26.Novembre

Come realizzare un data mining da 60 milioni di articoli

Il processo creativo va dalla complessità alla semplicità.

Quando **Iosif Gidiotis** ha iniziato quest'anno i suoi studi di dottorato in tecnologia educativa, è rimasto incuriosito dalle notizie secondo cui nuovi strumenti basati sull'intelligenza artificiale (AI) avrebbero potuto aiutarlo a digerire la letteratura nella sua disciplina, a realizzare un ***data mining***
estraendo informazioni attraverso metodi automatici o semi automatici

Con il numero di articoli in crescita - in tutta la scienza, l'anno scorso ne sono stati pubblicati quasi 3 milioni - un assistente di ricerca sull'intelligenza artificiale "sembra fantastico", afferma Gidiotis, che studia al KTH Royal Institute of Technology.

Sperava che l'intelligenza artificiale potesse trovare documenti più pertinenti rispetto ad altri strumenti di ricerca e riassumerne i punti salienti.

Ha vissuto un po' di delusione. Quando ha provato strumenti di intelligenza artificiale come quello chiamato

Elicit

Analyze research papers at superhuman speed
Automate time-consuming research tasks like summarizing papers,
extracting data, and synthesizing your findings.

ha scoperto che solo alcuni dei documenti restituiti erano rilevanti e che i riepiloghi di ***Elicit*** non erano abbastanza accurati da convincerlo. "Il tuo istinto è quello di leggere il documento reale per verificare se il riassunto è corretto, quindi non fa risparmiare tempo", dice. (***Elicit*** afferma che sta continuando a migliorare i suoi algoritmi per i suoi 250.000 utenti abituali, che in un sondaggio gli hanno attribuito il merito di aver risparmiato loro, in media, 90 minuti a settimana nella lettura e nella ricerca.)

Creato nel 2021 da un'organizzazione di ricerca senza scopo di lucro, ***Elicit*** fa parte di una scuderia crescente di strumenti di intelligenza artificiale che mirano ad aiutare gli scienziati a orientarsi nella letteratura.

“C'è un'esplosione di queste piattaforme”, afferma ***Andrea Chiarelli,*** che segue gli strumenti di intelligenza artificiale nell'editoria per la società ***Research Consulting***. Ma i loro sviluppatori devono affrontare delle sfide. Tra questi: i sistemi generativi che alimentano questi strumenti sono inclini a creare contenuti falsi “allucinanti” e molti dei documenti perquisiti sono dietro pagamento. Gli sviluppatori sono anche alla ricerca di modelli di business sostenibili; per ora, molti offrono un accesso introduttivo gratuito. ***“È molto difficile prevedere quali strumenti di intelligenza artificiale prevarranno, e c’è un certo livello di hype, ma sono molto promettenti”, afferma Chiarelli.***

Come **ChatGPT** e altri **modelli LLM** (Large Language Models), i nuovi strumenti ***vengono “addestrati”*** su un gran numero di campioni di testo, imparando a riconoscere le relazioni tra le parole.
Queste associazioni consentono agli algoritmi di riassumere i risultati della ricerca. Inoltre, identificano i contenuti pertinenti in base al contesto del documento, producendo risultati più ampi rispetto a una query che utilizza solo parole chiave.

Costruire e formare un **LLM** da zero è troppo costoso per tutti tranne che per le organizzazioni più ricche, afferma

Petr Knoth, direttore di CORE, il più grande archivio mondiale di documenti ad accesso aperto. Quindi Elicit e altri utilizzano **LLM** open source esistenti addestrati su una vasta gamma di testi, molti dei quali non scientifici.

Alcuni strumenti vanno oltre. ***Elicit,*** ad esempio, organizza gli articoli per concetto. Una domanda su un eccesso di caffeina porta a serie separate di documenti sulla riduzione della sonnolenza e sul deterioramento delle prestazioni atletiche. Una versione premium, che costa $ 10 al mese, utilizza una programmazione interna aggiuntiva per aumentare la precisione.

Una caratteristica dello strumento **SEMANTIC READER** creato dall'organnizzazione non profit **Allen Insitute for AI**, funziona come un evidenziatore automatico di inchiostro, che gli utenti possono personalizzare per applicare colori diversi alle dichiarazioni su novità, obiettivi e altri temi.
Fornisce "*una rapida diagnosi, un triage, per stabilire se vale la pena impegnarsi [in un articolo] il che "è molto prezioso",* afferma

Eytan Adar, uno scienziato informatico d*ell'Università del Michigan* che ha provato una versione iniziale prima di una versione ampliata. uno è stato presentato il mese scorso. Molti strumenti annotano anche riassunti con estratti di documenti su cui si basano, consentendo agli utenti di giudicarne da soli l'accuratezza.

Per cercare di evitare di generare risposte false, **l'Allen Institute** gestisce il Semantic Reader utilizzando una ***suite di LLM***, compresi quelli formati su articoli scientifici. Ma l'efficacia di questo approccio è difficile da misurare.

"*Si tratta di problemi tecnici difficili ai margini della nostra comprensione",* afferma ***Michael Carbin***, un informatico del *Massachusetts Institute of Technology* che ha contribuito a sviluppare un algoritmo per riassumere la letteratura medica.

Secondo **Dan Weld**, capo scienziato presso l'archivio di documenti **Semantic Scholar dell'Allen Institute,** "*In questo momento, lo standard migliore che abbiamo è quello di avere uno sguardo umano molto istruito [sull'output dell'intelligenza artificiale] e analizzarlo attentamente."*

L'istituto ha raccolto feedback da oltre 300 studenti laureati retribuiti e migliaia di tester volontari. I controlli di qualità hanno rivelato che l'applicazione di **Scim** a documenti non informatici ha prodotto problemi, quindi l'istituto offre attualmente **Scim** solo per circa 550.000 documenti di informatica.

Altri ricercatori sottolineano che gli strumenti di intelligenza artificiale raggiungeranno il loro potenziale solo se gli sviluppatori e gli utenti potranno accedere al testo completo dei documenti per informare i risultati di ricerca e l'analisi dei contenuti.

"*Se non possiamo accedere al testo, la nostra visione della conoscenza racchiusa in quei testi è limitata",* afferma **Karin Verspoor**, linguista computazionale presso la RMIT University di Melbourne.

Anche **Elsevier,** il più grande editore scientifico del mondo, limita i suoi strumenti di intelligenza artificiale agli abstract degli articoli. Ad agosto, l'azienda commerciale ha introdotto una funzione di ricerca assistita all'intelligenza artificiale del suo data base di **SCOPUS** i cui elenchi di 93 milioni di pubblicazioni di ricerca lo rendono uno dei più grandi per gli scienziati. In risposta a una query, i suoi algoritmi identificano gli abstract più rilevanti e utilizzano una versione di ChatGPT per fornire un riepilogo generale. (Lo strumento ristruttura le query degli utenti per ridurre le risposte inventate che ChatGPT a volte fornisce.) Scopus AI raggruppa anche gli abstract per concetto.

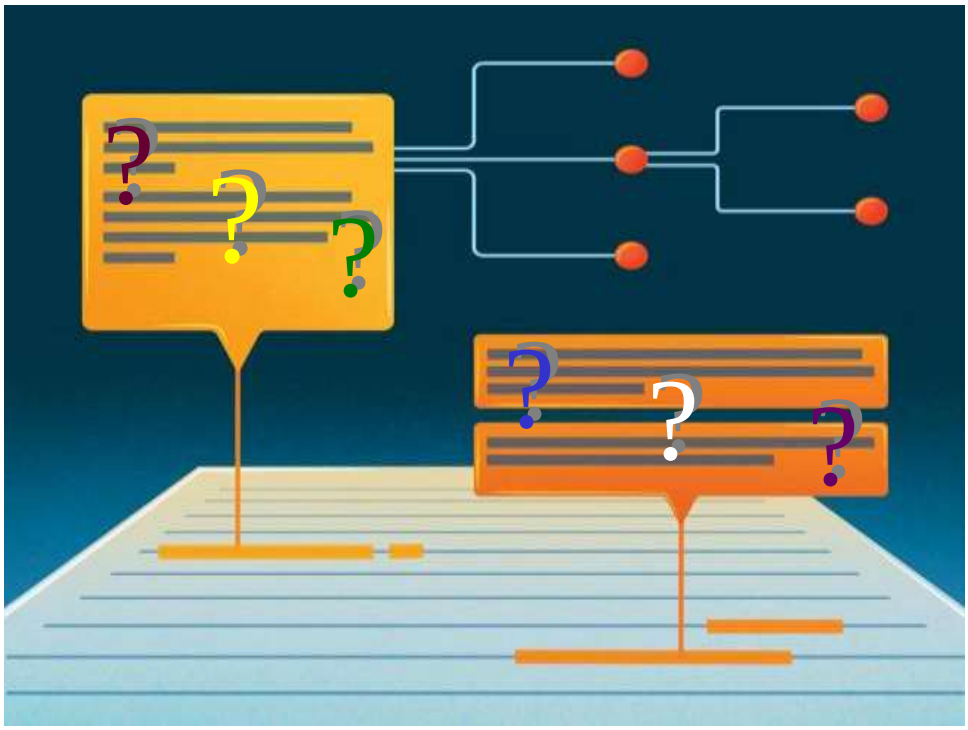
L'approccio basato sui soli abstract è coerente con i termini degli accordi di licenza di **Elsevier** con altri editori che consentono di elencare gli abstract dei loro articoli in Scopus, afferma **Maxim Khan**, vicepresidente senior per i prodotti di analisi e le piattaforme dati di **Elsevier**. Per ora, dicono gli utenti a Elsevier, questo approccio è sufficiente per "[aiutare] i ricercatori in campi interdisciplinari che cercano di comprendere rapidamente un particolare argomento", afferma.

L'Allen Institute ha adottato un approccio diverso: ha negoziato accordi con più di 50 editori che consentono ai suoi sviluppatori di estrarre dati dal testo completo dei documenti protetti da paywall. Weld afferma che quasi tutti gli editori hanno offerto l'accesso gratuitamente perché l'intelligenza artificiale indirizza loro il traffico.

Anche così, le restrizioni di licenza limitano gli utenti di **Semantic Reader** all'accesso al testo completo di ***soli 8 milioni dei 60 milioni di articoli a testo completo*** di Semantic Scholar. E Knoth

sostiene che tali trattative richiedono un tempo proibitivo per la sua organizzazione. "Difficilmente può essere visto come un terreno di gioco equo e paritario", afferma Knoth, il cui archivio finanziato dall'università lavora per sviluppare strumenti per aiutare gli scienziati a esplorarne il contenuto.
Abilitare il ***data mining*** su larga scala richiederà anche che più autori ed editori adottino formati non PDF che aiutino le macchine a digerire in modo efficiente i contenuti di un articolo. Una direttiva della Casa Bianca nel 2022 richiede che i documenti prodotti con finanziamenti federali siano leggibili dalle macchine, ma le agenzie devono ancora proporre i dettagli.

Nonostante le sfide, gli informatici stanno già cercando di sviluppare IA più sofisticate, in grado di raccogliere informazioni ancora più ricche dalla letteratura. Vogliono raccogliere indizi per migliorare la scoperta di farmaci e aggiornare continuamente le revisioni sistematiche. La ricerca sostenuta dalla **Defense Advanced Research Projects Agency** ha esplorato sistemi in grado di generare automaticamente ipotesi scientifiche, identificando le lacune nelle conoscenze esistenti come rivelato dagli articoli pubblicati.



Ma per ora, gli scienziati che utilizzano strumenti di intelligenza artificiale devono mantenere un sano livello di scetticismo, afferma **Hamed Zamani** *dell'Università del Massachusetts, Amherst,* che studia i sistemi interattivi di accesso alle informazioni. Gli **LLM** "miglioreranno sicuramente. **Ma in questo momento hanno molte limitazioni. Forniscono informazioni errate.** Quindi gli scienziati dovrebbero esserne molto consapevoli e ricontrollare i loro risultati".

A proposito dell'incompatibilità del pesce con il vino rosso

Per anni, ai commensali è stato detto che bere vino rosso mentre si mangiava pesce può produrre uno sgradevole retrogusto di pesce. ***La regola generale è stata il vino rosso con la carne***, ***il vino bianco con il pesce.*** Ma ogni regola ha le sue eccezioni.
I frutti di mare possono avere un buon sapore con alcuni rossi, mentre alcuni bianchi possono rovinare il pasto. Qual è il fattore comune?

I ricercatori della **Mercian Corp. di Fujisawa,** in Giappone, una divisione della quale produce vino e liquori, hanno deciso di scoprirlo e di pubblicare i risultati un recente numero del ***Journal of Agricultural and Food Chemistry*** **.**

Takayuchi et al

Iron Is an Essential Cause of Fishy Aftertaste Formation in Wine and Seafood Pairing

Hanno condotto un esperimento con sette degustatori esperti a cui sono state offerte 3***8 varietà di rosso*** e ***26 tipi di bianco***. Nel corso di quattro sessioni, i volontari hanno assaggiato i campioni, insieme a pezzi di capesante, i frutti di mare che con maggiore probabilità producono l'effetto pesce. Quindi i ricercatori hanno analizzato chimicamente i vini per un possibile collegamento con il retrogusto.

Il "colpevole" sembra essere il ferro.

Quando il contenuto dell'elemento superava i ***2 milligrammi per litro circa***, l'esperienza del pranzo a base di pesce diventava amara. Il team ha ricontrollato i risultati immergendo pezzi di capesante essiccate in campioni di vino. Le capesante inzuppate nel vino a basso contenuto di ferro avevano un odore normale, ma i pezzi immersi in campioni ad alto contenuto di ferro puzzavano di pesce.

Sono state trovate forti correlazioni positive tra l'intensità del retrogusto di pesce e la concentrazione sia del ferro totale che degli ioni ferrosi. Inoltre, l'intensità del retrogusto di pesce è stata aumentata dall'aggiunta di ioni ferrosi nel vino modello e soppressa dalla chelazione degli

ioni ferrosi nel vino rosso. In terzo luogo, potenti composti volatili dal retrogusto di pesce, come ***esanale, eptanale, 1-otten-3-one, (E , Z)-2,4-eptadienale, nonanale e decanale****, sono stati determinati mediante gascromatografia-olfattometria e gascromatografia. –spettrometria di massa su capesante essiccate* ***imbevute di vino rosso****. La formazione di questi composti dipendeva dalla dose di* ***ione ferroso*** *nel vino modello.*

Questi risultati suggeriscono che lo ***ione ferroso*** è un composto chiave nella formazione del retrogusto di pesce negli abbinamenti vino-frutti di mare entro l'intervallo di concentrazione comunemente riscontrato nel vino.

I ricercatori riferiscono di non aver ancora isolato il composto presente nelle capesante che reagisce con il vino, ma sospettano che si tratti di un ***acido grasso insaturo***, che potrebbe degradarsi rapidamente e rilasciare l'odore di pesce in decomposizione se esposto al *ferro.*

La *quantità di ferro* contenuta in un vino dipende dalla quantità nel terreno in cui è stata coltivata l'uva, nonché da altri fattori come il modo in cui l'uva viene raccolta e lavorata. ***Il vino rosso tende ad avere un contenuto di ferro*** più elevato, da qui l'avvertimento di non mescolarlo con i frutti di mare.
I identificato il "colpevole" sono scagionati i ***polifenoli o l'anidride solforosa*** indiziati come responsabili di produrre la sensazione spiacevole

Questi componenti rappresentano una percentuale maggiore del contenuto del vino rispetto al *ferro* e poiché il *ferro* non "induce cambiamento di colore, ossidazione accelerata o torbidità, i viticoltori tendono a ignorare il suo potenziale ruolo di "rovinatore" del pasto.
Ma le nuove scoperte, dice, offrono ai produttori di vino l'opportunità di riconsiderare gli aspetti negativi della contaminazione da ferro.

Considerazioni personali
Ci sono ragioni migliori per evitare il vino rosso con il pesce: qualsiasi vino rosso robusto, indipendentemente dal contenuto di ferro, probabilmente sopraffarebbe il sapore delicato e sottile di molti piatti a base di pesce. Il vino rosso si abbina meglio "con un grosso stufato o una succosa "fiorentina".