I bot dell'AI all’assalto delle biblioteche pubbliche

di **Antonella De Robbio**
26 LUGLIO 2025

https://ilbolive.unipd.it/it/news/societa/bot-dellai-allassalto-biblioteche-pubbliche

Immagine generata dall'AI

Nel cuore dell’University of North Carolina at Chapel Hill, tra scaffali ordinati e server ben protetti, è esplosa una guerra silenziosa e imprevista. Il caso delle biblioteche UNC racconta un paradosso: nel momento in cui le informazioni accademiche vengono messe a disposizione pubblicamente, sono proprio queste a diventare preda di un’accumulazione predatoria che minaccia la fruibilità del servizio per gli utenti reali. Un effetto collaterale della fame insaziabile dell’intelligenza artificiale, che trasforma cataloghi e bibliografie da strumenti per il sapere in caveau da depredare.

Era il 2 dicembre 2024 quando il primo allarme raggiunse il dipartimento IT delle biblioteche universitarie: una segnalazione innocua, una semplice richiesta di aiuto del personale che rilevava un errore anomalo nel catalogo online. “Questo sito web è sovraccarico (coda piena)”, riportava il messaggio. Nulla di insolito in un periodo d’esami, si pensò inizialmente. Eppure, già da quella prima avvisaglia, qualcosa strideva. A raccontare la vicenda oggi è David Romani, amministratore di sistema e responsabile della sicurezza che insieme a Tim Shearer e Jason Casden ha collaborato con il team IT della biblioteca e con i colleghi del campus per contrastare i bot che attaccavano il catalogo online della biblioteca.

Quello che sembrava un normale picco di traffico si stava rapidamente trasformando in un fenomeno anomalo, con il catalogo digitale incapace di reggere il carico. Non erano gli

studenti, i docenti o il personale a intasare il sistema, ma un'ondata di richieste continue e senza volto, capace di raggiungere oltre cinquecento interrogazioni simultanee, cinque volte il volume massimo normalmente gestito dalla piattaforma. In un primo momento il protocollo fu quello standard: individuare indirizzi IP sospetti, bloccare quelli provenienti da fonti malevole o già noti per comportamenti abusivi. La biblioteca disponeva già di una lista nera impressionante: più di quattro milioni di IP erano stati bloccati nel tempo per tentativi di accesso dannosi, a cui si sommavano altri milioni di indirizzi filtrati dall'amministrazione universitaria. Ma stavolta lo scenario era diverso e molto più insidioso. Le richieste non arrivavano da server sconosciuti o da Paesi con dubbia reputazione informatica, ma da indirizzi distribuiti capillarmente negli Stati Uniti, appoggiandosi a fornitori di servizi internet perfettamente legittimi come AT&T, Spectrum e Verizon. Ogni richiesta si presentava con la normalità disarmante delle consultazioni quotidiane che avvengono in una biblioteca di ricerca: il catalogo non riusciva a distinguere l'attacco dalle ricerche autentiche.

La lotta si fece serrata. Jason Casden, uno degli ingegneri IT coinvolti, descrive l'angoscia di quel momento come l'essere catapultati in un film di hacker: centinaia di richieste al minuto, un flusso ininterrotto e caotico. A rendere la faccenda più inquietante furono i primi dettagli emersi dalle analisi: un improvviso picco di ricerche sui temi più improbabili, come la musica finlandese, passata da poche decine a oltre undicimila interrogazioni in un solo giorno, provenienti da migliaia di sorgenti diverse. La caratteristica più spiazzante era la complessità delle interrogazioni. I bot combinavano opzioni di ricerca in modo inverosimile, selezionando più di venti filtri a volta, una strategia quasi impensabile per qualsiasi utente umana. In risposta, il team cercò di tamponare l'emorragia introducendo regole temporanee: bloccare chi effettuava interrogazioni troppo articolate in sequenza. Per qualche giorno funzionò, ma gli aggressori tornarono, ancora più elusivi.

Il secondo assalto giunse dalla Cina. Questa volta il vantaggio per i bibliotecari fu di natura tattica: era possibile bloccare ampie fasce di indirizzi IP cinesi in modo più rapido ed efficace. Ma bastarono pochi giorni perché il fenomeno mutasse ancora, rimescolandosi su scala globale e facendo ripartire la caccia ai fantasmi digitali.

Nel gennaio 2025, dopo settimane di tensione crescente, venne adottata una nuova linea difensiva: un sofisticato firewall alimentato da intelligenza artificiale. Era l'unico modo per distinguere con precisione il traffico legittimo da quello nocivo senza paralizzare completamente il servizio. Il risultato fu una stabilizzazione parziale: gli attacchi non cessarono, ma vennero finalmente contenuti, evitando il collasso del sistema.

Dietro questo attacco si celava un movente sempre più frequente nel mondo accademico: l'addestramento selvaggio dei modelli linguistici di grandi dimensioni (LLM). Era il fenomeno dello *scraping* sistematico a colpire le biblioteche: bot progettati per estrarre quantità massicce di dati bibliografici, utili a riempire i database di intelligenze artificiali sempre più voraci. Shearer, uno dei coordinatori del team IT, racconta come molti colleghi in altre università stessero affrontando lo stesso identico problema: un prelievo indiscriminato di contenuti pubblici da parte di attori ignoti per nutrire algoritmi destinati a mercati commerciali globali.

Eppure, per quanto sofisticati, questi assalti non erano necessariamente intelligenti. "I bot erano incredibilmente evasivi, ma estremamente inefficienti," osserva Casden. "Passavano giorni a recuperare gli stessi dati, con percorsi sempre più contorti, senza preoccuparsi dell'efficienza o della logica nell'interrogare il catalogo."

La biblioteca ha potuto reggere l'urto grazie a una cultura tecnologica consolidata: sistemi robusti, personale competente, un'infrastruttura difensiva affinata nel tempo. "Il nostro catalogo non è mai crollato, nonostante tutto," sottolinea Shearer, ricordando l'impegno di almeno sette tecnici IT impegnati a tempo pieno per settimane intere. "Molte istituzioni non dispongono di queste risorse, sono molto più esposte."

Se guardiamo a quanto accaduto oltreoceano con l'assalto al catalogo della biblioteca universitaria della University of North Carolina at Chapel Hill, è inevitabile chiedersi quanto le biblioteche accademiche italiane sarebbero pronte a gestire una situazione analoga. La risposta, purtroppo, non induce all'ottimismo.

Il primo elemento che salta agli occhi è la sproporzione strutturale. Negli Stati Uniti, almeno nei grandi atenei come UNC-Chapel Hill, esistono team IT dedicati, strutture informatiche robuste, collaborazioni trasversali con i dipartimenti tecnologici dell'università. Romani riflette sul futuro con pragmatismo: "Facciamo bene il nostro lavoro quando nessuno si accorge della nostra esistenza. Ma questo significa anche che l'AI generativa continuerà a bussare alla porta. E noi dovremo essere pronti a difendere la biblioteca, silenziosamente". Perché, nell'epoca dei bot e dei modelli linguistici, anche i bibliotecari si ritrovano in prima linea, soldati inconsapevoli di una nuova guerra digitale.

In Italia l'infrastruttura digitale delle biblioteche accademiche è spesso frammentata, con personale IT limitato, numericamente sottodimensionato rispetto alla mole di sistemi da gestire, o addirittura esternalizzato in convenzioni non sempre agili. La sicurezza informatica delle biblioteche, se esiste, è affidata in gran parte ai sistemi informatici centrali degli atenei, con poche risorse specializzate che conoscano in profondità le dinamiche dei cataloghi e delle piattaforme bibliografiche.

L'altro nodo critico è la scarsa consapevolezza del problema. In Italia si parla moltissimo di digitalizzazione, molto meno di protezione dei sistemi digitali bibliotecari dagli effetti collaterali delle nuove tecnologie. I cataloghi accademici italiani, così come gli OPAC (*On-line public access catalog*) di sistema bibliotecario nazionale, sono ampiamente accessibili dall'esterno, ma non esistono protocolli diffusi per intercettare *scraping* massivi. Il tema stesso del data scraping è anzi raramente oggetto di discussione nei contesti bibliotecari italiani, se non per la questione – molto più formale – dei diritti sui dati bibliografici.

Eppure anche qui, seppur più sommessamente, i segnali non mancano. Ci sono stati casi documentati di sovraccarico dei cataloghi durante periodi d'esame, rallentamenti inspiegabili, anomalie nei log degli accessi che vengono liquidate come "picchi di traffico": non esiste però un monitoraggio sistematico capace di distinguere un uso intensivo normale da un attacco strutturato. I *firewall* universitari si concentrano sugli attacchi più classici, come le intrusioni malevole e i *malware*; nessuno sembra interrogarsi su cosa accada quando decine di migliaia di query automatizzate simulano l'attività di centinaia di studenti contemporaneamente.

C'è poi un problema più profondo, culturale. Le biblioteche italiane, soprattutto quelle accademiche, vivono ancora una relazione ambivalente con la dimensione digitale: da un lato difendono con fatica la propria funzione pubblica di accesso aperto al sapere, dall'altro non hanno avuto finora gli strumenti per governare il confine tra accessibilità e sfruttamento. Su questo si innestano alcune considerazioni che riguardano più in generale il ruolo delle biblioteche nell'era dell'intelligenza artificiale. Il caso della UNC svela un aspetto trascurato nel dibattito pubblico: non sono solo i grandi portali giornalistici o le

piattaforme di social media a essere presi di mira dagli algoritmi di *scraping*, ma anche le più tradizionali istituzioni della conoscenza accademica. Quella che fino a pochi anni fa era una pratica marginale – estrarre metadati bibliografici per scopi di ricerca interna – oggi è diventata un gigantesco ingranaggio globale alimentato dalla necessità delle aziende IA di addestrare modelli linguistici sempre più vasti.

La cosa più significativa non è tanto la presenza dei bot, ormai quasi una normalità nell'ecosistema digitale, quanto la loro qualità "mimetica". Non parliamo di violazioni dirette o di furti di dati sensibili, ma di un assalto che si maschera da normale uso quotidiano, difficile da distinguere e quindi ancora più insidioso. È un attacco che non punta a distruggere ma a divorare: un consumo invisibile, silenzioso, ma devastante per la qualità del servizio.

Colpisce anche l'enorme disparità di risorse tra grandi istituzioni con team IT avanzati e realtà più piccole e non è difficile immaginare biblioteche minori che vedranno i propri sistemi collassare semplicemente per essere finite nel mirino di scraping massivo.

Infine, emerge un tema etico cruciale: se l'addestramento di LLM si regge sull'appropriazione sistematica di contenuti pubblici, a pagare il prezzo sono proprio le istituzioni che quei contenuti custodiscono e organizzano. È una versione aggiornata del colonialismo digitale: l'estrazione di valore da territori comuni (in questo caso il catalogo bibliografico) senza alcun ritorno per la comunità che lo ha costruito.

In questo senso la situazione italiana appare ancora più fragile. In un contesto in cui il personale bibliotecario si riduce, i fondi stagnano e l'età media cresce, chi sarà in grado di affrontare quella che gli statunitensi definiscono senza esitazioni una "corsa agli armamenti"? Chi, nelle nostre università, ha la responsabilità e le competenze per alzare la guardia contro i bot che saccheggiano i cataloghi per alimentare intelligenze artificiali lontane, magari anche a beneficio di mercati esterni al mondo accademico?

Il rischio è che il problema venga sottovalutato fino a quando non sarà troppo tardi con il rischio peggiore che si scelga la strada più semplice: restringere l'accesso, limitare l'apertura dei cataloghi, contraddicendo quella missione pubblica che le biblioteche hanno sempre difeso.

Non c'è dubbio che il mondo delle AI guardi ai dati bibliografici come a un enorme serbatoio di addestramento gratuito. La questione è se le biblioteche, in Italia, vogliano accettare di diventare inconsapevolmente la miniera di dati di qualcun altro o se vorranno rivendicare un protagonismo consapevole nella gestione del sapere digitale: dotarsi finalmente di infrastrutture e competenze tecnologiche adeguate e, contemporaneamente, sviluppare un discorso politico chiaro sul ruolo delle biblioteche nell'epoca dell'intelligenza artificiale.