Analisi dell'IRS ht://Dig

Gruppo e-scava

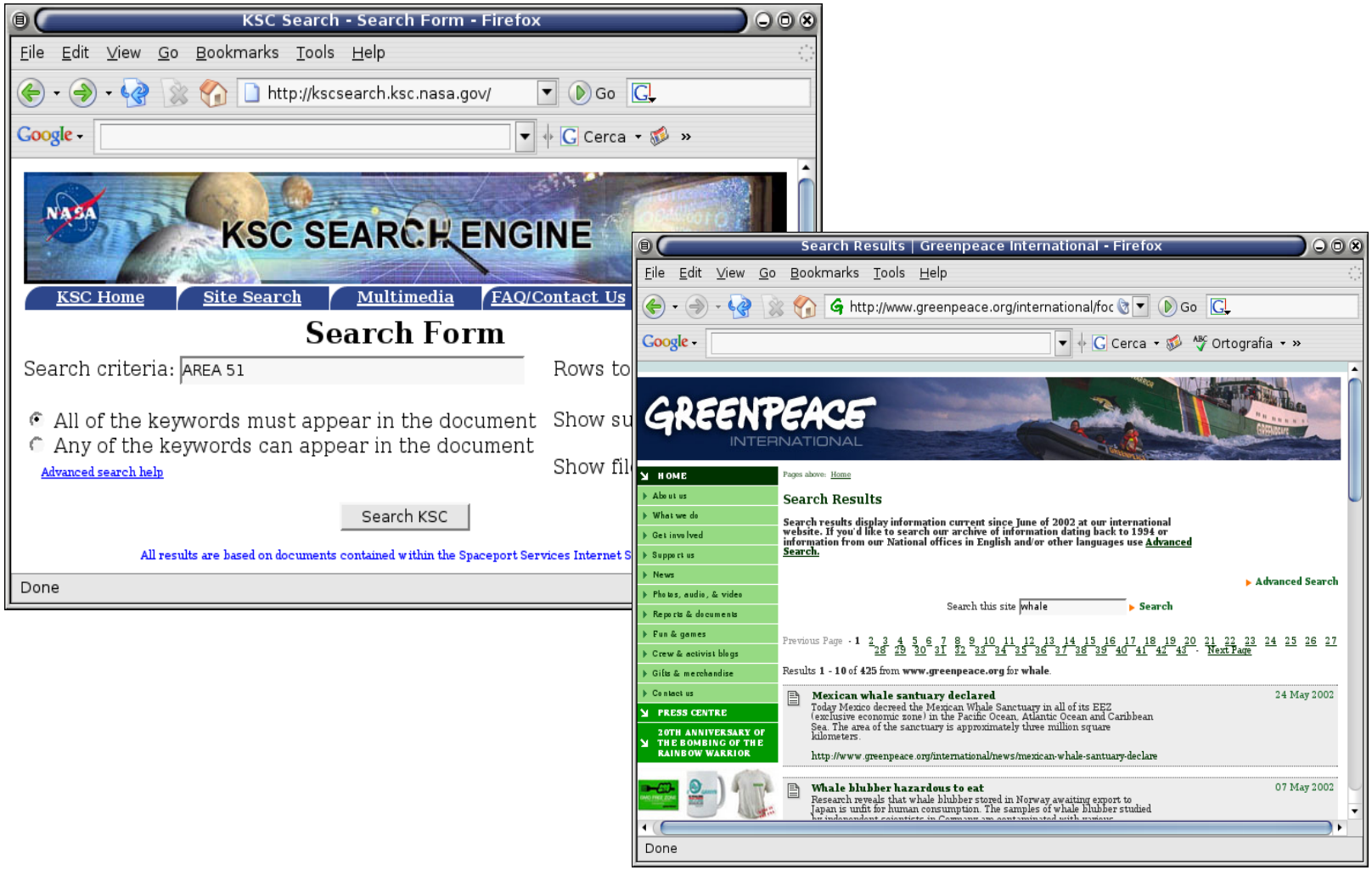
Fabio Dalla Libera, Fabrizio Lana, Daniele Masato
Luca Polin, Paolo Repele, Erik Squizzato
Michele Stecca, Federico Wegher

Introduzione (1/3)

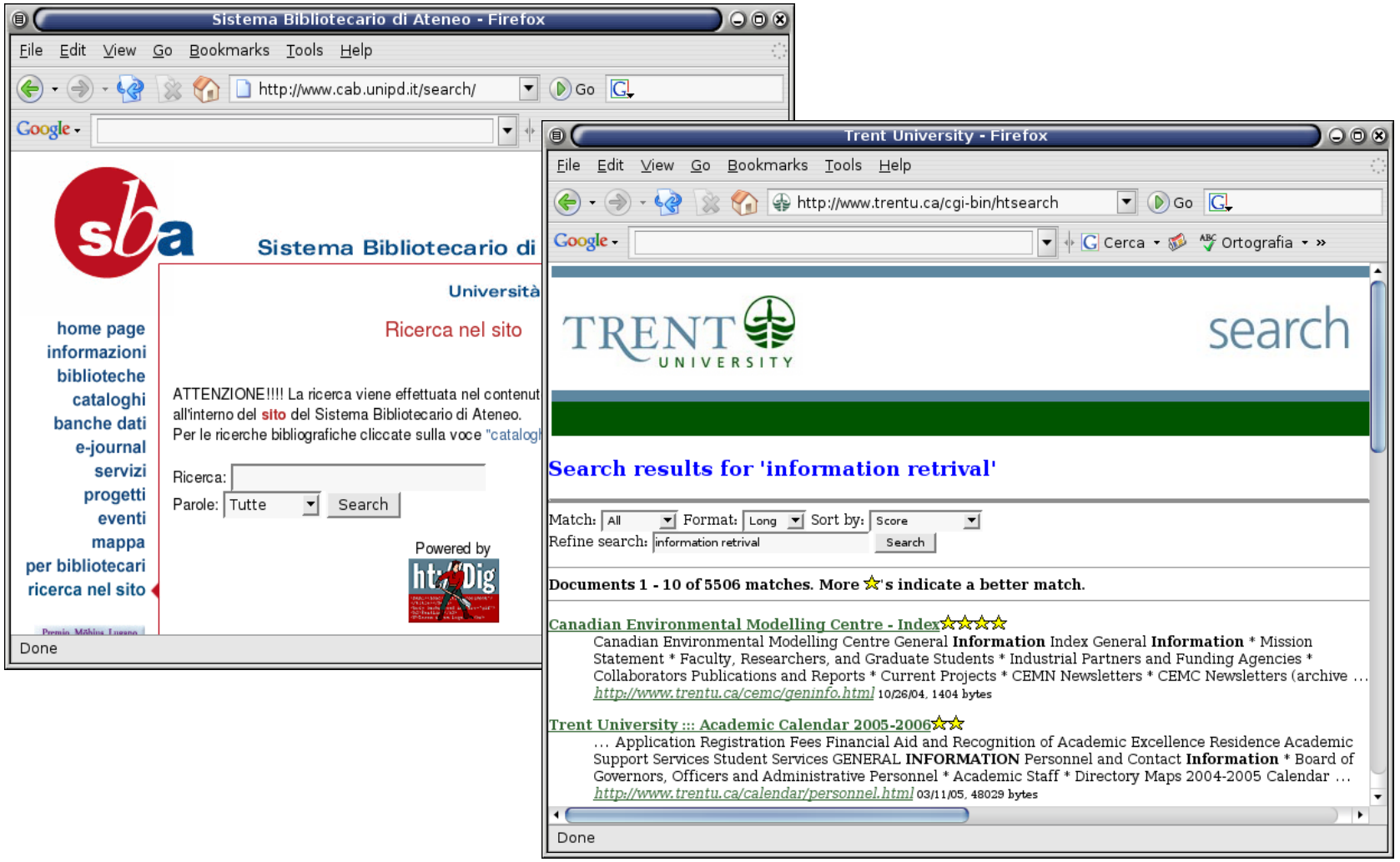
- Sistema: ht://Dig
- Versione: 3.2.0b6
- Produttore: The ht://Dig Group
- Licenza: GPL
- Sito Web: *http://www.htdig.org/*
- Utenza: multiutenza

- Usano ht://Dig: *(da http://www.htdig.org/uses.html)*
  - *http://kscsearch.ksc.nasa.gov/search/htdig/*
  - *http://www.cab.unipd.it/search/*
  - *http://www.trentu.ca/search.html*
  - *http://www.greenpeace.org/international/footer/search.htm*

Introduzione (2/3)



Introduzione (3/3)



Installazione

- Installazione: per ambiente Linux basata sulla classica terna di comandi

```
./configure
make
make install
```

- Dipendenze: Zlib (per compressione, facoltativa), libreria standard GNU per C (libc)
- Basato su più moduli (*htsearch*, *htdig*, *htfuzzy*, ...)
- Normalmente tutti i moduli utilizzano lo stesso file di configurazione
- E' possibile specificare più file di configurazione per gestire più collezioni

Documentazione

- “I don’t know what this is for. Does anyone?”

  http://htdig.sourceforge.net/dev/htdig-3.2/attrs.html

  Parametro: wordlist_cache_inserts
- Sufficiente per l’installazione del sistema
- Quasi assente per quanto riguarda il funzionamento e l’implementazione dei vari moduli
- Solo in inglese
- La più affidabile fonte di documentazione si è rivelata il codice

Configurazione

- Indicizzazione: incrementale

  E' comunque possibile effettuare un'indicizzazione non incrementale, specificando il parametro -i al modulo htdig

- Log applicazione: è possibile scegliere per qualsiasi modulo il livello di dettaglio del log, sino a giungere a *debug*, in cui ogni singola operazione viene segnalata
  - *logging: true/false, abilita il log su syslog*
  - *a seconda del numero di –v (max 4) passate da linea di comando varia la verbosità del log*

Collezione

- Formato documenti: solo in formato testuale e HTML

  Esistono comunque parser per PDF sviluppati da terzi che usano xpdf
- Codifica documenti: ASCII-7 ASCII-8 Latin-1 (ISO 8859-1)
- Dimensione: massimo 600 GB *
- Quantità: 500 milioni di documenti, 100 miliardi di parole nel dizionario *

  Server_max_docs (default: -1 = illimitati) permette di limitare ulteriormente il numero di documenti
- Compressione: *htdig* usa Mifluz che ha un suo sistema interno di compressione, oltre ai vari livelli di compressione di Zlib *
- Schemi Standard ID: non supportati
- Raccolta: nota o da Web
- Multicollezione: sì, occorre specificare più file di configurazione
- Formato resa: collegamento al documento indicizzato

* Da *http://www.gnu.org/software/mifluz/doc.en.html*

Indicizzazione (1/5)

- Viene effettuata dal modulo *htdig*
- Si specifica un insieme di pagine di partenza

  mediante il parametro start_url
- E' possibile indicare anche alcuni URL corrispondenti ad un percorso su disco

  mediante il parametro local_urls, ad esempio:

  local_urls: http://www.locale.com/collezione = /home/escava/coll
- Un crawler effettua ricorsivamente la visita di tutti i link dalla pagina iniziale; è possibile impostare dei vincoli:
  - *max_hop_count: imposta la distanza massima delle pagine da reperire, in termini di collegamenti dalle pagine iniziali*
  - *limit_urls_to: permette di specificare una serie di pattern di cui almeno uno deve essere contenuto nell'URL perché il documento venga reperito*
  - *exclude_urls: permette di specificare una serie di pattern che non devono essere contenuti nell'URL affinchè il documento venga reperito*

Indicizzazione (2/5)

- Dati strutturati:
  - Titolo del documento *nel tag <TITLE>*
  - Intestazioni contenute nel documento *nei tag <H1, H2, ..., H6>*
  - Descrizioni dei collegamenti alla pagina

    contenuto del tag <A HREF=...> descrizione </A>
  - Autore del documento

    contenuto nel tag <META> con name="author" e content="autore"
  - Due gruppi di tag *<META>* definibili dall'utente (tutti i tag di uno stesso gruppo vengono considerati come equivalenti)
  - Vengono salvati per ogni documento, ma non è possibile utilizzarli come criteri di ricerca:
    - URL
    - data del documento *nel tag <META> con name="date" content="data"*
    - numero di collegamenti alla pagina
    - numero di collegamenti contenuti nella pagina
    - *anchor* del documento *nel tag <A name=...>*
- Autority File: non supportato

Indicizzazione (3/5)

Flag (htcommon/HtWordReference.h)	**Fattore corrispondente**	**Tag associati** (htdig/HTML.cc)
FLAG_CAPITAL = 1	caps_factor	Mai impostato
FLAG_TITLE = 2	title_factor	*<TITLE>*
FLAG_HEADING = 4	heading_factor	*<H1>,<H2>, …, <H6>*
FLAG_KEYWORDS = 8	keywords_factor	Contenuto (content) di un tag *<META>* avente come nome una delle stringe indicate con il parametro *keywords_meta_tag_names*
FLAG_DESCRIPTION = 16	meta_description_factor	Contenuto (content) di un tag *<META>* avente come nome una delle stringe indicate con il parametro *description_meta_tag_names*
FLAG_AUTHOR = 32	author_factor	Contenuto di un tag *<META>* con nome (name) *author*
FLAG_LINK_TEXT = 64	description_factor	*<A HREF=…> text </A>*
FLAG_URL = 128	Non esistente	Mai impostato

Indicizzazione (4/5)

Caratteristiche Supportate

- Keyword: sì
- Termini: le parole separate da simboli a scelta dell'utente (es. "-") vengono salvate sia come parole distinte che come unica parola

  E' inoltre possibile la ricerca di frasi, racchiudendo le parole tra virgolette; le stopword vengono comunque eliminate anche in questo caso
- Stemming: non supportato (all'atto del reperimento è però implementato come meccanismo di query expansion)
- Pesi: per ogni parola viene salvato un flag che indica a quali tag appartiene (vedi tabella precedente)
- Localizzazione: viene salvata come intero
- Stoplist: presente e personalizzabile

  Ne vengono fornite due, quella di default è presente in /share/htdig/bad_words

  Il parametro bad_words_list permette di specificare una diversa stoplist
- Multimedia: non supportato

Indicizzazione (5/5)

Descrizione degli indici

- Dizionario: presente; in RAM viene mantenuta una cache con le entry del posting file usate più di recente
- Posting: utilizza Mifluz, una libreria basata su Berkeley DB; per memorizzare le posting list viene usato un BTree
- Vengono creati tre database: *(dimensioni riferite alla collezione CACM)*
  - un database dei documenti *(db.docdb, 400 KB)* nel quale per ogni pagina visitata viene salvato: ID numerico, titolo, la data, gli *anchor*, il numero di collegamenti contenuti nella pagina ed il numero di collegamenti che puntano alla pagina
  - un database delle citazioni *(db.excerpts, 1,3 MB)* in cui vengono salvati i primi byte di testo per documento in cui vengono evidenziate le parole chiave; *con il parametro max_head_length viene impostato il numero di byte da salvare*
  - un database delle parole *(db.words.db, 1,6 MB)* che memorizza per ogni occorrenza: l'ID del documento, la posizione all'interno del testo, l'*anchor* precedente ed un intero che rappresenta una serie di flag booleani indicanti dentro quali tag è contenuta la parola

La dimensione dei tre database è di 14 MB (!) se non si utilizza la compressione standard di Zlib

Funzione di Reperimento (1/3)

- Modello di reperimento: booleano con tecniche per la stima della rilevanza, permette il ranking
- Output a due livelli:
  - estratto, utilizza la citazione salvata nell'apposito database
  - collegamento al documento

  L'attributo matches_per_page imposta il numero di risultati per pagina
- Relevance feedback: assente
- Gestione sinonimia: presente
- Gestione polisemia: assente

Funzione di Reperimento (2/3)

- Mediante il file di configurazione è possibile specificare vari algoritmi di query expansion:
  - Soundex mappa parole con pronunce simili in uno stesso codice (specifico per cognomi inglesi);

    Algoritmo riportato in http://en.wikipedia.org/wiki/Soundex
  - Metaphone è simile a Soundex, ma specifico per la lingua inglese;

    Algoritmo riportato in http://www.wbrogden.com/phonetic/notice.html
  - Accents mappa tutte le lettere accentate nella loro controparte non accentata (e viceversa);
  - Synonyms dato un file di coppie di parole mappa ciascuna parola nell'altra;
  - Endings utilizza un dizionario ed un insieme di regole con la sintassi di *Ispell*, mappa ogni parola nelle sue varianti ed ogni variante nella parola radice;

Funzione di Reperimento (3/3)

- Prefix mappa ciascuna parola terminante con una stringa configurabile in tutte le parole del dizionario che hanno la parola come prefisso;
- Substring mappa ciascuna parola in tutte le parole del dizionario che la contengono come sottostringa;
- Regex data una parola la considera come un'espressione regolare e la mappa nelle corrispondenti parole del dizionario;
- Speling data una parola la mappa nelle parole del dizionario ottenute invertendo due lettere o togliendone una

- Per *soundex*, *metaphone*, *accents*, *synonyms*, *endings* vengono creati da *htfuzzy* dei database appositi prima che possano essere effettuate delle interrogazioni; i database per *synonyms* ed *endings* non dipendono dal dizionario
- Per *prefix*, *substring*, *regex* e *speling* l'elaborazione viene effettuata al momento dell'interrogazione; per *substring* e *regex* il confronto viene effettuato con tutto il dizionario della collezione causando un aumento del tempo di risposta

Interrogazione

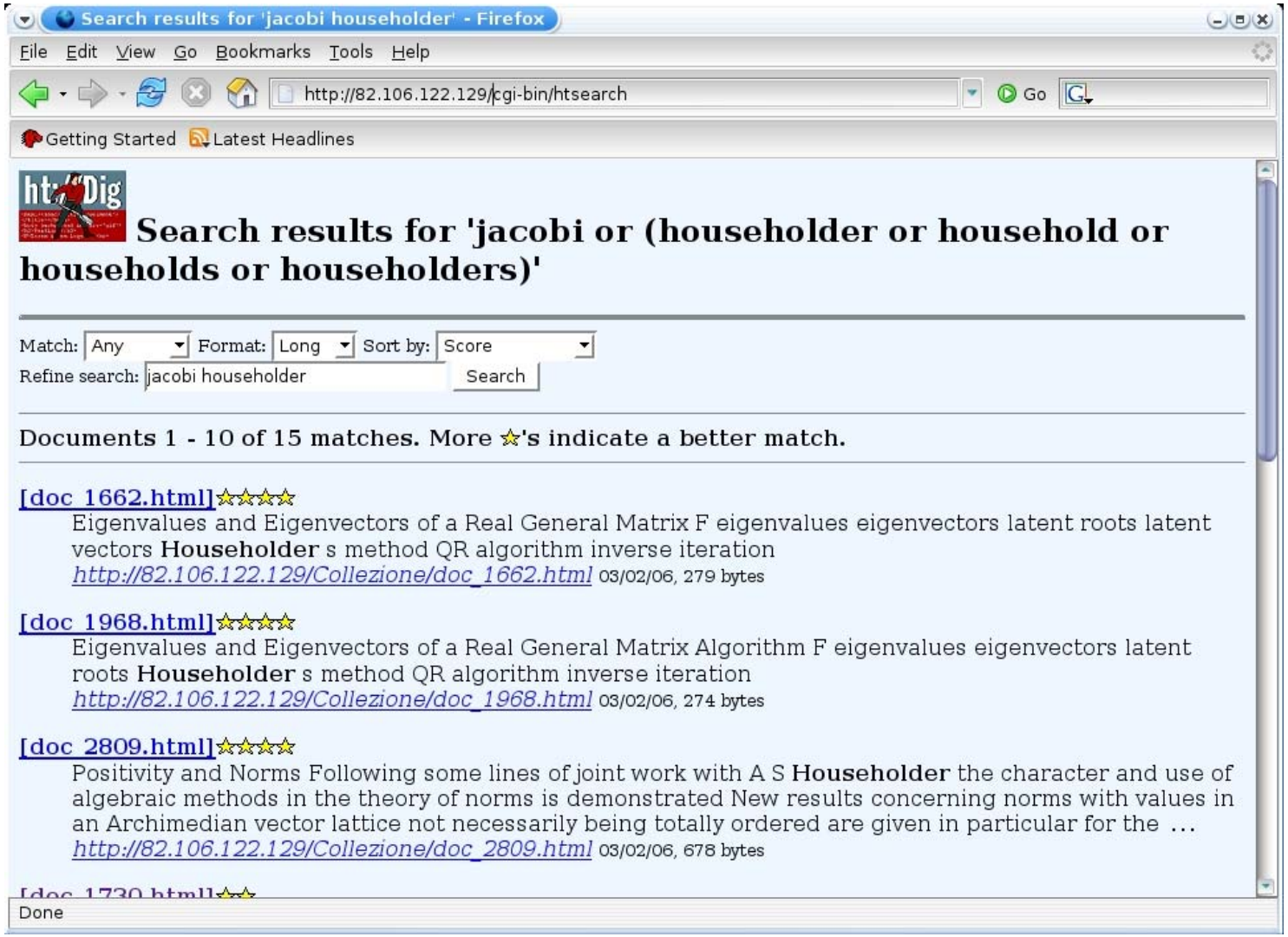
- Interfaccia: web, mediante CGI
- Linguaggio: naturale
- Help in Linea: no
- Avanzato: sì, quello Booleano
- Lingua: monolingua
- Sessione: no
- Personalizzazione: no

Interfaccia (1/2)

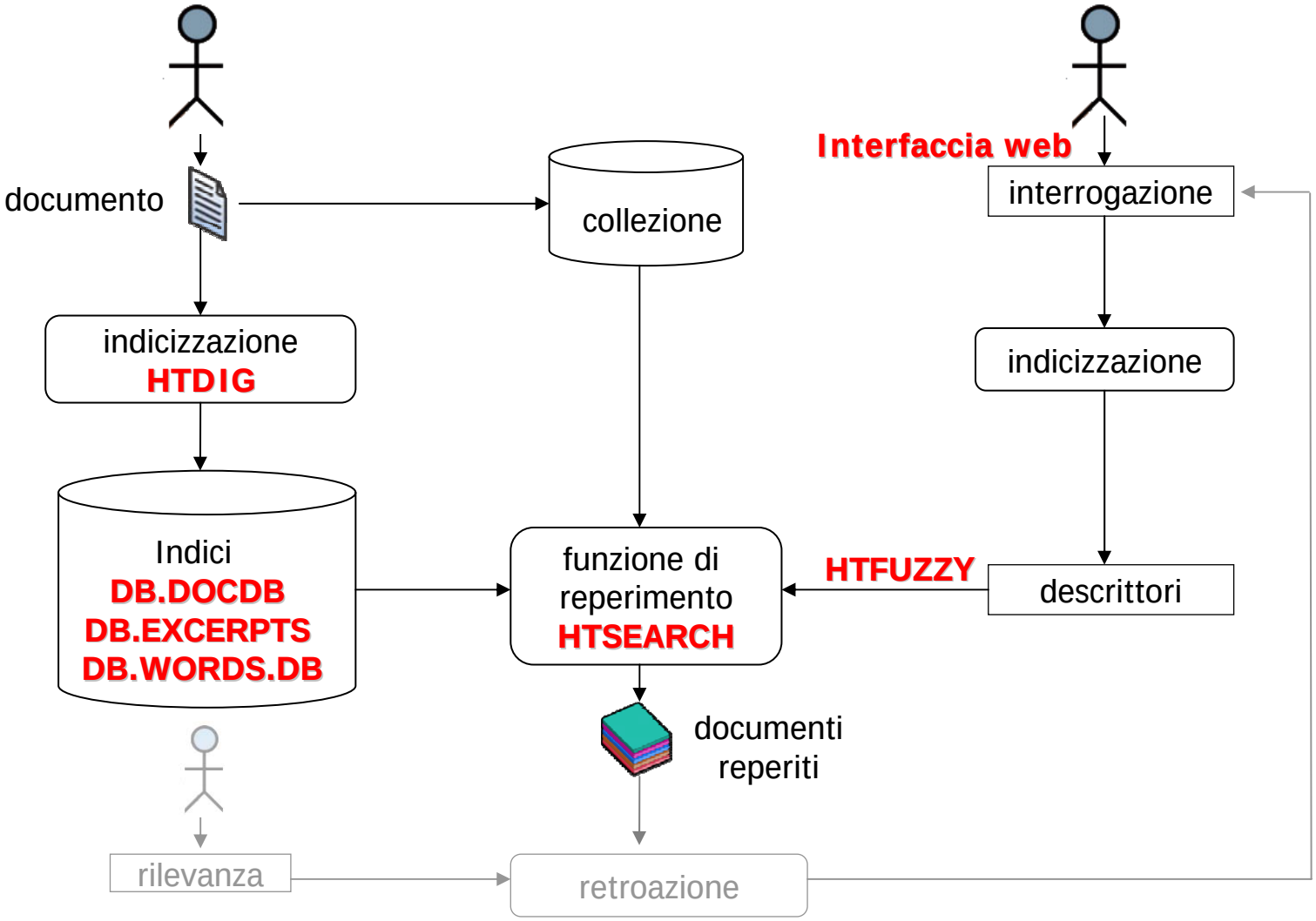
- L'interfaccia è semplice ed essenziale
- E' possibile specificare vari tipi di *match* per l'interrogazione:
  - All: esegue l'AND logico tra tutte le parole
  - Any: esegue l'OR logico tra tutte le parole
  - Boolean: è possibile specificare un'espressione booleana mediante gli operatori AND, OR, NOT

    Con il parametro boolean_keywords è possibile sostituire AND OR NOT con altre stringhe, ad esempio E, O, NON
- Sono disponibili vari tipi di ordinamento:
  - Score (Inverted Score)
  - Time (Inverted Time)
  - Title (Inverted Title)

Interfaccia (2/2)



Schema riassuntivo



Algoritmo di ranking (1/7)

- Quando si effettua un'interrogazione ciascuna parola viene espansa in una sequenza di parole legate mediante l'operatore OR, a seconda di quali degli algoritmi precedentemente descritti vengono attivati
- A ciascun algoritmo viene associato un peso che verrà assegnato alle parole trovate utilizzandolo
  - *Nel codice si nota che se una stessa parola viene ottenuta mediante diversi criteri di espansione non la si considera con peso pari alla somma dei pesi, ma ciascuna occorrenza viene trattata indipendentemente*
  - *I vocaboli specificati dall'utente nell'interrogazione assumono il peso di un algoritmo fittizio denominato exact, che associa ogni parola a se stessa*
- Per ciascuna parola viene recuperata una lista di documenti che la contiene
- Per ciascun documento viene salvato un punteggio *(score)* ed un contatore (inizialmente posto a 1), detto *orMatches*, utilizzato per indicare che lo stesso documento contiene più parole diverse ricercate, anche se queste erano legate da un OR e quindi non era necessaria la loro presenza contemporanea.

Algoritmo di ranking (2/7)

Rappresentazione della query *(file parser.cc)*

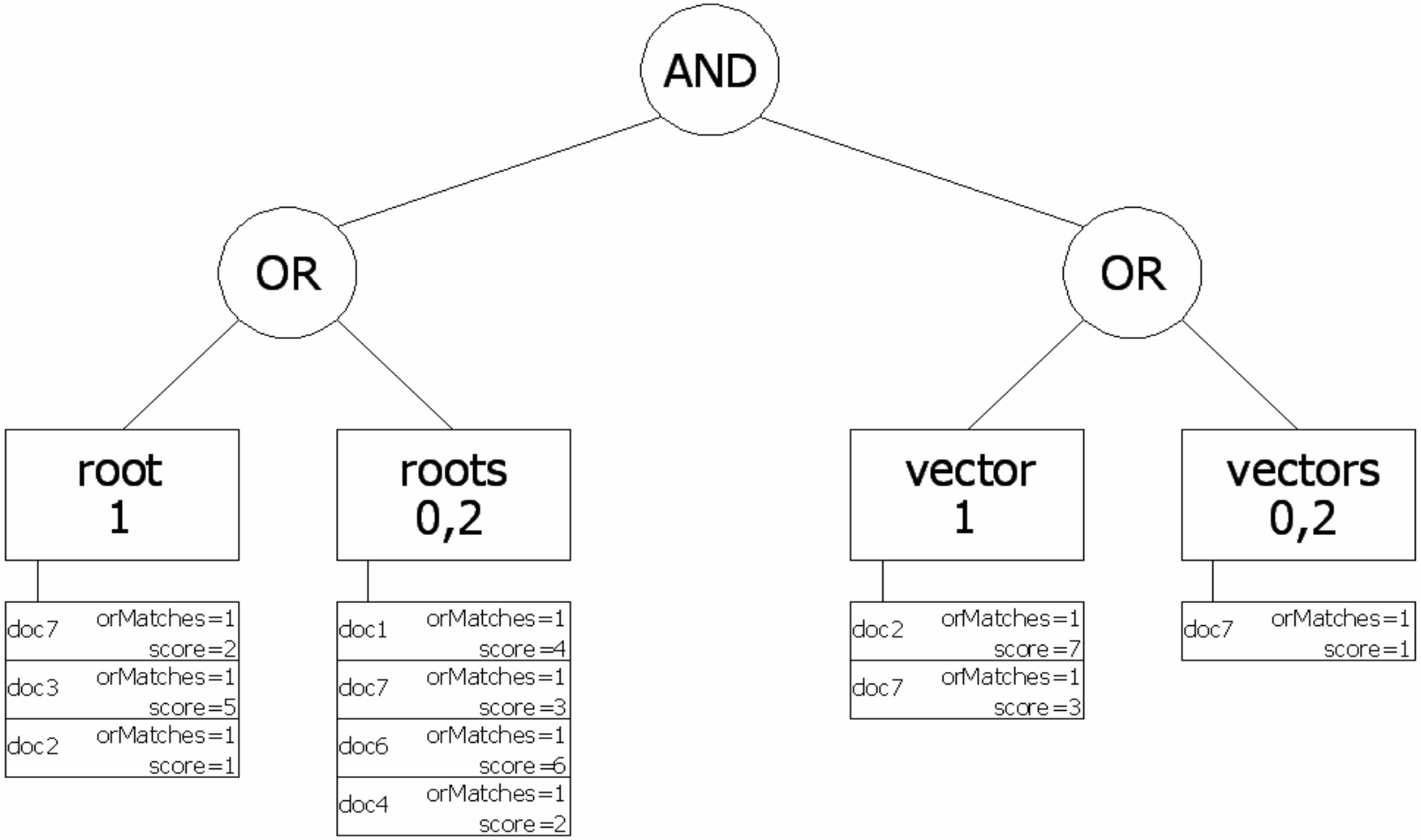
- Viene gestito uno stack che concettualmente corrisponde ad un albero, dove ogni foglia rappresenta una parola con la lista di documenti in cui compare e ogni nodo interno un operatore
- Il calcolo dello *score* per ogni documento avviene secondo la formula:
  - $$\frac{peso\_parola}{\#occorrenze} \cdot \sum_{i=1}^{n} factor_i \cdot flag_i$$

dove:

- i vari flag ed i factor sono quelli precedentemente riportati
- *peso_parola:* dipende dall'algoritmo che l'ha trovata
- *#occorrenze:* è il numero di occorrenze della parola su tutti i documenti della collezione

Algoritmo di ranking (3/7)

Esempio: interrogazione “root AND vector”, espansione mediante algoritmo *endings* con peso 0,2

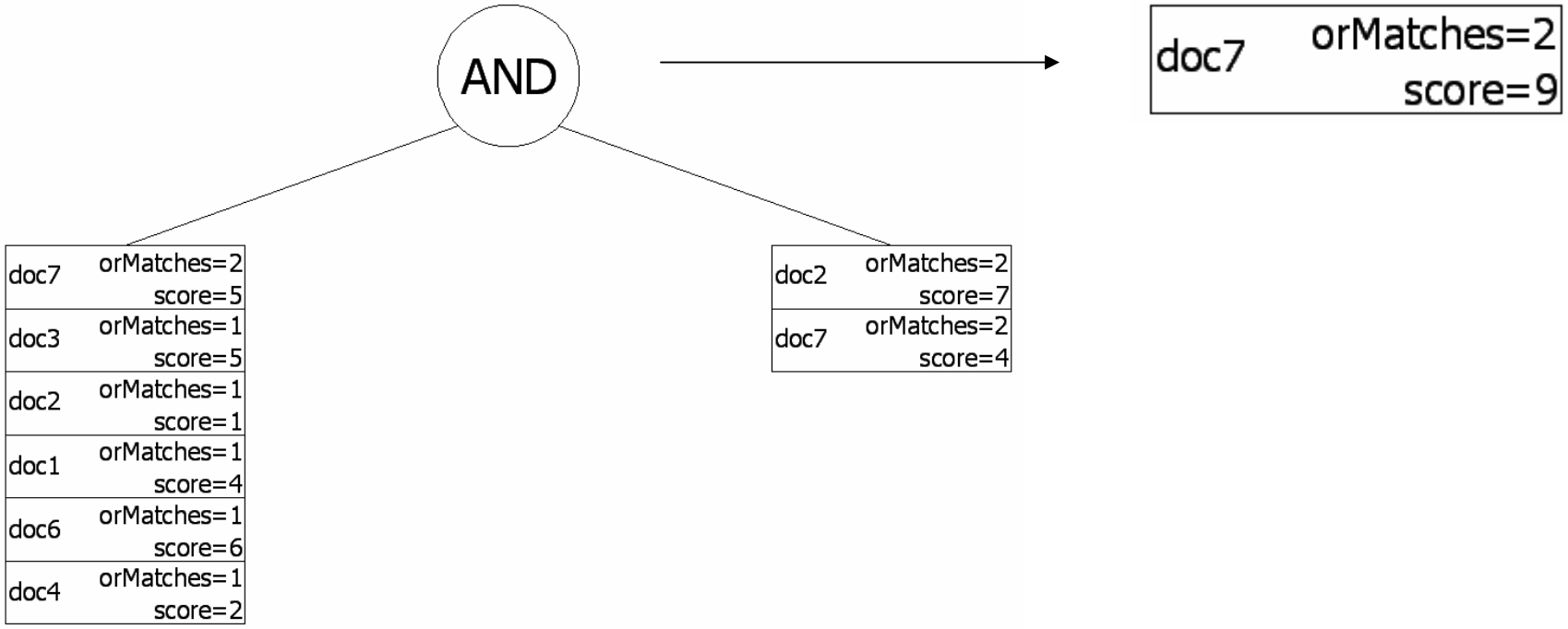


Algoritmo di ranking (4/7)

- Successivamente le liste di documenti associate alle varie parole vengono fuse, seguendo le regole di precedenza, secondo il significato insiemistico degli operatori: detta A la prima lista, B la seconda e C la lista risultante e detti $score_L^i$ ed $orMatches_L^i$ rispettivamente, il punteggio ed il valore del contatore *orMatches* del documento i nella lista L
- C=A OR B
  - $score_C^i = score_A^i + score_B^i$
  - $orMatches_C^i = orMatches_A^i + orMatches_B^i$
- C=A AND B
  - $score_C^i = score_A^i + score_B^i$
  - $orMatches_C^i = \left\lfloor \frac{orMatches_A^i + orMatches_B^i}{2} \right\rfloor$
- C=A NOT B (equivalente a A AND NOT B)
  - $score_C^i = score_A^i$
  - $orMatches_C^i = orMatches_A^i$

Algoritmo di ranking (5/7)

Esempio: ultimo passo nell'analisi dell'interrogazione "root AND vector"



Algoritmo di ranking (6/7)

- Si ottiene un'unica lista con tutti i documenti che soddisfano la query
- Se $orMatches > 1$ allora il punteggio del documento viene aggiornato come segue:
  - $score = score \cdot (1 + multimatch\_factor)$

multimatch_factor è un valore impostabile nel file di configurazione

- Per aumentare lo score dei documenti più recenti, allo score di ciascun documento viene aggiunto il termine
  - $$score = score + \frac{date\_factor \cdot 100}{1 + \frac{|data\_corrente - data\_documento|}{numero\_secondi\_anno}}$$

date_factor è un valore impostabile nel file di configurazione

Algoritmo di ranking (7/7)

- Per dare più peso ai documenti che vengono puntati da molti altri:
  - $score = score + \frac{backlink\_factor \cdot backlinks}{links}$

  backlinks e links sono rispettivamente il numero di link che puntano al documento e il numero di link contenuti nel documento

- E' possibile specificare una lista di fattori moltiplicativi e additivi per la presenza di pattern negli URL dei documenti *(mediante il parametro url_seed_score)*

- Il punteggio viene normalizzato secondo la formula:
  - $score = \log(1 + score)$

- Il numero di stelle presenti nell'interfaccia grafica viene calcolato sulla base dei punteggi ottenuti, come segue:
  - $n_{stelle} = \left\lfloor \frac{score - minscore}{maxscore - minscore} \cdot (max\_stars - 1) + \frac{1}{2} \right\rfloor + 1$

Valutazione delle prestazioni (1/15)

- Scomposizione collezione in più file mediante uno script bash
- Modifica del modulo *htsearch* affinché ricavi le query e relativi documenti rilevanti dai file *query.txt* e *qrels.txt*
- Applicazione delle query usando diverse stoplist
  - stoplist predefinita (17 parole)
  - stoplist aggiuntiva fornita da ht://Dig (350 parole)
  - *http://www.unine.ch/info/clef/englishST.txt* (571 parole)
  - stoplist effettuata analizzando la collezione, composta dalle parole che compaiono in più di 500 documenti e in meno di 2 (4342 parole)
  - stoplist ricavata dall'analisi delle interrogazioni (173 parole)
- Utilizzo di 26 diverse configurazioni diverse per il calcolo di precisione, richiamo e fallout, applicando vari algoritmi di ricerca e modificandone il peso associato
- Per ogni configurazione vengono considerati operatori AND e OR tra le parole, per un totale di 52 prove

Valutazione delle prestazioni (2/15)

LEGENDA

- **R, P, F =** richiamo, precisione e fallout medi ottenuti considerando tutti i documenti reperiti. Il primo valore in ciascuna cella indica la media di tipo macro, il secondo valore la media di tipo micro. Tutti i valori sono espressi in percentuale.
- **R_5, P_5, F_5 =** richiamo, precisione e fallout medi ottenuti considerando i primi cinque documenti reperiti. Il primo valore in ciascuna cella indica la media di tipo macro, il secondo valore la media di tipo micro. Tutti i valori sono espressi in percentuale.
- **R_{10}, P_{10}, F_{10} =** richiamo, precisione e fallout medi ottenuti considerando i primi dieci documenti reperiti. Il primo valore in ciascuna cella indica la media di tipo macro, il secondo valore la media di tipo micro. Tutti i valori sono espressi in percentuale.
- **N =** numero di interrogazioni usate per il calcolo della media di tipo macro.Il primo valore indica il numero di interrogazioni usate per il calcolo del richiamo, il secondo il numero di interrogazioni usate per il calcolo della precisione
- **t_{avg} =** tempo medio di risposta all'interrogazione in millisecondi
- **OR/AND =** indica il tipo di operatore usato
- **Sp =** stoplist predefinita
- **S2 =** seconda stoplist fornita con ht://Dig
- **Si =** stoplist trovata in Internet
- **Ss =** stoplist ottenuta mediante script
- **Sm =** stoplist ottenuta manualmente
- **L:x =** numero minimo di caratteri delle parole indicizzate, con x che ne indica il valore
- **Mf:x =** *multimatch_factor,* con x numero che ne indica il valore. Se non viene specificato il valore allora *multimatch_factor* assume il valore predefinito pari a 1
- **Al:x =** algoritmo di espansione, dove Al indica le prime due lettere del nome dell'algoritmo e x è un numero che indica il peso con cui è stato utilizzato (si sottintende che viene sempre usato exact con peso 1)

Valutazione delle prestazioni (3/15)

Configurazione di base

- Utilizzando l'operatore AND si ottiene alta precisione sia globale sia sui primi 5/10 documenti reperiti, ma il richiamo è molto basso (solo 6 interrogazioni trovano risposta)
- Utilizzando l'operatore OR si verifica la situazione opposta

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
OR Sp L:3	84,12	1,51	24,03	19,31	23,75	0,12	25,28	18,44	0,26	52	54
	82,91	1,33	24,02	9,55	23,75	0,12	14,82	18,44	0,26	64	
AND Sp L:3	1,78	89,58	0,00	1,43	96,67	0,00	1,78	93,33	0,00	52	25
	1,38	52,38	0,00	1,13	90,00	0,00	1,38	73,33	0,00	6	

Valutazione delle prestazioni (4/15)

Applicazione algoritmi di espansione con operatore OR

- Aumentano il richiamo globale, anche fino al 90%, ma riducono leggermente quello entro i primi 5/10 risultati
- Eccetto *synonyms* causano però decremento della precisione (incremento del fallout)
- Il peso degli algoritmi influenza in modo inversamente proporzionale il rank dei documenti rilevanti: aumentandolo diminuisce il richiamo nei primi 5/10 documenti, pur rimanendo globalmente elevato
- Si assiste comunque ad incremento dei tempi di risposta, in particolare per gli algoritmi che esaminano tutte le parole del dizionario: *synonyms* risulta il più veloce (incremento di pochi decimi di secondo), mentre *substring* richiede attesa oltre 150 volte superiore

Valutazione delle prestazioni (5/15)

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
OR Sp L:3	84,20	1,51	24,04	19,34	24,06	0,12	25,28	18,44	0,26	52	57
sy:0,2	83,17	1,33	24,03	9,67	24,06	0,12	14,82	18,44	0,26	64	
OR Sp L:3	84,20	1,51	24,04	19,34	24,06	0,12	25,13	17,97	0,26	52	56
sy:0,5	83,17	1,33	24,03	9,67	24,06	0,12	14,45	17,97	0,26	64	
OR Sp L:3	89,86	0,98	39,20	15,31	18,44	0,13	25,01	17,81	0,26	52	107
en:0,2	91,21	0,90	39,19	7,41	18,44	0,13	14,32	17,81	0,26	64	
OR Sp L:3	89,86	0,98	39,20	13,64	16,56	0,13	22,01	14,84	0,27	52	106
en:0,5	91,21	0,90	39,19	6,66	16,56	0,13	11,93	14,84	0,27	64	
OR Sp L:3	90,23	1,00	39,45	11,77	11,56	0,14	16,54	10,16	0,28	52	110
so:0,2	87,69	0,86	39,44	4,65	11,56	0,14	8,17	10,16	0,28	64	
OR Sp L:3	90,23	1,00	39,45	6,94	6,56	0,15	9,71	5,00	0,30	52	110
so:0,5	87,69	0,86	39,44	2,64	6,56	0,15	4,02	5,00	0,30	64	
OR Sp L:3	88,64	1,17	34,85	14,39	17,50	0,13	21,70	15,47	0,26	52	83
me:0,2	86,56	0,96	34,84	7,04	17,50	0,13	12,44	15,47	0,26	64	
OR Sp L:3	88,46	1,17	34,85	8,55	10,63	0,14	14,79	8,75	0,29	52	82
me:0,5	86,56	0,96	34,84	4,27	10,62	0,14	7,04	8,75	0,29	64	
OR Sp L:3	86,77	1,43	27,16	13,42	45,94	0,13	18,57	12,97	0,27	52	8617
su:0,2	87,44	1,24	27,15	6,41	15,94	0,13	10,43	12,97	0,27	64	
OR Sp L:3	86,77	1,43	27,16	9,48	11,25	0,14	15,64	9,84	0,28	52	8634
su:0,5	87,44	1,24	27,15	4,52	11,25	0,14	7,91	9,84	0,28	64	
OR Sp L:3	88,18	1,11	34,82	19,46	23,75	0,12	24,89	17,34	0,26	52	92
sp:0,2	88,57	0,98	34,80	9,55	23,75	0,12	13,94	17,34	0,26	64	
OR Sp L:3	88,18	1,11	34,82	18,70	21,56	0,12	22,86	16,72	0,26	52	93
sp:0,5	88,57	0,98	34,80	8,67	21,56	0,12	13,44	16,72	0,26	64	
OR Sp L:3	86,90	1,46	26,26	15,04	17,81	0,13	21,89	15,94	0,26	52	109
pr:0,2	87,19	1,28	26,25	7,16	17,81	0,13	12,81	15,94	0,26	64	
OR Sp L:3	86,90	1,46	26,26	13,74	14,38	0,13	19,40	12,97	0,27	52	108
pr:0,5	87,19	1,28	26,25	5,78	14,38	0,13	10,43	12,97	0,27	64	

Valutazione delle prestazioni (6/15)

Applicazione algoritmi di espansione con operatore AND

- In generale permettono di soddisfare più interrogazioni (fino a 8) rispetto alle 6 ottenute con la configurazione predefinita
- Eccetto *synonyms*, *prefix* e *substring* garantiscono un leggero incremento del richiamo sia globale, sia sui primi documenti
- Causano però visibile perdita di precisione rispetto alla configurazione base, fino al 10% nei primi risultati, fino al 20% globalmente
- Il peso degli algoritmi non influenza il rank dei documenti rilevanti
- Si assiste comunque ad incremento dei tempi di risposta, in particolare per gli algoritmi che esaminano tutte le parole del dizionario: *synonyms* risulta il più veloce (incremento di pochi decimi di secondo), mentre *substring* richiede attesa oltre 300 volte superiore

Valutazione delle prestazioni (7/15)

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
AND Sp L:3	1,81	83,93	0,01	1,46	90,00	0,00	1,81	87,14	0,00	52	26
sy:0,2	1,51	52,17	0,01	1,26	83,33	0,00	1,51	70,59	0,00	7	
AND Sp L:3	1,81	83,93	0,01	1,46	90,00	0,00	1,81	87,14	0,00	52	27
sy:0,5	1,51	52,17	0,01	1,26	83,33	0,00	1,51	70,59	0,00	7	
AND Sp L:3	2,24	76,00	0,01	1,54	82,86	0,00	2,06	81,43	0,00	52	61
en:0,2	1,88	45,45	0,01	1,38	84,62	0,00	1,76	77,78	0,00	7	
AND Sp L:3	2,24	76,00	0,01	1,54	82,86	0,00	2,06	81,43	0,00	52	60
en:0,5	1,88	45,45	0,01	1,38	84,62	0,00	1,76	77,78	0,00	7	
AND Sp L:3	2,22	78,21	0,01	1,34	82,38	0,00	2,04	83,81	0,00	52	64
so:0,2	1,88	42,86	0,01	1,26	71,43	0,00	1,76	73,68	0,00	7	
AND Sp L:3	2,22	78,21	0,01	1,34	82,38	0,00	2,04	83,81	0,00	52	64
so:0,5	1,88	46,86	0,01	1,26	71,43	0,00	1,76	73,68	0,00	7	
AND Sp L:3	2,16	83,14	0,01	1,46	90,00	0,00	1,99	88,57	0,00	52	43
me:0,2	1,76	43,75	0,01	1,26	83,33	0,00	1,63	76,47	0,00	7	
AND Sp L:3	2,16	83,14	0,01	1,46	90,00	0,00	1,99	88,57	0,00	52	43
me.0,5	1,76	43,75	0,01	1,26	83,33	0,00	1,63	76,47	0,00	7	
AND Sp L:3	1,78	89,58	0,00	1,25	93,33	0,00	1,78	93,33	0,00	52	8768
su:0,2	1,38	53,38	0,00	1,01	80,00	0,00	1,38	73,33	0,00	6	
AND Sp L:3	1,78	89,58	0,00	1,25	93,33	0,00	178	93,33	0,00	52	8603
su:0,5	1,38	52,38	0,00	1,01	80,00	0,00	1,38	73,33	0,00	6	
AND Sp L:3	2,22	72,75	0,01	1,69	81,25	0,00	2,04	77,50	0,00	52	53
sp:0,2	1,88	44,12	0,01	1,51	85,71	0,00	1,76	73,68	0,00	8	
AND Sp L:3	2,22	72,75	0,01	1,69	81,25	0,00	2,04	77,50	0,00	52	52
sp:0,5	1,88	44,12	0,01	1,51	85,71	0,00	1,76	73,68	0,00	8	
AND Sp L:3	1,78	89,58	0,00	1,25	93,33	0,00	1,78	93,33	0,00	52	78
pr:0,2	1,38	52,38	0,00	1,01	80,00	0,00	1,38	73,33	0,00	6	
AND Sp L:3	1,78	89,58	0,00	1,25	93,33	0,00	1,78	93,33	0,00	52	80
pr:0,5	1,38	52,38	0,00	1,01	80,00	0,00	1,38	73,33	0,00	6	

Valutazione delle prestazioni (8/15)

Applicazione stoplist con operatore OR

- E’ la tecnica che garantisce migliori risultati con questo operatore
- In tutti i casi si ottiene un miglioramento della precisione piuttosto evidente nei primi 5/10 risultati, più moderato globalmente
- Il richiamo rimane stabile su valori elevati (> 80%), come per la configurazione predefinita
- La stoplist migliore si è rivelata essere quella costruita analizzando le interrogazioni in linguaggio naturale
- Si assiste ad incremento o riduzione dei tempi di risposta in relazione alla dimensione della stoplist in quanto essa influisce sul numero di documenti reperiti e sulle operazioni di confronto con i termini dell’interrogazione

Valutazione delle prestazioni (9/15)

Applicazione stoplist con operatore AND

- In generale le stoplist non causano né un significativo aumento del richiamo, né una grossa perdita di precisione, sia globalmente, sia per i primi risultati
- Solo la stoplist ricavata dalle interrogazioni riesce a migliorare di qualche punto percentuale il richiamo globale (3,34% contro 1,78%), ma a scapito di una riduzione della precisione superiore al 25%; in questo caso vengono comunque soddisfatte 8 query
- Per i tempi di risposta valgono le stesse considerazioni effettuate per l'operatore OR

Valutazione delle prestazioni (10/15)

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
OR S2 L:1	84,93	1,58	22,74	21,65	25,00	0,12	25,70	18,75	0,25	52	53
	82,29	1,39	22,73	10,05	25,00	0,12	15,08	18,75	0,25	64	
OR Ss L:1	88,19	1,02	46,80	20,99	24,38	0,12	25,25	19,06	0,25	52	285
	86,81	0,72	46,80	9,80	24,38	0,12	15,33	19,06	0,25	64	
OR Si L:1	84,93	1,67	20,50	22,41	25,62	0,12	27,87	19,06	0,25	52	48
	82,29	1,54	20,49	10,30	25,62	0,12	15,33	19,06	0,25	64	
Or Sm L:1	83,13	3,93	17,07	22,34	27,94	0,11	28,03	21,11	0,24	52	40
	80,28	1,80	17,06	10,55	27,01	0,11	15,58	19,97	0,24	63	
AND S2 L:1	1,78	89,58	0,00	1,43	96,67	0,00	1,78	93,33	0,00	52	25
	1,38	52,38	0,00	1,13	90,00	0,00	1,38	73,33	0,00	6	
AND Ss L:1	1,88	81,25	0,01	1,53	88,33	0,00	1,88	85,00	0,00	52	233
	1,51	50,00	0,01	1,26	76,92	0,00	1,51	66,67	0,00	6	
AND Si L:1	1,78	89,58	0,00	1,43	96,67	0,00	1,78	93,33	0,00	52	23
	1,38	52,38	0,00	1,13	90,00	0,00	1,38	73,33	0,00	6	
AND Sm L:1	3,34	63,44	0,04	1,72	65,00	0,00	2,46	66,88	0,01	52	19
	2,76	23,16	0,04	1,26	50,00	0,00	2,14	51,52	0,01	8	

Valutazione delle prestazioni (11/15)

Variazione parametro *multimatch_factor (orMatches)*

- Per l'operatore OR, un aumento contenuto di questo parametro (inferiore a 100) causa un leggero aumento di precisione e richiamo solo nei primi 5/10 risultati
- Per l'operatore AND non si ottiene nessuna variazione degli indicatori in quanto i documenti reperiti presentano già tutte le parole specificate nell'interrogazione
- I tempi subiscono lievi variazioni, dovute principalmente a fattori estranei all'IRS (memoria, disco...)

Valutazione delle prestazioni (12/15)

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
OR Sp L:3	84,12	1,51	24,03	19,37	24,06	0,12	25,40	18,59	0,25	52	57
Mf:10	82,91	1,33	24,02	9,67	24,06	0,12	14,95	18,59	0,25	64	
OR Sp L:3	84,12	1,51	24,03	19,22	23,75	0,12	25,02	18,44	0,26	52	56
Mf:100	82,91	1,33	24,02	9,55	23,75	0,12	14,82	18,44	0,26	64	
AND Sp L:3	1,78	89,58	0,00	1,43	96,67	0,00	1,78	93,33	0,00	52	25
Mf:10	1,38	52,38	0,00	1,13	90,00	0,00	1,38	73,33	0,00	6	
AND Sp L:3	1,78	89,58	0,00	1,43	96,67	0,00	1,78	93,33	0,00	52	24
Mf:100	1,38	52,38	0,00	1,13	90,00	0,00	1,38	73,33	0,00	6	

Valutazione delle prestazioni (13/15)

Configurazioni a parametri misti con operatore OR

- L'effetto delle due stoplist utilizzate, nonostante siano combinate con diversi algoritmi di espansione, è essenzialmente quello che si ottiene facendole agire con il solo match esatto
- All'aumentare del numero di algoritmi utilizzati (> 2) si assiste a:
  - un maggior richiamo globale (fino ad oltre il 90%), a scapito di un progressivo peggioramento della precisione
  - un diminuzione del richiamo nei primi 5/10 documenti in quanto vengono reperiti più documenti non rilevanti (incremento del fallout)
  - una dilatazione dei tempi di risposta

Valutazione delle prestazioni (14/15)

Configurazioni a parametri misti con operatore AND

- La tecnica che garantisce migliori risultati dal punto di vista del richiamo consiste nell'adozione della stoplist ricavata dalle interrogazioni associata a più algoritmi di espansione
- All'aumentare del numero di algoritmi utilizzati (> 2) si assiste a:
  - quasi un raddoppio sia del richiamo globale, sia del numero di interrogazioni soddisfatte (fino a 11)
  - Un peggioramento molto evidente della precisione (fino ad oltre il 40%)
  - una dilatazione dei tempi di risposta, comunque più contenuta rispetto all'utilizzo dell'operatore OR
- La stoplist trovata su Internet garantisce un buon compromesso tra perdita di precisione (< 30%) ed aumento del richiamo

Valutazione delle prestazioni (15/15)

Configurazione	R	P	F	R_5	P_5	F_5	R_{10}	P_{10}	F_{10}	N	t_{avg}
OR Si L:1	85,00	1,68	20,51	22,45	25,94	0,12	27,87	19,06	0,25	52	50
sy:0,2	82,54	1,54	20,50	10,43	25,94	0,12	15,33	19,06	0,25	64	
OR Sm L:1	83,21	3,93	17,08	22,38	28,25	0,11	28,03	21,11	0,24	52	41
sy:0,2	80,53	1,80	17,07	10,68	27,33	0,11	15,58	19,97	0,24	63	
OR Sm L:1	88,52	3,49	22,89	22,02	26,98	0,11	27,36	20,32	0,25	52	63
sy:0,2 en:0,1	87,56	1,47	22,88	10,18	26,05	0,11	14,95	19,16	0,25	63	
OR Si L:1	92,59	0,96	39,76	14,29	12,50	0,14	19,60	11,09	0,28	52	131
sy:0,2 en:0,2 so:0,	91,46	0,89	39,75	5,03	12,50	0,14	8,92	11,09	0,28	64	
OR Sm L:1	90,48	1,64	27,35	16,74	15,94	0,13	22,58	13,28	0,27	52	86
sy:0,2 en:0,2 so:0.	88,82	1,25	27,34	6,41	15,94	0,13	10,68	13,28	0,27	64	
AND Si L:1	1,81	83,93	0,01	1,46	90,00	0,00	1,81	87,14	0,00	52	26
sy:0,2	1,51	52,17	0,01	1,26	83,33	0,0	1,51	70,59	0,00	7	
AND Sm L:1	3,38	61,94	0,04	1,76	63,33	0,01	2,50	65,00	0,01	52	21
sy:0,2	2,89	23,71	0,04	1,38	50,00	0,01	2,26	51,43	0,01	9	
AND Sm L:1	3,63	52,85	0,04	1,90	54,85	0,01	2,30	53,94	0,01	52	36
sy:0,2 en:0,1	3,27	23,42	0,04	1,51	44,44	0,01	2,14	40,48	0,01	11	
AND Si L:1	2,27	63,22	0,01	1,57	69,38	0,00	2,10	68,12	0,00	52	85
sy:0,2 en:0,2 so:0,	2,01	42,11	0,01	1,51	70,59	0,00	1,88	68,18	0,00	8	
AND Sm L:1	3,78	52,30	0,05	2,05	55,45	0,01	2,27	52,60	0,01	52	52
sy:0,2 en:0,2 so:0,	3,39	21,77	0,05	1,63	44,83	0,01	2,14	36,96	0,01	11	

Domande, chiarimenti... ?