Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le tecnologie e le scienze

RELAZIONE FINALE

Un'analisi del problema di imparzialità in information retrieval

Relatore Prof. Massimo Melucci
Dipartimento di ingegneria dell'informazione

Laureando: Michele Accorsi
Matricola N. 1193194

Anno Accademico 2021/2022

Indice

1. Introduzione

Information retrieval è un termine coniato nel 1950[1] da Calvin Mooers (Minneapolis, 1919-Cambridge, 1994) ed è oggi usato quasi esclusivamente in ambito informatico. Tuttavia l'*Information Retrieval* è un campo multidisciplinare capace di coinvolgere materie quali: psicologia cognitiva, architettura informativa, filosofia, design, comportamento umano sull'informazione, linguistica, semiotica, scienza dell'informazione e informatica.

L'espressione *Information Retrieval* (IR) identifica tutte le attività utilizzate per trovare le informazioni rilevanti per una specifica esigenza informativa di una particolare persona.

Un'esigenza informativa è l'insieme delle circostanze in cui una persona ha un problema da risolvere o un compito da svolgere e richiede informazioni importanti, utili o necessarie per la risoluzione del problema o lo svolgimento del compito.

Un *Information Retrieval System* (IRS) è un sistema informatico o una parte di un sistema informatico progettato e realizzato per svolgere in modo automatico compiti di IR.

Tuttavia, un compito di *Information Retrieval,* spesso richiede nozioni di tipo non razionale e per questo motivo i sistemi, essendo automatizzati, possono produrre risultati distorti.

Al giorno d'oggi, la presenza di *bias* negli IRS sta sollevando preoccupazioni sulla loro responsabilità sociale; l'imparzialità sta diventando un fattore estremamente importante nel momento della costruzione di sistemi per la ricerca di informazioni e per i suggerimenti dei contenuti online.

Un IRS, oltre ad eccellere nell'aiutare l'utente a trovare ciò di cui ha bisogno, ha anche la responsabilità sociale di essere imparziale.

La domanda che ci si pone è come si può ottimizzare la rilevanza delle informazioni sotto un qualche vincolo di imparzialità.

[1] Mooers, C. "The theory of digital non-numerical information and its implications to machine economics".

2. Obiettivi

- Analisi della procedura usata in *"How fair can we go: Detecting the Boundaries of Fairness Optimization in Information Retrieval"* [2] per risolvere il problema dell'imparzialità in IR.
- Definizione di un'estensione della procedura usata nell'articolo, nel caso In cui l'imparzialità non sia indipendente dalla rilevanza

[2] Gao, R. and Shah, C. (2019). "How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval".

3. Distorsione dei risultati e imparzialità in Information Retrieval

La rilevanza è una proprietà che rende l'informazione importante, utile o necessaria a soddisfare l'esigenza informativa.
Questa proprietà è fondamentale per la valutazione di un sistema di information retrieval, si vuole quindi ottimizzare la rilevanza del sistema.
Negli anni si è avuto modo di verificare che i sistemi di *information retrieval*, producono dei risultati distorti (*bias*).
Molti studi hanno dimostrato che gli IRS entrano in una *filter bubble* causando vari problemi all'utente:

- reperimento di informazioni non preferenziali
- distorsione della percezione di opinioni ed eventi
- rinforzamento di stereotipi sociali
- manipolazione
- reperimento non equo di opportunità e risorse

Per risolvere questi problemi di *bias* c'è il bisogno di considerare una rappresentazione imparziale dei risultati dell'IRS, senza sacrificare la loro rilevanza: c'è bisogno di trovare un compromesso tra le due proprietà.

Bias e imparzialità vengono spesso considerati come due facce della stessa medaglia.
Negli IRS l'imparzialità può essere vista in termini di imparzialità individuale o di gruppo:
La prima richiede che individui simili siano trattati in modo similare, la seconda richiede che un gruppo protetto (categorie demografiche, opinioni popolari, idee politiche) sia rappresentato in modo equo nell'IRS.
Il *bias* può provenire dai dati originali, può essere dell'algoritmo o del sistema, o può essere un *bias* cognitivo.
Lo studio di Ruoyuan Gao e Chirag Shah sviluppa una procedura generale che può essere implementata per vari sistemi di *Information Retrieval*.
In genere l'imparzialità viene modellata come un problema di ottimizzazione soggetto a delle costanti. Ci possono essere tre obiettivi di ottimizzazione:

- ottimizzazione della rilevanza
- ottimizzazione dell'imparzialità
- ottimizzazione di imparzialità e rilevanza congiuntamente
  La procedura di Gao, a differenza di altre procedure, non vuole risolvere solo uno di questi tre obiettivi, ma vuole trovare una procedura generale da utilizzare indipendentemente dal problema.

4. Procedura

Impostazione del problema

In un IRS, il sistema riceve un'interrogazione dall'utente, la elabora e restituisce un risultato. In particolare il sistema va ad esplorare tutte le possibili soluzioni (items) e raccoglie le soluzioni considerate rilevanti.

Si consideri dunque $D = \{a_1, a_2, \dots, a_N\}$ l'insieme di N items che definiscono il dataset.

Ogni item a_i è associato ad almeno $k \geq 2$ proprietà denotate dal vettore $< p_1, p_2, \dots, p_k >$.

Sia $f(D)$ una funzione che collega le proprietà ai dati.

Assumiamo che le proprietà siano indipendenti e identicamente distribuite provenienti da una distribuzione qualsiasi.

Si assuma che lo spazio delle possibili soluzioni è $S = \{S_1, S_2, \dots, S_j, \dots\} \subset D$, allora il minimo di $\{f_i(Sj)\}$ e il massimo di $\{f_i(Sj)\}$ sono il limite inferiore e quello superiore di S sulla proprietà i-esima.

L'obiettivo è quello di stimare la regione $R =< R_1, R_2, \dots, R_k >$ di S per ogni proprietà.

Ottenere gli intervalli

Si assuma che ogni item sia bidimensionale, $a_i =< r_i, g_i >$ in cui la prima dimensione rappresenta il punteggio della rilevanza, mentre la seconda dimensione rappresenta l'informazione che racchiude l'imparzialità dei gruppi e nel seguito sarà rappresentata dall'entropia.

Si assuma che ogni dimensione dell'item sia proveniente da una distribuzione di Bernoulli:

p_r sarà la probabilità che un item sia rilevante; $r_i \sim Be(p_r)$

p_g sarà la probabilità che un item appartenga al gruppo 1; $g_i \sim Be(p_g)$

Dato un insieme di $S \subset D$ di N items, sia:

f_r il punteggio medio di rilevanza dell'insieme S e

f_g l'entropia dei membri dei gruppi, $\overline{p_g}$ la proporzione di items del gruppo1

Allora:

$$f_r = \frac{\sum r_i}{n} \quad (1)$$

$$f_g = H(\vec{p}) = -\overline{p_g} log_2 \overline{p_g} - \left(1 - \overline{p_g}\right) log_2 \left(1 - \overline{p_g}\right)$$
$$= -\frac{\sum g_i}{n} log_2 \frac{\sum g_i}{n} - (1 - \frac{\sum g_i}{n}) log_2 (1 - \frac{\sum g_i}{n}) \quad (2)$$

Siano poi:
$\sum r_i \sim B(n, p_r), \sum g_i \sim B(n, p_g)$

Dalla diseguaglianza di Chebyshev[4] si ricavano:

$$f_r \in \left[p_r \pm \sqrt{\frac{p_r(1-p_r)}{n q_r}}\right] \quad (3)$$

e

$$f_g \in \left[p_g \pm \sqrt{\frac{p_g(1-p_g)}{n q_g}}\right] \quad (4)$$

Se si assume rilevanza e entropia indipendenti allora anche $\sum r_i$ e $\sum g_i$ lo sono.
In questo caso si può quindi definire la loro distribuzione congiunta per ottenere una regione per il range di $\sum r_i$ e $\sum g_i$,e conseguentemente anche per i valori di $< f_r, f_g >$.
Sia R_r il range di f_r con probabilità $1 - q_r$ e sia R_g il range di f_g con probabilità $1 - q_g$, allora:

$$P\left[f_r \in R_r, f_g \in R_g\right] \geq (1 - q_r)(1 - q_g) \quad (5)$$

Nei casi reali è tuttavia non è sempre possibile conoscere le distribuzioni di r_i e g_i, in questi casi si può procedere allo stesso modo tramite il teorema del limite centrale.

[3] Per ogni numero reale $t > 0, P(|x - \mu| \geq t\sigma) \leq \frac{1}{t^2}$

Punti ottimi

Una volta ottenuti gli intervalli per rilevanza ed entropia si possono stimare i valori ottimi per entrambe le dimensioni.

I valori ottimi dipenderanno dal peso che si dà ad ogni dimensione.

Siano w_r e w_g i pesi rispettivamente di f_r e f_g:

il valore ottimo si ottiene massimizzando f_r (o f_g) soggetta al vincolo:

$$\frac{f_r}{f_g} = \frac{w_g}{w_r}, f_r \in R_r, f_g \in R_g \quad (6)$$

5. Generalizzazioni

La procedura descritta nel capitolo precedente è utile in quanto è una procedura generale che può essere usata per dataset reali o dataset sintetici, inoltre a seconda delle esigenze informative dell'esperimento alcuni punti della procedura possono essere modificati ad-hoc.

Funzioni personalizzate

A seconda degli specifici scenari di applicazione, la procedura può essere implementata utilizzando funzioni personalizzate anziché le funzioni di rilevanza media e di entropia già descritte.
Un esempio di funzione di utilità diversa dalla rilevanza è il richiamo, mentre esempi di funzioni per la misura di imparzialità sono: *statistical parity* o *disparate impact.*

Dimensioni multiple

Un'altra caratteristica importante di questa procedura è che si può generalizzare per dimensioni multiple, questo è particolarmente utile quando si lavora con dataset reali perché dà la possibilità di considerare più di due fattori.
Un esempio può essere l'implementazione della procedura considerando alcune caratteristiche delle unità statistiche, ad esempio: il sesso, l'etnia o altre caratteristiche demografiche.

Rimozione dell'assunzione di indipendenza

In alcune situazioni reali la misurazione dell'imparzialità e quella di "utilità" potrebbero essere dipendenti, il capitolo successivo analizzerà una possibile estensione della procedura per questa casistica

6. Estensione della procedura rimuovendo l'assunzione di indipendenza

La rimozione dell'assunzione di indipendenza tra le componenti degli items, non rende più valida l'espressione (5), per trovare una soluzione a questo problema si può ragionare sulla definizione di probabilità condizionata.
In particolare, sullo spazio delle soluzioni si calcolano gli intervalli di confidenza per f_r e f_g. Il loro intervallo congiunto varierà per via della dipendenza tra le componenti.
Sia $S \subset D$ lo spazio delle soluzioni.
Sia N il numero totale di items.
Sia un item $a_i =< r_i, g_i >$ in cui le componenti rappresentano rispettivamente la rilevanza e l'entropia.
$P[f_r \in R_r, f_g \in R_g] \geq (1 - q_r)(1 - q_g)$ non è più valida.
Rimuovendo l'assunzione di indipendenza si ha che:
$P[f_r \in R_r, f_g \in R_g] \geq P[f_g \in R_g | f_r \in R_r] * P[f_r \in R_r]$ (7)

Il calcolo di $P[f_r \in R_r]$ è immediato, mentre per il calcolo di $P[f_g \in R_g | f_r \in R_r]$ si può procedere analizzando lo spazio delle soluzioni considerando unicamente gli items appartenenti a R_r.

7. Dimostrazione della procedura

Nel seguito si analizza come variano i risultati della procedura nel caso in cui l'assunzione di indipendenza tra le componenti sia ragionevole o meno.
Per l'analisi si è usato un dataset sintetico:
Si assume p_r=p_g=0.5 e si sono generati 100 items di parametri $< r_i, g_i >$, dove $r_i \sim Be(p_r)$ e $g_i \sim Be(p_g)$.

Assumendo indipendenza delle componenti nella procedura, si ottiene il risultato rappresentato in *figura1*, in cui sono rappresentati i limiti inferiori e superiori per le due componenti nello spazio delle soluzioni. Il rettangolo formatosi dall'intersezione degli intervalli rappresenta la regione di confidenza di livello 0.81 (0.9*0.9) di rilevanza e imparzialità congiuntamente.

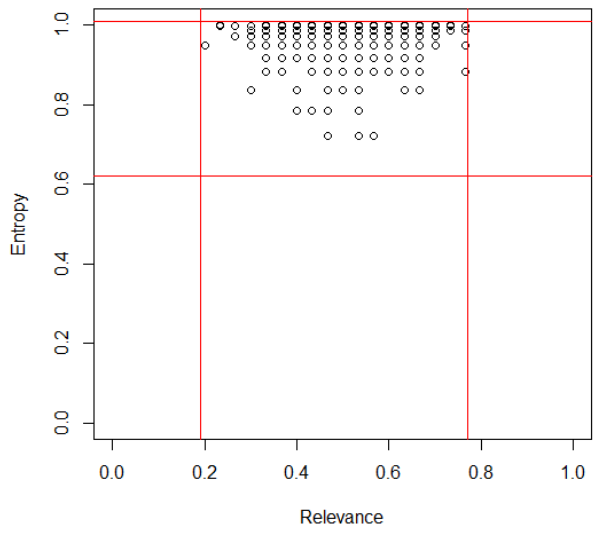

Figura 1: lo spazio delle soluzioni è illustrato dal grafico della densità ottenuto tramite simulazione (10^4 replicazioni). Le linee orrizzontali rappresentano l'intervallo di confidenza di livello 0.9 per l'imparzialità, mentre le linee verticali quello per la rilevanza.

Con la rimozione dell'assunzione di indipendenza tra le componenti varia la loro regione di confidenza congiunta:
Partendo dallo spazio delle soluzioni (figura2) si procede calcolando il limite inferiore e superiore sullo spazio delle soluzioni per la prima componente (figura3).

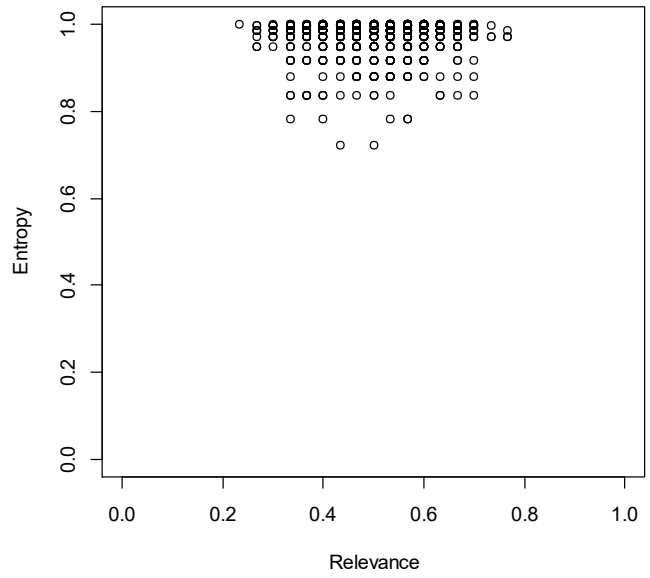

Figura2: Spazio delle soluzioni illustrato dalla densità ottenuta tramite simulazione

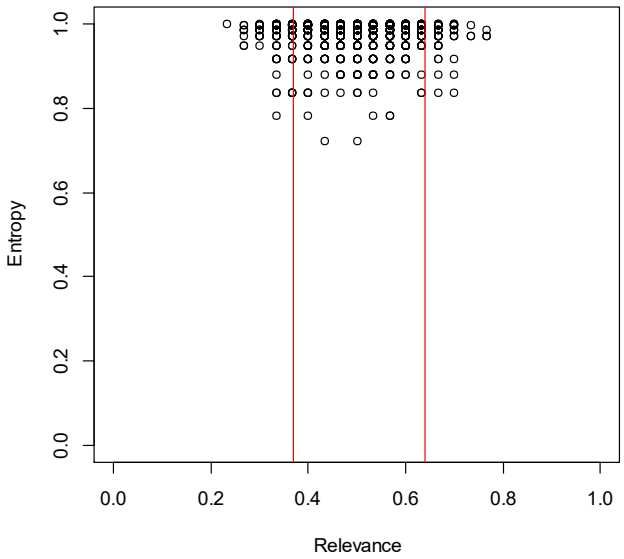

Figura3: Spazio delle soluzioni, le linee rosse delimitano la regione di confidenza per la rilevanza

Lo spazio delle soluzioni è ora formato dagli items contenuti in questo primo intervallo. Si procede calcolando il limite superiore e inferiore per la seconda componente sul nuovo spazio delle soluzioni (figura4).

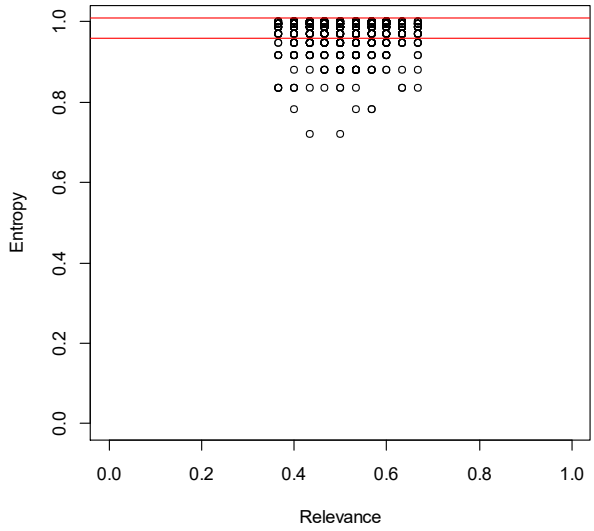

Figura 4: Spazio delle soluzioni rappresentato dagli items appartenenti alla regione di confidenza per la rilevanza, la regione di confidenza delimitata dalle linee rosse rappresenta la regione di confidenza congiunta

La regione finale rappresenta la regione di confidenza di rilevanza e imparzialità congiuntamente.

Approssimativamente si ottengono i seguenti risultati:

$P[f_r \in R_r]) \quad \cong \frac{9000}{10000} \cong 0.9$

$P\left[f_g \in R_g | f_r \in R_r\right] \quad \cong \frac{7200}{9000} \cong 0.8$

$P\left[f_r \in R_r, f_g \in R_g\right] \quad \cong 0.9 * 0.8 \cong 0.72$

8. Conclusioni

Con lo sviluppo di nuove tecnologie si stanno migliorando i sistemi di Information Retrieval, ma c'è la necessità di attuare nuove procedure che tengano conto anche dell'imparzialità dei risultati. Tramite l'uso di queste procedure si può fare uscire l'IRS dalla *filter bubble* garantendo risultati migliori per più utenti possibili.
La procedura descritta è semplice e facile da sviluppare, inoltre con delle estensioni può essere applicata in vari casi di studio.
Tuttavia, nell'impostazione della procedura bisogna assicurarsi la veridicità delle assunzioni, le quali, se non corrette, potrebbero portare a risultati fuorvianti.

9. Bibliografia

Croft, W., Metzler, D., and Strohman, T. (2009). Search Engines: Information Retrieval in Practise. Addison Wesley.

Sheldon M. Ross (2016). Calcolo delle probabilità.

Cicchitelli, D'Urso, Minozzo (2017). Statistica: principi e metodi

Ruoyuan Gao, Chirag Shah (2019). How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval