979] Fuhr N., "Probabilistic models in information retrieval", The Computer Journal, vol. 35, issue 3, pp. 243-255, 1992,

[Furnas1988] Furnas G. W., Deerwester S., Dumais S. T., Landauer T. K., Harshman R. A., Streeter L. A. and Lochbaum, K. E. (1988) "Information retrieval using a singular value decomposition model of latent semantic structure", In Proceedings of the 11th Annual international ACM SIGIR

Conference on Research and Development in information Retrieval (Grenoble, France). Y. Chiaramella, Ed. SIGIR '88. ACM Press, New York, NY, pp. 465-480

[Gaizauskas1997] Gaizauskas R. and Humphreys K., "Using a semantic network for information extraction", Journal of Natural Language Engineering, vol. 3, issue 2/3, 1997

[Garzanti1980] Dizionario della lingua italiana - Centri Garzanti, 1980

[Genesereth1987] Genesereth M. and Nilsson N. (1987), Logical Foundations of Artificial Intelligence, Morgan Kaufmann

[Greenfield2002] Greenfield D. (2002), Lost in cyberspace: the web at work, The Center for Internet Studies &Psychological Health Associates, University of Connecticut

[Gruber1993] Gruber T. R., "A translation approach to portable ontology specifications", Knowledge Acquisition , vol. 5, issue 2, pp. 199-220, 1993

[Guarino1995] Guarino N., and Giaretta P. (1995), "Ontologies and Knowledge Bases: Towards a Terminological Clarification" in N. Mars (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25-32 , IOS Press, Amsterdam.

[Guarino1999] Guarino N., et al., "OntoSeek: Content-based Access to the Web," IEEE Intelligent Systems, vol. 14, issue 3, pp. 70-80, 1999

[Halliday1976] Halliday M. A. K. and Hasan R. (1976), Cohesion In English, Longman

[Heylighen2002] Heylighen F. (2002), "Complexity and Information Overload in Society: why increasing efficiency leads to decreasing control", CLEA, Free University of Brussels

[Hirst1998] Hirst D., St-Onge G. (1998), "Lexical chains as representations of context for the detection and correction of malapropisms.", In C. Fellbaum, editor, WordNet: An electronic lexical database, chapter 13, pp. 305-332, The MIT Press, Cambridge, MA

[Jiang1997] Jiang J., Conrath D., "Semantic similarity based on corpus statistics and lexical taxonomy". In Proceedings on International Conference on Research in Computational Linguistics (ROCLING X), pp. 19-33, Taiwan, 1997

[Kerschberg2002] Kerschberg L., Kim W., Scime A., "Intelligent Web Search via Personalizable Meta-Search Agents", 2002

[Kim2001] Kim W., Kerschberg L., Scime A., "Personalization in a Semantic Taxonomy-Based Meta-Search Agent," presented at International Conference on Electronic Commerce 2001 (ICEC 2001), Vienna, Austria, 2001

[Klein1994] Klein D. A.(1994), Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition: Lawrence Erlbaum Associates

[Kozima1993] Kozima H., Furugori T., "Similarity between words computed by spreading activiation on an English dictionary", In Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL 93), pp. 232–239, 1993

[Kozima1997] Kozima H., Ito A. (1997). "Context-Sensitive Measurement of Word Distance by Adaptive Scaling of a Semantic Space", In Ruslan Mitkov and Nicolas Nicolov, editors, Recent Advances in Natural Language Processing: Selected Papers from RANLP'95, vol. 136, chapter 2, pp 111-124, John Benjamins Publishing Company, Amsterdam/Phildadelphia

[Krulwich1997] Krulwich B., "Lifestyle Finder," AI Magazine, vol. 18, issue 2, pp. 37-46, 1997

[Lancaster1993]Lancaster F.W. and Warner A. J. (1993), Information retrieval today, Information Resources Press, Arlington, VA

[Latiri2001] Latiri Ch. C., Yahia S. B., "Generating Implicit Association Rules from Textual Data", AICCSA 2001, pp. 137-143, 2001

[Leacock1998] Leacock C., Chodorow M. (1988), "Combining local context and WordNet similarity for word sense identification", In C. Fellbaum, editor, WordNet: An electronic lexical database, chapter 11, pp. 265-283, The MIT Press, Cambridge, MA

[Lee1993] Lee J. H., Kim M. H., Lee Y. J., "Information retrieval based on conceptual distance in IS-A hierarchies", Journal of Documentation, vol. 49, issue 2, pp. 188-207, 1993

[Li2000] Li Y., Bandar Z. A., McLean D., "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE

Transaction on Knowl.edge and Data Engineering, vol. 15, issue 4, pp. 871-882, 2000

[Lin1997] Lin D., "Using syntactic dependency as local context to resolve word sense ambiguity". In Proceedings of ACL/EACL-97, pp 64-71, Madrid, Spain, 1997

[Lin1998] Lin D., "An information-theoretic denition of similarity." In Proceedings of International Conference on Machine Learning, pp.296-304, Madison, Wisconsin, 1998

[Liu2003] Liu H., Monty Tagger 1.2, MIT Media Lab, 2003

[Maes1994] Maes P., "Agents that reduce work and information overload," Communications of the ACM, vol. 37, issue 7, pp. 30-40, 1994

[Marchionini,1993.] Marchionini G. et al., "Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise", Lib. Inf. Sci. Res., vol. 15, issue 1, 35-69, 1993

[Martin1991] Martin M. E. (1991), Analysis and Design of Business Information Systems. Englewood Cliffs, NJ, Prentice hall

[Martin2000] Martin P. and Eklund P. W., "Knowledge Retrieval and the World Wide Web," IEEE Intelligent Systems, vol. 15, issue 3, pp. 18-25, 2000

[Meadow1992] Meadow C.T. (1992), Text information retrieval system, Academic Press., San Diego

[Miller1995] Miller G. A., "WordNet a Lexical Database for English," Communications of the ACM, vol.38, pp. 39-41, 1995

[Moldovan2000] Moldovan D. I. and Mihalcea R., "Using WordNet and Lexical Operators to Improve Internet Searches", IEEE Internet Computing, vol.4, issue 1,pp. 34- 43, 2000

[Morris1991] Morris J. and Hirst G., "Lexical cohesion computed by thesaural relations as an indicator of the structure of text", Computational Linguistics, vol. 17, issue 1, pp. 21–48, 1991

[Navarro1995] Navarro G. and Baeza-Yates R. (1995), "A language for queries on structure and contents of textual databases,. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995),

E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95, pp. 93-101 ACM Press, New York, NY.

[Navarro1997] Navarro G. and Baeza-Yates R., "Proximal nodes: a model to query document databases by content and structure", ACM Trans. Inf. Syst. vol. 15, issue 4, 1997.

[Neches1991] Neches R., Fiches R. E., Finin T., Gruber T. R., Patil R., Senator T. and Swartout W., "Enabling Technology for Knowledge Sharing", AI Magazine, vol. 12. pp.36-56, 1991

[Nielsen1990] Nielsen J., Hypertext and Hypermedia, Academic Press, San Diego, CA, 1990.

[Ogawa1991] Ogawa Y., Morita T. and Kobayashi K., "A fuzzy document retrieval system using the keyword connection matrix and a learning method", Fuzzy Sets Syst., vol. 39, issue 2, 1991

[Okumura1994] Okumura M., Honda T., "Word sense disambiguation and text segmentation based on lexical cohesion", Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94), vol. 2, pp. 755-761, Kyoto, Japan, 1994

[Paisley1968] Paisley W., "Information needs and uses", Annual Review of Information Sience and Technology, 3, pp. 1-30, 1968

[PennTreebankProject2005] Penn Treebank Project: www.cis.upenn.edu/~treebank/, 2005

[Price1963] Price D. (1963), Little science, big science, Columbia University Press, New York

[Rada1989] Rada R., Mili H., Bicknell E., Blettner M., "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man and Cybernetics, vol. 19, issue 1, pp. 17-30, 1989

[Resnik1995] Resnik P., "Using information content to evaluate semantic similarity in a taxonomy", In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453, Montreal, Canada, 1995

[Resnik1999] Resnik P., "Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language", Journal of Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.

[Ribeiro1996] Ribeiro B. A. and Muntz R. (1996), "A belief network model for IR", In Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, pp. 253-260

[Richardson1995a] Richardson R., Smeaton A. F., "Using Word Net in a knowledge-based approach to information retrieval". Working paper CA-0395, School of Computer Applications, Dublin City University, 1995

[Richardson1995b] Richardson R., Smeaton A. F., " Automatic word sense disambiguation in a KBIR application", Working paper CA-0595, School of Computer Applications, Dublin City University, 1995

[Robertson1976] Robertson S. E. and Sparck Jones K., "Relevance weighting of search terms", Journal of the American Society for Information Science, vol. 27, pp. 129-146, 1976

[Rocha2004] Rocha C., Schwabe D., and de Aragao M. P., "A Hybrid Approach for Searching in the Semantic Web", World Wide Web Conference 2004, Semantic web applications, pag 374-383, 2004

[Roget1977] Roget P., (1977), Roget's International Thesaurus, Fourth Edition, Harper and Row Publishers Inc.

[Rubenstein1965] Rubenstein H., Goodenough J. B., "Contextual correlates of synonymy." Communications of the ACM, vol. 8, issue 10, pp. 627-633, 1965

[Saaty1980] Saaty (1980), The Analytic Hierarchy Process, McGraw-Hill, New York

[Salton1971] Salton G. and Lesk M. E., "Computer evaluation of indexing and text processing," in The SMART Retrieval System: Experiments in Automatic Document Processing (G. Salton, ed.), pp. 143-180, Prentice Hall, 1971

[Salton1983a] Salton G. and McGill M. (1983), Introduction to modern information retrieval. McGraw-Hill, New York

[Salton1983b] Salton G., Fox E.A., Wu H., "Extended Boolean Information Retrieval", Communications of the ACM, vol. 26, issue 11, pp. 1022-1036, 1983

[Santorini1990] Santorini B., "Part-of-speech tagging guidelines for the Penn Treebank Project", Department of Computer and Information Science, University of Pennsylvania, 1990

[Saracevic1996] Saracevic T., "Relevance reconsidered", In P. Ingwersen & N.O. Pors, (Eds.), Information science: integration in perspective, pp. 201-218, Copenhagen: Royal School of Library and Information Science, 1996

[Schutz1970] Schutz A. (1970), "Reflections on the problem of relevance", New Haven, CT: Yale University Press

[Shavlik1998] Shavlik J. and Eliassi-Rad T., "Building Intelligent Agents for Web-based Tasks: A Theory-Refinement Approach," Proceedings of the Conference on Automated Learning and Discovery: Workshop on Learning from Text and the Web, Pittsburgh, PA, 1998

[Shaw1991] Shaw D., "The human-computer interface for information retrieval", Journal Annual Review of Information Sience and Technology, vol. 26, pp. 155-195, 1991

[Shepard1987] Shepard R. N., "Towards a universal law of generalisation for psychological science", Science, vol. 237, pp. 1317-1323, 1987

[Sheth2002] Sheth A., Bertram C., Avant D., Hammond B., Kochut K., and Warke Y., "Managing Semantic Content for the Web", IEEE Internet Computing, vol. 6, issue 4, 2002

[SparkJones1997] Spark Jones K. and Willet. P. (1997), Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco

[SparkJones2000] Spark Jones K., "Further reflections on TREC", Information Processing & Management, vol. 36, issue 1, pp 37-85, 2000

[Staab1999] Staab S., et al., "A System for Facilitating and Enhancing Web Search", Proceedings of IWANN '99 - International Working Conference on Artificial and Natural Neural Networks, Berlin, Heidelberg, 1999

[Sussna1993] Sussna M., "Word sense disambiguation for free-text indexing using a massive semantic network", In Proceedings of the Second International Conference on Information and Knowledge Management, (CIKM-93),pp. 67–74, Arlington, Virginia, 1993

[Sussna1997] Sussna M., "Text Retrieval Using Inference in Semantic Metanetworks", PhD thesis, University of California, San Diego, 1997

[Swartout1999] Swartout W., Tate A., "Guest Editors' Introduction: Ontologies", IEEE Intelligent Systems, vol. 14, issue 1, pp. 18-19, 1999

[Terveen1997] Terveen L., et al., "PHOAKS: a System for Sharing Recommendations," Communications of the ACM, vol. 40, issue 3, pp. 59-62, 1997

[Turco2002] Turco M., "Messa a punto di un'ontologia per il web semantico. II caso del settore calzaturiero", Tesi di laurea in Ingegneria Informatica, Università degli Studi di Lecce, 2002

[Turtle1990] Turtle H.R. and Croft W.B. "Inference networks for document Retrieval" In Proceedings of ACM SIGIR, pp. 1-24 Brussels, Belgium, 1990

[Turtle1991] Turtle H. R. and Croft W. B., "Evaluation of an inference network-based retrieval model", ACM Transactions on Information Systems, vol. 9, issue 3, pp. 187-222, 1991

[Uschold1996] Uschold M., King M., "Building ontologies: Towards a unified method", AIAI-TR-197, 1996

[vanRijsbergen1979] van Rijsbergen C.J. (1979), Information Retrieval, 2nd edition, Butterworths, London

[Wellish1996] Wellish H. (1996),. Indexing From A-Z. 2nd Edition. H.W. Wilson

[West1953] West M. P., "A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology", Longman, Harlow, Sussex, 1953

[Wilkinson1991] Wilkinson R. and Hingston P. (1991), "Using the cosine measure in a neural network for document retrieval", In Proceedings of the 14th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Chicago, Illinois, United States, October 13 - 16, 1991). SIGIR '91. ACM Press, New York, NY, pp. 202-210

[Wong1985] Wong S. K. M., Ziarko W., Wong P. C. N., "Generalized vector space model in information retrieval", In ACM SIGIR '85), pp. 18-25, 1985

[Wong1995] Wong S., K. and Yao Y. Y., "On modeling information retrieval with probabilistic inference", ACM Trans. Inf. Syst., vol. 13, issue 1, pp. 38-68, 1995

[Wu1994]Wu Z., Palmer M., "Verb semantics and lexical selection", In Procedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133-138, Las Cruces, New Mexico, 1994

[Zhang2001] Zhang X. and Chignell M., "Assessment of the effects of user characteristics on mental models of information retrieval systems", J. Am. Soc. Inf. Sci. Technol., vol. 52, issue 6, pp. 445-459, 2001