Come farsi trovare sul Web

di Giuseppe Sturiale

Ingegnere, esperto di NLP, Natural Language Processing, e su tematiche di ingegneria cognitiva.

L'arte di galleggiare nel mare del web consiste nel capire quali sono i criteri usati dai motori di ricerca per determinare il *ranking* e sfruttarle. Il ranking è il punteggio assegnato ad ognuna delle pagine trovate e serve a stabilire il loro ordine di presentazione, in modo che le prime della lista siano le più utili. Utili a chi legge e utili a chi scrive, ma anche a chi pubblica. I soggetti che devono mettere in piedi il win-win-win sono dunque tre, ma i criteri e gli scopi del calcolo del ranking sono molti. Ad essere meticolosi, alla fine di queste considerazioni si potrebbe creare una tabella con tre colonne (i tre soggetti) e diverse righe (le tecniche utili) e stabilire che cosa si deve fare secondo il proprio ruolo. Queste considerazioni sono dedicate a chi scrive e ambisce a farsi trovare e leggere dal numero più alto possibile di lettori, ma sono utili anche quando si cerca. Chi indicizza e pubblica le conosce già perché questo è il suo mestiere e Google è di fatto il sistema di riferimento che detiene oltre il 90% del marcato globale.

Alle origini lo scopo del calcolo del ranking era la misura della pertinenza teorica di ogni pagina trovata rispetto all’argomento della ricerca dell'utente, confrontata con quella di tante altre pagine. Con le pagine più promettenti in testa alla lista di quelle trovate, la ricerca sarebbe stata veloce ed efficace. A cavallo fra gli anni ’80 e ’90 del secolo scorso lavoravo al miglioramento continuo del motore di ricerca Fulcrum, in un certo senso il papà di Google, un

motore di indicizzazione e di ricerca per vaste collezioni di documenti usato all'interno di grandi organizzazioni pubbliche e private. Ecco un primo aspetto da non trascurare: per quanto cresca la velocità dei computer moderni, il numero di documenti cresce ancora di più, quindi è bene che il motore di ricerca possa usare indici creati dal suo gemello, il motore di indicizzazione, meno visibile ma importantissimo, che legge instancabilmente tutto il web e aggiorna gli indici.

All'inizio dell'era dei motori di indicizzazione e ricerca il soggetto era solo uno, il ricercatore delle informazioni, e chi raccoglieva e indicizzava i documenti lavorava per lui e aveva i suoi stessi obiettivi. Non c'erano quindi conflitti da temperare. In questa situazione il ranking era ben calcolato se si teneva alto il richiamo (recall), cioè la capacità del sistema di non perdere documenti rilevanti, e di ridurre il rumore (noise), cioè il ritrovamento erroneo di documenti non pertinenti.

Siccome il motore di indicizzazione non capisce quello che legge e non sa creare abstract automatici che migliorino l'indicizzazione, il buon funzionamento del sistema è legato alla sua capacità di isolare le parole del documento e di creare un dizionario delle parole trovate dotato di rimandi alle liste dei documenti che le contengono, in modo che il simmetrico motore di ricerca possa creare la lista di documenti rilavanti a fronte di un insieme di parole di interesse. Il primo concetto importante che risulta da questo scenario è che chi cerca dovrebbe anche usare sinonimi alternativi, chi prepara i documenti per l'indicizzazione dovrebbe creare un tesauro di sinonimi destinato ad aiutare il motore di ricerca ad arricchire automaticamente la

richiesta, e chi scrive dovrebbe aiutare il motore di indicizzazione aggiungendo sezioni supplementari al documento con parole aggiuntive di soggettazione, che descrivano il documento e creino agganci potenziali per le future ricerche, anche se il suo testo non le contiene.

Questi elementi sul funzionamento dell'indicizzazione e della ricerca permettono di ricavare le prime regole importanti. Siccome anche il motore di ricerca non capisce quello che stiamo cercando ma si basa solo sulle parole, occorre aiutarlo, e dato che non sappiamo quanto è buono il tesauro dei sinonimi scritto dagli esperti che pubblicano, occorre usare diversi sinonimi insieme alle parole chiave che stiamo usando per orientare la ricerca. La mancata comprensione di grammatica e sintassi rende inutile (e spesso dannoso) un costrutto grammaticale o sintattico: "Vorrei sapere quali sono i diametri più diffusi per i dischi", una richiesta che produrrà molto rumore. Meglio scrivere "disco diametro centimetri cm millimetri mm", che troverà le pagine dei cataloghi che rispondono alla nostra domanda.

Morale: Chi cerca deve rimediare a quello che è stato trascurato da chi scrive (o indicizza) usando parole chiave alterative. Chi scrive gli articoli, se vuole essere trovato, deve prevedere gli errori e le dimenticanze di chi cerca, arricchendo il documento.

Resta da discutere il metodo di generazione del ranking e come alzarlo se siamo quelli che scrivono e che vogliono essere trovati.

Il criterio di base per un sistema di ricerca documentale interno ad una organizzazione (un solo soggetto, il ricercatore) era di generare il ranking, cioè la rilevanza di

un documento, sommando i numeri delle occorrenze presenti di ognuna delle parole cercate, moltiplicati per un indice di rarità della parola (misurato dal dizionario con la tavola delle occorrenze). Quindi a parità di parole presenti quelle che contribuiscono di più al peso sono quelle più rare.

Quando si usa un motore commerciale come Google, ci si trova a far parte di un sistema dove si distinguono e si temperano gli interessi dei tre soggetti (autore, ricercatore e intermediario ovvero fornitore del servizio) e il motore di ricerca aggiunge ai criteri di un sistema scientifico usato all'interno di una organizzazione una serie di altri fattori che contribuiscono al ranking, legati al gradimento della pagina mostrato dagli utenti (di cui è tenuta traccia) e alla reputazione acquisita dal sito che la ospita. Tutte queste considerazioni formano ormai una scienza chiamata *SEO*, cioè *Search Engine Optimization*, che guida la redazione del contenuto e determina la scelta delle parole da usare nel testo o da aggiungere nella sezione del documento che non è visibile se non al motore di indicizzazione. I video, per esempio, sono valutati da Youtube sulla percentuale realmente vista dagli utenti (minuti visti, cioè momento dell'interruzione rispetto alla durata totale del video) e i suoi *analytics* forniscono in una curva solo discendente le statistiche del livello di attenzione ad ogni minuto e secondo del video. Dopo averlo selezionato, tutti iniziano a vederlo, ma nella sua durata molti vanno interrompendo la visione. Quando si va a cercare la causa di un brusco gradino discendente nel grafico a segnalare una forte caduta di interesse a un certo minuto e secondo, si va poi a controllare e si scopre che c'è un colpo di tosse o una

affermazione un po’ troppo divisiva o altri problemi che hanno fatto “cambiare canale” all’ascoltatore. Per il momento è ancora impossibile capire quando invece il gradimento sale e quindi poter avere una curva che va su e giù.

Non si devono infine dimenticare le cruciali esigenze dell’editore, che per offrire il servizio gratuitamente deve guadagnare non solo sulla pubblicità che accompagna le liste o i documenti trovati ma anche sull’interesse pagante dell’autore (o del titolare del sito) a farsi leggere. Quindi un addendo che può alzare in modo decisivo il ranking di un articolo, oltre al suo valore scientifico e alla reputazione del sito ospite, è il livello di sostegno economico alla sua visibilità, cioè il denaro pagato al *publisher* per alzare il ranking ancora di più rispetto al valore già raggiunto con la pertinenza sulla richiesta, il valore scientifico e la reputazione del sito ospite. Questo è il motivo per cui i contenuti ad alta visibilità (cioè i primi della lista o quelli che stanno ancora più sopra della lista) in un motore di ricerca commerciale somigliano a quelli della televisione commerciale, cioè sono prodotti che risultano da investimenti più orientati al sostegno pubblicitario che alla qualità.

Inoltre, ogni contenuto che si pubblica ha una impennata di letture dovuta al peso della novità (un'altra componete del ranking è ricavata dalla data di pubblicazione, per cui una pagina nuova sale nel ranking rispetto a una vecchia). Altre impennate possono dipendere da motivi di attualità, che determinano un rinnovato interesse del pubblico per un certo argomento, ma questa componente pesa meno rispetto alla novità, quindi l’esortazione è quella a pubblicare

continuamente, che è poi il vecchio *publish or perish* del mondo accademico, un altro contesto dove si parla di ranking.

L'impennata iniziale negli analytics resi disponibili dal service provider responsabile del sito indica il richiamo iniziale di lettori e poi il successivo stabilizzarsi ad un valore asintotico che indica il vero pubblico degli specialisti interessati ai contenuti pubblicati. Il service provider che offre l'hosting del nostro dominio è in realtà un quarto soggetto che potrebbe avere servizi rilevanti da offrire o sulle politiche del quale poter fare leva, ma non è un elemento chiave per le nostre riflessioni.

Quindi occorrono eventi che suscitino altre impennate iniziali per poi guadagnare nuovi lettori, ma servono nuovi argomenti o un nuovo taglio o una revisione editoriale per cambiare il ranking di sito.

Di tutti gli stratagemmi descritti, quello importante, che non costa niente e che offre eccellenti risultati consiste nel farsi carico di quello che un bravo indicizzatore farebbe all'interno di un sistema privato, cioè scrivere un documento capace di farsi trovare, con un ricco abstract che contenga tutte le parole che potrebbero offrire agganci verso i lettori. Meglio usare molte parole specifiche e selettive (rare), anche solo una delle quali può alzarci di molto il *recall* che poche generali e potenzialmente ambigue e quindi fonte di rumore.

La scelta di queste keyword è fondamentale. Fra i tanti sinonimi conviene opportunisticamente scegliere anche quelli più alla moda e soprattutto, nello stile redazionale, tornare ad esprimere lo stesso concetto, o rifinirlo, usando

altre parole che potrebbero "agganciare" altri lettori che usano il motore di ricerca. Questa è davvero la regola numero uno. Lo stile nel condurre questa attività è in tutto parallelo al codice etico che ci si prefigge, indipendentemente dall'averlo esplicitato. Siti di prodotti commerciali non si fanno scrupoli nell'approfittare della non visibilità delle sezioni per le keyword inserendovi i nomi di prodotti e aziende concorrenti in modo da dirottare su di sé i clienti di altri produttori. È un modo per ottenere intenzionalmente il controllo del rumore, quando la pagina è trovata ma l'utente non capisce perché.