Facoltà di Ingegneria – Sede di Modena

Corso di Laurea in Ingegneria Informatica – *Nuovo Ordinamento*

TUCUXI
Un agente intelligente per la ricerca di sorgenti informative in Internet

Testo e codice sorgente disponibili presso: `http://dbgroup.unimo.it`

Relatore:
Prof. Sonia Bergamaschi

Candidato:
Daniele Gozzi

Anno Accademico 2003-2004

I^3 = Intelligent Integration of Information

Obiettivo di un framework di accesso e integrazione dell'informazione:

Fornire un sistema integrato di accesso a più sorgenti informative, tra loro eterogenee per organizzazione, modalità di accesso e contenuti.

Introduzione di una componente semantica nella memorizzazione di dati.

Classificazione

In linea di principio:

Le sorgenti di dati dovrebbero contenere informazioni riguardo alla natura del proprio contenuto.
(Paradigma del **Web Semantico**)

In realtà:

Informazioni di questo tipo non sono quasi mai presenti, indipendentemente dal tipo di sorgente.

In assenza di meta-informazioni, volendo integrare una sorgente di dati è necessario introdurre una classificazione basata sui soli dati.

L'obiettivo di un agente hunter è la ricerca di nuove sorgenti di dati da incorporare nel sistema di integrazione.

Nel caso specifico di TUCUXI, la ricerca ha per oggetto delle pagine web.

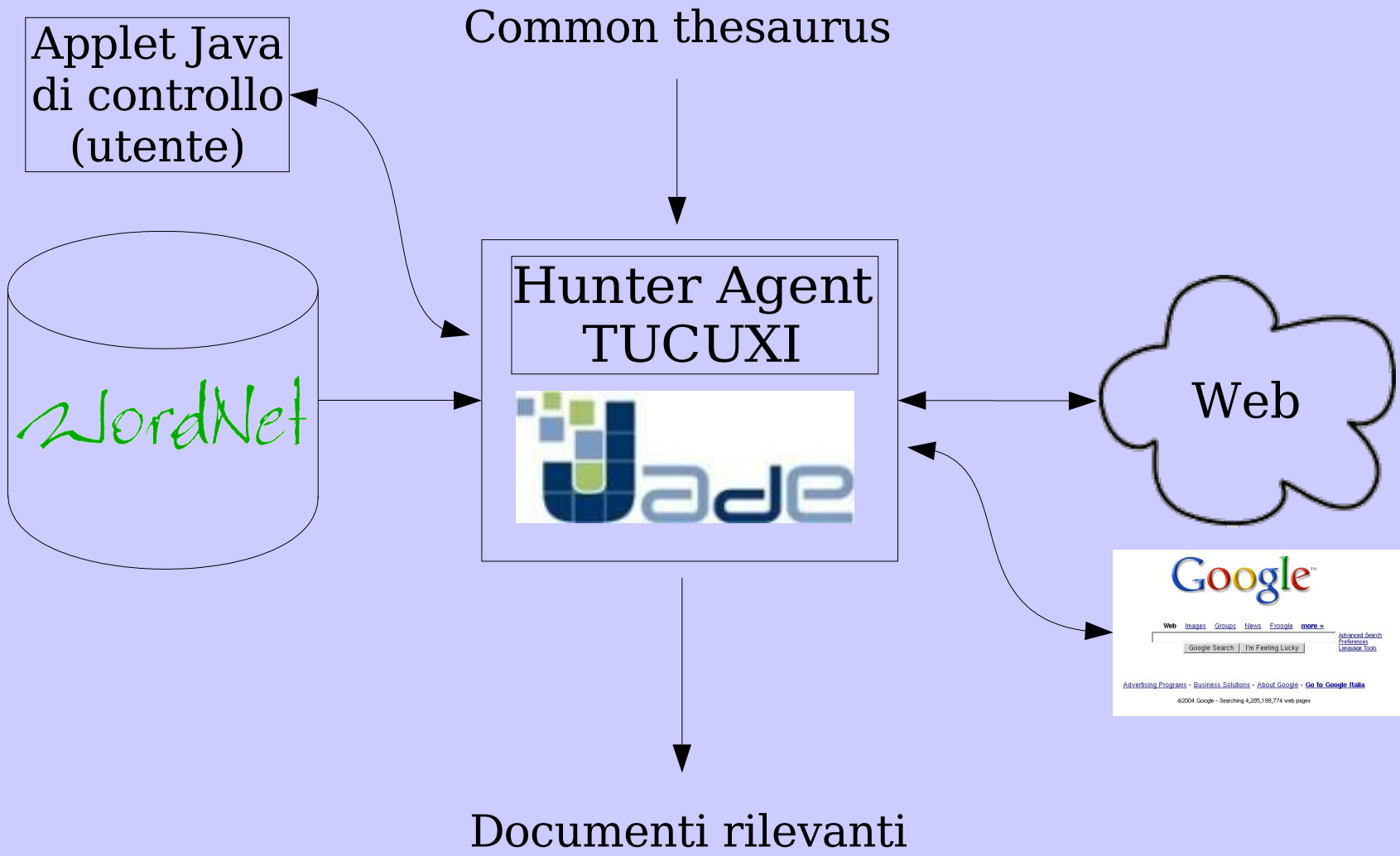
Algoritmi basati su quanto descritto in:

R. Benassi, S. Bergamaschi, M. Vincini, "TUCUXI: the intelligent hunter agent for concept understanding and lexical chaining", 2004

Nuova implementazione:

- 4603 righe di codice sorgente Java
- 22 classi
- Due distinti moduli (agente software e applet di controllo)

Organizzazione delle componenti



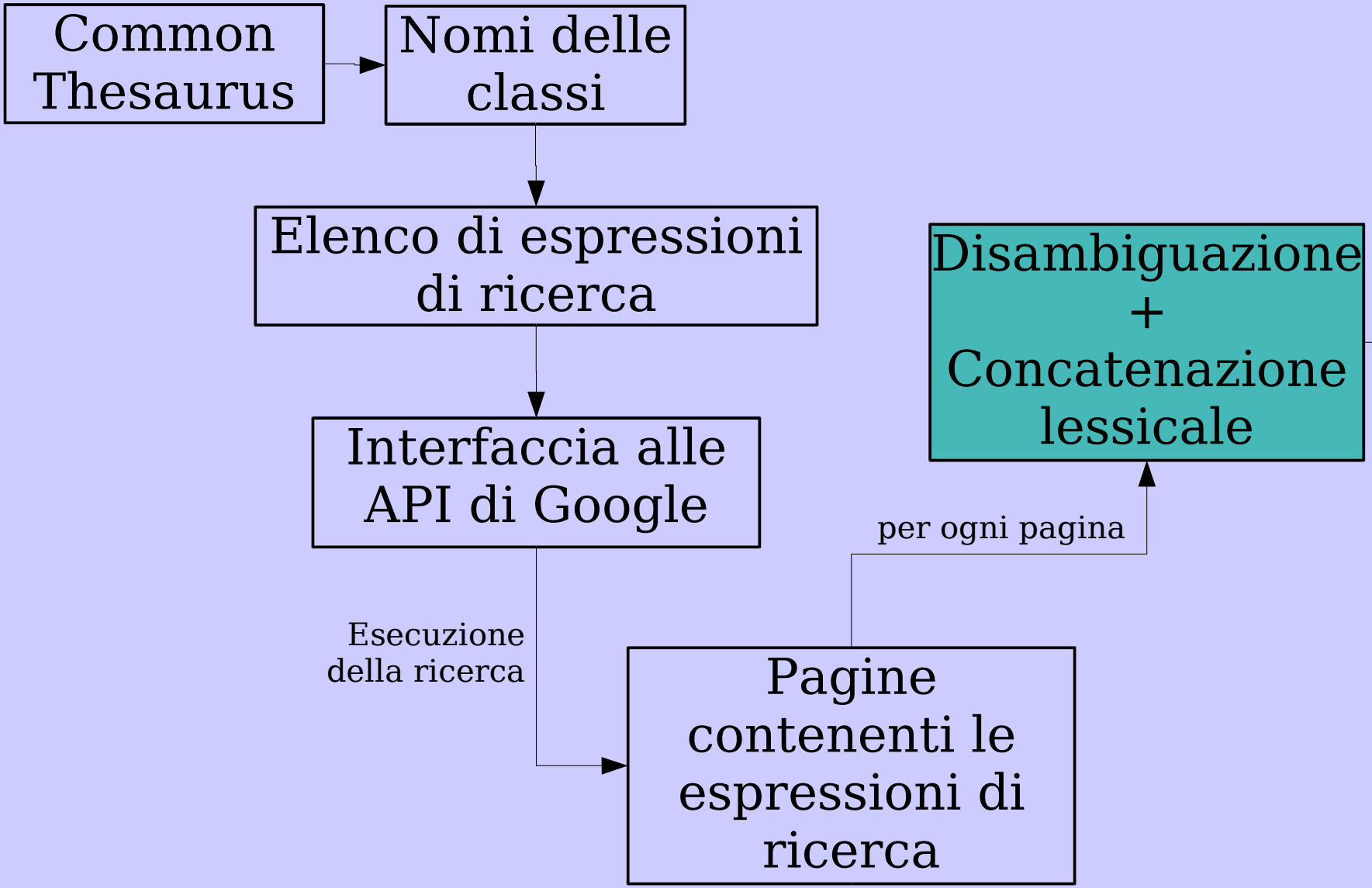
Common Thesaurus

- Ha il ruolo di ontologia condivisa tra sorgente di dati e sistema di integrazione.
- È costituito da un insieme di relazioni tra classi e attributi che descrivono rapporti esistenti tra diversi schemi
- Viene distribuito sotto forma di documento XML

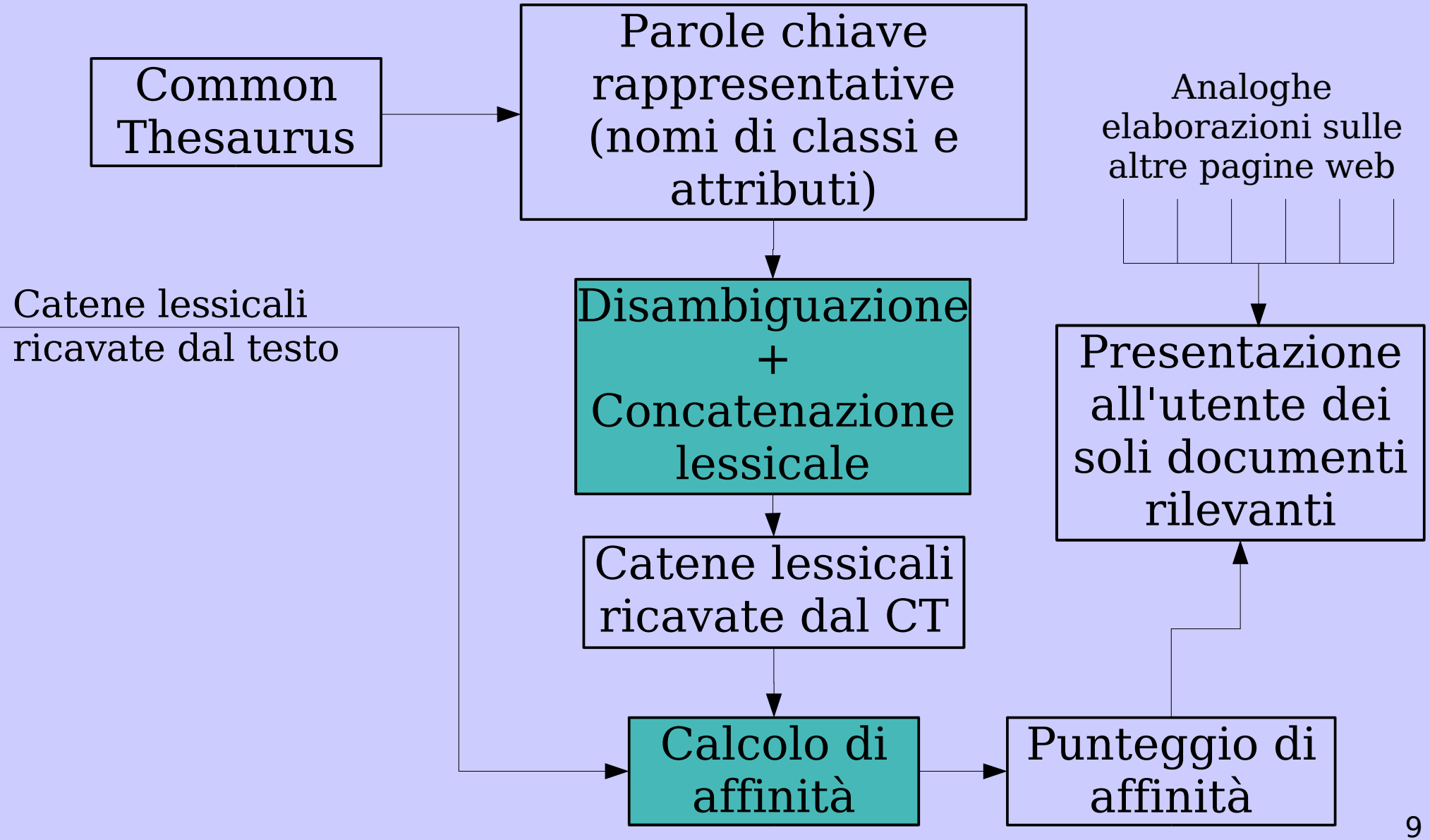
Funzioni implementate

- Estrazione dal Common Thesaurus di alcuni insiemi di parole chiave imprescindibili nel contenuto delle pagine cercate.
- Esecuzione di una ricerca *letterale* nel Web per ciascun insieme individuato, con metodiche tradizionali.
- Analisi dei risultati parziali con algoritmi di analisi del linguaggio naturale che determinano l'affinità nei confronti del Common Thesaurus.
- Eliminazione dai risultati delle pagine Web scarsamente significative.
- Presentazione di un insieme di risultati semanticamente rilevanti.

Estrazione di catene lessicali da ciascuna pagina individuata con metodiche tradizionali



Analisi semantica dei risultati



Estrazione delle catene lessicali

È un algoritmo di clustering, che viene applicato ai soli nomi presenti nel testo.

Prevede due fasi:

- Disambiguazione dei lemmi (Algoritmo #1)

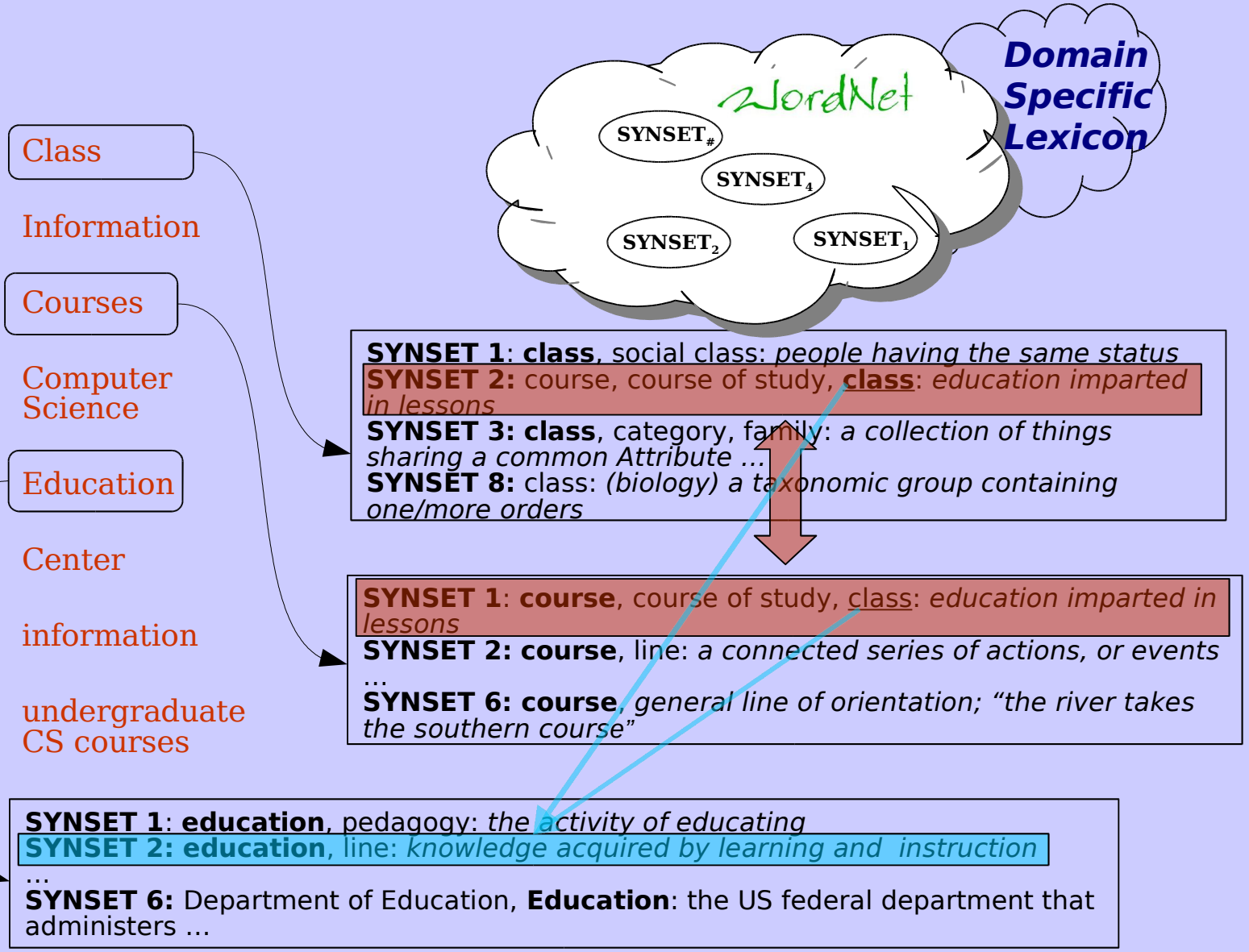
  (Individuazione del significato di ciascuna parola)

- Costruzione delle catene lessicali (Algoritmo #2)

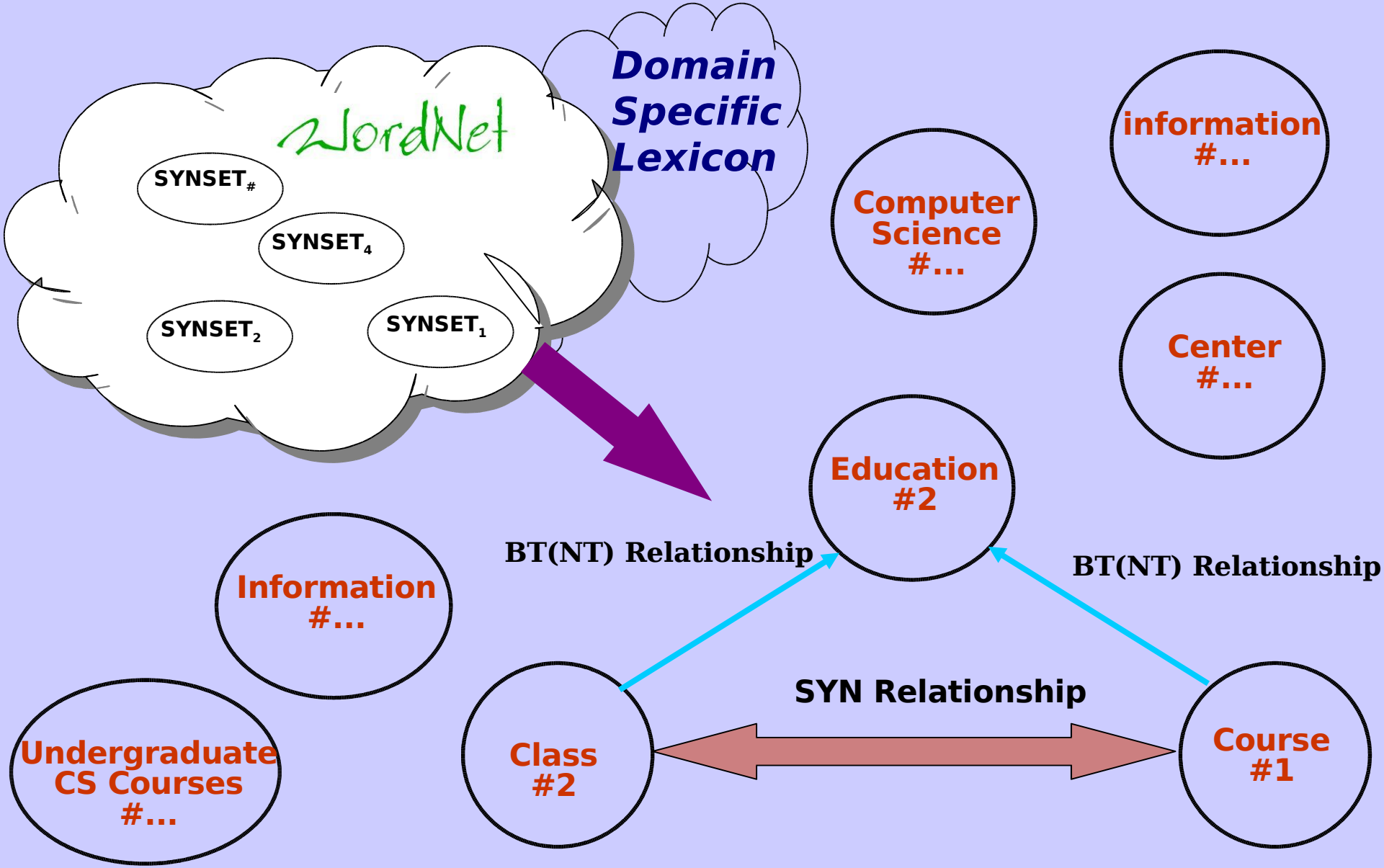
  (Raggruppamento dei lemmi secondo criteri di affinità semantica)

Esempio di applicazione dell'algoritmo di disambiguazione lessicale

"Class Information and **Courses**. The **Computer Science Education Center** has **information** on **undergraduate CS courses"**



Esempio di applicazione dell'algoritmo di concatenazione lessicale



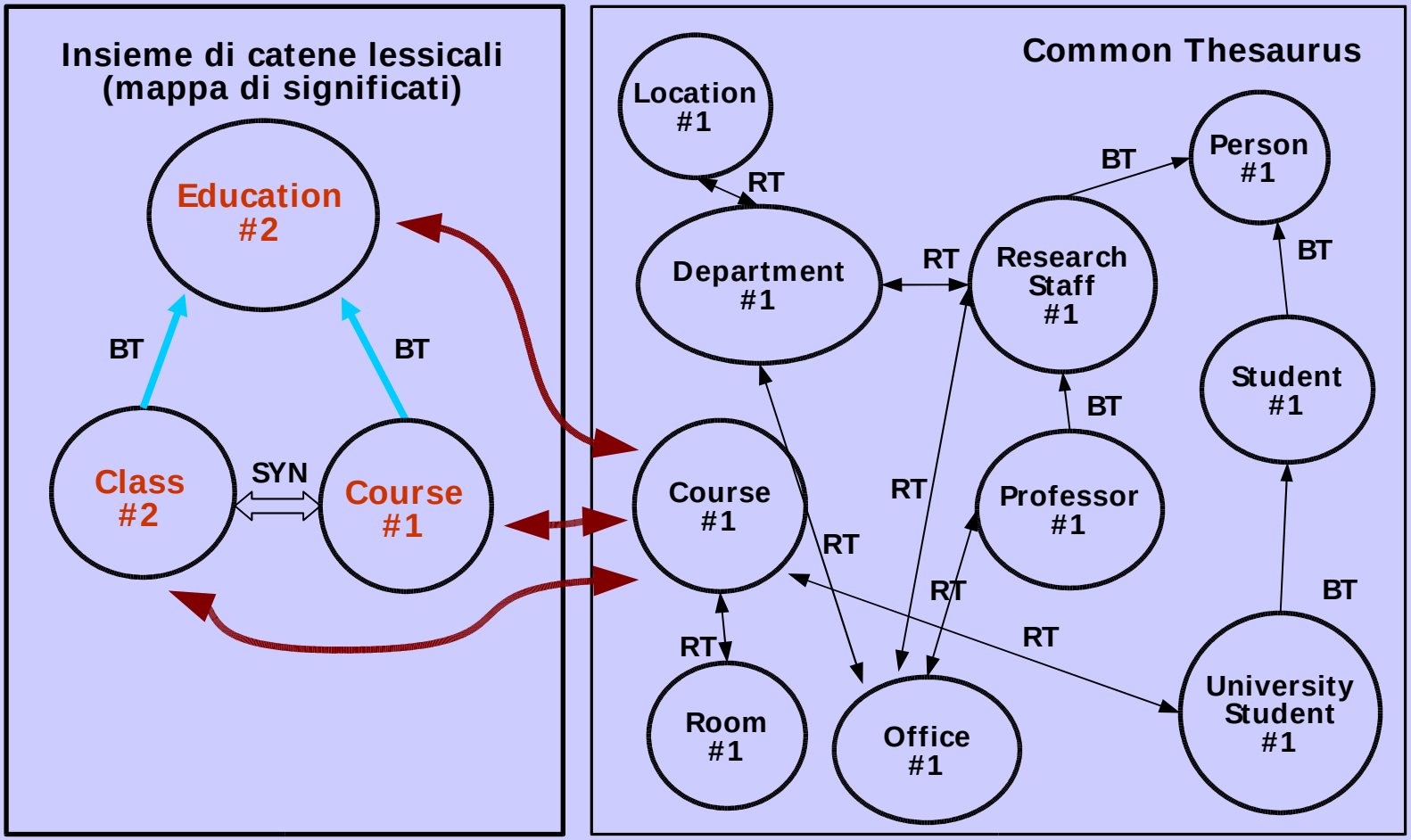
Calcolo dell'affinità lessicale

Valutazione dell'intensità delle relazioni esistenti tra due insiemi di catene lessicali, basata sulle strutture dati create contestualmente alla concatenazione lessicale.

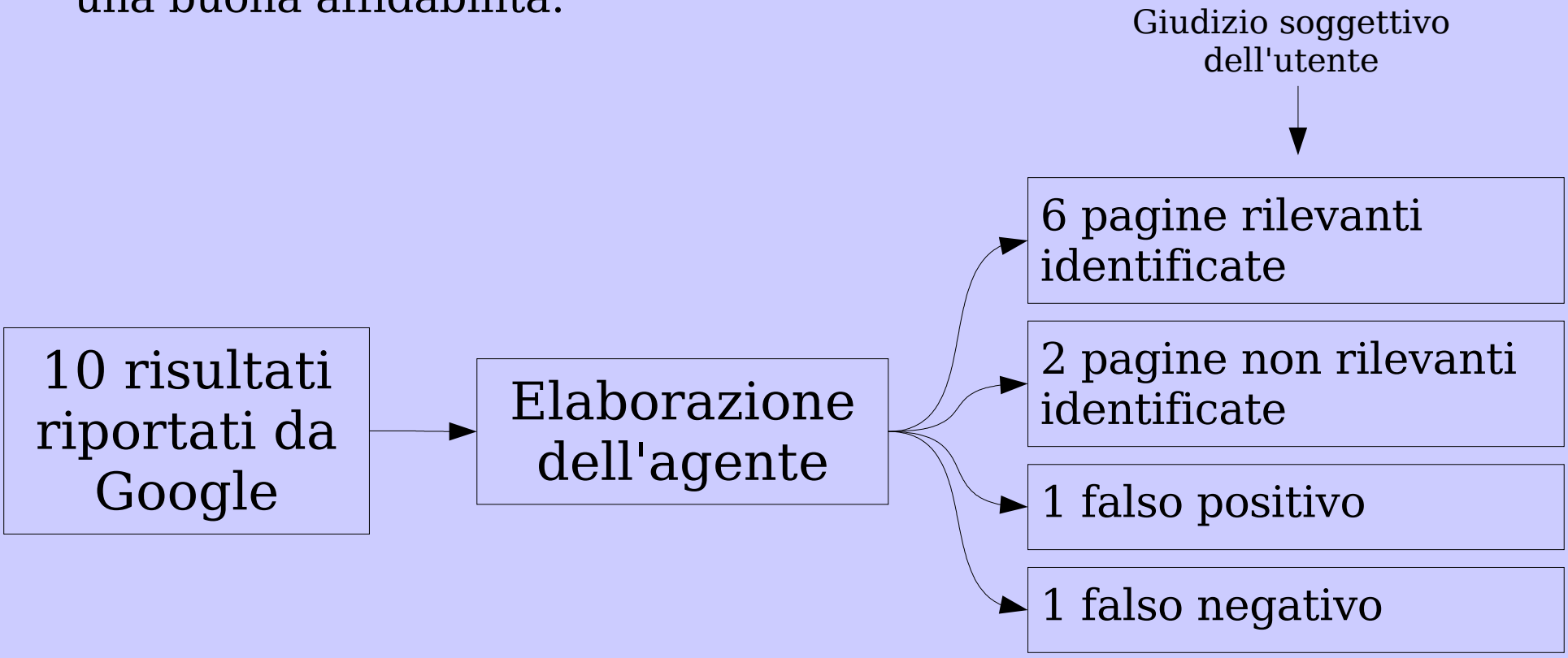
Relazione	***Tipi interessati***	***Simmetria***
Antonym	nomi, verbi, aggettivi, avverbi	√
Hypernym	nomi, verbi	√
Hyponym	nomi, verbi	√
Member Meronym	nomi	√
Substance Meronym	nomi	√
Part Meronym	nomi	√
Member Holonym	nomi	√
Substance Holonym	nomi	√
Part Holonym	nomi	√
Attribute	nomi, aggettivi	√
Entailment	verbi	X
Cause	verbi	X
Also see	verbi, aggettivi	X
Verb Group	verbi	X
Similar to	aggettivi	√
Participle of verb	aggettivi	X
Pertainym	riguarda nomi, aggettivi	X

- WordNet contiene tabelle che descrivono le relazioni lessicali tra i termini della lingua inglese.
- A ciascun tipo di relazione viene fatto corrispondere un punteggio.
- Si verifica il numero di relazioni esistenti tra i termini di due insiemi di catene lessicali.

Esempio di calcolo dell'affinità lessicale



Alcuni test eseguiti sull'agente TUCUXI hanno dimostrato una buona affidabilità:



In gran parte dei casi, una interpretazione soggettiva dell'utente (indipendente dalle analisi compiute) ha corrisposto al giudizio dell'agente circa il grado di interesse di un documento.