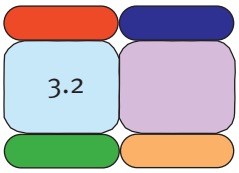
ESTRAZIONE AUTOMATICA D'INFORMAZIONE DAI TESTI

Sergio Bolasco
Bruno Bisceglia
Francesco Baiocchi



L'analisi automatica dei testi si è sviluppata soprattutto in virtù della crescente disponibilità di tecnologie informatiche e linguistiche e consente di dare una rappresentazione dei testi estraendone alcune proprietà essenziali, capaci di descrivere e interpretare il loro contenuto. Nell'ambito di aziende e istituzioni, è diventato dunque prioritario far fronte alla massa di materiali testuali da gestire quotidianamente, estraendo solo l'informazione d'interesse.

1.INTRODUZIONE

Con il termine *Text Analysis* (TA) s'intende un'analisi del testo "mediata" dal computer, ossia basata non sulla lettura del testo, bensì su un'analisi automatica, utile soprattutto quando i testi sono di ampia dimensione[1]. In questi casi, infatti, ogni lettura diretta sarebbe limitata, lunga e difficoltosa, mentre un'analisi automatica è veloce e aperta a "infiniti" confronti, resi possibili dall'uso del computer. Questo approccio ha come obiettivo, fra gli altri, quello di fornire alcune rappresentazioni del contenuto della collezione di testi oggetto di studio (*corpus*) e di estrarre da questi una *informazione*, ossia alcune proprietà, attraverso misurazioni di tipo quantitativo. In una logica di tipo statistico, in questo ambito, si parla anche di "analisi dei dati testuali" (ADT), sottolineando la possibilità di ricavare informazioni strettamente qualitative, a partire da risultanze quantitative, quali sono quelle tipiche della statistica.

L'evoluzione storica degli studi quantitativi su dati espressi in linguaggio naturale, ha visto a partire dagli anni Settanta forti cambiamenti strettamente legati all'evoluzione dell'informatica e alla crescente disponibilità di risorse linguistiche [30] e più recentemente all'enorme dimensione dei testi da consultare *on-line*. Nel corso degli anni, l'interesse per gli studi quantitativi della lingua[2] si è spostato da una logica di tipo *lin-*

[1] È opportuno distinguere la Text Analysis dall'*analisi testuale*, poiché con questa espressione si indica generalmente quella ampia area di ricerca che ha le sue radici in analisi non automatiche, basate su una lettura a più riprese del testo tendente a categorizzare brani, a studiare l'accezione dei termini fino a sviluppare, in taluni casi, un'analisi semiotica d'interpretazione del testo.

[2] Contributi significativi si trovano in riviste quali, fra le altre, Cahiers de Lexicologie, Computers and Humanities, ACM Computing Surveys, Journal of Quantitative Linguistics, Linguisticae Investigationes, Literary and Linguistic Computing, Mots, TAL.

guistico[3] (sviluppata fino agli anni Sessanta) a una di tipo *lessicale*[4] (intorno agli anni Settanta), per approdare negli anni Ottanta e Novanta ad analisi di tipo *testuale*[5] o ancor meglio *lessico-testuale*[6].
Fin dall'inizio le applicazioni hanno interessato tutti i campi disciplinari: dai linguisti specialisti negli studi stilometrici o di autenticità dell'autore agli psicologi e antropologi interessati alle analisi di contenuto sia su testi che su materiali provenienti da indagini sul campo (interviste, storie di vita, *focus group*), dai sociologi che si occupano di discorso politico o di indagini qualitative agli specialisti di comunicazione orientati al marketing e al linguaggio veicolato sui principali tipi di media.
In questi ultimi anni, suscita molto interesse, nel filone statistico dell'analisi dei dati testuali, un ulteriore approccio noto con il termine di *Text Mining* (TM): esso è tipico di applicazioni indirizzate alle aziende e istituzioni, le quali, dovendo interagire con enormi masse di materiali testuali spesso disponibili in rete, hanno il problema di selezionare, all'interno di queste fonti smisurate, i dati di loro interesse, per estrarne *informazione* capace di produrre valore. Si tratta di soluzioni orientate al *Knowledge Management* (KM) e alla *Business Intelligence* (BI) che, nella gran parte dei casi, consistono nel ricavare da testi non strutturati quei dati essenziali utili ad alimentare i *database* operativi aziendali con informazioni strutturate, più facili da gestire nei processi decisionali a fini strategici.
Nel seguito, verranno ripresi punti di vista, aspetti teorici ed esempi di applicazione dei vari approcci qui accennati. In questa prospettiva, è opportuno richiamare preliminarmente alcuni concetti di base e relative definizioni dei principali oggetti, visti come utensili da tenere nella "cassetta degli attrezzi" per estrarre informazione dai testi.

2. CONCETTI E DEFINIZIONI

La prima lettura automatica dei testi oggetto di studio da parte del computer comporta la cosiddetta *numerizzazione* del corpus: operazione che, a ogni *forma* o **parola** diversa che appare nel testo, fa corrispondere un codice numerico e l'elenco delle collocazioni di tutte le sue *occorrenze* (*token*) nel corpus (Tabella 1), ossia delle loro posizioni lungo lo sviluppo del testo (*discorso*).
Il risultato di questa fase si traduce nella costruzione della lista (*indice*) di tutte le parole diverse che figurano nel testo, il cosiddetto *vocabolario* del corpus, espresso in *forme*

Il termine ***parola*** non trova una definizione soddisfacente: le parole sono gli oggetti linguistici che costituiscono il lessico e sono raccolti nel dizionario. La parola è un segno che ha un senso, è un segno che è simbolo di un concetto o almeno espressione di una conoscenza; si tratta di adottare delle convenzioni precise e il più possibile corrette dal punto di vista linguistico, ma comunque in parte arbitrarie. Una parola può denotare: un oggetto (sostantivo), un'azione o uno stato (verbo), una qualità (aggettivo, avverbio), una relazione (preposizione). Qui nel seguito, per semplicità, si indica con il termine *parola* l'unità di analisi del testo, qualunque essa sia. Va osservato che a seconda degli obiettivi dell'analisi questa unità lessicale può essere una forma grafica, un lemma, un poliforme o una "forma testuale", ossia un'unità di tipo misto in grado di catturare al meglio i contenuti presenti nel testo. In realtà, ormai si sceglie sempre come unità d'analisi del testo una mistura di tipi; quindi nell'articolo con *parola* s'intenda in generale una *forma testuale*.

[3] Per cogliere i rapporti fra lingua e sue concrete possibilità d'analisi (Guiraud, Herdan) [18, 19], si potrebbe seguire un'immagine di Tournier [29], sintetizzata in Bolasco [6]. La dimensione illimitata della lingua fa sì che non sia possibile, per definizione, associare alle parole una qualche "frequenza" in senso statistico-probabilistico. Quest'ultima è invece misurabile su una raccolta di testi, intesi come spezzoni di lessici, ovvero come "campioni" particolari della lingua. È così che ci si limita a considerare le occorrenze delle parole in un testo come un'approssimazione delle frequenze in un lessico, a patto che il corpus sia sufficientemente ampio (almeno 50.000 occorrenze).

[4] Cfr. per esempio Muller [24] e Brunet [8].

[5] In questo approccio l'attenzione sulla testualità del contenuto privilegia l'analisi statistica in forme grafiche (cfr. Lebart *et al.*) [21, 22].

[6] Recentemente si è visto che l'analisi dei dati testuali migliora di gran lunga con l'apporto di meta-informazioni di carattere linguistico (dizionari elettronici, lessici di frequenza, grammatiche locali) e con alcuni interventi sul testo (normalizzazione, lemmatizzazione e lessicalizzazione), cioè attraverso un'analisi statistico-linguistica integrata di tipo lessico-testuale.

Occorrenze	Numero di volte in cui una parola appare nel corso del testo
Forma	Parola nella sua grafia originale nel testo (forma flessa assunta nel discorso dal corrispondente lemma): esempio *parlavo*
Lemma	Forma canonica corrispondente all'entrata del termine nel dizionario, che rappresenta tutte le flessioni con cui quell'unità lessicale può presentarsi nel discorso: esempio *parlare*
Tema	Famiglia lessicale di tutti i lemmi derivati da una medesima radice: esempio *parl>* (*parlare, parlato, parlatissimo, parlottante, parlucchiare, parlamentare, parlamento, parlamentarista,...*)

TABELLA 1

Definizione dei concetti basilari nell'analisi di un testo

grafiche (*type*) con relative occorrenze, come illustrato in tabella 2.

Se si applicano al testo altre operazioni - quali la *normalizzazione*[7], il *tagging grammaticale*[8], la *lemmatizzazione*[9], e/o la categorizzazione *semantica*[10], si produce un corpus che si potrebbe definire "annotato", la cui numerizzazione dà luogo a un vocabolario diverso, per numero di voci (*entrate*) e per quantità di occorrenze a esse associate. Il vocabolario di un testo annotato ha *voci* meno ambigue delle forme grafiche originarie ed è più ricco d'informazioni sul testo[11]. Queste operazioni non sono tutte indispensabili e dipendono dal tipo di analisi di contenuto e dai suoi obiettivi. Si osserva, per esempio, che, analizzando corpus di grandi dimensioni per forme grafiche o per lemmi, i risultati sono sostanzialmente gli stessi. Al contrario, per testi di minori dimensioni o per analisi dei concetti, la riduzione delle parole alla radice comune (*tema*) o al *lemma* fa guadagnare in scoperta di significati e cattura di informazione.

Data la mole d'informazioni presenti nella collezione di testi considerati (*corpus*), non

Antenne	1055	Contro	379
Antenna	792	Campi	367
Telefonia	590	Elettromagnetiche	365
Ripetitori	508	Telefonini	361
Cittadini	499	Radio	360
Tim	498	Inquinamento	354
Installazione	471	Ripetitore	339
Impianti	458	Cellulare	332
Salute	453	Omnitel	315
Onde	442	Metri	306
Cellulari	414	Legge	287
Elettrosmog	407	Elettromagnetico	283
Comune	402	Limiti	274
Sindaco	388	Wind	251
Mobile	386		

TABELLA 2

Esempio di vocabolario: parole piene più frequenti in una rassegna stampa sull'elettrosmog (Fonte: dati Elettra2000)

7 Per normalizzazione s'intende una serie di operazioni di *standardizzazione* del testo, effettuata sulle grafie attraverso il riconoscimento di nomi propri (persone, società, celebrità), toponimi, sigle, date, numeri (telefonici, prezzi, valute), percentuali, così come individuazione di locuzioni, di tipo avverbiale (in modo, per esempio), aggettivale (di massa, in via di sviluppo), o nominale (identificanti entità ricorrenti: per esempio, Capo dello Stato, Presidente del Consiglio, carta di credito).

8 Il *tagging* consiste nel marcare la *forma* con l'attribuzione della sua categoria grammaticale; per esempio: <parlavo> diventa *parlavo*_V.

9 Lemmatizzare significa trasformare la forma nel lemma corrispondente: per esempio <parlavo> diventa *parlare*_V.

10 L'attribuzione di una etichetta di tipo semantico permette di associare la forma ad altre appartenenti ad una stessa classe di equivalenza (per esempio, <cinema> categorizzato come spettacolo, sarà associabile a <circo>, <teatro> ecc.).

11 Per un riferimento sui corpus annotati si veda il sito: http://www.tei-c.org/.

Gli aspetti *grammaticali* sono risolti con i ***lemmatizzatori automatici***, strumenti per la lemmatizzazione del testo che raggiungono livelli di qualità superiori al 95% nella individuazione del giusto lemma. Questi tools sono basati su catene di Markov e/o sull'utilizzo di *grammatiche locali* che individuano nel testo strutture e regole sintattiche capaci di definire univocamente funzioni grammaticali diverse e quindi risalire correttamente al lemma di un termine (per l'italiano [17]). Questa funzione presuppone ovviamente la disponibilità di un *dizionario elettronico* (in grado di essere utilizzato dal computer) durante la lettura automatica del testo[1].
Gli aspetti *semantici* vengono risolti in parallelo con l'utilizzo di *basi di conoscenza*, dove sono inventariati via via per ogni vocabolo i diversi significati espressi nei dizionari (per esempio, il verbo <andare> prevede oltre 200 significati: "andare al Creatore", "andare a nozze", "andare a male", "andare a letto con i polli" ecc.). Per un riferimento generale, si veda Wordnet sviluppato da G. A. Miller presso la Princeton University (http://www.cogsci.princeton.edu/~wn/).

[1] Per approfondire questi aspetti, fra gli altri, si vedano i lavori di Elia [14, 16] e di Silberztein [28]. Al lettore interessato, per ampliare il glossario sulle nozioni fin qui introdotte [1], si consiglia un testo di "Linguistica elementare": per esempio, De Mauro [13].

è possibile tener conto letteralmente di tutto il testo: si pone, quindi, il problema di *estrarre l'informazione significativa* dal corpus, ovvero quella parte di linguaggio che fa la differenza, che contiene gli elementi caratteristici del contenuto o del discorso espresso nel corpus.
Non tutte le parole hanno, naturalmente, la stessa importanza; ma non è la frequenza l'unico elemento a determinare il peso di un termine in un testo. Anche le parole dette una sola volta (i cosiddetti *hapax*) possono essere molto importanti. Molte fra le parole più frequenti sono "parole vuote" (quali per esempio, <e>, <di>, <da>, <il> ecc., dette anche *stop word*), in quanto elementi necessari alla costruzione della frase; oppure sono parole strumentali con funzioni grammaticali e/o sintattiche (<hanno>, <questo>, <perché>,<non>, <tuttavia>), che non sono portatrici di significato autonomo.
Si considerano, al contrario, "parole piene" gli aggettivi, i sostantivi, i verbi e gli avverbi, in quanto termini che hanno un senso in sé (si veda a tal proposito il riquadro sulla parola); le parole più frequenti celano in sé molti usi e, quindi, molti significati (si pensi alla forma <fine> come nome può voler dire *termine*, *obiettivo* o *scopo*, come aggettivo può significare *raffinato* o *sottile*).
È indubbio, dunque, che il riconoscimento grammaticale, con relativo tagging, risolve non poche ambiguità. A tal fine, esistono strumenti di ***lemmatizzazione automatica***, sia grammaticale sia semantica[12].
Sia gli aspetti grammaticali, sia quelli semantici sono spesso risolvibili solo mediante lettura del *contesto locale*, definito da una, due,...*n* parole che precedono o seguono la parola in esame. È dunque evidente che l'analisi di *sequenze di parole* (o *segmenti*) permette di chiarire il significato presente nel testo, di togliere l'ambiguità ai termini ossia di *disambiguare* il testo. Il significato, per esempio, della sequenza <dato di fatto> è univoco, poiché deriva da una locuzione assai comune[13] che toglie l'ambiguità insita nelle parole semplici come <dato> e <fatto>, che in teoria potrebbero essere verbi, piuttosto che nomi. Queste sequenze, riconoscibili come frasi fisse (*multiword expression*), possono essere individuate già nella fase di *normalizzazione* del testo. Altre disambiguazioni sono, di fatto, realizzate con la lemmatizzazione.
Ma il problema di estrarre l'informazione dal testo non è risolta tanto dalla disambiguazione che semmai serve a non fraintendere un significato con un altro, quanto dal selezionare fra le unità di analisi quelle significative, quelle tipiche o caratteristiche dei contenuti di un testo. In generale, ciò avviene selezionando delle *parole chiave*.
È possibile estrarre queste parole in vari modi:
1. con un approccio *corpus based*, si può calcolare un indice, noto come *Term Frequency - Inverse Document Frequency* (TFIDF) [26], che si basa su due assunti:
a. tanto più un termine occorre in un documento tanto più è rappresentativo del suo contenuto;
b. tanti più documenti contengono un termine, tanto meno questo è discriminante [27].

[12] Per l'italiano, come software per la lemmatizzazione automatica dei testi, fra gli altri, si veda Lexical Studio, sviluppato da Synthema, http://www.synthema.it/english/documenti/Prodotti_LexicalStudio_i.pdf.

[13] Le strutture più comuni, ritrovabili nei gruppi nominali, sono del tipo <Nome_Prep_Nome>, <Nome_Agg>, <Agg_Nome>.

L'indice TFIDF è costruito, ponendo a rapporto queste due informazioni[14].

2. mediante confronti con informazioni *esterne al corpus*: sia attraverso criteri di *categorizzazione semantica* predisposti sulla base di specifici modelli di analisi, sia attraverso *lessici di frequenza* che costituiscano dei "riferimenti" rispetto ai quali il *sovra-uso* o il *sotto-uso* di un termine nel corpus può assumere un carattere di *specificità*. Nel software Taltac[15], fra le risorse statistiche sono disponibili, per esempio, vari lessici di frequenza che consentono di estrarre il *linguaggio peculiare* di un corpus, mediante il calcolo di uno scarto standardizzato fra le occorrenze d'uso nel corpus e quelle nel lessico prescelto[16].

3. L'EVOLUZIONE DEI METODI E DELLE TECNICHE DI ANALISI DEI TESTI: DALLA TEXT ANALYSIS AL TEXT MINING

Lo sviluppo delle tecniche di analisi dei testi ha subito profonde evoluzioni negli ultimi cinquant'anni passando da primordiali indagini semiautomatiche orientate allo studio della frequenza delle accezioni di singole parole in grandi raccolte di testi letterari, ad analisi completamente automatiche in grado di decifrare in profondità il senso di una frase all'interno di sterminate raccolte di materiali testuali quali quelle accessibili oggi dal web. Gli strumenti utilizzati per queste analisi dipendono ovviamente dagli obiettivi, ma si fondano essenzialmente su metodi per il trattamento del linguaggio naturale (*Natural Language Processing*) e su tecniche statistiche di tipo multidimensionale.

3.1. Analisi delle concordanze

Ogni metodo o tecnica utilizzata per l'analisi automatica dei testi ha in un modo o nell'altro l'obiettivo di fornire qualche "rappresentazione" del testo, tale da consentirne una lettura mirata.

I primi studi quantitativi sui testi si basavano soprattutto sull'*analisi delle concordanze* che - osservando tutti i contesti locali di una parola d'interesse - consente di discernere i diversi usi e significati di un termine (le sue concrete accezioni nel corpus), per poi confrontare e riunire tali conoscenze in un quadro più complessivo che in taluni casi arriva a definire il lessico di un autore (le prime concordanze furono applicate a studi biblici e risalgono a tempi remoti).

La tipologia delle concordanze presenta una casistica molto ampia. Una discriminazione radicale è espressa dal binomio concordanza verbale – concordanza reale: la prima è concordanza di parole, la seconda è concordanza di cose (concetti, temi, argomenti). Un esempio magistrale di analisi basate sulle concordanze è rappresentato dagli studi di R. Busa circa l'opera di S. Tommaso d'Aquino [10]. Nell'*Index Thomisticus*, la sintesi del lessico di Tommaso ha occupato i primi dieci volumi (su 56) per complessive 11.500 pagine. In questi volumi, sono presenti da una parte tutti i testi con ipertesti interni ed esterni, dall'altra il censimento classificato del vocabolario (il mappale panoramico, secondo l'espressione busiana).

3.2. Analisi delle corrispondenze

Per passare da un livello di studio "unidimensionale", quale può considerarsi quello dell'analisi delle concordanze, a uno "multidimensionale" si può utilizzare l'*analisi delle corrispondenze*. È una tecnica statistica proposta inizialmente negli anni Sessanta da J. P. Benzécri [3] come metodo induttivo

---

[14] L'indice *TFIDF* è espresso dalla seguente ponderazione:

$w_{t,d} = f_{t,d} \cdot \log N/f_t$

dove $w_{t,d}$ è il peso del termine t nel documento d, $f_{t,d}$ la frequenza del termine t nel documento d, N è il numero totale di documenti nel corpus, e f_t il numero di documenti contenenti questo termine.

[15] Software per il Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto: http://www.taltac.it

[16] Si definisce *peculiare* quella parte di linguaggio *tipica* del corpus sia perché è *sovra/sotto-utilizzata* rispetto alla "media" espressa dalle frequenze d'uso nel lessico, sia perché è così *originale* del testo oggetto di studio da non essere presente nel linguaggio assunto come riferimento. Per il calcolo delle occorrenze d'uso (indice d'uso), si veda Bolasco [6].

Obiettivi e strumenti delle tecniche statistiche di analisi multidimensionale dei testi

Al fine di analizzarne la variabilità linguistica e la struttura, il corpus viene in genere studiato per *frammenti* (spezzoni brevi di testo: proposizioni elementari o enunciati, singoli documenti, risposte, *e-mail* ecc.) o *per parti* (sub-testi o raggruppamenti dei frammenti per attributi: cronologici, tematici, caratteristiche socio-demografiche ecc.). In questa prospettiva, assume interesse la frequenza delle parole nelle parti o nei frammenti. Infatti, uno studio del testo fondato su base quantitativa, consiste sempre nel confronto di diversi *profili* lessicali, ossia di altrettante sub-distribuzioni statistiche generate dall'insieme delle frequenze delle parole in ciascuna parte e/o frammento. In quest'ultimo caso spesso la quantità di occorrenze viene ridotta a semplice "*presenza/assenza*".

Di fatto queste suddivisioni del corpus danno luogo a matrici di tre tipi diversi: una matrice "frammenti × parole", contenente dati booleani 0/1 (dove 1 indica la presenza della parola nel frammento e o la sua assenza); una matrice "parole × parti", contenente le frequenze con cui ogni parola ricorre nella parte (sub-testo); una matrice "parole × parole" che documenta l'associazione (*co-occorrenza*) di coppie di parole nei frammenti del corpus: qui il dato interno alla matrice può registrare la sola esistenza dell'associazione (0/1) o pesarne l'intensità con una misura di relazione.

Secondo l'algebra matriciale, ogni riga o colonna di queste matrici rappresenta un vettore, descrivente il profilo lessicale. Le tecniche utilizzate per l'analisi di tali matrici mirano alla sintesi o riduzione dei dati, attraverso lo studio della variabilità statistica.

In particolare, le *tecniche fattoriali* - attraverso una riduzione del numero di variabili del fenomeno (vettori colonna) - producono delle nuove variabili sintetiche, in grado di ricostruire i principali assi semantici che caratterizzano la variabilità dei contenuti del testo. L'*analisi delle corrispondenze* è la tecnica fattoriale utilizzata nel caso dei dati testuali. Essa visualizza le principali co-occorrenze fra parole presenti nel testo, sulla base della loro vicinanza nei piani cartesiani costituiti da coppie di assi fattoriali, ricostruendo in tal modo delle vere e proprie mappe del contenuto del testo, che forniscono spesso una rappresentazione globale del senso sottostante il discorso.

Le *tecniche di clusterizzazione* e di *segmentazione* mirano, invece, a ridurre la quantità delle unità statistiche (vettori riga), producendone una classificazione multidimensionale, in grado di definire delle tipologie attraverso le quali leggere simultaneamente le caratteristiche d'interesse.

La *cluster analysis*, come famiglia di metodi di raggruppamento (gerarchici e non, scissori o aggregativi), consente di individuare classi di parole o di frammenti di testo, caratterizzati da una forte omogeneità interna, tale da poter ricostruire i principali "mondi lessicali" presenti nel corpus, ossia i differenti "modi di parlare" del fenomeno studiato, contenuto nel testo.

Per maggiori riferimenti, anche ad altri metodi multidimensionali, si rimanda a Bolasco [6].

per l'analisi dei dati linguistici, assai efficace per trattare matrici di dati di ampie dimensioni, risultanti dalla descrizione di profili lessicali d'interesse. Per esempio, il "profilo" di nomi definito dalle associazioni che questi hanno con un insieme di verbi presenti in un corpus assai esteso di testi: questa informazione è raccolta in una matrice di dati che incrocia le co-occorrenze di nomi (in riga) con i verbi (in colonna). L'analisi delle corrispondenze di tale matrice produce una rappresentazione delle associazioni fra nomi e verbi in maniera tale da riprodurre per vicinanza dei punti su un piano cartesiano la similarità fra profili; questa tecnica si rivelò assai utile a ricavare induttivamente alcune regolarità linguistiche sulla base della cosiddetta distanza del chi-quadro[17]. Per maggiori dettagli si veda il riquadro di approfondimento.

Più recentemente, con il crescere della disponibilità dei testi da analizzare e per rispondere all'incremento esponenziale delle fonti quotidianamente da consultare/interrogare per esempio sul web in aziende o istituzioni, in parallelo alle tecniche di Text Analysis, si sono sviluppate procedure di *Text Mining* (TM) per estrarre informazioni da materiali espressi in linguaggio naturale, riassumibili sotto due "logiche": *Information Retrieval* (IR) e *Information Extraction* (IE)[18]. In maniera assai schematica si può dire che l'IR s'interessa al documento nella sua globalità, mentre l'IE seleziona le informazioni specifiche all'interno del documento, che in genere vanno a popolare un database strutturato.

3.3. Information Retrieval

Sono ormai molto diffusi software di *recupero di informazioni* in grado di effettuare ricerche su grandi collezioni di testi sulla base di richieste (*query*) formulate come singole parole o come frasi: l'esempio più comune può essere quello dei motori di ricerca sul web.

Impiegando software tradizionali, i singoli documenti d'interesse sono trattati come entità a sé stanti e, in particolare, non vengono prese in considerazione le possibili relazioni tra i documenti. Nei software sviluppati, invece, su *database*, ai singoli documenti vengo-

[17] Questo processo è definito in dettaglio da Benzécri ([3], p.102-105); svariati esempi di applicazione sono in Benzécri [2].

[18] Per una recente panoramica su questi due punti di vista Poibeau [25].

no di solito associati dei *metadati*. In tal modo, è possibile classificare e pesare in misura diversa i risultati della *query* sulla base di queste informazioni aggiuntive.
Questi software permettono di cambiare in tempo reale la visualizzazione delle informazioni in base alle esigenze del momento, utilizzando un approccio ai dati di tipo OLAP (*On-Line Analytical Processing*).
La fase di Information Retrieval si compone essenzialmente di due sottofasi: la *selezione delle fonti e il recupero dei testi* (unitamente alle eventuali informazioni relative ai metadati).
Ai fini del recupero dei documenti è necessario effettuare una scelta sul tipo di analisi delle parole e/o delle frasi.
Questa analisi può essere di tre tipi (o un insieme combinato dei tre): ortografico, semantico e statistico.
- *Ortografico*: riconoscimento delle parole in base alla loro grafia, senza alcun tentativo di correlarle al contesto.
- *Semantico*: associazione della parola al concetto che vuole esprimere. Parole diverse possono essere usate per esprimere concetti simili, pre-definiti in una base di conoscenza.
- *Statistico*: confronto della frequenza d'uso delle parole con una distribuzione di riferimento (lessico di frequenza).

Nella selezione delle fonti si individuano i soli documenti rilevanti, cioè compatibili con i criteri della richiesta[19]. Le fonti possono essere le più diverse: archivi che contengono informazioni espresse in linguaggio naturale, database strutturati che contengono informazioni già sintetizzate (con o senza metadati), immagini di documenti (in questo caso entra in gioco una componente successiva del processo che si occupa della scansione OCR (*Optical Character Recognition*) per trasformare l'immagine in testo).
Nella fase di IR si estraggono dai documenti selezionati quei frammenti di testo che contengono le parole o le frasi che costituiscono i criteri della richiesta. Soprattutto nel caso di frasi, la qualità dell'algoritmo di selezione e di estrazione è cruciale per ottenere buoni risultati. Per esempio è molto importante controllare la co-occorrenza delle parole e valutare la loro vicinanza all'interno del testo.
Individuate le parole (o le frasi), si calcola un peso per ciascun termine (si può usare la frequenza all'interno del documento, o funzioni più complesse, come l'indice TFIDF precedentemente definito).
Esistono diversi metodi per misurare questa rilevanza: vettoriale, probabilistico e booleano.
- Con il metodo *vettoriale* si rappresentano in spazi geometrici i documenti e le richieste che li hanno "generati": in tale spazio la vicinanza tra richiesta e documento misura la rilevanza di quest'ultimo rispetto alla prima.
- Con il metodo *probabilistico* un documento è tanto più rilevante quanto maggiore è il peso delle parole compatibili con la richiesta.
- Con il metodo *booleano* si valuta la presenza/assenza di parole tra documento e richiesta.

Con l'ultimo metodo si può solo dire se un documento è o no rilevante rispetto a una query, mentre con i primi due, oltre a determinare la presenza/assenza di rilevanza tra documento e richiesta, si genera anche una graduatoria di pertinenza, utile per filtrare i documenti.

3.4. Information Extraction

Dopo aver recuperato i documenti rilevanti, occorre sintetizzarne il contenuto informativo e renderlo disponibile per ulteriori analisi. Un compito molto impegnativo, le cui tecniche non sono del tutto standardizzate (nell'ambito del text mining).
La rappresentazione standard di un documento è quella di un vettore nello spazio geometrico definito da un numero di componenti pari all'ampiezza del vocabolario del corpus. Ma questo modo di rappresentare i documenti pone problemi di dimensione, perché cresce con l'ampiezza del vocabolario. Sono stati messi a punto diversi modi per ridurre la dimensione dei vettori-documento, tra i quali, per esempio, il considerare solo le parole significative del vocabolario e, quindi, utilizzare vettori-docu-

[19] A volte questa fase non può essere eseguita in modo automatico e viene affidata ad un esperto del settore.

mento di dimensione pari solo al numero di parole chiave.
La rappresentazione vettoriale dei documenti ha, peraltro, il difetto di non cogliere le relazioni tra parole, portando così potenzialmente a una rilevante perdita di informazione nel passaggio dal "*discorso*" alla sua formalizzazione vettoriale. Sono oggi disponibili varie tecniche per evidenziare l'informazione legata a queste relazioni che si basano in sostanza sullo studio delle co-occorrenze di parole nell'ambito della stessa frase[20]. Studiando le co-occorrenze che superano una soglia stabilita (in termini di frequenza), si cerca di derivare delle regole generali di associazione, che permettano, in relazione al contesto di analisi, di identificare delle sequenze significative di parole (non necessariamente adiacenti).
Un ulteriore passo molto importante, previa un'efficiente disambiguazione delle parole, è la *classificazione* dei documenti. Questa viene eseguita a partire dai metadati eventualmente associati ai documenti, e in genere mediante una lista pre-definita di categorie nelle quali far rientrare i documenti basandosi sulla presenza delle parole e/o delle sequenze più significative in essi contenuti. A tal fine, si utilizzano processi semi-automatici, che possono essere addestrati o che comunque sono in grado di migliorare la loro capacità di assegnazione in base alle operazioni precedenti. L'obiettivo perseguito con questi processi – definito da un punto di vista formale – consiste nell'attribuire un valore *vero* o *falso* a ciascuna coppia (documento, categoria) per tutti i documenti da analizzare e tutte le categorie presenti nelle liste di riferimento [27].
Operando sul versante della sintesi del contenuto, si riconducono le parole e le sequenze che caratterizzano i documenti a classi di significato derivate da una base di conoscenza esterna al corpus: in tal modo è possibile *concettualizzare* i documenti, producendone una rilevante riduzione in termini di dimensione, senza però perdere quantità significative di informazione. Questa fase, detta *summarization*, fornisce una rappresentazione astratta dei documenti che enfatizza i temi qualificanti del testo ed elimina gli altri[21]. La summarization è un pre-requisito per il popolamento di un eventuale database di concetti/azioni/parole d'interesse, che è strutturato in maniera più rigida rispetto a un testo espresso in linguaggio naturale.
Dopo aver classificato i documenti, si pone il problema della loro visualizzazione in un grafico sintetico di facile interpretazione. Normalmente questo problema viene risolto in due modi: mediante tecniche di *clustering*, che permettono di spostare l'attenzione dai singoli documenti a gruppi di documenti, in minor numero e quindi più facilmente rappresentabili; oppure mediante analisi di tipo multidimensionale (metodi fattoriali), che consentono la proiezione dei singoli documenti in spazi geometrici ridotti (tipicamente 2-3 dimensioni).

4. ESEMPI DI TEXT ANALYSIS

Al fine di ripercorrere alcune tra le fasi classiche della Text analysis, si illustrano in concreto due casi di studio. Il primo riguarda procedure e risultati di un monitoraggio sull'informazione relativa all'*elettrosmog*. Il secondo concerne un'analisi di messaggi *on-line* scambiati fra insegnanti incaricati di diverse funzioni obiettivo (FO). In entrambi i casi, l'obiettivo dell'analisi consiste nel conoscere il contenuto di fondo dei materiali oggetto di studio, al fine di valutare l'atteggiamento, le intenzioni e i diversi punti di vista degli "autori" dei testi (nella fattispecie, giornalisti o insegnanti).

4.1. Campi di applicazione

I campi di applicazione della Text analysis e le fonti di materiali testuali sono stati finora i più diversi. Fra questi, i testi tradizionalmente intesi (letterari, tecnico-scientifici o altra saggistica) rappresentano solo una mi-

[20] Anche se si stima che circa il 10-20% delle relazioni significative tra parole sia di tipo inter-frase, cioè a cavallo di frasi diverse.

[21] Si veda a tal proposito lo studio di Mani e Maybury [23].

nima parte delle applicazioni. Fra i tipi di corpus più studiati figurano quelli relativi a: discorsi politici (parlamentari, elettorali, dibattiti) e relazioni periodiche di pubbliche istituzioni (Banca d'Italia, Onu ecc.); rassegne stampa o intere annate di periodici; documenti tecnico-settoriali (archivi documentali, brevetti); collezioni o raccolte di *testi corti*: progetti, abstract, bibliografie, manifesti politici, messaggi pubblicitari, titolazioni di articoli di stampa, agenzie d'informazione. Ma sono assai frequenti anche analisi di testi prodotti a partire da indagini sul campo: indagini con risposte libere alle domande aperte nei questionari, interviste non direttive o semistrutturate, storie di vita o discussioni di gruppo (*focus group, forum in Internet, chat o news group*). Sono stati, anche, analizzati protocolli clinici, biografie, trascrizioni di messaggi non testuali (linguaggi visivi, musicali, gestuali/comportamentali ecc.), nonché "trascrizioni" del linguaggio parlato attraverso il riconoscimento vocale e infine, recentemente, studi su e-mail e sul lessico degli *sms*.
Queste applicazioni interessano psicologi, sociologi, medici, antropologi, storici, semiologi e specialisti della comunicazione.
La maggior parte delle analisi si fonda sull'interpretazione delle variazioni linguistiche con finalità psico/socio-linguistiche e sul riconoscimento del senso di fondo espresso nei testi. Esempi particolari di analisi, fra gli altri, sono alcuni studi sull'autenticità dell'autore di un documento o sulla dinamica del discorso nelle arringhe processuali oppure analisi del linguaggio in condizioni estreme di sopravvivenza quali quelle che si determinano ad alta quota, nelle profondità sottomarine.

4.2. Una rassegna stampa sull'elettrosmog

Lo studio era volto a misurare accuratamente nel tempo e nello spazio la presenza di temi e argomenti intorno al modo di trattare il fenomeno dell'inquinamento elettromagnetico sulla stampa quotidiana, a partire da un campione di testate giornalistiche a diffusione nazionale e locale, in un periodo di quattordici mesi, dall'ottobre 1999 al novembre 2000[22].
La rassegna è formata da 685 articoli[23] raccolti per individuare le caratteristiche generali del *linguaggio* presente nella stampa, con l'obiettivo di catturare la *terminologia* utilizzata e cogliere il livello di attenzione verso i vari aspetti del fenomeno e il loro tipo di percezione.
Le fasi di studio[24] hanno comportato: in primo luogo, una analisi generale del *vocabolario* utilizzato, in termini di forme testuali (parole e locuzioni) più frequenti; in secondo luogo, una evidenziazione dei *lemmi* più ricorrenti *per categorie grammaticali*; in terzo luogo, mediante una riduzione al tema[25] delle principali unità lessicali selezionate, l'individuazione del *linguaggio peculiare*, che ha permesso di quantificare le diverse percezioni del fenomeno.
Osservando il vocabolario già riportato in tabella 2, è interessante notare che il termine *elettrosmog* non è la parola-tema per eccellenza, ma è preceduta da antenne/a, telefonia, ripetitori-installazione-impianti, onde e cellulari. Ciò consente di definire subito l'ampio spettro del "tratto semantico" che ruota intorno all'argomento, ancor meglio inquadrabile dalle espressioni più ricorrenti riportate in tabella 3. Al di là del fenomeno in sé, in essa appaiono anche termini riguardanti il *Comune*, il *sindaco*, la *salute*, che costituiscono tracce importanti del rapporto che il fenomeno ha con l'opi-

[22] Rapporto Interno Consorzio Elettra 2000, Centro di Documentazione, http://www.elettra2000.it.

[23] Il corpus è pari ad un dossier di 750 pagine; la sua analisi ha prodotto un "vocabolario" di oltre 20.000 parole diverse, per un totale di 250.000 occorrenze.

[24] Per un maggior dettaglio su queste fasi si rimanda al Report, disponibile presso il Centro di Documentazione del Consorzio Elettra 2000.

[25] Per riduzione tematica s'intende un raggruppamento delle occorrenze di parole o espressioni secondo la loro *radice* riguardante il tema o significato comune. Ciò può concernere solo più flessioni di uno stesso lemma (<*cellular*+> sta per "cellulare" o "cellulari", siano essi aggettivi o sostantivi) o più entrate dello stesso lessema (<*controll*+> corrisponde alla fusione delle occorrenze di "controllo/i", di "controllore/i" e di varie voci e flessioni del verbo "controllare").

Telefonia			
Telefonia mobile	348	Gestori di telefonia	30
Telefonia cellulare	266		
Elettrosmog			
Campi elettromagnetici	292	Emissioni/radiazioni elettromagnetiche	81
Onde elettromagnetiche	274	Contro l'elettrosmog	54
Inquinamento elettromagnetico	231	Impatto ambientale	50
Istituzioni e Legislazione			
Il Sindaco	229	Ministero dell' Ambiente	33
Il Comune	195	In regola	40
Amministrazione comunale	114	Nulla osta	32
Consiglio comunale	105	Raccolta di firme	37
Concessione edilizia	75	Contro l'installazione	35
Legge quadro	52	All'unanimità	32
Legge regionale	36		
Salute			
La salute	334	Tutela della salute	53
Salute dei cittadini	77	Salute pubblica	67
Collocazione Impianti			
Stazioni radio	127	Centro abitato	91
Radio base	98	Centro storico	39
Stazioni radio base	76	In città	51
Sul tetto	103	Territorio comunale	36
Nuove antenne	47	Campo sportivo	34
Antenna selvaggia	33	In prossimità	36
Alta tensione	51	A ridosso	34
Ad alta frequenza	33	Nelle vicinanze	34
6 V	38	50 m	32
Volt per metro	31	Pochi metri	30

TABELLA 3
Tematiche più ricorrenti nella Rassegna Stampa

nione pubblica e con i problemi legati alla salute.
Successivamente si è proceduto all'analisi dei contenuti specifici degli articoli, al fine di cogliere le diversità di approccio al tema dell'elettrosmog delle diverse testate giornalistiche. Questa fase avviene confrontando i "profili lessicali" dei vari giornali mediante l'applicazione di un opportuno test statistico, che misura lo scarto tra la frequenza dei termini di ciascun giornale e la loro frequenza generale nel corpus. In tal modo, si estraggono le parole ed espressioni *specifiche* di ciascun giornale. Questa tecnica fa emergere i vari modi di percezione, i diversi livelli di attenzione e il tipo di "polemiche" sollevate nella stampa, le cui risultanze tematiche generali sono riassunte nella tabella 4.
La presenza dei temi presenti negli articoli si rileva anche attraverso i verbi, che possono essere raccolti nelle voci generali riportate nel riquadro a fianco.
Da quanto esposto finora emerge che esistono profonde diversità di trattazione del fenomeno, in gran parte dipendenti dall'area geografica d'appartenenza della testata, nonché

a = *impianti*	14%	installare, montare, spostare, smantellare, ...
b = *opinione pubblica*	53%	chiedere, spiegare, individuare, verificare, ...
c = *rischio*	33%	evitare, intervenire, bloccare, causare, ...

1	**impianti**: antenn+, install+, ripetitor+, impiant+, tralicci+, apparecch+, lavori, elettrodott+, stazion+ radio base, telefonic+, posiziona+, base, antenn per la tele>, posizione, emittent+, alta tensione, antenn selvagg+, cavi
2	**cittadini**: cittadin+, contro, chied+, richiest+, protest+, comitato, spieg+, abitant+, società, comitati, condomin+, bambini, ricors+, popolazione, quartiere, persone, denunc+, firma+, gente, battaglia, petizion+, associazione, assemblea, inquilin+, guerra, comunicazione, contro l elettrosmog, proprietari, lotta
3	**prodotti**: telefonin+, cellular+, telefon cellular+, telefonia mobile, concession+, auricolar+, telefonia, telecomunicazion+, radiotelevis+, televisiv+, consumatori, Gsm
4	**gestori**: Tim, Omnitel, Wind, gestor+, Telecom, milioni, Enel, Blu, mila, Umts, di proprietà, miliardi, gestor telefonia mobil>, Rai, licenze, licenza
5	**ambiente/elettrosmog**: ambient+, elettrosmog, problem+, onde elettromagn+, camp+ elettromagne>, emission+, radio, inquin+ elettromagn, Arpa+, territorio, inquin+, onde, esposizion+, camp+, emess+, frequenz+, Ambiente, elettric+, stazioni+, concession ediliz+, centro, camp magnetic+, elettromagnetic+, alta, elettronic+, ministero dell Amb>, impatto ambientale, magnetic+
6	**salute/rischi**: risch+, salute, pericol+, preoccup+, Asl, dann+, radiazion+, nociv+, provoc+, tutel+, sanitar+, allarm+, alla salute, cautel+, tumor+, sospen+, evit+, a rischio, salute dei cittadi>, conseguenze, Sanità, leucem+, protezione, salute pubblica, tranquill+, cancr+, Oms
7	**istituzioni**: sindaco+, Comune, legge, assessor+, comunale, autorizzazion+, approv+, consiglier+, Giunta, Tar, Consiglio Comunale, regionale, Amministrazione comun+, presidente, decreto, ordinanz+, Region+, parere, delibera, responsab+, commissione, Governo, amministrazione, comuni, intervento, Comuni, autorità, indagine, risposta, soluzione, Calzolaio, comunali, misure
8	**controllo/sicurezza**: controll+, norm+, regolament+, x metri, verific+, rilasc+, volt per metro, sul tett+, sicurezza, provvedimento, microtesla
9	**ricerca**: scientific+, siti, studi, studio, ricerca, risultati, monitora+, esperti, misurazion+, sito, mapp+, attenzione, ricerche, prevenzione, Università, ricercatori
10	**territorio**: zon+, limiti, vicin+, scuole, residenti, abitaz+, distanz+, case, città, palazz+, a poc distanz, edifici+, abitar+, limite, urbanistic+, metri, aree, ediliz+, centr abitat+, entro, sportiv+, fino a, vicinanz+, limiti previsti, scuola, livelli, luoghi, in città.

TABELLA 4
Sintesi delle principali radici lessicali della rassegna stampa, raggruppate per temi

dall'essere un quotidiano a carattere nazionale o regionale.
Dall'analisi si è potuto evincere tra l'altro come testate quali La Stampa, il Corriere della Sera, Il Sole 24 Ore, Italia Oggi, Il Messaggero e La Repubblica pongono un'attenzione maggiore, sia in assoluto sia rispetto agli altri giornali, a una trattazione del fenomeno in termini di tematiche generali sull'*ambiente*, l'*elettrosmog*, la *salute*, la *ricerca*, ma parlano anche significativamente dei prodotti. Al contrario, testate quali Il Tirreno, Il Secolo XIX, Corriere Adriatico e altri giornali regionali incentrano la loro attenzione su problemi locali e particolaristici, legati ai singoli impianti, al territorio e sono sensibili alle azioni dei cittadini e delle istituzioni di governo locali.

4.3. Un forum di discussione fra insegnanti per la formazione a distanza

L'analisi della discussione sulla formazione a distanza dell'Istituto Regionale di Ricerca Educativa per il Lazio era mirata a conoscere il *lessico praticato* in oltre 29.000 messaggi scambiati in un anno sul web fra insegnanti in diverse "conferenze" sulle funzioni obiettivo, nei vari forum provinciali predisposti dalla Biblioteca di Documentazione Pedagogica di Firenze (http://www.bdp.it/).
L'insieme dei materiali testuali, assai voluminoso (1,8 milioni di occorrenze, equivalenti a oltre 6.000 pagine di testo) è portatore di moltissime informazioni, che sono state via via estratte. A parte uno studio a sé stante sulla concatenazione dei messaggi, sono ri-

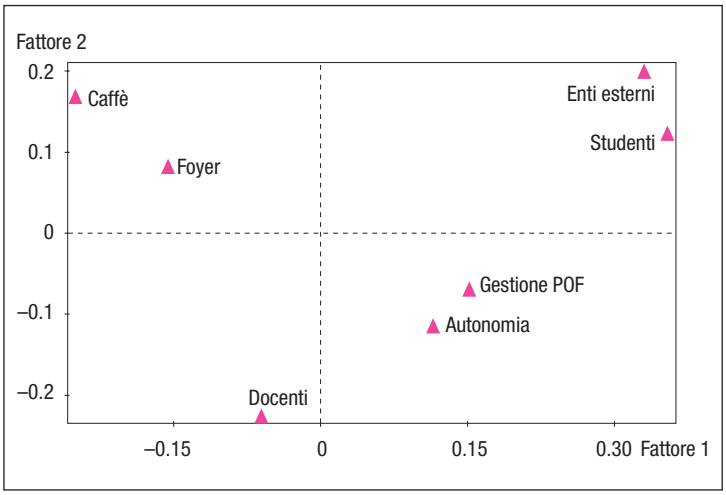


FIGURA 1

Mappa delle 7 conferenze sul sito delle FO

sultate molto interessanti l'analisi della punteggiatura, l'inizio del messaggio, l'analisi dei verbi e degli aggettivi.

Tutte queste sub-analisi hanno testimoniato l'entusiasmo degli insegnanti nel partecipare alla discussione e nella scoperta di poter comunicare a distanza e saper navigare in Internet, nonché il desiderio di portare la propria esperienza e raccontare eventi con molti particolari[26].

A titolo di esempio, si privilegia in questo articolo l'estrazione dell'informazione limitata ai verbi e i risultati emersi dall'applicazione dell'analisi delle corrispondenze.

Per quanto riguarda l'estrazione dei verbi peculiari, ottenuti confrontando la frequenza dei lemmi nel corpus dei messaggi con quella presente nel linguaggio standard, emergono i toni generali del <*ringraziare, sperare, confrontare, contattare, condividere, gradire, augurare, volere, collaborare, conoscere, piacere, imparare, incontrare, desiderare, coordinare, scusare*>, oltre che i riferimenti all'attività web come <*allegare, navigare, scambiare, inserire, scaricare, visionare, accedere, comunicare*>, che documentano le due principali dimensioni percepite della discussione nei forum.

L'applicazione dell'analisi delle corrispondenze produce invece una mappa della variabilità di linguaggio in funzione dei tipi di conferenze (Figura 1), secondo un *continuum* che dà il *senso* latente del "discorso" sviluppatosi nei forum. L'interpretazione degli assi di un piano fattoriale viene fatta a posteriori a partire dalla disposizione dei punti sul piano cartesiano, o meglio, asse per asse, secondo la graduatoria delle coordinate dei punti sull'asse (per approfondimenti si rimanda a Bolasco [6]). In particolare, nel nostro caso, l'interpretazione del posizionamento delle conferenze in figura 1, si dedurrà dal posizionamento delle parole in figura 2. In quest'ultima figura, seguendo l'asse orizzontale da sinistra a destra si osservano parole coerenti con un *gradiente* crescente di densità di *interessi e argomentazioni* che partendo da un livello minimo del grado di realizzazione, partecipazione, collaborazione nelle attività ("ci sono anch'io", "finalmente", "riuscita" "messaggi", "buon lavoro a tutti" espressioni tipiche delle conferenze "Caffè" e "Foyer", meno ricche di contenuti), passa via via secondo un crescendo del primo fattore nelle cinque conferenze sulle funzioni obiettivo. Dapprima fra le FO "Docenti", poi "Autonomia" e "Gestione del POF" si incontrano "sostegno al lavoro dei docenti", "corso integrativo", "bisogni formativi", "documentazione", "tecnologie", "informatica", "laboratori", "autonomia", "programmazione", fino ad arrivare ai "Progetti con Enti esterni" e ai "Servizi per gli Studenti", che risultano essere le conferenze più "dense" di messaggi argomentati e di attenzioni/sensibilità educative manifestate dagli insegnanti ("rapporti", "genitori", "educazione", "continuità", "dispersione scolastica", "interventi", "progetti", "educazione alla salute", "alternanza scuola lavoro"). Per brevità si lascia al lettore la scoperta di altri contenuti; per approfondimenti si rimanda a Bolasco [7]. Dalla lettura si confermerà l'impressione generale di un crescente grado di interesse nella partecipazione e nella comunicazione delle esperienze, che trova nelle aree di discussione quali il Foyer e il Caffè il suo livello minimo (coloro che sono rimasti "alla finestra", a guardare dall'esterno questo nuovo strumento di comunicazione *on-line*) e nelle aree relative alle

[26] Si rimanda il lettore interessato, all'analisi di dettaglio in Bolasco [7].

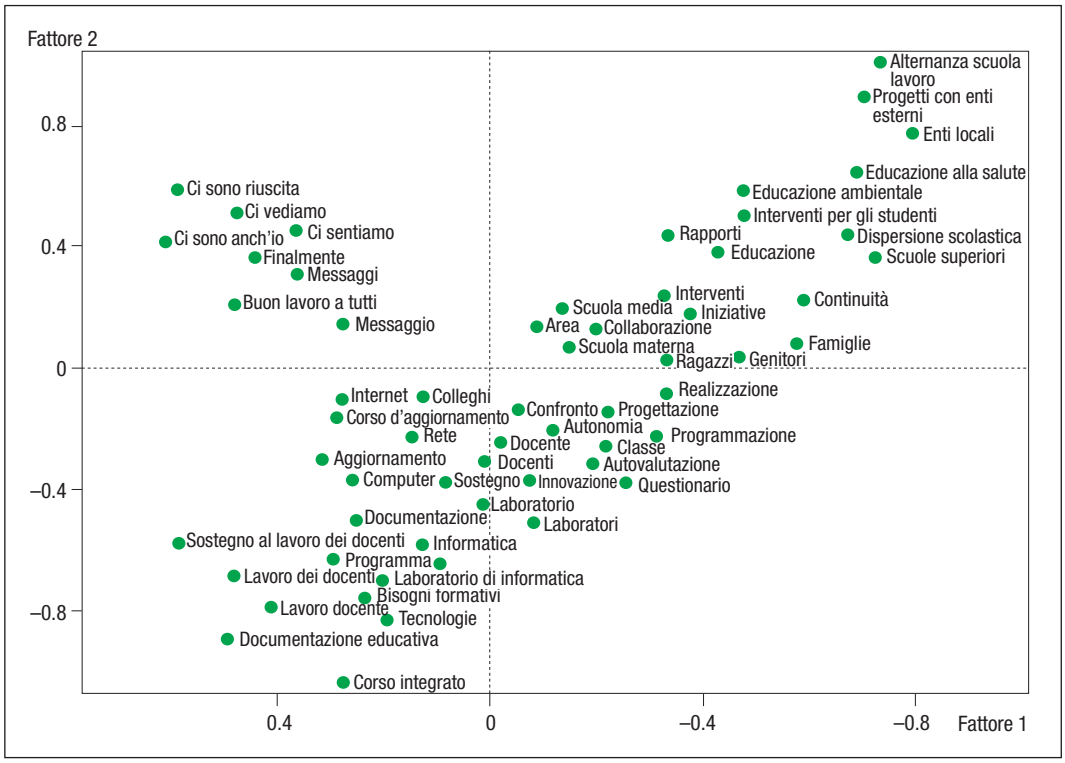


FIGURA 2

Piano fattoriale delle corrispondenze - Mappa delle unità lessicali (parole e sequenze più significative) associate ai forum sulle FO

FO su Studenti e su Attività con enti esterni il suo massimo (coloro che hanno già *svolto* molte attività e rendono conto delle esperienze già compiute).
Seguendo l'asse verticale, dall'alto verso il basso, si passa dalla dimensione del "progetto" a quella del "corso integrativo", ossia dal generale-ideale (sia Caffè, sia Enti esterni, ma con modalità opposte espresse dal primo asse fattoriale, rispettivamente "da realizzare"/"già realizzato") al particolare-concreto (Docenti). Le parole diverse in ciascun quadrante segnano dunque le differenze fra i linguaggi degli insegnanti.

4.4. Ulteriori esempi di applicazione

Da uno *studio sulle encicliche papali* svolto da Bisceglia e Rizzi [4] si rileva come anche solo le alte frequenze siano in grado di selezionare alcuni elementi essenziali dei documenti in esame. L'analisi delle prime 5 occorrenze più utilizzate (*top five*) dagli ultimi quattro pontefici nelle loro encicliche fornisce interessanti elementi che caratterizzano i pontificati (Figura 3). Il termine *fede* è presente solo in Pio XII, i termini *bene* e *sociale* in Giovanni XXIII, il termine *mondo* solo in Paolo VI e *uomo* solo in Giovanni Paolo II. Al contrario, il fatto che *Dio*, *chiesa* e *vita* siano comuni ai 4 papi rende questi termini meno significativi, perché prevedibili.
Nel campo del marketing e degli studi sociopsicologici, un contributo interessante all'analisi semantica nell'estrazione di informazione è dovuto all'*approccio semiometrico* [22]. La semiometria è una tecnica di descrizione dei legami semantici fra parole e si fonda sull'analisi di un insieme selezionato di termini che, al di là del loro significato, evocano ricordi e/o provocano sensazioni gradevoli-sgradevoli[27]. A partire da un campione di individui intervistati che reagisce all'insieme

[27] Si tratta di circa 200 sostantivi, verbi o aggettivi come per esempio: l'assoluto, l'ambizione, l'amicizia, l'angoscia, astuto, il coraggio, il pericolo, la disciplina, risparmiare, efficace, l'eleganza, la famiglia, la gioia, la gloria, la guerra, la giustizia, avventuroso, boheme, concreto, Dio, nobile, sublime, teatro ecc.

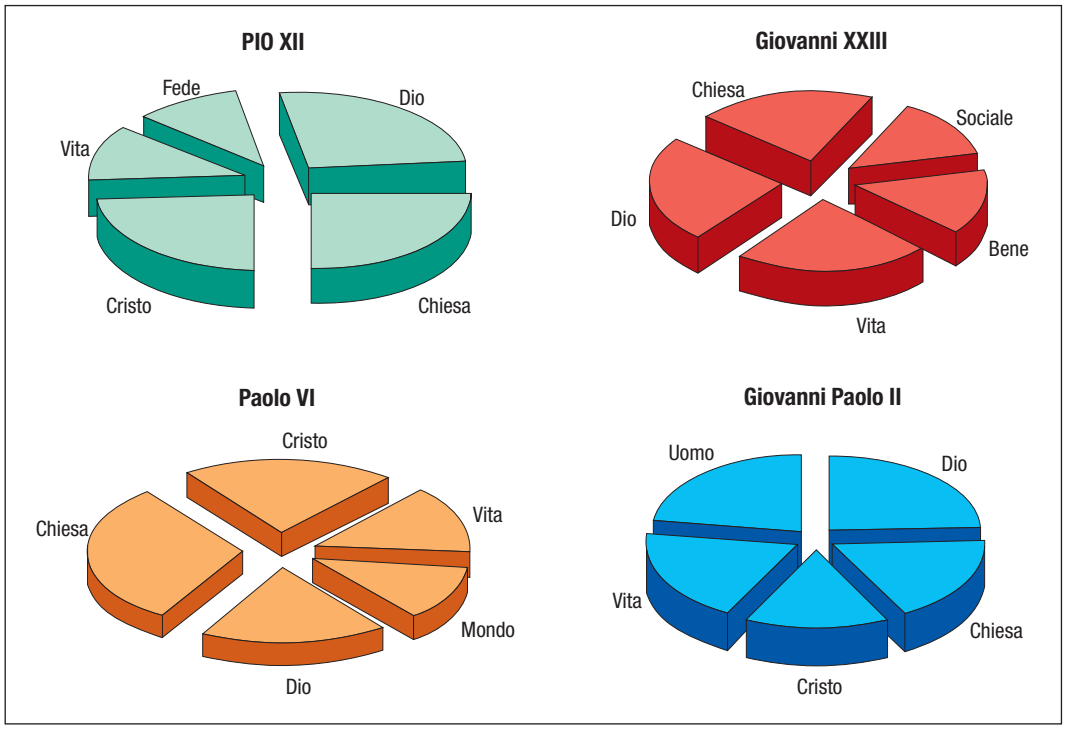


FIGURA 3
Le cinque parole maggiormente utilizzate da questi pontefici in tutte le encicliche da loro promulgate

di questi stimoli su una scala di sensazioni a 7 livelli da –3 a +3, secondo un crescendo di gradevolezza, si descrivono sistemi di valori e stili di vita. La tecnica di rappresentazione dell'informazione è quella dei piani fattoriali, ottenuti con l'analisi delle corrispondenze, che consentono il posizionamento degli individui su polarizzazioni semantiche quali: dovere/piacere, attaccamento/distacco, spirito/materia e così via.

Per **Technology Watch** (TW) si intende l'attività di monitoraggio della tecnologia, con lo scopo di evidenziare le caratteristiche delle tecnologie esistenti e le loro relazioni, nonché di identificare e descrivere le tecnologie emergenti.
Gli elementi fondanti del TW sono quindi la capacità da un lato di raccogliere tutte le informazioni sulle tecnologie consolidate che potrebbero non essere comunemente note, dall'altro di evidenziare sviluppi tecnologici ancora in uno stato embrionale, cogliendone potenzialità e campi di applicazione e analizzandone le relazioni con le tecnologie già note.

5. UN ESEMPIO DI TEXT MINING NEL TECHNOLOGY WATCH

I campi di applicazione del *Text Mining*[28] spaziano dal mining sul web (flussi di navigazione dei siti, comportamenti di visita nel sito), all'analisi delle banche dati sui brevetti (*Patent Analysis*) e più in generale al ***Technology Watch*** dalle azioni di *Customer Relationship Management* (gestione dei *call center, re-routing* di e-mail, analisi di *complaint*, monitoraggio sulla pubblicità dei prodotti, marketing diretto) alla gestione delle Risorse Umane (analisi dei *curriculum vitae* per la selezione di competenze specifiche, monitoraggio dell'Intranet aziendale); dalla classificazione automatica delle risposte in linguaggio naturale nelle indagini Istat (censimento, forze di lavoro, indagini multiscopo) alla categorizzazione dei *document warehouse* aziendali (database di case editoriali, basi documentali di istituzioni pubbliche ecc.).
Un esempio d'applicazione di Text Mining per il Technology Watch si trova all'interno del progetto europeo FANTASIE[29] (*Foreca-*

[28] Per un riferimento generale al TM si veda il sito del progetto europeo NEMIS (*Network of Excellence in Text Mining and its Applications in Statistics*): http://nemis.cti.gr

[29] Si veda il sito: http://www.etsu.com/fantasie/fantasie.html

sting and Assessment of New Technologies and TrAnsport System and their Impact on the Environment) sullo sviluppo tecnologico legato ai problemi dei trasporti e del traffico, volto a valutarne la situazione attuale e gli sviluppi a breve e medio termine. Le fonti per le analisi di text mining sono in questo caso interviste, documenti ricavati dalla letteratura e brevetti.

In una prima fase, sono stati intervistati esperti nei campi coinvolti nella ricerca. L'analisi delle interviste ha consentito di creare una prima base di conoscenza, estraendo la terminologia di riferimento in relazione a tecnologie, materiali, mezzi di trasporto e infrastrutture.

In un secondo momento, sulla base di questi riferimenti, è stata effettuata con tecniche di IR la selezione dei documenti rilevanti nella letteratura, sia quella pubblicata su carta che quella disponibile su Internet. Un problema tipico di questo genere di fonte è il ritardo temporale fra lo sviluppo delle ricerche e la loro effettiva pubblicazione, che la rende poco adatta per analizzare tecnologie emergenti. Al contrario, l'analisi su brevetti (*patent analysis*) nell'attività di TW è il metodo più efficace per estrarre conoscenza su argomenti in espansione.

Per questa ragione, la terza fase del progetto Fantasie ha previsto un'analisi di brevetti[30]. A differenza delle interviste e della letteratura, i brevetti sono documenti fortemente strutturati, dove il recupero di informazioni dipende dai campi analizzati (richiedente del brevetto, descrizione delle procedure, materiali brevettati, data ecc.). Sul testo contenuto in questi campi sono state applicate le tecniche di TM a livello morfologico, statistico e semantico per individuare le unità lessicali utili per l'*analisi concettuale*. Quest'ultima, a differenza dell'analisi per parole chiave, permette un *mining* più esteso, in quanto non è vincolata alla presenza di uno specifico termine bensì fa riferimento ad una base di conoscenza (rete semantica, thesaurus ecc.). Dopo aver ultimato la fase di IR e aver selezionato i brevetti pertinenti, sono state effettuate su di essi alcune analisi statistiche di clusterizzazione. A tal fine, sono stati costruiti degli indicatori quali: il conteggio dei brevetti pertinenti rispetto a una query, il numero di citazioni di un particolare brevetto (confrontato con una media di riferimento), il ciclo di vita di un brevetto misurato in termini di durata di una generazione di brevetti.

La combinazione dei risultati di questi tre momenti del progetto di ricerca ha permesso una classificazione dei vari documenti estratti, la cui similarità è servita ad evidenziare i temi più presenti e/o la co-presenza di tecnologie diverse nell'ambito dello stesso argomento, ossia è servita a estrarre l'informazione d'interesse, utile a orientare le future politiche del settore dei trasporti.

Un secondo esempio di TM proviene dal campo biomedico. Cineca (www.cineca.it) ha analizzato circa 400.000 pubblicazioni medico-scientifiche riguardanti il ciclo di vita delle cellule (fonte PubMed: http://www.pubmed.com), con l'obiettivo di individuare automaticamente pattern di parole in grado di selezionare documenti secondo citazioni dirette di nomi di geni o frasi descrittive di concetti altamente correlati con essi.

Nelle fasi di preparazione dei documenti, un particolare rilievo ha assunto l'identificazione dei nomi dei geni (l'analisi grammaticale li identifica come nomi propri), che nella fattispecie costituisce una operazione di IE effettuata utilizzando solo un dizionario con i nomi ufficiali e gli *alias* dei geni.

Sui documenti rappresentati in forma matriciale (indicando la presenza/assenza dei termini di interesse) è stato applicato un algoritmo di clustering i cui risultati vengono utilizzati come base per le successive operazioni di mining. Queste ultime vengono svolte direttamente on-line (in un'area ad accesso riservato), per permettere ai ricercatori di selezionare i documenti di interesse. In risposta a una query per ciascun cluster di documenti viene visualizzata in forma di istogramma la co-presenza dei singoli geni all'interno di ciascun documento.

[30] L'analisi dei brevetti richiede notevoli risorse, poiché le banche dati di brevetti sono consultabili a pagamento, soprattutto per le sezioni contenenti le informazioni più significative.

6. CONCLUSIONI

Questa panoramica sulle caratteristiche dell'analisi automatica dei testi ha fornito alcuni scorci sulle concrete possibilità di estrarre informazione d'insieme da un corpus, nella moderna tradizione degli studi di analisi del contenuto o nelle recenti attività di text mining a fini aziendali.

L'ambiguità teorica del linguaggio è fortunatamente assai ridimensionata da un forte effetto "contesto", che in ogni applicazione circoscrive tutte le possibili accezioni di un termine spesso a una sola alternativa o poco più.

Il confronto con lessici di frequenza, attraverso misure di contrasto capaci di valutare il sovra/sotto-utilizzo dei termini, consente l'estrazione del linguaggio peculiare di un corpus. Lo studio delle associazioni di parole completa il riconoscimento del senso da attribuire alle parole chiave di un testo; l'uso di tecniche statistiche multidimensionali ne permette una rappresentazione complessiva, in grado di rivelare anche i principali assi semantici che sono alla base della variabilità linguistica dei testi investigati o il cosiddetto "imprinting" che caratterizza la tipologia del testo (scritto/parlato, formale/informale ecc.).

In sostanza, attraverso la scelta di una appropriata unità di analisi del testo (diversa a seconda degli scopi), è oggi possibile raggiungere un buon livello nella cattura dei significati presenti in un testo, senza necessariamente doverlo leggere o ascoltare.

Questa capacità, fondata essenzialmente sui continui progressi della linguistica computazionale, come si può intuire, è di grande rilevanza. Per esempio, essa aprirà definitivamente la strada a una traduzione automatica, se non "fedelissima", almeno essenziale, che tuttavia non può fare a meno dell'inventariazione di ampie e approfondite risorse linguistiche, quali sono le cosiddette basi di conoscenza (grammatiche locali, reti semantiche, ontologie o altro) e i dizionari elettronici multilingue (non solo di semplici lemmi, ma anche di locuzioni e forme composte).

Le prospettive dei metodi di analisi dei testi (testi ormai quasi tutti disponibili sul web) si indirizzano verso la pratica di un reale multilinguismo, che consenta lo studio simultaneo di informazioni espresse in differenti lingue su uno stesso argomento. Se ne cominciano a vedere alcune concrete applicazioni nell'ambito del *technology watch* e della patent analysis.

Bibliografia

[1] Beccaria G.L.: *Dizionario di Linguistica*. Einaudi, 1994, Torino.

[2] Benzécri J.P.: *Pratique de l'Analyse des Données, tome 3: Linguistique et Lexicologie*. Dunod, 1981, Paris.

[3] Benzécri J.P.: *Histoire et préhistoire de l'analyse des données*. Dunod, 1982, Paris.

[4] Bisceglia B., Rizzi A.: *Alcune analisi statistiche delle encicliche papali*. Libreria Editrice Vaticana, 2001. Città del Vaticano.

[5] Bolasco S., Lebart L., Salem, A.: JADT 1995 - Analisi statistica dei dati testuali. *CISU*, Roma, Vol. 2, 1995b.

[6] Bolasco S.: *L'analisi multidimensionale dei dati*. Carocci ed., Roma, 1999, p. 358.

[7] Bolasco S.: Analisi testuale dei messaggi nel sito FO, in M. Radiciotti (ed.) *La formazione on-line dei docenti Funzioni Obiettivo. Indagine qualitativa sugli esiti dei forum attivati dalla Biblioteca di Documentazione Pedagogica*. Franco Angeli, Milano, 2001.

[8] Brunet E.: *Le vocabulaire de Jean Giraudoux: structure et évolution*. Ed. Slatkine, 1978, Genève.

[9] Brunet E.: *Le vocabulaire de Victor Hugo*. Slatkine-Champion, 1988, Genève-Paris.

[10] Busa R.: *Index Thomisticus: Sancti Thomae Aquinatis operum omnium Indices et Concordantiae*. Frommann – Holzboog, Stuttgart, Vol. 56, 1974-1980.

[11] Busa R.: *Fondamenti di Informatica Linguistica*. Vita e pensiero, 1987, Milano.

[12] Cipriani R., Bolasco S.: *Ricerca qualitativa e computer*. F. Angeli, 1995, Milano.

[13] De Mauro T.: *Linguistica elementare*. Bari, Editori Laterza, 1998, p. 144.

[14] Elia A.: *Dizionari elettronici e applicazioni informatiche*.In: Bolasco S., et al. (Eds.), 1995, p. 55-66.

[15] Elia A.: *Per una disambiguazione semi-automatica di sintagmi composti: i dizionari elettronici lessico-grammaticali*. In: Cipriani R. e Bolasco S. (Eds.), 1995b.

[16] Elia A.: *Tecnologie dell'informazione e della comunicazione*. In: Gensini S., (ed.) Manuale della comunicazione, Carocci Ed., Roma, 1999, p. 248-257.

[17] Grigolli S., Maltese G., Mancini F.: *Un prototipo di lemmatizzatore automatico per la lingua italiana*, 1992. In: Cipriani R. e Bolasco S. (Eds.), 1995, p. 142-65.

[18] Guiraud P.: *Problèmes et méthodes de la Statistique linguistique*. Presses Universitaires de France, 1960, Paris.

[19] Herdan G.:. *Quantitative Linguistics*, London, Butterworths, 1964. (trad. it.: *Linguistica quantitativa*, Il Mulino, Bologna, 1971).

[20] Lebart L., Piron M., Steiner F.: *La sémiométrie. Essai de statistique structurale*. Dunod, 2003, Paris.

[21] Lebart L., Salem A.: *Statistique textuelle*. Dunod, 1994, Paris.

[22] Lebart L., Salem A., Berry L.: *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht-Boston-London, 1998.

[23] Mani I., Maybury M.T.: *Advances in Automatic Text Summarization*. The MIT Press, 2001, Cambridge (Mass).

[24] Muller Ch.: *Principes et méthodes de statistique lexicale*. Hachzette, 1977, Paris (ristampa: Champion, 1992).

[25] Poibeau T.: *Extraction Automatique d'Information: du texte brut au web semantique*. Hermes - Lavoisier, 2003, Paris.

[26] Salton G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[27] Sebastiani F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, Vol. 34, n. 1, 2002, p. 1-47.

[28] Silberztein M.: *Dictionnaires électroniques et analyse automatique de textes.Le système INTEX*. Masson, 1993, Paris.

[29] Tournier M.: *Bilan critique in AA.VV. En hommage à Ch. Muller: Méthodes quantitatives et informatiques dans l'étude des textes*. Slatkine Champion, Genève-Paris, 1986, p. 885-9.

[30] Zampolli A., Calzolari N.: *Problemi, metodi e prospettive nel trattamento del linguaggio naturale: l'evoluzione del concetto di risorse linguistiche*, 1992. In: Cipriani R. e Bolasco S. (Ed.), 1995, p. 51-65.

Sergio Bolasco, statistico, professore ordinario di Statistica insegna "Metodi esplorativi per l'analisi dei dati" presso la Facoltà di Economia dell'Università degli studi di Roma "La Sapienza"; ha diretto ricerche anche a livello internazionale sulle metodologie per lo studio automatico dei testi e sulle tecniche di *Text Mining*; è autore di un manuale di "Analisi multidimensionale dei dati".
sergio.bolasco@uniroma1.it

Bruno Bisceglia, ingegnere elettronico, sacerdote della Compagnia di Gesù, ha conseguito il dottorato in Teologia presso la Pontificia Università Gregoriana, da anni si dedica allo studio dei testi mediante microanalisi ermeneutica computerizzata, ed è professore incaricato presso la Facoltà di Ingegneria delle Telecomunicazioni dell'Università del Sannio in Benevento.
bisceglia@unigre.it

Francesco Baiocchi, statistico, tecnico di ricerca nella direzione censimento della popolazione dell'Istat, già consulente del Censis; esperto in sviluppo software, è uno degli autori di Taltac, libreria di programmi per il trattamento automatico lessico-testuale per l'analisi del contenuto.
baiocchi@istat.it