Nicola De Bellis

Bibliometrics and Citation Analysis. From the Science Citation Index to Cybermetrics

Lanham, Md., Scarecrow press, 2009, p. 450

È con colpevole ritardo che presentiamo questo volume di eccezionale interesse al lettore di "Biblioteche oggi", il quale vi troverà non solo i principi fondamentali di una scienza – la bibliometria – oggi del tutto trascurata nella letteratura professionale italiana, ma anche un'accurata ricognizione dei futuri percorsi dell'*information retrieval* e delle istanze sciento-metriche in ambito biblioteconomico.

La bibliometria non gode di ottima stampa nella letteratura professionale per varie ragioni. Se nell'opinione comune l'economia è spesso tacitata come scienza "triste", si sarebbe tentati di affermare che la bibliometria è scienza "depressa", se è vero che propende, almeno a uno sguardo superficiale, a tralasciare l'analisi del contenuto – il ramo "nobile" della bibliografia e della biblioteconomia – per concentrarsi unicamente sulla misurazione su base statistica del successo di un'opera, di una scoperta, di un autore, o di un dipartimento universitario. Il potere della *doxa*, il successo quantitativo di un articolo o di un saggio, sembra prevalere sulla missione del bibliotecario, che è quella di mettere a disposizione del pubblico dei contenuti di qualità, sapientemente scelti nel magma delle conoscenze.

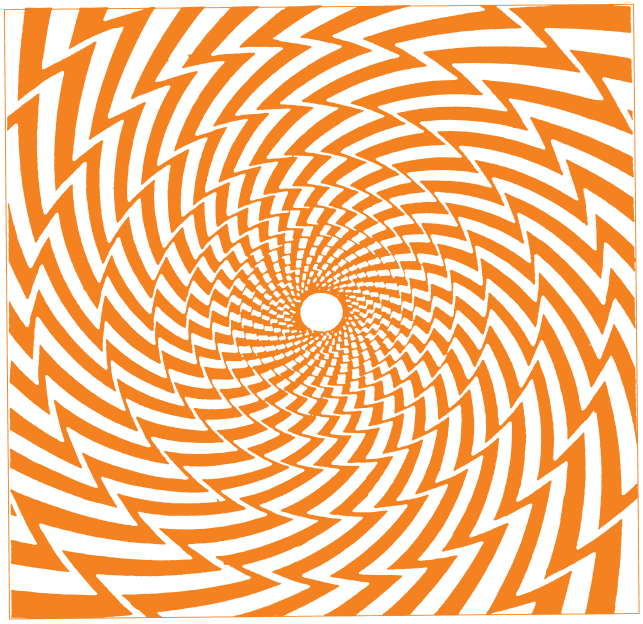
Ma è veramente questo la bibliometria o è, nel suo fondamento di analisi citazionale, una scienza avanzata dell'*information retrieval* suscettibile di rimodellare i fondamentali della biblioteconomia stessa, il suo modo di organizzazione delle risorse documentali e dunque anche la visione futura delle biblioteche?

Cominciamo col notare che essa è forse l'unica branca che, uscendo fuori dagli angusti confini disciplinari, ha esteso la sua influenza su scienze limitrofe alla biblioteconomia, come la sciento-metria e la gestione della conoscenza. L'Impact Factor, un approdo significativo, se non il più significativo della bibliometria, è infatti applicato alle riviste scientifiche, ma anche per valutare la validità della produzione delle comunità e dei settori accademici: ricercatori, discipline, università.

Nella sua ricerca De Bellis non si limita a esporre la vulgata bibliometrica dell'Impact Factor valutandone i pro e i contro (e più spesso i contro che i pro) come normalmente, e banalmente, fa la letteratura professionale italiana. Egli esplora con tecnica magistrale la dimensione metodologica della disciplina, documentandola nei suoi diversi "fondamenti" extra-bibliografici, ma sempre in costante tensione interdisciplinare.

Quali sono allora questi fondamenti? Uno dei primi approcci di base è la "sciento-metria" (o *naukometriya*, co-

me la chiamarono i teorici sovietici), una disciplina accademica nata con lo scopo di indirizzare "scientificamente" le politiche nazionali di ricerca e le risorse finanziarie verso le scoperte ritenute prioritarie dalla comunità accademica e dalla società.
In specie durante la guerra fredda, gli strumenti di misurazione di tali politiche nazionali hanno avuto forti connotazioni ideologiche. Con acutezza De Bellis mostra tuttavia come l'analisi citazionale di stampo anglosassone affondi le sue radici in una serie di approcci teorici piuttosto lontani dalla bibliografia, pur se ne condivide gli oggetti di indagine: libri, articoli, documenti. Un primo approccio, non del tutto scontato, è nella tradizione giuridico-prescrittiva dello *stare decisis*, il principio cioè di conformità a un giudizio precedente in cui si è soliti riconoscere uno dei caratteri fondamentali del sistema di *common law* nel diritto anglosassone. L'analisi citazionale fa proprio questo: stabilisce le regole del discorso scientifico (e/o le sue deviazioni) sulla base di un approccio cumulativo del sapere, in cui la legittimazione presente della ricerca si fonda sui risultati già acquisiti della letteratura scientifica passata.
Altro "fondamento" dell'analisi citazionale: l'*information retrieval*. È in questo ambito, infatti, che l'analisi citazionale muove i suoi primi passi, come tecnica di mediazione e di recupero dell'informazione complementare e, per certi versi, alternativa ai tradizionali sistemi di classificazione e di indicizzazione. Lo dimostra il successo degli algoritmi di Google come tecnica di recupero dell'informazione e il suo linguaggio di ricerca, tanto prossimo a quella che rimane la suprema ambizione di ogni linguaggio d'indicizzazione: la vicinanza con il linguaggio naturale.
L'analisi citazionale comincia ad essere applicata durante gli anni Cinquanta del secolo scorso in alcuni esperimenti bibliografici applicati alla genetica e alla brevettistica. Il suo creatore fu Eugene Garfield il quale, frustrato dall'approssimazione dei tradizionali sistemi di indicizzazione, aveva sviluppato una nuova idea di pertinenza (*relevance*) riferita alla rete dei riferimenti bibliografici. Comincia allora la strada maestra dell'analisi citazionale, la costruzione cioè di reti cognitive di articoli e di autori utile a misurare l'impatto di un lavoro individuale all'interno della comunità scientifica.
Il terzo fondamento della bibliometria – e non potrebbe essere altrimenti – è la matematica, e più precisamente la statistica quantitativa, che offre il quadro metodologico più adeguato e dal quale giungono gli apporti più produttivi. Apporti a dire il vero di tipo particolare, che descrivono i risultati dei processi comunicativi in termini più probabilistici che di predicibilità, per l'incapacità di questa disciplina di anticipare con certezza future occorrenze sulla base delle sue formule matematiche. De Bellis vi dedica il capitolo più cospicuo del libro, affrontando nel dettaglio le formule costitutive della bibliometria.
Una di esse è la "legge" di Lotka, di natura empirica, che si applica al tasso di produttività autoriale. Essa sostiene che il numero di autori che producono n lavori è pari a $1/n^2$ del numero di autori che produce solo un lavoro. Di matrice statistica è anche la legge di Bradford, che illustra la dispersione in circoli concentrici dei lavori su un determinato argomento all'interno delle riviste scientifiche: un primo cerchio è costituito da un nucleo essenziale di riviste; nel circolo immediatamente più ampio lo stesso ammontare di articoli è prodotto da un gruppo più numeroso di riviste e così via. Altra formula matematica è quella di Zipf, che descrive il numero di occorrenze delle parole nei testi di qualunque natura, giungendo alla conclusione che alcune di esse ricorrono con elevata frequenza, mentre assai esteso è il numero di parole che ricorre una sola volta. Tutte e tre le leggi presentano problematiche molto simili. Esse analizzano i rapporti di causalità dalle correlazioni grazie alle consuete tecniche regressive, ma la loro capacità di prevedere comportamenti possibili ricorrendo alle sequenze di dati è limitata. Fino a che punto, infatti, è possibile enunciare una correlazione tra consultazione e reputazione dell'autore o della rivista avendo come base il numero di referenze bibliografiche? Come trattare medie e varianze in bibliometria e come neutralizzare le fonti statistiche straordinarie o più produttive (quelle che in gergo sono denominate "code")? Come dare giustificazione matematica a quella regolarità di risultati, riassumibile nella formula detta paretiana (da Wilfredo Pareto, l'economista che l'ha applicata alla distribuzione della ricchezza), secondo cui l'80% della frequenza di consultazione riguarda il 20% delle riviste?
In termini filosofici la conseguenza del carattere "positivistico" di tali leggi è quella di giustificare una visione della scienza che tenderebbe a procedere in modo lineare, al punto che le nuove idee, per rivoluzionarie che siano, discendono sempre da un capitale precedente di conoscenze cumulate. Secondo De Bellis, tali leggi starebbero a dimostrare la fallacia della nozione di "paradigma" di Thomas Kuhn, per il quale invece i progressi della scienza procedono per salti e rivoluzioni e cancellano il carattere "normale" della teoria scientifica dominan-



te. Va detto, però, che il paradigma di Kuhn ha ben poco a che vedere con le tecniche bibliometriche, perché poggia su nozioni poco quantificabili, quali sono la conoscenza tacita e la negoziazione sociale.
Una delle tecniche più importanti di analisi citazionale è lo studio statistico delle co-occorrenze, vale a dire l'associazione di un lavoro ad un altro avente referenze bibliografiche simili (analisi degli accoppiamenti bibliografici) ovvero l'associazione di due lavori citati contemporaneamente da un terzo (analisi delle cocitazioni). C'è chi con l'accoppiamento bibliografico o citazionale ha tracciato la storia dei concetti di determinate discipline, costruendo cartografie fondate sulla copresenza di citazioni bibliografiche. C'è chi invece ha cercato di comprendere la storia e la dinamica dell'evoluzione scientifica. È vero però – lamenta De Bellis – che la matrice bibliometrica di natura scientifica, volta a documentare lo stato corrente e l'evoluzione delle discipline, si è oggi diluita nella logica commerciale delle classifiche mondiali delle università e dei ricercatori sul mercato globale dell'educazione.
E veniamo quindi all'Impact Factor, lo strumento, se non più importante, certamente più noto della bibliometria, cui De Bellis dedica l'intero capitolo sesto per una quantità di pagine pari a un quinto circa del volume. Normalmente una pubblicazione è valutata attraverso il *peer review*, effettuato a monte da un gruppo di esperti che giudicano non solo nel merito dei risultati della ricerca, ma anche sull'opportunità della loro comunicazione. Come è stato ampiamente dimostrato, il *peer review* ha alcune debolezze, che derivano non solo dai pregiudizi, positivi o negativi, dei certificatori, ma anche dall'impossibilità strutturale di emettere un parere universale e definitivo.
L'analisi citazionale, al contrario, offrirebbe il parametro di obiettività mancante. L'Impact Factor della rivista costituisce, a detta di De Bellis, una "scorciatoia", perché privilegia il momento quantitativo, applicando un algoritmo di calcolo così definito: dato un paniere di riviste accuratamente selezionate, al numeratore compare il numero di volte in cui una rivista è citata nei due anni precedenti; al denominatore è invece il numero di articoli pubblicati dalla rivista presa in considerazione in quegli stessi anni. Il denominatore serve da correttivo per bilanciare la dimensione fisica delle riviste prese in considerazione, i cui fascicoli non hanno un numero uniforme di articoli.
Per anni i bibliotecari hanno impiegato con successo l'Impact Factor come fonte di decisione per le acquisizioni bibliotecarie, malgrado le sue evidenti pecche e le indiscutibili distorsioni. I più importanti fattori di alterazione dei risultati sono l'autocitazione e il ritardo congenito (di almeno di due anni), che impedisce a nuove riviste, pur se valide, di entrare subito nel nucleo di quelle raccomandabili per l'acquisto. Altre manchevolezze sono la discutibile classificazione tra unità statistiche citabili (articoli) e unità non citabili (lettere, editoriali ecc.), la talvolta scarsa accuratezza del conteggio, con rischi di omissione o di plurimo conteggio delle citazioni, la densità delle referenze bibliografiche, differente da disciplina a disciplina e dipendente dal tempo, dal formato e dalla tipologia dell'articolo (una rassegna sullo stato dell'arte è di solito maggiormente citata di altri articoli). Depurato o neutralizzato da tali manchevolezze, l'Impact Factor ha però individuato negli anni, quasi sempre correttamente, il nucleo delle riviste più citate in determinate discipline.
Più discutibile è invece il suo uso da parte degli amministratori scientifici per valutare ricercatori singoli, gruppi di ricerca, dipartimenti, istituzioni e persino il progresso scientifico di alcuni paesi. Nel caso dei ricercatori, ad esempio, più valido è l'indice di Hirsch, dove si contano solo i lavori che hanno goduto di un certo numero di citazioni (il quale costituisce l'indice, appunto) mentre i lavori al di sotto di tale soglia non vengono presi in considerazione (a condizione di applicare un correttivo, proposto dallo stesso Hirsch, relativizzando l'indice al numero di anni di attività accademica).
Debole sul piano della valutazione dei ricercatori singoli, l'Impact Factor lo è ancora di più se applicato a organismi collettivi o ad aggregati di ricerca (come i gruppi scientifici, i dipartimenti ecc). In genere, le agenzie scientometriche preferiscono associare l'Impact Factor ad altri fattori qualitativi di natura istituzionale.
Un altro capitolo molto denso e lucido del volume di De Bellis è dedicato all'analisi citazionale sul web. A rigore, il web è una gigantesca macchina citazionale, in quanto non solo lega un documento all'altro, ma permette anche, a differenza della citazioni a stampa, il recupero del documento.
L'analisi citazionale applicata al web si è orientata verso due direzioni: la prima è la progettazione di sistemi d'indicizzazione automatici atti a catturare le fonti bibliografiche citate negli articoli, la seconda è l'applicazione di analisi statistiche alle strutture di hyperlink presenti nel web. Di conseguenza, la misurazione dell'impatto dei documenti può avvenire in vari modi: 1) contando il numero di volte in cui il documento è stato scaricato o visto; 2) inviando un questionario a un campione accuratamente selezionato di utilizzatori; 3) contando il numero di accessi al website dove è inserito il documento; 4) identificando le citazioni bibliografiche del documento, grazie alle tecniche di hyperlinking, anche se le citazioni stesse sono fuori dall'ambito ISI.
Le soluzioni 1) e 3), di tipo quantitativo, sono poco affidabili perché suscettibili di essere influenzate da fattori esogeni, come ad esempio la dimensione delle università o il vincolo di consultazione (ad esempio, una pubblicazione in open access raccomandata per lo svolgimento di un esame). La soluzione 2) utilizza l'apporto qualificato di esperti di riferimento, ma non tiene in conto il valore dinamico dell'analisi citazionale, il consenso "spontaneo" riscosso nella letteratura scientifica, al di fuori delle cappelle ufficiali e del paradigma corrente. La soluzione 4), l'analisi delle citazioni bibliografiche, registra meglio di altri metodi le condizioni d'influenza di un testo e per questo è stata estesa alle risorse in open access attraverso meccanismi quali Citebase e CiteSeer.
Citebase è un software che passa in rassegna le referenze bibliografiche dei lavori in full text contenuti in alcuni archivi ad open access e

segnala le corrispondenze tra referenze bibliografiche e i lavori citati presenti negli stessi archivi. Si tratta insomma di un indice delle citazioni, ma su scala ristretta, utile per identificare i lavori maggiormente citati o cocitati e estrarne le statistiche correlate. Citeseer è un sistema di gestione di una biblioteca digitale di maggiori ambizioni che sfrutta, oltre al conteggio delle citazioni, l'analisi statistica della copresenza di parole, delle cocitazioni e degli accoppiamenti bibliografici pescando i dati in una vasta mole di letteratura scientifica online (soprattutto di ambito informatico) disseminata in molteplici nodi della rete.

Con chiarezza De Bellis accenna alle tre direzioni di ricerca che contrassegnano la ricerca corrente sulla struttura del web: l'analisi delle reti complesse, lo studio degli hyperlink in rete e la webometria. L'analisi delle reti complesse mostra le relazioni esistenti all'interno di alcune reti (tecnologiche, sociali), le quali hanno aspetti topologici tipici, che non sono però né puramente regolari, né tipicamente casuali. Lo studio della rete di hyperlinks incrementa le *chances* per il ricercatore di estrarre l'informazione di cui ha bisogno dalla giungla digitale analizzando le connessioni tra siti web come simboli tecnologici di legami sociali tra individui, organizzazioni, nazioni. La webometria, infine, è la misurazione dei risultati ottenuti in seguito alle ricerche in rete.

Nel 1998 è stato proposto il WIF (web impact factor), esprimente la frazione tra il numero di pagine web legate a un certo sito e il numero di pagine web che lo stesso sito contiene. Sebbene la formula sia di limpida evidenza, essa si presta poco a conclusioni di carattere scientifico, per tre ragioni: a) la ricerca non copre il materiale "nascosto" nel web, come ad esempio i documenti tutelati dalla proprietà intellettuale, cui i *crawlers* non accedono; b) le ricerche condotte utilizzando diversi motori di ricerca portano a risultati differenti; c) si constata una marcata instabilità dei risultati dovuta alle fluttuazioni nel tempo dello statuto del materiale ospitato.

Insomma ogni analisi citazionale sul web, oltre ai tipici problemi di accuratezza, di copertura e di categorizzazione, mostra forti problemi a monte nella delimitazione del campo oggetto di indagine e nella raccolta di dati primitivi su cui fondare le deduzioni webometriche.

Vale la pena inoltre sottolineare che il volume di De Bellis non si ferma alla stretta analisi tecnica, ma riesce anche a tracciare il quadro teorico di filosofia della scienza entro cui si situano le indagini bibliometriche. L'analisi citazionale si sviluppa al tempo in cui Derek Price dava alle stampe *Little science, big science*, in cui si descriveva il modo in cui i grandi scienziati erano impiegati e come maturavano le idee più produttive. Sempre a quel tempo Robert Merton identificava l'"effetto San Matteo" e la fenomenologia del "successo che alimenta il successo", secondo cui alcuni autori godono di una visibilità forse sovradimensionata in rapporto al loro valore, procurata attraverso il numero di citazioni. Parallelamente Thomas Kuhn, nel suo *The structure of scientific revolutions*, introduceva i concetti di "paradigma" e di "scienza normale" e studiava i cicli periodici dell'emergere e del tramontare delle dottrine scientifiche e l'anatomia del "crollo", quando determinati paradigmi lasciano il posto ad altri.

I tre autori erano accomunati dall'ambizione di analizzare le logiche interne della scienza e, segno o sintomo dell'eccellenza di una ricerca, la citazione diventava, almeno nel caso di Price e Merton, lo strumento per antonomasia di misurazione e di identificazione di tali regole, leggi o cicli.

Non è questa la sola grande questione teorica adombrata da De Bellis. Ne citiamo alcune altre. Il progresso della scienza è costituito dal "democratico" sviluppo dovuto all'opera di una massa oscura di lavoratori della conoscenza (effetto Ortega, da Ortega y Gasset, il filosofo spagnolo che l'ha proposto) o è invece il risultato di una stratificazione sociale del sistema della scienza con una *élite* che, con conoscenze, attività e mezzi, orienta le direzioni maestre della ricerca scientifica? Le tecniche bibliometriche riescono a rendere conto di trasferimenti di conoscenza poco visibili (ad esempio, l'influenza della conoscenza "tacita")? Quanto studiata è la componente motivazionale, quanto grande è il fattore "emotivo" nella scelta delle citazioni? E l'autore scientifico, che, a differenza del letterario, è spesso un autore collettivo, non ha bisogno anch'egli di un ulteriore algoritmo perché siano valutati con precisione le tipologie di attività condotte dai singoli autori, o il contributo percentuale di ciascuno di essi alla produzione del lavoro, o il diverso credito ottenuto da ogni singolo autore?

Abbiamo menzionato solo alcune delle questioni teoriche che De Bellis affronta con dottrina e spigliatezza nel suo volume, che ha peraltro una storia singolare. *Bibliometrics and citation analysis* è, infatti, all'origine una tesi di dottorato sulla storia delle enciclopedie scientifiche rinascimentali. Vince poi il Premio "Biblioteche oggi" nel 2005. Approda, infine, agli onori della pubblicazione in una prestigiosa casa editrice internazionale. Un *cursus honorum* ampiamente meritato per questa pubblicazione elegantemente pensata e densamente documentata, da raccomandare a tutti i lavoratori dell'informazione che intendono dare una scossa alle proprie certezze teoriche. Inoltre, malgrado la materia non proprio romanzesca, l'inglese di De Bellis è avvincente e riesce persino a far trasparire in filigrana l'originale in italiano, per cui ci sentiremmo di consigliare il volume anche a quei bibliotecari desiderosi di migliorare le proprie abilità linguistiche professionali.

Giuseppe Vitiello

Nato Defense College, Roma
g.vitiello@ndc.nato.int