UNIVERSITÀ DEGLI STUDI DI UDINE

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Specialistica in Informatica

Tesi di Laurea

UNA METRICA DI VALUTAZIONE PER L'INFORMATION RETRIEVAL: ANALISI CRITICA E SPERIMENTAZIONI

Relatore:
Dott. STEFANO MIZZARO

Laureando:
GIANLUCA DEMARTINI

Correlatori:
Dott. VINCENZO DELLA MEA
Dott. LUCA DI GASPERO

ANNO ACCADEMICO 2004-2005

A Monica,
mio padre e mia madre

Indice

Capitolo 1

Introduzione

Questa tesi ha l'obiettivo di valutare concettualmente e sperimentalmente una nuova metrica per l'efficienza dei sistemi per il reperimento delle informazioni. In questo capitolo viene dapprima introdotto l'argomento delle metriche di valutazione (paragrafo 1.1), vengono poi delineati gli obiettivi (paragrafo 1.2) ed infine viene presentata la struttura della tesi (paragrafo 1.3).

1.1 Le metriche di valutazione per il reperimento delle informazioni

Il campo del Reperimento delle Informazioni (RI) ha avuto notevoli sviluppi negli ultimi anni, anche e soprattutto in conseguenza del clamoroso successo di internet. Si sono quindi sviluppati e consolidati diversi Sistemi per il Reperimento delle Informazioni (SRI) sul web, come ad esempio Google [24]. Diventa perciò molto importante avere a disposizione degli strumenti per valutare l'efficacia degli SRI al fine di comprendere e confrontare tra loro i diversi SRI che sono stati prodotti.

Nella letteratura scientifica sono state proposte numerose metriche per valutare l'efficacia degli SRI come vedremo nei prossimi capitoli, attualmente vi sono 45 metriche e quasi ogni anno, in occasione delle iniziative internazionali di valutazione, ne vengono proposte di nuove. Solamente nell'iniziativa INEX 2005 sono state proposte ed utilizzate 5 nuove metriche progettate appositamente per valutare il RI su documenti in formato XML. Questo fa comprendere come il campo della valutazione degli SRI sia, di questi tempi, un tema molto caldo per la comunità scientifica del RI.

Tra tutte le metriche presenti ne è stata definita una chiamata Average Distance Measure (ADM) che ha come caratteristica fondamentale, e si differenzia dalle altre per questo, il modo in cui considera il concetto di pertinenza, sia quella percepita dall'utente sia quella assegnata dall'SRI. ADM considera la pertinenza una misura nel continuo a differenza di molte metriche classiche che ritengono questa dimensione binaria [41].

Fino ad ora la metrica ADM è stata valutata solamente in modo parziale e, per questo motivo, non si capisce ancora appieno qual è il suo vero potenziale. Il motivo per cui non è facile valutare sperimentalmente l'efficacia di ADM è la mancanza di dati di pertinenza e di reperimento di tipo continuo. Per questo motivo è necessario effettuare delle approssimazioni e considerare il calcolo di ADM su giudizi di pertinenza e dati di reperimento di tipo binario o al più a categorie, e quindi comunque discreto. Sono a disposizione della comunità scientifica collezioni di test con giudizi di pertinenza a 3 o 4 livelli e dati di reperimento che danno informazioni solamente sull'ordinamento con cui i documenti sono stati reperiti e non una stima sull'effettiva pertinenza del singolo documento.

1.2 Obiettivi della tesi

In questa tesi si intende affrontare il problema della valutazione degli SRI analizzando ADM valutandone sperimentalmente la sua efficacia e confrontandola con le metriche correntemente usate in tutti gli ambiti di valutazione nel campo del RI. Si vuole fare ciò proponendo una metodologia standardizzata di confronto che sia possibile riutilizzare ogni volta che si vuole effettuare una valutazione sperimentale della metrica.

Gli obiettivi di questa tesi sono dunque:

- effettuare un'analisi critica di ADM collocandola in una classificazione delle metriche proposte in letteratura basata sulle diverse definizioni possibili di pertinenza e di reperimento;
- effettuare delle valutazioni sperimentali sull'efficacia della metrica ADM al fine di comprendere meglio le sue prestazioni;
- proporre una metodologia standardizzata che può essere adottata nuovamente per ogni collezione di test su cui si desidera valutare l'efficacia di ADM;
- dopo aver compreso meglio le caratteristiche e le peculiarità della metrica ADM si vuole proporre una sua estensione che si adatti meglio al reperimento di documenti XML.

Alcuni dei risultati ottenuti nella tesi sono stati raccolti nei due lavori [15, 18] sottoposti per la pubblicazione al convegno *28th European Conference on Information Retrieval* (ECIR06) [19].

1.3 Struttura della tesi

Questa tesi è strutturata in due parti. La prima descrive alcuni concetti fondamentali come il concetto di SRI, la pertinenza e il processo di valutazione, presenta le caratteristiche delle metriche presenti in letteratura, effettua alcune considerazioni

sull'efficacia delle metriche e descrive ADM. La seconda parte presenta una classificazione originale, assente in letteratura, delle metriche di valutazione, descrive e discute i risultati ottenuti dalle sperimentazioni su ADM, definisce ed utilizza nelle sperimentazioni le curve ADP/ADR, che sono l'analogo delle curve Precision/Recall utilizzando concetti di pertinenza e reperimento continui, valuta e confronta sperimentalmente il tasso d'errore della metrica ADM, che indica quanto stabile è la metrica, e definisce e valuta preliminarmente una possibile estensione di ADM nel caso di pertinenza a due dimensioni quali l'esaustività e la specificità.

In particolare nel capitolo 2 viene data una definizione di cosa si intende per SRI, si definisce il concetto di pertinenza e si descrive il processo di valutazione di un SRI. In questo capitolo viene inoltre descritto come il processo di valutazione viene messo in opera in diverse iniziative internazionali di valutazione.

Nel capitolo 3 vengono descritte le 45 metriche proposte finora in letteratura, suddivise in metriche classiche, metriche orientate all'utente, metriche alternative alle precedenti e metriche per la valutazione del reperimento di documenti XML. Sono trattate anche le metriche adottate nell'iniziativa di valutazione INEX 2005 (tutt'ora in corso di svolgimento), proposte pochi mesi fa.

Nel capitolo 4 vengono descritti alcuni concetti che permettono di valutare l'efficacia delle metriche di valutazione e che sono necessari per effettuare un'analisi critica delle metriche di valutazione. Vengono definiti i concetti di sensitività e specificità mostrando come delle nozioni del campo dell'Informatica Medica siano analoghi ai concetti utilizzati nel campo del RI. Viene poi illustrato un metodo per valutare la stabilità delle metriche di valutazione calcolando il tasso d'errore commesso nel definire un SRI migliore rispetto ad un altro. Infine vengono discussi dei modelli di distribuzione dei valori di pertinenza nelle collezioni di documenti utilizzate per valutare l'efficacia degli SRI.

Nel capitolo 5 viene definita formalmente la metrica ADM e vengono presentati i risultati delle prime valutazioni sperimentali fatte su ADM su due diverse collezioni di test.

La seconda parte, in cui sono presentati e discussi i risultati ottenuti, inizia con il capitolo 6, nel quale vengono richiamati i concetti definiti nella prima parte della tesi e descritti gli obiettivi e la metodologia utilizzata per le valutazioni sperimentali di ADM.

Nel capitolo 7 viene proposta una classificazione originale delle 45 metriche di valutazione basata sulla nozione di pertinenza (se e quanto un documento è pertinente) e di reperimento (se e quanto un documento è reperito) che le metriche utilizzano, per comprendere meglio il potenziale di ADM;

Nei tre capitoli successivi vengono presentati i risultati delle diverse sperimentazioni effettuate utilizzando tre diverse collezioni di test. I risultati vengono analizzati e discussi per effettuare un'analisi critica della metrica ADM. Nel capitolo 8 viene usata la collezione di test TREC8 unitamente ai giudizi di pertinenza su 4 livelli effettuate da Sormunen [52]. Nel capitolo 9 viene usata la collezione di test TREC13 TeraByte. Viene inoltre calcolato il tasso d'errore delle metriche utilizzando la me-

desima collezione di test. Nel capitolo 10 vengono descritti i problemi riscontrati nell'utilizzo della metrica ADM per valutare l'efficacia degli SRI partecipanti all'iniziativa INEX 2004. Viene quindi definita una possibile estensione della metrica ADM ed effettuata una valutazione sperimentale preliminare di questa metrica.

Il capitolo 11, infine, conclude la tesi riassumendo il lavoro svolto e delineando i possibili sviluppi futuri.

Parte I

Prerequisiti

Capitolo 2

La valutazione dei sistemi per il reperimento delle informazioni

In questo capitolo vengono introdotti i concetti di SRI (paragrafo 2.1), pertinenza (paragrafo 2.2) e valutazione degli SRI (paragrafo 2.3). Nel paragrafo 2.4 viene mostrato come il processo di valutazione viene effettivamente messo in pratica in diverse iniziative internazionali organizzate per valutare l'efficacia degli SRI.

2.1 I Sistemi per il Reperimento delle Informazioni

Dovendo trattare la valutazione degli SRI è necessario innanzitutto dare una definizione di cosa si intende per SRI. Un SRI è progettato per svolgere al meglio il processo di RI. Esso dovrà portare a termine al meglio tutti i passi di questo processo (si veda la figura 2.1).

Inizialmente, prima che il processo di reperimento abbia inizio, è necessario disporre di una base di dati testuale. È quindi necessario specificare i documenti da utilizzare, le operazioni da effettuare sul testo per rendere più veloce la ricerca dei documenti ed il modello (ad esempio un tipo di struttura particolare che ha il testo e quali sono gli elementi che possono essere reperiti). Ai documenti vengono applicate diverse procedure per semplificare i termini che contengono. Vengono eliminate le "stopwords", cioè i termini molto frequenti nella lingua in cui è scritto il testo che non sono quindi significativi per capire l'argomento trattato dal testo (ad esempio gli articoli e le congiunzioni). Inoltre, per i termini rimasti, vengono estratte le radici mediante algoritmi di *stemming* (ad esempio l'algoritmo di Porter [2]) per non fare distinzioni tra singolari, plurali, verbi coniugati. Queste operazioni permettono di ottenere una visione logica dei documenti originali.

Una volta che è stata definita la visione logica dei documenti, si costruisce un indice del testo. Un indice è la struttura dati principale per un SRI in quanto esso permette una rapida ricerca all'interno di grosse quantità di dati. Le risorse, in termini di tempo e spazio di memorizzazione, impiegate nella fase di indicizzazione

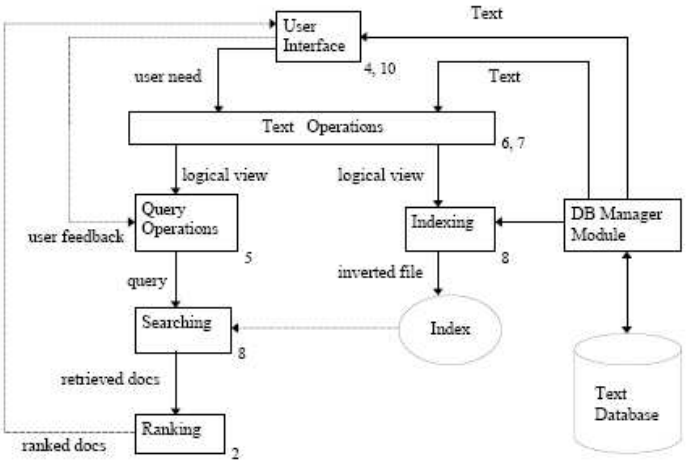


Figura 2.1: Il processo di reperimento messo in opera da un SRI (da [2])

sono giustificate dalle minor risorse necessarie per le (notevolmente più numerose) operazioni di interrogazione.

Data una collezione indicizzata, il processo di reperimento può essere svolto nel modo seguente. L'utente specifica il suo bisogno informativo che viene trasformato nello stesso modo in cui sono stati trasformati i documenti presenti nella collezione. Vengono cioè estratte le radici dei termini e inseriti termini simili a quelli specificati dall'utente. A questo punto l'interrogazione viene processata per ottenere quelli che sono i documenti reperiti. Come visto, un reperimento dei documenti rapido è possibile grazie alla precedente fase di indicizzazione.

Prima che i documenti vengano mostrati all'utente, l'SRI li ordina confrontandoli e stimando quali documenti sono più pertinenti rispetto ad altri assegnando ad ognuno un *Retrieval Status Value* (RSV) detto anche *score*. Questi valori assegnati vengono poi trasformati in un ordinamento (rank) che viene presentato all'utente. A questo punto l'utente analizza l'insieme dei documenti che sono stati reperiti dall'SRI ed indica ad esso quali documenti sono, a suo giudizio, pertinenti all'interrogazione. In questo modo l'SRI può eseguire delle operazioni di *relevance feedback* per riformulare l'interrogazione al fine di renderla più corretta.

2.2 La Pertinenza

I documenti reperiti da un SRI sono quelli che vengono giudicati pertinenti da esso. Per essere in grado di valutare quanto bene un SRI svolge il compito per cui è stato progettato, è necessario capire cosa si intende con il termine *pertinenza* (in inglese *relevance* [23]):

> "The capability of a search engine or function to retrieve data appropriate to a user's needs." da Answers.com

Il concetto di pertinenza è fondamentale e sta alla base di tutti gli aspetti del RI [6, 39, 40, 50]. È un concetto difficile da definire con precisione: ancora oggi non è del tutto chiaro e compreso a fondo e per questo rimane uno dei principali problemi del campo del RI. Si potrebbe definire la pertinenza come "la proprietà di un dato documento (inteso come testo, immagine, pezzo musicale, ecc.) di soddisfare il bisogno informativo dell'utente". Questa interpretazione non risolve il problema ma lo sposta sulla definizione del concetto di "bisogno informativo". Si intende per bisogno informativo la necessità di informazione mancante che l'utente sa di non possedere e che non sempre riesce a esprimere con chiarezza a parole o tramite parole chiave in una interrogazione all'SRI. Il documento dovrebbe essere definito pertinente nel momento in cui soddisfa questo tipo di bisogno.

Sperber e Wilson [54] hanno proposto un ulteriore punto di vista, più cognitivo, di quello che dovrebbe significare "pertinenza". Il principio di pertinenza proposto punta l'attenzione sulla comunicazione: i comportamenti dovrebbero essere svolti tenendo conto che il mittente crede di fornire informazioni pertinenti per il destinatario. Quindi il destinatario dovrebbe porre attenzione al messaggio al fine di aumentare le proprie conoscenze. L'informazione più pertinente che ci possiamo aspettare dovrà avere due proprietà: dovrà essere *nuova*, altrimenti non abbiamo migliorato il nostro stato di conoscenza. Ma dovrà anche essere *connessa* ad altre informazioni, altrimenti il fattore novità non aggiungerà praticamente nulla di interessante. Sperber e Wilson pongono l'attenzione anche sull'importanza del contesto entro cui la comunicazione avviene. Un aspetto di questo contesto è la *mutua conoscenza* che il mittente ed il destinatario devono avere, in modo che la comunicazione avvenga efficientemente. La mutua conoscenza è quella conoscenza k che è nota sia al mittente che al destinatario. Nel campo del RI l'atto comunicativo fondamentale che garantisce la mutua conoscenza è quello conosciuto con il nome di "relevance feedback".

2.2.1 Pertinenza binaria

Abbiamo cercato di descrivere diversi modi per definire il concetto di pertinenza; ora ci occupiamo di descrivere come assegnare ad un dato documento un *valore numerico* che misuri la pertinenza rispetto al bisogno informativo.

Storicamente, la prima proposta, che è anche la prima che viene in mente, è stata di tipo binario. L'utente trasforma il proprio bisogno informativo in una richiesta informativa, in modo tale che l'SRI possa utilizzarla per reperire dei documenti. Ad un documento viene assegnato un valore binario in base alla sua pertinenza rispetto una certa richiesta informativa: 1 se esso risulta essere pertinente, 0 nel caso contrario. Quindi è necessario, da parte degli esperti che devono giudicare un documento per una collezione, decidere, in base al contenuto, se un certo documento contiene informazioni che riguardano quello che si sta cercando oppure no. Quest'idea è ba-

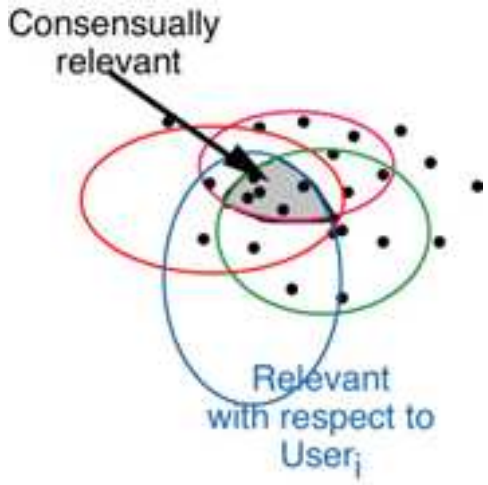


Figura 2.2: Pertinenza consensuale (da [3])

sata sul fatto che, storicamente, i primi SRI dovevano reperire un certo documento oppure no, e quindi dovevano compiere una valutazione di tipo binario sulla pertinenza di quel documento rispetto l'interrogazione effettuata dall'utente [41]. Infatti, le misure classiche di valutazione degli SRI sono basate sul concetto di pertinenza binaria e di reperimento binario.

In questo caso vale l'assunzione che ci si fida di un singolo esperto in grado di fornire valutazioni di pertinenza affidabili. Un diverso punto di vista della pertinenza di tipo binario è dato dall'idea di "pertinenza consensuale" che propone di utilizzare l'opinione di diversi utenti combinando la valutazione di molteplici giudici umani [21]. Piuttosto che avere una singola valutazione booleana di pertinenza data da un esperto, si potrebbe considerare la pertinenza come un'opinione consensuale degli utenti degli SRI. Se gli utenti considerano un certo documento pertinente allora lo è, anche se qualche esperto del settore sostiene che il documento non dovrebbe essere reperito. In figura 2.2 è riportato un esempio di pertinenza consensuale rappresentato come una serie di insiemi ognuno dei quali rappresenta i giudizi di pertinenza di un singolo utente. Gli elementi dell'insieme rappresentano i documenti giudicati pertinenti dal singolo utente, quindi i documenti contenuti nell'intersezione di tutti gli insiemi saranno i documenti consensualmente pertinenti.

Negli anni si sono sviluppati diversi modelli, tra i quali citiamo il *vector space* [47] e modelli probabilistici [44], che hanno portato all'implementazione di SRI non binari che assegnano ad ogni documento dell'insieme una misura pesata di similarità (la RSV) con l'interrogazione. Sulla base di questo valore è possibile classificare i documenti in livelli di similarità decrescente. Nel prossimo paragrafo vedremo valutazioni di pertinenza che non dividono in due l'insieme dei documenti ma assegnano ad ogni documento una misura pesata.

2.2.2 Pertinenza non binaria

Non avendo una definizione certa e formale del concetto di pertinenza, non si può sapere se la pertinenza è di tipo binario. Infatti non è escluso che un utente possa vedere un documento reperito non soltanto del tutto attinente alla ricerca da lui fatta o per nulla attinente, ma è possibile che ci sia una scala di valori di pertinenza in cui un documento si possa trovare. Potrebbe essere addirittura vero che esistano infiniti livelli di pertinenza con i quali un documento possa essere classificato.

Un caso pratico di pertinenza non binaria utilizzato nelle iniziative internazionali di valutazione è quello che prevede quattro distinti livelli di pertinenza. Ad esempio in NTCIR-4 (NII-NACSIS Test Collection for IR Systems) [42] i giudici umani devono associare ai documenti reperiti uno dei seguenti valori:

- S (completamente pertinente)
- A (pertinente)
- B (parzialmente pertinente)
- C (non pertinente)

Non risulta comunque chiaro come associare dei valori numerici ai vari livelli così come accade nel caso di pertinenza binaria dove ad un documento pertinente viene associato il valore 1 e ad un documento non pertinente viene associato il valore 0. La tendenza è comunque quella di utilizzare intervalli di uguale misura. In [16], ad esempio, si utilizzano i valori $S = 7/8, A = 5/8, B = 3/8, C = 1/8$ al fine di avere degli intervalli di uguale misura. Un altro possibile assegnamento di valori potrebbe essere $S = 1, A = 2/3, B = 1/3, C = 0$.

Con l'utilizzo di più di due livelli sorge quindi il problema di definire la distanza tra i livelli. Nell'esempio precedente gli intervalli sono di ugual misura, ma questo non è detto che sia il modo giusto di assegnarli. Potrebbe, ad esempio, essere più corretto assegnare degli intervalli di grandezza crescente. L'uguale distanza tra i vari livelli di pertinenza è un'assunzione fatta dagli sperimentatori, ma al fine di assegnare le distanze tra i livelli si potrebbe ad esempio utilizzare una particolare distribuzione di valori in cui gli intervalli non hanno la stessa grandezza.

Nel caso limite, si potrebbe ottenere una pertinenza ad infiniti livelli in cui ad ogni documento si associa un valore r tale che $r \in \mathbb{R}$ e $r \in [0, 1]$.

Oltre al problema di esplicitare con chiarezza il bisogno informativo dell'utente, un ulteriore concetto che complica la definizione di pertinenza è la sua visione su più dimensioni, come avviene in certi congressi di valutazione degli SRI.

Ad esempio è possibile considerare la pertinenza come un concetto composto da due separati elementi: la *specificità* e l'*esaustività*. In INEX [22] si classificano i documenti su due scale diverse, a 4 livelli ciascuna, per quanto riguarda questi due concetti. Per *specificità* si intende quanto il singolo elemento reperito sia focalizzato sull'argomento, ovverosia se parli solamente e in particolare di ciò che si è

richiesto. Per *esaustività* si intende quanto il singolo elemento reperito copra l'intero argomento, ovvero se parli completamente di ciò che si è richiesto.

2.3 Il processo di valutazione

Ora che abbiamo definito cosa si intende per SRI e descritto il concetto di pertinenza, è possibile descrivere come avviene il processo di valutazione di un SRI.

Nel momento in cui volessimo costruire uno strumento che reperisca informazioni, o nel momento in cui dovessimo scegliere quale strumento usare tra i diversi disponibili, sentiremmo il bisogno di poter valutare quale strumento è il migliore. Bisogna quindi stabilire una metodologia attraverso la quale valutare rigorosamente le prestazioni di un sistema.

Esistono due diversi approcci alla valutazione degli SRI. La principale metodologia di valutazione è l'esecuzione automatica di test utilizzando delle "collezioni di test". L'altro approccio è la scelta di non automatizzare il processo di valutazione effettuando degli studi utente. Le collezioni di test sono formate da un insieme di documenti, detto anche "document collection", da un insieme di interrogazioni da fare all'SRI e da un insieme di giudizi di pertinenza per ogni documento su ogni interrogazione.

È possibile anche effettuare degli esperimenti nella vita reale degli utenti utilizzando delle situazioni di RI che accadono senza il bisogno di crearle artificialmente nei laboratori. Comunque la metodologia attualmente più utilizzata per effettuare valutazioni degli SRI è di usare delle sperimentazioni in laboratorio con metodi automatici o non automatici [2, Cap. 3.2].

2.3.1 Ricostruzione storica della valutazione

Prima di analizzare l'attuale processo di valutazione di un SRI, vediamo come tale processo si è evoluto nel tempo [3, Cap. 4].

A causa delle limitate dimensioni dei dischi e della lentezza dei computer nel passato, inizialmente le valutazioni venivano svolte in modo diverso da oggi. Gli insiemi di documenti su cui si svolgevano gli esperimenti erano necessariamente di dimensioni limitate. Tuttavia, il fatto di avere un piccolo insieme di documenti, permetteva di avere interrogazioni che potevano essere confrontate con ogni parte di ogni documento a disposizione. Alcuni dei primi esperimenti effettuati sono stati i "Cranfield experiments" [36] svoltisi nel 1968, in cui 1400 documenti di metallurgia, sono stati ricercati rispetto a 221 interrogazioni costruite da alcuni degli autori dei documenti. Un importante contributo è stato dato dagli esperimenti di Salton, svolti nel 1980, in cui 82 articoli, pubblicati nel 1963, sono stati reperiti attraverso 35 interrogazioni ed i risultati del reperimento sono stati valutati da studenti ed esperti [47].

È anche possibile effettuare un campionamento di un piccolo sottoinsieme del corpo dei documenti da reperire e in seguito farlo analizzare da un gruppo di valu-

tatori umani. Questa operazione permette di estrapolare dal numero di documenti pertinenti trovati nel sottoinsieme il numero atteso di documenti pertinenti sull'intera collezione. Questa tecnica è stata utilizzata da Blair e Maron con l'SRI *STAIRS* [4] proposta dall'IBM nei primi anni '80.

Con l'aumentare delle capacità dei calcolatori, le valutazioni sono state eseguite su collezioni di documenti di maggiori dimensioni. Un nuovo standard per la valutazione degli SRI è stato imposto dall'iniziativa TREC, che si è tenuta per la prima volta nel 1992. Per prima cosa si evita di valutare l'intera collezione di documenti utilizzando la metodologia nota con il nome di *pooling*. L'idea di base è di usare ogni SRI partecipante indipendentemente dagli altri, e quindi mettere assieme i risultati di tutti gli SRI al fine di formare un insieme di documenti che hanno almeno una possibilità di essere pertinenti. Tutti gli SRI reperiscono documenti fornendo una lista ordinata di k documenti potenzialmente pertinenti. I primi n ($n < k$) documenti reperiti da ogni SRI vengono uniti in un unico insieme che compone il pool. Questi documenti sono presentati ad alcuni giudici umani al fine di effettuare i giudizi di pertinenza. Con questa tecnica solamente i documenti reperiti da almeno uno degli SRI sono sottoposti alla valutazione di un giudice umano. Esiste quindi la possibilità che un documento pertinente venga valutato come non pertinente. Infatti vale la regola per la quale i documenti non valutati dai giudici umani sono considerati non pertinenti.

2.3.2 Esecuzione automatica di valutazioni

Dopo aver descritto l'evoluzione che ha avuto il processo di valutazione nel tempo, vediamo le varie metodologie con cui attualmente avvengono le valutazioni degli SRI.

Come visto, i due principali approcci alla valutazione degli SRI sono l'esecuzione automatica di test e l'effettuazione di studi utente. I metodi automatici, sui quali poniamo attenzione in questa tesi, prevedono di costruire delle collezioni di test (in inglese *test collection* [23]) che sono composte da un insieme di documenti e da un insieme di richieste collegate all'elenco di documenti pertinenti nella collezione relativamente alla singola richiesta. A questo punto, per confrontare le prestazioni di diversi SRI, si fa eseguire ad ognuno la ricerca dei documenti per ogni richiesta e si misurano le prestazioni utilizzando le metriche tipiche del RI.

2.3.3 Esecuzione non automatica di valutazioni

L'altra metodologia di valutazione consiste nell'eseguire degli studi utente [45]. In questo caso è necessario avere degli utenti reali a cui si richiede di utilizzare un SRI. Durante e in seguito all'utilizzo si misurano le prestazioni dell'SRI e la soddisfazione dell'utente mediante la misurazione di tempi e la compilazione di questionari. Queste valutazioni vengono svolte in laboratorio utilizzando dei bisogni informativi realistici indotti o reali. È facile capire che l'utilizzo di utenti reali durante la valutazione di un SRI ha dei vantaggi rispetto all'esecuzione automatica di test. Infatti, avendo

a disposizione utenti reali è possibile capire meglio se un documento soddisfa il bisogno informativo dell'utente, mentre nelle collezioni di test si utilizza il giudizio di pertinenza di esperti esterni, non dell'utente. Un problema reale degli studi utente è, invece, oltre al maggior costo in termini di risorse, le difficoltà in termini di ripetitibilità. Se lo studio utente non è eseguito in modo corretto, è possibile che, nel caso in cui si ripeta l'esperimento in un momento successivo, i risultati non siano gli stessi.

2.4 Le iniziative internazionali di valutazione

Come già accennato, in questa tesi verrà trattata esclusivamente la metodologia di valutazione automatica. Le iniziative internazionali di valutazione utilizzano questa metodologia al fine di avere, annualmente, dei risultati da analizzare.

Esistono numerose iniziative internazionali di valutazione che hanno come obiettivo quello di mettere a disposizione dei partecipanti delle "collezioni di test" al fine di valutare l'efficacia degli SRI che vi partecipano. Le collezioni di test sono composte da una collezione di documenti, un insieme di interrogazioni e, per ogni interrogazione, un insieme di documenti pertinenti (utilizzando eventualmente diversi tipi di pertinenza).

Dato che negli ultimi anni le collezioni di documenti hanno raggiunto dimensioni notevoli, diventa difficile (o impossibile) riuscire a valutare il grado di pertinenza di tutti quanti i documenti e quindi riuscire a trovare tutti i documenti pertinenti all'interno della collezione. Per venire incontro a questo problema è stata introdotta la tecnica del *pooling*. Come abbiamo visto, essa consiste nel prendere i primi n documenti reperiti dagli SRI e far valutare solo questi dagli esperti umani. Questa tecnica, quindi, presuppone il fatto che se un documento non è reperito da nessun SRI allora non è pertinente. La pertinenza dei documenti nel pool viene valutata mentre gli altri documenti sono giudicati non pertinenti.

Come visto, l'organizzazione mette a disposizione dei partecipanti le collezioni di documenti che, a seconda del tipo di iniziativa, varia nella tipologia. Vengono messi a disposizione articoli scientifici, articoli di settori ed argomenti specifici, documenti XML, ma anche filmati e file con contenuto multimediale.

Oltre alla collezione di documenti vengono distribuiti i giudizi di pertinenza. Ad ogni documento viene associato un valore di pertinenza, che può essere 0 o 1 nel caso si utilizzi una pertinenza binaria, una categoria di pertinenza in caso di pertinenza a più livelli o anche più di un valore nel caso di pertinenza definita su più di una dimensione (si veda il paragrafo 2.2).

Al termine dell'iniziativa, solitamente annuale, l'organizzazione rende pubblici ai partecipanti i risultati. Vengono calcolati per tutti gli SRI partecipanti le prestazioni ottenute sulla collezione di test utilizzando diverse metriche di valutazione ed i relativi grafici (come ad esempio le curve precision/recall).

Tipicamente le iniziative di valutazione forniscono ai loro partecipanti diverse categorie in cui è possibile partecipare, in modo tale che tutti i tipi di SRI possano

trovare il modo di testare le proprie prestazioni e di confrontarsi con altri SRI. Queste categorie solitamente prevedono specifiche collezioni di test per SRI multimediali, per sistemi di Question-Answering, per SRI progettati per collezioni di documenti di dimensioni elevate.

Possiamo riassumere dicendo che solitamente un'iniziativa di valutazione mette a disposizione della comunità scientifica numerose risorse. Queste sono la collezione di documenti su cui viene effettuato il reperimento, un insieme di interrogazioni di reperimento, le classificazioni effettuate dagli SRI partecipanti dei vari documenti sulle varie interrogazioni, i giudizi di pertinenza dei documenti sulle interrogazioni ed, infine, i risultati di efficacia degli SRI misurati con certe metriche di valutazione scelte opportunamente. Nei paragrafi seguenti si presenteranno alcune iniziative di valutazione che vengono svolte annualmente.

2.4.1 Text REtrieval Conference (TREC)

La prima iniziativa di valutazione, in termini storici, è TREC. Nel 1992 viene tenuta la prima "Text REtrieval Conference" [56], facente parte del "TIPSTER Text program", co-sponsorizzata dal "National Institute of Standards and Technology" (NIST) e dal dipartimento della difesa degli Stati Uniti d'America. Lo scopo di TREC è quello di supportare la ricerca nel campo del RI fornendo le infrastrutture necessarie per valutazioni su larga scala di metodologie di reperimento di testi. In particolare, i congressi di TREC si pongono i seguenti obiettivi:

- incoraggiare la ricerca nel reperimento delle informazioni basata su collezioni di grandi dimensioni;
- incrementare la comunicazione tra industrie, università e governo al fine di creare un forum per lo scambio di idee;
- velocizzare il trasferimento tecnologico dai laboratori di ricerca verso i prodotti commerciali, dimostrando quali sono i miglioramenti ottenuti nelle metodologie di reperimento sui problemi del mondo reale;
- incrementare la disponibilità di appropriate tecniche di valutazione alle industrie e alle università, compreso lo sviluppo di nuove tecniche di valutazione più adatte ai sistemi odierni.

TREC è supervisionato da un comitato che comprende rappresentanti del governo americano, dell'industria e delle università. Per ogni edizione di TREC, il NIST fornisce un insieme di documenti di prova e le relative interrogazioni. I partecipanti devono utilizzare i propri SRI su questi dati e fornire al NIST un elenco dei documenti meglio classificati. Il NIST elabora questi risultati e giudica i documenti reperiti. Il ciclo di TREC si conclude con un workshop in cui i partecipanti condividono le proprie esperienze.

I documenti presenti in una collezione dovrebbero riflettere le stesse caratteristiche (argomento, termini usati, stile letterario, formato) dei risultati dati dal

<num> Number: 409
<title> legal, Pan Am, 103

<desc> Description:
What legal actions have resulted from the destruction
of Pan Am Flight 103 over Lockerbie, Scotland, on
December 21, 1988?
<narr> Narrative:
Documents describing any charges, claims, or fines
presented to or imposed by any court or tribunal are
relevant, but documents that discuss charges made in
diplomatic jousting are not relevant.

Figura 2.3: Un topic di TREC (da [60])

reperimento in condizioni reali. In TREC, ad ogni documento usato viene assegnato un identificativo univoco ed il suo contenuto viene mantenuto il più possibile uguale a quello originale (non vengono corretti errori di sintassi, formattazioni poco consone, ecc.). TREC definisce con il termine *topic* il bisogno informativo che l'ipotetico utente necessita di soddisfare, descritto in linguaggio naturale, mentre per *query* si intende l'interrogazione strutturata che viene passata all'SRI per reperire dei documenti in grado di soddisfare il bisogno informativo dell'utente. Solitamente la struttura di un topic comprende quattro sezioni: identificativo, titolo, descrizione, e narrativa (si veda la figura 2.3 per un esempio).

TREC utilizza principalmente giudizi di pertinenza binari. Quindi un documento viene considerato pertinente o meno, senza vie di mezzo, riguardo ad un relativo topic. I giudizi di pertinenza sono in genere effettuati direttamente dagli autori dei singoli topic. Al fine di decidere se un documento va considerato pertinente o meno, l'organizzazione di TREC indica ai giudici di considerarsi dei potenziali autori di un articolo che riguarda il soggetto espresso nel topic e, se il singolo documento contiene qualche informazione utile allo scopo, l'intero documento va considerato pertinente a quel topic.

TREC utilizza la metodologia del pooling al fine di creare un sottoinsieme di documenti che devono essere giudicati. Ogni documento contenuto nel *pool* viene giudicato pertinente o meno dall'autore del singolo topic. I documenti non facenti parte del pool sono assunti come non pertinenti. Il pool viene creato inserendo al suo interno i documenti maggiormente reperiti dagli SRI partecipanti. Quindi si suppone che i documenti non inseriti, e quindi considerati dagli SRI non pertinenti, abbiano valore di pertinenza pari a zero.

TREC considera tipi di compiti o *task*, ossia modi di reperire le informazioni, diversi per ogni topic. Il principale task considerato è quello conosciuto sotto il nome "ad hoc". e corrisponde all'attività effettuata da un bibliotecario: l'ambiente di ricerca, cioè la collezione di documenti, è noto a priori, mentre il particolare argomento che si cerca viene reso noto solamente all'inizio della ricerca. Lo stesso

tipo di reperimento è svolto, ad esempio, da un navigatore di Internet che utilizza un motore di ricerca, da un avvocato che cerca precedenti nella giurisprudenza, e così via.

Un diverso tipo di task che TREC considera è quello chiamato *known-item search*. In questo caso l'obiettivo della ricerca è un particolare documento (o un piccolo insieme di documenti) di cui colui che sta effettuando la ricerca conosce l'esistenza.

TREC utilizza inoltre un ulteriore tipo di task, detto *filtering*, in cui il topic è noto, mentre la collezione di documenti è in continuo aggiornamento.

Abbiamo visto i diversi tipi di task che vengono considerati in TREC, passiamo ora alla valutazione dei task di tipo "ad hoc" viene effettuata dall'organizzazione utilizzando il pacchetto *trec_eval*. Esso calcola 85 differenti valori, tra cui le metriche di valutazione classiche.

Ogni edizione dell'iniziativa di valutazione TREC è composta da una serie di "track". Esse sono aree di interesse in cui sono definiti particolari compiti di reperimento. I track servono da incubatori per nuove aree di ricerca: la prima edizione di un track spesso serve principalmente per definire con chiarezza il problema e per fornire un'infrastruttura (collezioni di documenti, metodologie di valutazione, ...) a supporto della ricerca. I track inoltre rendono TREC interessante per le comunità di ricerca fornendo compiti che incontrano gli interessi di svariati gruppi di ricerca.

Ogni track ha una propria mailing list il cui scopo primo è di discutere i dettagli dei compiti legati ad un track. Comunque, la mailing list svolge anche il ruolo di luogo virtuale in cui discutere dei problemi metodologici relativi ai compiti di reperimento. L'insieme dei track che vengono proposti per una data edizione di TREC sono determinati dal comitato organizzatore di TREC.

Di seguito sono riportati i track di TREC 2005:

- *Enterprise Track*: il suo scopo è di studiare le ricerche aziendali. L'idea è che l'utente stia cercando informazioni riguardo ad un'organizzazione.
- *Genomics Track*: il suo scopo è di studiare il reperimento nel dominio di interesse dei dati genetici.
- *HARD Track*: il suo scopo è di raggiungere un elevato grado di accuratezza nel reperimento di documenti sfruttando informazioni addizionali riguardo colui che necessita del reperimento e il contesto in cui il reperimento si svolge.
- *Question Answering Track*: Questo track è stato progettato per dirigersi maggiormente al reperimento delle informazioni invece che al reperimento di documenti.
- *Robust Retrieval Track*: Questo track include il tradizionale compito di reperimento "ad hoc", rivolgendo però maggiormente l'attenzione sull'efficacia sui singoli topic piuttosto che l'efficacia media dei sistemi.
- *SPAM Track*: il suo scopo è di fornire una valutazione dei sistemi correnti di filtraggio della posta elettronica.

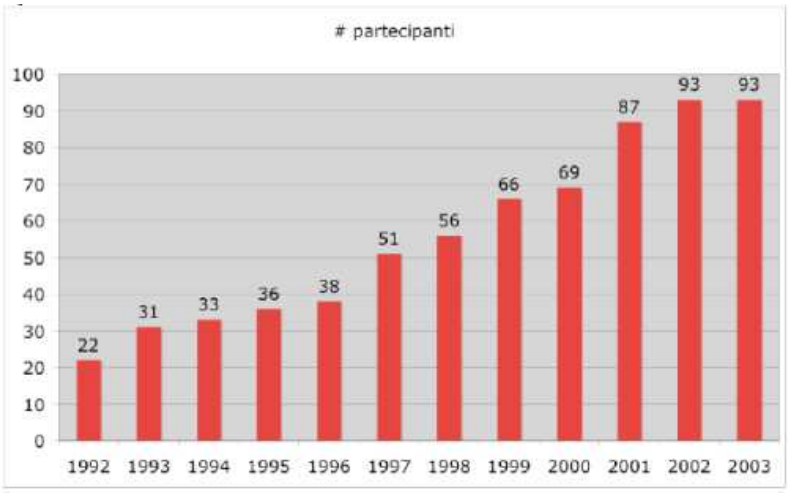


Figura 2.4: I partecipanti a TREC

- *Terabyte Track*: il suo scopo è di capire se e come è possibile trasferire la valutazione tradizionale a delle collezioni di documenti significativamente più grandi. La collezione di documenti ha, in questo caso, dimensioni di un terabyte di pagine web.

TREC è, ad oggi, la più importante iniziativa internazionale di valutazione e continuerà in futuro, annualmente, a valutare gli SRI che vi parteciperanno (si veda la figura 2.4 in cui è presente un istogramma con il numero di SRI partecipanti a TREC in ogni sua edizione).

2.4.2 NII-NACSIS Test Collection for IR Systems (NTCIR)

Un'altra iniziativa internazionale di valutazione di SRI è NTCIR che viene organizzata dal *National Center for Science Information Systems*. La principale caratteristica che contraddistingue NTCIR è il fatto che valuta il reperimento di documenti scritti in lingue orientali. NTCIR consiste in una serie di workshop di valutazione, congressi annuali in cui i ricercatori presentano i loro risultati, progettati per stimolare la ricerca in diversi campi tra cui quello del reperimento delle informazioni. Gli scopi principali di NTCIR sono:

- incoraggiare la ricerca delle tecnologie per l'accesso all'informazione fornendo una collezione di test di grandi dimensioni, per svolgere esperimenti, ed una infrastruttura di valutazione che permette confronti incrociati;

- mettere a disposizione un forum in cui i gruppi di ricerca possano scambiarsi idee in un'atmosfera informale;

- studiare i metodi di valutazione degli SRI e le tecniche per costruire insiemi di dati di notevoli dimensioni al fine di utilizzarli per sperimentazioni.

Fin dall'inizio del progetto NTCIR, si è posto lo sguardo sia sulla valutazione di SRI di tipo tradizionale, sia sulla valutazione delle tecnologie più innovative. NTCIR ha come caratteristica principale il fatto che concentra l'attenzione sul RI utilizzando documenti di lingue tra loro diverse (cinese, coreana, giapponese ed inglese). Inoltre si studiano le metodologie di valutazione in ambienti realistici e metodologie di valutazione con giudizi di pertinenza non binari.

Come visto, NTCIR ha come obiettivo quello di fornire una collezione di documenti che contenga una buona varietà di generi e di lingue al fine di fornire un'infrastruttura di valutazione quanto più possibile realistica. I documenti sono sia di tipo scientifico che non scientifico. È una collezione ben bilanciata tra documenti presi da articoli giornalistici di lingua cinese, coreana, giapponese ed inglese.

La struttura dei topic per le collezioni di test è simile a quella usata nelle iniziative di TREC. I topic sono definiti in linguaggio naturale come richieste che l'utente medio potrebbe fare. Queste stringhe vengono, dopo un'eventuale elaborazione, sottoposte agli SRI.

La scala utilizzata per i giudizi di pertinenza, a differenza di TREC e di altre iniziative internazionali di valutazione, è a quattro livelli: *highly relevant, relevant, partially relevant, irrelevant.*

Come in TREC, anche in NTCIR ci sono diversi tipi di task di reperimento a cui è possibile partecipare. Ad esempio nell'edizione NTCIR-4, erano presenti 5 diverse categorie:

- CLIR (Cross-Lingual Information Retrieval) composto da 3 sotto-task: Multilingual CLIR, Bilingual CLIR, Single Language IR
- PATENT (Patent Retrieval Task)
- QAC (Question Answering Challenge)
- TSC (Test Summarization Challenge)
- WEB (WEB Task)

Per valutare l'efficacia degli SRI partecipanti a NTCIR vengono usate metriche di valutazione diverse per ogni task.

2.4.3 Initiative for the Evaluation of XML Retrieval (INEX)

I contenuti di oggi stanno diventando sempre di più un miscuglio di testo, dati multimediali e metadati. Un modo di formattare questi contenuti è di utilizzare lo standard XML (eXtensible Markup Language), proposto dall'organizzazione W3C. L'utilizzo sempre maggiore del linguaggio XML ha portato alla nascita di diversi sistemi per la memorizzazione ed il reperimento di documenti di questo tipo. L'XML

offre l'opportunità di sfruttare la struttura logica del documento in modo da fornire un reperimento più preciso. In questo caso la pertinenza dovrà comprendere requisiti sia sul contenuto che sulla struttura.

La valutazione degli SRI che lavorano su documenti XML dovrà utilizzare specifici documenti e specifici giudizi di pertinenza. La Initiative for the Evaluation of XML Retrieval (INEX) [26], organizzata dal *DELOS Network of Excellence for Digtal Libraries*, si occupa dal 2002 di fornire questi strumenti ai progettisti degli SRI per documenti XML.

Per l'edizione 2005, INEX mette a disposizione una collezione di 12.107 documenti (in formato XML) della IEEE Computer Society proveniente da 6 riviste pubblicate nel periodo 1995-2002. Mediamente, ogni articolo contiene 1.532 nodi XML ed è ha una profondità di 6.9 livelli. Le interrogazioni sono proposte dai gruppi partecipanti e possono essere riferite solamente al contenuto (content-only, CO) oppure riferite sia al contenuto che alla struttura (content-and-structure, CAS).

Il principale task di reperimento utilizzato in INEX è quello "ad hoc". L'unica differenza è che le interrogazioni utilizzate possono contenere sia condizioni sul contenuto che sulla struttura dei documenti da reperire. Oltre agli altri task di reperimento, per il 2005 è prevista l'introduzione di un task di reperimento su dati multimediali (multimedia track) che utilizza la collezione "Lonely Planet".

I giudizi di pertinenza vengono effettuati utilizzando un sistema on-line. Questo sistema permette al giudice di evidenziare, per ogni documento, le parti giudicate almeno parzialmente pertinenti da un punto di vista della significatività. Il livello di significatività del frammento evidenziato viene stimato automaticamente. Successivamente il giudice deve valutare il livello di esaustività di ogni porzione di testo evidenziata. Il sistema assiste il giudice al fine di non permettergli di commettere errori di coerenza.

Ho usato questo sistema on-line per giudicare due topic al fine di ottenere dati utili per le valutazioni sperimentali qui presentate. Infatti l'organizzazione di INEX garantisce a coloro che giudicheranno i documenti di almeno 2 topic l'accesso all'intera collezione di documenti di INEX (altrimenti protetta da copyright).

2.4.4 Cross-Language Evaluation Forum (CLEF)

Un'ulteriore iniziativa internazionale di valutazione è CLEF [12]. Questa iniziativa di livello europeo, organizzata dal *DELOS Network of Excellence for Digtal Libraries*, ha come obiettivo la valutazione dell'efficacia degli SRI che utilizzano documenti di lingue europee diverse tra loro.

CLEF mira anche a stabilire delle relazioni con iniziative di valutazione multilingua che si tengono al di fuori del territorio europeo al fine di stimolare lo sviluppo di SRI europei e garantirgli un buon livello di competitività sul mercato mondiale.

Nel 2005 il track principale è composto da tre diversi task di reperimento: monolingua, bilingua (interrogazioni in una lingua e reperimento di documenti un un'altra lingua) e multilingua (interrogazioni in una lingua e reperimento di documenti in diverse altre lingue). Per i task monolingua e bilingua sono state utilizzate collezioni

di documenti in bulgaro, francese, portoghese e ungherese. Per il task di reperimento multilingua sono stati utilizzati documenti in francese, inglese, italiano, olandese, russo, spagnolo e svedese.

2.5 Conclusioni

In questo capitolo sono stati definiti i concetti di SRI, di pertinenza e di processo di valutazione degli SRI. Inoltre si è visto come questo processo viene messo in opera nelle iniziative internazionali di valutazione.

Per SRI si intende un sistema che, in seguito ad un'interrogazione effettuata dall'utente in modo da rispecchiare un proprio bisogno informativo, restituisce una lista ordinata di documenti pertinenti all'interrogazione.

Il concetto di pertinenza può avere molteplici interpretazioni ma è possibile definirlo come la proprietà, che un documento (inteso come testo, immagine, canzone, ecc.) possiede, di soddisfare il bisogno informativo dell'utente.

Il processo di valutazione dell'efficacia di un SRI considerato qui è quello effettuato mediante l'utilizzo di collezioni di test che permettono di automatizzare la valutazione al fine di effettuare esperimenti ripetibili e non eccessivamente costosi. Questo processo viene messo in pratica in iniziative internazionali quali TREC, NTCIR, INEX e CLEF.

Nel prossimo capitolo verranno descritte le metriche utilizzate nel processo di valutazione dell'efficacia degli SRI.

Capitolo 3

Le metriche di valutazione

Nel capitolo precedente è stato definito il concetto di SRI, è stata discussa la nozione di pertinenza ed è stato descritto il processo di valutazione degli SRI e come esso viene effettivamente messo in opera in diverse iniziative internazionali di valutazione.

Dopo aver mostrato quali sono i procedimenti solitamente svolti per valutare gli SRI, è necessario analizzare quali siano le metriche in base alle quali un SRI venga giudicato più efficace di un altro.

In questo capitolo è presentata una rassegna delle metriche per valutare l'efficacia di un SRI proposte in letteratura. Le metriche sono stata suddivise in metriche classiche (basate sul concetto di pertinenza binaria, paragrafo 3.1), metriche orientata all'utente (che utilizzano dei parametri in modo da meglio adattarsi all'utente, paragrafo 3.2), metriche alternative (che considerano un concetto di pertinenza non binaria, paragrafo 3.3) ed infine metriche definite per la valutazione di sistemi di reperimento di documenti XML (paragrafo 3.4).

3.1 Le metriche classiche

Data una strategia di reperimento S, una metrica quantifica la *similitudine* tra l'insieme dei documenti reperiti da S e l'insieme dei documenti considerati pertinenti da esperti o da utenti[1]. Questa misura fornisce una stima della *bontà* della strategia S.

Come abbiamo visto in precedenza, nelle iniziative internazionali di valutazione vengono utilizzati dei giudizi di pertinenza effettuati, a seconda della modalità di valutazione, da esperti dell'argomento oppure ad utenti. Questi giudizi vengono poi utilizzati dalle metriche per valutare l'efficacia degli SRI.

Andrebbero, inoltre, considerate le prestazioni del sistema su un ampio gruppo di interrogazioni, e non solamente su di una in particolare, per poi fare la media delle varie misurazioni in quanto è naturale pensare che ci sia un'elevata variabilità nel tipo di interrogazioni che gli utenti degli SRI possono generare.

[1] Questa definizione va bene per situazioni di pertinenza e reperimento binario

3.1.1 Precision e Recall

La prima metodologia per definire delle metriche di valutazione messa a punto nel campo dell'RI è la seguente. Si focalizza l'attenzione su di una particolare interrogazione, rispetto la quale si identifica un insieme *Rel* di documenti pertinenti per essa. Un buon sistema sarà quello che reperirà tutti e soli i documenti contenuti in *Rel*. Se indichiamo con *Retr* l'insieme dei documenti reperiti dal sistema, il numero di documenti contenuti in $Rel \cap Retr$, cioè il numero di documenti pertinenti che sono stati reperiti, sarà la misura chiave per determinare la bontà del sistema.

Le prime due metriche proposte nel campo del RI confrontano $|Rel \cap Retr|$ con altri valori [58]. Se si è particolarmente interessati al fatto che il sistema debba reperire ogni documento considerato pertinente, è giusto confrontare l'intersezione con il numero dei documenti pertinenti $|Rel|$. Questa metrica è nota con il nome di *Recall*:

$$Recall = \frac{|Rel \cap Retr|}{|Rel|}$$

In alternativa, se si è interessati a sapere quanto di quello che viene presentato all'utente dall'SRI è realmente pertinente, una possibile alternativa è quella di confrontare l'intersezione con il numero dei documenti reperiti $|Retr|$. Questa metrica è noto con il nome di *Precision*:

$$Precision = \frac{|Rel \cap Retr|}{|Retr|}$$

Gli utenti possono spesso considerare più importante avere SRI con un alto valore di Precision oppure con un alto valore di Recall. Ad esempio un avvocato che sta cercando informazioni riguardo tutti i precedenti casi simili al suo sarà più interessato ad un alto valore di Recall. Invece, uno studente che deve fare una veloce ricerca sul web per un compito scolastico sa bene che ci sono molti documenti pertinenti alla sua ricerca, e vorrà soltanto che i risultati della sua ricerca siano tutti pertinenti, sarà quindi interessato ad un alto valore di Precision.

3.1.2 Fallout, Generality Factor, Classification Accuracy, E-measure

Una metrica strettamente collegata a Precision e Recall, ma meno utilizzata nelle valutazioni degli SRI, è chiamata *Fallout* [58, Cap. 7]. In questa metrica si focalizza l'attenzione sui documenti non pertinenti e la loro proporzione rispetto a tutti quelli reperiti. Se definiamo con $\overline{Rel}$ l'insieme dei documenti che non sono stati giudicati pertinenti, possiamo definire la Fallout come:

$$Fallout = \frac{|\overline{Rel} \cap Retr|}{|\overline{Rel}|}$$

Questa metrica misura la percentuale dei documenti non pertinenti che sono stati reperiti, quindi, in altre parole, ci dice quanti errori ha commesso l'SRI che stiamo valutando (cioè quanti documenti ha erroneamente reperito). Possiamo anche vederla come la probabilità per un elemento reperito di non essere pertinente [58, Cap. 7].

Un altro aspetto che risulta naturale analizzare è quanto *ampia* risulta essere l'interrogazione che si sta analizzando, cioè quanto essa è generale. Si definisce quindi la *Generality Factor* [58] nel seguente modo:

$$\text{Generality Factor} = \frac{|Rel|}{|Doc|}$$

dove $|Doc|$ indica il numero totale di documenti nella collezione.

Questa metrica indica la percentuale dell'intera collezione che è stata considerata pertinente. Se il valore di questa metrica è elevato vorrà dire che l'interrogazione che si sta effettuando è generale, cioè che molti documenti sono pertinenti ad essa. Nel caso di molteplici interrogazioni si definisce il Generality Factor come il numero medio di documenti pertinenti che possiamo trovare nella collezione [49, Cap. 5]. Si può notare che però questo indicatore non è una misura dell'efficacia di un SRI, bensì un indicatore del tipo di interrogazioni considerate.

Un'ulteriore metrica proposta in letteratura è la *Classification Accuracy* [3, Cap. 4], che misura la frequenza con cui la classificazione della pertinenza di un documento, eseguita dall'SRI, è corretta:

$$\text{Classification Accuracy} = \frac{|Rel \cap Retr| + |\overline{Rel} \cap \overline{Retr}|}{|Doc|}$$

La Classification Accuracy ci dice se l'SRI ha reperito i documenti che sono pertinenti e se ha evitato di reperire quelli non pertinenti .

La metrica *Utility* [49] permette di associare un costo o un valore ad un particolare insieme di documenti. La formula dell'Utility è:

$$\begin{aligned} Utility = W_1(|Rel \cap Retr|) + C_1(|\overline{Rel} \cap Retr|) + \\ + C_2(|Rel \cap \overline{Retr}|) + W_2(|\overline{Rel} \cap \overline{Retr}|) \end{aligned}$$

dove W_1 è un peso positivo dato alla corretta classificazione dell'SRI che ha reperito i documenti pertinenti, C_1 è un peso negativo dato ai documenti che sono stati reperiti pur non essendo pertinenti, C_2 è un peso negativo dato ai documenti che non sono stati reperiti pur essendo pertinenti e W_2 è un peso positivo dato ai documenti non pertinenti che, correttamente, non sono stati reperiti dall'SRI.

Una misura che combina assieme Recall e Precision è chiamata *E-measure* [58]. L'idea è di permettere all'utente di specificare se è più interessato ad ottenere un elevato valore di Recall od un elevato valore di Precision. La definizione di E-measure è la seguente:

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{R(j)} + \frac{1}{P(j)}}$$

dove $R(j)$ è il valore di Recall per il j-esimo documento nell'ordinamento, $P(j)$ è il valore di Precision per il j-esimo documento nell'ordinamento, $E(j)$ è la misura relativa ai valori $R(j)$ e $P(j)$ ed infine b è un parametro specificato dall'utente che riflette l'importanza che ricoprono Recall e Precision. Per $b = 1$, il valore della misura $E(j)$ assume il valore della media armonica tra Recall e Precision. Per valori di b maggiori di 1 si intende un maggior interesse per alti valori di Precision, mentre per valori di b minori di 1 si intende un maggior interesse per alti valori di Recall.

3.1.3 Curve Precision/Recall e Precision at N

Le metriche Precision e Recall sono molto elementari e considerano solo certi aspetti del reperimento. Sono misure relative all'intero insieme di elementi reperiti, quindi non considerano la qualità della classificazione degli elementi fatta da un SRI. Naturalmente vanno presi in considerazione anche altri aspetti per definire una buona metrica di valutazione, in quanto gli utenti degli SRI vogliono che i documenti reperiti siano ordinati in base al loro grado di pertinenza e non soltanto forniti sotto forma di insieme.

I documenti con un maggiore grado di pertinenza devono essere presentati in testa ai documenti restituiti per un'interrogazione. Per considerare l'ordine in cui i documenti sono reperiti dal sistema è possibile, ad esempio, utilizzare una misura di correlazione (Kendall o Spearman [8]) che indichi quanto simili tra loro sono due ordinamenti. In questo caso gli ordinamenti saranno quello proposto dall'SRI che stiamo valutando e quello che l'utente ha costruito dando dei giudizi di pertinenza.

Un ulteriore modo per tenere conto dell'ordinamento di documenti reperiti è di calcolare il valore di Precision a diversi livelli [2]. Per esempio se i primi 10 documenti sono tutti pertinenti rispetto all'interrogazione ed i secondi 10 sono tutti non pertinenti, si ha il 100% di Precision ad un livello di 10 documenti (P@10), ma il 50% di Precision ad un livello di 20 documenti (P@20).

Un'ultima cosa che è possibile costruire quando si hanno a disposizione i valori delle misure Precision e Recall (si veda il paragrafo 3.1.1) sono le curve Precision/Recall [58]. Questi grafici sono forse le metriche più usate e vengono pubblicati dalle maggior parte delle iniziative di valutazione degli SRI.

Questi grafici mostrano, sugli assi x ed y, rispettivamente i valori di Recall e di Precision. Solitamente vengono costruiti mostrando il valore di Precision ad 11 livelli standard di Recall (che sono 0, 0.1, 0.2, ... 1.0). Il risultato che si ottiene è un grafico simile a quello riportato in figura 3.1.

Con questo tipo di grafico è facile confrontare le prestazioni di due SRI perché è sufficiente vedere se la curva di uno stia sopra a quella dell'altro. Nel caso le due curve si incrocino in un punto vorrà dire che un SRI ha prestazioni migliori ad alti valori di Precision, mentre l'altro funziona meglio su alti valori di Recall. Può anche

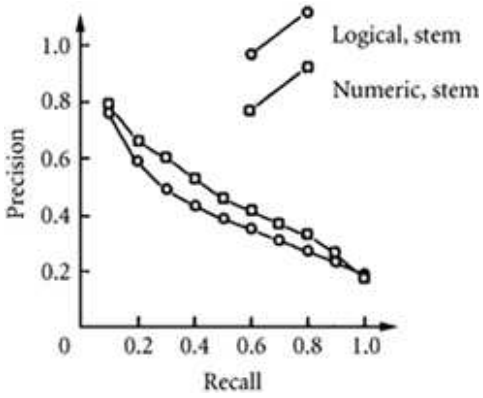


Figura 3.1: Curve Precision/Recall (da [3, Cap. 4])

capitare che le curve si incrocino in diversi punti. In questo caso non è possibile giudicare i due SRI.

Analogamente alle curve Precision/Recall, possono venire prodotte, come risultato delle valutazioni, anche delle curve Recall/Fallout [58].

3.1.4 Average Precision, R-Precision

Sulla base delle misure di base Precision e Recall (paragrafo 3.1.1), si possono definire ulteriori misure che vengono solitamente utilizzate nelle iniziative internazionali per valutare gli SRI. Le due metriche principalmente utilizzate sono *Average Precision* ed *R-Precision*.

Nel calcolo dell'Average Precision [2] l'idea è di combinare assieme il valore di Precision e l'ordinamento fatto dall'SRI. Sia $n = |Retr|$ il numero di documenti reperiti e $h[i]$ l'i-esimo documento reperito. Definiamo $Rel[i]$ uguale a 1 se $h[i]$ è pertinente e 0 altrimenti. Infine, sia $|Rel|$ il numero totale di documenti pertinenti rispetto all'interrogazione. Possiamo, con questa notazione, formulare la definizione di Precision al documento j come:

$$P@j = \sum_{k=1..j} Rel[k]/j$$

Possiamo quindi definire l'*Average Precision*:

$$\text{Average Precision} = \sum_{j=1..n} (P@j * Rel[j])/|Rel|$$

Average Precision risulta quindi essere la somma dei valori di Precision calcolati quando ogni documento pertinente è reperito, normalizzata per il numero totale di documenti pertinenti presenti nella collezione.

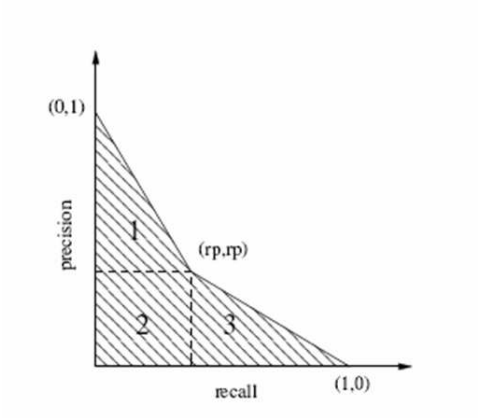


Figura 3.2: Interpretazione geometrica di R-Prec (da [1])

La misura R-Precision (R-Prec) [2] è data dalla Precision ad R ($P@R$), dove R è il numero di documenti pertinenti per quell'interrogazione nella collezione.

Si può osservare che la metrica Average Precision è riferita ad un'unica interrogazione. Se si calcola la media dei valori di Average Precision su tutte le interrogazioni utilizzate per la valutazione, si ottiene la metrica *Mean Average Precision* (MAP). È stata definita anche l'*Interpolated Mean Average Precision* come la media dei valori di Average Precision calcolata ad 11 livelli standard di Recall (che sono 0, 0.1, 0.2, ... 1.0).

È possibile interpretare geometricamente la metrica MAP nel modo seguente: il valore di MAP è approssimativamente l'area presente sotto la curva Precision/Recall dell'SRI valutato. In [1] viene mostrato come è possibile interpretare geometricamente anche la metrica R-Prec e come il suo valore è fortemente legato a quello di MAP. La curva in figura 3.2 è ottenuta unendo i punti (1,0) (R-Prec,R-Prec) e (0,1). Di conseguenza l'area sottesa da questa curva è pari a R-Prec. Come visto, MAP approssima l'area sotto la curva Precision/Recall. Se si suppone che la curva presente in figura 3.2 approssima una curva Precision/Recall, dato che l'area sotto questa curva è uguale al valore di R-Prec, è possibile concludere che MAP approssima il vaore di R-Prec e viceversa.

3.2 Le metriche orientate all'utente

Le metriche fin qui viste assumono il fatto che l'utente faccia parte di una categoria omogenea. Naturalmente si tratta di un'assunzione semplificatrice che non rispecchia appieno la realtà. Le misure Precision e Recall assumono il fatto che l'insieme dei documenti pertinenti è indipendente dall'utente che giudica la pertinenza. È naturale pensare, però, che diversi utenti possano dare diverse interpretazioni di quanto pertinente sia un certo documento rispetto ad una certa interrogazione. Per

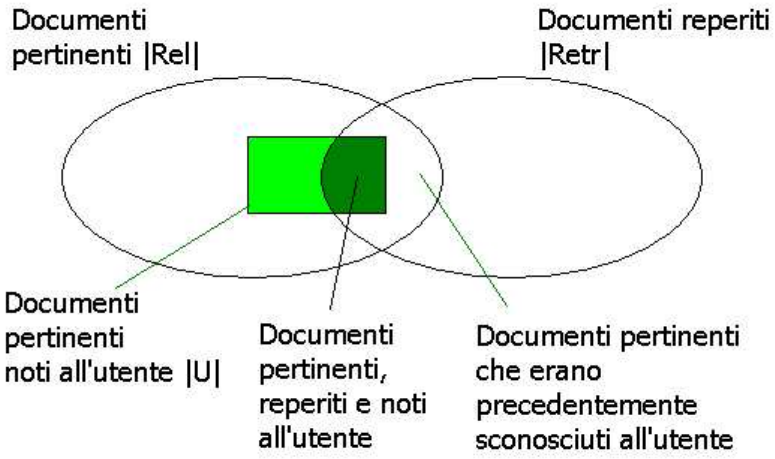


Figura 3.3: Coverage e novelty

cercare di considerare questi aspetti soggettivi all'interno delle metriche di valutazione sono state proposte diverse metriche raggruppate nella categoria di metriche "orientate all'utente" [35, Cap. 8].

Se indichiamo con $|U|$ il numero di documenti pertinenti già noti all'utente prima di sottoporre l'interrogazione corrente all'SRI, possiamo definire il *Coverage Ratio* [35] come la percentuale di documenti pertinenti già noti all'utente che sono stati reperiti con questa interrogazione:

$$\text{Coverage Ratio} = \frac{|Retr \cap U|}{|U|}$$

(si veda la figura 3.3). Un alto valore di Coverage Ratio indica che l'SRI rileva come maggiormente pertinenti i documenti che l'utente si aspetta di trovare all'interno dei documenti reperiti. Quindi questa metrica ci indica quanto l'SRI è in grado di reperire i documenti che l'utente ha già potuto giudicare pertinenti in passato.

Un'altra metrica orientata all'utente è il *Novelty Ratio* [35], che è definito come la percentuale dei documenti pertinenti che sono stati reperiti e che, in precedenza, erano sconosciuti all'utente. Se indichiamo con $|\overline{U}|$ il numero di documenti pertinenti che erano sconosciuti all'utente prima di sottoporre l'interrogazione, abbiamo che:

$$\text{Novelty Ratio} = \frac{|\overline{U} \cap Retr|}{|Rel \cap Retr|}$$

(si veda nuovamente la figura 3.3). Un alto valore di Novelty Ratio indica quanto l'SRI è in grado di mostrare (all'utente) documenti pertinenti che prima gli erano sconosciuti.

La *Relative Recall* [35] è una metrica che si dirige più verso il fattore soggettivo dato da quanti documenti pertinenti l'utente vuole reperire. Supponendo che l'utente voglia 5 documenti pertinenti e che l'SRI reperisca 20 documenti tra cui ce ne sono 3 pertinenti, il valore di Relative Recall sarà dato da 3/5: l'utente ha ottenuto solamente 3 dei 5 documenti che desiderava. Questa metrica quindi misura il rapporto tra il numero di documenti pertinenti che l'utente avrebbe voluto analizzare ed il numero di documenti pertinenti e reperiti.

Possiamo notare che se il valore di Relative Recall è pari a 1, questa metrica non dà nessuna indicazione di quanto sforzo è stato necessario all'utente nell'individuare i documenti che desiderava. Infatti, se nell'esempio precedente la Relative Recall è pari a 1, è possibile che l'utente abbia trovato i 5 documenti che voleva all'interno dei primi 5 o 6 documenti reperiti oppure è possibile che abbia dovuto analizzare tutti e 20 i documenti reperiti per trovarne 5 pertinenti.

Per questo motivo è stata definita un'ulteriore metrica orientata all'utente. Supponendo che la collezione di documenti contenga il numero di documenti pertinenti desiderato dall'utente e che l'SRI permetta all'utente di cercarli, è possibile definire il *Recall Effort* [35] come il rapporto tra il numero di documenti desiderato dell'utente ($NDes$) ed il numero di documenti che l'utente ha dovuto analizzare per trovarli (NAn):

$$\text{Recall Effort} = NDes/NAn$$

In questo modo il Recall Effort sarà uguale ad 1 se i documenti desiderati dall'utente sono i primi documenti reperiti dall'SRI, mentre sarà prossimo a 0 se l'utente ha dovuto analizzare centinaia di documenti prima di trovare tutti i documenti pertinenti che desiderava.

3.3 Le misure alternative

Precision, Recall e le metriche derivabili da queste (paragrafo 3.1) sono sicuramente quelle più utilizzate nell'ambito della valutazione. Nonostante ciò esse non sono sempre le misure di valutazione più appropriate per valutare le prestazioni di un SRI. Per questo motivo sono state proposte diverse metriche alternative.

3.3.1 Expected Search Length, Normalized Recall e Precision, bpref

Gli SRI che partecipano alle iniziative internazionali di valutazione ordinano l'insieme dei documenti reperiti in modo decrescente in base ad un valore assegnato ad ogni documento (denominato Retrieval Status Value, RSV). Come abbiamo visto, l'ipotetico utente dell'SRI dovrà scorrere la lista dei documenti reperiti alla ricerca di quelli pertinenti. Così come la metrica Recall Effort tiene conto dello sforzo che l'utente deve fare nell'analizzare i documenti reperiti per trovare quelli pertinenti, esiste un'altra metrica che tiene conto di questo fattore. La misura proposta nel 1968 da Cooper, e chiamata *Expected Search Length* (ESL) [13], considera la lunghezza del cammino che l'utente deve fare lungo la lista ordinata di documenti reperiti. Si misura, quindi, quanti documenti non pertinenti l'utente deve analizzare prima di trovare ogni documento pertinente.

Si può anche vedere ESL come una funzione e definire $ESL(n)$ come il numero di documenti che bisogna esaminare seguendo il rank per trovare n documenti pertinenti. Utilizzando la notazione riferita alla metrica Recall Effort, possiamo dire che $ESL(NDes) = NAn$. ESL non è quindi un singolo valore ma una funzione di N. È possibile calcolare $ESL(n)/n$ per avere un singolo valore che rappresenta il numero medio di documenti letti per ogni documento pertinente.

Oltre ad ESL esistono altre metriche che considerano l'ordine con cui i documenti vengono reperiti e presentati all'utente. Un sistema ideale reperirà tutti i documenti pertinenti prima di quelli non pertinenti. È possibile definire quindi delle metriche che indicano quanto le prestazioni dell'SRI che si sta analizzando differiscono da quelle di un SRI ideale.

La misura *Normalized Recall*, proposta da Rocchio nel 1966 [35], misura alcune differenze che ci sono tra l'SRI ideale e quello che si sta valutando. Si misurano le differenze dell'ordine dei documenti reperiti dall'ordine ideale che mette per primi i documenti pertinenti e si normalizza il risultato tra 0 ed 1. Un'ulteriore metrica, sempre proposta da Rocchio nel 1966, che tiene conto delle differenze tra l'SRI attuale e quello ideale è la misura *Normalized Precision* [35] che misura la deviazione tra la curva di Precision attuale da quella ideale.

Nel 2004 è stata proposta da Buckley e Voorhees [11] una metrica chiamata *bpref*. Questa metrica ha come obiettivo quello di basarsi solamente sui documenti giudicati pertinenti o non pertinenti, e quindi solamente sui documenti contenuti nel pool. Per un'interrogazione con R documenti pertinenti è possibile definire bpref come:

$$bpref = \frac{1}{R} \sum_{r} 1 - \frac{|n \text{ più in alto nel rank di } r|}{R}$$

dove r è un documento pertinente e n sono i primi R documenti giudicati non pertinenti e reperiti dall'SRI.

Per come è definita, la metrica bpref considera solamente pochi documenti. Per questo motivo sono state definite varianti [11] che considerano un numero maggiore di documenti (almeno 10).

3.3.2 Sliding Ratio, Satisfaction, Frustration, Total

Dopo aver descritto le metriche Normalized Recall e Normalized Precision vediamo una metrica che è concettualmente simile alle due precedenti. Lo *Sliding ratio* (SR) [35] proposto da Pollack nel 1968 rispetto le precedenti metriche, è basato su giudizi di pertinenza pesati e sul reperimento di n documenti invece che di tutti quelli pertinenti. L'analoga assunzione è che il SRI ideale reperisce n documenti con valore decrescente di pertinenza. L'SRI che si sta valutando probabilmente non reperirà questi n documenti in ordine. È possibile quindi comparare questi due ordinamenti:

$$SR(n) = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} W_i}$$

dove i W_i rappresentano i pesi dei documenti come sono stati reperiti dal sistema che si sta valutando, mentre i w_i rappresentano i pesi dei documenti in ordine decrescente come l'SRI ideale li avrebbe ordinati.

È possibile notare che lo SR è calcolato considerando solamente i documenti reperiti. A differenza della metrica Normalized Recall, lo SR ha il vantaggio di utilizzare dei pesi di pertinenza e di considerare solamente i documenti reperiti e non quelli dell'intera collezione.

La misura SR è stata in seguito perfezionata con la definizione di tre metriche [35] simili a SR che considerano separatamente i documenti pertinenti e quelli non pertinenti. La *Satisfaction* che considera solamente i documenti pertinenti, la *Frustration* che considera solamente i documenti non pertinenti e la *Total* che effettua una combinazione pesata di Satisfaction e di Frustration.

3.3.3 Relative Relevance, Ranked Half-Life

Un tipico problema nella valutazione degli SRI è dato dal fatto che si utilizzano dei valutatori umani che non possono essere totalmente obiettivi nella loro opera. Sono state proposte delle metriche che considerano il reperimento effettuato da un SRI alla pari dei giudizi di pertinenza di giudici umani o di utenti. Sono quindi disponibili diverse classificazioni di pertinenza effettuate da giudici, da utenti e da SRI. La metrica *Relative Relevance*, proposta nel 1998 in [7], misura il grado di accordo che esiste tra due diversi giudizi di pertinenza. Il calcolo della misura Relative Relevance (RR) è effettuato combinando il risultato di due giudizi di pertinenza effettuati in un contesto di pertinenza non binario. A questo scopo si utilizza la "Jaccard measure" che si pone l'obiettivo di quantificare la relazione esistente tra due tipi diversi di giudizi (R_1 e R_2). Questi due giudizi sono costituiti dai valori attribuiti dai giudici, da utenti o da SRI ai vari documenti della collezione.

$$Associazione(R_1, R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$$

I giudizi di pertinenza sono assegnati in base al tipo di pertinenza che si sta utilizzando. Ad esempio si può usare la scala a tre livelli: Fortemente pertinenti

(1.0), Debolmente pertinenti (0.5) e Non pertinenti (0.0). Il numero di giudizi di pertinenza che sono in accordo tra loro, che definisce il nominatore della formula di Jaccard, viene diviso per la cardinalità dell'unione dei giudizi di pertinenza che sono stati effettuati sui documenti. Con questa misura è possibile capire che tipo di giudizi è meglio utilizzare per valutare l'efficacia di un SRI.

Possiamo concludere dicendo che la RR non è una misura di prestazione simile alle precedenti, ma che, più che altro, essa è in grado di permettere una maggiore comprensione delle caratteristiche delle prestazioni che hanno gli SRI.

Dopo essere riusciti a comparare due o più diversi tipi di giudizi di pertinenza, diventa importante riuscire a comparare tra loro i ranking risultanti dal reperimento svolto da due SRI. Avendo a disposizione queste liste ordinate si possono utilizzare due diversi tipi di informazioni:

- l'ordine risultante dal reperimento, che rappresenta una lista di valori decrescenti di pertinenza stimata dall'SRI;
- i valori dei giudizi di pertinenza, che rappresentano le interpretazioni dell'utente o del valutatore rispetto la pertinenza dei documenti rispetto le interrogazioni.

L'indicatore *Ranked Half-Life*, proposto nel 1998 in [7], utilizza entrambe queste informazioni. Il valore di Ranked Half-Life (RHL) è dato dal caso "mediano" di pertinenza. La dimensione che si considera è quella del ranking dei documenti prodotto da un SRI. L'idea è di sfruttare i punteggi di pertinenza (che sono più alti per i documenti in testa alla lista) assegnati dai giudici. In questo modo, se ci sono dei documenti fortemente pertinenti che non si trovano all'interno dei primi reperiti, il caso "mediano" scenderà verso il basso della lista degli oggetti reperiti dall'SRI.

Questa metrica quindi ci indica il grado con cui i documenti pertinenti sono posizionati in testa alla lista dei documenti reperiti.

In confronto alla normale metrica di Precision, l'indicatore RHL sfrutta le informazioni riguardo all'ordine in cui i documenti sono reperiti. Se Precision e Recall vedono i documenti reperiti come un insieme, e quindi come dei gruppi di elementi non ordinati tra loro, RHL vede i documenti reperiti come una lista, e quindi con un ordinamento, e valuta l'ordinamento dei documenti oltre che il loro livello di pertinenza.

3.3.4 NDPM, Usefulness, ASL

La metrica che abbiamo appena descritto (cioè RHL) sfrutta l'ordinamento effettuato dall'SRI che valuta. Esistono altre metriche che tengono in considerazione questo ordinamento. La metrica NDPM, proposta da Yao nel 1995 [62], misura la distanza tra l'ordinamento effettuato da un SRI e quello effettuato dall'utente e la divide per la distanza massima tra l'ordinamento effettuato dall'utente e tutti quelli effettuati

dai SRI che si stanno valutando. In questo modo, se la classificazione dell'SRI è perfetta, la metrica NPDM avrà valore 0, mentre nel caso peggiore avrà valore 1.

Un'ulteriore metrica che tiene in considerazione l'ordinamento nella valutazione è la *Usefulness* proposta da Frei e Schäuble nel 1991 [20]. Questa metrica permette di confrontare tra loro due SRI e di stabilire quale fornisce informazioni più utili all'utente. Per prima cosa è necessario confrontare le coppie di documenti in cui il primo ha un valore di pertinenza, in base ai giudizi dell'utente, minore del secondo con tutte le coppie analogamente formate secondo la classificazione di pertinenza fatte da un SRI. Si effettua lo stesso confronto con la classificazione fatte da un altro SRI. Per capire quale dei due SRI è migliore si misura quante volte, in media su più interrogazioni, le classificazioni dei documenti sono in accordo con quelle fatte dal giudice umano.

Un'altra metrica che considera l'ordinamento fatta dagli SRI e che è derivata dal ESL è l'*Average Search Length* (ASL) di Losee [37]. Questa metrica misura il numero medio di documenti che è necessario analizzare lungo la lista ordinata di elementi reperiti, per raggiungere la posizione media tra tutte le posizioni dei documenti pertinenti.

3.3.5 Discounted Cumulative Gain

Le maggior parte delle metriche fin qui viste considerano un modello di pertinenza binaria. Può capitare che nelle valutazioni si utilizzino modelli di pertinenza a più di due livelli. Tipicamente, quando questo succede, Precision e Recall vengono calcolate collassando i vari livelli di pertinenza a soli due (ad esempio in NTCIR [42]). Per i casi di test con pertinenza non binaria, esistono altri tipi di metriche. La Discounted Cumulative Gain (DCG) [28, 29, 34] proposta da Järvelin e Kekäläinen nel 2000 è una metrica che sfrutta gradi multipli di pertinenza e che punta a valutare la capacità degli SRI di classificare come primi della lista dei documenti reperiti, quelli fortemente pertinenti. Quando si esamina la lista ordinata dei documenti restituiti da un SRI è facile notare che:

- i documenti fortemente pertinenti sono quelli con maggiore valore per l'utente rispetto a quelli debolmente pertinenti;
- più grande è la posizione nel rank di un documento pertinente, più piccolo è il suo valore per l'utente perché difficilmente esaminerà i documenti con posizioni alte nel rank.

Dati questi fatti, si può pensare di valutare gli SRI in base all'accumulo di pertinenza man mano che si analizzano i documenti presenti nel rank. È possibile, cioè, trasformare la lista dei documenti in una lista di valori via via crescenti utilizzando i valori di pertinenza dei vari documenti. Ad esempio con una pertinenza a 4 livelli (0, 1, 2 o 3) possiamo ottenere un vettore di questo tipo:

$$G =< 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \ldots >$$

Questo vettore indica in posizione i il valore di pertinenza del documento reperito nella medesima posizione dall'SRI. Da questo vettore è possibile derivare una lista di valori debolmente crescenti in cui alla posizione i è data dalla somma dei valori del vettore G dalla posizione 1 alla posizione i. Possiamo definire il vettore CG ricorsivamente in questo modo:

$$CG[i] = \begin{cases} G[1] & \text{se i=1} \\ CG[i-1] + G[i] & \text{altrimenti} \end{cases}$$

Nel caso dell'esempio precedente otterremmo il seguente vettore:

$$CG =< 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \ldots >$$

Il secondo punto dell'elenco precedente porta alla conclusione che è possibile comparare le prestazioni degli SRI in modo tale che, tanto più grande è il rank di un documento, tanto più piccolo dovrebbe essere il valore aggiunto al vettore CG in quanto più è grande il rank, tanto meno utile è il documento per l'utente. È necessaria quindi una funzione che riduca progressivamente il valore dei documenti all'aumentare del rank. Ad esempio dividendo il valore del documento per il logaritmo del suo rank. Con la scelta della base del logaritmo, l'utente potrà scegliere quanto dovrà pesare la posizione nel rank dei documenti. A questo punto possiamo definire il vettore DCG ricorsivamente:

$$DCG[i] = \begin{cases} G[1] & \text{se i=1} \\ DCG[i-1] + G[i]/log_b i & \text{altrimenti} \end{cases}$$

Nell'esempio, se consideriamo $b = 2$ otterremo il seguente vettore:

$$DCG =< 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \ldots >$$

La capacità di un SRI di classificare come primi della lista dei risultati i documenti fortemente pertinenti viene così visualizzata nei vettori CG e DCG. Facendo una media su più interrogazioni delle prestazioni dei SRI è possibile valutarli e compararli.

Alcune caratteristiche interessanti di questa misura sono che riesce a combinare il grado di pertinenza dei documenti con la loro posizione nel rank e che ad ogni posizione del rank fornisce una stima del valore accumulato di pertinenza.

3.3.6 Average Weighted Precision, Weighted R-Precision, Q-Measure e R-Measure

Dopo aver visto la metrica DCG vediamo alcune metriche che sono state derivate da essa. Ad esempio l'*Average Weighted Precision* (AWP) [30], del 2001, calcola il vettore CG e lo confronta con il vettore CG di un SRI ideale che posiziona ai primi posti i documenti maggiormente pertinenti. Per effettuare il confronto si effettua la media dei rapporti tra i due vettori calcolati ad ogni documento pertinente che

si incontra nella lista dei documenti reperiti. Nel caso di pertinenza binaria questa metrica diventa analoga all'Average Precision. Possiamo quindi dire che il rapporto tra i vettori CG dell'SRI che si valuta e dell'SRI ideale è una sorta di Precision pesata in base ai livelli di pertinenza che si considerano.

Analogamente ad AWP è possibile calcolare la *R-Weighted Precision* [30] dividendo tra loro i vettori CG calcolati al R-esimo documento reperito, dove R è il numero di documenti pertinenti presenti nella collezione:

$$AWP = 1/R \sum_{r:g(r)>0} CG(r)/CGI(r)$$

dove $g(r)$ è maggiore di 0 per i documenti pertinenti, $CG(r)$ è il vettore CG calcolato su r documenti reperiti e $CGI(r)$ è il vettore CG dell'SRI ideale calcolato su r documenti reperiti.

AWP non riesce a penalizzare la presenza di documenti pertinenti oltre la posizione R dell'ordinamento. Per risolvere questo problema è stata proposta una metrica simile a questa, la *Q-Measure* [46]:

$$Q - Measure = 1/R \sum_{r:g(r)>0} CBG(r)/(CGI(R) + R)$$

dove $CBG(r)$ è analogo al vettore $CG(r)$ tranne per il fatto che riceve un bonus se il documento in posizione r è pertinente. In questa metrica il denominatore cresce di almeno r scendendo nella lista di documenti reperiti anche dopo la posizione R a differenza di AWP.

È stata definita anche la metrica *R-Measure* [46] definita come controparte di Q-Measure così come R-Precision lo è per Average Precision:

$$R - Measure = \sum CBG(R)/CGI(r)$$

3.4 Le metriche per documenti XML

Come abbiamo visto nel paragrafo 2.4.3 il reperimento di documenti XML è diverso da quello di normali documenti testuali. Per questo motivo sono state appositamente proposte metriche che possano cogliere le caratteristiche degli SRI che si occupano del reperimento di documenti XML. Queste metriche, per la maggior parte, sono state proposte e vengono utilizzate nell'iniziativa di valutazione INEX (paragrafo 2.4.3).

3.4.1 Tolerance to Irrelevance, Expected Ratio of Relevant Documents

Nel caso di reperimento di documenti XML, si considera il fatto che l'SRI deve fornire all'utente un *entry-point* nel documento che sia il più possibile vicino all'informazione giudicata pertinente. L'SRI deve quindi produrre un elenco ordinato di

entry-point. L'utente, quindi, legge il documento partendo dall'entry-point fino a che la sua tolleranza a informazioni non pertinenti (parametro da indicare) termina. A questo punto passerà al successivo elemento reperito. Questa metrica [14] valuta il reperimento focalizzato e favorisce gli SRI che portano l'utente vicino alle informazioni pertinenti evitando di reperire frammenti di testo troppo grandi.

Un'ulteriore metrica per valutare SRI di documenti XML è l'*Expected Ratio of Relevant Documents* (ERR) [43]. In questo caso si stima il numero di elementi pertinenti che un utente trova analizzando la lista dei primi k elementi reperiti, diviso il numero di elementi pertinenti che l'utente avrebbe trovato analizzando l'intera collezione.

3.4.2 Normalized eXtended Cumulative Gain, MAnxCG, Effort-Precision

Nell'edizione di INEX del 2005 sono state proposte delle nuove metriche per valutare l'efficacia degli SRI che vi hanno partecipato. Una metrica derivata da DCG è chiamata *Normalized eXtended Cumulative Gain* (nXCG) [32]. Essa misura l'accumulo dei valori di pertinenza lungo la lista ordinata di elementi reperiti. Il valore $xCG(i)$ è dato dalla somma dei valori di pertinenza fino al rank i. Si può denotare con xCI il vettore CG ideale con i valori di pertinenza più elevati ai primi posti dell'ordinamento. Confrontando questi due vettori si ottiene il valore di nXCG.

Gli SRI possono essere confrontati tra loro a diversi livelli di documenti reperiti. Per questo motivo è stata definita la *Mean Average nXCG* [32] data da:

$$MAnXCG[i] = \frac{\sum_{j=1}^{i} nXCG[j]}{i}.$$

Un ultima metrica proposta in INEX 2005 è l'*Effort-Precision* (EP) [32]. Questa metrica vuole misurare lo sforzo necessario all'utente per raggiungere un certo livello di soddisfazione. L'EP è definita come:

$$EP[r] = e_{ideal}/e_{run}$$

dove e_{ideal} è la posizione dell'ordinamento in cui il valore di CG(r) è ottenuto dall'SRI ideale, e_{run} invece è la posizione in cui il valore CG(r) è ottenuto dall'SRI che si sta valutando.

3.5 Conclusioni

In questo capitolo è stata presentata una panoramica delle metriche proposte in letteratura. Sono state illustrate per prime le metriche basate sul concetto di pertinenza binaria come Precision, Recall e Fallout (paragrafi 3.1.1 e 3.1.2). Sono state poi definite le metriche che tengono conto dell'ordinamento dei documenti effettuato dagli SRI come le curve Precision/Recall e le metriche Average Precision e R-Precision (paragrafi 3.1.3 e 3.1.4). Sono state descritte le metriche orientate all'utente e le

metriche alternative che utilizzano concetti di pertinenza e di reperimento diversi da quello binario (paragrafi 3.2 e 3.3). Infine sono state presentata le metriche utilizzate per valutare il reperimento di documenti XML (paragrafo 3.4).

Nel prossimo capitolo vedremo alcuni meccanismi per comprendere meglio le metriche descritte in questo capitolo e per valutarne l'efficacia calcolandone la stabilità.

Capitolo 4

La valutazione dell'efficacia delle metriche

Nei capitoli precedenti abbiamo esaminato alcuni aspetti fondamentali quali la definizione di SRI, il concetto di pertinenza, lo svolgimento del processo di valutazione degli SRI, come esso viene adottato in diverse istanze reali e quali sono le metriche utilizzate per valutare l'efficacia degli SRI.

In questo capitolo vengono presentate alcune considerazioni sui comportamenti delle metriche e degli SRI per capire dove è possibile intervenire per valutare, ed eventualmente migliorare, l'efficacia delle metriche. Vengono descritti i concetti di sensitività e specificità mostrando come delle nozioni del campo dell'Informatica Medica siano analoghe ai concetti utilizzati nel campo del RI (paragrafo 4.1). Si illustra poi come delle ipotesi fatte sui giudizi di pertinenza possano essere riportate analogamente riguardo alle metriche di valutazione (paragrafo 4.2). Viene poi mostrato un metodo per valutare la stabilità delle metriche calcolando il loro tasso d'errore (paragrafo 4.3). Infine vengono mostrati i modelli di distribuzione degli score dei documenti utilizzati per valutare l'efficacia degli SRI (paragrafo 4.4).

4.1 Sensitività e specificità

Nel campo dell'Informatica Medica [27, 51] vengono spesso utilizzati i concetti di sensitività e specificità [51] per valutare la correttezza dei test diagnostici. Vediamo come questi due concetti dell'Informatica Medica sono in realtà molto simili a due concetti utilizzati nel campo del RI.

Un test diagnostico dà come risultato un valore continuo ma il medico deve optare tra uno solo di due stati: sano o malato. Si possono quindi verificare due tipi di errore (si veda la figura 4.1): nel caso in cui un malato non venga riconosciuto come tale, si parla di falso negativo (FN), mentre nel caso in cui un sano viene scambiato per malato si parla di falso positivo (FP).

Analogamente ai concetti di FN e di FP è possibile definire le categorie dei veri positivi (VP) e dei veri negativi (VN) che sono i soggetti il cui stato di malato o sano

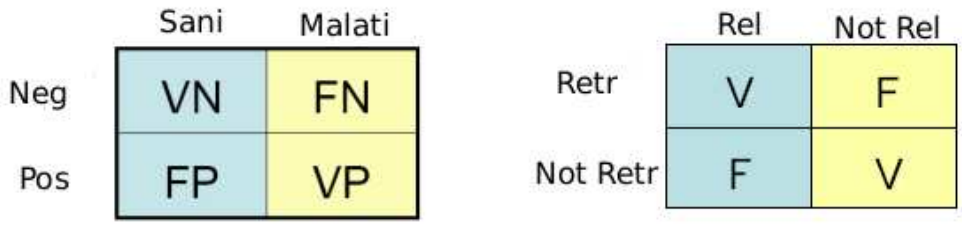


Figura 4.1: Falsi negativi e falsi positivi

è stato correttamente individuato dal test diagnostico. A questo punto è possibile definire la sensitività e la specificità. Per *sensitività* si intende la percentuale di positivi al test diagnostico tra i pazienti malati, mentre per *specificità* la percentuale di pazienti malati sul totale dei positivi al test diagnostico. Quindi, utilizzando la notazione precedente, la sensitività sarà data dalla formula $Sens = VP/(VP+FN)$, mentre la specificità dalla formula $Spec = VP/(VP + FP)$.

Possiamo ora vedere come queste due misure siano le stesse utilizzate nel campo della valutazione degli SRI. Nella figura 4.1 possiamo vedere descritti i concetti di falsi negativi e di falsi positivi appena descritti. Si possono anche vedere i concetti analoghi nel campo del RI. Se consideriamo il concetto di malato come un documento pertinente ed il concetto di positivo come un documento reperito possiamo facilmente arrivare alla trasformazione tra le due tabelle in figura 4.1. Con la stessa similitudine possiamo vedere che il concetto di sensitività è quindi dato dalla percentuale di documenti reperiti tra tutti quelli pertinenti, che corrisponde alla definizione di Recall e che il concetto di specificità è dato dalla percentuale di documenti pertinenti su tutti quelli reperiti, che corrisponde alla definizione di Precision (si veda il paragrafo 3.1.1).

In Informatica Medica si costruiscono diversi grafici dai dati risultati dall'utilizzo di un test su di una popolazione. Nella figura 4.2 è possibile vedere come le popolazioni di sani e di malati vengono distribuite in base all'esito del test diagnostico. Il valore di cut-off per discriminare tra sani e malati è determinato dai ricercatori che determinano qual è la soglia più adatta in base a studi statistici sulle popolazioni. Nel campo del RI si costruiscono grafici analoghi per discriminare tra documenti pertinenti, non pertinenti, reperiti e non reperiti [58, Cap. 7] [49, Cap. 5]. Queste figure, data la precedente similitudine con l'RI, sono esattamente le stesse che si utilizzano nella valutazione degli SRI. Le curve ROC (Receiver Operating Characteristic) sono l'analogo delle curve Precision/Recall (di veda il paragrafo 3.1.2) nel campo del RI. Entrambe le curve utilizzano una scala da 0 a 1, ma la differenza tra le due è che le curve ROC pongono gli assi farli incontrare entrambi nel valore 1, mentre nelle curve Precision/Recall gli assi si incontrano nell'origine.

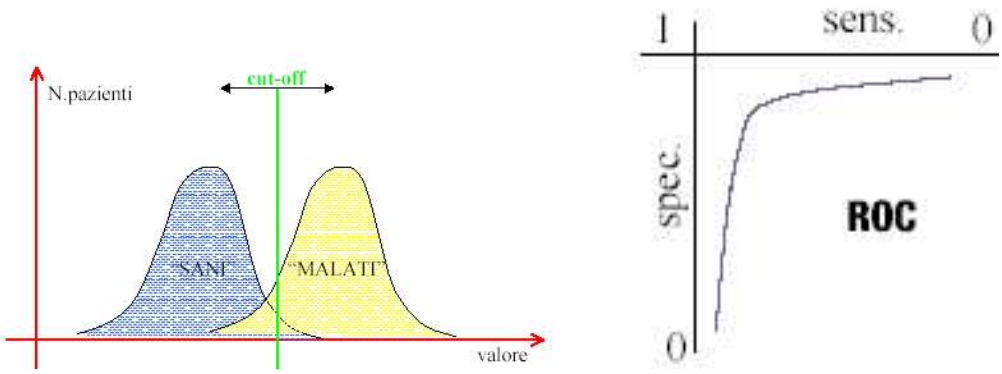


Figura 4.2: Curve ROC

4.2 Ipotesi debole e forte

Lesk e Salton nel 1968 [48] hanno supposto (e confermato sperimentalmente) che i cambiando giudici di pertinenza non cambiano i risultati della valutazione dei SRI. In particolare hanno proposto un'ipotesi debole (*weak hypothesis*) in cui affermano che differenze nei giudizi di pertinenza non influiscono sui risultati del confronto tra due SRI. Quindi se un SRI risulta migliore di un altro utilizzando certi giudizi di pertinenza, allora questo risulterà migliore anche con giudizi effettuati da giudici diversi. La seconda proposta di Lesk e Salton è l'ipotesi forte (*strong hypothesis*) con la quale affermano che eventuali cambiamenti dei giudici di pertinenza non si trasmettono nella valutazione della pertinenza dei documenti utilizzati per la valutazione.

Le idee proposte da Lesk e Salton si possono facilmente trasformare in modo da descrivere le metriche di valutazione. In particolare l'ipotesi debole dice che differenze nei singoli valori delle metriche non modificano l'ordinamento finale degli SRI. L'ipotesi forte afferma che se si utilizzano metriche diverse il sistema migliore secondo una metrica resterà migliore anche secondo le altre metriche. In questa nuova interpretazione l'ipotesi forte non è più vera in quanto metriche diverse misurano caratteristiche degli SRI diverse; quindi è possibile che un SRI risulti migliore di un altro con una metrica ma peggiore utilizzandone un'altra.

4.3 La stabilità delle misure

Vediamo ora alcune caratteristiche comuni delle metriche di valutazione degli SRI. Alcuni studi [9, 10, 59, 61, 63] quantificano l'errore che viene commesso quando, a seguito di un test su una collezione di documenti, si arriva ad una conclusione del tipo "l'SRI A è migliore dell'SRI B". Un primo fatto è che il tasso di errore di una conclusione del genere cresce al diminuire del numero delle interrogazioni

	INQa	INQe	INQp	Saba	Sabe	Sabm	acs	pir
APL	18 0 3	2 11 8	19 0 2	11 0 10	0 19 2	3 11 7	21 0 0	0 19 2
INQa		0 21 0	4 6 11	0 14 7	0 21 0	0 21 0	21 0 0	0 21 0
INQe			21 0 0	19 0 2	1 16 4	4 4 13	21 0 0	0 17 4
INQp				0 15 6	0 21 0	0 21 0	21 0 0	0 21 0
Saba					0 21 0	0 21 0	21 0 0	0 21 0
Sabe						21 0 0	21 0 0	2 4 15
Sabm							21 0 0	0 19 2
acs								0 21 0

Tabella 4.1: Matrice per calcolare il tasso d'errore (da [9])

effettuate per testare gli SRI. Quindi se si fa una media su un maggior numero di interrogazioni si ottiene un risultato meno affetto da errore. Il valore di Precision dopo 30 documenti reperiti (P@30, si veda il paragrafo 3.1.1) ha almeno il doppio del tasso d'errore dell'Average Precision (si veda il paragrafo 3.1.4) che viene, invece, calcolata su tutti i documenti. Ma naturalmente questo non vuol dire che le metriche con un alto tasso d'errore non devono venire utilizzate, in quanto differenti metriche valutano differenti aspetti degli SRI e quindi vanno scelte in base agli obiettivi del test che si sta effettuando.

Un metodo per calcolare questo tasso d'errore è il seguente. Si considerano delle interrogazioni e si calcola la media dei valori delle metriche, per ogni metodologia di reperimento, o SRI, che si vuole comparare. È possibile ottenere così una matrice quadrata con i vari SRI sulle righe e sulle colonne. All'interno di questa matrice, per ogni elemento, si avrà una tripla di numeri che darà informazioni riguardo quante volte un SRI è stato migliore, peggiore o uguale all'altro per ogni insieme di interrogazioni utilizzato (si veda la tabella 4.1).

Se si suppone che nel confronto tra due SRI la risposta corretta è data dal valore maggiore in una singola casella della matrice e che l'errore è commesso nella scelta dell'SRI che ha avuto prestazioni migliori meno volte, è possibile definire il tasso d'errore che si verifica nella scelta tra due SRI, come il numero totale di errori di valutazione delle prestazioni commesso tra tutti gli SRI diviso il numero totale di scelte effettuate sugli SRI (cioè la somma della tripla di numeri presenti nella matrice moltiplicato il numero di interrogazioni eseguite):

$$\text{Tasso d'errore} = \frac{\sum Min(|A > B|, |B > A|)}{\sum(|A > B| + |A < B| + |A == B|)}$$

dove $|A > B|$ è il numero di volte che il metodo A risulta migliore del metodo B.

Con questa formula di errore è, però, possibile ottenere un basso tasso d'errore semplicemente perché una metrica conclude raramente che un SRI è migliore dell'altro. Va quindi considerata anche la porzione di pareggi, definita come il totale del numero di volte che si ottiene un pareggio diviso il numero totale di decisioni che si sono prese.

Measure	Error Rate (%)
Prec(1)	14.3
Prec(10)	3.6
Prec(30)	2.9
Prec at .5 R	2.2
Prec(100)	1.8
AvgPrec	1.5
R-Prec	1.3
Prec(1000)	1.0
Recall(1000)	0.6

Tabella 4.2: Tassi d'errore calcolati in [9]

In questo modo si ottiene che metriche che si basano su pochi documenti fortemente pertinenti hanno un tasso d'errore maggiore di metriche che tengono conto di un numero maggiore di documenti. Un altro risultato che si può evidenziare con questo tipo di tasso d'errore è che esso decresce all'aumentare del numero di interrogazioni considerate nell'esperimento. I tassi d'errore di alcune metriche, calcolati in [9], sono riportati nella tabella 4.2.

Sono state proposte leggere modifiche nel calcolo del tasso d'errore [10, 59, 61] che non modificano i risultati ottenuti. Il risultato di questi studi è che il tasso d'errore decresce all'aumentare del numero di interrogazioni considerate nell'esperimento e all'aumentare del numero di documenti che la metrica tiene in considerazione.

4.4 I modelli di distribuzione degli score dei documenti

Dopo aver visto come è possibile calcolare il tasso d'errore associato alle metriche di valutazione degli SRI e quali sono i fattori che lo influenzano, vediamo come la pertinenza si distribuisce all'interno della collezioni di documenti e quindi come gli SRI dovrebbero assegnare i punteggi di SRS.

A partire da Swets [55], un certo numero di ricercatori [5, 58] negli anni '60 e '70 ha proposto di utilizzare una distribuzione normale per simulare l'andamento dei documenti ordinati dall'SRI ed i seguito di utilizzare tecniche statistiche per definire una soglia che separi i documenti pertinenti da quelli non pertinenti.

In [38], invece, si dimostra che, in seguito ad esperimenti effettuati su alcune collezioni di TREC, l'insieme dei punteggi dei documenti pertinenti può essere ragionevolmente rappresentato da una distribuzione di tipo gaussiano quando si hanno a disposizione un numero sufficiente di documenti pertinenti (solitamente sono necessari almeno 60 documenti). È anche stato mostrato che una distribuzione esponenziale decrescente si adatta bene all'insieme dei documenti non pertinenti. Va quindi utilizzata una composizione delle due distribuzioni per rappresentare l'andamento dei documenti di una collezione (figura 4.3). Quindi un SRI dovrebbe

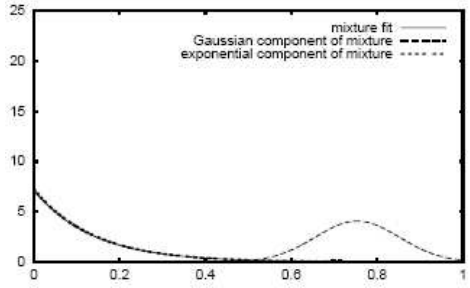


Figura 4.3: Distribuzione dei documenti (da [38])

utilizzare questo tipo di distribuzione per assegnare i valori di SRS ai documenti che reperisce.

È possibile quindi utilizzare un modello misto che comprende le caratteristiche di entrambe le distribuzioni. Si definisce il modello $p(x)$ formato dalle rispettive componenti $p(x|j)$:

$$p(x) = \sum_j P(j)p(x|j)$$

dove j identifica il singolo componente e $P(j)$ soddisfa la relazione $\sum_{j=1}^{2} P(j) = 1, 0 \leq P(j) \leq 1$. I due componenti sono dati dalla distribuzione esponenziale con media λ

$$P(x|1) = \lambda^{-\lambda x}$$

e dalla distribuzione gaussiana con media μ e varianza σ^2

$$P(x|2) = \frac{1}{\sqrt{2\pi}\sigma}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In conclusione si può dire che il modello esponenziale si adatta bene alla distribuzione dei documenti non pertinenti, mentre il modello gaussiano si adatta alla distribuzione dei documenti pertinenti. Nei casi in cui mancano informazioni riguardo la distribuzione di pertinenza all'interno della collezione di documenti, ad esempio nelle iniziative internazionali di valutazione, si può utilizzare un modello misto per avere una stima della distribuzione dello score dei documenti, tenendo però in considerazione che il numero di documenti non pertinenti è, in genere, molto maggiore di quello dei documenti pertinenti.

4.5 Conclusioni

In questo capitolo sono stati definiti i concetti di sensitività e specificità (paragrafo 4.1) ed è stato mostrato come questi concetti corrispondano a concetti utilizzati

nel campo del RI. È poi stato mostrato come è possibile valutare la stabilità delle metriche calcolando il tasso d'errore commesso nella scelta di un SRI rispetto ad un altro (paragrafo 4.3). Infine sono stati mostrati i modelli di distribuzione che meglio si adattano alla distribuzione degli score nei documenti di una collezione di test.

Viste queste considerazioni, nel prossimo capitolo definiremo una nuova metrica di valutazione che si pone l'obiettivo di avere una maggiore stabilità delle metriche fin'ora proposte.

Capitolo 5

Una nuova metrica: Average Distance Measure

Dopo aver proposto una rassegna di tutte le metriche proposte in letteratura con le loro caratteristiche, viene descritta ora una nuova metrica che ha come obiettivo quello di andare oltre al concetto di pertinenza e reperimento binario utilizzando un modello continuo per entrambe le dimensioni.

In questo capitolo viene dapprima esposta la definizione formale della metrica ADM (paragrafo 5.1) e di alcune sue possibili estensioni (paragrafo 5.2). Vedremo poi i risultati delle prime valutazioni sperimentali effettuate su ADM utilizzando due diverse collezioni di test (paragrafo 5.3).

5.1 La definizione di ADM

La metrica *Average Distance Measure* (ADM) è stata proposta nel 2001 in [41] (si vedano anche [16, 17]) e si propone di valutare le prestazioni degli SRI adottando un modello di pertinenza continuo. Introduciamo la terminologia utilizzata in ADM. Indichiamo con il termine *User Relevance Score* (URS) un valore (che può essere visto come un numero reale compreso tra 0 e 1) assegnato dall'utente ad ogni documento su una certa interrogazione che indica il valore di pertinenza.

Indichiamo con il termine *System Relevance Score* (SRS) il valore (che può essere visto come un numero reale compreso tra 0 e 1) che un SRI assegna ad un certo documento relativamente ad una certa interrogazione. L'SRI, poi, ordinerà i documenti reperiti in base a questo punteggio in modo da avere una lista ordinata di elementi da restituire all'utente.

È possibile dire che SRS ed URS sono delle misure dell'ammontare di pertinenza, e, per questo motivo, vi è un'importante differenza con l'RSV (si veda paragrafo 2.1). Mentre l'RSV è un punteggio utile assegnato ai documenti solamente con lo scopo di creare un ordinamento, l'SRS è una vera e propria stima dell'ammontare di pertinenza dei documenti.

Con l'utilizzo di queste due definizioni è facile capire che ADM utilizza il concetto di pertinenza continua e, per simmetria, di reperimento continuo. È anche immediato il passaggio alla situazione classica con pertinenza e reperimento binari semplicemente con l'utilizzo di una soglia (ad esempio a 0.5) che permette di separare l'insieme di documenti pertinenti da quelli non pertinenti e l'insieme di documenti reperiti da quelli non reperiti.

ADM è una metrica che misura la distanza media tra i valori URS ed i valori SRS. In modo più formale possiamo dire che, per ogni interrogazione q, è possibile definire due diversi pesi ($SRS_q(d_i)$ e $URS_q(d_i)$) per ogni documento d presente nella collezione D relativamente all'interrogazione q. ADM sull'interrogazione q è definita come la distanza media tra $SRS_q(d_i)$ e $URS_q(d_i)$:

$$ADM_q = 1 - \frac{\sum_{d_i \in D} |SRS_q(d_i) - URS_q(d_i)|}{|D|}$$

dove $|D|$ è il numero di documenti presenti nella collezione. ADM_q avrà valori compresi tra 0 ed 1 e, facendo la media di ADM_q su più interrogazioni, si ottiene una metrica di valutazione dell'efficacia di un SRI.

È possibile dare anche una spiegazione grafica di questa metrica. Consideriamo ogni documento della collezione con i suoi valori SRS ed URS (compresi nell'intervallo [0..1]) e visualizziamoli su di un piano cartesiano $[0..1]^2$. Così facendo ogni documento sarà rappresentato da un punto del piano cartesiano (figura 5.1). Un SRI ideale posizionerà tutti i documenti sulla linea in cui i valori di SRS coincidono con quelli di URS (cioè sulla bisettrice del piano cartesiano). A questo punto è necessario disporre di una misura di distanza tra un punto e la linea di classificazione ideale. In ADM non viene utilizzata la distanza convenzionale tra punto e retta (ottenuta misurando la lunghezza delle linea ortogonale tra il punto e la retta) in quanto si ipotizza che i valori di URS siano sempre corretti e quindi non vadano cambiati. Per questo motivo ADM utilizza la distanza tra il punto che rappresenta il documento ed il punto sulla retta di classificazione ideale con lo stesso valore di ascissa.

5.2 Estensioni di ADM

Vediamo ora alcune possibili estensioni della metrica ADM.

5.2.1 ADM@N e QADM

È anche possibile definire la metrica $ADM@N$ che corrisponde al valore di ADM calcolata solamente sui primi N documenti reperiti analogamente a come viene fatto per altre metriche come ad esempio $P@N$ (si veda paragrafo 3.1.3).

È possibile definire una variante di ADM utilizzando come misura di distanza la somma del quadrato delle distanze (QADM). Ciò vuol dire che si considera la distanza come la lunghezza delle linea ortogonale tra il punto e la retta elevata

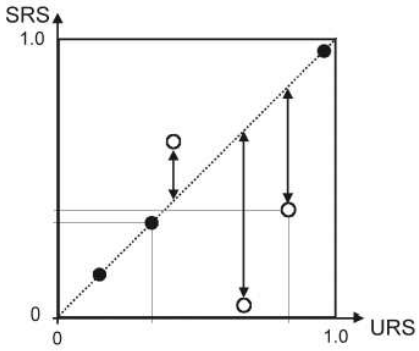


Figura 5.1: Piano SRS/URS (da [16])

al quadrato. Questa scelta influenzerebbe la metrica in modo tale da favorire gli SRI che sono più precisi nel valutare tutti quanti i documenti sfavorendo quelli che valutano molto bene certi documenti e peggio altri (avendo ADM uguale). Si favoriscono quindi gli SRI che commettono molti errori meno grossolani rispetto a quelli che sbagliano palesemente anche solo qualche volta.

In figura 5.2 è riportato un esempio della differenza di caratteristiche valutate tra ADM tradizionale e QADM. Si può vedere che, per quanto i due SRI abbiano lo stesso valore di ADM, l'SRI che effettua errori meno grossolani nella classificazione dei documenti risulta avere un valore maggiore di QADM.

5.2.2 Average Distance Precision e Average Distance Recall

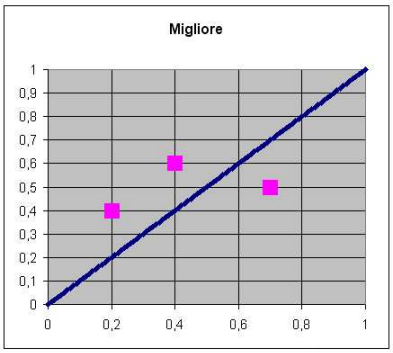
Il piano cartesiano usato per definire ADM consente di avere una definizione grafica su anche per le metriche Precision e Recall. È sufficiente dividere in quattro parti il piano in modo da avere i documenti reperiti ma non pertinenti in alto a sinistra (che indichiamo con A), i documenti reperiti e pertinenti nel settore in alto a destra (che indichiamo con B), i documenti non pertinenti e non reperiti nel settore in basso a sinistra (che indichiamo con C) ed i documenti pertinenti ma non reperiti in basso a destra (che indichiamo con D) (figura 5.3).

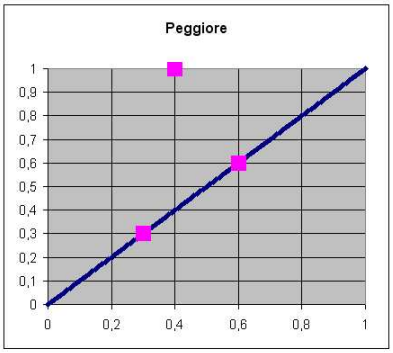
A questo punto possiamo calcolare Precision e Recall in base al numero di documenti presenti in queste aree. Si ottengono le seguenti formule:

$$Precision = \frac{|B|}{|A \cup B|}$$

$$Recall = \frac{|B|}{|D \cup B|}.$$

Inoltre è possibile dire che i documenti entro una certa distanza dalla retta ideale in cui URS=SRS sono stati correttamente giudicati dall'SRI, mentre quelli oltre





Peggiore			Migliore	
URS	**SRS**		**URS**	**SRS**
0,30	0,30		0,2	0,4
0,40	1,00		0,4	0,6
0,60	0,60		0,7	0,5

ADM(Peggiore)=1-(0+0,6+0)/3=0,8
QADM(Peggiore)= 1-(0,6^2)/3=**0,88**

ADM(Migliore)=1-(0,2+0,2+0,2)/3=0,8
QADM(Migliore)= 1-(0,2^{2+} 0,2^2+ 0,2^2)/3=**0,96**

Figura 5.2: Un esempio del confronto tra ADM e QADM

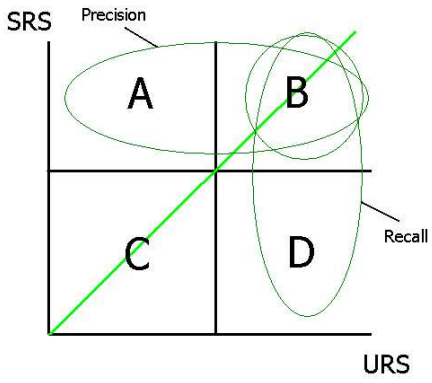


Figura 5.3: Un ulteriore modo per definire Precision e Recall

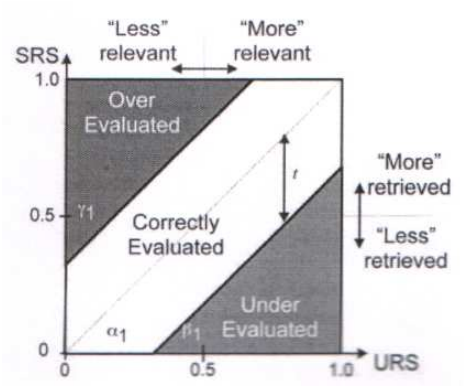


Figura 5.4: Documenti sopra e sottovalutati su cui vengono calcolati ADP ed ADR (da [17])

una certa distanza non sono stati correttamente giudicati (si veda la figura 5.4). In particolare quelli che si trovano nella parte superiore alla retta ideale saranno stati sopravvalutati dall'SRI, mentre quelli che si trovano al di sotto saranno stati sottovalutati dall'SRI.

Viste le considerazioni su Precision e Recall sulla sopravvalutazione e sottovalutazione dei documenti è possibile definire due misure derivate da ADM in modo da approssimare le caratteristiche degli SRI valutate da Precision e Recall. Average Distance Precision (ADP) è una metrica che calcola ADM solamente sui documenti sopravvalutati, mentre Average Distance Recall (ADR) calcola ADM solamente sui documenti sottovalutati (si veda la figura 5.4).

$$ADP_q = 1 - \frac{\sum_{d_i \in DO} |SRS_q(d_i) - URS_q(d_i)|}{|D|}$$

$$ADR_q = 1 - \frac{\sum_{d_i \in DU} |SRS_q(d_i) - URS_q(d_i)|}{|D|},$$

dove DO è l'insieme dei documenti sopravvalutati e DU è l'insieme dei documenti sottovalutati.

5.2.3 Curve ADP/ADR

Utilizzando le metriche Precision e Recall è possibile calcolare le curve Precision/Recall (si veda paragrafo 3.1.3). È possibile vedere le metriche ADP ed ADR come delle estensioni, in un modello di pertinenza continuo, di queste due metriche fondamentali. Sulla base di queste considerazioni si definisce un indicatore che mantiene questo stesso tipo di analogia con le curve Precision/Recall. Le *curve ADP/ADR* sono costruite su un piano cartesiano in cui poniamo i valori di ADR sull'asse x ed i valori di ADP sull'asse y. Si calcolano i valori di $ADP@N$ ed $ADR@N$, cioè

le metriche calcolate dopo N documenti reperiti, e si posizionano i corrispondenti valori nel piano.

Questo indicatore, proposto per la prima volta in questa tesi, è adatto a valutare al meglio l'andamento di queste due metriche rispetto al numero di documenti considerati.

5.3 Le valutazioni sperimentali su ADM

Dopo aver descritto la metrica ADM, vediamo i risultati di alcune valutazioni sperimentali che, in passato, sono state effettuate per comprendere l'efficienza di ADM rispetto alle metriche classiche.

L'obiettivo di queste valutazioni è di scoprire se ADM è più "sensibile" rispetto alle altre metriche, cioè se è in grado di valutare in modo corretto l'efficacia degli SRI, almeno alla pari di altre metriche, utilizzando però meno informazioni di queste ultime. Ad esempio si può voler scoprire se il calcolo di ADM effettuato utilizzando meno documenti o meno interrogazioni dà gli stessi risultati del calcolo di MAP.

5.3.1 Le valutazioni sulla collezione TREC8

La prima valutazione che è stata compiuta su ADM [17] utilizza la collezione di test (paragrafo 2.4) utilizzata nell'ottava edizione dell'iniziativa TREC.

Sono stati utilizzati, per misurare ADM, i giudizi di pertinenza, che in TREC sono binari, ed i risultati di classificazione dei singoli SRI partecipanti a TREC riguardo ai documenti per ogni topic.

Con queste informazioni è possibile calcolare i valori di ADM per ogni SRI su ogni topic. Si può notare quindi che non è necessario avere a disposizione l'intera collezione di documenti per effettuare delle valutazioni sperimentali di ADM: servono solo i giudizi di pertinenza ed i risultati di classificazione.

Il procedimento è il seguente. Si calcolano i valori di ADM per ogni esecuzione (run) di ogni SRI. Per run si intende una singola esecuzione di un SRI sulla collezione di documenti in quanto ogni SRI può essere testato più di una volta con piccole modifiche su esso. Dopo aver calcolato i valori di ADM, si ordinano i run in base alla misura di valutazione utilizzata e si calcola la correlazione di Kendall tra gli ordinamenti ottenuti con metriche diverse.

La similarità tra due ordinamenti è definita dalla τ di Kendall. Questa misura calcola il minimo numero di scambi necessario a trasformare un ordinamento nell'altro normalizzata in modo tale che due ordinamenti uguali producono una correlazione di 1.0 e che la correlazione attesa di due ordinamenti casuali è 0.

È stata calcolata la correlazione di Kendall per capire se l'andamento di ADM è simile a quello delle misure standard utilizzate, e calcolate, in TREC8.

Possiamo definire la correlazione come la tendenza di due grandezze a variare in modo concomitante. In questo caso le due grandezze analizzate saranno i valori di ADM e, ad esempio, i valori di Average Precision.

	ADM
MAP	0.876
R-Prec	0.844

Tabella 5.1: Correlazione tra ADM e le metriche standard in TREC8

Un problema nell'utilizzo dei dati di TREC8 è il seguente. Gli SRI partecipanti a TREC8 non stimano l'ammontare di pertinenza (SRS), utilizzato da ADM per valutare le prestazioni, ma semplicemente calcolano un RSV (si vedano i paragrafi 2.1 e 5.1) utilizzato solamente per ottenere un ordinamento dei documenti (rank). Per questo motivo ogni trasformazione degli RSV che preserva l'ordine è equivalente. Inoltre, nei dati di TREC8 è stata riscontrata un'inconsistenza tra i valori di RSV assegnati ai vari documenti e l'ordinamento proposto dagli SRI. Per questo motivo, nelle sperimentazioni, è stato deciso di utilizzare le informazioni derivanti dall'ordinamento dei documenti (quello che è stato usato in TREC8 per valutare gli SRI), e di assegnare ad ogni documento un valore di SRS con una distribuzione lineare. Quindi, se ad esempio un SRI ha reperito 1000 documenti, al primo verrà assegnato un valore di SRS pari a 1, al secondo un valore pari a 0.999 fino all'ultimo che avrà valore 0.

Questa scelta ha probabilmente portato a risultati non precisi in quanto è più corretto supporre una distribuzione non lineare dei valori di SRS nei documenti reperiti (si veda il paragrafo 4.4).

Comunque, nonostante questi problemi, i risultati indicano che ADM correla bene con le metriche standard utilizzate in TREC8 e, quindi, si propone come valida alternativa a queste. I risultati in tabella 5.1 mostrano come ADM correla con le metriche MAP ed R-Prec utilizzate in TREC. Questa correlazione ha valori buoni se paragonata con quella tra le metriche MAP ed R-Prec che è pari a 0.902. Naturalmente si suppone che ADM abbia prestazioni migliori in situazioni con pertinenza e reperimento non binario a differenza di quanto avviene in TREC.

5.3.2 Le valutazioni sulla collezione NTCIR

Per valutare ADM in una situazione che considera un modello di pertinenza non binario, è stata effettuata una seconda sperimentazione su ADM utilizzando la collezione di test di NTCIR-4 (si veda il paragrafo 2.4.2). Analogamente ai test svolti con i dati dell'iniziativa TREC, sono state calcolate le correlazioni di ADM con le metriche utilizzate nell'iniziativa NTCIR (si veda [16]).

In NTCIR la pertinenza è a 4 livelli, a differenza di TREC dove è binaria. Questo fatto dovrebbe favorire le prestazioni di ADM in quanto ci si avvicina di più al caso continuo.

Sulla collezione NTCIR sono stati calcolati valori di $ADM@N$ sia utilizzando le informazioni dell'ordinamento dei documenti fatto dagli SRI (come fatto con la collezione di TREC), sia utilizzando i valori di SRS assegnati dai sistemi partecipanti.

Le 4 possibili categorie di pertinenza sono state assegnate a valori di URS compresi tra 0 ed 1. Come visto in 2.4.2, le 4 categorie di pertinenza sono:

- S (completamente pertinenti)
- A (pertinenti)
- B (parzialmente pertinenti)
- C (non pertinenti)

Queste categorie sono state assegnate rispettivamente ai valori $S = 7/8$, $A = 5/8$, $B = 3/8$, $C = 1/8$ al fine di poter sopravvalutare e sottovalutare anche nel caso delle categorie estreme S e C.

Oltre che utilizzare quattro categorie di pertinenza, è stata calcolata ADM su due sole categorie di pertinenza ottenute collassando le quattro originali in due modi diversi analogamente a quanto viene fatto in NTCIR. La versione rigida porta la categoria S ad un valore di URS pari a 1 e le altre tre categorie al valore 0. La versione rilassata del mapping, invece fa corrispondere le categorie S e A al valore 1 di URS e le categorie B e C al valore 0.

Le sperimentazioni portano a diverse conclusioni. Utilizzando le informazioni sull'ordinamento dei documenti restituito dagli SRI, e quindi assegnando valori di URS decrescenti in modo lineare (come nelle valutazioni fatte con i dati di TREC8), le correlazioni tra $ADM@N$ e le metriche utilizzate in NTCIR danno dei valori più che accettabili (si veda nella tabella 5.2 un confronto con le metriche MAP e R-Prec) considerando che la correlazione tra MAP e R-Prec è, in questo caso, pari a 0.90.

Nonostante la bontà della correlazione, è stato notato un fenomeno particolare che non è stato ancora possibile spiegare in modo soddisfacente. Il valore della correlazione di $ADM@100$ e $ADM@200$ con le metriche standard utilizzate in NTCIR-3 è molto basso (anche più basso di quella con ADM calcolata utilizzando tutti i documenti reperiti). In altre parole, ci si aspetterebbe che il valore di correlazione cresca al crescere di n, ma questo è vero per tutti gli n esclusi $n = 100$ ed $n = 200$.

I risultati che considerano solo due livelli di pertinenza, e che quindi fanno il mapping dei quattro livelli su una scala binaria danno delle correlazioni più basse rispetto al caso che considera quattro livelli (si veda la tabella 5.3).

Un ulteriore risultato è quello ottenuto calcolando i valori di ADM basandosi sui valori di SRS assegnati dagli SRI. In questo caso le correlazioni con le metriche utilizzate nell'iniziativa non sono buone. Questo risultato è probabilmente dovuto al fatto che i progettisti degli SRI si sono concentrati sull'avere degli ordinamenti significativi dei documenti reperiti e non si sono preoccupati del valore di RSV da assegnare ai vari documenti in quanto non veniva valutato in NTCIR.

	MAP	R-Prec
$ADM^{rank}_{(4)}$@5	0.75	0.76
$ADM^{rank}_{(4)}$@10	0.79	0.80
$ADM^{rank}_{(4)}$@20	0.80	0.82
$ADM^{rank}_{(4)}$@50	0.79	0.80
$ADM^{rank}_{(4)}$@100	0.72	0.72
$ADM^{rank}_{(4)}$@200	0.13	0.13
$ADM^{rank}_{(4)}$	0.35	0.37

Tabella 5.2: Correlazione tra ADM(rank)@N su 4 livelli e le metriche standard in NTCIR-3

	MAP	R-Prec
$ADM^{rank}_{(2)}$@5[rilassata]	0.51	0.51
$ADM^{rank}_{(2)}$@10[rilassata]	0.54	0.54
$ADM^{rank}_{(2)}$@5[rigida]	0.52	0.53
$ADM^{rank}_{(2)}$@10[rigida]	0.58	0.59

Tabella 5.3: Correlazione tra ADM(rank)@N su 2 livelli e le metriche standard in NTCIR-3

5.4 Conclusioni

In questo capitolo è stata definita una nuova metrica — ADM — che utilizza un modello continuo di pertinenza e di reperimento e sono state definite le curve ADP/ADR che sono proposte in questa tesi per la prima volta. Sono stati inoltre presentati i risultati delle valutazioni preliminari effettuate su questa metrica utilizzando due diverse collezioni di test, che dimostrano come ADM si comporti similmente a misure tradizionali quali R-Prec e MAP, su condizioni di pertinenza binaria e a 4 livelli.

Nella seconda parte di questa tesi vedremo ulteriori valutazioni, concettuali e sperimentali, su questa metrica.

Parte II

Risultati

Capitolo 6

Gli obiettivi delle sperimentazioni e la metodologia utilizzata

Nella prima parte di questa tesi sono stati presentati i concetti di base della valutazione dell'efficacia degli SRI, sono state descritte le metriche proposte in letteratura, sono state esaminate alcune caratteristiche di queste metriche ed è stata illustrata ADM, una recente metrica che utilizza modelli continui di pertinenza e di reperimento.

In questo capitolo vengono riassunti i concetti fondamentali definiti nella prima parte della tesi (paragrafo 6.1) e delineati gli obiettivi delle sperimentazioni e delle analisi effettuate su ADM utilizzando diverse collezioni di test (paragrafo 6.2). Viene inoltre definita la metodologia utilizzata nelle valutazioni sperimentali effettuate (paragrafo 6.3).

6.1 I concetti fondamentali

I dati che servono per effettuare una valutazione sperimentale di una metrica sono reperibili dalle iniziative di valutazione di SRI. Un'iniziativa di valutazione mette a disposizione della comunità scientifica numerose risorse; in genere queste risorse sono le seguenti:

- la collezione di documenti su cui viene effettuato il reperimento;
- un insieme di interrogazioni di reperimento;
- le classificazioni, effettuate dagli SRI partecipanti, dei vari documenti sulle varie interrogazioni;
- un insieme di giudizi di pertinenza, effettuati da umani, per tutti i documenti su ogni interrogazione;

- i risultati di efficacia degli SRI misurati con certe metriche di valutazione scelte opportunamente.

Per effettuare delle valutazioni sperimentali è necessario utilizzare i dati di una collezione di test che è composta dalla collezione di documenti, dalle interrogazioni e dai giudizi di pertinenza. Le iniziative internazionali di valutazione (si veda il paragrafo 2.4) mettono a disposizione delle comunità scientifica le collezioni di test utilizzate annualmente per valutare l'efficacia degli SRI partecipanti all'iniziativa. Per le valutazioni sperimentali qui presenti sono state utilizzate tre collezioni di test: TREC8, TREC13 TeraByte e INEX 2004.

La metrica che viene valutata nelle presenti sperimentazioni è l'ADM, che utilizza il concetto di pertinenza continua e di reperimento continuo misurando la distanza media tra i valori URS ed i valori SRS (si veda il paragrafo 5.1).

Per confrontare la metrica ADM con le metriche standard calcolate nelle iniziative di valutazione sono utilizzate la correlazioni di Kendall e di Spearman tra gli ordinamenti degli SRI prodotto dalle metriche. È possibile definire la correlazione come la tendenza di due grandezze a variare in modo concomitante. Questi due tipi di correlazioni confrontano due ordinamenti calcolando il numero necessario di spostamenti per far coincidere i due ordinamenti.

6.2 Gli obiettivi delle valutazioni

In questa tesi ci si propone di valutare ulteriormente a quanto è stato fatto in passato la metrica ADM. La valutazione è duplice: si intende valutare la metrica da un punto di vista concettuale analizzandone le caratteristiche e le peculiarità e da un punto di vista sperimentale calcolando il grado di correlazione con altre metriche e la sua stabilità.

Quindi, sulla base dei risultati ottenuti nelle precedenti valutazioni, viene proposta una nuova classificazione delle metriche proposte in letteratura basata sui diversi tipi di pertinenza. L'obiettivo è quello di illustrare le peculiarità di ADM e come essa è in grado di adattarsi a diverse situazioni di pertinenza e di reperimento.

Dopo aver analizzato la metodologia utilizzata per effettuare le valutazioni passate di ADM, un ulteriore obiettivo è quello di proporre una metodologia standardizzata che può essere adottata nuovamente per ogni collezione di test su cui si desidera valutare l'efficacia di ADM.

Le ulteriori valutazioni sperimentali effettuate hanno come obiettivo quello di ottenere una migliore comprensione dell'efficacia della metrica ADM. Si suppone che ADM sia una metrica più sensibile delle altre proposte, perciò si vuole capire il suo livello di sensibilità analizzando quanti documenti è necessario considerare nel calcolo di ADM per ottenere un valore di significatività pari alle metriche standard.

Nelle valutazioni sperimentali si è preferito utilizzare la metrica ADM piuttosto che la metrica QADM (si veda paragrafo 5.2.1). Questa scelta è stata fatta per poter confrontare i risultati ottenuti con quelli trovati in passato [16, 17].

Si vuole inoltre che l'effettuazione delle valutazioni porti alla proposta di un'estensione della metrica ADM che possa colmare le possibili mancanze riscontrate.

6.3 La metodologia utilizzata per le valutazioni

Abbiamo visto che esistono iniziative di valutazione diverse, ma che ognuna di queste rende disponibile alla comunità scientifica lo stesso tipo di informazioni. Quindi sarebbe desiderabile disporre di una metodologia che sia possibile utilizzare per valutare ADM ogniqualvolta un'iniziativa di valutazione rende disponibili i suoi dati.

Per standardizzare la valutazione di ADM si è deciso di utilizzare una base di dati in cui memorizzare tutti i dati necessari: sia quelli provenienti da un'iniziativa di valutazione, sia quelli calcolati in seguito. In questo modo è possibile importare i dati grezzi provenienti dalle iniziative di valutazione in tabelle con struttura sempre uguale. È così possibile avere dei dati di partenza omogenei su cui è possibile effettuare sempre le stesse interrogazioni e routine.

Al fine di confrontare le prestazioni di ADM con quelle delle metriche che vengono utilizzate in una particolare collezione di test è necessario effettuare i seguenti passi:

- importare in una base di dati i giudizi di pertinenza;
- importare in una base di dati le classificazioni effettuate dagli SRI;
- costruire una tabella contenente i nomi e gli identificativi degli SRI;
- costruire una tabella (o una vista) nella base di dati che faccia il join dei giudizi di pertinenza con le classificazioni effettuate dagli SRI (per ogni documento su ogni interrogazione);
- nella tabella appena costruita è necessario impostare a 0 la colonna del giudizio di pertinenza per i record che hanno valore NULL. Questa operazione è derivata dall'utilizzo della tecnica del pooling. Utilizzando questa tecnica, che sceglie opportunamente i documenti da far giudicare ad un giudice umano, si assume che i documenti non classificati da un giudice vadano considerati non pertinenti;
- analizzare i dati per trovare eventuali SRI che svolgono il compito in maniera non ottimale (ad esempio che non reperiscono il numero richiesto di documenti per ogni interrogazione) in modo da effettuare le valutazioni utilizzando solamente dati significativi;
- calcolare i valori di ADM per ogni sistema su ogni interrogazione utilizzando i valori dei giudizi di pertinenza (normalizzati tra 0 ed 1) come URS e i valori di classificazione degli SRI come SRS;
- importare in una base di dati i i valori delle metriche, calcolati durante l'iniziativa di valutazione, per ogni SRI su ogni interrogazione;

- calcolare le correlazioni tra i valori di ADM ed i valori delle metriche utilizzate nella valutazione.

Naturalmente la procedura appena descritta è molto generale e va adattata a seconda di ciò che si vuole valutare e dei dati che si stanno utilizzando. Ad esempio è possibile calcolare, oltre ad ADM, anche i valori di $ADM@N$ (si veda paragrafo 5.1) per certi valori di N, oppure è possibile calcolare anche i valori di ADP e di ADR (si veda paragrafo 5.1) per poterli confrontare poi con le metriche classiche.

Questa metodologia è utile per effettuare valutazioni sperimentali di metriche in quanto questa pratica risulta essere altrimenti complessa. La difficoltà è data dalla notevole quantità di dati presenti in una collezione di test e dalle numerose computazioni che è necessario considerare.

6.4 Conclusioni

In questo capitolo sono stati richiamati dalla prima parte i concetti necessari per discutere dei risultati delle valutazioni sperimentali effettuate in questa tesi, sono stati definiti gli obiettivi delle valutazioni sperimentali effettuate ed è stata descritta la metodologia proposta ed utilizzata per le valutazioni.

Capitolo 7

Una nuova classificazione delle metriche di valutazione

In questo capitolo, dopo aver definito i criteri di classificazione (paragrafo 7.1), sarà presentata una nuova classificazione delle 45 metriche proposte in letteratura (descritte nel capitolo 3) basata sul concetto di pertinenza e di reperimento che le metriche utilizzano per valutare gli SRI (paragrafo 7.2).

7.1 I criteri di classificazione

Nel capitolo 3 abbiamo presentato una breve descrizione di tutte le metriche che sono state proposte in letteratura. Vediamo ora una classificazione di tutte le metriche note che possa aiutare nella scelta di quale misura usare per valutare l'efficacia di un SRI.

Questa nuova classificazione delle metriche di valutazione è stata sottoposta alla European Conference of Information Retrieval [19] del 2006 (ECIR06) ed attualmente è in fase di revisione.

La classificazione è basata sulla nozione di pertinenza (se e quanto un documento è pertinenti) e di reperimento (se e quanto un documento è reperito). L'obiettivo è posizionare ogni metrica nota all'interno di una griglia che indichi quanto essa si adatta alle varie possibili nozioni di pertinenza e di reperimento.

Le descrizioni delle metriche e gli opportuni riferimenti bibliografici sono già state fornite nel capitolo 3. La tabella 7.1 presenta l'anno in cui la metrica è stata proposta, il riferimento bibliografico in cui è possibile trovare una definizione completa della metrica, il nome delle metrica e a quale/i categoria/e la metrica è adatta. Inoltre viene indicato se la metrica è o meno adottata nelle principali iniziative di valutazione. La maggior parte delle metriche è descritta nei più utilizzati libri di testo ([35, Cap. 8], [58, Cap. 7], [49, Cap. 5], [3, Cap. 4],[2, Cap. 3]).

Le nozioni di pertinenza e di reperimento che una metrica può adottare sono divise in binaria, a categorie oppure continua. Quindi, a seconda di come è stata progetta la valutazione, certe metriche saranno più adatte di altre a valutare

l'efficacia degli SRI. Le varie categorie che ho previsto per classificare le metriche sono:

- Pertinenza binaria e reperimento binario;
- Pertinenza binaria e reperimento a categorie;
- Pertinenza binaria e reperimento continuo;
- Pertinenza a categorie e reperimento binario;
- Pertinenza a categorie e reperimento a categorie;
- Pertinenza a categorie e reperimento continuo;
- Pertinenza continua e reperimento binario;
- Pertinenza continua e reperimento a categorie;
- Pertinenza continua e reperimento continuo.

7.2 La classificazione delle metriche

È possibile notare (si veda la tabella 7.1) che, man mano le metriche sono state proposte negli anni, è cambiato il concetto di pertinenza e di reperimento che utilizzano per valutare l'efficacia degli SRI. Le prime metriche considerano, forse semplificando, una pertinenza ed un reperimento di tipo binario, mentre quelle che sono state proposte in seguito hanno cercato di adattarsi alle caratteristiche di reperimento degli SRI che valutano e hanno utilizzato una pertinenza ed un reperimento di tipo non binario.

Una possibile variante nella classificazione è di sostituire la pertinenza ed il reperimento a categorie con le voci *ordinamento parziale* e *ordinamento totale*. Si considera l'ordinamento totale nel caso in cui i giudizi di pertinenza o il reperimento ponga ogni documento ad un livello diverso dagli altri, mentre si considera l'ordinamento parziale nel caso in cui dei documenti si possano trovare allo stesso livello di pertinenza o di reperimento. Nella classificazione qui proposta è stato deciso di fondere assieme queste due categorie e considerare se la singola metrica considera o meno una forma di ordinamento dei documenti che non sia quello a valori continui.

7.3 Conclusioni

In questo capitolo sono stati definiti i criteri di una nuova classificazione delle metriche di valutazione dei SRI e sono state classificate le 45 metriche proposte in letteratura (descritte nel capitolo 3) in base al concetto di pertinenza e di reperimento che le metriche utilizzano per valutare gli SRI.

In questa classificazione la metrica ADM, a differenza delle altre, è risultata essere in grado di adattarsi ad ogni situazione di pertinenza e di reperimento.

Year	Relevance: / Retrieval:	Binary B	Binary R	Binary C	Rank B	Rank R	Rank C	Cont. B	Cont. R	Cont. C	TREC	INEX	NTCIR	CLEF
1955	Precision [58]	•												×
1955	Recall [58]	•												×
1965	Generality Factor [58]	•												
1966	Fallout [58]	•												
1966	Normalized Recall [35]		•											
1966	Normalized Precision [35]		•											
1967	R/P curve [58]		•								×	×	×	×
1967	R/fallout curve [58]		•											
1968	Expected Search Length [13]		•											
1968	Sliding Ratio [35]								•					
1969	E-Measure F-measure [58]	•											×	
1971	Novelty Ratio [35]	•												
1971	Coverage Ratio [35]	•												
1971	Relative Recall [35]	•												
1971	Recall effort [35]	•												
1973	Utility [49]	•												
1975	MAP [2]		•								×	×	×	
1975	Interpolated MAP [2]		•											
1975	P@N [2]		•								×	×	×	
1975	R-Precision [2]		•								×		×	
1990	Satisfaction [35]								•					
1990	Frustration [35]								•					
1990	Total [35]								•					
1991	Usefulness measure [20]					•								
1994	Average Search Length [37]		•											
1995	NDPM [62]					•								
1998	Ranked Half Life [7]									•				
1998	Relative Relevance [7]					•								
2000	Classification accuracy [3]	•												
2000	DCG [29]		•			•			○				×	
2001	AWP [30]		•			•								
2001	Weighted R-Precision [30]		•			•								
2001	ADM [17]	•	•	•	•	•	•	•	•	•				
2003	XCG [31]		•			•			○			×		
2004	bpref [11]		•											
2004	Q-measure [46]		•			•						×		
2004	R-measure [46]		•			•						×		
2004	Tolerance to Irrelevance [14]		•									×		
2004	Estimated Ratio of Relevant [43]		•									×		
2005	Kendall, Spearman [8]					•								
2005	Normalized xCG [32]		•			•						×		
2005	Mean average nxCG at rank n [32]		•			•						×		
2005	Effort-precision/gain-recall @ std. gain-recall p. [32]		•			•						×		
2005	Non-interpolated mean average effort-precision [32]		•			•						×		
2005	Interpolated mean average effort-precision [32]		•			•						×		

Tabella 7.1: Una classificazione delle metriche per il RI (ordinate per anno)

Capitolo 8

Le valutazioni sperimentali su TREC8

In questo capitolo vengono presentati e discussi i risultati delle valutazioni sperimentali effettuate su ADM utilizzando la collezione di test TREC8 e i giudizi di pertinenza a 4 livelli. Nel paragrafo 8.1 viene descritto l'esperimento effettuato ed i dati utilizzati. Nel paragrafo 8.2 vengono presentati i risultati ottenuti che vengono poi discussi nel paragrafo 8.3.

8.1 L'esperimento

Come visto in precedenza (si veda il paragrafo 5.3.1), sono già state effettuate delle valutazioni sperimentali di ADM utilizzando i dati provenienti da TREC8, in cui si utilizza una nozione di pertinenza binaria.

Nel 2004 il gruppo di ricerca di Sormunen ha riclassificato la collezione di documenti di TREC8 su una scala a quattro livelli di pertinenza [52]. Questa riclassificazione è stata fatta utilizzando una metodologia analoga a quella utilizzata in TREC. Nonostante questo i risultati della classificazioni hanno dato diverse inconsistenze. Documenti giudicati pertinenti in TREC sono stati giudicati non pertinenti durante le classificazioni di Sormunen e documenti giudicati non pertinenti in TREC sono stati giudicati almeno parzialmente pertinenti dai giudici di Sormunen. In seguito a queste inconsistenze i documenti sono stati sottoposti nuovamente ai giudici per essere classificati e le inconsistenze sono diminuite ma, sostanzialmente, sono rimaste analoghe.

Sono state utilizzati assieme i dati del lavoro di Sormunen ed i dati provenienti da TREC8 per valutare nuovamente ADM e vedere come varia ADM e le sue correlazioni con le metriche standard, quando si passa da 2 a 4 livelli di pertinenza.

Avendo a disposizione dei giudizi di pertinenza su quattro livelli è stato possibile effettuare le stesse valutazioni che sono state compiute nel caso di NTCIR-3 (si veda il paragrafo 5.3.2). Sono stati quindi calcolati i valori di ADM utilizzando i quattro

		Sormunen						
		ADM[2rig]	ADM[2rel]	ADM[4]	R-Prec[rig]	MAP[rig]	R-Prec[rel]	MAP[rel]
TREC8	MAP	0.79	0.85	0.82	0.72	0.43	0.82	0.79
	R-Prec	0.79	0.84	0.80	0.68	0.46	0.78	0.79
	ADM	0.80	0.94	0.86	0.69	0.39	0.82	0.66

Tabella 8.1: Correlazioni tra le metriche calcolate su TREC e quelle calcolate con i giudizi di pertinenza di Sormunen

livelli di pertinenza (0, 1, 2, 3), ma anche mappando i quattro livelli a due soli livelli in due modi diversi:

- Mapping rigido: i livelli 0 e 1 diventano 0, i livelli 2 e 3 diventano 1;
- Mapping rilassato: il livello 0 resta 0, i livelli 1, 2 e 3 diventano 1.

Le linee guida di TREC [57] affermano che:

> "Only binary judgments ("relevant" or "not relevant") are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)."

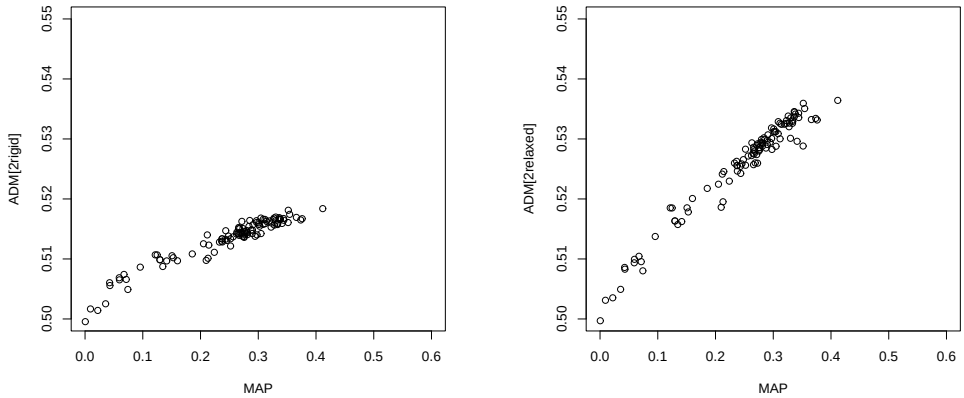
Di conseguenza è possibile immaginare, ancor prima di disporre dei risultati della sperimentazione, che ADM calcolata sui due livelli di pertinenza mappati nel modo rilassato, correli meglio di ADM calcolata utilizzando il mapping rigido.

8.2 I risultati

I risultati sono presentati nella tabella 8.1. Qui è possibile vedere i valori della correlazione di Kendall tra le metriche calcolate con la collezione di test TREC8 e le metriche calcolate con le riclassificazioni di Sormunen.

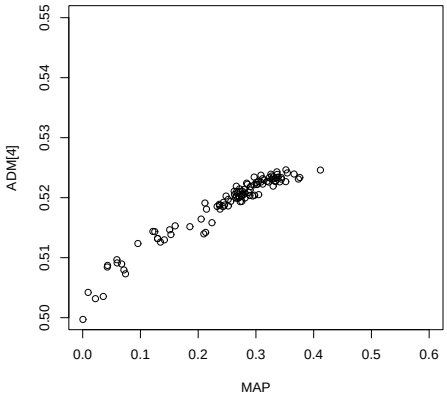
Nelle figure 8.1, 8.2 e 8.3 sono riportati i grafici di correlazione tra i valori di ADM, MAP ed R-Prec calcolati con i giudizi di pertinenza di TREC8 con i valori di ADM calcolata con i 4 livelli di pertinenza, con 2 livelli mappati in modo rilassato e con 2 livelli mappati in modo rigido. Il grafico di correlazione tra ADM e ADM[2rilassato] in figura 8.3(c) evidenzia il valore molto alto di correlazione pari a 0.94.

Sui dati di TREC8 unitamente alle ri-classificazioni di pertinenza sono stati calcolati i valori di ADP e di ADR a certi valori di N documenti reperiti per cercare di costruire un grafico (si veda il paragrafo 5.2.3) che approssimi il classico grafico Precision/Recall (si veda il paragrafo 3.1.2). È possibile vedere l'andamento di $ADP@N$ ed $ADR@N$ al variare di N in figura 8.4 ed in tabella 8.2.



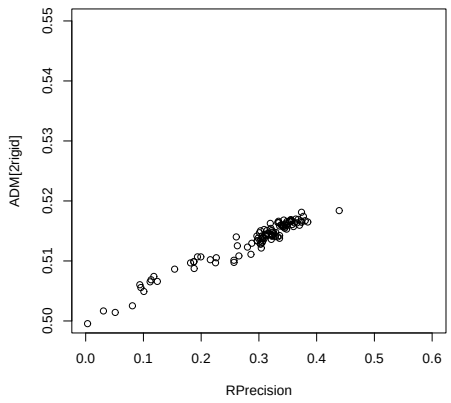
(a)

(b)

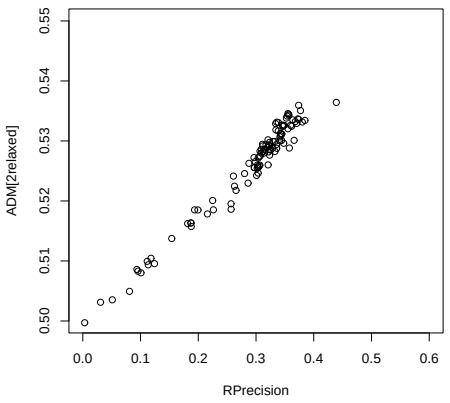


(c)

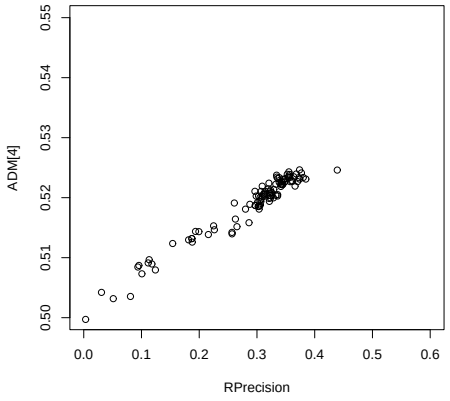
Figura 8.1: Correlazioni tra MAP calcolate su TREC e ADM calcolata con i giudizi di pertinenza di Sormunen



(a)

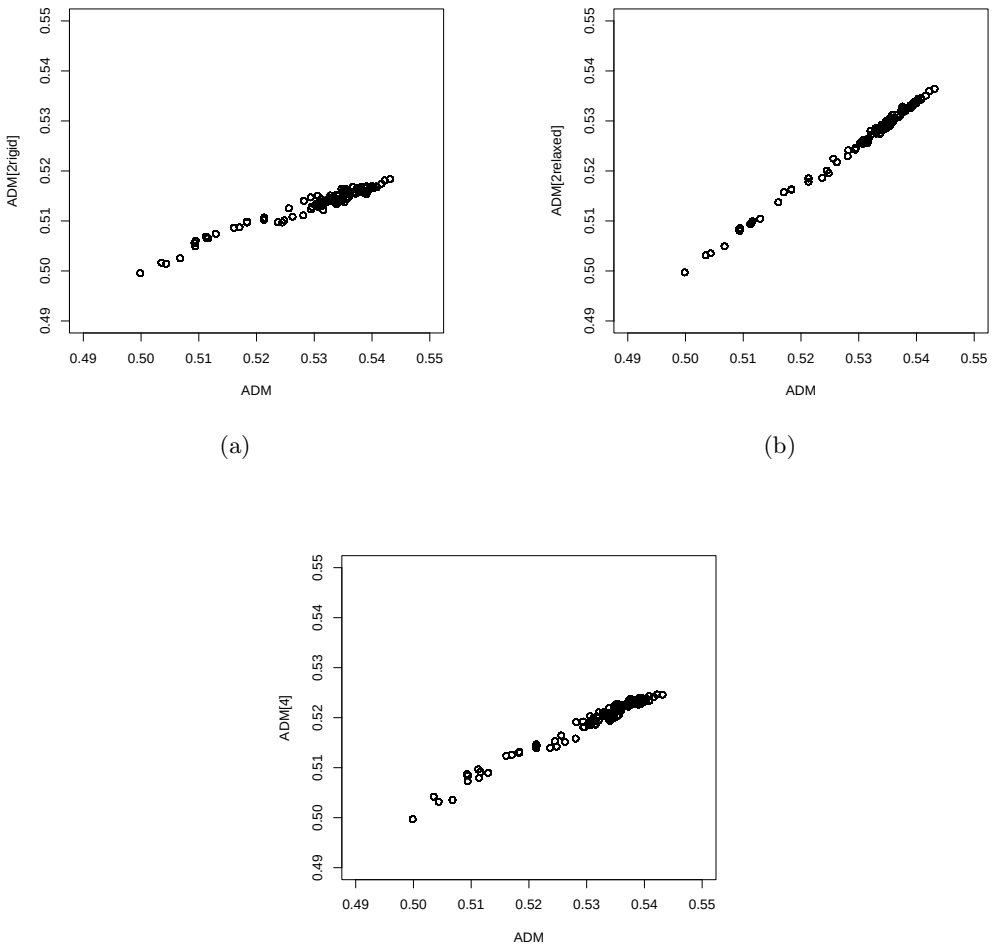


(b)



(c)

Figura 8.2: Correlazioni tra R-Prec calcolate su TREC e ADM calcolata con i giudizi di pertinenza di Sormunen



(a)

(b)

(c)

Figura 8.3: Correlazioni tra ADM calcolate su TREC e ADM calcolata con i giudizi di pertinenza di Sormunen

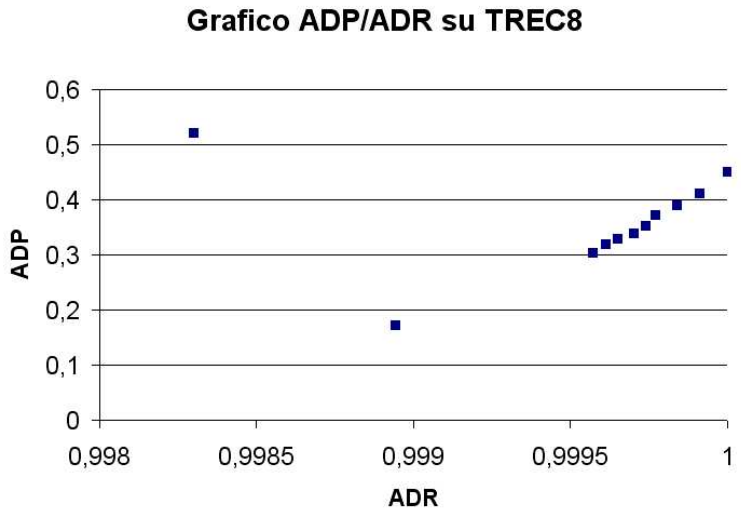


Figura 8.4: Andamento di ADP e ADR al variare di N su TREC8

N	ADR	ADP
1	1	0,45147
3	0,99991	0,41177
4	0,99984	0,38927
5	0,99977	0,37254
6	0,99974	0,35352
7	0,9997	0,33885
8	0,99965	0,32891
9	0,99961	0,32022
10	0,99957	0,30335
100	0,99894	0,17264
1000	0,9983	0,52063

Tabella 8.2: Andamento di ADP e ADR al variare di N su TREC8

8.3 Discussioni sui risultati

Come visto in 8.1, è stata utilizzata la collezione di test di TREC8 unitamente ai giudizi di pertinenza di Sormunen [52] per valutare l'efficacia di ADM.

Dai risultati è possibile vedere che, in generale, il calcolo di ADM su 2 livelli di pertinenza mappati in modo rilassato correlano, con le metriche calcolate in TREC8, meglio degli altri modi di calcolare ADM. Questo è naturale visto il fatto che i giudizi di pertinenza in TREC8 sono stati dati in un modo che viene meglio approssimato dal mapping rilassato. Ai giudici di TREC viene detto di giudicare pertinente un documento anche se contiene una minima parte di informazione pertinente. E questo è proprio quello che fa il mapping rilassato: porta al livello pertinente ogni documento che è anche solo parzialmente pertinente. Per questo motivo le correlazioni delle metriche calcolate con il mapping rilassato correlano meglio con le metriche calcolate in TREC8.

Anche dai grafici di correlazione (figure 8.1, 8.2, 8.3) è possibile vedere che i valori di ADM calcolati con due livelli di pertinenza mappati in modo rilassato correlano con le metriche standard meglio di ADM calcolata con quattro livelli di pertinenza.

Un ulteriore dato che va analizzato è la correlazione tra le metriche calcolate in TREC8 (metriche standard) e le stesse metriche ricalcolate (con un mapping rilassato) utilizzando i giudizi di pertinenza di Sormunen. Queste correlazioni sono di circa 0.79 (ultime due colonne della tabella 8.1). Il motivo per cui sono più basse di quanto ci si aspetta è, presumibilmente, dovuto al fatto che durante i giudizi di pertinenza di Sormunen [52] sono state rilevate notevoli incoerenze rispetto ai giudizi di pertinenza originali dei giudici di TREC8. La differenza nei giudizi di pertinenza è descritta, per l'appunto, dalla non perfetta correlazione tra, ad esempio, i valori di R-Precision calcolati in TREC8 e quelli ricalcolati utilizzando i giudizi di pertinenza di Sormunen.

Sui stessi dati sono, inoltre, stati calcolati i valori di ADP ed ADR ed è stato costruito un grafico ADP/ADR (figura 8.4). Si può subito notare che i valori di $ADR@N$ sono, per ogni N, prossimi a 1. Questo risultato è spiegabile dal fatto che, nelle sperimentazioni, è stata utilizzata l'informazione dell'ordinamento dei documenti effettuato dagli SRI da cui sono stati derivati dei valori di SRS. Come visto in precedenza (paragrafo 5.3.1), i valori di SRS sono stati derivati in modo lineare dall'ordinamento effettuato dagli SRI. Questo modo di derivare i valori di SRS ha fatto sì che i documenti siano, in maggior parte, sopravvalutati, in quanto, da un punto di vista dei valori di URS, ci saranno pochissimi documenti con valore 1 e moltissimi con valore 0. Dato che ADR viene calcolata solamente sui documenti sottovalutati (che sono pochi), le differenze tra SRS ed URS saranno minime in quanto questi documenti si troveranno principalmente a livelli di pertinenza prossimi a 1.

Un ulteriore risultato da sottolineare è che per valori piccoli di N (si veda la parte finale della curva in figura 8.4) i valori di ADP ed ADR crescono invece che diminuire come succede nelle curve Precision/Recall.

8.4 Conclusioni

In questo capitolo sono stati presentati e discussi i risultati delle valutazioni sperimentali effettuate su ADM utilizzando la collezione di test TREC8 e i giudizi di pertinenza a 4 livelli di Sormunen.

I risultati hanno mostrato come ADM correli meglio quando è calcolata sulla versione rilassata del mapping di 4 livelli su 2 livelli in quanto le metriche di TREC sono state calcolate utilizzando giudizi di pertinenza che utilizzano una classificazione analoga al mapping rilassato. La metrica ADM, in questo caso, ha comunque dimostrato di adattarsi a diversi modelli di pertinenza (binaria e a categorie).

Nel prossimo capitolo si vedranno e verranno discussi i risultati ottenuti utilizzando la collezione TREC13 TeraByte.

Capitolo 9

Le valutazioni sperimentali su TREC13 TeraByte

In questo capitolo vengono presentati gli esperimenti per valutare ADM utilizzando la collezione di test TREC13 TeraByte (paragrafo 9.1. Vengono poi presentati e discussi i risultati ottenuti nelle sperimentazioni (paragrafo 9.2). Infine si analizzando i risultati ottenuti calcolando il tasso d'errore della metrica ADM e della metrica MAP (paragrafo 9.3).

9.1 L'esperimento ed i risultati

Dopo aver utilizzato la collezione di test di TREC8, abbiamo testato ADM utilizzando un'ulteriore collezione di test. Abbiamo scelto di utilizzare una collezione di documenti di dimensioni notevoli (quella del track TeraByte (TB) di TREC13, si veda il paragrafo 2.4.1) al fine di capire se ADM ha prestazioni migliori su collezioni che mirano ad approssimare le dimensioni del web. Un'ulteriore caratteristica della collezione TREC13 TB è che i giudizi di pertinenza sono dati su una scala a 3 livelli.

La collezione di test di TREC13 TB utilizza dati proveniente da 70 SRI. Per queste valutazioni sono stati utilizzati i dati dei 57 SRI che hanno reperito almeno 1000 documenti per ogni interrogazione.

L'obiettivo del test è di capire se ADM riesce a valutare l'efficacia degli SRI utilizzando meno informazioni rispetto quelle necessarie alle metriche tradizionali. A questo scopo sono state calcolate le correlazioni tra $ADM@N$ (ADM calcolata dopo N documenti reperiti) e le metriche di riferimento utilizzate in TREC13 TeraByte al variare di N (si veda la tabella 9.1). Inoltre è stata calcolata la correlazione tra le due metriche classiche usate per il confronto (Mean Average Precision e R-Precision) per capire che livello di correlazione è da considerare come soglia per la significatività del risultato (si veda la figura 9.1).

Sono state, inoltre, calcolate le correlazioni tra i valori di $P@N$ ed $ADM@N$ per vedere se, come ci si aspetta, le correlazioni più alte si verificano a pari valore di N. I risultati sono presentati in tabella 9.2 ed in figura 9.2.

N	**1000**	**900**	**800**	**700**	**600**	**500**	**400**	**300**	**200**	**100**	**90**
MAP	0.90	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.92	0.90	0.88
R-Prec	0.91	0.91	0.92	0.91	0.91	0.92	0.92	0.93	0.94	0.90	0.88
N	**80**	**70**	**60**	**50**	**40**	**30**	**20**	**10**	**5**	**1**	
MAP	0.87	0.87	0.88	0.87	0.87	0.86	0.82	0.76	0.73	0.57	
R-Prec	0.86	0.87	0.87	0.85	0.85	0.84	0.81	0.76	0.72	0.59	

Tabella 9.1: Correlazioni tra $ADM@N$ e le due metriche standard MAP e R-Prec

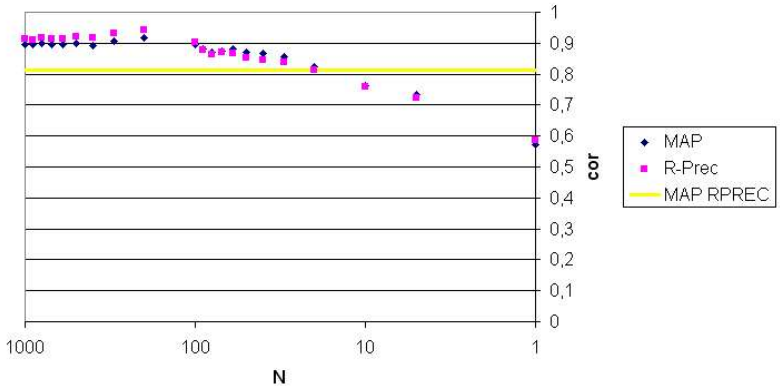


Figura 9.1: Correlazioni tra $ADM@N$ e le due metriche standard MAP e R-Prec

		ADM							
	N	5	10	20	30	100	200	500	1000
Precision	5	**0.89**	0.88	0.88	0.86	0.82	0.76	0.75	0.75
	10	0.84	**0.88**	0.91	0.91	0.88	0.83	0.80	0.80
	20	0.82	0.85	**0.92**	0.94	0.89	0.85	0.82	0.82
	30	0.81	0.83	0.91	**0.94**	0.87	0.81	0.78	0.77
	100	0.72	0.74	0.81	0.85	**0.94**	0.93	0.90	0.90
	200	0.71	0.75	0.79	0.82	0.91	**0.98**	0.94	0.93
	500	0.67	0.71	0.75	0.77	0.85	0.92	**0.99**	0.97
	1000	0.66	0.68	0.72	0.75	0.83	0.87	0.92	**0.92**

Tabella 9.2: Correlazioni tra $ADM@N$ e $P@N$

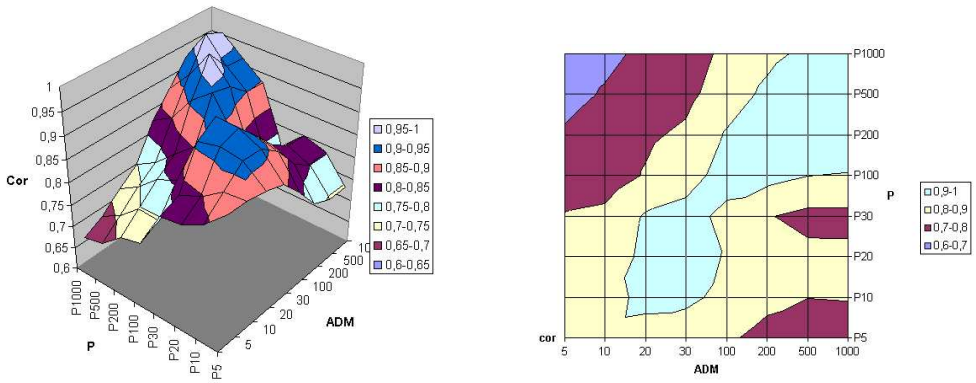


Figura 9.2: Correlazioni tra $ADM@N$ e $P@N$

La definizione di ADP è fatta per approssimare la metrica Precision. È stata calcolata la correlazione tra $ADP@1000$ e $P@1000$ per vedere se, anche sperimentalmente, l'approssimazione è buona. Questa correlazione è risultata essere pari a 0.9608.

Dopo aver testato le prestazioni di ADM calcolata sui primi N documenti reperiti dall'SRI, si è passati a valutare le prestazioni delle metriche ADP ed ADR derivabili da ADM (si veda il paragrafo 5.1).

Sono stati calcolati i valori di ADP e di ADR a certi valori di N documenti reperiti per cercare di costruire un grafico che approssimi il classico grafico Precision/Recall (si veda il paragrafo 3.1.2). È possibile vedere l'andamento di $ADP@N$ ed $ADR@N$ al variare di N in figura 9.3.

Dopo aver visto l'andamento medio di ADP ed ADR si è cercato di capire come differiscono tra loro gli andamenti dei singoli sistemi. A questo proposito sono stati calcolati i grafici ADP/ADR per i due SRI con prestazioni migliori secondo la metrica ADM e per i due SRI che, secondo ADM, sono i peggiori (si veda la figura 9.4).

I giudizi di pertinenza su tre livelli che sono stati fatti non hanno valori compresi tra 0 ed 1, è stato quindi necessario normalizzarli. Sono state provate diverse normalizzazioni al fine di vedere se, lasciando dei margini per la sopravvalutazione e per la sottovalutazione, i valori di ADP ed ADR subiscono cambiamenti. Sono state testate le seguenti normalizzazioni:

- 0, 0.5 ed 1
- 0.1, 0.5 ed 0.9

Nei test appena descritti la normalizzazione adottata prevedeva che i giudizi di pertinenza corrispondessero ai valori di URS pari a 0, 0.5 ed 1. Sono stati calcolati nuovamente ADP e ADR utilizzando una diversa normalizzazione dei giudizi di pertinenza.

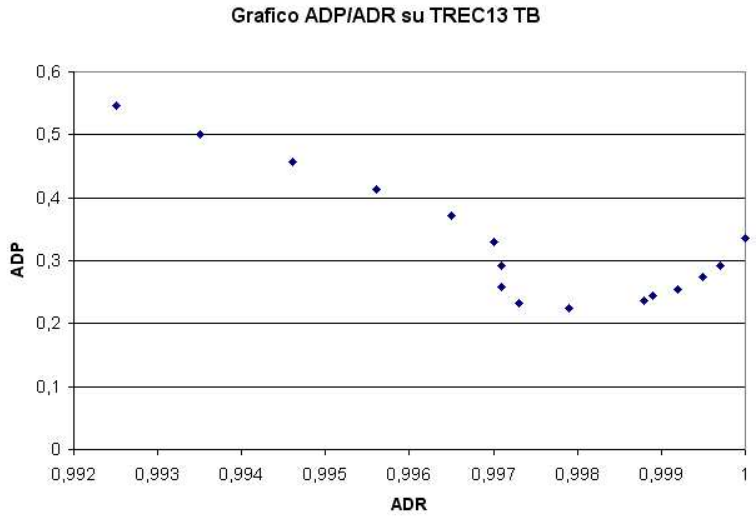


Figura 9.3: Andamento di ADP e ADR al variare di N su TREC13 TB

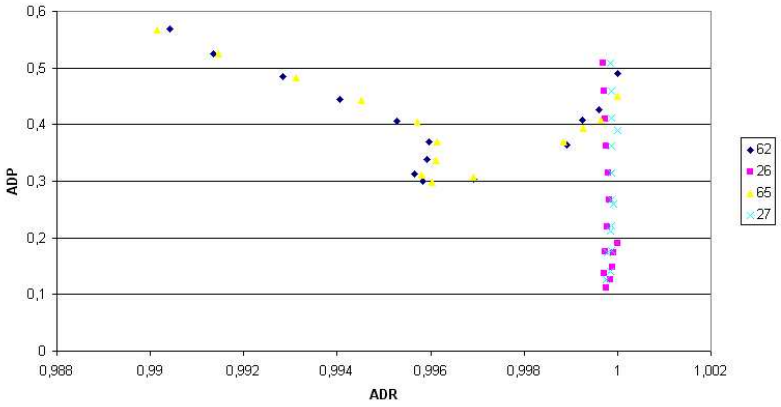


Figura 9.4: Andamento di ADP e ADR per i sistemi migliori e peggiori

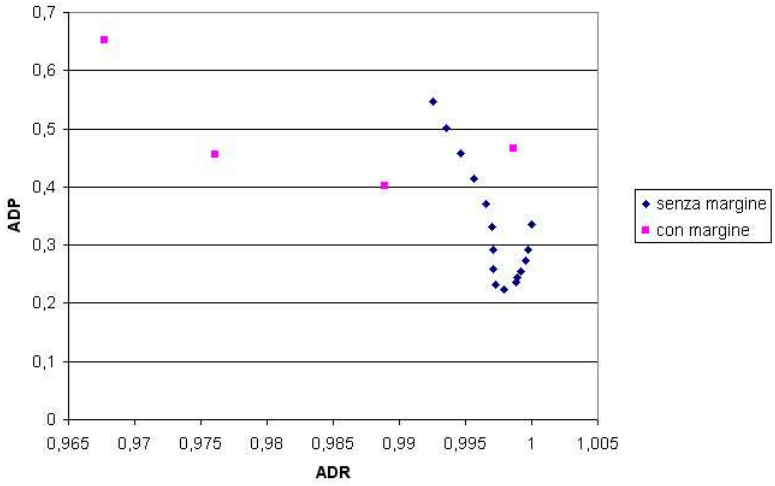


Figura 9.5: Andamento di ADP e ADR utilizzando o meno un margine per la sopravvalutazione/sottovalutazione

Si è deciso di normalizzare facendo corrispondere alle tre categorie di pertinenza dei valori di URS pari a 0.1, 0.5 oppure 0.9. L'obiettivo è di capire se, permettendo agli SRI di sopravvalutare e di sottovalutare le classificazioni dei documenti con dei valori di SRS che possono essere maggiori o minori del dei valori di URS, i valori di ADP e di ADR cambiano (si veda la figura 9.5).

Analogamente a quanto fatto in precedenza, sono stati calcolati i grafici contenenti le curve ADP/ADR per i due SRI con prestazioni migliori secondo la metrica ADM e per i due SRI che, secondo ADM, sono i peggiori, considerando la diversa normalizzazione dei gradi di pertinenza e paragonando i risultati con la prima normalizzazione utilizzata (si veda figura 9.6).

9.2 Discussioni dei risultati

Dai risultati descritti in tabella 9.1 ed in figura 9.1 è possibile apprendere diverse caratteristiche di ADM. La correlazione tra le metriche MAP ed R-Prec calcolate in TREC13TB è pari a 0.82. È possibile vedere dai risultati che le correlazioni con $ADM@N$ sono maggiori di questo valore per $N \geq 20$. Questo risultato può suggerire che ADM calcolato su solo 20 documenti fornisce le stesse informazioni che danno R-Prec e MAP utilizzando l'intero insieme di documenti reperiti. Si può vedere che al decrescere di N le correlazioni di ADM scendono.

In [17] gli autori suppongono che utilizzando ADM al posto di Precision e Recall, l'efficacia della metrica su interrogazioni in cui si considerano pochi documenti

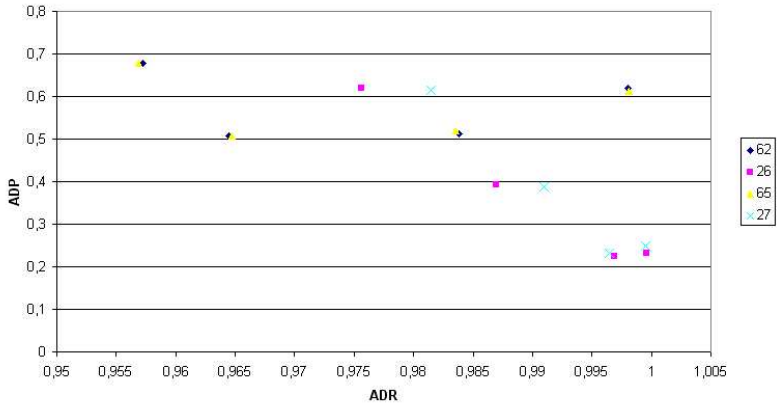


Figura 9.6: Andamento di ADP e ADR per i sistemi migliori e peggiori utilizzando un margine per la sopravvalutazione/sottovalutazione

reperiti è maggiore. I risultati di queste valutazioni supportano l'ipotesi che ADM funziona bene anche considerando solamente pochi documenti.

Dal calcolo delle correlazioni tra i valori di $P@N$ ed $ADM@N$ è stato possibile vedere che, come ci si aspetta, le correlazioni sono più alte a pari valore di N, mentre scendono al crescere della differenza di N. Nella tabella 9.2 possiamo quindi vedere le correlazioni più elevate sulla diagonale e più basse man mano che ci si allontana da essa (si veda figura 9.2). Possiamo quindi concludere che la metrica $ADM@N$ misura caratteristiche degli SRI analoghe a quelle misurate dalla metrica $P@N$.

Il calcolo delle metriche ADP ed ADR (si veda paragrafo 5.2.2) sui dati provenienti da TREC13TB hanno dato ulteriori risultati interessanti. Per prima cosa si può subito notare che i valori di $ADR@N$ sono, per ogni N, prossimi a 1 come nei risultati ottenuti sulla collezione TREC8. Questo risultato è spiegabile in modo analogo al precedente, dal fatto che nelle sperimentazioni è stato utilizzato l'ordinamento dei documenti fatto dagli SRI per derivare i valori di SRS (si veda paragrafo 5.3.1). I valori di SRS sono stati derivati, anche in questo caso, in modo lineare dall'ordinamento effettuato dagli SRI. Questo modo di derivare i valori di SRS ha fatto sì che i documenti siano, in maggior parte, sopravvalutati, in quanto, da un punto di vista dei valori di URS, ci saranno pochissimi documenti con valore 1 e moltissimi con valore 0. Dato che ADR viene calcolata solamente sui documenti sottovalutati (che sono pochi), le differenze tra SRS ed URS saranno minime in quanto questi documenti si troveranno principalmente a livelli di pertinenza prossimi a 1. I valori di ADP sono prossimi a 0.5 in quanto, su molti documenti, il valor medio delle differenze è vicino ad 1/2 (si veda figura 9.7).

Dalle curve ADP/ADR è possibile notare che, come nel caso delle curve Precision/Recall, i valori di ADP scendono al crescere dei valori di ADR. Questo, però, è vero solamente per valori di N (numero di documenti considerati nel calcolo di

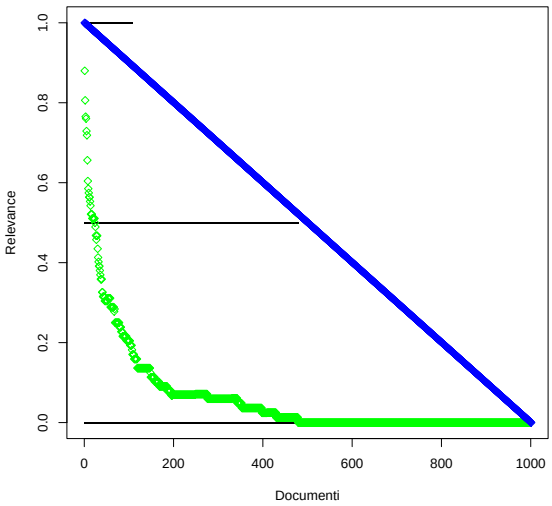


Figura 9.7: Andamento di URS e di SRS assegnato

ADP ed ADR) alti. Infatti è possibile notare che per valori di $N \in [1, 40]$ i valori di ADP crescono al crescere dei valori di ADR e sono più alti tanto più scende N. Questo comportamento delle curve ADP/ADR su valori piccoli di N non è ancora stato compreso.

Dopo aver calcolato i grafici ADP/ADR come media di tutti i sistemi, sono stati calcolati gli stessi grafici considerando solamente i sistemi con prestazioni migliori e peggiori secondo la metrica ADM. In questo caso i risultati mostrano che l'andamento dei grafici resta lo stesso, ma cambia l'intervallo dei valori. Nel caso dei sistemi peggiori i valori di ADP variano su un intervallo maggiore ($[0.2, 0.5]$) mentre i valori di ADR sono compresi in un intervallo minore ($[0.999, 1]$). Considerando i sistemi migliori, invece, gli l'intervallo di valori per ADP si restringe a $[0.45, 0.55]$ ed i valori di ADR si allargano a $[0.99, 1]$.

Infine, i risultati della mappatura dei tre livelli di pertinenza a (0.1 0.5 0.9) (si veda la figura 9.5) mostrano che ADR assume valori in un intervallo più ampio ($[0.96, 0.99]$), ma comunque sempre prossimi a 1. Anche in questo caso i valori alti di ADR sono dovuti all'utilizzo dei comunque pochi documenti sottovalutati.

9.3 Il tasso d'errore

Sulla collezione di test di TREC13 TeraByte è stata svolta un'ulteriore valutazione sperimentale. Sono stati calcolati i tassi d'errore (si veda il paragrafo 4.3), in base alle definizioni in [9], per le metriche $ADM@N$ e MAP con l'obiettivo di vedere se all'aumentare dei documenti considerati il tasso d'errore di $ADM@N$ scende.

Inoltre si vuole vedere se esiste un valore di N per cui $ADM@N$ e MAP hanno lo stesso tasso d'errore.

Si ricorda che il tasso d'errore viene calcolato in base alla media dei valori delle metriche, per ogni SRI che si vuole comparare. In questo caso sono stati considerati i 57 SRI che hanno reperito almeno 1000 documenti per ogni interrogazione. A questo punto, per ogni coppia possibile di SRI, si calcola il numero di volte che un SRI è risultato migliore su una interrogazione. In questo caso sono state considerate le 49 interrogazioni presenti nella collezione di test, anche se uno studio [53] mostra come siano appena accettabili collezioni di test contenenti 250 interrogazioni. Si calcola poi il tasso d'errore tenendo conto di quante volte l'SRI peggiore dei due (cioè quello che la metrica ha giudicato migliore meno volte) è stato valutato migliore su un'interrogazione. Infine è stata calcolata la media dei tassi d'errore su tutte le interrogazioni e le coppie di SRI considerate per avere un unico valore per la metrica sulla singola collezione di test.

Nel calcolo si considera un margine di errore, ad esempio del 5%, che viene applicato al valore delle metrica. Se la differenza tra i due valori è inferiore a questo margine, allora si considera che i due SRI sono stati valutati alla pari. Le sperimentazioni sono state effettuate sia senza considerare un margine d'errore sia considerando un margine del 5% come in [9]. I risultati sono presentati in tabella 9.3.

Metrica	Tasso d'errore senza margini
ADM@1000	0.2212
ADM@500	0.2121
ADM@300	0.2224
ADM@200	0.2298
ADM@100	0.2588
ADM@50	0.2740
ADM@20	0.2817
MAP	0.2264

Metrica	Tasso d'errore con 5% di margine
ADM@1000	0.0186
ADM@500	0.0812
ADM@300	0.1361
ADM@200	0.1697
ADM@100	0.2174
ADM@50	0.2639
ADM@20	0.2605
MAP	0.2060

Tabella 9.3: Tassi d'errore delle metriche

Dai risultati è possibile notare che il tasso d'errore della metrica MAP è quasi

equivalente al tasso di ADM calcolata su 200 documenti se non si considera un margine nel calcolo del tasso d'errore e al tasso di ADM calcolata su 100 documenti nel caso in cui si considera un margine del 5%. Da questi risultati si può concludere che ADM calcolata su 200 documenti effettua gli stessi errori di valutazione delle metrica MAP che utilizza tutti i documenti reperiti per il suo calcolo.

I tassi di errore di $ADM@N$ con $N > 300$ sono minori di quelli di MAP. Questo indica che ADM è una metrica più stabile di MAP. Si può ulteriormente notare che, come ci si aspetta, il tasso di errore cresce al diminuire dei documenti considerati per il calcolo della metrica come constatato anche in [9]. In particolare il tasso scende fino a 500 documenti considerati per poi risalire nel caso in cui si considerano 1000 documenti.

I valori numerici dei tassi d'errore non sono comparabili con quelli ottenuti dalle sperimentazioni descritte in [9] (si veda la tabella 4.2) in quanto, in quel caso, sono stati utilizzati complessivamente 1050 interrogazioni, raggruppate in 21 gruppi da 50. È quindi comprensibile che i valori dei tassi d'errore calcolati qui siano notevolmente più elevanti in quanto sono stati considerati solamente 49 interrogazioni (si veda paragrafo 4.3).

Quindi è possibile concludere che ADM è più sensibile di MAP in quanto necessita di meno documenti per effettuare le valutazioni degli SRI.

9.4 Conclusioni

In questo capitolo sono stati presentati i risultati ottenuti valutando ADM utilizzando la collezione di test TREC13 TeraByte. Sono stati discussi i risultati ottenuti e presentati i risultati ottenuti calcolando il tasso d'errore della metrica ADM e della metrica MAP.

I risultati mostrano che ADM calcolato su solo 20 documenti fornisce informazioni analoghe alle metriche R-Prec e MAP che usano l'intero insieme di documenti reperiti. Si può vedere, inoltre, che al crescere di N la correlazione di $ADM@N$ con le metriche standard aumenta.

Dalle curve di ADP/ADR è possibile notare che i valori di ADP scendono al crescere dei valori di ADR tranne che per valori di ADP ed ADR calcolate su N documenti con $N \in [1, 40]$ in cui i valori di ADP crescono al crescere dei valori di ADR e sono più alti tanto più scende N.

Il calcolo del tasso d'errore ha mostrato che ADM calcolato su 200 documenti ha la stessa sensibilità della metrica MAP. Quindi l'analisi del tasso d'errore ha dato indicazione che ADM è una metrica più stabile di MAP e che $ADM@N$ è una metrica utilizzabile per le valutazioni degli SRI anche per valori di N bassi.

Nel prossimo capitolo si vedranno e verranno discussi i risultati ottenuti utilizzando la collezione INEX 2004.

Capitolo 10

Le valutazioni sperimentali su INEX 2004

In questo capitolo vengono mostrati i risultati delle valutazioni sperimentali basate sulla collezione INEX 2004. In questo caso è stato deciso di utilizzare una versione modificata della metrica ADM che si adatta meglio alla collezione di test. Nel capitolo quindi si vedranno per prime le motivazioni che hanno portato alla scelta di utilizzare un'estensione di ADM per le valutazioni sperimentali (paragrafo 10.1). Si vedrà poi (paragrafo 10.2) la definizione delle nuova metrica ed infine (paragrafo 10.3) le valutazioni sperimentali utilizzando questa nuova metrica.

10.1 L'iniziativa INEX 2004

L'iniziativa di valutazione INEX (si veda il paragrafo 2.4.3) considera il concetto di pertinenza definito su due diverse dimensioni (si veda il paragrafo 2.2.2). Questo impone l'utilizzo di funzioni che trasformino le due dimensioni in un'unica dimensione di pertinenza se si vuole calcolare i valore delle metriche standard come MAP ed R-Prec, ma anche se si vuole calcolare ADM, in quanto ADM considera una sola dimensione, anche se continua, di URS. Per portare le due dimensioni di pertinenza su una sola sono state proposte diverse funzioni [31]. Ad esempio:

$$f_{strict}(e, s) := \begin{cases} 1 & \text{se e=3 e s=3,} \\ 0 & \text{altrimenti.} \end{cases}$$

dove e ed s sono i valori di esausitività e di specificità rispettivamente (si veda paragrafo 2.2.2). Queste funzioni portano a un livello binario il concetto di pertinenza in modo da poterlo utilizzare per calcolare i valori delle metriche che assumono una pertinenza di tipo binario. Una funzione di tipo rilassato, al contrario di quella appena vista (f_{strict}), è quella che porta al livello 1 ogni componente che abbia almeno una dimensione di pertinenza diversa da 0.

Un ulteriore problema che si verifica nell'utilizzo della collezione di test INEX 2004 è il fatto che in questa iniziativa gli SRI non devono reperire documenti interi, bensì delle parti di documenti contenenti le informazioni pertinenti [22]. Questa diversità dalle altre iniziative di valutazione porta al fatto che le metriche devono considerare, oltre alla quantità di informazione pertinente contenuta nei documenti reperiti, anche la dimensione dei componenti, altrimenti basterebbe reperire componenti più grandi per reperire un maggior numero di informazioni pertinenti all'interrogazione.

Un'altra complicazione che si presenta dovendo valutare il reperimento di componenti di documenti è il problema della sovrapposizione [22, 25, 33]. Questo problema è dato dalla possibile sovrapposizione dei componenti reperiti dall'SRI. I componenti infatti, possono essere reperiti più di una volta come parte di un componente più grande che lo contiene. L'ipotetico utente dell'SRI non vorrà analizzare due volte lo stesso componente in cerca di informazione pertinente e quindi la sovrapposizione dei componenti reperiti è un fattore da tenere in considerazione nella valutazione degli SRI.

10.2 Un'estensione di ADM su due dimensioni di URS

Abbiamo visto che nell'iniziativa di valutazione INEX il concetto di pertinenza è definito su due dimensioni diverse: esaustività e specificità. Per venire in contro a questa nozione di pertinenza multidimensionale, propongo una possibile modifica alla metrica ADM (si veda il paragrafo 5.1) che considera solamente una dimensione continua di pertinenza.

La misura ADM considera due dimensioni continue: quella di URS e quella di SRS. Nell'iniziativa INEX la pertinenza è definita su due diverse dimensioni in quanto, per documenti XML e per filmati, il concetto di pertinenza è diverso. Infatti in questi casi ad essere reperito non è più un intero documento, bensì dei frammenti di documento o di un filmato. Per questo motivo è necessario valutare in modo distinto, per questi frammenti, qual è la loro esaustività e la loro specificità. I giudici umani quindi devono dare due distinti giudizi per ogni frammento che considerano pertinente. Gli SRI partecipanti ad INEX continuano comunque a reperire i frammenti assegnando un unico valore di SRS.

La misura qui proposta, che chiamo ADM^3, utilizza due dimensioni di URS. In questo modo non è necessario utilizzare delle funzioni per il mapping di specificità ed esaustività su un'unica dimensione di pertinenza come viene fatto in INEX e come è stato fatto per valutare ADM (si veda il paragrafo 10.3).

Analogamente ad ADM, verrà misurata la distanza tra il valore di SRS assegnato dall'SRI ed i valori di esaustività e di specificità. Geometricamente è possibile rappresentare il tutto in uno spazio 3D in cui sul piano xy sono rappresentati tutti i valori di URS, mentre sull'asse z vengono rappresentati i valori di SRS.

In ADM i documenti erano rappresentati da punti nel piano, in questo caso avremo punti nello spazio. Se in ADM la classificazione corretta dei documenti

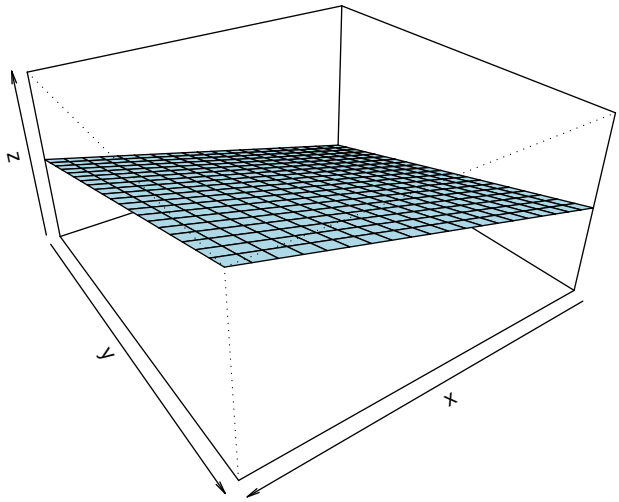


Figura 10.1: Una visione 3D del piano ideale di classificazione dei documenti

deve essere sulla retta $y = x$, in questa estensione di ADM i documenti dovranno essere posizionati sul piano $z = (x + y)/2$ (si veda la figura 10.1). Verrà quindi misurata la distanza tra i punti e questo piano. Come in ADM, non viene misurata la distanza tra punto e piano, cioè quella ortogonale, perché si suppone che i giudizi di esaustività e specificità siano corretti, e quindi si andrà a misurare la distanza del punto dal piano lungo l'asse z.

La formula di ADM^3 risulta quindi essere la seguente:

$$ADM^3_q = 1 - \frac{\sum_{d_i \in D} |SRS(d_i) - (\alpha E(d_i) + \beta S(d_i))|}{|D|}$$

dove $|D|$ è il numero di documenti presenti nella collezione e $E(d_i)$, $S(d_i)$ sono i valori di esaustività e specificità del documento d_i rispettivamente e q è un'interrogazione. I valori α e β sono dei coefficienti che permettono di dare un peso maggiore ad esaustività e specificità e sono tali che $\alpha + \beta = 1$. Nel caso che questi due coefficienti siano entrambi pari a 0.5 si considerano alla pari le due dimensioni di pertinenza. ADM^3 avrà valori compresi tra 0 ed 1 e, facendo la media di ADM^3 su più interrogazioni si ottiene una metrica di valutazione dell'efficacia di un SRI.

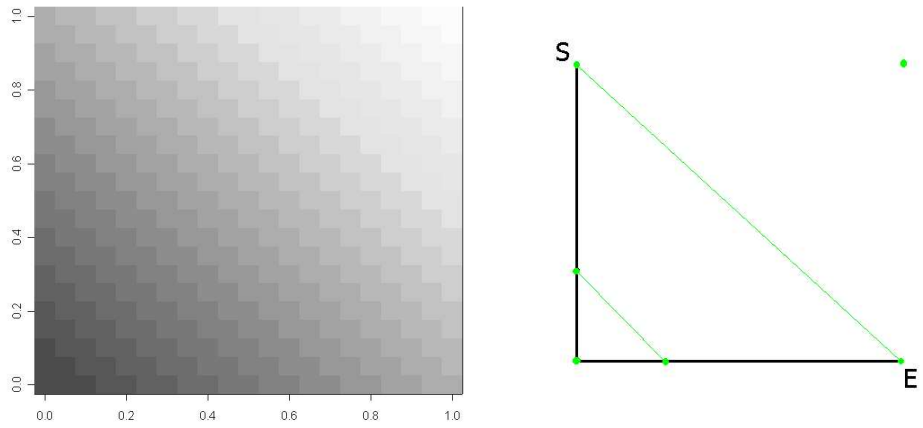


Figura 10.2: Una visione del piano URS

Vediamo con maggior dettaglio il motivo per cui si sceglie di far corrispondere all'esatta classificazione dei documenti il posizionamento sul piano $z = (x + y)/2$. Utilizzando due dimensioni di pertinenza è presumibile che esse siano indipendenti l'una dall'altra e che siano ugualmente importanti. Per questo motivo è corretto pensare che uno stesso valore di SRS sia ugualmente adatto per tutte le combinazioni di esaustività e specificità che si trovano sulla linea immaginaria che collega punti sugli assi x ed y equidistanti dall'origine (si veda la figura 10.2a). Utilizzando il piano $z = (x+y)/2$ e la distanza lungo l'asse z, si giudicano allo stesso modo i valori di SRS che si trovano sulla stessa linea del piano xy (si veda la figura 10.2b).

10.3 Una prima valutazione sperimentale

È stata fatta una prima valutazione sperimentale di ADM^3 utilizzando i dati di INEX 2004 e confrontando i valori di ADM^3 con quelli di ADM calcolata sempre sulla stessa collezione di test.

I dati utilizzati sono stati ottenuti in seguito alla partecipazione ad INEX 2005 come giudice di pertinenza (2 settimane di lavoro). Questi dati sono stati resi disponibili dal 15 ottobre 2005 e, per il limitato tempo a disposizione, i risultati qui presenti sono preliminari.

Sulla collezione di test INEX 2004 sono stati calcolati i valori di ADM considerando i singoli elementi reperiti. Non si è quindi tenuto conto delle sovrapposizioni di componenti e della loro grandezza in quando ADM non prevede questi fattori nella valutazione dell'efficacia degli SRI. Allo stesso modo sono stati calcolati i valori di ADM^3 con valori $\alpha = 0.5$ e $\beta = 0.5$.

Le metriche ADM ed ADM^3 sono state confrontate tra loro utilizzando le correlazioni di Kendall e di Spearman ottenendo rispettivamente valori di correlazione

	MAP	ADM^3
Kendall ADM	0.3103	0.7943
Spearman ADM	0.5690	0.9379
Kendall ADM^3	0.4535	1
Spearman ADM^3	0.6815	1

Tabella 10.1: Correlazioni tra ADM, ADM^3 e MAP

pari a 0.7943 e 0.9379. Questi valori indicano che l'estensione di ADM considerando due dimensioni di URS ha un andamento prossimo a quello della metrica ADM classica.

Sulla collezione di test INEX 2004 sono, inoltre, stati calcolati i valori di MAP utilizzando lo strumento `inex_eval`, considerando quindi tutti i fattori di sovrapposizione e dimensione dei componenti reperiti, sia i valori di MAP considerando i singoli elementi reperiti senza tenere conto delle dimensioni e delle sovrapposizioni ed utilizzando una funzioni di mapping di tipo rilassato, cioè portando ad un livello pertinente i componenti con almeno una dimensione di pertinenza (esaustività o specificità) diversa da 0. I valori di questi due modi diversi di calcolare la metrica MAP hanno dato risultati pessimi di correlazioni (pari a -0.4923) indicando la notevole differenza che si presenta nel considerare o meno le informazioni riguardo la dimensione e la sovrapposizione dei componenti.

Le correlazioni tra i valori di ADM e di ADM^3 con i valori di MAP calcolati senza considerare la dimensione e la sovrapposizione di componenti non risulta essere comunque buona (si veda la tabella 10.1)

10.4 Conclusioni

In questo capitolo sono state descritte le motivazioni che hanno portato alla scelta di utilizzare un'estensione di ADM per le valutazioni sperimentali. Si è vista la definizione delle nuova metrica ed infine sono state presentate le valutazioni sperimentali effettuate utilizzando questa nuova metrica.

I risultati hanno mostrato che questa metrica ha un andamento prossimo a quello della metrica ADM classica.

Il prossimo capitolo conclude la tesi riassumendo il lavoro svolto e delineando possibili sviluppi futuri.

Capitolo 11

Conclusioni e sviluppi futuri

In questo capitolo viene dapprima riassunto il lavoro svolto (paragrafo 11.1) e vengono poi delineati possibili sviluppi futuri (paragrafo 11.2).

11.1 Il lavoro svolto

In questa tesi si è trattato il problema della valutazione dell'efficacia degli SRI. Nella letteratura scientifica sono state proposte numerose metriche di valutazione. Questa tesi ha l'obiettivo di valutare concettualmente e sperimentalmente una nuova metrica di valutazione. È stata proposta una classificazione di tutte le metriche di valutazione per comprendere le peculiarità di una particolare metrica per la valutazione dell'efficacia degli SRI: è stata effettuata un'analisi critica e delle valutazioni sperimentali di questa metrica.

La tesi è organizzata in due parti. Nella prima parte sono stati descritti i concetti di base necessari per poter effettuare delle valutazioni di efficacia di una metrica.

Nel secondo capitolo sono stati definiti i concetti fondamentali quali gli SRI, la pertinenza di tipo binario, a categorie e continua, il processo di valutazione ed è stato descritto come questo viene messo in pratica in diverse iniziative internazionali di valutazione quali CLEF, INEX, NTCIR e TREC.

Nel terzo capitolo sono state definite le metriche di valutazione proposte negli anni nella letteratura scientifica, suddivise tra metriche classiche, orientate all'utente, alternative a quelle classiche e metriche per documenti XML.

Nel quarto capitolo si sono visti alcuni metodi per valutare l'efficacia delle metriche. È stato descritto come è possibile valutare il tasso d'errore delle metriche per capirne la stabilità al variare del numero di documenti e di interrogazioni considerate per valutare l'efficacia degli SRI. Sempre nel terzo capitolo si sono visti i modelli di distribuzione dei valori di pertinenza nei documenti che un SRI reperisce. Questi modelli andrebbero tenuti in considerazione dagli SRI nell'assegnare dei punteggi di pertinenza ai documenti reperiti.

Nel quinto capitolo si è introdotta la nuova metrica di valutazione ADM che

è stata analizzata e valutata sperimentalmente in questa tesi. Sono stati inoltre illustrati i risultati delle valutazioni eseguite in passato su ADM.

Nella seconda parte della tesi è stata proposta un'analisi critica della metrica ADM e sono stati presentati e discussi i risultati ottenuti dalle valutazioni sperimentali.

Nel capitolo 6 è stata presentata una metodologia per effettuare valutazioni sperimentali di una metrica. La metodologia proposta è stata utilizzata per valutare l'efficacia di ADM ed è risultata essere in grado di adattarsi a collezioni di test diverse tra loro come TREC8, TREC13 TeraByte e INEX 2004.

Per analizzare quali sono le peculiarità di ADM rispetto le altre metriche, nel capitolo 7 è stata presentata una nuova classificazione delle metriche basata sul concetto di reperimento e pertinenza binaria, a categorie e continua. Per ogni metrica è stato individuato il contesto di pertinenza e reperimento più adatto affinché la metrica venga utilizzata per valutare l'efficacia degli SRI e sono state indicate le iniziative di valutazione che la utilizzano. In questa classificazione ADM è risultata essere una metrica che, a contrario delle altre, è in grado di adattarsi ad ogni situazione di pertinenza e di reperimento.

Riguardo all'obiettivo di effettuare ulteriori valutazioni sperimentali, nel capitoli 8 e 9 sono stati illustrati i risultati sperimentali delle valutazioni sulla collezione TREC8 unitamente a delle riclassificazioni di pertinenza e sulla collezione TREC13 TeraByte. I risultati presentati sono stati discussi cercando di trarre delle conclusioni su ADM e paragonandole con le ipotesi effettuate. La metrica ha dato buoni valori di correlazione con le metriche utilizzate nelle iniziative internazionali di valutazione. Inoltre è risultata essere più stabile di altre metriche in quanto richiede meno documenti di altre per ottenere valori significativi.

Nel capitolo 10 si è valutata l'efficacia di ADM usando la collezione INEX 2004. Per prima cosa è stata presentata un'estensione (ADM^3) della metrica ADM che considera due dimensioni di pertinenza (come nell'iniziativa INEX) la quale è stata poi valutata sui dati di INEX 2004 in quanto la metrica ADM non si adatta bene a questa collezione di test. La metrica ADM^3 è risultata essere una metrica con caratteristiche paragonabili a quelle di ADM, ma che meglio si adatta ad una situazione di pertinenza a due dimensioni.

Complessivamente possiamo dire che ADM è risultata essere una metrica in grado di adattarsi a diversi modelli di pertinenza e di reperimento: sperimentalmente è stato mostrato che si adatta a situazioni in cui si utilizza una pertinenza binaria e reperimento a categorie (come in TREC8) e situazioni con pertinenza e reperimento a categorie (in TREC13 TeraByte). I risultati sulle correlazioni con le altre metriche ottenuti in questa tesi hanno confermato i risultati ottenuti in precedenza: ADM correla bene con le altre metriche, a volte presentando anche valori superiori rispetto alle correlazioni fra le metriche standard. Anche usando un minor numero di documenti ($ADM@N$) le correlazioni sono risultate essere buone.

Questa metrica inoltre ha dimostrato una maggiore stabilità rispetto alle altre metriche. Questo è stato dimostrato sia dal calcolo del tasso d'errore sia dal fatto

che ADM correla bene con le altre metriche anche considerando un minor numero di documenti reperiti per valutare l'efficienza degli SRI. Quindi possiamo dire che ADM è una metrica che è possibile utilizzare in situazioni in cui si hanno a disposizione un numero ridotto di informazioni.

ADM sembra quindi essere una metrica flessibile in quanto si può adattare, con minime modifiche, a diverse situazioni (vari tipi di pertinenza e di reperimento, considerazione di pochi documenti, modelli di pertinenza a due dimensioni come in INEX, ecc.)

11.2 Sviluppi futuri

Vediamo ora alcuni dei possibili sviluppi futuri di questa tesi:

- effettuare maggiori valutazioni sperimentali su ADM per confermare i risultati ottenuti con le valutazioni effettuate in questa tesi;
- rifare le valutazioni fatte fino ad ora utilizzando la metrica QADM e confrontare l'efficacia di questa metrica con quella di ADM;
- analizzare le metriche ADP ed ADR per comprendere l'andamento delle curve ADP/ADR calcolata su pochi documenti reperiti dove i valori di ADP e di ADR crescono al diminuire del numero di documenti considerati (si veda la figura 8.4);
- calcolare i valori di ADM, ADP ed ADR utilizzando una distribuzione di SRS non lineare (come si può vedere in figura 9.7) che possa assomigliare di più, rispetto a quella lineare, alla distribuzione dei documenti pertinenti all'interno di quelli reperiti dagli SRI;
- effettuare ulteriori test di valutazione per confrontare ADM^3 (si veda il paragrafo 10.2) su 2 dimensioni di pertinenza (esaustività e specificità) con le altre metriche utilizzate per valutare le prestazioni degli SRI utilizzando 2 dimensioni di pertinenza utilizzando la collezione di test INEX 2005;
- utilizzare una metodologia per il calcolo del tasso d'errore assodata in campo statistico come ad esempio il test di Friedman;
- effettuare maggiori valutazioni sperimentali sulla metrica ADM^3 (si veda il paragrafo 10.2) utilizzando diverse collezioni di test (INEX 2005), diversi valori per i parametri α e β e le k-statistiche;
- calcolare il tasso d'errore di ADM^3 e confrontarlo con quello di ADM e delle metriche standard utilizzate nella collezione di test usata (si veda il paragrafo 4.3);

- capire come mai la metrica ADM calcolata con i giudizi di pertinenza a 4 livelli correli, con le metriche standard, peggio di ADM calcolata su due livelli di pertinenza mappati in modo rilassato (si veda il capitolo 8);
- progettare ed implementare un SRI che stima il valore di pertinenza per ogni documento e sfrutta le distribuzioni degli score nei documenti (si veda il paragrafo 4.4) utilizzando un modello di pertinenza continua;
- applicare idee simili all'ADM nel campo dell'Informatica Medica (si veda il paragrafo 4.1).

Ringraziamenti

Vorrei racchiudere in questa pagina i ringraziamenti alle persone che, durante i passati 5 anni, hanno permesso e mi hanno aiutato nel raggiungimento dell'obiettivo della Laurea Specialistica.

Per primo vorrei ringraziare il mio relatore, dott. Stefano Mizzaro, per la pazienza, per il tempo dedicato, per i consigli dati anche al di fuori della tesi e per la notevole quantità di caffeina che mi ha gentilmente offerto al bar del campus universitario.

Per avermi assistito nel lavoro di tesi e per la loro simpatia, vorrei inoltre ringraziare il dott. Vincenzo Della Mea e il dott. Luca di Gaspero.

Ringrazio i miei colleghi, che prima di tutto sono stati amici, (elencati in ordine alfabetico per non fare preferenze) Adolfo, Andrea, Daniele, Dante, Elisabetta, Emanuele, Fabio, Francesca, Ingrid, Luca, i 3 Marco, Mauro, Monica, per il supporto morale e professionale che hanno saputo darmi in questi 5 anni.

Ringrazio i miei genitori per non avermi tagliato i viveri.

Ringrazio Chiara, per essermi stata vicina nei momenti più difficili di questi 5 anni.

E per ultima, ma non per importanza, vorrei ringraziare colei che mi ha introdotto e che mi ha fatto innamorare del campo dell'Informatica: la prof.ssa Rossana Dell'Andrea, mia insegnante all'ITC "E. Fermi" di Gorizia per il periodo 1997-2000.

Bibliografia

[1] E. Aslam, J. A. Yilmaz e V. Pavlu. A geometric interpretation of Rprecision and its correlation with average precision. In *28th SIGIR*, pp. 573–574, 2005.

[2] R. Baeza-Yates e R. Neto. *Modern Information Retrieval*. ACM Press, 1999.

[3] R.K. Belew. *Finding Out About*. Cambridge Univ. Press, 2000.

[4] D. C. Blair e M. E. Maron. An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

[5] A. Bookstein. When the most "pertinent" document should not be retrieved - An analysis of the Swets model. *Information Processing and Management*, 13:377–383, 1977.

[6] A. Bookstein. Relevance. *Journal of the American Society for Information Science*, 30(5):269–273, 1979.

[7] P. Borlund e P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *21st SIGIR*, pp. 324–331, 1998.

[8] R. Brache. Personal communication, 2005.

[9] C. Buckley e E. Voorhees. Evaluating evaluation measure stability. In *23rd SIGIR*, pp. 33–40, 2000.

[10] C. Buckley e E. Voorhees. The effect of topic set size on retrieval experiment error. In *25th SIGIR*, pp. 316–323, 2002.

[11] C. Buckley e E. Voorhees. Retrieval evaluation with incomplete information. In *27th SIGIR*, pp. 25–32, 2004.

[12] Cross-Language Evaluation Forum (CLEF). http://www.clef-campaign.org/ (ultima visita novembre 2005).

[13] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science and Technology*, 19:30–41, 1968.

[14] A.P. de Vries, G. Kazai, e M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference Proceedings*, pp. 463–473, 2004.

[15] V. Della Mea, G. Demartini, L. Di Gaspero, e S. Mizzaro. Experiments on Average Distance Measure. In fase di revisione all'ECIR06.

[16] V. Della Mea, L. Di Gaspero, e Mizzaro S. Evaluating ADM on a four-level relevance scale document set from NTCIR. In *Proceedings of NTCIR Workshop 4 Meeting - Supplement Vol. 2*, pp. 30–38, 2004.

[17] V. Della Mea e S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):530–543, 2004.

[18] G. Demartini e S. Mizzaro. A classification of IR effectiveness metrics. In fase di revisione all'ECIR06.

[19] 28th European Conference on Information Retrieval (ECIR06). http://ecir2006.soi.city.ac.uk/ (ultima visita novembre 2005).

[20] H. Frei e P. Schäuble. Determining the effectiveness of retrieval algorithms. *Information Processing and Management*, 27(2):153–164, 1991.

[21] T. J. Froehlich. Relevance reconsidered-Towards an agenda for 21st century: Introduction. *Journal of the American Society for Information Science*, 45(3), 1994.

[22] S. Fuhr, M. Lalmas, e S. Malik. Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003. In *INEX 2003 Workshop Proceedings*, 2003.

[23] *Il nuovo dizionario inglese Garzanti*. Garzanti, 1991.

[24] Google. http://www.google.com (ultima visita novembre 2005).

[25] N. Gövert, G. Kazai, N. Fuhr, e M. Lalmas. Evaluating the effectiveness of content-oriented XML retrieval. Relazione tecnica, 2003.

[26] INitiative for the Evaluation of XML Retrieval (INEX). http://inex.is.informatik.uni-duisburg.de/ (ultima visita novembre 2005).

[27] M. A. J. Van Bemmelm. *Handbook of Medical Informatics*. Springer-Verlag, 2a edizione, 2002.

[28] K. Järvelin e J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *23rd SIGIR*, pp. 41–48, 2000.

[29] K. Järvelin e J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20:422–446, 2002.

[30] N. Kando, K. Kuriyama, e M. Yoshioka. Information retrieval system evaluation using multi-grade relevance judgments. In *IPSJ SIGNotes*, 2001.

[31] G. Kazai. Report of the INEX 2003 metrics working group. In *Proceedings of the 2nd Workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, pp. 184–190, 2004.

[32] G. Kazai e M. Lalmas. INEX 2005 evaluation metrics, 2005. http://inex.is.informatik.uni-duisburg.de/2005/inex-2005-metricsv4.pdf.

[33] G. Kazai e M. Lalmas. Notes on what to measure in INEX. In *INEX Workshop on Element Retrieval Methodology*, 2005.

[34] J. Kekäläinen. Binary and graded relevance in IR evaluations - Comparison of the effects on ranking of IR systems. *Information Processing and Management*, 41:1019–1033, 2005.

[35] R. R. Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, 1997.

[36] F. W. Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. J. Wiley, 1968.

[37] R. M. Losee. Upper bounds for retrieval performance and their use measuring performance and generating optimal boolean queries: Can it get any better than this? *Information Processing and Management*, 30(2):193–204, 1994.

[38] R. Manmatha, T. Rath, e F. Feng. Modeling score distribution for combining the output of search engines. In *24th SIGIR*, pp. 267–275, 2001.

[39] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[40] S. Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322, 1998.

[41] S. Mizzaro. A new measure of retrieval effectiveness (or: What's wrong with precision and recall) In *International Workshop on Information Retrieval (IR'2001)*. A cura di Ojala, pp. 43–52, 2001.

[42] NTCIR workshop 4. http://research.nii.ac.jp/ntcir-ws4/ (ultima visita novembre 2005).

[43] B. Piwowarski e P. Gallinari. Expected ratio of relevant units: A measure for structured information retrieval. In *INEX'03 proceedings*, pp. 158–166, 2004.

[44] S. E. Robertson e K. Spark Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[45] I. Ruthven. Users and Information Retrieval - Lezione alla 5a European summer school in Information Retrieval, 2005. http://www.cdvp.dcu.ie/ESSIR2005/index.html (ultima visita novembre 2005).

[46] T. Sakai. New performance metrics based on multigrade relevance: Their application to question answering. In *NTCIR 4 Meeting Working Notes*, 2004.

[47] G. Salton e M. Lesk. Computer evaluation of indexing and text processing. *Journal of the Association for Computing Machinery*, 15:8–36, 1968.

[48] G. Salton e M. Lesk. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(3):343–359, 1968.

[49] G. Salton e M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1984.

[50] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[51] E. H. Shortliffe, L. M. Fagan, G. Wiederhold, e L. E. Perreault. *Medical Informatics: Computer Applications in Health Care and Biomedicine*. Springer-Verlag, 2a edizione, 2000.

[52] E. Sormunen. Liberal relevance criteria of TREC - Counting on negligible documents? In *25th SIGIR*, pp. 324–330, 2002.

[53] K. Spark Jones e C. J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[54] D. Sperber e D. Wilson. *Relevance — Communication and Cognition*. Harvard University Press, 1986.

[55] J. A. Swets. Information retrieval systems. *Science*, 141:245–250, 1963.

[56] Text REtrieval Conference (TREC). http://trec.nist.gov/ (ultima visita novembre 2005).

[57] English Relevance Judgments (TREC). http://trec.nist.gov/data/reljudge_eng.html (ultima visita novembre 2005).

[58] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2a edizione, 1979.

[59] E. Voorhees. Measuring ineffectiveness. In *27th SIGIR*, pp. 562–563, 2004.

[60] E. Voorhees e D. Harman. Overview of the eighth Text REtrieval Conference (TREC-8). In *Proceedings of the eighth Text REtrieval Conference (TREC-8)*, pp. 1–24, 2000.

[61] E. M. Voorhees. Evaluation by highly relevant documents. In *24th SIGIR*, pp. 74–82, 2001.

[62] Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.

[63] J. Zobel e Sanderson M. Information retrieval system evaluation: effort, sensitivity and reliability. In *28th SIGIR*, pp. 162–169, 2005.