Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Un meta-motore semantico per la ricerca di dati su internet

Stefano Gristina, Giovanni Pilato, Filippo Sorbello, Giorgio Vassallo

RT-ICAR-PA-03-17 **dicembre 2003**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: www.icar.cnr.it
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: www.na.icar.cnr.it
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: www.pa.icar.cnr.it

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Un meta-motore semantico per la ricerca di dati su internet

Stefano Gristina[2], Giovanni Pilato[1],
Filippo Sorbello[2], Giorgio Vassallo[2]

Rapporto Tecnico N.17: ***Data:***
RT-ICAR-PA-03-17 **dicembre 2003**

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sezione di Palermo Viale delle Scienze edificio 11 90128 Palermo
[2] Università degli Studi di Palermo Dipartimento di Ingegneria Informatica Viale delle Scienze 90128 Palermo

Indice

INTRODUZIONE **3**

1 ANALISI DEL PROBLEMA **5**

1.1 INTRODUZIONE 5

1.4 METODOLOGIE DI RAGGRUPPAMENTO DI TESTI E DOCUMENTI WEB 6

2 CARATTERISTICHE DEL DIZIONARIO SEMANTICO WORDNET **14**

2.1 INTRODUZIONE 14

2.2 MATRICE LESSICALE E SUA IMPLEMENTAZIONE 15

2.3 RELAZIONI SEMANTICHE E LESSICALI 16

3 SOLUZIONE PROPOSTA **19**

3.1 INTRODUZIONE 19

3.2 RAPPRESENTAZIONE SEMANTICA DI UN DOCUMENTO 21

3.3 RAPPRESENTAZIONE LESSICALE DI UN DOCUMENTO 38

3.4 METRICHE UTILIZZATE 40

3.5 USO DELL'ALGORITMO DI SAMMON 42

3.6 ALGORITMI DI RAGGRUPPAMENTO 43

4 RISULTATI SPERIMENTALI **46**

4.1 INSIEME DI TEST 46

4.2 PROVE SPERIMENTALI 47

4.3 CONCLUSIONI 50

BIBLIOGRAFIA **51**

Introduzione

Lo sviluppo sempre più crescente di internet ha reso necessario il potenziamento della ricerca e dell'organizzazione delle risorse informative sparse nella rete. I motori di ricerca, pur assistendo gli utenti nell'esplorazione del web, rendono difficile la selezione delle informazioni di proprio interesse poiché i risultati restituiti sono, generalmente, semanticamente eterogenei. Nasce quindi l'esigenza di migliorare l'organizzazione e la visualizzazione semantica dei documenti restituiti.

Il lavoro proposto si pone l'obiettivo di creare un sistema che organizza in maniera "intelligente" i risultati ottenuti da vari motori di ricerca, al fine di mostrare, su un piano cartesiano, raggruppamenti semanticamente omogenei. In questa rappresentazione la distanza euclidea tra due documenti visualizzati sul piano è proporzionale alla differenza esistente tra i loro contenuti informativi. Per rappresentare il contenuto lessicale e semantico di un documento web è stato associato ad ogni documento un insieme lessicale, costituito da parole, opportunamente filtrate, presenti nel corpo del riassunto restituito dal motore di ricerca, e un insieme semantico, costituito dai corretti significati semantici che i sostantivi esprimono nel contesto lessicale in cui sono presenti.

La rappresentazione semantica nasce dalla constatazione che un'espressione lessicale della stessa parola può assumere diversi significati in contesti differenti. Diventa fondamentale riuscire a cogliere i diversi comportamenti semantici della stessa forma lessicale e ciò è stato raggiunto utilizzando il dizionario semantico WordNet.

In caso di assenza di parole nel dizionario semantico che potrebbero essere significative per descrivere il contenuto informativo di un documento, è essenziale la presenza di una rappresentazione lessicale da opporre a quella semantica.

Tramite l'uso di una metrica, definita come una combinazione lineare, con coefficienti opportunamente variabili, della distanza esistente tra gli insiemi lessicali e semantici, è stato definito un modo per confrontare i contenuti lessicali e semantici dei documenti. Diventa a questo punto possibile una visualizzazione, su un piano cartesiano, caratterizzata dalla presenza di gruppi di contenuto lessicale e semantico pressochè simile.

Il problema affrontato riguardante gruppi di documenti si contrappone al caso in cui il dominio d'interesse è rappresentato da testi. La differenza fondamentale esistente nei

due approcci deriva dalla natura differente dell'insieme che costituisce l'ingresso. Il web, costituito da documenti non strutturati ed eterogenei, si contrappone ad un insieme di informazioni testuali strutturate e dislocate, generalmente, in un database. Ciò rende le tecniche utilizzate nel raggruppamento di documenti differenti rispetto a quelle adoperate nel raggruppamento di testi.

1 Analisi del problema

1.1 Introduzione

Il problema di raggruppare testi o documenti web in gruppi di contenuto tematico simile va sotto il nome di raggruppamento di testi o documenti ed è meglio noto come *text clustering* o *document clustering*. Le più comuni soluzioni a questo problema sono caratterizzate dai seguenti due passi:

1. *Trasformazione del documento o testo in un modello matematico*. Generalmente viene effettuata una trasformazione in un vettore dove ogni componente è associata ad una parola e il suo valore memorizza l'importanza posseduta dalla parola nella rappresentazione del contenuto informativo del documento o testo. Un'altra tecnica sviluppata è la trasformazione in un grafo o albero [46].
2. *Applicazione di opportuni algoritmi di raggruppamento, meglio noti come algoritmi di clustering, ai modelli matematici ottenuti*. Possono essere utilizzati algoritmi appositamente creati, o fare ricorso ad approcci standard nel *text* o *document* clustering come l'utilizzo di Sammon [3], delle reti neurali SOM [47] o dell'algoritmo K-Means [42].

Il raggruppamento di documenti web, pur avendo in comune queste fasi col raggruppamento di testi, presenta delle differenze dipendenti dalla sorgente che caratterizza il dominio d'ingresso.

Per il raggruppamento di documenti il dominio d'interesse è rappresentato da documenti web, mentre il raggruppamento di testi è interessato ad un database costituito da un insieme di testi strutturati. Inoltre i settori di ricerca in cui si inquadrano sono differenti. Il raggruppamento di documenti fa parte dell'area di ricerca, che si occupa di estrazione di conoscenze e risorse dal web, nella quale numerose ricerche sono state fatte, come calcoli linguistici, statistici ed informativi.

Il raggruppamento di testi, invece, si colloca nel settore che ha come obiettivo la deduzione di informazioni e conoscenze da un insieme di testi strutturati ed omogenei dislocati in genere su un database.

1.4 Metodologie di raggruppamento di testi e documenti web

1.4.1 Estrazioni di caratteristiche da testi

Il primo passo da affrontare nel raggruppamento di testi consiste nell'estrazioni di caratteristiche, elementi ad alto potere discriminatorio che permettono di associare ad un testo un modello matematico fondamentale per una successiva applicazione di un algoritmo di raggruppamento. In genere, le caratteristiche sono rappresentate dalle parole presenti all'interno del testo. Si presenta il problema che una collezione di testi può contenere un numero elevato di parole.

In generale due vettori casuali in un ipercubo ad alta dimensionalità tendono ad avere una distanza costante uno dall'altro. Detto in altri termini, si ha che all'aumentare della dimensionalità dei vettori, diminuisce la varianza della funzione di distanza normalizzata e di conseguenza i gruppi, che prima erano ben definiti, tendono ad avvicinarsi.

L'eliminazione delle parole comuni (stop words) e successivamente un riporto alla loro radice comune, (comput, informat, etc..) mediante l'applicazione dell'algoritmo di Porter [22], aiuta, ma non risolve del tutto il problema. Diventa necessario applicare opportune tecniche di estrazione delle parole.

Comunemente vengono utilizzati due tipi di approcci.

1. *Metodo di Zipf*: consiste nell'eliminare le parole aventi una frequenza, calcolata sull'intera collezione di documenti, maggiore e minore di opportuni valori. Parole ad alta frequenza, comuni a molti testi, e di bassa frequenza, poco presenti, non sono rilevanti nell'individuazione dei vari gruppi tematici.

2. *Metodo entropico*: deriva dalla formula base della teoria dell'informazione e consiste nell'estrarre le parole aventi massima entropia, che viene calcolata come prodotto tra la probabilità di occorrenza di una parola e il logaritmo della stessa. Maggiore è l'entropia associata ad una parola maggiore risulta il suo contenuto informativo.

Superata la fase di estrazione delle caratteristiche è possibile avere a disposizione una forma di rappresentazione matematica del documento che permette una successiva applicazione di un algoritmo di raggruppamento.

1.4.2 Estrazioni di caratteristiche dal web

L'estrazione di caratteristiche da documenti web, pur avendo molti punti in comune con il caso in cui il dominio d'interesse è rappresentato da semplici testi, presenta delle caratteristiche uniche che la differenziano. Le diversità derivano dalla struttura intrinseca del web, che risulta costituita da un grafo consistente di nodi, rappresentati da documenti, e collegamenti tra essi, dati dalla presenza di *hyperlinks*. Questa struttura permette la possibilità di sfruttare, per rappresentare il contenuto informativo di un documento, non solo il suo insieme di parole, come viene fatto nella fase di estrazione di caratteristiche da testi, ma anche la struttura di una pagina web, che risulta divisa, come descritto dalla figura 2, in struttura esterna ed interna.

La struttura esterna tiene conto dei collegamenti esistenti tra pagine web, mentre la struttura interna è determinata dal numero di immagini, email e links, etc, presenti all'interno di un documento web.

Un approccio interessante consiste nel rappresentare l'informazione di un documento mediante due insiemi che sintetizzano il contenuto lessicale e strutturale di una pagina web e combinarli in modo opportuno per definire una metrica che permetta di effettuare un confronto tra due documenti. Gli insiemi, funzione del contenuto lessicale e strutturale, sono una rappresentazione matematica del documento, che viene dettagliatamente descritta nel paragrafo seguente.

1.4.3 Rappresentazione matematica

1.4.3.1 Introduzione

Esistono sostanzialmente due rappresentazioni matematiche da associare ad un documento o testo. Una rappresentazione basata su un vettore, opportunamente pesato, di parole, contrapposta ad una rappresentazione costituita da un grafo semantico, dove i nodi, associati alle parole contenute nel testo, sono connessi da relazioni semantiche.

E' possibile, determinato il modello, scegliere l'opportuno algoritmo di *text* o *document clustering*. Prima di descrivere i vari algoritmi di clustering comunemente

utilizzati, viene effettuata, nel prossimo paragrafo, una descrizione delle due rappresentazioni matematiche.

1.4.3.2 Rappresentazione basata su vettori

Il modello matematico più utilizzato nel settore del *text* o *document clustering* è rappresentato da un vettore in cui ogni componente è associata ad una parola precedentemente filtrata. Nasce il problema di stabilire quale valore attribuire alle componenti del vettore. Si possono utilizzare due metodi, uno basato su un vettore opportunamente pesato, l'altro su un vettore binario.

Per quanto riguarda il primo metodo, ogni componente del vettore viene pesata con un valore proporzionale al potere discriminatorio che la parola ad essa associata possiede all'interno del documento.

Una formula utilizzata al fine di assegnare un peso ad una componente risulta:

$$Wij=Tfij*log(N/Dfj) \qquad \textbf{(4)}$$

Dove:

Wij- peso del termine *Tj* nel documento *Di*

Tfij- frequenza del termine *Tj* nel documento *Di*

N- numero totale di documenti nella collezione

Dfj- numero di documenti contenenti il termine *Tj*

Uno schema di pesi di questo tipo favorisce i termini che sono presenti frequentemente all'interno di un documento, ma raramente nell'intera collezione. Normalmente i documenti possono essere di dimensione variabile, e di conseguenza per permettere il loro confronto è necessaria una opportuna normalizzazione, che può essere effettuata utilizzando la formula seguente:

$$Xnormalizedi = Xi/(\sqrt{X1 + X2 + X3 + ...Xn} \qquad \textbf{(5)}$$

Dove:

Xi-componente del vettore di posizione i

La presenza di una parola, ad alto potere discriminatorio per il documento nel quale è presente, riduce il peso di tutte le altre ai minimi termini.

La normalizzazione basata sulla formula precedente, da un punto di vista matematico, trasferisce il vettore delle parole su una sfera unitaria in uno spazio a n dimensione.

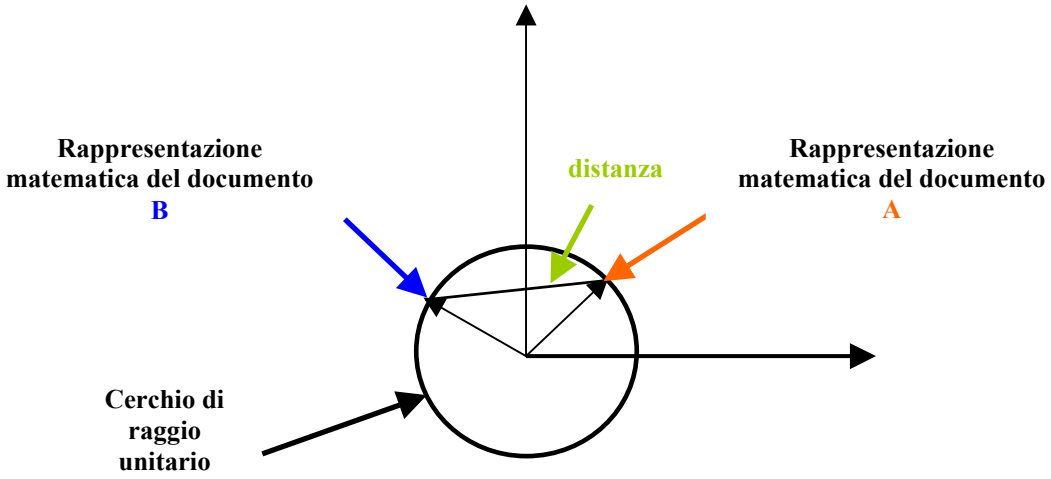

Figura 3: Esempio di rappresentazione matematica normalizzata di due documenti in uno spazio a due dimensioni.

Nel secondo metodo si utilizza un modello matematico rappresentato da un vettore binario, in cui ogni componente è associata ad una parola precedentemente filtrata. Ogni componente del vettore assume il valore 0 o 1 in funzione del fatto che, la parola associata ad essa, sia presente o meno nel documento. Questo modello matematico è stato utilizzato nel lavoro proposto per determinare la rappresentazione lessicale e semantica (in questo caso si parla di concetti) da associare ad ogni riassunto.

Un lavoro interessante, in quest'ottica, fatto da Vlajic e Card [14], consiste nell'associare due vettori ad un documento. Il primo viene calcolato come descritto nel primo metodo, l'altro è un vettore di 12 dimensioni (detto vettore *ipertestuale*). Ciascuna dimensione di quest'ultimo, è definita come una funzione che riceve, come ingresso, una delle caratteristiche seguenti:

profondità dell'indirizzo url, dimensione della pagina web, numero di collegamenti, percentuale collegamenti esterni ed interni, numero di immagini, email, presenza di file (java, exe, etc..).

Possiamo inquadrare l'approccio utilizzato sia nel settore del *document mining*, che nel *structure mining.*

La metrica utilizzata per confrontare due vettori e quindi due documenti è data dalla seguente formula:

$$\alpha * DistWord + \beta * DistHyper \quad \textbf{(6)}$$

Dove:

DistWord- Rappresenta la distanza euclidea tra due vettori delle parole.

DistHyper - Rappresenta la distanza euclidea tra due vettori ipertestuali.

α, β- Parametri modificabili per dare un peso diverso alle due distanze.

Scegliendo in modo opportuno α e β è possibile effettuare un raggruppamento che tiene conto in modo diverso del contenuto e della struttura interna di un documento web. Per esempio, se $\alpha \neq 1$ e $\beta=0$, si produce un raggruppamento basato sul contenuto testuale, mentre se α e β sono entrambi diversi da zero, sia il vettore delle parole che quello ipertestuale influenzano il risultato finale. Il rapporto α/β determina se l'informazione testuale o strutturale è più o meno decisiva nel determinare la formazione dei vari gruppi tematici.

1.4.4 Algoritmi di raggruppamento

1.4.4.1 Mappe di Kohonen

Una mappa di Kohonen o Self Organization Map[42] è una rete neurale non supervisionata che ha l'obiettivo di inserire vettori simili in regioni dello spazio relativamente vicine. Il modello consiste in un griglia di unità di elementi di processo detti neuroni. A ciascuna delle unità è assegnata un *m* vettore i cui pesi hanno la stessa dimensionalità dei vettori d'ingresso. Il processo di raggruppamento può essere descritto in termini di presentazione dei vettori d'ingresso e adattamento dei pesi del *m* vettore.

Si inizia con la selezione casuale di uno degli ingressi. Viene presentato alla rete e ciascuna unità determina la sua attivazione, calcolata come distanza euclidea tra il vettore m e il corrispondente ingresso. I pesi dell'unità con l'attivazione più alta, chiamato neurone vincente, e dei suoi vicini sono adattati in modo tale da ridurre la differenza tra le componenti corrispondenti del vettore d'ingresso e i vettori pesati dell'unità stessa. In questo modo il neurone vincente e i suoi vicini avranno una maggiore tendenza a vincere per una presentazione di ingressi simili, e ciò condurrà ad una raggruppamento spaziale di vettori d'input simili in regioni vicine della griglia di neuroni.

Una variante, utilizzata nel raggruppamento di testi gerarchico (hierarchical text clustering),), è data dalle Hierarchical Self Organization Map (HSOM) [38]. Una hierarchical SOM utilizza un algoritmo di raggruppamento simile alle SOM, ma inizia al top level con tutti i documenti a disposizione per il training. Quando il top level è finito, procede nei livelli più bassi. Solo i documenti che sono presenti nei nodi genitori al top level sono usati per ciascuna griglia nel livello più basso. Allora lo stesso processo è ripetuto per i livelli seguenti, ottenendo un raggruppamento gerarchico di un'intera collezione di documenti.

1.4.4.2 Algoritmo K-Means

K-Means[42] è una semplice procedura iterativa che viene utilizzata per partizionare *N* punti, definiti in uno spazio N-dimensionale, in *K* disgiunti insiemi *Sj*, contenenti *Nj* punti, al fine di minimizzare la seguente metrica:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \|x_n - \mu_j\|^2, \tag{7}$$

Dove *xn* è un vettore rappresentante l'ennesimo punto, e μ_n è il centroide geometrico dell'insieme dei punti in *Sj*.

Il centroide è un punto artificiale che rappresenta la locazione media di un ben determinato gruppo di punti. Le sue coordinate sono ottenute effettuando la media aritmetica delle coordinate di tutti i punti appartenenti al gruppo. I passi fondamentali che costituiscono l'algoritmo risultano :

1. Selezionare casualmente k punti e assumerli come semi per i centroidi di k gruppi.
2. Assegnare ciascun punto al centroide più vicino, formando in questo modo k gruppi mutuamente esclusivi.
3. Calcolare i nuovi centroidi dei gruppi mediando aritmeticamente le coordinate dei punti appartenenti allo stesso gruppo.
4. Controllare se sono cambiate le coordinate dei centroidi dei gruppi. Se si, ritornare al passo 2, altrimenti, l'individuazione dei gruppi è terminata e di conseguenza i punti, che costituiscono l'insieme d'ingresso, sono assegnati ciascuno ad un ben determinato gruppo.

Il problema che rimane aperto è legato alla scelta del *k* ottimale.

I risultati che si ottengono non sono soddisfacenti se il parametro k non è scelto in modo tale da venire incontro alla struttura dei dati in ingresso. Un metodo per alleviare il problema consiste nel sperimentare con diversi *k*. In linea di principio, il miglior k esibisce la più piccola distanza all'interno di un gruppo e la più grande tra gruppi differenti.

1.4.4.3 Algoritmo di Sammon

L'obiettivo dell'algoritmo di Sammon [3] è, considerato un set di *n* punti in uno spazio *m* ad alta dimensionalità, trovarne altrettanti in uno spazio *d*-dimensionale con *d* minore di *m*, in modo tale che le corrispondenti distanze approssimino le originali il più possibile.

Denotiamo con:

- d_{ij} $\forall$ i,j=1…n, la distanza tra due punti in uno spazio a *m* dimensioni
- δ_{ij} $\forall$ i,j=1…n, la distanza tra due punti in uno spazio a *d* dimensioni

Senza perdere di generalità ci possiamo soffermare solo sulle proiezioni negli spazi a due dimensioni (d=2).

Fatta questa non limitativa ipotesi, scaturisce il problema di decidere se una configurazione è migliore di un'altra. A tal fine, la funzione d'errore E, che misura la differenza tra la presente configurazione di *n* punti nello spazio *d*-dimensionale e la configurazione degli stessi nello spazio originale, è considerata:

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}} \qquad \textbf{(8)}$$

Il problema di trovare la giusta configurazione in uno spazio a bassa dimensione diventa un classico problema di ottimizzazione, che va risolto riducendo la differenza $(\delta_{ij} - d_{ij})$, fino a quando l'errore varia di una quantità piccola fissata a priori.

2 Caratteristiche del dizionario semantico WordNet

2.1 Introduzione

WordNet è un dizionario lessicale costituito da sostantivi, verbi ed aggettivi organizzati in insiemi di sinonimi, ciascuno rappresentante un ben determinato concetto lessicale [11]. Relazioni semantiche differenti tra i vari concetti completano la struttura e l'organizzazione del dizionario. Questo nuovo modo di gestire la conoscenza rende WordNet più ricco di informazione e più utile di un dizionario convenzionale.

Un esempio permette di caratterizzare le omissioni e mancanze di un normale dizionario. Un dizionario convenzionale definisce uno dei significati del sostantivo *tree* con un'espressione del tipo:

a plant that is large, woody, perennial, and has a distinct trunk.

Nella definizione non viene inserito che un albero ha radici, che è un organismo vivente. Manca l'informazione circa quale senso di pianta ci si riferisce. Un lettore, che vorrebbe determinare altri tipi di piante, dovrebbe scandire il dizionario dalla lettera A alla Z e cercare nelle definizioni la presenza della parola *plant*. Lo stesso ragionamento vale per chi è interessato alla conoscenza di differenti tipi di albero. Inoltre non è possibile accedere ad informazioni, che normalmente non vengono inserite nella definizione, ma che interessano molto da vicino il concetto di albero. Non è possibile sapere che gli alberi hanno una corteccia e rami, che provvedono ombra e protezione dal vento, che crescono in foreste selvaggie, etc…

Tutte queste informazioni, assenti in un normale dizionario, sono presenti nel dizionario semantico WordNet grazie alla presenza di relazioni semantiche esistenti tra insiemi di sinonimi. Le relazioni semantiche permettono, per esempio, di relazionare il concetto *tree,* decritto precedentemente, con i seguenti concetti:

- Il concetto immediatamente generico *plant.*
- I concetti che rappresentano diversi tipi di albero.
- I concetti che rappresentano elementi facenti parte della struttura fisica di *tree*

Nei paragrafi seguenti si descrivono le caratteristiche principali di WordNet e il modo in cui sono state implementate.

2.2 Matrice lessicale e sua implementazione

Comunemente una parola è utilizzata per riferirsi sia alla sua espressione che al suo significato semantico. Per ridurre l'ambiguità ci si riferisce con *word form* alla espressione fisica della parola, con *word meaning* al concetto che può essere associato ad essa. Nella figura 16 viene descritto in modo intuitivo e semplice la nozione di matrice lessicale.

Le *word form* sono presenti come indici delle colonne, mentre le *word meaning* come indici per le righe. Un elemento inserito in una cella implica che la *word form* in quella colonna può essere utilizzata per esprimere il concetto di quella riga. L'elemento *E1.1* rappresenta che la *word form F1* può essere usata per esprimere il *word meaning M1*. Se ci sono due elementi nella stessa colonna, la *word form* ha più sensi o significati. Invece, se ci sono due elementi nella stessa riga, le due *word form* sono sinonime.

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				. . .	
M_m					$E_{m,n}$

Figura 6: Illustra il concetto di una matrice lessicale: F1 e F2 sono sinonimi, F2 è polisensa

La corrispondenza tra espressioni lessicali e concetti è molti a molti. Esistono parole che hanno diversi significati, e concetti che possono essere espressi da diverse parole. Le relazioni che coinvolgono i concetti sono note come relazione semantiche, mentre quelle che interessano le parole relazioni lessicali. Al presente, il dizionario semantico WordNet le rappresenta in modo differente.

I concetti vengono implementati elencando i sinonimi che possono essere utilizzati per rappresentarli. Il significato *M1* della figura precedente viene rappresentato dai sinonimi *F1* e *F2*, compresi tra parentesi graffe ({*F*1,*F2* }). Se una parola è polisensa, sarà presente in più insiemi di sinonimi, chiamati in WordNet *synset*.. Ad ogni *synset*, per descrivere il concetto rappresentato e per differenziare concetti diversi caratterizzati dalla stessa forma lessicale, è associato una breve descrizione (*glos*s).

Per esempio, uno dei sensi della parola *board* è definito dalla seguente descrizione:

{*board,(a person's meals,provided regularly for money*}

La gloss serve, oltre per esplicare il concetto, per distinguerlo dagli altri sensi di *board* con un solo membro. La sinonimia è una relazione lessicale che, come visto precedentemente, viene rappresenta utilizzando le parentesi graffe, mentre tutte le altre relazioni lessicali, che hanno in WordNet un ruolo inferiore, vengono modellate includendo le parole interessate tra parentesi quadre.

Le relazioni semantiche esistenti tra concetti vengono implementate utilizzando dei puntatori.

In seguito vengono descritte le relazioni, lessicali e semantiche, presenti in WordNet.

2.3 Relazioni Semantiche e lessicali

2.3.1 Sinonimia

Due parole sono considerate sinonime se la sostituzione una nell'altra non modifica il valore di verità della frase nella quale la sostituzione è avvenuta. Questa definizione implica che i sinonimi veri, se esistono, sono rari. Una versione più debole considera i sinonimi relativi al contesto. Due parole sono sinonimi relativi ad un contesto *C* se la sostituzione una nell'altra non cambia il valore di verità in *C*. In base a quest'ultima definizione si rende necessario suddividere WordNet in sostantivi, verbi aggettivi e avverbi. Infatti, se i concetti sono rappresentati da synset, e i sinonimi devono essere sostituibili, allora parole in differenti categorie sintattiche non possono essere sinonime (non possono formare synset).

Un'importante proprietà della relazione lessicale, esistente tra sinonimi, è la simmetria. Se x è sinonimo di y allora è conveniente e naturale assumere y sinonimo di x.

2.3.2 Antinomia

L'antinomo di una parola *x* è qualcosa che è non *x*, ma non sempre. Per esempio, ricco e povero sono antinomi, ma qualcuno che è non ricco non implica che è assolutamente povero.

E' una relazione lessicale tra parole, non una relazione semantica tra concetti. Infatti, i concetti {*rise,asce*nd} e {*fall,descend*} possono essere concettualmente opposti ma non sono antinomi. Non è comunemente accettato l'antinomia tra *rise* e *descend* o *fall* e *ascend*.

Come verrà descritto, la relazione di antinomia assume un ruolo fondamentale nella descrizione degli aggettivi.

2.3.3 Iponimia e Iperonimia

A differenza della sinonimia e antinomia, che sono relazioni lessicali tra parole, l'iponimia/iperonimia (note come *hyponymy/ hypernymy*) è una relazione semantica che coinvolge i concetti, ovvero le *word meaning*.

Un concetto rappresentato dal synset {X_1, X_2,…} è detto essere un iponimo del concetto associato al synset {Y_1, Y_2,…} se si accettano espressioni del tipo:

An x is (kind of) a y.

La relazione di iperonimia è inversa a quella di iponimia.

Un concetto rappresentato dal synset {X_1, X_2,…} è detto essere un iperonimo del concetto associato al synset {Y_1, Y_2,…} se si accettano espressioni del tipo:

An y is (kind of) of y.

Le relazioni sono implementate includendo nel primo synset un puntatore al suo concetto generico (hypenymy), e nel secondo un puntatore al concetto direttamente più specifico (hyponymy). Le relazioni sono asimmetriche e generano una struttura gerarchica semantica, nella quale un iponimo si trova al di sotto del suo diretto concetto generico. Un iponimo eredita tutte le caratteristiche del concetto generico, aggiungendone delle nuove che lo differenziano da esso e da tutti gli altri iponimi dello stesso. Per esempio, *maple* eredita le caratteristiche del suo concetto generico *tree,* ma è distinto da esso e da tutti gli altri alberi dalla leggerezza del suo tronco, dalla forma delle sue foglie, etc…

Come verrà descritto in seguito, queste relazioni giocano un ruolo fondamentale nell'organizzazione dei sostantivi in WordNet.

2.3.4 Meronimia e Olonimia

E' la relazione parte-intero (part-whole), comunemente nota ai linguistici come meronimia/olonimia (*meronymy/holonymy)*.

Un concetto rappresentato dal synset {X_1, X_2,…} è un meronimo del concetto associato al synset {Y_1, Y_2,…} se si accettano delle espressioni del tipo:

A y has a x (as a part) o A x is a part of y.

La relazione di olonimia è inversa a quella di meronimia.

Un concetto rappresentato dal synset {X_1, X_2,…} è un olonimo del concetto associato al synset {Y_1, Y_2,…} se si accettano delle espressioni del tipo:

A x has a y (as a part) o A y is a part of x.

Si tratta di una relazione asimmetrica e viene rappresentata in WordNet da puntatori che collegano un synset ad un altro.

Le relazioni semantiche di meronimia e olonimia, unite alle precedenti, già fanno intuire che WordNet non è altro che una vasta rete semantica, dove i concetti o significati semantici, costituiti parole sinonime, sono connessi da differenti tipi di relazioni.

2.3.5 Relazioni Morfologiche

Un' importante classe di relazioni lessicali è data dalle relazioni morfologiche esistenti tra parole. Inizialmente gli interessi sono stati concentrati sulle relazioni semantiche, ma nessun sostanziale cambiamento è stato fatto per includere le relazioni morfologiche. Sono stati aggiunti dei programmi che permettono di individuare ed eliminare le variazioni a cui vanno soggette le parole delle diverse categorie sintattiche.

Ne deriva che l'utente può cercare nel dizionario la parola *trees,* ed ottenere le stesse informazioni che avrebbe ottenuto se avesse cercato la sua forma singolare *tree.*

3 Soluzione proposta

3.1 Introduzione

L'obiettivo del lavoro proposto consiste nel migliorare l'organizzazione dei documenti restituiti da vari motori di ricerca attraverso una loro visualizzazione, su un piano cartesiano, caratterizzata dalla presenza di gruppi semanticamente omogenei. In questo modo gli utenti possono focalizzare la loro attenzione all'interno dei gruppi semantici di loro interesse, e ridurre il tempo e lo sforzo impiegato nella consultazione di documenti non attinenti alle proprie necessità.

L'obiettivo posto viene raggiunto tramite la creazione di un sistema che organizza in maniera "intelligente" i risultati ottenuti da vari motori di ricerca, al fine di mostrare, su un piano cartesiano, raggruppamenti semanticamente omogenei. In questa rappresentazione la distanza euclidea tra due documenti visualizzati sul piano è proporzionale alla differenza esistente tra i loro contenuti informativi. Per dare una visione chiara ed intuitiva dei gruppi visualizzati sul piano, i documenti appartenenti ad essi vengono contraddistinti da un ben determinato colore. Inoltre, è permessa la possibilità di scegliere, tramite interfaccia grafica, diverse metriche, utilizzate per confrontare i contenuti informativi dei documenti, che fanno uso, in modo più o meno combinato, di approcci lessicali e semantici.

Per rappresentare il contenuto lessicale e semantico di un documento web è stato associato ad ogni documento un insieme lessicale, costituito da parole, opportunamente filtrate, presenti nel corpo del riassunto restituito dal motore di ricerca, e un insieme semantico, costituito dai corretti significati semantici che i sostantivi esprimono nel contesto lessicale in cui sono presenti.

L'introduzione di un insieme semantico nasce dalla constatazione che un'espressione lessicale della stessa parola può assumere significati diversi in contesti differenti. Per esempio, l'espressione lessicale *Giove* ha in alcuni contesti il significato inerente a pianeta, in altri quelli relativo al Dio greco. Diventa necessario riuscire a cogliere i diversi comportamenti semantici della stessa forma lessicale e ciò è stato raggiunto utilizzando il dizionario semantico WordNet.

Sono proposti, per la disambiguazione delle parole, diversi approcci che sfruttano le relazioni di sinonimia, iperonimia, iponimia, meronimia e olonimia.

L'insieme lessicale, che è costituito da parole opportunamente filtrate presenti in ogni riassunto, ha un ruolo di fondamentale importanza. Infatti, a causa dell'incapacità del dizionario semantico WordNet di relazionare semanticamente sostantivi propri {*Bush, president*} e causa dell'assenza di parole che potrebbero essere significative per descrivere il contenuto informativo di un documento {*perl, UML, lex, yacc*), è necessaria una rappresentazione lessicale da opporre a quella semantica.

Il modello matematico complessivo che ne deriva è costituito da due insiemi, che rappresentano rispettivamente il contenuto lessicale e semantico.

Resta da definire la metrica utilizzata per confrontare i contenuti lessicali e semantici tra due documenti e l'algoritmo di raggruppamento da adottare. La metrica adoperata è una combinazione lineare, con coefficienti variabili dall'utente, della distanza esistente tra gli insiemi lessicali e semantici associati ai documenti, e ciò comporta la possibilità di avere un raggruppamento in cui l'informazione lessicale o semantica può risultare più o meno decisiva.

Per quanto riguarda l'algoritmo di raggruppamento, è stato del tutto naturale, vista la natura del problema, utilizzare l'algoritmo di Sammon[3].

3.2 Rappresentazione semantica di un documento

3.2.1 Introduzione

La rappresentazione semantica di un documento viene ottenuta tramite un insieme costituito dai corretti significati semantici che i sostantivi possiedono nel contesto lessicale in cui sono presenti. L'introduzione di un insieme semantico nasce dalla constatazione che un'espressione lessicale della stessa parola può assumere significati diversi in contesti differenti. Per esempio, l'espressione lessicale *Giove* ha in alcuni contesti il significato inerente a pianeta, in altri quelli relativo al Dio greco. Diventa necessario riuscire a cogliere i diversi comportamenti semantici della stessa forma lessicale e ciò è stato raggiunto utilizzando il dizionario semantico WordNet.

Sono proposti due algoritmi di disambiguazione che sfruttano le relazioni di iperonimia/iponimia presenti tra insiemi di sinonimi. La motivazione dietro gli approcci utilizzati nasce dalla constatazione che in un qualsiasi testo i corretti significati delle parole che lo costituiscono sono in un certo qual modo relazionati semanticamente mediante la relazione di iperonimia/iponimia. Il significato semantico di una parola maggiormente relazionato con i sensi delle parole del contesto lessicale in cui è presente permette di determinare il suo corretto significato.

Nasce il problema di disambiguare le parole che, pur essendo relazionate semanticamente, risultano in WordNet non connesse. Le parole *plant* e *leaf*, per esempio, appartengono entrambi al contesto *flora*, ma in esso WordNet include solo *plant*, mentre *leaf* è assegnato al contesto object. Tutto ciò porta all'esistenza di parole che non possono essere disambiguate utilizzando la relazione di iperonimia. Deriva la necessità di apportare dei miglioramenti ai due algoritmi di disambiguazione e ciò viene ottenuto sfruttando il contenuto lessicale presente nella descrizione dei sinonimi, meronimi, olonimi e iperonimi dei sensi della parola da disambiguare. Infatti può verificarsi che uno dei sensi di una parola, pur non essendo relazionato tramite la relazione di iperonimia con le parole del contesto lessicale in cui è presente, possieda, nella descrizione (gloss) dei suoi olonimi, meronimi e iperonimi, qualche parola del contesto lessicale. Sfruttando queste informazioni è possibile riuscire ad associare il corretto significato semantico a parole rimaste ambigue dall'applicazione di uno dei due algoritmi di disambiguazione.

Le varie fasi che, partendo dai risultati ottenuti dai motori di ricerca, portano alla formazione degli insiemi semantici associati ai documenti risultano:

1. Memorizzazione dei risultati ottenuti alla query sottoposta a vari motori di ricerca.
2. Applicazione della fase di estrazione delle caratteristiche all'insieme dei sostantivi presenti nel riassunto di ogni documento.
3. Applicazione di uno dei due algoritmi di disambiguazione proposti .
4. Applicazione degli algoritmi che sfruttano il contenuto lessicale presente nella descrizione dei sinonimi, meronimi, olonimi e iperonimi dei sensi delle parole rimaste ambigue.
5. Eliminazione dei sostantivi rimasti ambigui.

3.2.2 Coesione Lessicale

Un insieme semantico, associato ad ogni riassunto, è costituito da una sequenza di concetti relazionati semanticamente nel testo. Per esempio, in un documento relativo ad autovetture, un probabile insieme semantico può essere rappresentato dai concetti espressi dalle parole appartenenti alla lista, {*engine, vehicle, wheel, car,automobile*}, dove ciascuna parola nell'insieme, o meglio il concetto ad essa associato, è direttamente o indirettamente legato mediante relazioni come iperonimia, iponimia, meronimia e olonimia.

Le relazioni semantiche tra unità lessicali presenti in un testo sono state descritte e opportunamente inquadrate in modo formale da Halliday e Hasan [30].

Un qualsiasi testo non è costituito da frasi non relazionate, ma queste sono in effetti connesse a ciascun altra in modo coeso. Per coesione si intende una relazione semantica tra frasi in un testo ed è stata divisa in tre classi: *reference, conjuction* e *lexical cohesion.*

1. *Conjuction*: è la sola classe che mostra esplicitamente le relazioni tra due espressioni. Un esempio è dato dalla seguente frase: *I have a cat and his name is felix*.
2. *Reference*: si verifica soprattutto quando si utilizzano pronomi per riferirsi a parole presenti in qualche espressione. Un esempio è dato dalla seguente frase: *Get inside now!" shouted the teacher. When nobody moved, he was*

furious. In questo caso i pronomi sono il modo per scoprire le relazioni semantiche all'interno di un contesto lessicale.

3. *Lexical cohesion*: si basa sull'individuazione di elementi lessicali e delle loro relazioni semantiche. Per esempio nella frase, *I parked outside the library, and then went inside the building to return my books,* la coesione è rappresentata dalle relazioni semantiche tra le unità lessicali: *library, building* e *books*.

Reference e *lexical cohesion* indicano, a differenza della *conjunction*, relazioni tra frasi in termini di parole semanticamente simili. L'ultimo tipo di relazione è funzione delle relazioni semantiche esistenti tra i significati delle parole. Sfruttando queste relazioni è possibile attribuire ad ogni parola di un testo il corretto significato semantico. Per risolvere il problema, che nella letteratura scientifica va sotto il nome di *word sense disambiguation,* viene utilizzato il dizionario semantico WordNet.

3.2.3 Uso del dizionario semantico WordNet

In Word-Net, sostantivi, aggettivi, verbi e avverbi sono raggruppati in synset (gruppi di parole sinonime) organizzati, per ogni categoria sintattica, in due file lessicali. L'uso che viene fatto di WordNet consiste nel determinare le relazioni semantiche tra la parola candidata ad essere disambiguata e le altre presenti nel suo contesto lessicale. Si è solo interessati ai sostantivi perché i verbi non hanno relazioni semantiche con le altre categorie sintattiche e gli aggettivi hanno solo relazioni unidirezionali.

Le relazioni semantiche che sono state sfruttate, al fine di assegnare ad ogni sostantivo il corretto significato semantico posseduto nel contesto lessicale in cui è presente, sono quelle che caratterizzano l'organizzazione semantica dei sostantivi e quindi risultano essere le relazioni di iponimia, iperonimia, meronimia e olonimia.

Ogni parola che rimane ambigua o che è non relazionata semanticamente con le altre parole del contesto lessicale in cui è presente, viene considerata irrilevante a descrivere il contenuto semantico del testo ovvero non partecipa alla sua struttura coesa [30] e di conseguenza viene scartata.

3.2.4 Estrazione delle caratteristiche

Il processo di estrazione delle caratteristiche è il primo passo da affrontare in un lavoro di raggruppamento di documenti. Da esso dipende la bontà e la correttezza dei

risultati sperimentali. Si è interessati soltanto all'estrazione dei sostantivi presenti nel corpo del riassunto, dal momento che i verbi non sono relazionati semanticamente con essi, e gli aggettivi possiedono solo relazioni unidirezionali.

Tra i sostantivi presenti ci possono essere parole con basso potere discriminatorio che portano ad inevitabili errori nella successiva fase di disambiguazione. Diventa necessario ridurli al minimo e a tal fine verte l'eliminazione delle parole di senso comune (*stop words*) e l'applicazione del metodo di *Zipf*[45], che permette di eliminare le parole poco rilevanti nel determinare la tematica tipica di un documento.

Le fasi principali che caratterizzano il processo di estrazione delle parole risultano:

1. Eliminazioni delle parole di senso comune, cosiddette *stop words*, che non contribuiscono a caratterizzare la tematica tipica associata ad un riassunto.
2. Eliminazione delle parole che non appartengono alla categoria sintattica rappresentata dai sostantivi.
3. Trasformazione dei sostantivi plurali nella loro forma singolare.
4. Applicazione del metodo di Zipf. Vengono eliminate le parole aventi una frequenza di occorrenza nell'intero insieme di documenti scaricati maggiore di 0.005 e minore di 1. I sostantivi, con frequenza di occorrenza elevata, pur possedendo un basso potere discriminatorio, assumono un ruolo decisivo nel raggruppamento poiché possono assumere nei diversi contesti lessicali in cui sono presenti differenti significati semantici.

I sostantivi, che si ottengono dopo aver superato la fase di estrazione delle caratteristiche, costituiscono l'insieme d'ingresso dato all'algoritmo di disambiguazione prescelto, che assegna ad essi il corretto significato semantico posseduto nel contesto lessicale in cui sono presenti. In questo modo è possibile associare ad ogni documento un insieme, costituito da concetti o synset di WordNet, che rappresenta il suo contenuto semantico.

3.2.5 Disambiguazione con uso della relazione di iperonimia

3.2.5.1 Introduzione

L'algoritmo in questione si pone l'obiettivo di assegnare ad ogni sostantivo il corretto significato semantico che possiede nel contesto lessicale in cui è presente. In questo modo diventa possibile, sfruttando le relazioni semantiche che intercorrono tra i concetti espressi dalla parola da disambiguare e quelli associati alle parole presenti nel suo contesto lessicale, associare ad ogni riassunto un insieme chiamato ***Xsemantico*** che contiene una lista di concetti (*synset numbers*) da contrapporre alla lista delle parole che costituisce una rappresentazione lessicale.

La relazione semantica utilizzata dall'algoritmo è quella di iperonimia. Per disambiguare una parola si determinano le relazioni tra gli *iperonimi* dei suoi sensi, e gli iperonimi dei sensi delle parole presenti nel contesto. La motivazione dietro l'approccio utilizzato nasce dalla constatazione che i concetti di una parola in un ben determinato contesto sono in un certo qual modo relazionati per mezzo della relazione di *iperonimia.*

La seguente figura agevola nella comprensione di quanto detto.

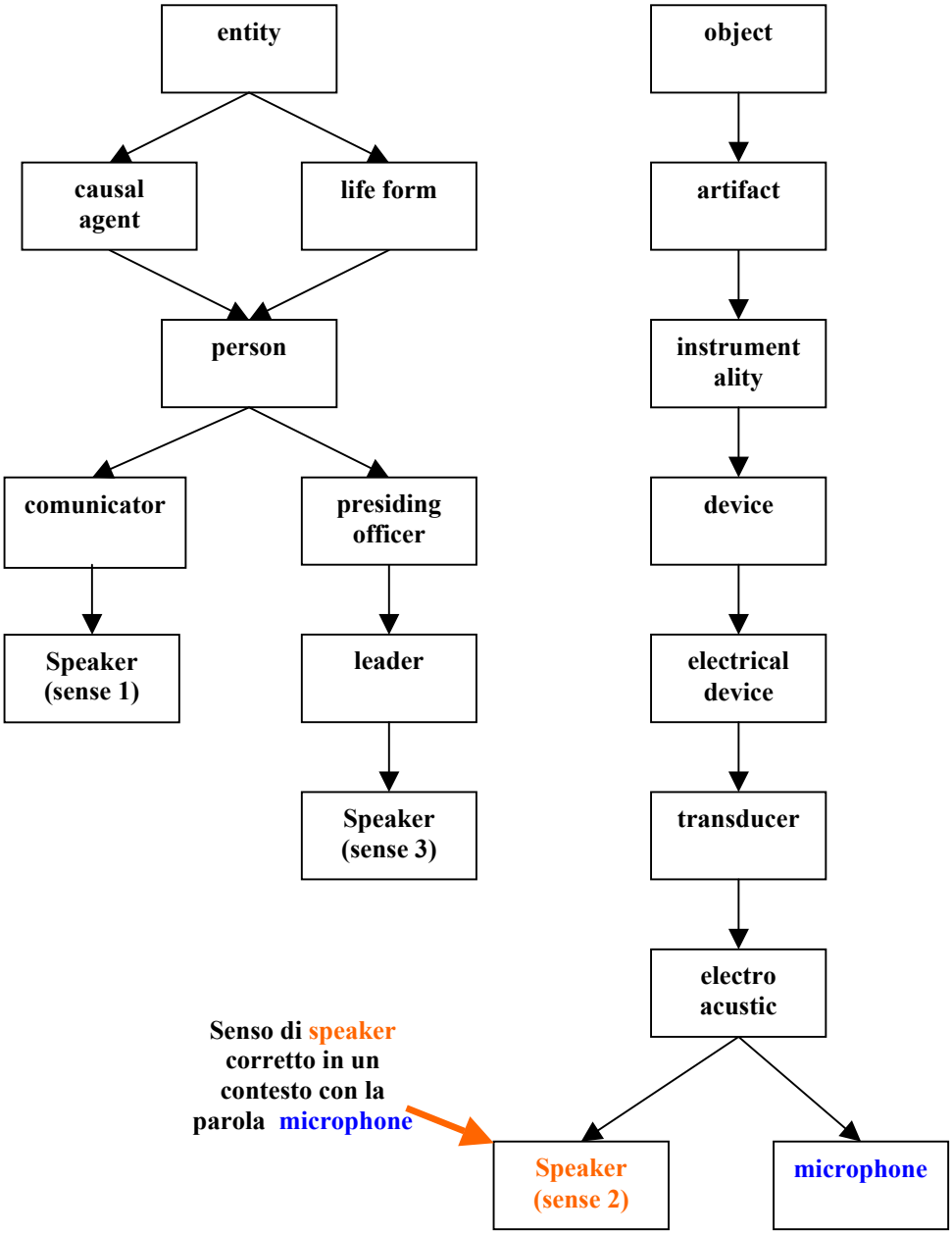

Figura 14: Collocazione dei sensi di speaker nella gerarchia del dizionario semantico WordNet

Come mostrato dalla figura, dei tre sensi della parola *speaker*, gli *iperonimi* di ciascuna parola (*speaker e microphone*) sono solo relazionati nel secondo senso. Se entrambe le parole sono presenti in un testo è logico pensare che questo significato è il più corretto in quel contesto lessicale.

3.2.5.2 Implementazione

Il metodo presentato consiste nell'automatizzare l'assegnamento del corretto significato semantico ai sostantivi presenti all'interno del corpo del riassunto.

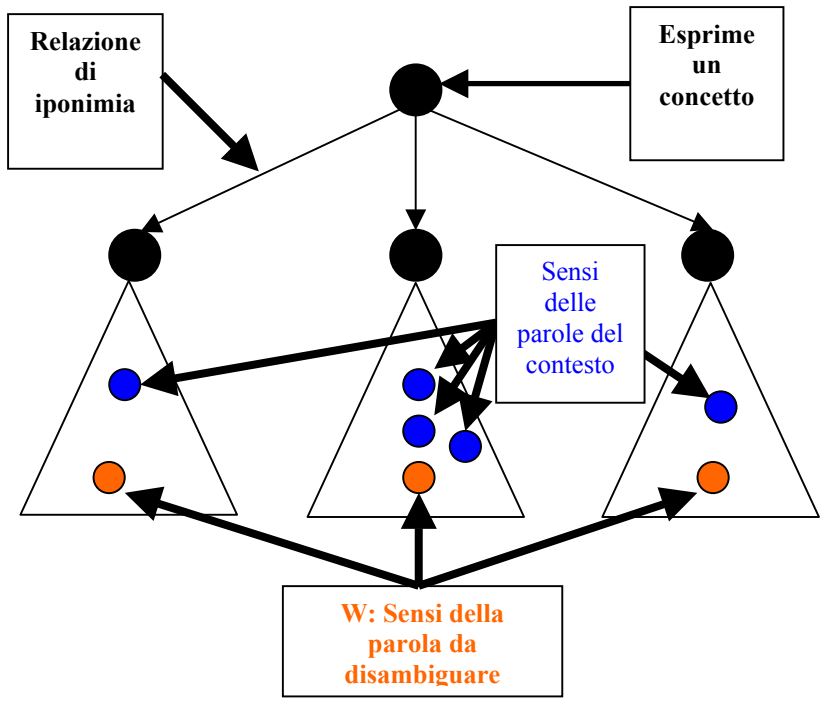

Figura 15: Esempio di struttura contenente i sensi della parola W da disambiguare e quelli associati alle parole del contesto.

E'opportuno, prima di descrivere la soluzione implementa, fornire un esempio esplicativo.

La figura sopra mostra la collocazione dei sensi della parola **W** in una struttura gerarchica di WordNet. I sensi delle parole che formano il contesto da analizzare sono rappresentati tramite cerchi blu. Ciascun senso della parola appartiene ad una ben determinata sottogerarchia. Il senso della parola da disambiguare contenuto nella sottogerarchia con più alto numero di sensi delle parole del contesto viene scelto come suo corretto significato semantico.

Descriviamo in che modo procederebbe l'algoritmo se dovesse disambiguare la parola **W** dell'esempio precedente.

Per ogni senso della parola **W** viene analizzata la gerarchia più in alto nella rete di WordNet che contiene quel senso, ovvero la gerarchia che ha come radice l'iperonimo più in alto. Le prime gerarchie analizzate per ogni senso sono caratterizzate dall'avere la stessa radice che coincide con quella dell'albero. Poiché esse contengono tutte lo stesso

numero di parole, allora, per ogni senso, si considerano le gerarchie ottenute abbassando di un livello. La sottogerarchia alla quale appartiene il senso 2 contiene il maggior numero di sensi delle parole del contesto e di conseguenza viene scelto come senso di disambiguazione della parola **W**.

L'idea che sta dietro l'algoritmo è di prendere, come corretto significato semantico di una parola, quello che è maggiormente relazionato semanticamente con le parole che formano il suo contesto lessicale.

In seguito viene data una descrizione più formale dell'algoritmo proposto:

Passo 1:

Si estraggono i sostantivi da disambiguare dal corpo del riassunto. Questi sostantivi costituiscono il contesto lessicale d'ingresso. Per esempio, Context= {plant, tree, perennial, leaf}.

Passo 2:

Per ciascuna parola

- *Vengono memorizzati in una lista tutti i suoi possibili sensi Si={ Si1, Si2, Si3, Si4,..}.*

Passo 3*:*

Per ciascuna parola.

Per ciascun senso.

- *Si memorizzano in uno stack tutti i suoi iperonimi.*

Passo 4:

Per ciascuna parola.

Per ciascun senso.

Per ciascun iperonimo.

- *Tutti i sensi delle parole del contesto nei quali è presente come iperonimo vengono memorizzati in una lista.*

Passo 5:

Per ciascuna parola

Per ciascun senso

- *Deve essere localizzato il più basso iperonimo presente nello stack associato e il numero dei sensi memorizzati nella lista creata al passo 3 viene assegnato.*

Passo 6:

Per ogni parola si seleziona il senso caratterizzato dal numero più grande.

- *Se c'è un solo senso avente il numero più grande è scelto ragionevolmente.*
- *Se c'è più di un senso si ritorna al passo 4 abbassando di un livello all'interno della gerarchia fino a quando un solo senso è ottenuto.*

3.2.6 Disambiguazione tramite densità concettuale

3.2.6.1 Introduzione

L'obiettivo che si pone l'algoritmo in questione è fornire una soluzione, alternativa a quella descritta precedentemente, per assegnare ai sostantivi il corretto significato semantico che possiedono nel contesto lessicale in cui sono presenti.

L'approccio utilizzato è legato all'uso della tassonomia dei sostantivi di WordNet ed alla nozione di distanza concettuale che si traduce in una formula di densità concettuale sviluppata da Rigau [16].

Il metodo si basa sull'assegnamento di un peso ad ogni nodo della gerarchia di WordNet che rappresenta un iperonimo di uno dei sensi della parola da disambiguare. Il senso della parola che si trova nella sottogerarchia con maggior punteggio viene scelto come corretto significato semantico della parola considerata. L'algoritmo, pur essendo simile a quello descritto precedente, presenta delle caratteristiche uniche. Infatti, l'algoritmo precedente, in una prima fase, cerca di scegliere come corretto significato semantico di una parola quello che è contenuto nella gerarchia avente maggior numero di sensi delle parole del contesto. Si può facilmente verificare che due o più sensi della parola considerata si trovino sotto due gerarchie contenenti entrambe lo stesso numero di sensi delle parole del contesto e ciò porterebbe la parola a rimanere ambigua.

Un esempio esplicativo di quanto detto viene in seguito descritto tramite il seguente diagramma.

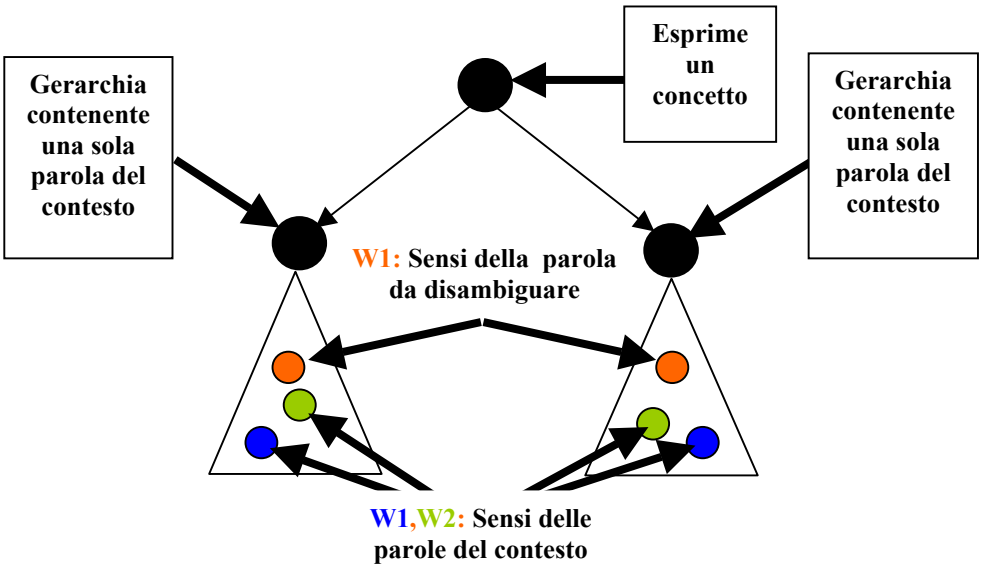

Figura 16: Esempio di due gerarchie contenenti lo stesso numero di sensi delle parole del contesto da disambiguare

Come si evince dalla figura, i due sensi della parola da disambiguare sono contenuti in due gerarchie aventi lo stesso numero di sensi delle parole del contesto. L'algoritmo precedente non riesce a disambiguare la parola considerata.

Al contrario, in questo algoritmo, il peso che viene attribuito ad ogni nodo di una gerarchia, dipende, oltre che dai sensi delle parole del contesto contenuti in essa, anche da fattori come la sua profondità e il numero dei suoi discendenti. L'algoritmo potrebbe scegliere, al fine di disambiguare la parola in questione, uno dei due sensi se, per esempio, le due gerarchie contenessero un numero differente di discendenti.

3.2.6.2 Densità concettuale

La misura del legame tra concetti può essere una valida sorgente di conoscenza per decisioni nel *Processo di Linguaggi Naturali* (*Natural Language Processing*). Per esempio, il legame semantico del senso di una parola con il contesto lessicale in cui è presente permette di selezionare quel senso su tutti gli altri, e di conseguenza individuare il suo corretto significato semantico.

Come detto da Miller e Tebel [17], il legame tra concetti può essere misurato attraverso la loro distanza concettuale in una rete semantica come WordNet. Ciò permette di scoprire la coesione lessicale (*vedi paragrafo 4.3.1*) di un dato insieme di parole.

La distanza concettuale tra due concetti è definita in Rada[18] come lunghezza del cammino più breve che connette i concetti nella rete semantica, che in questo caso è

data dalla gerarchia dei sostantivi di WordNet. Altri hanno cercato (Sussna[19]) di migliorare la definizione di distanza concettuale, e tutti concordano sul fatto che deve essere sensibile a:

- La lunghezza del cammino più breve che connette i concetti che vengono esaminati. Minore è la distanza, misurata come numero minimo di nodi che li divide, più semanticamente i due concetti risultano vicini.
- La profondità della gerarchia: concetti in regioni più profonde della gerarchia sono semanticamente più vicini. Quest'ultima deduzione può essere esplicata meglio tramite un esempio. Supponiamo due coppie di concetti tali che ogni coppia derivi dallo stesso nodo. Consideriamo i concetti *pine* e *abete,* che derivano dal concetto *tree,* e i concetti *living thing* e *non living thing*, che derivano dal concetto *thing, entity*. I concetti *pine* e *abete,* essendo collocati abbastanza profondamente nella gerarchia di WordNet, risultano più vicini semanticamente dell'altra coppia, che anzi possono essere considerati uno opposto all'altro.
- La densità dei concetti nella gerarchia: concetti in regioni più dense della gerarchia sono relativamente più vicini rispetto a quelli in regioni più sparse. Concetti che si trovano in gerarchie caratterizzate da molti discendenti, cioè a bassa densità, sono semanticamente meno vicini rispetto a quelli che si trovano in gerarchie caratterizzate da pochi discendenti.

Partendo dalle caratteristiche che deve possedere la distanza concettuale deriva, come è indicato in Agirre[16], una formula di densità concettuale che ci permette di confrontare le gerarchie di WordNet.

Per illustrare come la densità concettuale può aiutare a disambiguare una parola ci si serve della figura 14 dove la parola *w1* è candidata ad essere disambiguata in un contesto costituito dalle parole *w2* e *w3*. Ciascun senso della parola w1 appartiene ad una sottogerarchia di WordNet.

Ad ogni iperonimo dei due sensi della parola *w1* viene associato un peso (densità concettuale) che dipende, oltre che dal numero di sensi delle parole del contesto presenti tra i suoi discendenti, anche dalla sua profondità e dal numero dei suoi discendenti. In questo caso, i nodi colorati rossi e blu, supposti aventi lo stesso numero di discendenti, contengono lo stesso numero di sensi delle parole del contesto, ma il nodo blu, poiché esprime un concetto più specifico, risulta vincente.

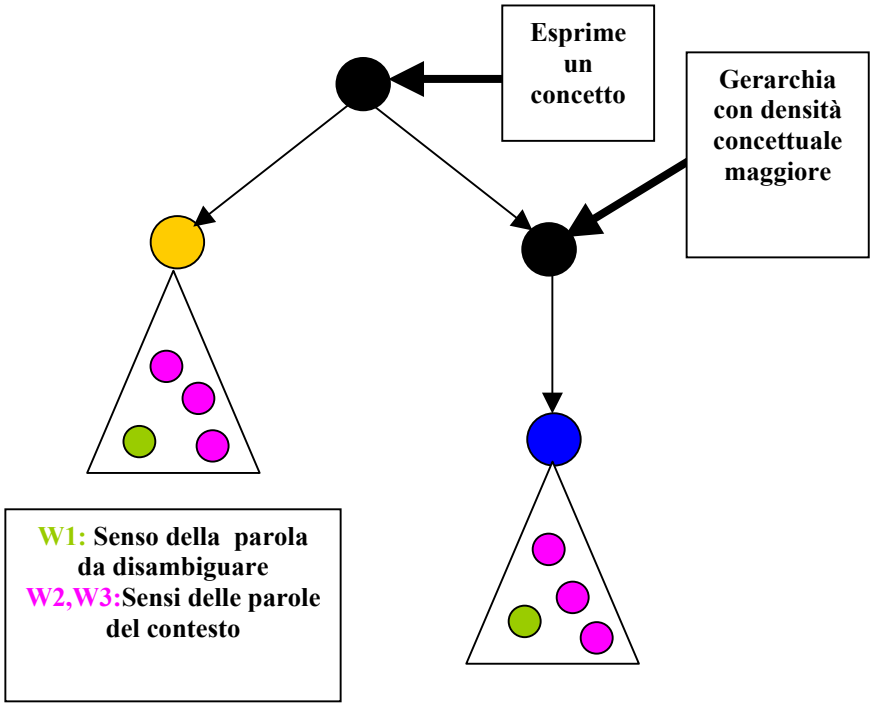

Figura 16: Esempio di due gerarchie contenenti lo stesso numero di sensi delle parole del contesto da disambiguare

La formula esatta che definisce la densità concettuale risulta:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i:0.20}}{descendants_c} \quad \textbf{(9)}$$

Dove:

nhyp: profondità del concetto considerato,

CD(c,m): densità concettuale per *c* quando la sua sottogerarchia contiene un numero di sensi *m* delle parole del contesto lessicale.

descendants: numero di discendenti del concetto *c*.

La formula mostra un parametro che è stato calcolato sperimentalmente. Il valore 0.2 cerca di smorzare l'esponenziale i. Vari valori sono stati provati da Agirre in [16] e quello che ha dato migliori performance è proprio 0.2. Esaminando la formula possiamo fare alcune importanti considerazioni. A parita dei parametri *nhyp* e *discendents*, il metodo dà un maggior punteggio alle gerarchie contenenti il maggior numero di sensi del contesto. Mentre a parità di *m* e *discendents* l'algoritmo privilegia

le gerarchie più specifiche, ed a parità di m e *hypo* l'algoritmo privilegia maggiormente le gerarchie che contengono una densità di concetti del contesto maggiore.

3.2.6.3 Implementazione

Viene data una descrizione più formale dell'algoritmo di disambiguazione appena descritto.

I passi fondamentali che lo costituiscono risultano:

Passo 1:

Si estraggono i sostantivi da disambiguare dal corpo del riassunto. Questi sostantivi costituiscono il contesto lessicale d'ingresso. Per esempio, Context= {plant, tree, ***perennial, leaf}.***

Passo 2:

Per ciascuna parola

- *Vengono memorizzati in una lista tutti i suoi possibili sensi Si={ Si1, Si2, Si3, Si4,..}.*

Passo 3*:*

Per ciascuna parola.

Per ciascun senso.

- *Si memorizzano in uno stack tutti i suoi iperonimi.*

Passo 4:

Per ciascuna parola.

Per ciascun senso.

Per ciascun iperonimo

- *Si calcola e si memorizza la sua densità concettuale*

Passo 5:

Per ciascuna parola.

- *Si seleziona come corretto significato semantico quello che si trova nella sottogerarchia avente una densità concettuale maggiore.*

3.2.7 Miglioramenti agli algoritmi di disambiguazione

Gli algoritmi di disambiguazione descritti fanno l'assunzione che la parola da disambiguare è relazionata con i sensi delle parole del contesto lessicale in cui è presente. Nasce il problema di disambiguare le parole che, pur essendo relazionate semanticamente, risultano in WordNet non connesse.

Le parole *plant* e *leaf*, per esempio, appartengono entrambi al contesto *flora*, ma WordNet include in esso solo *plant*, mentre *leaf* è assegnato al contesto object. Tutto ciò porta all'esistenza di parole che non possono essere disambiguate utilizzando la relazione di iperonimia. Deriva la necessità di apportare dei miglioramenti agli approcci di disambiguazione descritti sopra. Il primo miglioramento è di utilizzare la euristica, in seguito descritta, nel caso in cui una o più parole rimangono ambigue.

Euristica di Definizione

Con questa euristica la *word sense disambiguation* è ottenuta usando la definizione (*gloss* in Word-Net) della parola da disambiguare.

Per assegnare il corretto significato semantico ad una parola viene verificato se la definizione dei suoi sensi contiene parole presenti nel contesto lessicale analizzato. Per ciascuna parola presente viene incrementato un peso che viene associato allo stesso senso. Il senso con il peso maggiore viene scelto come corretto significato semantico della parola.

In seguito viene fornita una descrizione più dettagliata dell'algoritmo.

Per ogni parola rimasta ambigua

Per ogni senso

Weigth=0 *Ciascun senso ha un peso iniziale nullo.*

- *Vengono estratte le parole presenti nella sua definizione.*
- *Se qualche parola del contesto è presente nelle parole estratte al passo precedente:*

$$Weigth = Weigth + 1$$

Il senso con il peso più alto è scelto e viene dato come soluzione.

Se continuano ad esserci più di un senso di una parola ancora ambigui, si ricorre ad un'altra euristica basata sulla relazione di iponimia.

Un senso di una parola può avere presente nella definizione di qualche suo iperonimo una o più parole appartenenti al contesto lessicale analizzato. E' il caso del primo senso della parola *leaf,* che ha come suo diretto iperonimo il concetto *plant organ.* Quindi *leaf* e *plant* possono essere disambiguati sfruttando proprio questa informazione.

In seguito è descritta in modo formale l'euristica.

Euristica di Iperonimia

Per assegnare il corretto significato semantico ad una parola viene verificato se la definizione degli iperonimi dei suoi sensi contiene la presenza di una o più parole del contesto lessicale analizzato. Ogni volta che una parola viene rilevata viene assegnato al senso un peso, in funzione della distanza del cammino esistente nella rete semantica di WordNet con l'iperonimo nella cui definizione la parola è presente.

La distanza semantica tra un senso e il suo iperonimo, calcolata come numero minimo di nodi interposti tra essi, risulta importante perchè, se due sensi contengono nella definizione di un loro iperonimo una parola del contesto lessicale, è intuitivamente più probabile che il senso corretto, tra i due, è quello che ha una distanza minore dall'iperonimo considerato.

Come corretto significato semantico da attribuire ad una parola viene scelto quello a cui è associato il peso più alto.

Una descrizione più formale dell'euristica risulta:

Per ogni parola rimasta ambigua

Per ogni senso

Weigth=0 *Ciascun senso ha un peso iniziale nullo*

Per ogni iperonimo.

Per il resto delle parole appartenenti al contesto.

- *Si verifica se qualche parola è presente nella definizione dell'iperonimo*
- *Se una parola è presente, il peso del senso è incrementato secondo la seguente formula*

$$Weigth = Weigth + Distanza(senso, iperonimo)/profondità(senso) \quad \textbf{(10)}$$

Il senso con il peso più alto è scelto e viene dato come soluzione. E' stata pesata la distanza per la profondità del senso preso in considerazione. In questo modo, a parità di distanza vince il nodo più profondo, che rappresenta un concetto più specifico.

Se si continua ad avere ancora la presenza di parole ambigue, vengono applicate due ultime euristiche che sfruttano le relazioni semantiche di meronimia/olonimia possedute da WordNet.

Per esempio, tra gli olonimi di *car*, intesa nel senso di autovettura, abbiamo *door*, *pedal, window, engine, floor*, etc.., e di conseguenza possiamo sfruttare questo bagaglio di conoscenza per disambiguare la parola *car* in un contesto in cui sono presenti i vocaboli poc'anzi menzionati.

Euristica di Olonimia

Con questa euristica la *word sense disambiguation* è ottenuta utilizzando un approccio basato sulla relazione di *olonimia,* (has-part, tree/branch).

Per assegnare ad una parola il suo corretto significato viene verificato, per ciascuno dei suoi sensi, se la definizione (gloss in WordNet) o i sinonimi dei suoi *olonimi* contengono parole presenti nel suo contesto lessicale. Per ciascuna parola presente viene incrementato un peso associato al senso.

In seguito viene fornita una descrizione più dettagliata dell'algoritmo.

Per ogni parola rimasta ambigua

Per ogni senso

Weigth=0 *Ciascun senso ha un peso iniziale nullo.*

Per ogni olonimo relazionato semanticamente.

- *Vengono estratte le parole presenti nei sinonimi e nella sua definizione.*
- *Se contengono parole presenti nel contesto:*

Weigth= Weigth+1

Il senso con il peso più alto è scelto e viene dato come soluzione. Resta da descrivere l'ultima euristica che si basa sull'uso della relazione di meronimia:

Euristica di Meronimia

Con questa euristica la *word sense disambiguation* è ottenuta utilizzando un approccio basato sulla relazione di *meronimia,(* is a *part ot*).

Per assegnare ad una parola il suo corretto significato, viene controllato, per ciascuno dei suoi sensi, se la definizione (gloss in WordNet) o i sinonimi dei suoi *meronimi* contengono parole presenti nel suo contesto lessicale.

Per ciascuna parola presente viene incrementato un peso associato al senso.

In seguito viene fornita una descrizione più dettagliata dell'algoritmo.

Per ogni parola rimasta ambigua

Per ogni senso

Weigth=0 *Ciascun senso ha un peso iniziale nullo.*

Per ogni meronimo relazionato semanticamente

- *Vengono estratte le parole presenti nei sinonimi e nella sua definizione.*
- *Se contengono parole presenti nel contesto:*

Weigth= Weigth+1

Il senso con il peso più alto è scelto e viene dato come soluzione.

Se rimane qualche parola ambigua non viene considerata a far parte del contenuto semantico, cioè viene considerata irrilevante nella descrizione semantica del riassunto preso in considerazione.

3.3 Rappresentazione lessicale di un documento

3.3.1 Introduzione

A causa dell'incapacità del dizionario semantico WordNet di relazionare semanticamente sostantivi propri {*Bush, president*} e a causa dell'assenza di parole che potrebbero essere significative per descrivere il contenuto informativo di un documento {*perl, UML, lex, yacc*}, è necessaria una rappresentazione lessicale da opporre a quella semantica e quindi un modello matematico, che contenga il contenuto lessicale di un documento, da poter essere successivamente elaborato. Il modello matematico è costituito da un insieme di parole, presenti nel testo, opportunamente filtrate da una fase di estrazione delle caratteristiche, che consente di eliminare le parole con basso potere discriminatorio, ovvero irrilevanti nel contribuire a determinare la tematica tipica del documento.

Le varie fasi che, partendo dai risultati ottenuti dai motori di ricerca, portano alla formazione degli insiemi lessicali associati ai documenti risultano:

1. Memorizzazione dei risultati ottenuti alla query sottoposta a vari motori di ricerca.
2. Estrazione e memorizzazione, dai risultati precedentemente ottenuti, delle informazioni, associate ad un documento, che risultano essere: titolo, url e riassunto del contenuto del documento.
3. Applicazione della fase di estrazione delle caratteristiche all'insieme delle parole presenti nel riassunto di ogni documento.

3.3.2 Estrazione delle caratteristiche

Si è interessati, a differenza del processo di estrazione delle caratteristiche che coinvolge l'aspetto semantico, non solo ai sostantivi, ma anche alle altre categorie sintattiche.

Tra le forme lessicali, presenti nel corpo del riassunto, ci possono essere parole con basso potere discriminatorio che portano ad inevitabili errori nella successiva fase di raggruppamento. Diventa necessario ridurli al minimo e a tal fine verte l'eliminazione delle parole di senso comune (*stop words*) e l'applicazione del metodo di *Zipf*[45], che

permette di eliminare le parole poco rilevanti nel determinare la tematica tipica di un documento.

Le fasi fondamentali che caratterizzano il processo di estrazione delle caratteristiche risultano:

1. Eliminazioni delle parole di senso comune, cosiddette stop words, che non contribuiscono a caratterizzare la tematica tipica del riassunto.
2. Eliminazione delle parole che appartengono alla categoria sintattica rappresentata dagli avverbi.
3. Trasformazione delle forme lessicali nella loro radice comune mediante l'applicazione dell'algoritmo di Porter [43].
4. Applicazione del metodo di Zipf. Vengono eliminate le parole aventi una frequenza di occorrenza nell'intero insieme di documenti maggiore di 0.005 e minore di 0.35.

3.4 Metriche utilizzate

3.4.1 Introduzione

Ogni riassunto ha associato, per rappresentare il suo contenuto, un insieme lessicale, costituito da parole, opportunamente filtrate, presenti nel corpo del riassunto restituito dal motore di ricerca, e un insieme semantico, costituito dai corretti significati semantici che i sostantivi esprimono nel contesto lessicale in cui sono presenti.

Resta da definire in che modo sfruttare questa duplice rappresentazione al fine di confrontare il contenuto informativo esistente tra due documenti. L'approccio che è stato utilizzato è stato quello di definire, come distanza tra due documenti, una combinazione lineare della distanza esistente tra gli insiemi lessicali e semantici.

Le distanze tra gli insiemi sono state calcolate utilizzando la formula di Tanimoto, che viene descritta in seguito.

3.4.2 Distanza di Tanimoto

La distanza di Tanimoto è stata utilizzata per confrontare due insiemi. La formula, in seguito descritta, non dipende dal fatto che gli insiemi considerati siano legati alla rappresentazione lessicale o semantica.

Siano X e Y due insiemi d'ingresso . La distanza di Tanimoto è data dal valore che si ottiene applicando la seguente formula:

$$|X \cap Y| / |X U Y| \qquad \textbf{(11)}$$

Si tratta di una semplice metrica che tiene conto del numero di elementi che due insiemi hanno in comune rapportato al loro numero totale.

Maggiore è la distanza di Tanimoto tra due insiemi più i due insiemi risultano simili. Se la distanza è uguale a 0, gli insiemi considerati non hanno nessun elemento in comune, se è uguale a 1, i due insiemi sono perfettamente identici.

3.4.3 Formula per il calcolo della distanza tra due documenti

La metrica utilizzata, definita come una combinazione lineare, con coefficienti opportunamente variabili, della distanza esistente tra gli insiemi lessicali e semantici, permette di definire un modo per confrontare i contenuti lessicali e semantici dei documenti. In questo modo diventa possibile una visualizzazione, su un piano

cartesiano, caratterizzata dalla presenza di gruppi di contenuto lessicale e semantico pressochè simile.

La metrica utilizzata è formalmente espressa dalla seguente formula:

$$Dist(d1,d2)=\alpha*DistLessicale(d1,d2)+\beta*DistSemantica(d1,d2) \quad \textbf{(12)}$$

Dove le funzioni utilizzate hanno il seguente significato:

- **DistLessicale(d1,d2):** rappresenta la distanza lessicale tra il riassunto *d1* e il riassunto *d2*. E' stata calcolata sottraendo ad 1 il risultato proveniente dall'applicazione della formula di Tanimoto, in cui gli insiemi d'ingresso sono dati dagli insiemi lessicali associati ai due documenti.
- **DistSemantica(d1,d2):** rappresenta la distanza semantica tra il riassunto *d1* e il riassunto *d2*. E' stata calcolata sottraendo ad 1 il risultato proveniente dall'applicazione della formula di Tanimoto, in cui gli insiemi d'ingresso sono dati dagli insiemi semantici associati ai due documenti.

Una scelta opportuna di α e β permette un raggruppamento che tiene conto in modo diverso del contenuto lessicale e semantico del riassunto. Per esempio, se $\alpha \neq 1$ e $\beta=0$, si produce un raggruppamento basato sul contenuto lessicale, mentre, se α e β sono entrambi diversi da zero, sia il vettore delle parole che quello semantico influenzano il raggruppamento dei documenti.

Il rapporto α/β determina se l'informazione lessicale o semantica è più o meno decisiva nel determinare le coordinate dei documenti web sul piano cartesiano.

3.5 Uso dell'algoritmo di Sammon

La scelta dell'algoritmo di *clustering* è il passo finale da affrontare in un lavoro di raggruppamento di documenti. Tra i diversi algoritmi proposti nel paragrafo 1.3.3, visto l'obiettivo del lavoro, è stato scelto l'algoritmo di Sammon [3].

E' stata necessaria una sua modifica poiché riceve in ingresso un insieme di vettori piuttosto che le loro distanze. Il formato dei dati, che rappresentano le distanze semantiche e lessicali esistenti tra i riassunti, è dato da una matrice triangolare superiore, dove ogni elemento contiene la distanza tra i riassunti associati agli indici corrispondenti.

La matrice delle distanze lessicali e semantiche viene data in ingresso all'algoritmo di Sammon [3], che determina su un piano cartesiano le coordinate dei documenti.

La struttura risultante dell'intero sistema viene descritto in modo esauriente dalla figura seguente.

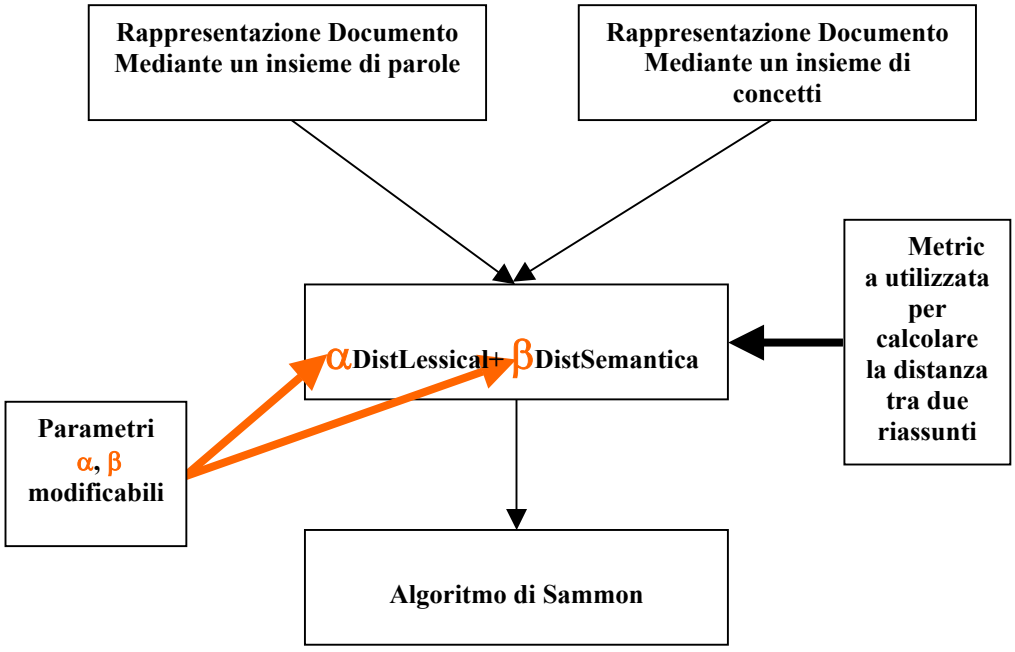

Figura 18: Descrizione grafica del sistema risultante

3.6 Algoritmi di raggruppamento

3.6.1 Approccio con uso dell'algoritmo K-Means

E' necessario, al fine di facilitare la visione dei gruppi visualizzati sul piano cartesiano, attribuire ad ognuno di essi dei colori differenti. L'obiettivo è stato raggiunto utilizzando, in una prima fase, un algoritmo di stima dei gruppi presenti, e successivamente, l'algoritmo k-means[42].

La necessità di utilizzare un algoritmo di stima nasce dalla constatazione che l'approccio utilizzato da *k-means[42]* fa uso di un numero *k* di gruppi predeterminato, e come è intuibile, ciò contrasta con l'impossibilita di prevedere a priori il numero di gruppi visualizzati sul piano. Diventa necessario un algoritmo che stimi il numero di gruppi presenti, calcolando il parametro *k* che deve essere dato come ingresso all'algoritmo *k-means[42]*.

In seguito sono elencati e descritti i passi fondamentali che costituiscono l'algoritmo:

1. Applicazione dell'algoritmo di stima per calcolare il numero k di gruppi.
2. Applicazione dell'algoritmo k-means. Riceve in ingresso il parametro k determinato al passo precedente più le coordinate cartesiane dei punti, ognuno rappresentante un documento, che devono essere assegnati ai k gruppi.

Dal momento che l'algoritmo k-means[42] è stato ampiamente descritto, in seguito ci si sofferma solamente sul primo dei punti elencati.

3.6.1.1 Algoritmo per la stima dei gruppi

E' opportuno, prima di entrare nei dettagli del metodo, dare una definizione di cluster. Per cluster si intende un insieme, costituito da almeno due punti, tale che per ogni punto *p* di esso esista sempre almeno un punto *t*, appartenente sempre allo stesso, che dista da *p* una distanza minore di un certo ε.

In termini più formali, indicato con *S* il cluster considerato, e con p o t un punto appartenente ad esso, si ha che:

$$\surd \text{ p appartenente a S } \exists \text{ t}\neq\text{p appartenente a S : dist(t,p)} < \varepsilon$$
$$\text{cardinalità(S)} >= 2$$

Data la definizione di cluster, è possibile descrivere i passi fondamentali che portano alla stima del parametro k da dare come ingresso all'algoritmo k-means[42].

Per la comprensione dell'algoritmo si ricorda che con *IS* ci si riferisce all'insieme dei punti d' ingresso.

1. Inizializza k ad 1.
2. Associa ad ogni punto l'etichetta NON_ASSEGNATO.
3. Per i:1 to cardinalità(IS) fai:
   - 3.1 Inizializza P al punto i-esimo.
   - 3.2 Inizializza VICINI a 0.
   - 3.3 Se P è NON_ASSEGNATO.
     - 3.3.1 Associa l'etichetta ASSEGNATO a P.
     - 3.3.2 Determina l'insieme J costituito da tutti i punti, etichettati. NON_ASSEGNATO, aventi distanza da P minore di certo ε.
     - 3.3.3 Setta VICINI=VICINI+ cardinalità(J).
     - 3.3.4 Associa l'etichetta ASSEGNATO a tutti i punti J.
     - 3.3.5 Per j:1 to cardinalità(J) fai:
       - 3.3.5.1 Inizializza a P al punto j-simo.
       - 3.3.5.2 Vai al passo 3.3.2.
   - 3.4 Se VICINI>=1
     - 3.4.1 Incrementa k

Si tratta di un algoritmo ricorsivo, che incrementa il parametro *k* ogni volta che vengono determinati dei punti appartenenti ad un insieme che rispetta la definizione di cluster data precedentemente. I punti appartenenti al cluster individuato vengono inizializzati al valore *ASSEGNATO* per evitare di riprenderli in considerazione. Un punto isolato non contribuisce all'incremento del parametro *k*.

Il parametro *k* e le coordinate dei punti che hanno contribuito a stimarlo vengono date in ingresso all'algoritmo *k-means[42],* che associa ogni punto d'ingresso al sistema uno dei k cluster.

Il passo finale, di semplice implementazione, consiste nell'assegnare, ad ogni gruppo individuato, un ben determinato colore.

3.6.2 Approccio con uso dell'algoritmo di Sammon

E' proposto all'utente la possibilità di utilizzare, al fine di facilitare la visione dei gruppi visualizzati sul piano cartesiano, un algoritmo che utilizza un approccio differente rispetto a quello precedente.

L'algoritmo di Sammon[3] ha l'obiettivo di trasferire un set di vettori, definiti in un spazio ad alta dimensionalità, in uno a bassa dimensionalità. La dimensione dello spazio d'uscita, vista la natura del problema, è 2.

L'idea dietro questo nuovo approccio consiste nell'aumentare la dimensionalità dello spazio d'uscita a cinque, in modo tale che, due di queste dimensioni rappresentino le coordinate sul piano cartesiano, le rimanenti, invece, le componenti RGB. A tal fine diventa necessario modificare opportunamente la formula che permette di calcolare la distanza tra due vettori nello spazio d'uscita.

Considerando uno spazio d'uscita di dimensioni 5, la distanza tra un vettore X=[X1, X2, Rx, Gx, Bx] e Y=[*Y1, Y2, Ry, Gy,* By] risulta modificata in questo modo:

$$Dist(X,Y) = \sqrt{(X1-Y1)\text{^}2+(X2-Y2)\text{^}2+(Rx-Ry)\text{^}2+(Gx-Gy)\text{^}2+(Bx-By)\text{^}2}$$

Per dare all'utente la possibilità di dare un peso più o meno maggiore alle componenti RGB sono stati introdotti tre parametri: α, β, e γ.

La formula precedente diventa:

$$Dist(X,Y)=\sqrt{(X1-Y1)\text{^}2+(X2-Y2)\text{^}2+\alpha(Rx-Ry)\text{^}2+\beta(Gx-Gy)\text{^}2+\gamma(Bx-By)\text{^}2}$$

I parametri introdotti possono essere modificati dall'utente, tramite interfaccia grafica, in modo tale da variare, a proprio piacere, il peso delle componenti RGB rispetto alle prime due che rappresentano le coordinate del punto da visualizzare sul piano cartesiano.

Le tre componenti RGB calcolate dall'algoritmo di Sammon[3] assumono valori compresi in un intervallo [0,x]. E' stato necessario una loro espansione nell'intervallo [0,255], dominio di interesse del modello di colore RGB.

4 Risultati sperimentali

4.1 Insieme di Test

Il sistema è valutato con i riassunti dei documenti ottenuti effettuando delle query ai seguenti motori di ricerca: *Google*, *Lycos*, *HotBoot*. E' anche possibile, tramite interfaccia, scegliere un loro qualsiasi sottoinsieme.

Viene effettuato, anche, un confronto nel caso in cui l'input al sistema viene dato da interi documenti piuttosto che da riassunti.

Ogni riassunto ha associato un insieme lessicale contrapposto ad uno semantico. L'insieme lessicale è costituito dalle parole, ridotte a radice comune (*stemming*), presenti nel documento, alle quali sono state eliminate parole di uso comune dette *stop words*. Si è potuto notare che si ottengono risultati migliori eliminando gli avverbi dalle parole d'ingresso.

L'insieme semantico è costituito dai corretti significati semantici che i sostantivi, considerati nella loro forma singolare, possiedono nel contesto lessicale in cui sono presenti.

Ottenuta questa duplice rappresentazione del riassunto, è possibile, utilizzando la metrica descritta nel paragrafo 4.5.3, definire un modo per confrontare i contenuti lessicali e semantici dei documenti. e ottenere una visualizzazione, su un piano cartesiano, caratterizzata dalla presenza di gruppi di contenuto lessicale e semantico pressochè simile.

I parametri α e β, presenti nella metrica utilizzata, sono stati, nelle varie prove, inizializzati ai seguenti valori:

1. $\alpha=1$, $\beta=0$: per verificare i risultati sperimentali derivanti dall'utilizzo di un approccio basato solo su una rappresentazione lessicale.
2. $\alpha=0$, $\beta=1$; per verificare i risultati sperimentali di un approccio basato solo su una rappresentazione semantica. Vengono effettuate due prove, dal momento che sono stati proposti due algoritmi di disambiguazione e quindi due possibili insiemi semantici da poter associare ad un riassunto.

3. $\beta=\alpha=0.5$; per verificare i risultati sperimentali derivanti dall'utilizzo di un approccio basato sulla combinazione della rappresentazione lessicale e semantica.

4.2 Prove sperimentali

I risultati in seguito mostrati e discussi derivano dall'applicazione degli approcci sviluppati dopo che è stata sottoposta ai motori di ricerca la query '*language*'. In questo modo possiamo verificare meglio la capacità di individuazione dei vari gruppi tematici. Infatti, i risultati dati da un motore di ricerca alla query considerata sono costituiti da documenti di svariata natura semantica. Alcuni documenti trattano di linguaggi nella eccezione informatica, altri nell'eccezione letteraria, sociale, etc…. Successivamente l'applicazione viene testata con la query *'web mining'*, sicuramente più specifica.

Il primo risultato, in seguito visualizzato, si riferisce all'applicazione della metrica sviluppata dove α è uguale ad 1 e β a 0. Un approccio di tale tipo utilizza soltanto il contenuto lessicale di un riassunto. E' da mettere in evidenza che il risultato ottenuto è riferito ad una distanza di *Tanimoto* esaltata in modo esponenziale.

La formula utilizzata risulta essere:

$$Distanza(i,j)=exp(\gamma * DistanzaTanimoto(i,j))$$

Dove:

- *DistanzaTanimoto(i,j):* Rappresenta la distanza semantica e lessicale esistente tra due documenti. E' stata calcolata utilizzando la formula descritta nel paragrafo 3.4.3.
- Distanza(i,j): Rappresenta la distanza semantica e lessicale esistente tra due documenti esaltata esponenzialmente. Il suo valore è presente nella matrice delle distanze, che è utilizzata dall'algoritmo di Sammon [3] per determinare le loro coordinate sul piano cartesiano.

Dalle prove sperimentali si è ottenuto che si ottengono risultati migliori ponendo γ uguale a 20.

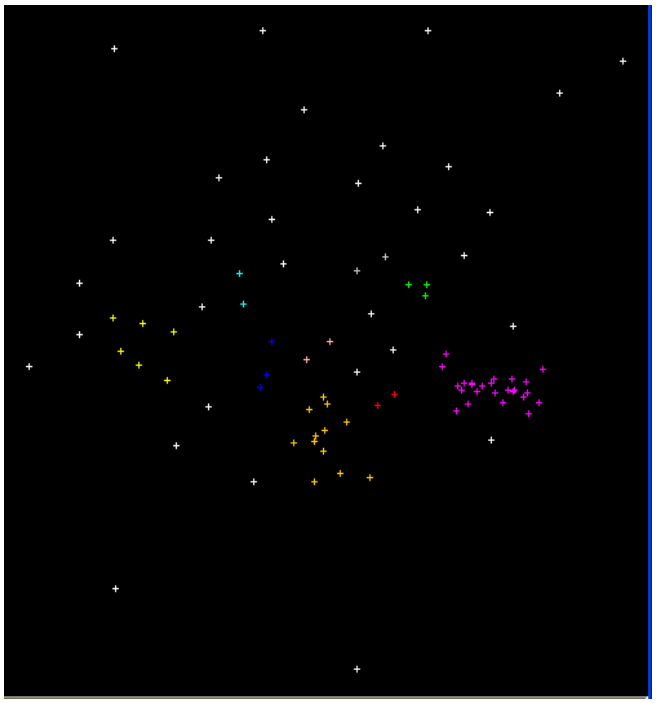

Figura 19: Approccio lessicale

Sono visibili tre gruppi, di colore rispettivamente arancio, viola e giallo. Nel gruppo giallo sono inseriti documenti che trattano di linguaggi di programmazione. I titoli di questi risultano:

Hypertext Preprocessor, HTML Home Page, The Rexx Language, Python Language Website, Perl The Source for Perl development perl, The Source for Java TM Technology, Common Lisp the Language, Contents Prev Next Index Java Language Specification Second, Website META Language WML Title, Perl.

Il gruppo di colore viola contiene documenti che trattano di linguaggi inseribili nel settore delle scienze sociali, ovvero che propongono corsi di apprendimento di linguaggi, traduttori automatici, etc.. I titoli dei documenti, effettivamente di contenuto simile, appartenenti a questo gruppo risultano:

The Modern Language Journal, Language Journal of the LSA, Language Learning Lab, Language Learning Technology, Language International, Yamada Language Center Language Guides, Psychology of Language Page of Links, Language Sites, Language Box Many Cultures One Planet.

Il gruppo arancio scuro si distingue dal celeste perché contiene documenti che hanno a che fare con dei dizionari. Alcuni tra i titoli dei documenti particolarmente simili risultano:

Animated ASL Dictionary, NetLingo Dictionary of Internet Words Glossary of Online, Online Dictionaries and Translators, The American Heritage Dictionary of the English Language Fourth, Foreignword com The Language Site Online dictionaries free, travlang Translating Dictionaries.

Una parte dei documenti rappresentati da punti bianchi tratta di tematiche differenti da quelle associate ai gruppi descritti prima, un'altra, pur trattando di argomenti tali da essere inseriti in quei gruppi, è formata da documenti che non sono riusciti ad avere una distanza tale da rientrarci..

4.3 Conclusioni

L'obiettivo del lavoro proposto è stato di creare un sistema che organizza in maniera "intelligente" i risultati ottenuti da vari motori di ricerca, al fine di mostrare, su un piano cartesiano, raggruppamenti semanticamente omogenei. In questa rappresentazione la distanza euclidea tra due documenti visualizzati sul piano è proporzionale alla differenza esistente tra i loro contenuti informativi. Per rappresentare il contenuto lessicale e semantico di un documento web è stato associato ad ogni documento un insieme lessicale, costituito da parole, opportunamente filtrate, presenti nel corpo del riassunto restituito dal motore di ricerca, e un insieme semantico, costituito dai corretti significati semantici che i sostantivi esprimono nel contesto lessicale in cui sono presenti.

Dai dati sperimentali preliminari ottenuti si evince che non si può fare a meno della rappresentazione lessicale. Le ragioni che portano il metodo basato sull'analisi semantica a dare risultati meno attraenti sono da ricercare, oltre all'assenza di alcune parole nel dizionario semantico WordNet, nei risultati dati dall'algoritmo di disambiguazione, che può non portare a termine correttamente la sua attività per l'assenza di legami semantici tra parole, in realtà relazionate.

In futuro sempre più motori di ricerca cercheranno di offrire ai loro utenti servizi e funzionalità che possano essere di valido aiuto per la ricerca di documenti web. L'applicazione progettata, anche se ancora non soddisfa interamente le esigenze e i bisogni dell'utente, si pone, assieme ad altri lavori fatti in tale settore, come tappa fondamentale per l'estrazione di conoscenza e risorse da quel mondo variegato, eterogeneo e non strutturato che è il web.

Bibliografia

[1] G.Pilato, F.Sorbello and G.Vassallo: *"An Innovative Way to Measure the Quality of a Neural Network without the Use of the Test Set"* - IJACI International Journal of Advanced Computational Intelligence - Vol. 5 No 1, 2001, pp:31-36

[2] A.Cirasa, G.Pilato, F.Sorbello and G.Vassallo: *"EaNet: A Neural Solution for Web Pages Classification"* - Proc. of 4th World MultiConference on Systemics, Cybernetics and Informatics - SCI'2000 - 23-26 July 2000, Orlando - Florida U.S.A.

[3] G.Pilato, F.Sorbello and G.Vassallo: "Ordering Web Pages through the Use of the Sammon Formula and the CGRD Algorithm" - Proc. of AICA Congress, 27-30 Oct, 2000 Taormina (ME) -Italy - pp.495-503

[4] F. Fukumoto, Y. Suzuki, Event Tracing based on Domain Dependency."*An Investigation of the preconditions for effective data fusion in IR: A pilot study*," In the Proceedings of the 61th Annual Meeting of the American Society for Information Science 1998.

[5] E. Fox, G. Nunn, W. Lee," *Coefficients for combining concept classes in a collection"*. In the proceedings of the 11 th ACM SIGIR Conference, pp. 291-308, 1988.

(ICML), pp. 1167-1182, 2000.

[6]Katzer, M. McGill, J. Tessier, W. Frakes, P. DasGupta,"*A study of the overlap among document representations*,".Information Technology: Research and Development, 1(4):261-274, 1982.

[7] S. Mukherjea, J. Foley, and S. Hudson. "*Interactive clustering for navigating in hypermedia systems*". In Proc. of European Conference on Hypertext Technology (ECHT'94), pages 136– 145. ACM, Sep. 1994.

[8]R. Weiss, B. V´elez, M. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. Gifford. Hypursuit*:"A hierarchical network search engine that exploits content - link hypertext clustering*." In Proc. of the 7th ACM Conference on Hypertext'96, pages *180–193. ACM, Mar. 1996.*

[9] V. V. Raghavan and J. S. Deogun. *"User-oriented Document Clustering"*. In Proc. of ACM conference on Research and development in information retrieval (*SIGIR'86)*, pages 157–163. ACM, Sep. 1986.

[10] R. Barzilay, M. Elhadad.”*Using Lexical Chains for Text Summarization*,”In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS’97), ACL, Madrid, 1997.

[11] George A.Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katerine Miller: *Introduction to WordNet: An On-line Lexical Database.*

[12] Nicola Stokes, Joe Carthy: “*First Story Detection using a Composite Document Rapresentation”*- Department of Computer Science, University College Dublin, Ireland.

[13] A Montoyo,M Palomar: ”*Word Sense Disambiguation with Specification Marks in Unrestricted Texts*”-Department of Software and Computing systems, university of Alicant, Alicante, Spain

[14] D Moldovan, R Mihalcea: “ *Using WordNet and Lexical Operators to Improbe Internet Searches”*- Proc. 5th Message Under-standing Conf., Morgan Kaufmann, San Francisco, 1997, pp. 305-320.

[15] E. Agirre, G Rigau*: “Word Sense Disambiguation Using Conceptual Density”.*

[16] Rigau G. and Agirre E: “*Disambiguating bilingual nominal entries against WordNet*”-Seventh European Summer School in Logic, Language and Information, ESSLLI'95,Barcelona, August 1995.

[17] 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. Arlington, Virginia USA. 1993.

[18] Voorhees E: ”*Using WordNet to Disambiguate Word Senses for Text Retriva”.*In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA, June 1993.

[19]Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B: “*Providing Machine Tractablle Dictionary Tools”.I*n Semantics and the Lexicon (Pustejovsky J. ed.), 341-401, 1993.

[20] Yarowsky: “D. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora”.* In Proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.

[21] Yarowsky D:”*One sense per Collocation*”. In proceedings of ARPA Workshop on Human Language Technology, 266-271, Plainsboro, New Jersey, March, 1993.

[22] Miller G, Teibel D: *“A proposal for Lexical Disambiguation”*

[23]Miller G. and Teibel D: "*A proposal for Lexical Disambiguation".* In Proceedings of DARPA Speech and Natural Language Workshop, 395-399, Pacific Grave, California. February, 1991

[24] Miller G. Leacock C, Randee T. and Bunker R. "*A Semantic Concordance". I*n proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey, March, 1993.

[25] Miller G., Chodorow M., Landes S., Leacock C. and Thomas R: "*Using a Semantic Concordance for sense Identification".* In proceedings of ARPA Workshop on Human Language Technology, 232-235, 1994.

[26] Rada R., Mili H., Bicknell E. and Blettner M: "*Development an Applicationof a Metric on Semantic Nets".*In IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1,17-30. 1989.

[27] Resnik P. "*Disambiguating Noun Groupings with Respect to WordNet Senses".* In Proceedings of the Third Workshop on Very Large Corpora, MIT, 1995.

[28] Ribas F:"*On learning more Appropriate Selectional Restrictions*". In proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, 112- 118, Belfield, Dublin, Ireland. 1995.

[29J. Morris, G. Hirst "*Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Tex",* Computational Linguistics 17(1), March 1991.

[30]M. Halliday, R. Hasan:"*Cohesion in English".* Longman:1976

[31]V. Hatzivassiloglou, L. Gravano, A. Maganti:"*An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering*". In the proceedings of the 23 rd ACM SIGIR Conference, Athens, pp. 224-231, 2000.

[32] S. J. Green: "*Automatically Generating Hypertext By Comparing Semantic Similarity*". University of Toronto, Technical Report number 366, October 1997.

[33] J Jiang, d Conrath: "*Multi Word Complex Concept Retrieval via lexical Semantic Similarity"*

[34] M. A. Stairmand, W. J. Black: "*Conceptual and Contextual Indexing using WordNet-derived Lexical Chains*".In the Proceedings of BCS IRSG Colloquium, pp. 47-65, 1997.

[35] M. Okumura, T. Honda:"*Word sense disambiguation and textsegmentation based on lexical cohesion*". In Proceedings of the Fifteen Conference on Computational Linguistics (COLING-94), volume 2, pp. 755-761, 1994.

[36] Huang Yuan, Wang Jicheng, Wn Gangshan:"*Web Mining: Knowledge Discovery on the web"*

[37] Natalija Vlajic, Howard C. Card:"*An adaptive Neural Network Approach to Hypertext Clustering".* State key Laboratory for Novel Software tecnology. Department of Computer Science and Tecnology, Naujing University.

[38] Michael Dittembach. Dieter Merkl, Andreas Rauber: "*The Growing Hierarchical Sel-Organizing-map".* Institut fur Softwaretecnick. Tecnische Universitat Wien

[39] Dieter Merkl, Andreas Ruber: "*Automatic Labelig of Self-Organizing Maps for Information Retrieval"*

[40] Savio L.y Sam, Dik Lun Lee:"*Feature Reduction for Neural Network Based Text Categorization".* Department of Computer Science Hong Kong University of Science and Technology Clear Water Bay, Hong Kong

[41] Natalija Vlajic, Howard C. Card: "*Categorizing web pages using modified Art".* Internet Innovation Centre. Departmental of Electrical and Computer Engineering university of Manitoba Winnipeg, Manitoba, Canada

[42] J.J Rocchio: "Docu*ment Retrieval System-Optimization and Evaluation".* Ph.d Thesis, Harward University

[43] Y Yang :" *Expert Network: Effective and efficient learning from human decisons in Text categorization and Retrieval*". In procedings of the 17th Annual International ACM SIGIR Conference on Reasearch and Development in information Retrieval (SIGIR 94).

[44] Tom Michaell: "*Machine Learning"* . McGraw-Hill., 1996

[45] M. F. Porter *:"An algorithm for suffix stripping*." Program, 14(3):130–137, 1980.

[46] H. Uchida, M. Zhu, Senta T. Della**.** *UNL:" A Gift for aMillennium"* The United Nations University

[47] Tuve Kohonen: "*The Self-Organizing Maps"*- Springer Verlag, berlin 1995

[48]Tuve Kohonen: "*Self-Organization of very large document collections: State of Arts".* In proc Conference Artificial Neural Networks 1998

[49]D. D. Lewis: "*Evaluating text categorization"*. In Proceedings of the Speech and Natural Language Workshop, 1991.

[50]G. Salton and C. Buckley*:" Term-weighting approaches in automatic text retrieval"*. Information Processing and 24(5):513–523, 1988.

[51]T. Honkela, S. Kaski, K. Lagus, and T. Kohonen.."*WEBSOM-Self-Organizing Maps of Document Collections".* In Proc. of the Workshop on Self-Organizing Maps (WSOM'97), Jun.1997.

[52]K. Tajima, Y. Mizuuchi, M. Kitagawa, and K. Tanaka. *"Cut as a Querying Unit for WWW, Netnews, and E-mail*". In Proc. Of the 9th ACM Conference on Hypertext and Hypermedia, pages 235–244. ACM, Jun. 1998.

[53] A. Duda, and D. Gifford. Hypursuit. "A hierarchical network search engine that exploits content - link hypertext clustering."

In *Proc. of the 7th ACM Conference on Hypertext'96*, pages 180–193. ACM, Mar. 1996.

[54] G. Salton, J. Allan, and C. Buckley. "Automatic structuring and retrieval of large text files". Communicasions of the ACM, 37(2):97–108, Feb. 1994.