Indicizzazione semantica in pillole

Il criterio della 'distanza semantica' nell'analisi concettuale dei documenti

Alberto Cheti

L'espressione 'distanza semantica' evoca uno spazio di significati (parole e concetti) tra loro collegati e separati da distanze che si presume possano dipendere dalle loro relazioni paradigmatiche (come le relazioni fondamentali di un thesaurus) e/o sintagmatiche (come quelle tra le parole in un testo). Nell'uso che se ne fa qui, lo spazio è il testo del documento, i significati sono i temi di cui tratta, la distanza si riferisce al loro grado di vicinanza/lontananza sotto l'aspetto semantico. Gli effetti di questa distanza sono considerati in termini di probabilità che un tema possa essere inferito a partire da un altro tema, con conseguenti ripercussioni sulla loro selezione ai fini dell'indicizzazione per soggetto.

Si tratta, dunque, di un'applicazione della nozione di 'distanza semantica' nel contesto dell'analisi concettuale dei documenti. In generale, questo criterio può concorrere all'obiettivo di affinare i metodi e le procedure relative alla prima fase del processo di indicizzazione, la cui importanza è sottolineata da Maria Teresa Biagetti nella parte conclusiva di un articolo sugli sviluppi delle biblioteche digitali:

> l'analisi concettuale dei documenti realizzata da esseri umani, ma impostata su metodologie di indicizzazione semantica più raffinate di quelle ordinariamente adottate nelle biblioteche tradizionali, deve costituire l'obiettivo principale, allo scopo di consentire la realizzazione di metadati semanticamente più ricchi e quindi elevare il livello delle funzionalità della ricerca nelle biblioteche digitali. […] Le diverse proprietà mostrate da un documento, oggettivamente rilevabili, possono avere significati diversi in ambiti disciplinari diversi e all'interno di finalità scientifiche diverse. […] L'analisi semantica dei documenti dovrebbe quindi considerare un ventaglio oggettivamente sostenibile di possibili soggetti facendo emergere una pluralità di argomenti, ritagliati sulle diverse necessità informazionali delle diverse utenze cui i documenti si rivolgono[1].

Il criterio della 'distanza semantica', come formulato qui, fa parte di un insieme di criteri di selezione dei concetti – indicativo, non certo esaustivo, – che si è andato via via formando

[1] Maria Teresa Biagetti, *Sviluppi e trasformazioni delle biblioteche digitali: dai* repositories *di testi alle* semantic digital libraries, «AIB studi», 54 (2014), n. 1, p. 11-34 (p. 31-32).

durante la sperimentazione relativa all'indicizzazione per soggetto di opere antiche, tutt'ora in corso, condotta dalla Biblioteca nazionale di Firenze e dalla Biblioteca dell'Accademia della Crusca. La specificazione di questi criteri – che troverà posto all'interno di linee guida sulla soggettazione di opere antiche, da sottoporre alla comunità bibliotecaria – intende, appunto, aiutare l'indicizzatore nella verifica dei possibili fattori di rilevanza dei concetti identificati nel corso dell'esame del documento, così da non trascurare temi potenzialmente utili per gli utenti. A questo proposito, non è un caso che una delle prime chiarificazioni, nel gruppo di lavoro che si sta occupando delle linee guida, abbia riguardato la distinzione tra due nozioni tradizionalmente presenti nell'indicizzazione semantica: 'coestensione' ed 'esaustività'. Le due misure, infatti, si applicano, direttamente, a oggetti diversi: la prima al contenuto concettuale di un tema, la seconda all'intero contenuto concettuale del documento. È al primo, ossia al contenuto concettuale di un tema, che si riferisce la 'stringa unica coestesa', raccomandata dal GRIS[2] e nel *Nuovo soggettario*[3]: una stringa che abbia la stessa estensione semantica del tema individuato attraverso l'analisi; sebbene, quando si tratta del 'tema di base', essa ci dia anche una comprensione globale del contenuto del documento. Dunque, se è preferibile esprimere uno stesso tema con un'unica stringa di soggetto, così da accrescere il 'grado di precisione' nel recupero, nulla impedisce di selezionare più temi e, dunque, di costruire più stringhe di soggetto per uno stesso documento, così da accrescere il 'grado di richiamo'.

I fattori che influenzano il giudizio di rilevanza possono essere qualificati come 'testuali' o 'contestuali'. Gli uni si basano sull'analisi delle relazioni tra i temi all'interno della struttura tematica del testo, allo scopo di individuare il tema principale ('tema di base') e i temi secondari ('temi particolari'). Gli altri sull'analisi del contesto (intertestuale, storico, epistemologico, socio-culturale) di produzione/ricezione del documento, allo scopo di individuare i temi che, anche indipendentemente dal loro ruolo all'interno dell'organizzazione tematica del testo, risultano rilevanti rispetto a determinati interessi informativi o alla luce di particolari prospettive scientifiche, disciplinari, ecc.

Il criterio della 'distanza semantica' appartiene al primo tipo di fattori e può essere enunciato in questo modo: un documento, avente come soggetto un tema generale, può contenere delle parti – opere autonome o parti distinte di una stessa opera – in cui sono trattati con un certo rilievo anche singoli temi specifici, i quali, pur essendo inclusi nel tema generale, necessitano per la loro distanza semantica da esso di una segnalazione autonoma, che ne favorisca il recupero. La

[2] Associazione italiana biblioteche. GRIS – Gruppo di ricerca sull'indicizzazione per soggetto, *Guida all'indicizzazione per soggetto*. Roma: AIB, 1996 (ristampa con correzioni 2001), in part. p.10-11.
[3] Biblioteca nazionale centrale di Firenze, *Nuovo soggettario: guida al sistema italiano di indicizzazione per soggetto. Prototipo del thesaurus*. Milano: Editrice Bibliografica, 2006, in part. p. 39-40 e p. 102.

questione, dunque, è se e in quali circostanze è opportuno selezionare e segnalare, in aggiunta a un tema generale, un tema specifico in esso compreso, a motivo della distanza semantica tra i due.

Prima di precisare queste circostanze, si dà un cenno sui significati della nozione di 'distanza semantica', che ricorre in vari ambiti: linguistica computazionale, information retrieval, semantica dei motori di ricerca, ontologie, ecc. Due, in particolare, le accezioni:

a) distanza lessicale o concettuale, quale si verifica tra i nodi di una rete semantica, sulla base di determinate relazioni semantiche fondamentali (p.e., gerarchiche);

b) distanza fisica o distribuzionale, quale si verifica tra le parole che co-occorrono nei testi, sulla base dell'ipotesi che le parole fisicamente vicine tra loro in un testo abbiano una relazione più stretta rispetto a quelle più lontane e, dunque, che la similarità tra due parole dipenda dalla frequenza con cui ricorrono in contesti simili.

Per esempio, nel caso a), i seguenti enunciati mostrano una distanza semantica crescente tra la 'specie' e il 'genere':

- i canarini sono fringillidi
- i canarini sono passeriformi
- i canarini sono uccelli
- i canarini sono animali

Nel caso b), la frequenza con cui in un corpus di documenti la parola 'presidente' ricorre nel contesto di 'repubblica', rispetto a quella di parole come 'torta' o 'panino', determina la misura della loro similarità sul piano semantico[4].

Questa duplice valenza si ritrova anche nel criterio della 'distanza semantica': si tratta della distanza concettuale tra un tema specifico e uno più generale che lo comprende, come pure di una distanza fisica, essendo il tema specifico trattato in parti distinte di una stessa opera o in opere autonome contenute nello stesso documento. Tuttavia, diverso è l'oggetto dell'analisi (non singoli concetti, né singole parole, ma temi trattati in un documento), come pure il contesto (non l'analisi del loro significato, ma della loro rilevanza in un documento ai fini del recupero).

[4] L'esempio è tratto da Alessandro Lenci, *Spazi di parole: metafore e rappresentazioni semantiche*, «Paradigmi», 2009, p. 83-100. Dello stesso autore cfr. anche *Semantica distribuzionale: un modello computazionale del significato*. In: *Compter parler soigner: tra linguistica e intelligenza artificiale*, a cura di Edoardo Maria Ponti, Marco Budassi. Pavia: Pavia University Press, 2016, p. 39-53. I due testi forniscono una chiara esposizione dell'approccio distribuzionale al significato delle parole, comprensibile anche ai non specialisti.

Tornando al dunque, la distanza semantica di un tema da quello più ampio di cui fa parte implica due condizioni concomitanti: a) il grado di specificità del tema, ossia il numero di temi intermedi che lo separano dal tema più ampio, supponendo i temi organizzati in una gerarchia che va dal più generico al più specifico; b) la sua unicità o singolarità, ossia l'essere trattato nel documento come tema singolo, in assenza di co-occorrenze relative a temi simili, tutti riconducibili a un tema minimo comune che li comprenda. Per esempio, 'coltivazione dei cardi' ha una distanza da 'agricoltura' maggiore rispetto a 'orticoltura', che rappresenta un tema intermedio tra il più specifico ('coltivazione dei cardi') e il più ampio ('agricoltura'). Tuttavia, se il documento che tratta della coltivazione dei cardi riguardasse anche la coltivazione di altri tipi di ortaggi, i singoli temi specifici sarebbero sussunti sotto il tema comune 'orticoltura', riducendosi così la loro distanza da 'agricoltura'. La misura della distanza semantica è, dunque, direttamente proporzionale all'ampiezza del tema generale rispetto al tema specifico, o al 'dislivello' tra i due, e inversamente proporzionale al numero/importanza di co-occorrenze di temi specifici simili, ossia con analogo grado di specificità, riconducibili a un tema comune più ampio. La distanza semantica minima si verifica quando i temi specifici trattati nel documento costituiscono una parte consistente/importante – se non la maggior parte – dei temi attinenti al tema più ampio.

La 'distanza semantica', così intesa, può influenzare la selezione dei concetti, ai fini dell'indicizzazione. Infatti, quanto maggiore è la distanza tanto minore è il grado di implicazione del tema specifico nel tema generale e, dunque, la probabilità che possa essere inferito a partire da quest'ultimo. Perciò, una volta accertato il rilievo e l'interesse dei singoli temi specifici trattati, può essere opportuno indicizzare anche ciascuno di essi, creando altrettante stringhe di soggetto.

Per esempio, l'opera *Ragionamenti del dottor Giovanni Targioni Tozzetti sull'agricoltura toscana*[5] è una raccolta di opere – perizie, consulenze, memorie, preparate in occasioni diverse – «concernenti l'agricoltura». Alcune di esse affrontano aspetti generali, come lo studio dell'agricoltura, le caratteristiche dei terreni agrari, il modo di lavorarli, le colture più idonee a ciascuno di essi ecc. In altre, sono trattati temi molto specifici rispetto al tema generale 'agricoltura', sebbene con esso coerenti, quali 'coltivazione dei cardi' e 'produzione della resina del lentisco in Maremma', che nell'intenzione dell'autore «potrebbero riuscire utili ad alcuno». La cancellazione di quest'ultimi, in quanto sussunti nell'ambito del tema più ampio 'agricoltura', potrebbe pregiudicare la possibilità, per chi vi fosse interessato, di reperire quest'opera a partire dalla sola intestazione di soggetto **Agricoltura**.

[5] Giovanni Targioni Tozzetti, *Ragionamenti del dottor Giovanni Targioni Tozzetti sull'agricoltura toscana*. In Lucca: nella stamperia di Jacopo Giusti alla colonna del Palio, 1759.

Tuttavia, l'indicizzatore dovrebbe applicare con cautela il criterio della 'distanza semantica', verificando sempre prima la possibilità di raggruppare i singoli temi specifici trattati nel documento sotto un tema più ampio che li comprenda, secondo il criterio generale della 'sommarizzazione specifica', in base al quale si seleziona, come soggetto, il tema inclusivo più specifico ('tema di base'), essendo i temi inclusi ('temi particolari') desumibili da esso con ragionevole probabilità.

Al criterio della distanza semantica si può ricorrere anche quando nel documento sono trattate più entità individuali riconducibili a una classe comune. Per esempio, per scegliere se indicizzare un'opera biografica con il nome delle singole persone o con quello della classe cui sono attribuibili, si può considerare la distanza semantica delle prime dalla seconda, tenendo conto del numero di soggetti biografici e dell'estensione della classe: quanto maggiore è l'uno e minore l'altra, tanto più piccola è la distanza semantica tra ciascun individuo e la classe di appartenenza, cosicché quest'ultima può essere considerata il soggetto dell'opera; viceversa, saranno i singoli individui a prevalere. Quando la classe e i singoli individui si bilanciano, ossia hanno lo stesso peso – la classe è sufficientemente specifica da costituire un probabile accesso ai singoli individui; questi, a loro volta, sono in numero ridotto da valere per sé, indipendentemente dalla classe, – possono essere indicizzati entrambi, la classe e ciascun individuo.

Per esempio, sempre in riferimento a opere antiche, l'opera *Cominciamento e progresso dell'arte dell'intagliare in rame colle vite di molti de' più eccellenti maestri della stessa professione* di Filippo Baldinucci[6] raccoglie diciotto biografie di altrettanti incisori operanti tra il 15. e il 17. secolo. L'enunciato di soggetto 'biografie di incisori, sec. 15.-17.' definisce una classe non troppo larga («maestri della stessa professione» in un determinato arco di tempo); d'altra parte, il numero di soggetti biografici presenti nell'opera costituisce un insieme non troppo ristretto («molti de' più eccellenti maestri della stessa professione»): entrambe le condizioni sono favorevoli alla scelta di indicizzare l'opera al nome della classe e non a quello dei singoli individui.

Del resto, il criterio della 'distanza semantica' è implicito anche nella scelta di stabilire un limite numerico prefissato (p.e., fino a cinque) oltre il quale indicizzare l'opera alla classe invece che ai singoli individui, ferma restando l'esigenza di accertarsi, in questo caso, che il nome della classe e le eventuali altre specificazioni presenti nel soggetto (p.e., luogo e tempo) ne restringano il più possibile l'estensione.

[6] Filippo Baldinucci, *Cominciamento e progresso dell'arte dell'intagliare in rame colle vite di molti de' più eccellenti maestri della stessa professione* […]. 2. ed., accresciuta di annotazioni del sig. Domenico Maria Manni. In Firenze: per Gio. Batista Stecchi e Anton-Giuseppe Pagani, 1767.

Il contesto nel quale si è manifestata l'occasione per formulare il criterio della 'distanza semantica' e gli esempi citati sopra fanno riferimento a opere antiche, ma ovviamente tale criterio è applicabile anche a quelle moderne, sebbene sia possibile che le condizioni che danno luogo alla sua applicazione si verifichino più frequentemente nelle prime.