Approcci probabilistici basati su Continuous Active Learning (CAL) per le revisioni sistematiche in ambito medico

Giorgio Maria Di Nunzio

Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Padova
giorgiomaria.dinunzio@unipd.it

Abstract

Le revisioni sistematiche in ambito medico vengono utilizzate per raccogliere in modo affidabile e completo i risultati della ricerca scientifica. Vincoli di tipo temporale o risorse limitate possono compromettere la qualità di una revisione a tal punto da invalidarne i risultati. In questo articolo, presentiamo alcuni studi relativi ad un approccio probabilistico per la costruzione di sistemi automatici di revisione assistita basati su Continuous Active Learning (CAL).

1 Introduzione

Il processo di revisione sistematica della letteratura scientifica, in particolare in ambito medico, è rigorosamente suddiviso in fasi di lavoro (ricerca della letteratura, selezione dei lavori, valutazione critica, sintesi, ecc.) al fine di giungere a delle conclusioni accurate e corrette.[1] Tuttavia, la crescita esponenziale del numero di pubblicazioni scientifiche rende complesso e dispendioso, in termini finanziari e temporali, il recupero di documenti rilevanti per il completamento delle revisioni sistematiche [O'Mara-Eves *et al.*, 2015]: "it is unlikely that [there will be enough] time, skills and resources to find, appraise and interpret all this evidence and to incorporate it into healthcare decisions."[2] Difatti, nella fase preliminare del processo di revisione, i medici (o gli esperti nel settore) dedicano una quantità di tempo significativa all'interrogazione manuale di uno o più database al fine di raccogliere, se non tutti, la maggior parte dei documenti rilevanti per l'oggetto della ricerca. Spesso la difficoltà di recupero dei documenti è dovuta al motore di ricerca a disposizione. Nel caso specifico, i principali database di pubblicazioni scientifiche in ambito medico[3] supportano interrogazioni di tipo booleano che vengono costruite in modo incrementale: partendo da semplici parole chiave, si arriva a *query* estremamente complesse che possono coinvolgere centinaia di termini.

In questo contesto, i sistemi di revisione assistita, o *Technology Assisted Reviews* (TAR), aiutano l'utente che deve compilare una revisione a recuperare quante più informazioni rilevanti possibili limitando i tempi e lo sforzo cognitivo [Cormack e Grossman, 2017]. Ad esempio, nel caso della ricerca basata su keyword, sono stati condotti degli studi relativi ad approcci che combinano una strategia di ricerca booleana ad algoritmi di ordinamento per restituire dei risultati in maniera più efficiente [Karimi *et al.*, 2010]. I sistemi TAR più efficaci affrontano il problema addestrando un classificatore automatico attraverso un approccio di *Continuous Active Learning* (CAL): ogni volta che un utente legge un nuovo documento e lo giudica rilevante o meno ai fini della revisione, il sistema utilizza questo feedback per riordinare i rimanenti documenti in base alla nuova informazione acquisita [Singh *et al.*, 2018].

2 Sistemi TAR probabilistici

In questo contributo, presentiamo i risultati ottenuti da un approccio che segue il metodo *AutoTAR Continuous Active Learning* proposto da [Cormack e Grossman, 2016] e che si basa su un'interpretazione bidimensionale di uno dei modelli più efficaci di *Information Retrieval* (IR) chiamato BM25 [Robertson e Zaragoza, 2009]. Il BM25 associa ad ogni termine presente in un documento un 'peso', denominato *relevance weight*, che stima la probabilità che quel termine appaia nei documenti giudicati rilevanti e non. In questo modello, la probabilità che un documento sia rilevante rispetto ad una *query* dell'utente è proporzionale alla somma dei *relevance weight* dei termini del documento che sono presenti nella stessa *query*.

In particolare, l'interpretazione bidimensionale del BM25 [Di Nunzio, 2014] permette di visualizzare le probabilità che un documento sia rilevante o meno per la revisione sistematica. In questo approccio, il *relevance weight* di un termine t_i, indicato con w_i^{BM25}, può essere decomposto in due parti, $w_i^{BM25,\mathcal{R}}$ e $w_i^{BM25,\mathcal{NR}}$, in maniera da distinguere il contributo del peso del termine nell'insieme dei documenti rilevanti, R, o non rilevanti, NR. Dato un documento d, la probabilità che esso sia rilevante o meno è data da:

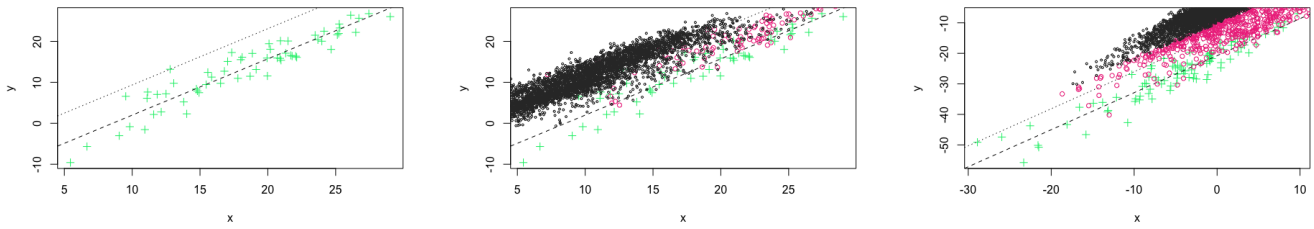
$$P(d|\mathcal{R}) \propto \sum_{w_i \in d} w_i^{BM25,\mathcal{R}} \quad (1)$$

$$P(d|\mathcal{NR}) \propto \sum_{w_i \in d} w_i^{BM25,\mathcal{NR}} \quad (2)$$

[1] https://training.cochrane.org/handbook

[2] Cochrane Handbook for Systematic Reviews of Interventions http://handbook-5-1.cochrane.org

[3] https://www.ncbi.nlm.nih.gov/pubmed/



(a) (b) (c)

Figura 1: Visualizzazione di alcune fasi di una revisione sistematica assistita interattiva. Le croci verdi rappresentano i documenti rilevanti (1a), i cerchi rossi i documenti non rilevanti e i cerchi neri i documenti non ancora letti (1b). Il sistema suggerisce automaticamente quando terminare la revisione se non ci sono più documenti da leggere nell'area d'interesse compresa tra le due linee tratteggiate (1c).

In questo modo, è possibile trasformare il problema della classificazione o scelta dei documenti da presentare all'utente come un problema di selezione della migliore retta di separazione di un insieme di punti (in generale non linearmente separabili) su un piano bidimensionale [Di Nunzio, 2017]:

$$\underbrace{P(d|\mathcal{NR})}_{y} < m \underbrace{P(d|\mathcal{R})}_{x} + q \tag{3}$$

dove i parametri m e q possono essere ottimizzati in base al tempo alle risorse a disposizione [Di Nunzio, 2018a; Di Nunzio *et al.*, 2018]. In Figura 1, mostriamo una sequenza di visualizzazione di questo sistema durante una sessione di lavoro interattiva [Di Nunzio, 2018b].

3 Conclusioni

In questo articolo, abbiamo discusso il problema delle revisioni sistematiche in ambito medico e abbiamo presentato alcuni studi relativi ad una possibile soluzione che utilizza un sistema TAR basato su Continuous Active Learning. Questo approccio interattivo ha il vantaggio di poter essere visualizzato graficamente in maniera da decidere, insieme all'esperto in ambito medico, la miglior strategia di impiego delle risorse di tempo e/o economiche al fine di non compromettere la qualità della revisione da portare a termine.

Riferimenti bibliografici

[Cormack e Grossman, 2016] Gordon V. Cormack e Maura R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1039–1048, New York, NY, USA, 2016. ACM.

[Cormack e Grossman, 2017] Gordon V. Cormack e Maura R. Grossman. Technology-assisted review in empirical medicine: Waterloo participation in CLEF ehealth 2017. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[Di Nunzio *et al.*, 2018] Giorgio Maria Di Nunzio, Maria Maistro, e Federica Vezzani. A gamified approach to naïve bayes classification: A case study for newswires and systematic medical reviews. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1139–1146, 2018.

[Di Nunzio, 2014] G. M. Di Nunzio. A New Decision to Take for Cost-Sensitive Naïve Bayes Classifiers. *Information Processing & Management*, 50(5):653 – 674, 2014.

[Di Nunzio, 2017] Giorgio Maria Di Nunzio. Interactive Text Categorisation: The Geometry of Likelihood Spaces. *Studies in Computational Intelligence*, 668:13–34, 2017.

[Di Nunzio, 2018a] Giorgio Maria Di Nunzio. Finding all the needles in the haystack. A system to estimate the costs of e-discovery and systematic reviews. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018.*, page 106, 2018.

[Di Nunzio, 2018b] Giorgio Maria Di Nunzio. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 672–677, 2018.

[Karimi *et al.*, 2010] Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, e Justin Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, 10:58–58, 2010.

[O'Mara-Eves *et al.*, 2015] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, e Sophia Ananiadou. Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Systematic Reviews*, 4(1):5, 2015.

[Robertson e Zaragoza, 2009] Stephen E. Robertson e Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[Singh *et al.*, 2018] Gaurav Singh, James Thomas, e John Shawe-Taylor. Improving active learning in systematic reviews. *CoRR*, abs/1801.09496, 2018.