Corso di

Web Mining & Retrieval

Operazioni sulle Interrogazioni

Relevance Feedback & *Query Expansion*

(a.a. 2010-2011)

Roberto Basili

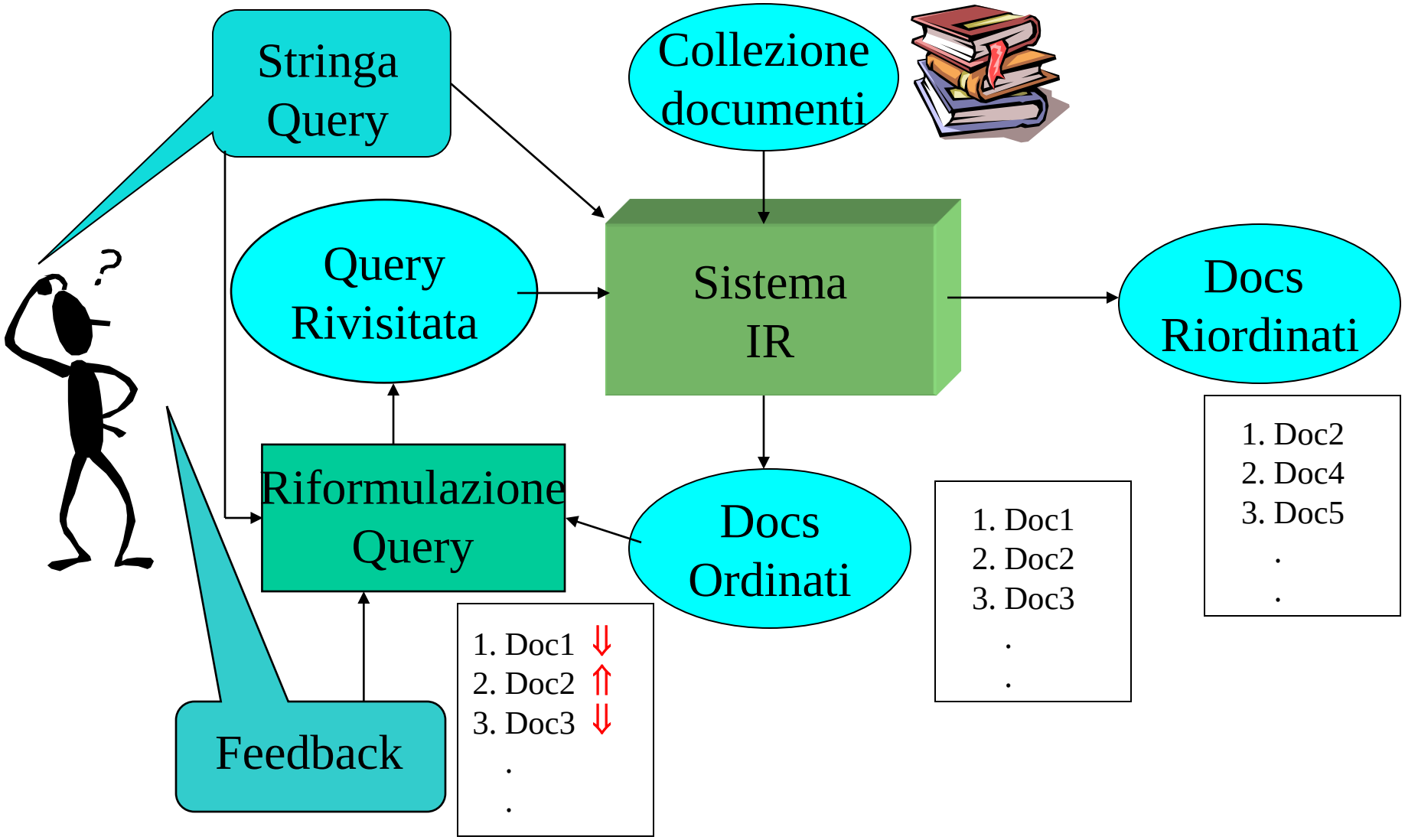
Outline

- *Operazioni sulle interrogazioni*
- Relevance Feedback:
  - User Relevance Feedback
  - Pseudo Relevance Feedback
- Query Expansion
  - Basata su Tesauri
  - Automatizzata

Relevance Feedback

- Dopo che i risultati iniziali vengono presentati consente all'utente di inserire la sua valutazione riguardo alla rilevanza di uno o piu' documenti ritrovati.
- Questa valutazione preliminare viene utilizzata per riformulare la interrogazione.
- Questa tecnica produce un processo interattivo, possibilmente a piu' passi.

Relevance Feedback: Architettura



Riformulazione della Interrogazione

- Rivisita la *query* per tenere conto del *feedback*:
  - Espansione della *Query*:Aggiungi nuovi termini alla query tratti dai documenti rilevanti.
  - Ripesatura dei termini: Aumenta il peso dei termini dei documenti rilevanti (segnale) e diminuisci il peso dei termini dei docs irrilevanti (rumore).
- Sono numerosi gli algoritmi per la riformulazione della interrogazione.

Riformulazione della *Query* per il *VSR*

- Modifica il vettore della query usando l'algebra dei vettori.
- **Aggiungi i vettori dei documenti rilevanti al vettore della** *query*.
- **Sottrai i vettori dei documenti irrilevanti dal vettore della** *query*.
- Effetto:
  - Aggiungi termini nuovi alla *query* con pesi positivi e negativi
  - Ripesa i termini inziali della *query*.

Query Ottima

- Nel caso (ideale) in cui tutti i documenti rilevanti, C_r, siano noti.
- Allora la migliore interrogazione che fornisce tutti e soli i documenti rilevanti in alto nell'ordinamento e':

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j$$

con N pari al numero totale dei documenti.

Metodo: Standard Rocchio

- Poiche' l'insieme dei documenti rilevanti e' sconosciuto, allora si approssima tale insieme usando gli insiemi **noti** rilevanti (D_r) e irrilevanti (D_n) ed includendo anche la *query* iniziale *q*.

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

α: Peso parametrico per la query iniziale.
β: Peso parametrico per i documenti relevanti.
γ: Peso parametrico per i documenti irrilevanti.

Metodo *Ide Regular*

- Se il feedback accresce il grado di riformulazione allora e' il caso di non normalizzarne il modello rispetto alla sua intensità:

$$\vec{q}_m = \alpha \vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

α: Peso parametrico per la query iniziale.

β: Peso parametrico per i documenti relevanti.

γ: Peso parametrico per i documenti irrilevanti.

Metodo: *Ide "Dec Hi"*

- Polarizzazione verso il solo documento irrilevante a massimo score secondo il primo ordinamento:

$$\vec{q}_m = \alpha\vec{q} + \beta \sum_{\forall d_j \in D_r} \vec{d}_j - \gamma \max_{non-relevant}(\vec{d}_j)$$

α: Peso parametrico per la query iniziale.
β: Peso parametrico per i documenti relevanti.
γ: Peso parametrico per i documenti irrilevanti.

Confronto tra i metodi

- Tutti i metodi di IR generalmente migliorano la performance (*recall* e *precision*) grazie allo *user feedback*.
- I risultati sperimentali mostrano che nessun metodo di *user feedback* specifico e' superiore agli altri in ogni scenario applicativo.
- I valori piu' utilizzati per le costanti parametriche sono il valore $\alpha=1$ o $\alpha=\beta=\gamma=1$.
- A volte $\gamma=0$ ha dato buoni risultati (feedback positivo).

La valutazione delle prestazioni

- Per costruzione nei metodi di *relevance feedback* (*RF*) la *query* riformulata dipende dai docs esplicitamente definiti rilevanti ed irrilevanti dall'utente (i.e. peso piu' alto o piu' basso rispettivamente)
- Poiché il loro comportamento **dopo** il *feedback* e' **noto**, i metodi di RF non dovrebbero quindi essere misurati rispetto ai miglioramenti relativi a tali documenti
- In *machine learning*, questo e' l'errore di *verificare la prestazione sugli stessi dati di addestramento*.
- Al meglio, questo produce un test di consistenza del metodo di apprendimento ma mai una sua proprietà universale (*bias* eccessivo rispetto ai dati di addestramento)
- La valutazione quindi dovrebbe riguardare esclusivamente i documenti non segnalati dall'utente.

Valutazione Corretta dei metodi *RF*

- Eliminare dalla collezione tutti i documenti per i quali e' stato fornito *feedback* dall'utente
- Misurare prestazioni come *recall/precision* sulla *collezione residua di documenti*.
- Rispetto al corpus completo di testi, i valori di *recall/precision* possono diminuire poiche' molti documenti rilevanti sono stati rimossi.
- In ogni caso, le misure **relative** alla collezione residua sono piu' affidabili riguardo alle prestazioni del *relevance feedback*.

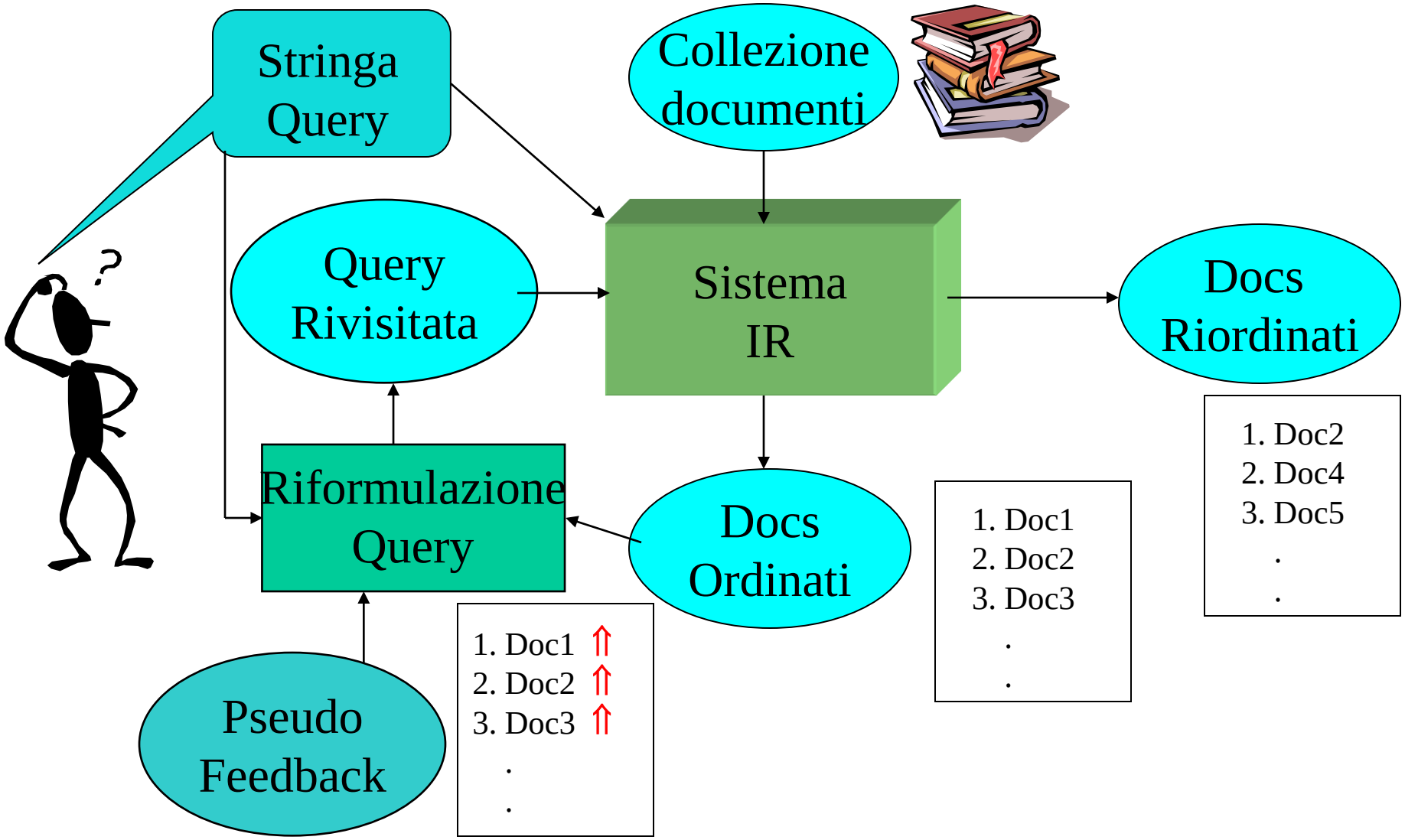
Problemi del *feedback* diretto

- Gli utenti non forniscono facilmente un feedback esplicito.
- La riformulazione puo' produrre interrogazioni lunghe inadatte a molti scenari
  - piu' calcolo durante il *retrieval*,
  - impatto su *Web search engines* dove l'elevato parallelismo delle *query* impedisce la applicazioni di metodi pesanti.
- In presenza di *feedback* e' difficile giustificare il perché un documento e' stato trovato.

Pseudo Feedback

- Simula i metodi relevance feedback senza accedere esplicitamente all'input utente.
- Basta **assumere** che i primi *m* documenti trovati siano rilevanti, ed usare esclusivamente questi per riformulare la interrogazione.
- Supporta una espansione della query che include termini correlati (ma anche diversi) da quelli presenti nella query.

Relevance Feedback: Architettura



PseudoFeedback: Results

- Migliora le prestazioni (valutazione in TREC nei task di *ad-hoc retrieval*).
- Se i primi documenti sono garantiti da criteri logici (come nel caso di ricerca booleana) allora l'impatto delle tecniche di *pseudo feedback* e' persino migliore.

Thesaurus

- Un tesauro fornisce informazioni su sinonimi e parole semanticamente correlate
- Example:

```
medico
  syn: ||dottore, doc, doctor, professore
  rel: medicina, professionista,
                                chirurgo,
```

Espansione basata sul *thesaurus*

- Per ogni termine, *t*, in una *query q*, espandi *q* con i sinonimi e la parole correlate di *t* dal *thesaurus*.
- E' possibile pesare i termini aggiunti meno di quelli originali in *q*.
- In genere aumenta la capacità di *matching* e quindi migliora la *recall* …
- Anche se purtroppo puo' ridurre significativamente la *precision*, a causa dei termini ambigui:
  - "imposta comunale" → "imposta comunale porta infissi edilizia"

Tesauri: esempi

- **MeSH (Medical Subject Heading)**
  - MeSH è il vocabolario controllato usato dalla U.S. National Library of Medicine.
  - Esso supporta la indicizzazione di articoli per la biblioteca virtuale MEDLINE/PubMed.
  - La terminologia di MeSH fornisce metodi sistematici per recuperare le informazioni da tali sorgenti che usano terminologie diverse per gli stessi concetti medici
  - Il linguaggio controllato del thesaurus rappresenta il punto di incontro tra l'indicizzatore e l'utente che interroga la base di dati.
  - I MeSH di diverse lingue sono stati resi disponibili e sono revisionati annualmente

MeSH

National Library of Medicine - Medical Subject Headings

2009 MeSH

MeSH Descriptor Data

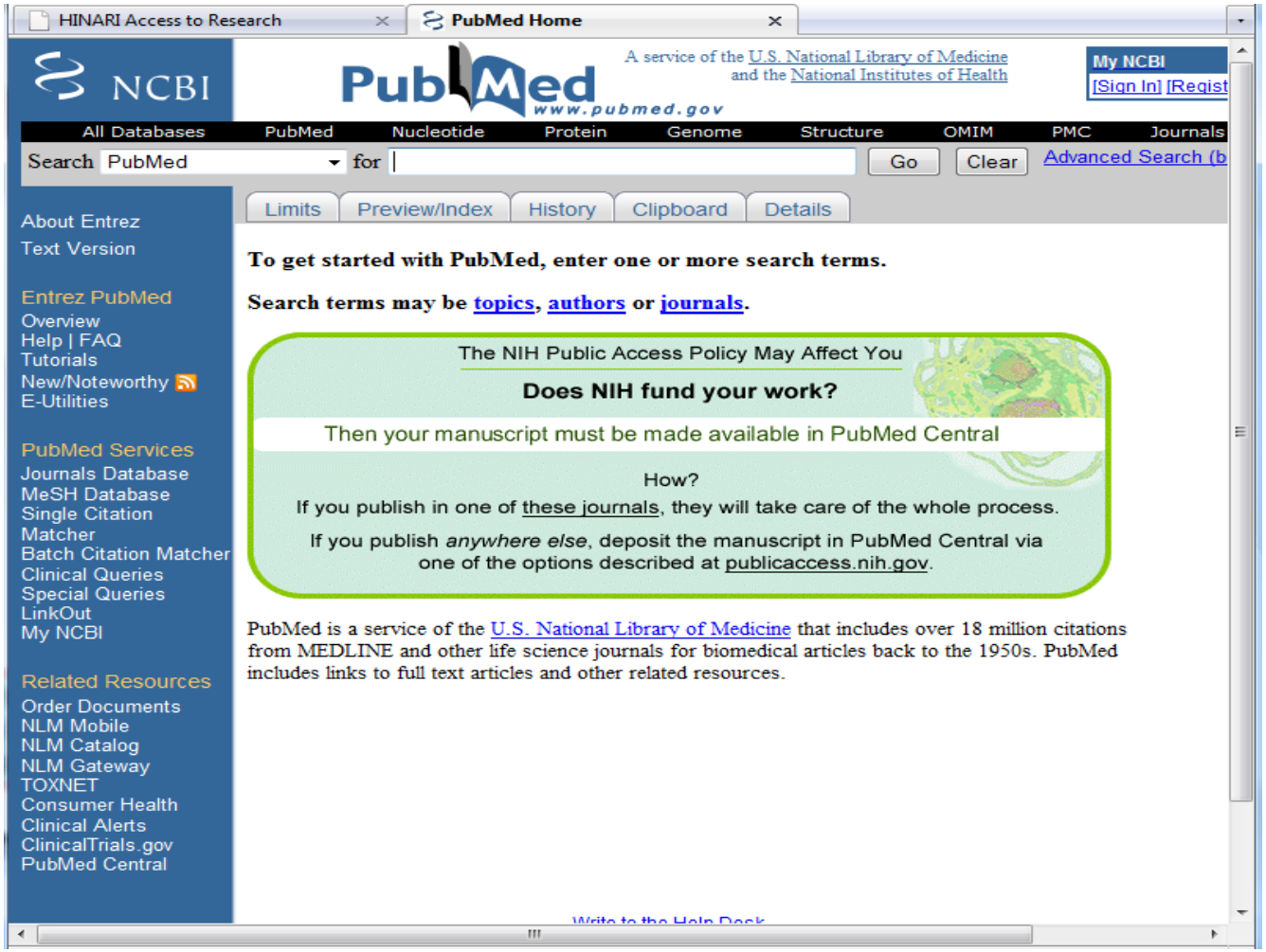
Return to Entry Page

Concept View. Go to Standard View

Expanded Concept View. Go to Standard Concept View

MeSH Heading	Bronchitis		
Tree Number	C08.127.446		
Tree Number	C08.381.495.146		
Tree Number	C08.730.099		
Annotation	tuberculous bronchitis: index under TUBERCULOSIS, PULMONARY & not also under BRONCHITIS unless particularly discussed; BRONCHIOLITIS is also available		
Concept 1 (Preferred)	**Bronchitis**		
	Concept UI	M0002972	
	Scope Note	Inflammation of the large airways in the lung including any part of the BRONCHI, from the PRIMARY BRONCHI to the TERTIARY BRONCHI.	
	Semantic Type	T047 (Disease or Syndrome)	
	Term (Preferred)	Bronchitis	
		Term UI	T005668
		Date	01-JAN-1999
		Lexical Tag	NON
		Thesaurus	NLM (1966)
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI		
Date of Entry	19990101		
Unique ID	D001991		

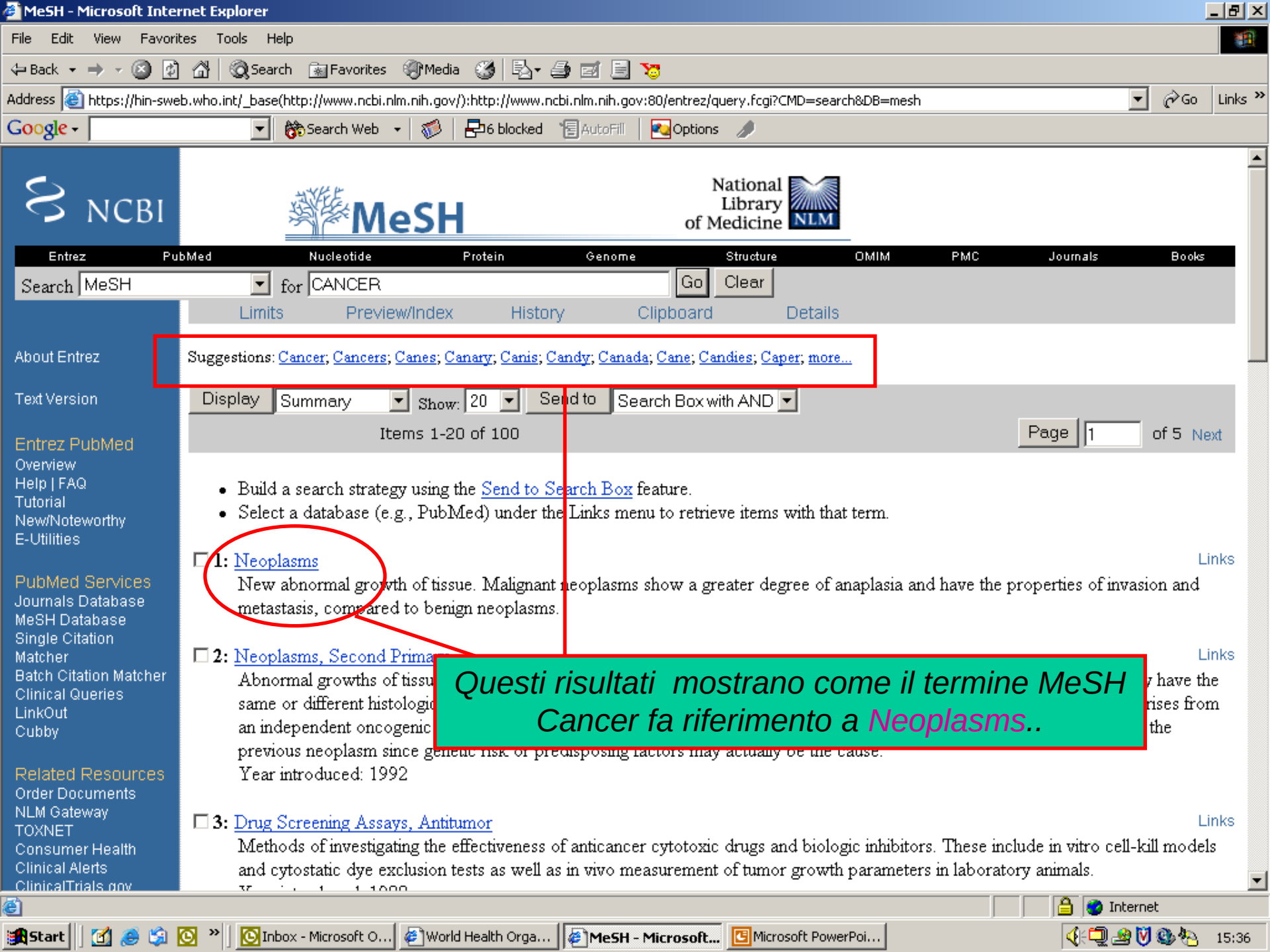
MeSH

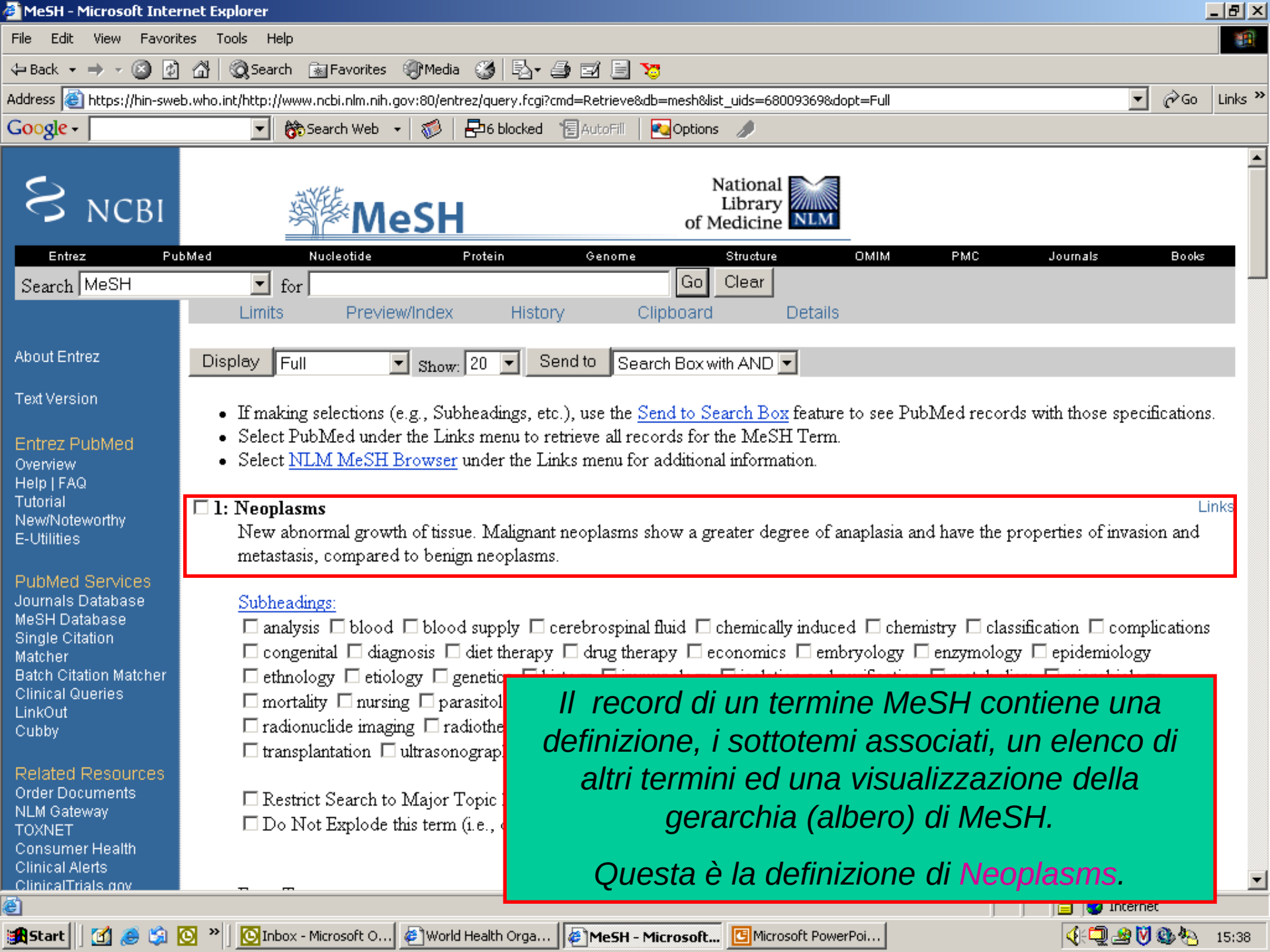

HINARI Access to Research
PubMed Home
NCBI
PubMed
www.pubmed.gov
A service of the U.S. National Library of Medicine and the National Institutes of Health
My NCBI
[Sign In] [Regist
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals
Search PubMed for
Go
Clear
Advanced Search (b
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI
Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
To get started with PubMed, enter one or more search terms.
Search terms may be topics, authors or journals.
The NIH Public Access Policy May Affect You
Does NIH fund your work?
Then your manuscript must be made available in PubMed Central
How?
If you publish in one of these journals, they will take care of the whole process.
If you publish anywhere else, deposit the manuscript in PubMed Central via one of the options described at publicaccess.nih.gov.
PubMed is a service of the U.S. National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

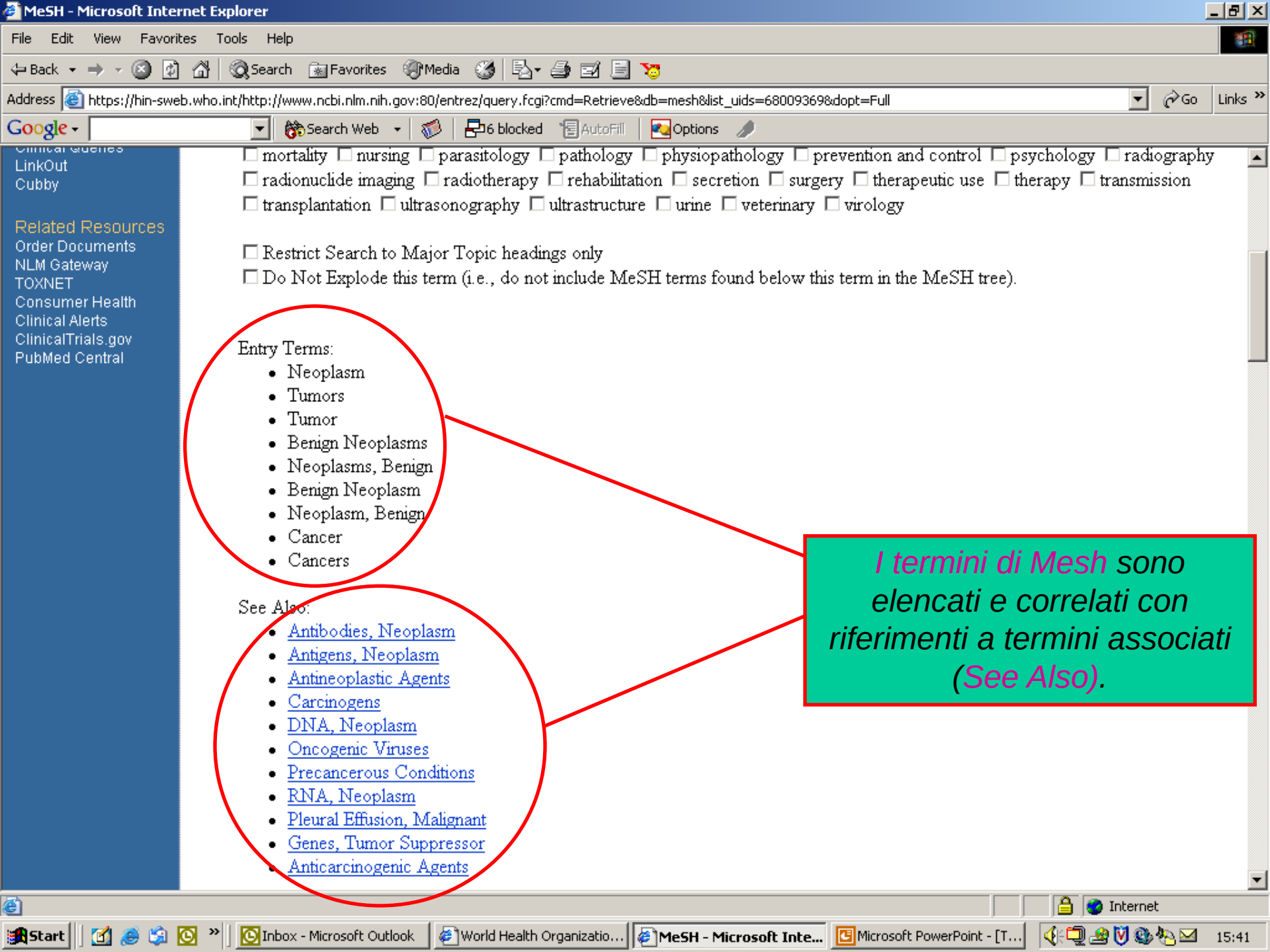
MeSH

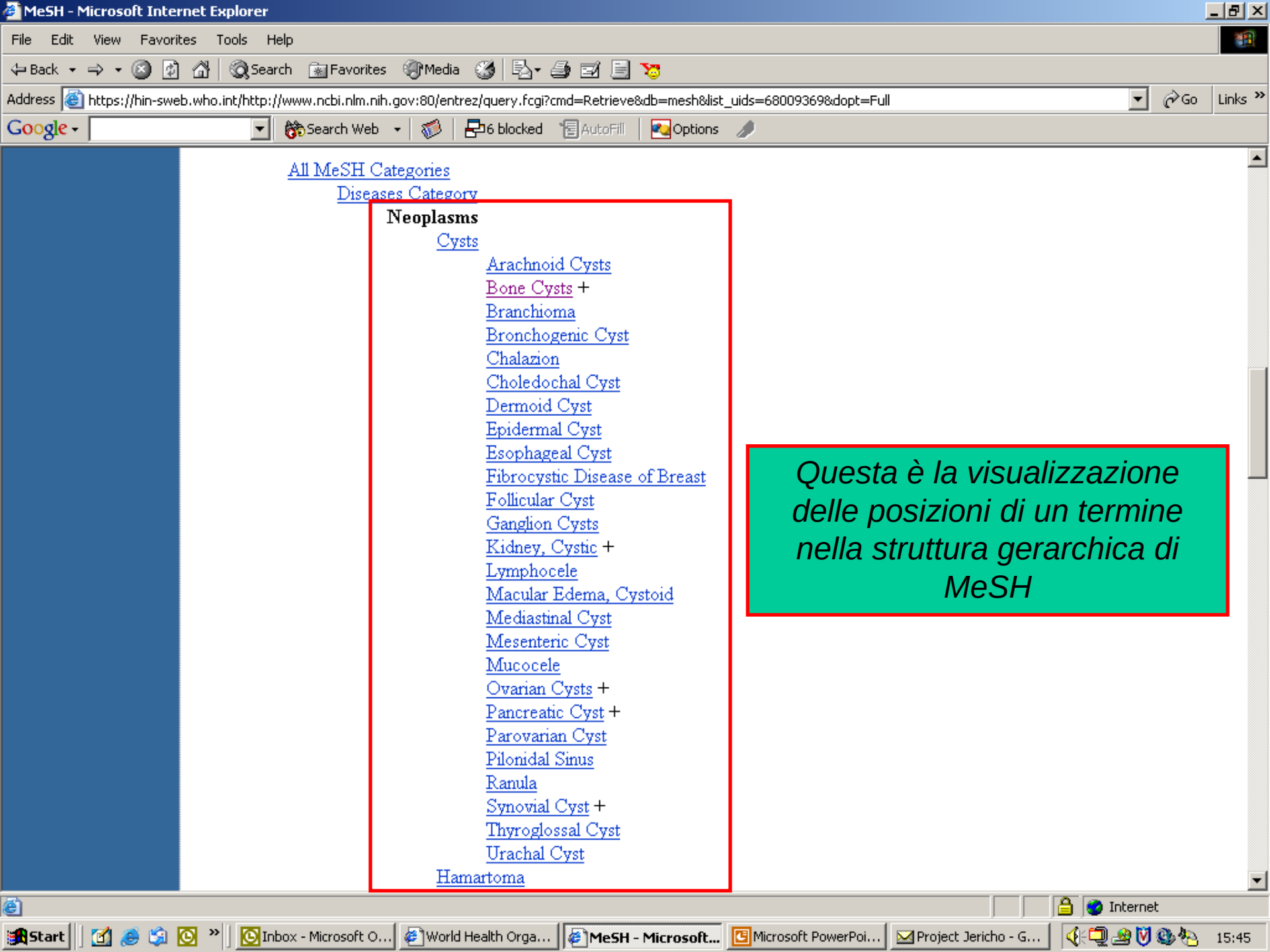
Il thesaurus MeSH è composto da:

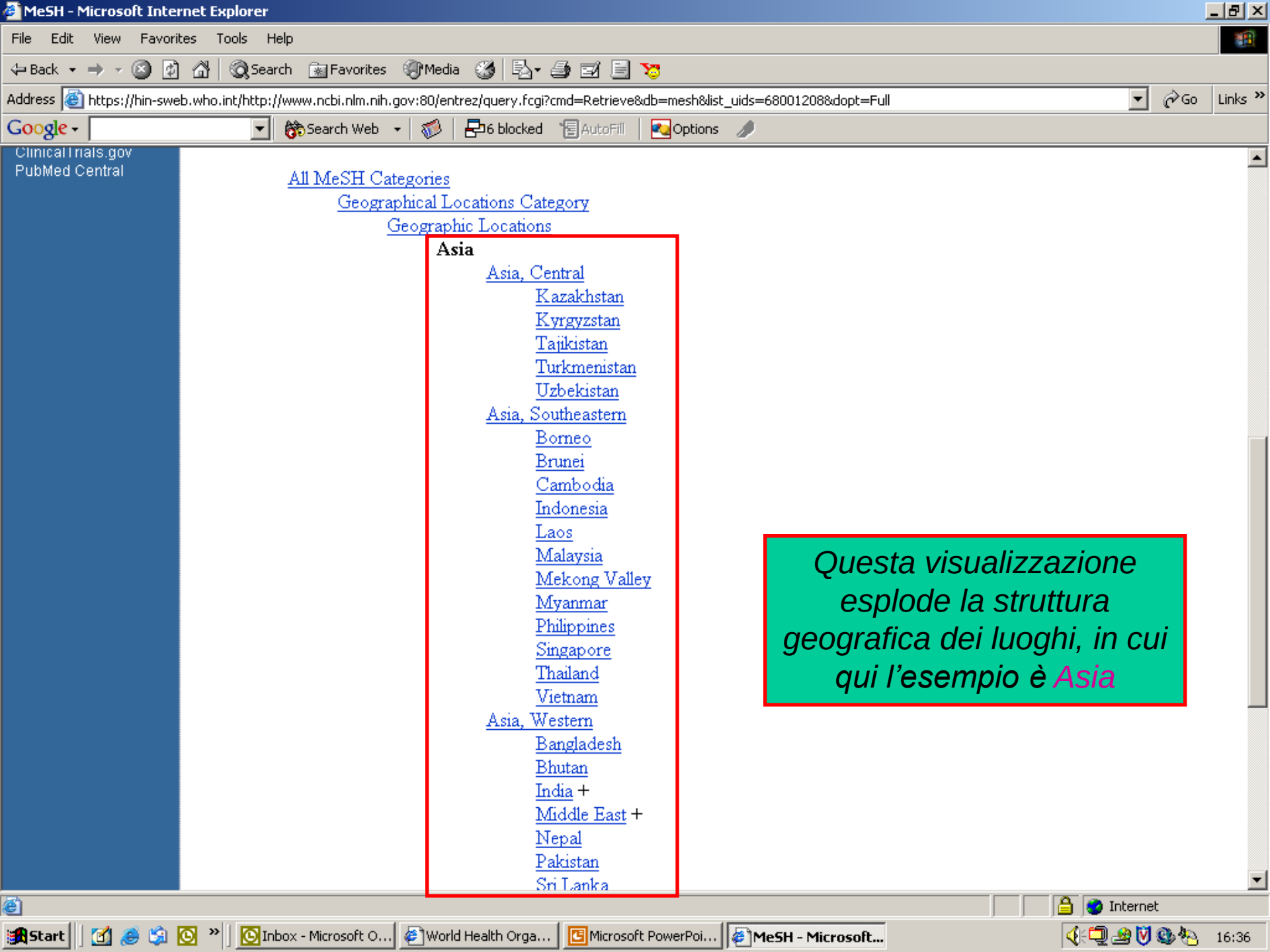
- oltre 22,000 descrittori (*main headings*)
- 81 sottodescrittori (*subheading* o *qualifiers*)
- oltre 100,000 voci supplementari (*Supplementary Concept Records*) , comprendenti nomi di sostanze chimiche, numeri di registro CAS, etc.

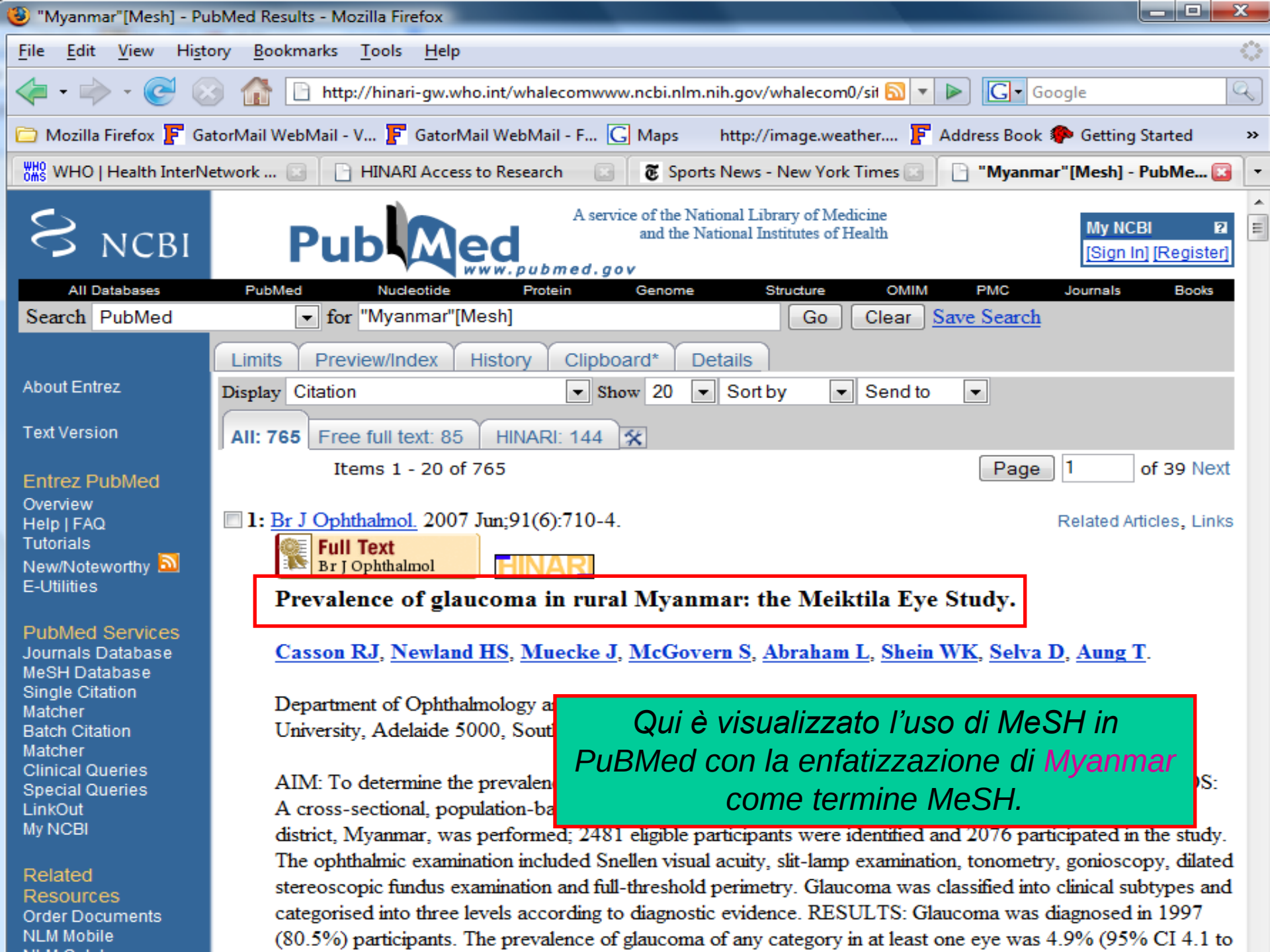

MeSH - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://hin-sweb.who.int/_base(http://www.ncbi.nlm.nih.gov/):http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=search&DB=mesh
Go Links
Google Search Web 6 blocked AutoFill Options
NCBI
MeSH
National Library of Medicine NLM
Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search MeSH for CANCER Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
Suggestions: Cancer; Cancers; Canes; Canary; Canis; Candy; Canada; Cane; Candies; Caper; more...
Display Summary Show: 20 Send to Search Box with AND
Items 1-20 of 100
Page 1 of 5 Next
Build a search strategy using the Send to Search Box feature.
Select a database (e.g., PubMed) under the Links menu to retrieve items with that term.
1: Neoplasms Links
New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.
2: Neoplasms, Second Prima Links
Abnormal growths of tissu y have the same or different histologic rises from an independent oncogenic the previous neoplasm since genetic risk or predisposing factors may actually be the cause.
Year introduced: 1992
3: Drug Screening Assays, Antitumor Links
Methods of investigating the effectiveness of anticancer cytotoxic drugs and biologic inhibitors. These include in vitro cell-kill models and cytostatic dye exclusion tests as well as in vivo measurement of tumor growth parameters in laboratory animals.
Questi risultati mostrano come il termine MeSH Cancer fa riferimento a Neoplasms..
Internet
Start
Inbox - Microsoft O...
World Health Orga...
MeSH - Microsoft...
Microsoft PowerPoi...
15:36


MeSH - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://hin-sweb.who.int/http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=mesh&list_uids=68009369&dopt=Full
Go Links
Google Search Web 6 blocked AutoFill Options
NCBI
MeSH
National Library of Medicine NLM
Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search MeSH for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
Display Full Show: 20 Send to Search Box with AND
If making selections (e.g., Subheadings, etc.), use the Send to Search Box feature to see PubMed records with those specifications.
Select PubMed under the Links menu to retrieve all records for the MeSH Term.
Select NLM MeSH Browser under the Links menu for additional information.
1: Neoplasms
Links
New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.
Subheadings:
analysis blood blood supply cerebrospinal fluid chemically induced chemistry classification complications
congenital diagnosis diet therapy drug therapy economics embryology enzymology epidemiology
ethnology etiology
mortality nursing
radionuclide imaging
transplantation
Restrict Search to Major Topic
Do Not Explode this term (i.e.,
Il record di un termine MeSH contiene una definizione, i sottotemi associati, un elenco di altri termini ed una visualizzazione della gerarchia (albero) di MeSH.
Questa è la definizione di Neoplasms.
Internet
Start
Inbox - Microsoft O...
World Health Orga...
MeSH - Microsoft...
Microsoft PowerPoi...
15:38


MeSH - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://hin-sweb.who.int/http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=mesh&list_uids=68009369&dopt=Full
Go Links
Google Search Web 6 blocked AutoFill Options
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
mortality nursing parasitology pathology physiopathology prevention and control psychology radiography
radionuclide imaging radiotherapy rehabilitation secretion surgery therapeutic use therapy transmission
transplantation ultrasonography ultrastructure urine veterinary virology
Restrict Search to Major Topic headings only
Do Not Explode this term (i.e., do not include MeSH terms found below this term in the MeSH tree).
Entry Terms:
Neoplasm
Tumors
Tumor
Benign Neoplasms
Neoplasms, Benign
Benign Neoplasm
Neoplasm, Benign
Cancer
Cancers
See Also:
Antibodies, Neoplasm
Antigens, Neoplasm
Antineoplastic Agents
Carcinogens
DNA, Neoplasm
Oncogenic Viruses
Precancerous Conditions
RNA, Neoplasm
Pleural Effusion, Malignant
Genes, Tumor Suppressor
Anticarcinogenic Agents
I termini di Mesh sono elencati e correlati con riferimenti a termini associati (See Also).
Internet
Start
Inbox - Microsoft Outlook
World Health Organizatio...
MeSH - Microsoft Inte...
Microsoft PowerPoint - [T...
15:41


MeSH - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://hin-sweb.who.int/http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=mesh&list_uids=68009369&dopt=Full
Go Links
Google Search Web 6 blocked AutoFill Options
All MeSH Categories
Diseases Category
Neoplasms
Cysts
Arachnoid Cysts
Bone Cysts +
Branchioma
Bronchogenic Cyst
Chalazion
Choledochal Cyst
Dermoid Cyst
Epidermal Cyst
Esophageal Cyst
Fibrocystic Disease of Breast
Follicular Cyst
Ganglion Cysts
Kidney, Cystic +
Lymphocele
Macular Edema, Cystoid
Mediastinal Cyst
Mesenteric Cyst
Mucocele
Ovarian Cysts +
Pancreatic Cyst +
Parovarian Cyst
Pilonidal Sinus
Ranula
Synovial Cyst +
Thyroglossal Cyst
Urachal Cyst
Hamartoma
Questa è la visualizzazione delle posizioni di un termine nella struttura gerarchica di MeSH
Internet
Start
Inbox - Microsoft O...
World Health Orga...
MeSH - Microsoft...
Microsoft PowerPoi...
Project Jericho - G...
15:45


MeSH - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://hin-sweb.who.int/http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=mesh&list_uids=68001208&dopt=Full
Go Links
Google Search Web 6 blocked AutoFill Options
ClinicalTrials.gov
PubMed Central
All MeSH Categories
Geographical Locations Category
Geographic Locations
Asia
Asia, Central
Kazakhstan
Kyrgyzstan
Tajikistan
Turkmenistan
Uzbekistan
Asia, Southeastern
Borneo
Brunei
Cambodia
Indonesia
Laos
Malaysia
Mekong Valley
Myanmar
Philippines
Singapore
Thailand
Vietnam
Asia, Western
Bangladesh
Bhutan
India +
Middle East +
Nepal
Pakistan
Sri Lanka
Questa visualizzazione esplode la struttura geografica dei luoghi, in cui qui l'esempio è Asia
Internet
Start
Inbox - Microsoft O...
World Health Orga...
Microsoft PowerPoi...
MeSH - Microsoft...
16:36


"Myanmar"[Mesh] - PubMed Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://hinari-gw.who.int/whalecomwww.ncbi.nlm.nih.gov/whalecom0/sit
Google
Mozilla Firefox GatorMail WebMail - V... GatorMail WebMail - F... Maps http://image.weather.... Address Book Getting Started
WHO | Health InterNetwork ... HINARI Access to Research Sports News - New York Times "Myanmar"[Mesh] - PubMe...
NCBI
PubMed
www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for "Myanmar"[Mesh] Go Clear Save Search
Limits Preview/Index History Clipboard* Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI
Related Resources
Order Documents
NLM Mobile
Display Citation Show 20 Sort by Send to
All: 765 Free full text: 85 HINARI: 144
Items 1 - 20 of 765
Page 1 of 39 Next
1: Br J Ophthalmol. 2007 Jun;91(6):710-4.
Related Articles, Links
Full Text
Br J Ophthalmol
HINARI
Prevalence of glaucoma in rural Myanmar: the Meiktila Eye Study.
Casson RJ, Newland HS, Muecke J, McGovern S, Abraham L, Shein WK, Selva D, Aung T.
Department of Ophthalmology a
University, Adelaide 5000, Sout
AIM: To determine the prevalen
A cross-sectional, population-ba
district, Myanmar, was performed; 2481 eligible participants were identified and 2076 participated in the study.
The ophthalmic examination included Snellen visual acuity, slit-lamp examination, tonometry, gonioscopy, dilated stereoscopic fundus examination and full-threshold perimetry. Glaucoma was classified into clinical subtypes and categorised into three levels according to diagnostic evidence. RESULTS: Glaucoma was diagnosed in 1997 (80.5%) participants. The prevalence of glaucoma of any category in at least one eye was 4.9% (95% CI 4.1 to
Qui è visualizzato l'uso di MeSH in PuBMed con la enfatizzazione di Myanmar come termine MeSH.

MeSH

- L’organizzazione gerarchica in MeSH descrive termini e le loro generalizzazione/specializzazioni ed è stata tradotta in piu’ lingue

MeSH - Medical Subject Headings

Traduzione Italiana

- ⊞ A01 (Regioni del corpo)
- ⊞ A02 (Sistema muscoloscheletrico)
- ⊞ A03 (Apparato digerente)
- ⊟ A04 (Apparato respiratorio)
  - ⊞ A04.329 (Laringe)
  - ⊟ A04.411 (Polmone)
    - ⊟ A04.411.125 (Bronchi)
    - ⊟ A04.411.300 (Fluido extravascolare del polmone)
    - ⊟ A04.411.715 (Alveoli polmonari)
      - ⊟ A04.411.715.025 (Barriera tra sangue e aria alveolare)

WordNet

- Un modello mentale del lessico dell'americano moderno (anche in italiano, spagnolo, inglese, basco, …)
- Motivato psicologicamente (George Miller e il suo team alla Princeton University).
- Circa *144,000* parole dell'Inglese Americano.
- Nomi, aggettivi, verbi e avverbi organizzati in unita' semantiche dette *synsets* (ca. 109,000 synonym sets).

Relazioni tra i synset in WordNet

- Antonym: front → back
- Attribute: benevolence → good (nomi vs. aggettivi)
- Pertainym: alphabetical → alphabet (aggettivi vs. i nomi)
- Similar: unquestioning → absolute
- Cause: kill → die
- Entailment: breathe → inhale
- Holonym: chapter → text (part-of)
- Meronym: computer → cpu (whole-of)
- ✓ Hyponym: tree → plant (specialization)
- ✓ Hypernym: fruit → apple (generalization)

Introduzione a Wordnet

- Vedi Lezione 17
  - "*Word Sense Disambiguation as a Machine Learning Task*"

Espansione della query e WordNet

- Aggiungi i sinonimi dello stesso *synset*.
- Aggiungi gli *hyponyms* per aumentare la informazione piu' specifica.
- Aggiungi *hypernyms* per generalizzare la *query*.
- Usa altre relazioni (come nel caso dei *related terms)* per espandere la *query*.
- Problemi aperti:
  - Come individuare il senso corretto dato un contesto (cioe' un documento o una short query in IR)
  - Come pesare i termini (sinonimi, hyponims, hyperonims) nella espansione?

Sensi ed IR

- I sensi delle parole (come in Wordnet) consentono la modellazione di
  - Documenti (*bag-of-senseID* puttosto che *bag-of-words*)
  - Interrogazioni
- Il problema è che:
  - I token (le parole) sono osservazioni *oggettive*
  - I sensi invece debbono essere derivati dai token e questa relazione è *N-a-M* (in generale)
- *WSD: Word Sense Disambiguation*

Wordnet: Semantic tagging

WSD: Word Sense Disambiguation

- E' il *task* di assegnamento ad una parola *w* in un contesto *C* il suo senso *s(w)* appropriato
- Richiede:
  - la disponibilità di un catalogo di sensi $s_i(w)$ (ad es. WN)
  - la disponibilità di una *metrica* in grado di misurare la correttezza di un senso $s_i(w)$ per la parola *w* in un contesto *C*

Un esempio

Sviluppo Statistico dei *Thesaurus*

- I tesauri compilati a mano non esistono e sono difficili e costosi da ottenere
  - Domini specifici
  - Lingue *non-English*
- I tipi di associazione semantica nei tesauri costruiti manualmente sono pochi (sinonimi, *hypernyms* e *related terms*).
- Associazione semantiche tra i termini possono essere scoperte automaticamente dalla analisi statistica di grandi collezioni di testi (>10^7 parole).

Automatic Global Analysis

- Determina la similitudine tra due termini (similarità semantica) attraverso una analisi statistica dell'intero corpus.
- Calcola una matrice di associazione che misura le correlazioni tra i termini in base alla frequenza dell'evento di loro co-occorrenza nei docs
- La espansione delle *query* quindi avviene aggiungendo i termini piu' simili statisticamente.

Matrice di Co-occorrenza

	d_1	d_2	d_3	.....	d_m
w_1	ω_{11}	ω_{12}	ω_{13}	.....	ω_{1m}
w_2	ω_{21}				
w_3	ω_{31}				
.	.				
.	.				
w_n	ω_{n1}				

ω_{ij}: metrica di pesatura tra il termine i ed il documento j

$$c_{ij} = (WW^T)_{ij} = \sum_{d_k \in D} \varpi_{ik} \times \varpi_{jk}$$

c_{ij}: è una metrica di associazione tra i termini i e j

OSS: WW^T è simmetrica e a valori positivi

Matrice di Associazione

	w_1	w_2	w_3	.....	w_n
w_1	c_{11}	c_{12}	c_{13}	.....	c_{1n}
w_2	c_{21}				
w_3	c_{31}				
.	.				
.	.				
w_n	c_{n1}				

c_{ij}: Fattore di correlazione tra il termine *i* e il termine *j*

$$c_{ij} = \sum_{d_k \in D} f_{ik} \times f_{jk}$$

f_{ik} : Frequenza di un termine *i* nel documento *k*

Matrice di Associazione Normalizzata

- Un fattore di correlazione basato sulla sola frequenza favorisce i termini più frequenti.
- Normalizzazione dei fattori di associazione:

$$s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}$$

- Un fattore di associazione normalizzato e' 1 se i due termini hanno la stessa frequenza in tutti i documenti.

Matrici di Correlazione Metriche …

- La correlazione di tipo associazione non e' sensibile alla prossimità dei termini nei documenti
- Le correlazioni metriche includono la prossimità tra i termini:

$$c_{ij} = \sum_{k_u \in V_i} \sum_{k_v \in V_j} \frac{1}{r(k_u, k_v)}$$

V_i: Insieme delle occorrenze del termine i in qualsiasi documento.
$r(k_u,k_v)$: Distanza in numero di parole tra le due occorrenze k_u e k_v (∞ se k_u e k_v occorrono in diversi documenti).

… Normalizzate

- I fattori vengono normalizzati per bilanciare la influenza delle parole con frequenze piu' alte:

$$s_{ij} = \frac{c_{ij}}{|V_i| \times |V_j|}$$

Matrici di Correlazione ed Espansione

- Per ogni termine *i* nella query *q*, espandi *q* con gli *n* termini, *j*, con il valore più alto di c_{ij} (s_{ij}).
- Questo aggiunge termini semanticamente correlati nell'interno dei termini originali della query.

Problemi con la Analisi Globale

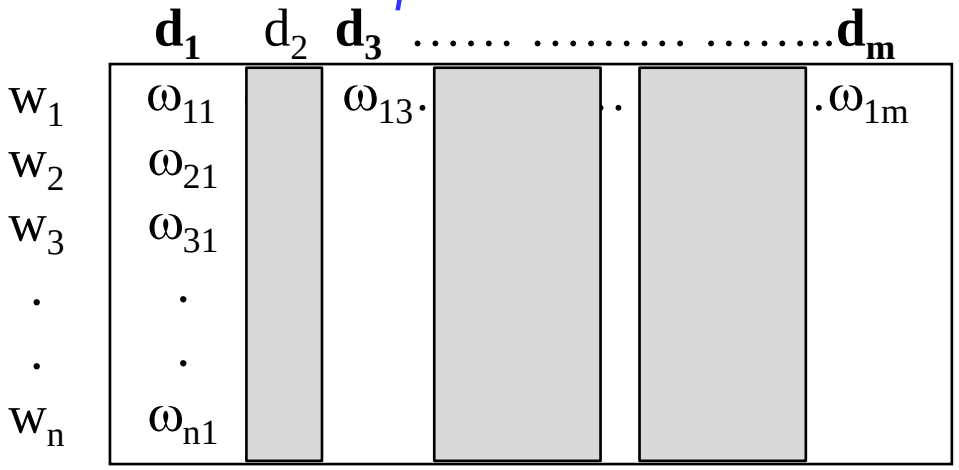
- Ambiguità dei termini può introdurre correlazioni irrilevanti ma statisticamente valide:
  - "*Apple computer*" → "*Apple red fruit computer*"
  - OSS: nota le possibili ambiguità di senso
- Tutti i termini altamente correlati (i piu' sicuri) sono già contenuti nei documenti rilevanti ed il loro uso può non fornire alcun nuovo documento utile.

Automatic Local Analysis

- Al *query time*, determina dinamicamente i termini simili analizzando i soli documenti ritrovati e più in alto nel *ranking*.
- Opera la analisi delle correlazioni localmente, cioè solo sull'insieme dei documenti ritrovati per una certa *query*.
- Elimina sorgenti di ambiguità perché confinata a soli documenti rilevanti
  - "*Apple computer*" → "*Apple computer Powerbook laptop*"

Automatic Local Analysis

- La matrice delle correlazioni è ridotta ai soli documenti rilevanti recuperati D_r al primo *run*, cioe' a tutti e soli i $d_i \in D_r$
  - *Sia W_r la proiezione di W nei soli documenti di D_r*



$$c_{ij} = (W_r^T W_r)_{ij} = \sum_{d_k \in D_r} \varpi_{ik} \times \varpi_{jk}$$

$\Rightarrow$ WW^T deve essere ri-calcolato per ogni query

Global vs. Local Analysis

- La *global analysis* richiede grandi moli di calcolo solo una volta, cioè nella fase di sviluppo (indicizzazione).
- La *local analysis* richiede calcolo pesante per la correlazione tra termini ad ogni query a *run-time* (sebbene la complessità locale e' funzione di dimensioni del problema molto più piccole).
- … la *local analysis* da' i risultati migliori.

Raffinamento della *Global Analysis*

- Espandi solo i termini della query con i termini che sono simili a tutti gli altri termini della query.

$$sim(k_i, Q) = \sum_{k_j \in Q} c_{ij}$$

  - "*fruit*" non valido per "*Apple computer*" poiche' non correlato con "*computer*."
  - "*fruit*" espanso per "*apple pie*" poiche' "*fruit*" e' correlato sia a "*apple*" che a "*pie*."
- Funzioni di pesatura più complesse (rispetto alla sola frequenza) per il calcolo della correlazione tra termini.

Query Processing: Conclusioni

- La espansione delle *query* con termini correlati migliora significativamente le prestazioni, specialmente la *recall*.
- Comunque, la selezione dei termini “simili” deve essere molto accurata per il rischio di una caduta significativa della *precision*.
- Abbiamo visto come i tesauri rappresentao una risorsa “semantica” per i domini
- Lo sfruttamento dei tesauri richiede l’estensione del modello base di ad hoc retrieval con metodi di *disambiguazione semantica (WSD)*

Sommario

- Alcune operazioni sulla *query* son utili a definire una approssimazione migliore del concetto di *relevance*
- Nel *relevance feedback* la *query* viene manipolata sulla base di assunzioni sulla rilevanza dei documenti rilevanti al primo passo
  - Coinvolgimento dell'utente vs. *pseudo feedback*
  - Le misure di prestazione non debbono essere effettuate sui documenti già giudicati dall'utente
  - Lo *pseudo relevance feedback* ha numerose analogie con il *re-ranking* probabilistico

Sommario (2)

- La espansione della query puo' essere ottenuta anche attraverso l'uso di tesauri
- I tesauri sono dizionari specializzati in domini specifici o repertori di sononimi
- Abbiamo visto il caso di Wordnet
  - alternative esistono in domini specifici come quello medico di cui un esempio illustre è il Medical Subject Heading (MeSH)

Sommario (3)

- L’uso di tesauri generalisti è sensibile al problema della ambiguità dei termini
- La ricerca in Word Sense Disambiguation nell’ambito dell’IR si occupa dello sviluppo di algoritmi per la selezione del senso delle parole in contesti brevi (ad es. query o frasi nei documenti)
- Infine sono state discusse le tecniche automatiche per la creazione dei tesauri

Sommario (4)

- I processi per la creazione automatica dei tesauri differiscono per la architettura del processo di generazione automatica di termini correlati
- La *global analysis* viene computata a priori su tutta la collezione (quindi in modalità *off-line*) e non dipende da un specifica *query*
  - *Piu' efficiente*
  - *Meno accurata*
- La *local analysis* insiste sulla collezione evocata da una query
  - Meno *efficiente (ricalcolo ad ogni query)*
  - *Risultati sono migliori*