ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA
SEDE DI CESENA

FACOLTA' DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Scienze e Tecnologie Informatiche

PROGETTAZIONE E REALIZZAZIONE DI UN SISTEMA DI SOCIAL BUSINESS INTELLIGENCE BASATO SUL MOTORE SYNTHEMA

Relazione finale in
Data Mining

Docente:
Chiar.mo Prof.
Matteo Golfarelli

Presentata da:
Luca Gaia

Correlatore:
Dott. **Matteo Francia**

III Sessione
Anno Accademico 2011/2012

Dedico il raggiungimento di questo traguardo a mia madre e a mio padre, mia unica ragione di vita.

"L'esperienza è il tipo di insegnante più difficile. Prima ti fa l'esame, poi ti spiega la lezione"

Oscar Wilde

Indice

Elenco delle figure

Elenco delle figure

Elenco delle tabelle

Introduzione

Con l'esplosione dei *Social Media* e la conseguente moltiplicazione delle informazioni disponibili sul *Web* sotto forma di *User Generated Content*, la quantità di informazioni potenzialmente interessanti di natura destrutturata è cresciuta in modo più che esponenziale. In virtù di questo fatto, le aziende hanno incominciato a nutrire un sempre più ampio desiderio di attingere conoscenza ed informazioni di interesse per il proprio business a partire da questa nuova miniera di testi ed opinioni. Si è incominciata dunque a diffondere nel mondo IT l'esigenza di avere a disposizione un nuovo tipo di tecnologie; strumenti che permettessero di gestire in automatico un insieme esteso di testi non trattabili con tecniche tradizionali se non a costi insostenibili, e che consentissero di analizzare dati non strutturati, estrarre da questi le informazioni più rilevanti, classificarli sulla base dell'argomento trattato, risultando in questo modo di estremo aiuto ne processo decisionale dell'impresa. Nasce la *Social Business Intelligence* (SBI): un nuovo paradigma di business volto a fare dei contenuti destrutturati presenti sulla rete Internet una ulteriore fonte di informazione e conoscenza per l'impresa.

Introduzione

Il progetto di tesi trattato in questo documento è stato incentrato sulla progettazione e l'implementazione di un sistema di SBI realizzato per mezzo del motore semantico *SyNTHEMA*, la cui funzionalità principale è quella di consentire l'analisi linguista di documenti eseguita su base morfologica, sintattica, logico-funzionale e semantica. Il lavoro è stato incentrato sulle attività di recupero dal Web delle informazioni di interesse per il *case study* scelto, sulla loro successiva lavorazione ed analisi eseguita ad opera dal motore semantico (le cui risorse linguistiche sono state a loro volta oggetto di verticalizzazione orientata al dominio di lavoro), ed infine *mining* dei risultati offerti in output.

La tesi di articola in cinque capitoli (Introduzione e Conclusioni sono fuori numerazione). In ognuno di essi viene trattato approfonditamente un particolare aspetto dello sviluppo del sistema di Social BI. Il primo capitolo è dedicato all'*environment* socio-tecnologico e culturale la cui affermazione ed espansione su scala globale ha portato al cambiamento di una serie di consuetudini profondamente radicate nell'azienda come siamo abituati a conoscerla ed al sorgere di nuovi bisogni informativi ed organizzativi ai quali questo progetto si propone di dare risposta. Un paragrafo è riservato all'illustrazione dell'architettura funzionale alla base del sistema. Nel secondo capitolo, di carattere più tecnico rispetto al precedente, si è cercato di presentare nel modo più chiaro possibile i mezzi tecnici che hanno permesso la realizzazione del sistema di SBI e le loro funzionalità principali con le quali si ha avuto a che fare. Il documento prosegue poi con un capitolo in cui vengono illustrate le funzionalità della Social Business Intelligence messe a disposizione degli utilizzatori, più un'ampia parte dedicata all'illustrazione di come è stato costruito il database dedicato al progetto. Proseguendo, nel penultimo capitolo viene presentata la metodologia di verticalizzazione del sistema di SBI pensata ed adottata nel progetto, e di ogni suo passo sviscerati i punti fondamentali attraversati e le azioni intraprese per gestirli/eseguirli al meglio. La trattazione si conclude con una parte dedicata alla descrizione dei test e delle analisi effettuate, delle operazioni

eseguite per la verticalizzazione delle risorse linguistiche e delle considerazioni sui risultati offerti in output dal motore semantico.

Capitolo 1

La Social Business Intelligence

1.1 User Generated Contents: la nuova voce del Web

L'avvento dell'IT e la sua prepotente affermazione ha influenzato tanto usi e abitudini dei singoli, quanto quelli di mercati, imprese, organizzazioni: la rivoluzione digitale ha intensamente sconvolto e sta sconvolgendo tutt'ora il mercato dalle fondamenta. Di pari passo con il progresso informatico, i mercati moderni vanno evolvendosi di giorno in giorno, e tale trasformazione è stata soggetta nell'ultimo decennio ad una forte accelerazione dovuta agli *user generated contents* (UGC): termine che dal 2005 negli ambienti del *web publishing* e dei *new media* sta ad indicare il materiale disponibile sul Web prodotto da utenti invece che da società specializzate (foto e video digitali, blog, podcast, wiki; siti Web che si basano su questa filosofia sono *Flickr*, *Friends Reunited*, *FourDocs*, *OpenStreetMap*, *YouTube*, *Second Life* e *Wikipedia*) ("Contenuto," n.d.). La generazione degli UGC è iniziata con l'arrivo del Web 2.0, è proseguita con la diffusione di forum e blog ed ha raggiunto l'apice con la nascita dei social network: infatti la partecipazione di uno strabiliante numero di utenti a piattaforme come *Facebook*, *Google+*, Twitter e affini

ha fatto di essi un ottimo mezzo di propagazione di idee, notizie e informazioni. Le imprese hanno immediatamente colto le potenzialità di tutto ciò, di come questi social network potessero effettivamente essere sfruttati tanto per instradare i propri potenziali clienti attraverso *fanpage* dedicate, quanto per raccogliere opinioni e pareri espressi e sfruttare tali informazioni a proprio vantaggio come strategia di mercato vera e propria. In questo modo il Web assume un'accezione che permette alle aziende di ottenere profitti di grande entità (Phneah, 2012).

1.2 La transizione verso il Social Business Model

Per ottenere il massimo da queste condizioni, le imprese devono inevitabilmente ristrutturare il proprio modello di business estendendolo a un nuovo mercato caratterizzato da un inedito modo di comunicare, da nuove forme di competizione e da un nuovo tipo di consumatore. Tale modello è spesso denominato *Social Business Model* (SBM), ad accentuare come i processi aziendali siano innescati o comunque ispirati dal comportamento degli utenti del Web che si cattura e si influenza esaminando e producendo UGC.

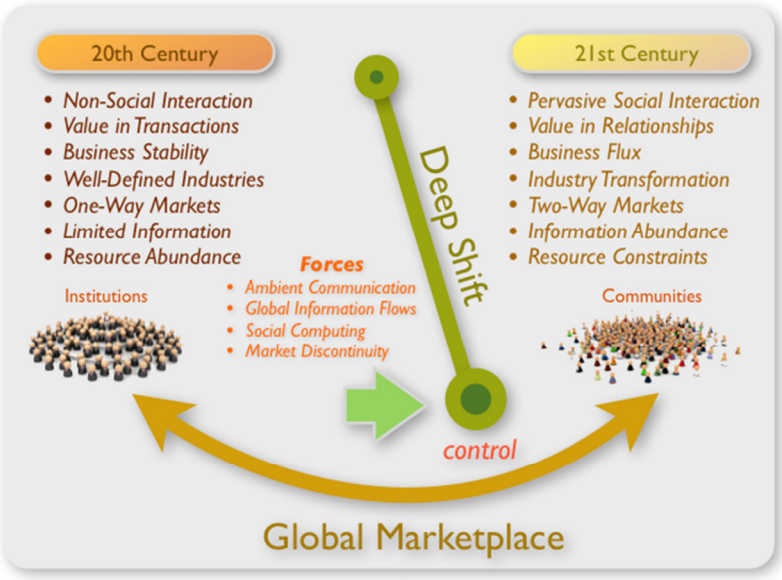

Fig 1.1: La transizione verso il Social Business Model (Hinchcliffe, 2010)

Le informazioni di rilievo che da questi contenuti destrutturati che possono essere ricavate sono tanto fondamentali quanto eterogenee: possono concernere la comprensione del contesto (il cosiddetto "parlato"), all'interno del quale l'impresa agisce, l'individuazione delle *best practice* di settore, l'identificazione degli *unmet needs* dei propri consumer e dei consumer in generale, le occasioni che in specifiche circostanze possono venirsi a creare e le possibili criticità che l'impresa deve essere in grado di amministrare. Le ripercussioni che l'ottenimento ed il giusto utilizzo di questi dati vi sono sul *core business* dell'azienda sono di fondamentale rilievo, come indicato in Figura 1.2, nella quale sono rappresentate le principali attività che fanno capo al SBM e l'impatto che esse hanno sulle spese dell'impresa. I pro che conseguono dall'uso dei social media si raggruppano sostanzialmente in due aspetti: l'incremento

- della velocità di accesso al *know-how*
- della velocità di identificazione del personale con il giusto *background* di conoscenze e maturità
- della collaborazione
- della velocità di assistenza al cliente e conseguentemente a ciò l'aumento nel cliente di soddisfazione e *loyalty*

e l'abbattimento di un insieme di costi a cui l'impresa deve far fronte, come i costi di viaggio, di comunicazione interna ed esterna.

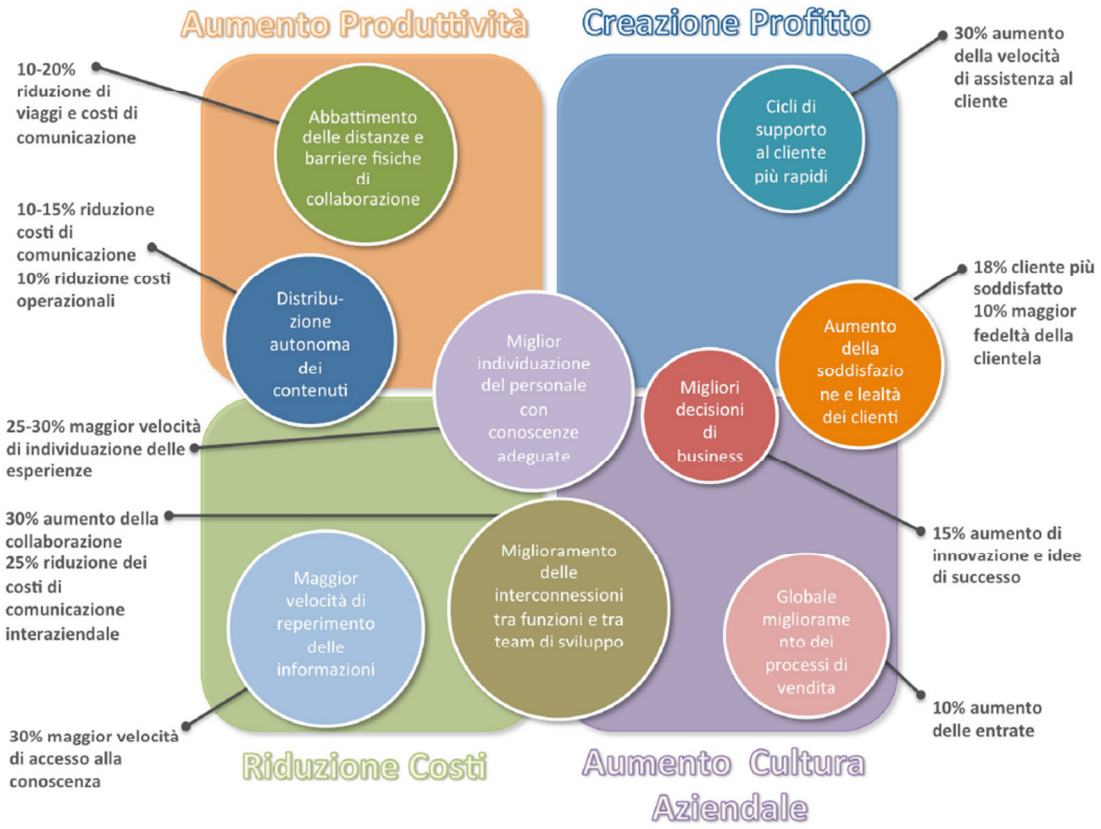

Fig 1.2: I vantaggi di una corretta gestione ed interpretazione dei dati derivanti dai processi di SBM

La facoltà di poter raccogliere le opinioni espresse in rete da ogni consumer porta si a vantaggi, ma anche ad oneri organizzativi e tecnologici a cui l'impresa deve far fronte. Organizzativamente parlando, il Social Business interseca tutti i processi aziendali e i vari livelli decisionali, richiedendo di conseguenza un mutamento della cultura aziendale ed una crescente attenzione ai fenomeni sociali in tutte le fasi del business: sviluppo di un servizio o prodotto, reclutamento delle risorse umane, elaborazione di campagne e strategie di marketing con la derivante stima della loro efficacia, infine la gestione delle criticità inattese ove si presentassero. Un aggiuntivo aspetto positivo del SBM è senz'altro il breve intervallo temporale che intercorre tra l'immissione dell'informazione sul Web e il suo raccoglimento (*information retrieval*) da parte dell'impresa; questo è senz'altro un punto di forza se paragonato alle tradizionali metodologie di analisi del mercato, ma allo stesso tempo obbliga

l'azienda ad avere dei tempi di reazione fulminei al fine di non perdere il vantaggio competitivo accumulato.

1.3 I Social Media e l'esplosione della quantità di dati: nascono i Big Data

I dati assumono un'importanza strategica fondamentale in un'economia come quella del mondo moderno in cui la conoscenza è alla base del valore creato nei mercati. Diventa quindi fondamentale avere la possibilità di trasformare il più velocemente possibile il mare di dati, di informazioni di cui si dispone, in conoscenza sulla base della quale prendere decisioni ed interpretare trend di mercato.

Tecnologicamente parlando, il SBM richiede l'archiviazione e l'analisi di grandi aggregazioni di dati destrutturati: i cosiddetti Big Data, la cui grandezza e complessità richiede strumenti più avanzati rispetto a quelli tradizionali, in tutte le fasi del processo (dalla gestione, alla *curation*, passando per condivisione, analisi e visualizzazione). Il progressivo aumento della dimensione dei data set è legato alla necessità di analisi su un unico insieme di dati correlati rispetto a quelle che si potrebbero ottenere analizzando piccole serie con la stessa quantità totale di dati ottenendo informazioni che non si sarebbero potute ottenere sulle piccole serie.

Big Data rappresenta anche l'interrelazione di dati provenienti potenzialmente da fonti completamente differenti. Non solo quindi dalle fonti tradizionali sino ad oggi concepite ed utilizzate ma anche attraverso l'impiego di informazioni provenienti dai Social Network come Facebook e Twitter e da qualsiasi forma di informazione collaterale che può incidere sui consumi o sulle abitudini. L'insieme di tutti questi dati, sia di origine convenzionale che di origine social e statistica generano quel che si chiama Big Data consentendo a chi li analizza di ottenere una plusvalenza legata ad analisi più complete che sfiorano anche gli "umori" dei mercati

e del commercio e quindi del *trend* complessivo della società e del fiume di informazioni che viaggiano e transitano attraverso internet. Con i Big Data si arriva a parlare di *Zetta-Byte* ovvero di una mole di Byte dell'ordine di 10^{21} e quindi di miliardi di Terabyte (già 10^{12}) ("Big," n.d.). La gestione dei Big Data non è per nulla semplice, e richiede lo svolgimento accurato delle seguenti attività:

- estrazione da fonti eterogenee (social network, blog, forum, portali web, *repository* documentali interni, ecc.)
- *tuning* e memorizzazione
- analisi, che concerne
  - estrazione di nuovi *pattern*
  - riconoscimento di trend già noti
  - enucleazione di informazioni rilevanti come nomi, entità di interesse, relazioni tra entità

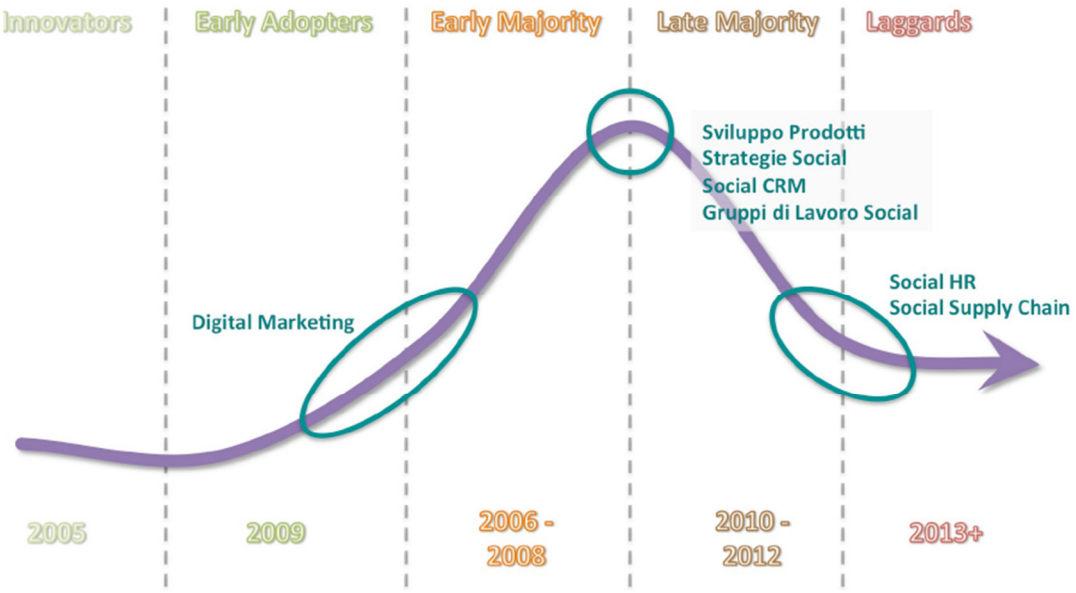

Fig 1.3: Curva di adozione del Social Business negli USA

In Figura 1.3 illustrata nella pagina seguente viene riportato il trend di adozione delle soluzioni di Social Business negli USA. Tale rappresentazione mostra come il

mercato statunitense sia avanti rispetto a quello italiano. Vantaggio che trova spiegazione nell'esistenza consolidata di un numero maggiore di soluzioni di mercato ed in una più grande consapevolezza e interesse del management d'impresa verso i fenomeni social.

1.4 Abbracciare il Social Business ed i sistemi di Social BI

Come già evidenziato il Social Business tange numerosi processi aziendali, per questa ragione l'approccio più naturale nell'abbracciare questo genere di tecniche avviene di consueto per step successivi in un iter di tipo *bottom-up*. L'inizio del percorso di adozione combacia frequentemente con quel settore aziendale che generalmente riesce a convertire in modo celere lo sfruttamento dei social media in valore competitivo sul mercato: il marketing. Questo spiega la grande propagazione nel mercato nazionale di quel gruppo di attività svolte dalla divisione Marketing incentrate sull'interazione con il cliente e la sua fidelizzazione attraverso gli strumenti di comunicazione propri del Web: il Digital Marketing (DM). Nelle prime fasi l'adozione è contraddistinta da una tendenza alla parcellizzazione delle applicazioni e dell'informazione social in silos; con il passare del tempo e con l'utilizzo crescente dei social media nell'amministrazione dei processi aziendali, la tendenza sembra essere quella di abbracciare, anche per i dati social, un approccio di distribuzione e di utilizzo tipico dei dati nell'enterprise tradizionale (Hinchcliffe, 2013). Si parla in questo caso di **Social Business Intelligence**, ossia di un'architettura software che consenta di avere un quadro integrato dei dati social che diventa conoscenza usufruibile a tutti i livelli aziendali e che procura un maggiore ausilio alle attività decisionali proprie di ogni funzione. Obiettivo fondamentale di un sistema di SBI è inoltre quello di unire i dati social con i dati strutturati aziendali (*data warehouse* aziendale) per consentire di rilevare la misura di correlazione

esistente tra i fenomeni che si verificano in ambiente social ed i fenomeni del "mondo reale" (Souza, 2012) (Savitz, 2012).

Ad oggi, un argomento piuttosto scottante è il modo in cui gli effetti del Social Business influenzano processi e performance aziendali. Tale connessione dipende senza dubbio alcuno dal settore commerciale di riferimento poiché direttamente da esso è determinata la maggiore o minore percentuale di web customer presenti rispetto al totale dei clienti di un'impresa. Non si può tuttavia non tenere conto di come in alcune eventualità questa influenza possa essere fortemente influenzata dal verificarsi o meno di singoli eventi come campagne pubblicitarie persuasive o fallimentari, la scarsa o eccessiva ricettività del mercato dovuto a particolari condizioni storiche. Nel panorama italiano lo sviluppo dei sistemi di SBI si trova in uno stato molto arretrato e di rado si trovano aziende che abbiano avviato attività di estensione del sistema interno di Business Intelligence con un modulo Social. Sul mercato USA cominciano invece a venire a galla le prime conferme che le imprese più consolidate nel settore del Social Business, essendo riuscite già in parte ad integrare le nuove attività nel sistema Business Intelligence interno, riescono ad estrapolare informazioni spesso importanti, a volte fondamentali, per il Business Process Management (Pearson, 2013).

1.5 Il Digital Marketing

Il Digital Marketing è quel particolare tipo di marketing che fa uso di fonti digitali per entrare in contatto con i consumer ed altri partner di business ("Digital," n.d.). Nel mondo reale la divisione di Digital Marketing gestisce e controlla le attività di marketing sui canali digitali. La recente diffusione dei social network è solo l'ultima sorgente informativa aggiunta, in ordine cronologico, ad una funzione che già si serviva di strumenti come la presenza sui portali web, messaggi diretti e mailing list per massimizzare la visibilità e l'efficacia delle proprie campagne pubblicitarie. Il

ciclo operativo del DM non si diversifica da quello del marketing consueto, ma aggiunge alle attività da questa già svolte, una serie di task specifici (in nero in Figura 1.4). L'utilizzo del canale social amplia ulteriormente l'insieme di funzioni eseguite dalla divisione di DM (in blu in Figura 1.4). Con la divisione Digital l'attività di marketing nel suo insieme rinforza la sua figura di punto di contatto azienda-consumatore nell'ottica di un rapporto comunicativo bidirezionale in cui l'approccio comunicativo tradizionale viene ribaltato a favore di un modello basato sull'interazione.

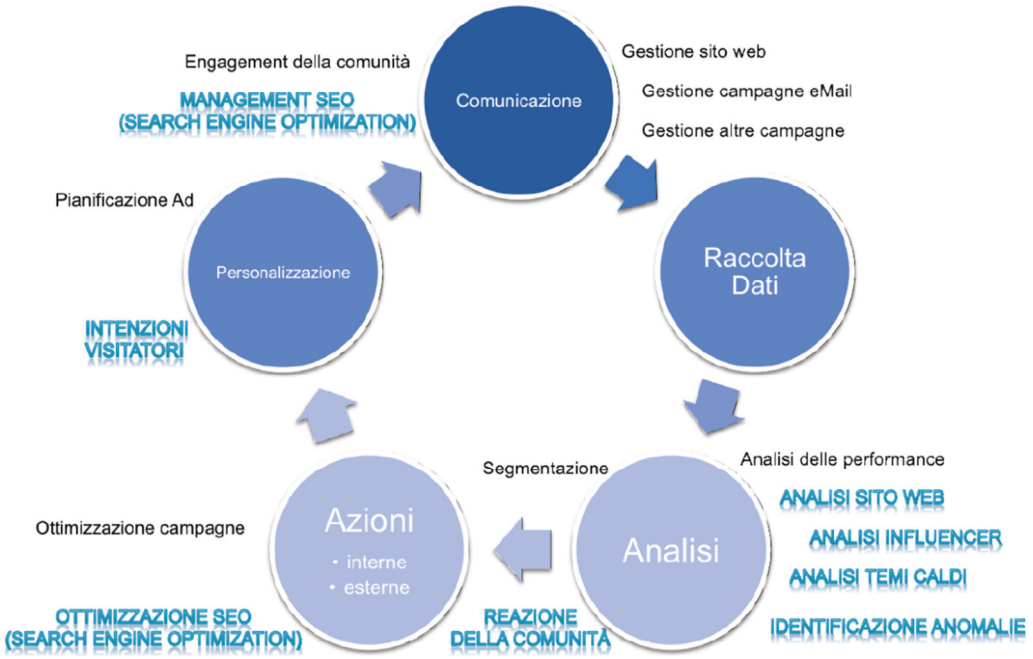

Fig 1.4: Il *task breakdown* del Digital Marketing (Malloy, 2012)

La nuova ottica sotto cui il Marketing viene ad visto nel panorama del Social Business va al di là della mera gestione delle campagne pubblicitarie d'impresa divenendo di fatto il principale collettore di dati, per un'azienda che si trova nella prima fase di adozione discussa in precedenza (vedi Figura 1.3). Quanto spiegato colloca il Marketing in una posizione privilegiata dalla quale cominciare lo sviluppo e la pianificazione di strategie ed architetture di SBI. Si evidenzia che la SBI non combacia con il DM, ma ne supporta le fasi di raccolta dati e analisi: a valle del

percorso di "comunicazione", in cui l'azienda gioca un ruolo di primaria importanza nella creazione di stimoli verso web customer, vi è una fase di "raccolta dati" molto più articolata e fertile rispetto a quanto si sia potuto operare fino ad oggi. Durante questo processo l'azienda gioca invece un ruolo passivo di ascolto del web recuperando le informazioni concernenti i feedback relativi alle attività svolte, i fenomeni collegati al dominio di ascolto che accadono indipendentemente da esse nel Web nonché altri dati, al di fuori del dominio di ascolto ma comunque rilevanti per il sistema azienda (i cosiddetti *ultraviolet data*) (Phneah, 2012).

1.6 Text Mining e Sentiment Analysis

La Sentiment Analysis, od Opinion Mining, può essere considerata come l'applicazione delle tecniche proprie della Text Analytics: dal *Natural Language Processing* (NLP) (affrontato in dettaglio nel paragrafo 3.1.3) alla linguistica computazionale, per l'identificazione sistematica, l'estrazione e l'elaborazione (tipicamente la valutazione) di informazioni soggettive dai documenti sorgenti dell'analisi. Essendo i documenti sorgente Big Data provenienti prevalentemente dal Web, possono essere di vario tipo ed avere contenuti assai differenti tra loro: dal testo ai contenuti audio visivi passando per le immagini (Grimes, 2012). Poiché quella della classificazione dell'opinione è un'attività che lascia ampio spazio all'interpretazione personale, valutare le prestazioni un sistema di Social BI che sfrutta tecniche di Sentiment Analysis è un'opera quanto mai ardua. Si è deciso perciò di considerare una misura della bontà delle prestazioni del sistema, individuata nella percentuale di volte in cui la classificazione del sentimento coincide con il giudizio dato da una persona. Attualmente la Sentiment Analysis è un argomento scottante nell' ambito della ricerca, e sul mercato si trovano già soluzioni industriali. È però d'obbligo chiarire che si sta parlando di un ambito fortissimamente legato alla lingua del testo che si analizza. Ciò ne ha ritardato la diffusione sul mercato italiano e per questo le soluzioni impiegate oggi su larga scala

presentano un livello di maturità sicuramente inferiore a quelle in uso nei paesi anglofoni. Essendo un argomento che spazia dall'informatica alla linguistica toccando tutti gli aspetti dell'analisi del linguaggio naturale, la Sentiment Analysis è caratterizzata da una complessità molto elevata e questo ha fatto si che nel passato si svolgesse sull'argomento relativamente poca attività di ricerca (Liu, 2012). Oggi l'attività ha conosciuto un rinnovato impulso grazie ai progressi effettuati, alla consapevolezza delle imprese che le opinioni influenzano fortemente il comportamento dei consumatori e non ultimo grazie ai social media che, come esposto precedentemente, offrono un canale di accesso facilitato a grandi quantità di informazioni ed opinioni esposte direttamente dai consumatori (UGC). Risulta ormai chiaro come l'Opinion Mining o Sentiment Analysis, specializzazione del Text Mining a cui ci si riferisce oggi anche con denominazioni quali *Sentiment Mining*, *Subjectivity Analysis*, *Emotion Detection* ed O*pinion Spam Detection* (Liu, 2012), abbia l'obiettivo di elaborare un qualsiasi documento testuale, o corpus di documenti, ed individuare al loro interno tutte le opinioni presenti. Per poter valutare il raggiungimento di tale risultato è però necessario definire il concetto di "opinione". In generale un'opinione è un giudizio soggettivo, individuale o collettivo, su di un qualcosa o riguardo a qualcuno. In altri termini possiamo dire che un'opinione sia la manifestazione di un sentimento positivo o negativo espresso su di un'entità o su un aspetto di un'entità, da un *opinion holder* (colui che formula il giudizio). L'opinione comporta quindi l'espressione di un giudizio, giudizio che può essere positivo, negativo o neutrale (in questo caso non esiste un'opinione) e determina quindi quella che in gergo viene chiamata polarizzazione dell'opinione (o del sentimento).

Una grande sfida per i software di Opinion Mining è rappresentata dal sarcasmo e dall'ironia. Queste sono caratteristiche squisitamente umane alle quali persino le persone reagiscono in modo differente in base alla propria sensibilità personale ed al proprio senso dell'umorismo: in virtù di questo fatto è molto difficile che una macchina possa darne un'interpretazione universalmente corretta ("Ironìa," n.d.). Per quanto la ricerca sia attiva in questo campo, in merito a questa difficile

sfida che la Sentiment Analysis si trova ad affrontare ogni volta che un testo viene processato si è ancora lontani da una soluzione accettabile. Per questa ragione, quello che fino ad ora si è rivelato l'approccio migliore all'Opinion Mining su larga scala dal punto di vista dei risultati è senza dubbio quello ibrido, dove per ibrido intendiamo quell'approccio in cui il lavoro della macchina viene controllato e corretto dall'essere umano, dove i risultati dell'analisi automatizzata svolta dall'applicativo vengono visionati ed il comportamento dell'applicativo stesso modificato al fine di ottenere performance migliori. Complessivamente, afferma Seth Grimes, un tool *general purpose* con il quale non è stato realizzato nessun training o *tuning* specifico, classifica l'opinione rispetto all'interpretazione di un essere umano, con una precisione che si aggira intorno al 50% (Grimes, 2011). Se da un lato questo dato mette in luce come debba ancora essere fatta molta strada, evidenzia altresì quanto fondamentale sia eseguire sull'applicativo un training specifico relativo all'ambito di utilizzo e dotare lo strumento di mining di dizionari e basi d conoscenza proprie del *vertical* in cui questo verrà impiegato.

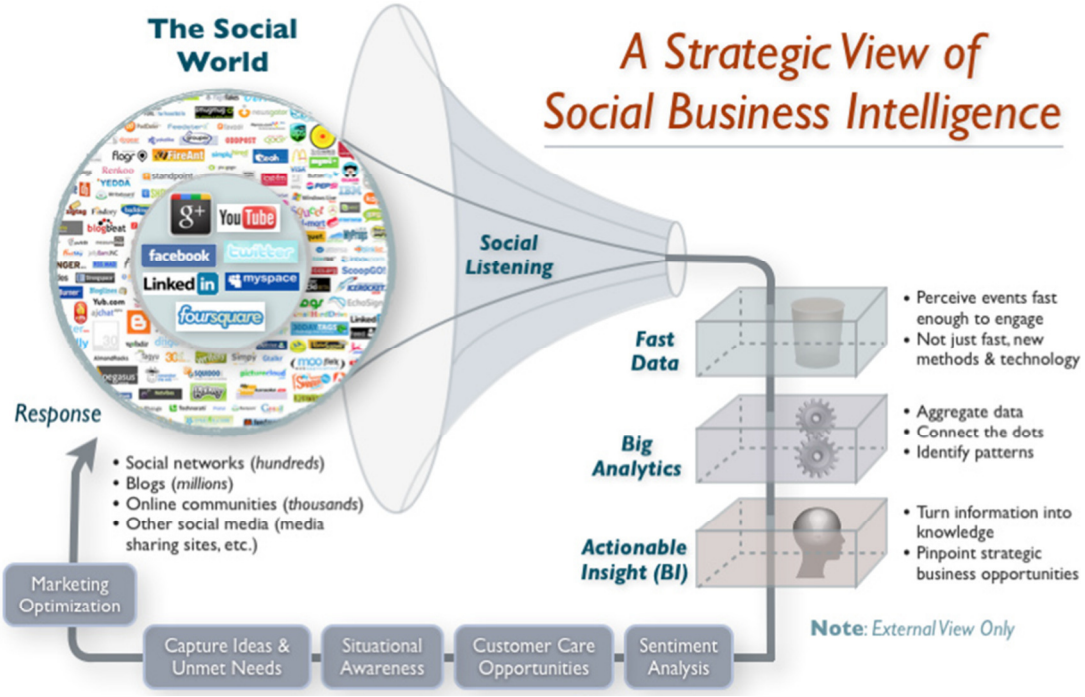

Fig 1.5: Ascoltare, analizzare, rispondere: il virtuoso ciclo del Social Business - *http://goo.gl/i77ZZ*

1.7 Architettura funzionale del sistema di Social BI

Quella che andremo a descrivere nel seguito è l'architettura funzionale del sistema di Social BI oggetto della tesi, ovvero una *overview* dei moduli che sono stati progettati ed implementati perché si riuscisse, partendo da informazioni destrutturate presenti sul Web, ad ottenere informazioni strutturate ed adeguatamente organizzate in modo tale da poter essere oggetto di analisi, studio e statistiche di varia natura.

Per riuscire a capire come deve essere costruita una soluzione, il primo passo da fare è comprendere quali sono i bisogni informativi del "cliente" ed a quali domande questo si aspetta di riuscire a dare una risposta. Lo strumento di cui noi avevamo bisogno era un mezzo tramite il quale poter ascoltare ciò che si dice in rete in merito all'ambito della politica, comprendendolo e sapendolo interpretare correttamente. Le possibilità che uno strumento del genere fornisce sono moltissime: permette, per esempio, non solo di sapere che all'ipotetico elettore piace o non piace il lavoro svolto da un certo politico e/o partito ma può portare a conoscenza di quali sono gli aspetti che nel dettaglio sono più o meno apprezzati. Poter ascoltare la rete permette quindi di entrare in possesso di informazioni di rilievo relative all'opinione che media ed elettori si sono fatti o si stanno facendo in merito ad un certo politico, partito o fatti recentemente accaduti che vedono coinvolti personaggi della sfera governativa.

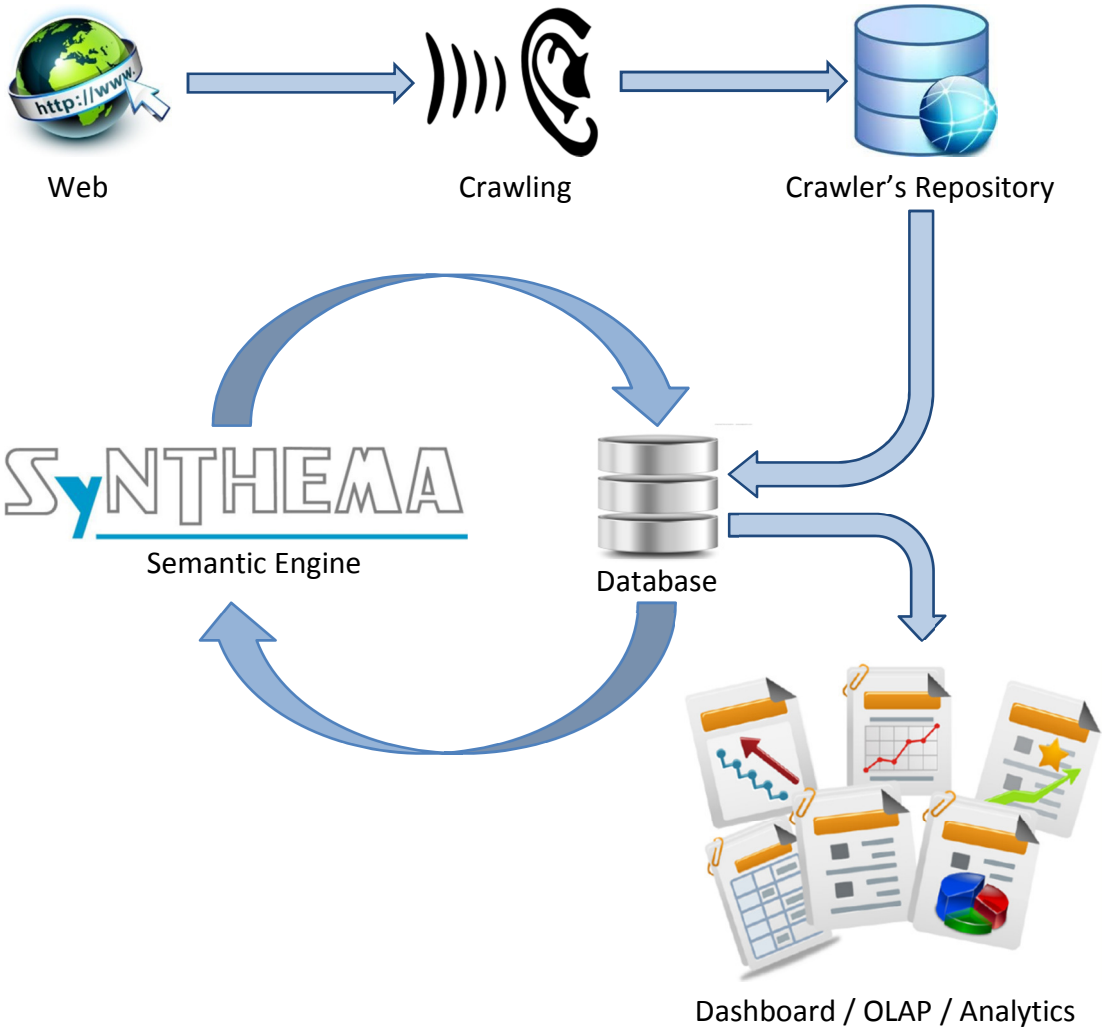

Fig 1.6: Architettura funzionale del sistema di Social BI

L'architettura funzionale illustrata in Figura 1.6 può essere scomposta in 4 macro-processi:

1. **Crawling** All'interno di questo processo vengono racchiuse le attività volte al recupero dal web di quei testi in cui sono contenute informazioni interessanti per il case study sulla politica. La fase di *collecting* di queste informazioni avviene tipicamente grazie a dei meccanismi che effettuano un *crawling* sistematico della rete recuperando informazioni destrutturate grazie a delle ricerche basate su degli insiemi di *keyword*.

2. **Popolamento Database** Si occupa di trasferire le clip raccolte in fase di crawling sul database predisposto per il sistema di Social BI attraverso l'utilizzo del modulo di *data integration* fornito dalla suite di prodotti *Talend Open Studio*.
3. **Autoanello di analisi** Questa fase è senza dubbio quella che mi ha impegnato maggiormente nel corso di questo lavoro di tesi. Si occupa, con l'ausilio del modulo di data integration, di prelevare le clip dal database e trasferirle sul motore semantico SyNTHEMA, che si prenderà carico del compito di eseguire sulle clip una approfondita analisi atta a scovare temi e topic di interesse, le relazioni che li legano e le opinioni su di essi espresse (positive, negative o neutrali). Terminata l'analisi viene compiuta una parcellizzazione della clip in frasi e delle frasi in gruppi di parole; infine si riscrive, in forma strutturata, il risultato completo di questa analisi nelle corrispondenti tabelle del database.
4. **Dashboarding & Reporting** Qui ci si occupa della rielaborazione dell'informazione strutturata, output del passaggio precedente, utilizzandola per fornire (come output finale dell'elaborazione) delle dashboard ed in generale un servizio di reportistica interattiva creata specificatamente sulle esigenze dell'utenza di business di questo servizio.

Capitolo 2

Tecnologie

2.1 Tecnologie per la ricerca di informazioni non strutturate

A partire dalla metà degli anni ‘80, con la disponibilità di maggiore capacità di calcolo e la prepotente affermazione delle reti aziendali, si è sentita l’esigenza di strumenti che consentissero un accesso semplice ed intuitivo alle informazioni, sfruttando la tempestività delle risorse e delle conoscenze disponibili per l’azienda. Risultati sostanziali sono stati raggiunti solo in situazioni dove era possibile poter organizzare le informazioni in dati, memorizzandole in database relazionali. Per le informazioni disponibili in forma non strutturata, che mediamente rappresentano l’80% di tutte le informazioni aziendali rilevanti (Plummer & Gartner, 2006), gli sforzi intrapresi hanno prodotto solo risultati parziali. Questo a causa di problemi sia qualitativi, collegati alla difficoltà di automatizzare il recupero e la gestione delle informazioni, sia quantitativi, dipendenti dal volume documentale sempre crescente. Il numero di documenti disponibili in formato elettronico è cresciuto nel tempo in

modo quasi esponenziale, mentre la nostra capacità di lettura e di analisi è rimasta praticamente immutata nel tempo. I motori di ricerca aziendali, invece di semplificare il recupero delle informazioni, lo hanno reso se possibile più complesso, restituendo alle nostre interrogazioni lunghe liste di documenti, di cui spesso non è chiara né la pertinenza, né la rilevanza.

2.1.1 Aumentare l'efficacia aziendale migliorando la capacità di ricerca

In generale, l'inefficienza della gestione della conoscenza e delle informazioni in un'azienda si riverbera sempre sulla sua produttività ed efficacia sul mercato. La necessità di integrare i dati con informazioni non strutturate, nascoste e quindi spesso prive di valore, perché non facilmente accessibili o collegate ad un processo specifico, complica non poco la *governance* aziendale. L'informazione più rilevante si trova spesso "codificata" all'interno di testi ed è possibile coglierla solo "leggendo tra le righe". La maggior parte dei documenti elettronici è disponibile sul Web, nelle reti locali aziendali o dipartimentali. I documenti sono scritti in formati diversi tra loro, in lingue diverse, per utenze diverse. Diverse per interessi e competenze sono le persone che li leggono. E-mail, telefonate o pagine Web (trasformate in conoscenza) possono costituire una fonte inesauribile, continuamente aggiornata, su cui poter basare le proprie decisioni, formulare ipotesi in politica, come in economia. Tuttavia, mediamente, il 50% del tempo dedicato ad attività di trattamento delle informazioni è assorbito dalla ricerca e dalla consultazione di documenti, il 10% è legato ad attività di ricerca infruttuose, un altro 20% alla riscrittura totale o parziale di testi.

2. 1.2 Il Knowledge Mining: una tecnologia al servizio delle imprese efficienti

L'idea è quella di considerare questo insieme esteso di testi come una miniera; esplorarla con l'obiettivo di trovare l'informazione di interesse e di scoprire eventuali relazioni a prima vista nascoste o inaspettate, e che constituiscono la vera vena d'oro. Il *Knowledge Mining*:

- È una tecnologia linguistica e matematica che permette di gestire in automatico un insieme estesi di testi, non trattabili con tecniche tradizionali se non a costi insostenibili
- Permette di analizzare testi non strutturati, estrarre da questi le informazioni più rilevanti, classificarli sulla base dell'argomento trattato, risultando in questo modo di estremo aiuto ne processo decisionale

Nella gestione delle informazioni, il Knowledge Mining è la risposta al sovraccarico cognitivo, meglio conosciuto come *Information Overloading*. Esso si verifica quando si ricevono troppe informazioni per riuscire a prendere una decisione o sceglierne una specifica sulla quale focalizzare l'attenzione. Lo sviluppo della tecnologia ha contribuito alla diffusione e alla riconoscibilità di questo fenomeno. La grande quantità di informazioni che si ottengono con un'interfaccia mal progettata (poco ergonomica) o su siti Internet egualmente scadenti, possono inibire la capacità di scremarle. Ad esempio nel caso della Internet dipendenza vi sono soggetti che passando in continuazione da un sito web all'altro, non riescono a fermarsi né a ricordare le informazioni ricevute, poiché viene percepito tutto come rumore (in termini cognitivi) (Lavenia, 2007). Tornando al Knowledge Mining, questa nuova metodologia di analisi prevede due fasi che si succedono in automatico l'una all'altra (Fig. 2.1). L'analisi linguistica permette di cogliere per ciascun testo i suoi concetti chiave, operando elaborazioni di tipo morfologico, sintattico, logico-funzionale e semantico.

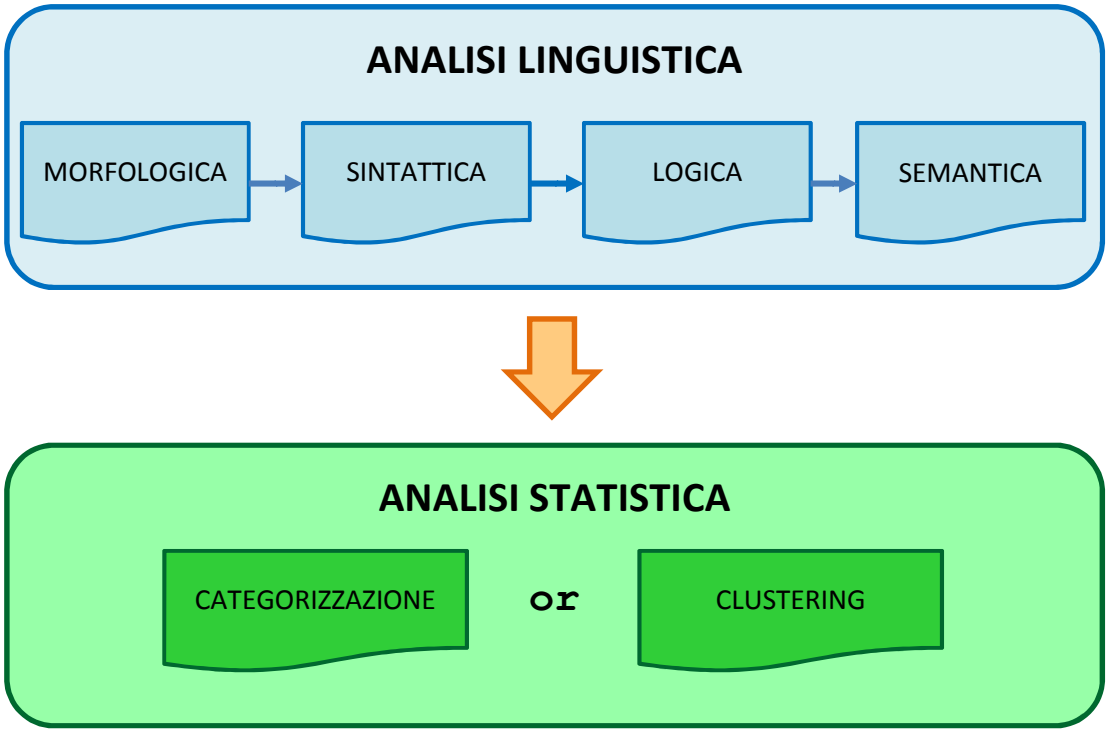

Fig 2.1: Le due fasi di analisi del Knowledge Mining

L'analisi morfo-sintattica permette di rimuovere le eventuali ambiguità presenti nel testo, classificando ogni parola da un punto di vista grammaticale, riconducendola alla sua entrata nel dizionario. Nel caso di un aggettivo, ad esempio, la parola viene ricondotta con l'analisi alla sua forma maschile singolare; nel caso di un verbo, questo viene ricondotto alla forma infinita. L'analisi morfo-sintattica riduce sensibilmente il numero di possibili concetti candidati a descrivere un testo: in un testo di politica o di economia, ad esempio, i nomi propri di persona, luogo o organizzazione permettono una facile identificazione del tema trattato; in una mail ricevuta da un cliente o in un post di un blog o di un forum, invece, gli aggettivi possono connotare o graduare la valenza positiva o negativa di un servizio. L'analisi logica consente poi l'identificazione automatica del ruolo funzionale svolto nel testo da ciascuna parola (soggetto, complemento oggetto, complemento di luogo, di tempo o di modo, ecc.), permettendo di capire chi fa cosa, come, quando e dove. L'analisi semantica coglie inoltre il significato profondo di ogni parola. Proiettare queste

informazioni nel tempo, permette di valutare con efficacia il grado di soddisfazione dei clienti di un'azienda a fronte di sue iniziative commerciali. Collocare il nome di un'azienda nello "spazio delle parole", valutarne la distanza da concetti come buono, bello, giovane, rassicurante, ecc., permette di cogliere la percezione che di lei hanno i clienti o i potenziali fruitori di un suo servizio (Semiometria e *Brand Analysis*). Se l'analisi linguistica si presenta più facile per l'inglese, lingua strutturalmente più semplice rispetto alle lingue latine, questo non avviene certo per lo spagnolo, il portoghese, l'italiano o il francese, lingue cariche di rimandi letterali. Poche sono le aziende o i centri di eccellenza in grado di operare efficientemente in questo settore. L'analisi statistica assegna poi i documenti "concettualizzati" sia a categorie tematiche predefinite e personalizzate (categorizzazione), che non note a priori (*clustering*). Nel caso della categorizzazione, ad esempio, l'analisi può associare il lancio di un'agenzia di stampa a temi come politica, cronaca, esteri, economia, sport e spettacolo; oppure ancora un'e-mail, genericamente inviata all'azienda, al reparto (commerciale, tecnico) più idoneo al suo tempestivo trattamento. Nel caso del clustering, la classificazione dei testi avviene mediante loro aggregazione spontanea, secondo uno schema di classificazione non noto a priori. Questo permette di aggregare lamentele o suggerimenti riguardo a prodotti e servizi secondo prospettive diverse, fornendo nuove chiavi di lettura per i dati dell'azienda.

Nei portali di ricerca Knowledge Mining, la ricerca delle informazioni può essere effettuata non solo tramite parole chiave ma anche in linguaggio naturale, semplificando il linguaggio di interrogazione. Il processo di disambiguazione lessicale e semantica permette di perfezionare il processo di ricerca, ottenendo migliori risultati di un normale motore di ricerca. Queste applicazioni sono in grado di dare risposte precise e puntuali alle richieste formulate; possono dare risposte multilingua, trovando risposte disponibili in un un'altra lingua mediante una sofisticata tecnologia di traduzione automatica estremamente precisa e puntuale.

2.1.3 Il Natural Language Processing

Per riuscire ad ottenere un'analisi automatizzata del documento che sia robusta e sicura, la soluzione è quella di eseguire un esame approfondito del significato del testo stesso. Questa considerazione ha portato ad una fusione delle tecniche matematiche e statistiche proprie del Text Data Mining con le tecniche di analisi linguistica sviluppatesi nell'ambito del NLP (Natural Language Processing). Il NLP è una disciplina da cui sono nate una serie di tecnologie il cui scopo è quello mettere un calcolatore nella posizione di elaborare correttamente testi scritti in linguaggio naturale, ovvero il linguaggio usato quotidianamente dalle persone. Diretta conseguenza dell'acquisizione di questa capacità è la possibilità di identificare le informazioni rilevanti contenute nei blocchi di testo, dove per informazioni rilevanti si intende quella parte del testo pertinente rispetto a specifici interessi (Bolasco, 2005). L'obiettivo ultimo del NLP è quindi quello di permettere ad una macchina di comprendere a fondo quanto espresso da una persona, conseguire quindi quella che in gergo è chiamata *Natural Language Understanding* (NLU), la comprensione del linguaggio naturale.

Una delle fasi fondamentali eseguite dal NLP è il *Part-Of-Speech Tagging* (POS Tagging), che riguarda l'analisi lessicale. Con questo termine si indica il procedimento tramite il quale vengono riconosciuti gli elementi grammaticali di un periodo ed ad ognuno viene assegnato un identificatore chiamato appunto tag che riporta informazioni sul token (parola) al quale è associato, per questo si può considerare il POS Tagging un procedimento di classificazione. Un esempio di tagging potrebbe essere il seguente:

- Luca Gaia, Tag : R
- quaderno, Tag: N
- scrivere, Tag: V
- precipitosamente, Tag: A

- fantastico, Tag: G

Dove R sta per "nome proprio", N per "nome" (*noun*), V per "verbo" (*verb*), A per "avverbio" (*adverb*), G per "aggettivo" (*grader*) e U per "sconosciuto" (*unknown*). Questo procedimento rende possibile una rapida selezioni delle parti di testo più utili al fine di un'analisi (Robin, 2009). Il livello successivo, quello sintattico, è il livello nel quale la frase viene vista dai software di NLP come un insieme di parole delle quali è necessario individuare in modo esatto tutte le relazioni al fine di estrapolare la struttura del periodo e poterne esplicitare il significato. Questo è di fondamentale importanza poiché generare correttamente l'albero sintattico della frase, cioè individuare correttamente le dipendenze tra tutti gli elementi che la compongono, è necessario per risolvere l'analisi delle figure retoriche proprie del linguaggio. Inoltre è necessario saper classificare le diversità tra due frasi come "il poliziotto insegue il ladro" e "Il ladro insegue il poliziotto" che sono morfologicamente e lessicalmente identiche ma nelle quali la posizione delle varie parole porta a grosse differenze relazionali tra i vocaboli. L'ultimo livello per noi rilevante, quello semantico, rappresenta lo stadio dell'analisi in cui viene determinato il verosimile significato di una frase sulla base dei legami e delle interazioni tra le parole che compongono un periodo. In quest'ultimo livello le regole si fanno più lasche poiché per sua natura la semantica di un linguaggio è maggiormente mutabile ed influenzata dal senso comune (Liddy, 2003).

2.1.4 Introduzione a SyNTHEMA Semantic Center

Internet rappresenta uno spazio di condivisione e confronto sempre più diffuso, una riproduzione del mondo reale sempre più fedele ed efficace. Per un'azienda o un'istituzione, la possibilità di individuare, decifrare, classificare o "misurare" la percezione che di lei hanno i clienti o i potenziali fruitori di un suo servizio/prodotto, misurando con efficacia il loro grado di soddisfazione (*customer satisfaction*), costituisce un formidabile *asset* competitivo. Ad esempio, un'azienda può scoprire

facilmente cosa si dice riguardo ai suoi prodotti o servizi, coglierne facilmente i difetti o rilevare inaspettati punti di forza da valorizzare; può cogliere la risposta dei clienti rispetto a campagne pubblicitarie in atto e, eventualmente, modificarle in tempo reale. In generale, le attività di *Web Sentiment* e *Web Reputation* si espletano nell'analisi semantica e statistica di blog, gruppi di discussione, chat ed articoli giornalistici. Ma perché Web Sentiment e Web Reputation possano essere eseguite, le informazioni destrutturate di interesse presenti nell'ecosistema Web devono essere raccolte ed analizzate. È qui che entra in gioco SyNTHEMA.

SyNTHEMA Semantic Center (iSyN SC) è un sistema integrato per la gestione e la condivisione della conoscenza in grado di analizzare le richieste formulate dagli utenti in linguaggio naturale e, sulla base di una analisi linguistica profonda, fornire loro una risposta in modo automatico. È anche un sistema di Knowledge Mining per l'indicizzazione semantica dei contenuti testuali (*feature* principale per le nostre finalità), in grado di facilitare la ricerca di informazioni non strutturate contenute all'interno di documenti, siti web, database ed altre fonti documentali. Le funzionalità che questo sistema ha e le attività che permette di svolgere sono molteplici: non le andremo ad illustrare nella loro totalità, ma ci concentreremo solo su quelle che abbiamo utilizzato per svolgere il lavoro di tesi: la parte di crawling, e quella legata al motore semantico.

2.2 iSyN SC: il crawler

In letteratura, un web crawler è un *bot* che analizza in maniera automatica i materiali presenti nel web, acquisendo ed indicizzando quei contenuti i quali risultino essere di interesse. Il crawler presente in iSyN SC si basa su insiemi di keyword fornite dall'utente che determinano con la loro presenza o assenza nel testo analizzato il recupero o meno del contenuto. Esso esegue una scansione delle fonti documentali definite dall'utente finalizzata al caricamento dei contenuti nel database del server di

indicizzazione. In iSyN SC il crawling può essere di tipo interno o esterno. Il crawling interno è realizzato mediante funzioni interne al sistema, mentre quello esterno richiede un componente software di terze parti (*Searchbox*) che interagisce con il server di indicizzazione mediante *webservice*. La scansione delle fonti può essere pianificata e ripetuta a determinati intervalli di tempo in modo da aggiornare regolarmente l'indice e consentire la ricerca di contenuti variabili. Il crawler messo a disposizione da iSyN SC ci da grosse libertà di parametrizzazione e definizione. Vediamo le principali.

2.2.1 Definizione delle fonti documentali

Le fonti documentali sono repository da cui vengono prelevati i contenuti da indicizzare (documenti, e-mail, pagine web, ecc). Possono essere interne (accesso diretto e permanente dal filesystem del server di indicizzazione) o esterne (accesso tramite collegamento su richiesta ad un altro server mediante un protocollo di comunicazione quale FTP, SMB, HTTP, POP3, IMAP, ecc). Per ogni fonte standard scelta in fase di *source selection*, è necessario definire una fonte documentale.

Fig 2.2: Scheda di aggiunta fonte

Per ciascuna fonte è doveroso:

- Selezionare l'area
- Selezionare il tipo di fonte
- Specificare la descrizione (tipicamente qui si indica il nome della fonte)
- Attivare/disattivare l'opzione per la determina della data del documento a partire dal testo: questa opzione va attivata se si desidera che nel documento estratto dal crawling venga utilizzata la data presente nel documento. In questo caso la data verrà inserita non appena il sistema avrà analizzato linguisticamente il documento. Se questa opzione non è attiva, verrà utilizzata come data quella di importazione del documento estratto
- Attivare/disattivare l'opzione per il filtro sui contenuti: attivare questa opzione per non estrarre nel documento le informazioni non rilevanti per il contenuto. Questa opzione risulta particolarmente utile per le fonti di tipo Web in cui tipicamente le pagine contengono informazioni pubblicitarie o informazioni ripetute in tutte le pagine in base al tipo di layout del sito. I contenuti sono filtrati in base ai parametri presenti nella sezione Crawling e Clustering dei Parametri di configurazione
- Scegliere la modalità di gestione dei duplicati tra
  - Duplica sezione in ogni documento
  - Mantieni sezione solo nel primo documento
  - Elimina sezione da tutti i documenti

  In caso di più sezioni identiche in più documenti, ai fini dell'indicizzazione del documento è possibile scegliere se duplicare la sezione in ogni documento, mantenere la sezione solo nel primo documento oppure eliminare la sezione da tutti i documenti. Anche questa opzione risulta particolarmente utile per le fonti di tipo Web
- Scegliere "si" o "no" dall'elenco a discesa per attivare o disattivare l'analisi del sentiment su tutti i contenuti della fonte specificata

- Definire il grado di affidabilità della fonte tra
  - Completamente attendibile
  - Normalmente attendibile
  - Abbastanza attendibile
  - Di solito non attendibile
  - Non attendibile
  - Non si è in grado di valutarne l'attendibilità

Alcune delle opzioni disponibili nella finestra AGGIUNTA FONTE variano in base al tipo di fonte scelto (sono richiesti i campi *Host*, *Utente* e *Password* per le fonti di tipo Filesystem, FTP, Mailbox, Database).

2.2.2 Scelta della tipologia di fonte

Il crawler di iSyN SC permette scegliere quali tipologia di fonte documentale utilizzare da un insieme finito di tipi possibili:

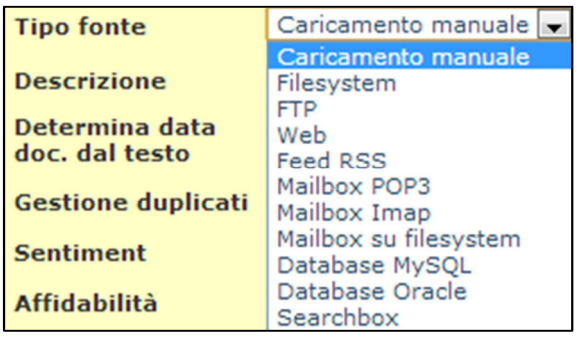

Fig 2.3: Tipi di fonti documentali

- Caricamento manuale (mediante upload da web browser o da webservice)
- Filesystem (anche su altro server mediante mount)
- FTP
- Web (siti Web o ricerche mediante motori di ricerca)
- Feed RSS

- Mailbox (POP3, IMAP, su filesystem)
- Database (MySQL, Oracle)
- Searchbox (componente esterno di crawling)

2.2.3 Classi di documenti recuperabili

I documenti sul Web si presentano in svariate forme. La scelta della tipologia di documento da recuperare va effettuata in funzione di come i contenuti sono strutturati all'interno della fonte scelta. Le possibilità offerte dal crawler di iSyN SC sono le seguenti:

- NOSEZIONI: tipo di documento che non tiene conto della suddivisione in sezioni del testo identificato nel crawling
- RILASCIO: tipo di documento che tiene conto delle informazioni sugli stili di formattazione e raccoglie ogni sezione in base alle informazioni di appartenenza al settore e all'argomento. Si tratta di un unico modello multi-settore e multi-argomento utilizzato per la ricerca di contenuti estratti dal crawling e classificati in sezioni per settori e argomenti
- STILETITOLO: tipo di documento più semplice rispetto al precedente che tiene conto delle informazioni sugli stili di formattazione senza raccogliere ogni sezione in base alle informazioni di appartenenza al settore e all'argomento. Questo tipo di documento viene utilizzato in particolare per fonti di tipo web e crea una sezione distinta per ogni stile titolo identificato
- CLASSEDIV: tipo di documento strutturato in base a tag di suddivisione HTML. Questo tipo risulta particolarmente utile per la classificazione di fonti di tipo Web e feed RSS per escludere o includere esclusivamente i testi racchiusi tra tag di tipo `<DIV>` di determinate classi.

2.2.4 Schedulazione del crawling

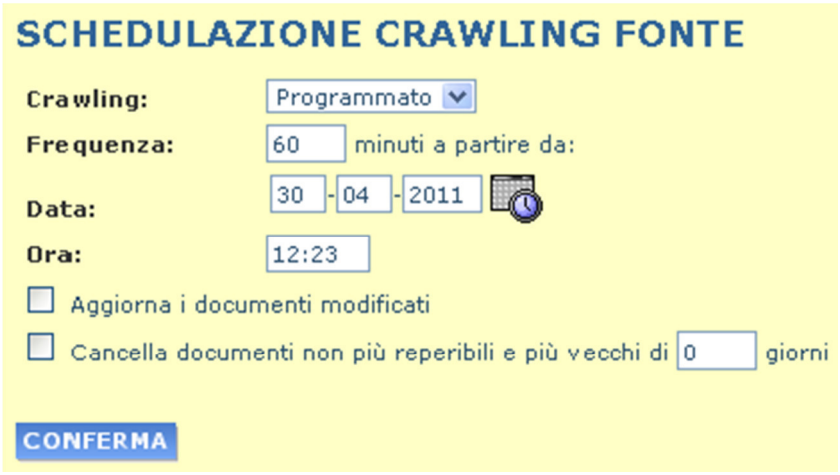

Fig 2.4: Schedulazione del crawling

Per ogni fonte documentale definita, viene data la possibilità all'utente di schedulare temporalmente il crawling (esegui il crawling sulla fonte ogni *n* minuti a partire dal giorno *k*), oppure mantenere inattivo il crawling per una fonte specifica (quindi impostare lo stato di schedulazione su "inattivo") fino a quando l'utente lo desidera.

2.3 iSyN SC: il motore semantico

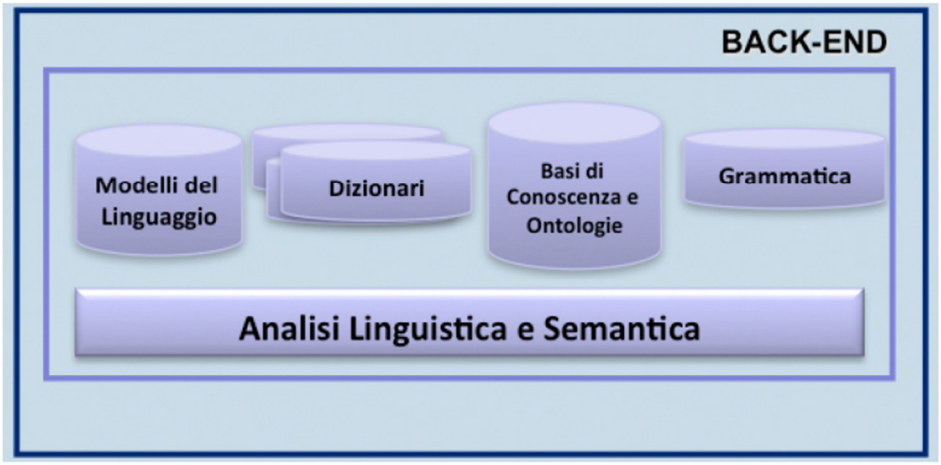

Fig 2.5: Componenti del motore semantico SyNTHEMA

Il motore semantico di iSyN SC viene chiamato in figura *back-end* in quanto la soluzione dell'azienda per la sentiment analysis comprende anche un *front-end* di presentazione dei risultati che però non è di pertinenza del progetto trattato in questo documento. Andando ora ad analizzare quella che è la struttura del software SyNTHEMA è necessario tenere a mente che all'interno di questo progetto è stato utilizzato come una *black box* (eccezion fatta per gli interventi realizzati sulla base di conoscenza che successivamente andremo a presentare) per cui non sarà possibile presentare il motore con un livello di dettaglio molto approfondito. Con riferimento alla Figura 2.5 osserviamo che, per portare a termine il processo di NLU l'analisi si avvale di una serie di risorse linguistiche.

2.3.1 Dizionari e modelli del linguaggio

I dizionari sono risorse che si costituiscono di elenchi di termini ai quali possono essere associati differenti informazioni come per esempio:

- La lingua cui un vocabolo appartiene
- Uno o più tipi i quali possono servire a definire l'ambito di appartenenza di un termine
- Un'informazione preliminare sul POS del lemma (in modo da poter distinguere "buona" aggettivo da "buona" sostantivo per esempio)
- Un elenco di sinonimi di un termine

SyNTHEMA utilizza la lingua associata ad ogni termine poiché tenta di compiere in fase di elaborazione un'analisi *cross language* per analizzare correttamente le frasi ibride in cui più linguaggi sono mescolati e poter così distinguere vocaboli come "come" avverbio della lingua italiana e "come" terza persona singolare del verbo *comer* (mangiare) in spagnolo. Per poterli cosi correttamente tradurre e gestire.

Riguardano dizionari e set di regole grammaticali da applicare ai documenti. Come gli strumenti precedentemente presentati anche i modelli del linguaggio sono strumenti logico-matematici grazie ai quali si possono definire generiche strutture, ovviamente strutture di linguaggi, e le relazioni che intercorrono tra più strutture. Sono quindi mezzi fondamentali per gestire ed organizzare l'uso delle risorse linguistiche precedentemente esposte

2.3.2 Base di Conoscenza e Ontologia di Dominio

Tutti i documenti estratti tramite crawling vanno ad arricchire una ontologia di dominio. iSyN SC consente di gestire ed organizzare al meglio la base di conoscenza (*Knowledge Base,* KB), denominata anche ontologia di dominio tramite la rappresentazione formale dei concetti e delle relazioni semantiche all'interno di un dominio di interesse. I diversi sensi dei termini, che costituiscono i concetti del dominio, sono memorizzati nella base di conoscenza di dominio insieme alle relazioni concettuali che intercorrono con gli altri concetti. Un'ontologia costituisce una concettualizzazione assiomatica di un dominio di interesse ed in quanto tale è esprimibile in una logica descrittiva. Nel campo dell'intelligenza artificiale, e successivamente della linguistica computazionale, questo strumento cominciò ad essere utilizzato per descrivere il modo in cui diversi *pattern* (schemi) vengono combinati in una struttura dati contenente tutte le entità rilevanti (e le loro relazioni) di uno specifico dominio ("Ontologia," n.d.). Le basi di conoscenza sono database particolari progettati esplicitamente per la gestione della conoscenza in ambito aziendale o universitario ("Knowledge," n.d.). La KB sfruttata da SyNTHEMA permette la rappresentazione formale dei concetti e delle relazioni semantiche all'interno di un dominio di interesse. L'ontologia di dominio contiene inoltre le informazioni necessarie per la risoluzione delle ambiguità di senso e delle anafore. I concetti e le relazioni semantiche possono essere facilmente definiti, mediante l'editor della base di conoscenza, attingendo alle relazioni semantiche trovate automaticamente dal sistema durante l'analisi dei testi (estrazione automatica di relazioni semantiche dal corpus dei documenti). I termini vengono inseriti

automaticamente nell'ontologia di dominio la prima volta che il sistema li rileva durante l'analisi linguistica dei testi. In assenza di una specifica di senso, il termine viene inserito nella sua accezione comune (definizione del senso vuota) e definito come elemento di tipo standard (STD): sarà poi compito dell'utente con ruolo terminologico definire i contenuti di questa base di conoscenza. Se durante l'analisi dei testi il motore linguistico non è in grado di classificare correttamente un termine perché non è presente né nei dizionari generali della lingua, né nel dizionario di dominio, il termine viene inserito comunque nell'ontologia di dominio, marcato con la categoria grammaticale "Sconosciuto". Mediante l'editor della base di conoscenza è possibile definire sensi diversi per i termini ambigui nel dominio. Con l'analisi viene alimentata la base di conoscenza con nuovi termini, avviene il ricalcolo automatico delle frequenze dei termini e di conseguenza anche il ricalcolo del modello di categorizzazione in base alle modifiche apportate dall'ultima analisi. Cambiando la base di conoscenza, infatti, vengono introdotti nuovi concetti che hanno relazioni tra loro e quindi possono cambiare le classificazioni presenti. I diversi sensi dei termini, che costituiscono i concetti del dominio, vengono quindi memorizzati insieme alle relazioni concettuali che intercorrono con gli altri concetti.

Fig 2.6: Interfaccia di gestione della Base di Conoscenza in SyNTHEMA

Vediamo chiaramente in Figura 2.6 che ad ogni concetto sono associati una serie di attributi, questi definiscono univocamente i nodi dell'ontologia di dominio, ovvero i concetti stessi, e sono:

- **Termine** Contiene il termine indicizzato ed è associato a una specifica icona in base al tipo di concetto (serve a definire l'importanza da dare al concetto

nell'ambito del dominio o a definirlo come parte di specifiche classi), che può essere

- `Standard (STD)` è il valore di default
- `Topic Of Interest (TOI)` concetto particolarmente rilevante per il dominio
- `Deprecated (DEP)` concetto non desiderato o da penalizzare
- `Abbreviation (ABR)` classe abbreviazioni, termini che un punto finale
- `Named Entity (ENT)` classe entità, in genere nomi propri

Il tipo di concetto può influire sulla scelta del senso in caso di ambiguità. In assenza di altre informazioni contestuali (relazioni concettuali specifiche) vengono scelti i concetti di tipo TOI. Il tipo di concetto influisce anche sul peso attribuito durante la categorizzazione dei contenuti e durante la ricerca in linguaggio naturale. I concetti di tipo TOI sono i più significativi, quelli di tipo DEL vengono ignorati. La definizione dei concetti di tipo ABR è determinante per una corretta segmentazione dei testi in presenza di abbreviazioni. Solo i concetti contrassegnati dai tipi TOI, ENT e ABR vengono passati al motore linguistico come "dizionario di dominio" e vengono utilizzati durante l'analisi. Le espressioni poli-lessicali o *multiword expression* (lemma composto da più parole) se contrassegnate come TOI possono avere un effetto determinante sull'analisi linguistica delle frasi in quanto tali termini vengono riconosciuti come entità unica, non più come concetti dei singoli costituenti

- **POS** La Part Of Speech come abbiamo precedentemente accennato serve a distinguere termini lessicalmente uguali ma semanticamente differenti
- **Sent** La base di conoscenza prevede la possibilità di associare ad un concetto un sentiment che, nel caso specifico riportato in figura, non è stato esplicitato

- **Alias** Sono l'equivalente dei sinonimi dei quali abbiamo già parlato precedentemente, i termini associati ad un target specifico (in figura il target è "Silvio Berlusconi"). Nel caso dovessero essere identificati all'interno del testo verrebbero sostituiti con il loro lemma di riferimento per facilitare l'analisi. SyNTHEMA inoltre prevede un meccanismo di gestione dei sinonimi attraverso tecniche di stemming e lemmatizzazione finalizzate alla flessione e declinazione corretta dei concetti inseriti nella base di conoscenza come alias. Diventa quindi possibile per il sistema, nel caso si presenti una relazione di sinonimia tra i termini "Lista Civica con Monti per l'Italia" e "Lista Monti", riconoscere come alias di "Lista Civica con Monti per l'Italia" anche "Lista Monti" e non solamente "Lista Civica con Monti per l'Italia", cosa che sarebbe successa se invece gli alias fossero trattati come semplici stringhe da ricercare, per corrispondenza, nel testo. La gestione che SyNTHEMA attua degli alias si basa sulla loro lemmatizzazione in questo modo.
- **Senso** È un campo che l'utente può valorizzare a suo piacimento per facilitare la lettura dell'output di analisi
- **Lingua** Linguaggio di appartenenza del termine
- **Frq** Indica la frequenza, il totale delle volte, in cui il termine in esame è stato individuato dal motore semantico durante l'analisi
- **Relazioni** Campo che contiene le informazioni sulle relazioni esistenti tra i concetti che compongono la base di conoscenza, ovvero sugli archi che collegano i nodi (concetti) che compongono la KB. Il sistema mette a disposizione dell'utente una serie di relazioni gestite che vanno da quelle semantiche di gerarchia come la relazione *ISA*: "Silvio Berlusconi **ISA** imprenditore" a relazioni più squisitamente sintattiche, delle quali si avrà evidenza dall'output di analisi, come la relazione *AGENT* vigente tra soggetto e predicato verbale. Infatti in fase di elaborazione della seguente frase: "Il bambino gioca con il cane" il sistema segnala correttamente la relazione tra il

primo ed il secondo elemento del periodo *AGENT [1:giocare, 2:bambino]* (il che significa: l'azione *giocare* viene svolta da *bambino*). Nello specifico le relazioni concettuali (tra due concetti A e B) a disposizione dell'utente sono:

- **ISA** A è un iponimo di B
- **PARTOF** A è parte di B
- **MEMBEROF** A è membro di B
- **SYNONYM** A è un sinonimo di B
- **ANTONYM** A è il contrario di B
- **TRANSLATION** A è la traduzione di B
- **ABSTRACT** A afferisce semanticamente a B
- **NEARTO** A specifica il contesto di B
- **AGENT** A compie l'azione B
- **OBJ** l'azione A si compie su B
- **IOBJ** l'azione A ha complemento di termine B
- **QUAL** A viene qualificato da B
- **HOW** l'azione A si svolge nel modo B
- **WHEN** l'azione A si svolge nel momento B
- **WHERE** l'azione A si svolge nel luogo B
- **COMP** l'azione A ha complemento B

Con l'analisi dei testi, il motore linguistico fornisce una proposta di classificazione della relazione in base alle regole di derivazione semantica definite nel sistema. Queste relazioni "proposte" sono indicate nelle relazioni con i seguenti tipi di relazione:

- **P_ABSTRACT** utilizzata per identificare concetti pseudo-sinonimi tra POS diverse, ovvero lo stesso concetto per termini con categorie semantiche diverse unite da una relazione di afferenza semantica

- **P_ISA** utilizzata per identificare concetti in una relazione gerarchica tra loro
- **P_SYNONYM** utilizzata per identificare concetti che per il motore linguistico che li analizza risultano diversi ma sono sinonimi e hanno la stessa POS. Si tratta di *single word* e *multi word*: ad esempio "mass-media" è in relazione P_SYNONYM con "mass media". Grazie alla creazione di questa relazione, il sistema consente di individuare tutti i termini con o senza trattino nella ricerca

Un insieme di relazioni gerarchiche costituisce una tassonomia (ad esempio iponimia/iperonimia tra i nomi). Re relazioni sono inoltre ereditarie, pertanto le relazioni concettuali definite sui concetti "padre" valgono anche i concetti "figlio" (non è necessario ridefinire le relazioni sul figlio)

2.3.3 Grammatica

Sono strumenti utilizzati per descrivere un linguaggio formale. In ultima analisi le grammatiche sono degli insiemi di regole grazie alle quali è possibile, utilizzando formalismi appositi, delineare insiemi di sequenze di simboli (i simboli utilizzabili da una grammatica ne costituiscono l'alfabeto) dette stringhe che possono essere generate grazie alle regole (produzioni) che la grammatica prevede. Per quanto non sia stato possibile nell'ambito di questo progetto visionare le grammatiche utilizzate da SyNTHEMA è noto che l'approccio all'NLP adottato faccia uso di grammatiche generative. Queste sono insiemi di regole che permettono, in modo iterativo, la generazione delle cosiddette "stringhe ben formate" che costituiscono un linguaggio. In altre parole, più precisamente quelle dell'ideatore di questo strumento ovvero il linguista Noam Chomsky[1], le grammatiche sono la rappresentazione matematica di tutte le conoscenze che una persona deve avere per poter riconoscere un'espressione

[1] *http://en.wikipedia.org/wiki/Noam_chomsky*

come grammaticale. E' in pratica la formalizzazione di un algoritmo per la generazione di stringhe linguistiche

2.3.4 Processo di analisi

Le risorse linguistiche sopra riportate vengono sfruttate dal motore semantico nel corso dell'attività di analisi, il cui *workflow* si compone degli step riportati in figura 2.7. SyNTHEMA offre la possibilità di svolgere questi step secondo due modalità differenti, le quali si diversificano tra loro per il livello di profondità dell'analisi svolta:

- Un'analisi linguistica **superficiale** che richiede meno risorse e garantisce una miglior efficienza a fronte di un risultato meno preciso. Il sistema infatti in questo caso non è in grado di risolvere le ambiguità semantiche nonostante utilizzi comunque tecniche di lemmatizzazione e stemming per distinguere casistiche come per esempio "i detti" (popolari) ed "io detti" (prima persona singolare del passato remoto del verbo dare)
- Un'analisi linguistica **approfondita** richiede invece maggior tempo, comporta quindi minor efficienza, ma a fronte di un maggior impiego di risorse aumenta l'efficacia del sistema. Nel caso venga impiegata questa tecnica il motore semantico cercherà di risolvere ogni ambiguità presente nel testo ed individuare il ruolo svolto da ogni concetto all'interno del periodo non limitandosi agli elementi principali portatori di significato ovvero il soggetto di un verbo ed il suo complemento oggetto

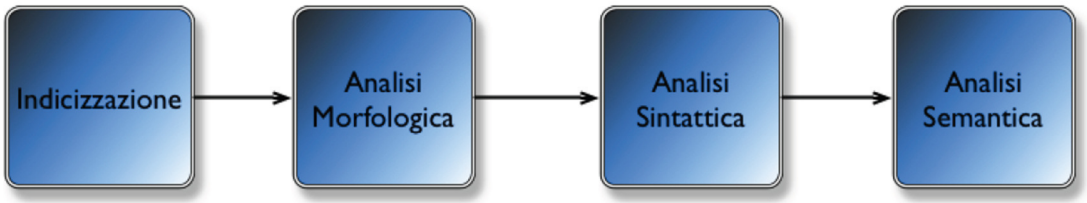

Fig 2.7: Gli step di analisi compiuti dal motore semantico SyNTHEMA

Sinteticamente, le attività svolte durante queste fasi sono le seguenti:

1. **Indicizzazione** durante questa prima fase elaborativa il motore ricerca, sfruttando i dizionari di cui è dotato e basandosi quindi su una serie di keyword, i concetti presenti nel testo analizzato. Qui non viene assolutamente prestata attenzione alla relazione sintattica esistente tra i vari concetti, l'elaborazione è focalizzata unicamente alla corretta individuazione dei concetti ed alla loro estrazione funzionalmente agli step successivi
2. **Analisi Morfologica** La fase successiva è quella dell'analisi morfologica durante la quale il sistema cerca di ricondurre ogni parola ad una possibile classificazione grammaticale, non semantica, che quindi non consente la rimozione delle ambiguità. Per identificare le varie classificazioni il motore confronta i concetti estratti con le possibili corrispondenze presenti nei dizionari dopo aver, per ogni termine, determinato: la forma base, la categoria grammaticale, la flessione o coniugazione (grazie ad algoritmi di stemming e lemmatizzazione dei vocaboli)
3. **Analisi Sintattica** Questa fase prevede un'elaborazione dei concetti estratti dai testi analizzati volta a rimuovere tutte le ambiguità linguistiche (ma non quelle semantiche). Il sistema deve quindi essere in grado in questo step di analizzare correttamente figure linguistiche, estremamente diffuse nel linguaggio italiano, come l'anafora: per esempio analizzato il periodo "Ho trovato 50 euro per terra e li ho raccolti" il motore deve essere in grado di capire che la particella pronominale "li" si riferisce ai 50 euro trovati per terra. Questa fase è profondamente influenzata dalla correttezza del testo e dalla sua attinenza alle regole grammaticali italiane poiché il motore usa tecniche volte a facilitare la comprensione della frase come la trasformazione di tutti i periodi passivi in attivi per agevolare il compito di individuazione dei soggetti di verbi passivi trasformandoli nei complementi oggetti di forme attive. L'uso di queste tecniche potrebbe complicare l'analisi invece di semplificarla se il periodo che si va a trasformare è scritto in modo scorretto.

Sarebbe quindi comodo poter disporre di dizionari e risorse specifiche legati ai contesti di utilizzo: un dizionario da utilizzare quando l'utente sta scrivendo una mail, un altro per i social network e così via. In questo step del workflow di analisi l'applicativo crea *n* alberi sintattici grazie ai quali è possibile eseguire il *parsing* della frase. Nel caso esista più di un albero che risulti essere corretto ci troveremmo in una situazione di ambiguità che verrà risolta in fase di analisi semantica

4. **Analisi Semantica** Questo ultimo passaggio elaborativo deve risolvere ogni ambiguità residua che impedisca la corretta analisi della frase. Con un processo iterativo di confronto il motore cerca di attribuire, tra i vari alberi sintattici calcolati in fase di analisi sintattica, quello corretto rispetto al periodo in esame. È proprio questa la fase in cui diventa più importante disporre di una base di conoscenza che, contrariamente all'approccio top-down adottato da un'ontologia, sia stata costruita e parametrizzata specificamente sul *vertical* di applicazione dello strumento analitico in modo da poter correttamente identificare le convergenze di significato che possano aiutare nella disambiguazione. Un'analisi semantica svolta con le giuste risorse linguistiche potrebbe per esempio capire che in un particolare contesto come quello della politica italiana, il "cavaliere" (che per un'ontologia della lingua italiana altro non è se non un membro di ordini cavallereschi) possa in realtà essere sinonimo di Silvio Berlusconi. È sempre in quest'ultima fase di analisi che SyNTHEMA, grazie alla tipizzazione dei concetti realizzata ovviamente nella base di conoscenza, effettua il calcolo della polarizzazione del sentimento. Benché i testi vengano analizzati in funzione delle loro unità costituenti, i singoli periodi, il motore restituisce all'utente anche un calcolo complessivo del sentimento espresso in un testo organico. Questo sentimento viene riportato in forma numerica tramite un valore che è il risultato della somma algebrica delle singole opinioni posta la scala di valori riportata nella Tabella 2.1.

Sentimento espresso	Valore associato
molto positivo	2
positivo	1
neutro	0
negativo	-1
molto negativo	-2

Tab 2.1: Valori utilizzati da SyNTHEMA per il calcolo del sentiment espresso in un testo

L'output finale restituito dal sistema contiene tutti gli elementi risultanti dalle varie fasi di analisi riportate e si presenta come l'esempio riportato in Figura 2.8, relativo alla frase: "Sarebbe un grande traguardo se un partito politico innovativo come il M5S vincesse le elezioni".

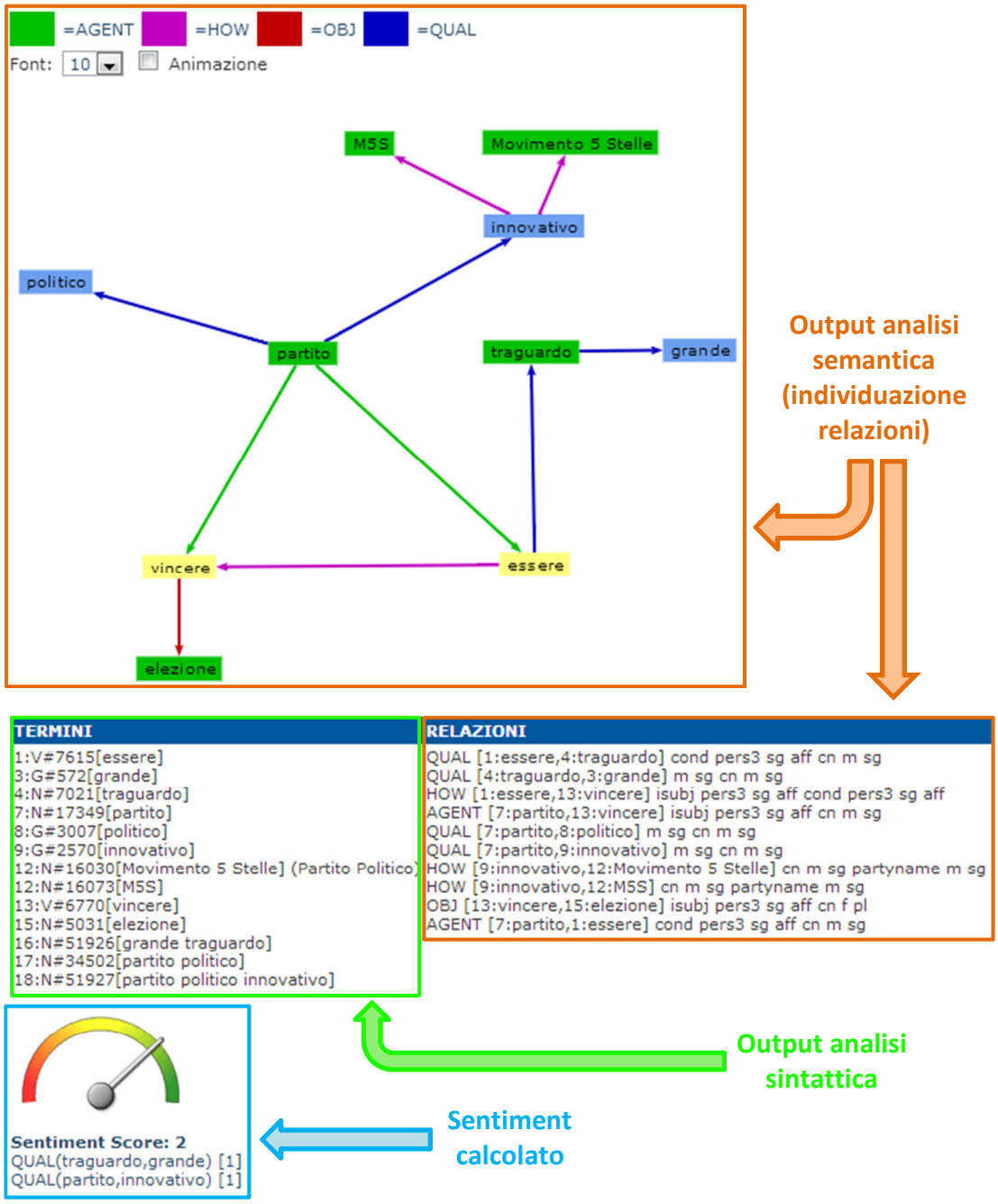

Fig 2.8: Esempio dell'output fornito da SyNTHEMA

2.4 Talend Open Studio

Talend Open Studio (TOS) è un potente e versatile set di prodotti open source per sviluppo, testing, deploy, amministrazione, gestione dei dati e dei progetti di integrazione tra applicazioni. TOS fornisce l'unica piattaforma unificata che rende più semplice il *data management* e l'integrazione di applicazioni fornendo un solo ambiente per gestire l'intero ciclo di vita che attraversa un'azienda. Gli sviluppatori raggiungono vasti guadagni di produttività attraverso un interfaccia grafica *easy-to-use* ed eclipse[2]-based che combina *data integration*, *data quality*, *master data management* (MDM), integrazione di applicazioni e big data.

I prodotti di Talend abbassano drasticamente la barriera di adozione per quelle imprese che desiderano potenti soluzioni "pacchettizzate" verso le sfide operative come il *data cleansing*, MDM e *enterprise service bus* deployment (un Enterprise Service Bus (ESB) è un'infrastruttura software che fornisce servizi di supporto ad architetture SOA (*Service-Oriented Architecture*) complesse) ("Enterprise," n.d.). Sfruttando ed estendendo le principali tecnologie di *Apache*[3], le soluzioni open source ESB e SOA di Talend aiutano le imprese a costruire architetture aziendali flessibili ed altamente performanti.

Tutti i prodotti TOS sono 100% open source e liberi di esser scaricati ed utilizzati. La *community* di Talend provvede un set di servizi di supporto e di *tutoring* inclusi forum di discussione (che ci hanno salvato la vita in più occasioni), *bug tracking*, codici sorgenti ed *add-on* dei prodotti (Talend, 2006a).

2.4.1 Data Integration

I prodotti Talend di data integration forniscono un'integrazione potente e flessibile, in modo tale da far smettere alle aziende di preoccuparsi su come i database e le

[2] *http://www.eclipse.org/*
[3] *http://www.apache.org/*

applicazioni comunicano gli uni con le altre, e di converso massimizzare il valore dell'uso dei dati.

Il modulo di data integration fornisce un estensibile e altamente performante set di strumenti open source per l'accesso, la trasformazione e l'integrazione di dati da qualsiasi sistema aziendale in real-time o batch per soddisfare sia le esigenze operative che quelle analitiche di integrazione di dati. Con più di 450 connettori, integra più o meno ogni possibile sorgente di dati. L'ampia gamma di casi d'uso gestiti include: integrazione su larga scala (big data/NoSQL), ETL (*Extraction, Transformation and Loading*) per business intelligence e data warehousing, data migration, data sharing e data services. TOS Data Integration fornisce:

- un **business modeler**, ovvero uno strumento visuale per il design della logica di business per una applicazione
- un **job designer**, uno strumento visuale per diagrammi funzionali, gestione dei metadati, storing e gestione di ogni progetto di metadati, inclusi dati contestuali come i dettagli di connessione a database e i percorsi dei file

Talend si connette in modo nativo a database, applicazioni pacchettizzate (ERP, CRM, ecc), applicazioni Cloud e SaaS (*software-as-a-service*), mainframe, file, webservice, data warehouse, data mart, e applicazioni OLAP. Offre avanzati componenti integrati per l'ETL, inclusi manipolatori di stringhe, *Slowly Changing Dimensions* (nell'ambito del data warehousing, si intendono dimensioni i cui attributi hanno valori che possono variare lentamente nel tempo), gestione automatica del *lookup* e del *bulk loading* (azione di inserimento di un grosso set di righe all'interno di una tabella). L'integrazione diretta è fornita con data quality, data matching, MDM e le relative funzionalità. Talend si connette inoltre alle applicazioni di clouding più popolari incluse *salesforce.com*[4] e *sugarCRM* [5]. Il repository condiviso

[4] *http://www.salesforce.com/it/*

messo a disposizione da Talend consolida tutte le informazioni del progetto e i metadati aziendali in un repository centralizzato condiviso da tutti gli *stakeholders*: utenti di business, sviluppatori, staff IT. Gli sviluppatori possono facilmente fare il *versioning* dei job in quanto dispongono di una funzione di roll-back che li può portare immediatamente alla versione precedente. Talend include infine potenti strumenti di testing, debugging, gestione e tuning dotati di monitoraggio in tempo reale delle statistiche di esecuzione dei dati, ed una modalità di analisi avanzata. Può essere eseguito il deploy dei processi attraverso sistemi aziendali e di rete come *data services* utilizzando lo strumento di esportazione (Talend, 2006b).

Nel nostro lavoro di tesi, TOS Data Integration è stato utilizzato in due momenti: quello di passaggio delle clip dal repository del motore semantico al database, e viceversa. Facendo riferimento alla figura 2.6, il lavoro svolto dal modulo di data integration si colloca:

- nel passaggio tra il nodo "crawler's repository" e "database"
- nel passaggio da "database" a "SyNTHEMA", e in quello da "SyNTHEMA" a "database"

[5] *http://www.sugarcrm.com/*

Capitolo 3

Funzionalità e tecniche

Quali sono le caratteristiche più desiderate in un sistema di Social BI? È uno studio della società Alta Plana (Grimes, 2011), i cui risultati sono riportati in Figura 3.1, a mostrarcele. È innanzitutto interessante notare che di tutte le caratteristiche elencate la metà siano di natura spiccatamente tecnica mentre il restante 50% siano invece di carattere più generale. Tra queste la caratteristica più ricercata nasce direttamente dalla consapevolezza che il valore aggiunto di uno strumento di Social BI è direttamente proporzionale alla sua capacità di raccogliere la maggior quantità possibile di dati ed informazioni rilevanti dalla Rete (copertura del parlato). A ciò si legano con un rapporto quasi causale le altre voci presenti in Figura 3.1 a cominciare dalla possibilità di elaborare informazioni riportate in varie lingue ed in maniera quasi scontata la capacità di rilevare le opinioni presenti. Le caratteristiche di natura spiccatamente tecnica riguardano in maniera sostanziale quella che è la capacità di un sistema, o meglio del motore di analisi semantica alla base del sistema stesso, di lavorare in maniera efficace:

- utilizzando tassonomie e dizionari specifici per il core business dell'azienda o comunque relativi all'ambito di utilizzo del sistema rendendolo verticalizzabile e parametrizzabile sulla base delle esigenze informative
- offrendo la possibilità di creare e decidere a quali argomenti prestare maggior attenzione

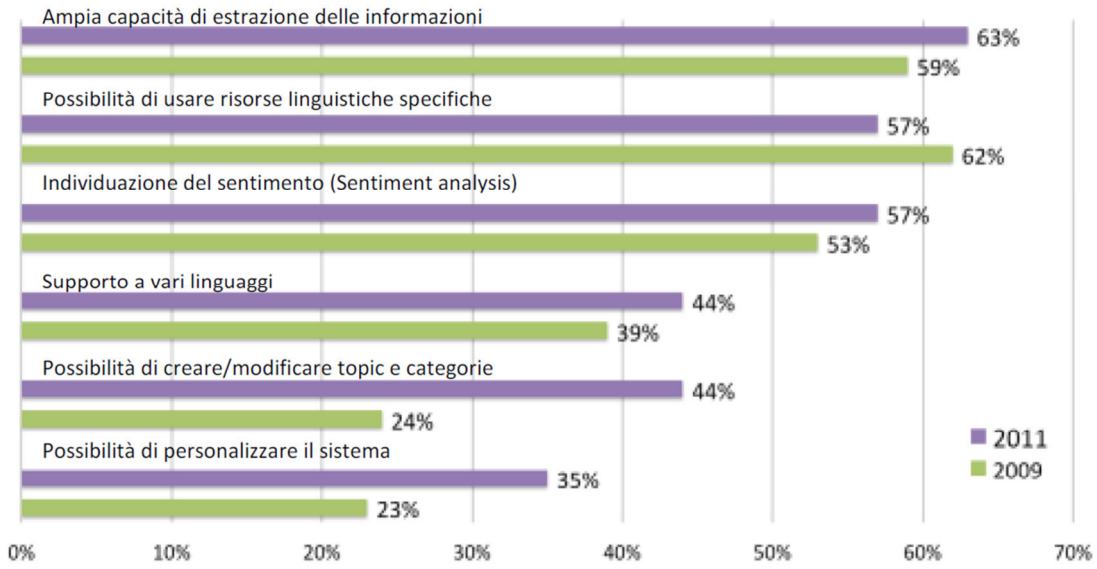

Fig 3.1: Caratteristiche di un sistema di Social BI con maggior rilevanza

3.1 Gli approcci all'analisi

I dati riportati in Figura 3.1 confermano come nella scelta dell'architettura di un sistema di Social BI giochi un ruolo fondamentale l'aspetto tecnico, fattore caratterizzante del motore di analisi semantica che costituisce l'elemento principale del processo di Text Analytics. Negli anni sono state sviluppate varie soluzioni di analisi automatica del testo ed ognuna svolge questo compito in un modo specifico. Nonostante i passi dell'analisi possano differire da soluzione a soluzione, l'approccio generale al problema prevede una serie di step consolidati ognuno dei quali con un obiettivo ben preciso (stiamo parlando di analisi morfologica, lessicale, sintattica e semantica, che abbiamo già illustrato nel secondo capitolo). La differenza tra le varie

modalità di svolgere l'analisi risiede quindi non tanto nelle fasi specifiche, che rimangono sostanzialmente quelle appena elencate, quanto nella modalità adottata e nelle tecniche sfruttate per portare a termine le fasi di analisi sintattica ed analisi semantica. Allo stato dell'arte gli approcci utilizzati si dividono in due grandi famiglie:

- Gli approcci **linguistici** (che è il caso del motore semantico da noi utilizzato), si pongono come scopo quello di compiere un'analisi linguistica approfondita del testo applicando gli step presentati ed avendo come principale riferimento la grammatica standard del linguaggio analizzato
  - PRO: quando i testi analizzati rispettano le regole grammaticali della lingua questi metodi forniscono spesso un'analisi corretta
  - CONTRO: i risultati sono fortemente influenzati dalla qualità dei dati in *input* che spesso si rivela un problema quando la fonte di acquisizione dei testi è l'ambiente social
- Gli approcci **statistici**, si distinguono dalle precedenti nello svolgimento delle ultime due fasi di analisi (sintattica e semantica). Vengono impiegate metodologie statistiche e di data mining con l'intento di estrarre pattern linguistici noti o identificarne e apprenderne di nuovi
  - PRO: queste tecniche si dimostrano più elastiche di quelle prettamente linguistiche e meglio adattabili a dati scarsamente attinenti alla grammatica standard della lingua
  - CONTRO: in mancanza di tecniche linguistiche di analisi, anche un training intenso potrebbe non portare al corretto riconoscimento di strutture standard della lingua italiana

Per completezza è bene riportare che oggi sul mercato è possibile trovare una terza famiglia di soluzioni che implementano solamente la fase di analisi morfologica con l'obiettivo di fornire, grazie allo sfruttamento di algoritmi di categorizzazione automatica, funzionalità di estrazione di concetti noti e polarizzazione di termini

conosciuti a priori perdendo così la possibilità di dare evidenza all'utente dei nuovi argomenti discussi in rete.

3.2 Le funzionalità

Una prima distinzione tra le funzionalità messe a disposizione da un sistema di Social BI è tra quelle di front-end e quelle di back-end. Le prime sono destinate agli utenti finali, le seconde sono destinate agli utenti tecnici e sono necessarie per il corretto funzionamento di quelle di front-end. A un livello più approfondito le funzionalità di front-end, normalmente fruite tramite report dinamici e dashboard, possono essere classificate in base al livello di profondità delle analisi linguistiche a cui i testi devono essere sottoposti per estrarre le informazioni necessarie. Le funzionalità per cosi dire "standard" di un sistema di SBI possono essere raggruppate all'interno di tre macro-categorie che assumono appellativi differenti a seconda della soluzione commerciale che li implementa:

- **Conteggio** identifica la capacità di tenere traccia di tutte le occorrenze di un termine o di un concetto di interesse (dove per concetto viene inteso un argomento al quale si può fare riferimento utilizzando parole diverse)
- **Co-occorrenza** fa riferimento all'individuazione delle occorrenze di due termini distinti all'interno della stessa frase, si chiamano per questo co-occorrenze
- **Topic discovery**, ovvero catalogazione di documenti estratti dal web in base agli argomenti da questi trattati ed ai concetti in essi contenuti. Richiede normalmente l'utilizzo di tecniche di clustering
- **Sentiment Analysis (Opinion Mining)** si intende la disciplina tramite la quale un motore di analisi semantica riesce, analizzando un testo, ad individuare le opinioni espresse in esso ed estrarre i concetti trattati, le relazioni in cui essi sono coinvolti ed interpretare ad un livello semantico tali

termini ed implicazioni al fine di coglierne la positività o negatività, in ultima istanza quindi il significato. Le strategie implementative adottate sul mercato italiano sono estremamente varie. Questo conferma come quella della Sentiment Analysis sia ancora una disciplina giovane in cui manca un approccio consolidato e universalmente riconosciuto come migliore

Queste macro-funzionalità danno infine vita ad un insieme di funzionalità di dettaglio che elaborano ulteriormente il dato grezzo. Il conteggio viene utilizzato per fornire funzionalità di:

- **Counting**, cioè il conteggio delle occorrenze dei termini o di un insieme di topic noti. Per essere efficace richiede un'operazione di "normalizzazione semantica" in cui tutti i termini usati per identificare uno specifico topic vengono riconosciuti come equivalenti al topic stesso riuscendo così a tenere traccia delle reali occorrenze. Per esempio, in ambito politico, ci si potrebbe riferire a Mario Monti dicendo "Monti", "Super Mario", "il Professore" o usando altri nomignoli. L'informazione così estratta può essere ulteriormente rielaborata per evidenziarne specifici aspetti come:
  - **Top topic**, ovvero gli argomenti più discussi in un certo intervallo temporale scelto arbitrariamente
  - **New topic**, cioè i nuovi argomenti di cui si parla nei testi recuperati ma che non erano stati mai trattati precedentemente
  - quei topic che, in particolari momenti storici, diventano **trendy** ovvero particolarmente discussi e dibattuti (potrebbero essere argomenti del tutto nuovi o argomenti che mai come nel lasso temporale considerato avevano attirato l'attenzione degli utenti web)

L'utilizzo di base della funzione di co-occorrenza è quello di identificare un legame, sintattico, semantico o semplicemente di vicinanza, tra due termini finalizzato a

capire cosa si dice di un certo soggetto. Ascrivibili alla categoria di sentiment analysis sono state individuate le seguenti funzionalità:

- **Opinion Mining di base** permette di identificare complessivamente se il giudizio su un concetto è positivo, negativo o neutro differenziando eventualmente sulla base del periodo temporale di riferimento, sulla sorgente dell'informazione o su qualsiasi altro attributo con cui sono stati arricchiti i testi
- Coadiuvata con la funzione di co-occorrenza permette di individuare i **punti deboli e punti di forza** di un singolo concetto evidenziando tutte le sue relazioni (sintattiche e semantiche) presenti nei testi analizzati. In questo caso l'opinione su un concetto è differenziata rispetto al suo legame con altri concetti.

Una classificazione delle funzionalità di back-end può essere basata sulle modalità a disposizione dell'utente di influenzare l'analisi compiuta dai motori. Questa seconda tipologia di funzionalità, definita di "verticalizzazione", si compone di azioni vincolate alla natura specifica del software (iSyN SC) ed all'approccio analitico (linguistico) da questo adottato. La possibilità di influenzare l'analisi del motore adattando il sistema al dominio specifico di utilizzo (quindi la verticalizzazione), può compiersi grazie a:

- **Arricchimento del dizionario**, che prevede di aggiungere alle risorse linguistiche sulle quali si basano le varie fasi di estrazione dell'informazione i vocaboli di interesse propri del vertical di utilizzo del sistema
- **Modifica della polarizzazione** di un termine, che è quel processo con il quale si va ad attribuire un'accezione ad una parola presente nel dizionario nel caso si possa di questa identificare a priori la positività o negatività

- **Creazione di relazioni sintattiche**, un'attività applicabile al motore che, utilizzando un approccio basato sull'analisi linguistica, può in questo modo "apprendere" costrutti sintattici nuovi e funzionali alla corretta interpretazione della frase

3.3 Realizzazione di un database adatto alla Social BI

È stato necessario progettare e realizzare una base di dati costruita ad hoc per le funzionalità e le operazioni della Social BI ove memorizzare l'output, in forma strutturata, dell'analisi fornita dal motore semantico su cui poi eseguire considerazioni, studi, report e creazione di dashboard. L'implementazione fisica è avvenuta su di un database MySQL: un *Relational database management system* (RDBMS), composto da un client con interfaccia a riga di comando e un server, entrambi disponibili sia per sistemi *Unix* o *Unix-like*. Possiede delle interfacce per diversi linguaggi, compreso un driver ODBC e due driver Java (JDBC, che abbiamo sfruttato ampiamente in fase di integrazione ("MySQL," n.d.).

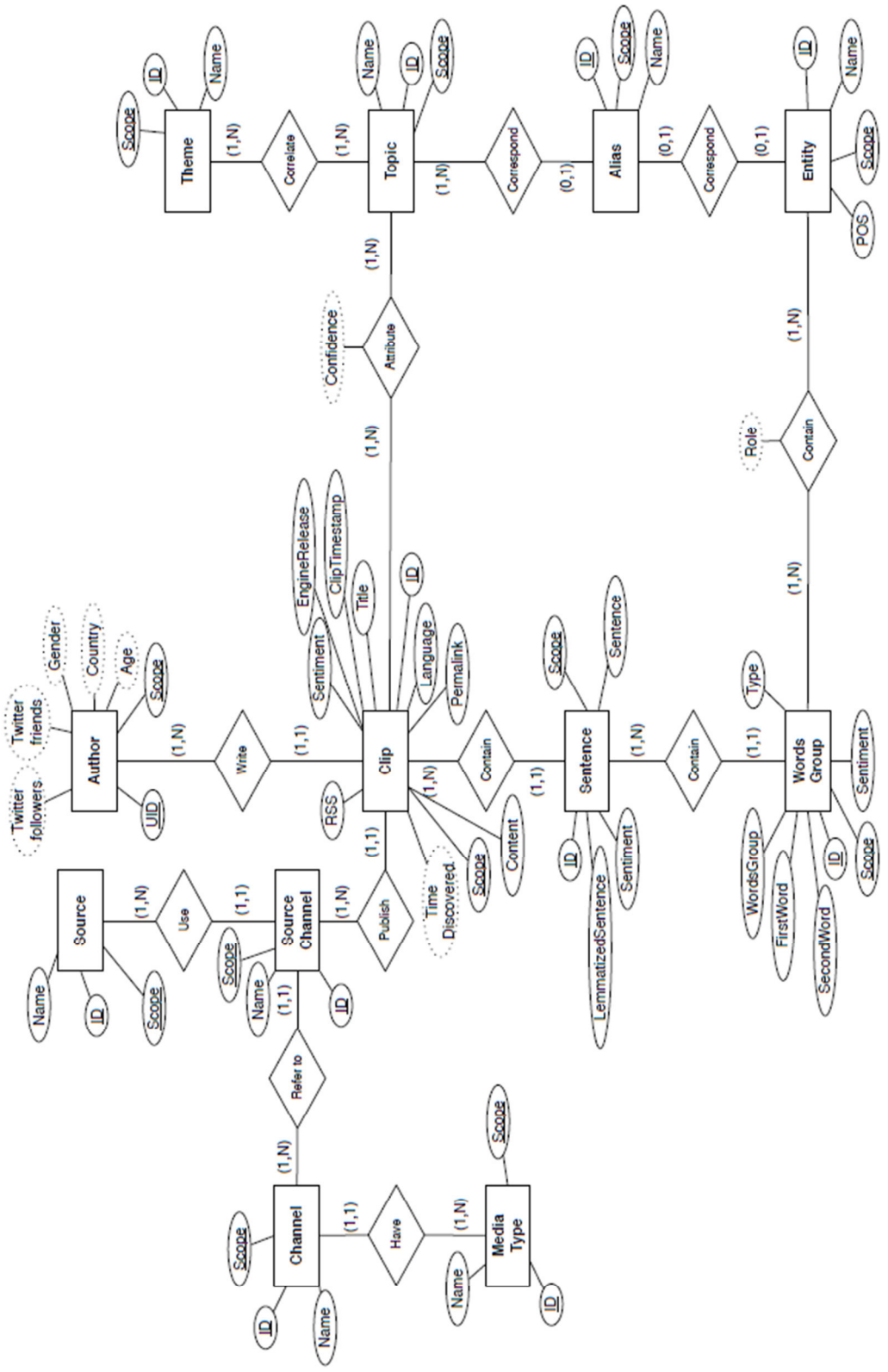

Fig 3.2: Schema E/R tipico di un database per la Social BI

Per poter meglio comprendere il significato delle entità, potrebbe esser necessario ricorrere ad esempi. Ogni volta che sarà necessario, prenderemo come riferimento esplicativo quanto scritto di seguito. Supponiamo che il crawler abbia recuperato la clip "Mario Monti si è candidato per le elezioni nonostante avesse sempre detto il contrario" da un commento presente sulla fanpage di Facebook del quotidiano "Il Sole 24 Ore".

- **Media Type**, all'interno di questa tabella andranno ad essere memorizzate le tipologie generiche di sorgente informativa dalle quali vengono recuperate le clip: esempi di *media type* sono "sito web", "social network", "blog", "forum", ecc. Facendo richiamo all'esempio di riferimento, il media type è "social network", in quanto l'informazione è stata prelevata da Facebook che è un social network.
- **Channel**, in questa tabella vengono memorizzati i canali su cui sono veicolate le informazioni: Facebook, Twitter, Blog, sito web, ecc. Ogni sorgente informativa può aver associati ad essa 1 o più canali ("Il Sole 24 Ore" pubblica informazioni su più fronti: facebook, twitter, sito web, ecc). Facendo richiamo all'esempio di riferimento, il channel è "Facebook", in quanto l'informazione è stata prelevata da Facebook.
- **Source**, qui si memorizzano le sorgenti scelte in fase di source selection, che sono soggette all'operazione di crawling. Nell'esempio di riferimento, la sorgente o source è "Il Sole 24 Ore".
- **Source Channel**, in questa tabella sono salvate le associazioni tra la source specifica ed il channel che sta utilizzando per diffondere quella specifica notizia.
- **Clip**, ogni record di questa tabella memorizza le informazioni relative alle clip prelevate in fase di crawling dalle sorgenti selezionate, nel senso che il crawler ha trovato all'interno di ognuno dei documenti salvati in questa tabella una o più parole chiavi tra l'insieme di keyword definite dall'utente.

Di tutte le informazioni memorizzate per ciascuna clip in fase di scrittura del record su database, è rilevante segnalare la presenza attributi come "Content", che verrà riempito con il contenuto testuale della clip recuperata, e "sentiment", contenente il sentiment associato alla clip dal motore semantico, ed "EngineRelease" contente la versione del motore semantico che ha analizzato la clip (es: se "EngineRelease" contiene il valore "syn_v0" significa che quella clip è stata analizzata dalla versione base del motore (syn_v0 sta per "versione originale della KB del motore semantico") semantico SyNTHEMA, prima e priva di qualsiasi intervento di modifica sulla base di conoscenza).

- **Author**, memorizza le informazioni relative agli autori della clip. Un autore scrive da 1 a *n* clip, e una clip è scritta da uno ed un solo autore. La maggior parte degli attributi di questa tabella sono opzionali, e l'opzionalità è legata al fatto che alcune informazioni possono non esservi (ad esempio, se la clip non è prelevata da Twitter, non verranno valorizzati attributi come *Twitter friends* o *Twitter followers*).
- **Sentence**, tra le tante funzioni svolte dal motore semantico, vi è anche quella del partizionamento di una clip nelle frasi che la compongono, che avviene seguendo particolari regole proprie del motore. Ogni riga di questa tabella contiene una frase, legata alla clip di appartenenza grazie alla chiave importata dalla tabella clip. Ogni frase, e quindi ogni record di questa tabella, appartiene ad una ed una sola clip, mentre una clip contiene al suo interno da 1 a *n* sentence. Tra i suoi attributi, la tabella Sentence contiene "Sentence" e "Sentiment": il primo è valorizzato con il testo da cui la frase è formata, mentre il secondo contiene il sentiment della frase, associata ad essa dal motore in fase di analisi semantica.
- **Wordsgroup**, ogni clip è formata da frasi, ogni frase è formata da gruppi di parole (wordsgroup). Con "gruppo di parole" indichiamo due singleword o multiword legate l'una all'altra da una relazione sintattica. Nella frase "Heidi

saluta le caprette", è presente il wordsgroup AGENT[salutare, Heidi], che indica come le due parole siano legate tra loro da una relazione di tipo AGENT: l'azione "salutare" viene svolta da "Heidi". Tutti i wordsgroup presenti all'interno di una sentence vengono memorizzati all'interno di questa tabella: un record per ciascun wordsgroup. Di ogni gruppo di parole viene memorizzato il tipo di relazione, la prima parola, la seconda parola, ed il sentiment associato dal motore semantico (oltre ai valori delle chiavi importate). Un wordsgroup appartiene ad una ed una sola frase, mentre ad una frase appartengono uno o più gruppi di parole.

- **Entity**, in questa tabella viene fatto corrispondere un record ad ogni entità (verbo, nome, aggettivo, nome proprio, avverbio) rilevata da SyNTHEMA in fase di analisi. Entità può essere sia singleword che multiword. Per ogni entità, viene memorizzata la sua POS e il suo lemma.
- **Contain**, relazione molti-a-molti che collega le entità Entity e Wordsgroup. Ogni entità è presente in uno o più gruppi di parole, ed un gruppo di parole contiene per forza di cose una o più entità (nel nostro caso al massimo due entità: firstword e secondword).
- **Theme**, sono le macro-aree del vertical politico definite a monte del lavoro di tesi. Ad ogni record corrisponde un tema (politica interna, politici e partiti, istruzione, sanità, ecc).
- **Topic**, sono gli argomenti (persone, società, partiti, leggi, ecc) di interesse maggiormente discussi sul web, che emergono a valle del processo di analisi eseguito dal motore semantico sui documenti prelevati dal crawler. Ad ogni record corrisponde un topic, di cui viene memorizzato il. Un topic può appartenere a più temi (almeno uno), ad un tema possono appartenere uno o più topic. Le due tabelle sono dunque legate da una associazione molti-a-molti.
- **Alias**, questa tabella memorizza i sinonimi dei topic salvati nella tabella corrispondente. Nasce il concetto di alias principalmente per il fatto che

quando si vuole avere un'analisi precisa su quanto un argomento sia più o meno discusso, non si vogliono tralasciare dati. Ad esempio, se al topic "Silvio Berlusconi" non vi facessero capo gli alias "il cavaliere", "il nano" e similari, verrebbero tralasciate dall'analisi tutte le considerazioni e le opinioni che utilizzano al posto di "Silvio Berlusconi" gli alias citati. In questo modo non si avrebbe una visione a 360° di quel che si dice riguardo ad un certo topic. Anche gli alias, come i topic, emergono a valle del processo di analisi. Un alias fa capo ad uno ed un solo topic, mentre ad un topic fanno capo da uno a molti alias.

Capitolo 4

Metodologia di verticalizzazione

La complessità di un progetto di Social BI dipende principalmente dal modello di business adottato e dalle possibilità di verticalizzazione offerte dalla piattaforma utilizzata. A meno che non si utilizzi un servizio di base (tipicamente in modalità *as-a-service*) con un numero limitato di report e di keyword di ricerca, un sistema di Social Business Intelligence si incentra sul processo iterativo mostrato in Figura 4.1. Le fasi possono essere cosi descritte:

1. **Analisi del dominio di ascolto**: viene svolta con gli utenti di business ed è finalizzata a circoscrivere la porzione di Web da analizzare, individuando i topic di interesse (ed i loro alias) che saranno in seguito classificati in temi, e scegliendo le sorgenti informative da scandagliare per il recupero di dati interessanti per l'analisi. La possibilità di identificare classificazioni complesse determina poi la capacità di svolgere analisi sofisticate sui dati.

2. **Definizione condizioni di ricerca**: finalizzata ad individuare le parole chiavi e le condizioni di ricerca che saranno utilizzate dal crawler per acquisire le clip dal web e dai social network.
3. **Creazione della conoscenza di dominio**: finalizzata a inserire nel motore di arricchimento semantico dei testi le specificità del linguaggio del dominio di ascolto al fine di migliorare la corretta interpretazione dei testi. Questa è certamente la fase più delicata del processo di verticalizzazione e la sua ampiezza dipende principalmente dal livello di sofisticazione del motore semantico utilizzato per il sistema di SBI.
4. **Acquisizione ed analisi dei dati**: in questa fase il sistema è messo in esecuzione, le clip vengono raccolte dal crawler ed in seguito elaborate ed arricchite dal motore semantico.
5. **Diagnosi dei risultati**: questa fase può essere svolta solo dopo che, nella fase 4, il sistema è stato messo in funzione ed ha raccolto ed elaborato una serie di dati. Il suo obiettivo è quello di permettere di identificare gli aggiustamenti da apportare a chiavi di ricerca e a conoscenza di dominio al fine di migliorare ulteriormente le performance

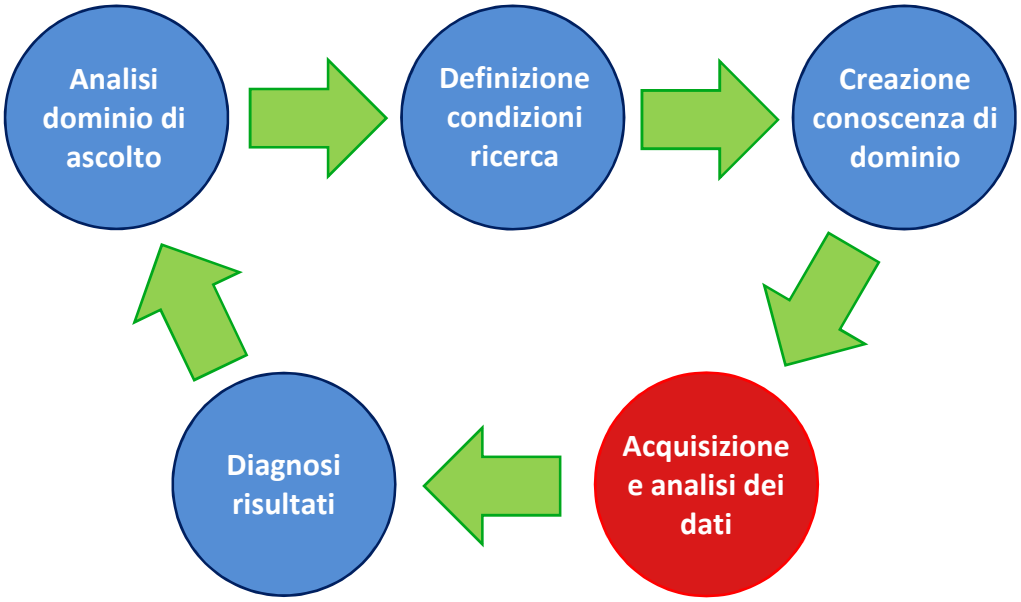

Fig 4.1: Ciclo di verticalizzazione di un sistema di Social BI

4.1 Analisi del dominio di ascolto

Dietro l'espressione "analisi del dominio di ascolto" si cela un preciso insieme di attività che deve essere negoziato, gestito e organizzato in modo congiunto dagli sviluppatori del sistema di Social BI e dall'ipotetico cliente che ha demandato i lavori. Analisi del dominio di ascolto significa comprendere, cercare, scovare e definire univocamente cosa è importante e di rilievo per il vertical oggetto del sistema di social BI e, una volta messi a fuoco questi punti, scovare e delineare le fonti di informazione che si ritengono più prolifere e rilevanti per estrapolare opinioni ed notizie di interesse. Questo lavoro viene svolto in primissima istanza in modo "manuale": a tavolino, sviluppatori e cliente si accordano, colloquiano, cercano cosa (temi, topic, alias) è importante per il dominio e da dove (*source selection*) devono essere prelevati i documenti relativi a quanto ritenuto di interesse per il dominio. Nelle istanze successive alla prima, gli attori del processo possono scegliere se cambiare le sorgenti informative, se arricchire/impoverire l'insieme di tali sorgenti con l'aggiunta o la cancellazione di fonti documentali, e se aggiungere/modificare/eliminare temi/topic/alias in virtù dei risultati forniti dall'analisi eseguita dal motore semantico verticalizzato. Tutto questo è "analisi del dominio di ascolto", ed è il primo step della metodologia di verticalizzazione.

4.1.1 Topic Discovery

Come accennato precedentemente, la prima volta che un sistema di Social BI viene messo in funzione, non ha dati da analizzare dai quali estrapolare informazioni utili per la scelta di nuovi temi, argomenti, alias. La prima iterazione di *topic discovery* generalmente viene fatta a partire dalla conoscenza propria di base, e da quella che si accumula parlando con gli esperti del dominio e navigando in rete.

Dopo aver identificato con il cliente una serie di temi influenti ed interessanti del vertical politico (ambiente, sanità, politica interna, ecc), ci si è adoperati per individuare gli argomenti più caldi ad essa facenti capo. Data dunque questa lista di

temi,. Tra i vari scenari di ricerca possibili nei quali ci si poteva calare, si è scelto di utilizzare due servizi targati Google: *AdWords*[6] e *Trends*[7]. Google AdWords è il sistema *pay per clic* (ppc) più diffuso in Italia, dove con ppc non si intende altro che un' azione di marketing online la quale genera immediatamente traffico altamente targhettizzato; Trends invece è uno strumento che permette di scoprire quali sono le keywords maggiormente ricercate relative ad una parola chiave immessa come riferimento, rendendo inoltre disponibili grafici e statistiche relativi ai risultati della ricerca.

Ci siamo forniti di Google AdWords per individuare, dato un certo tema, quale fosse il volume di traffico generato sulla rete, e le tematiche correlate.

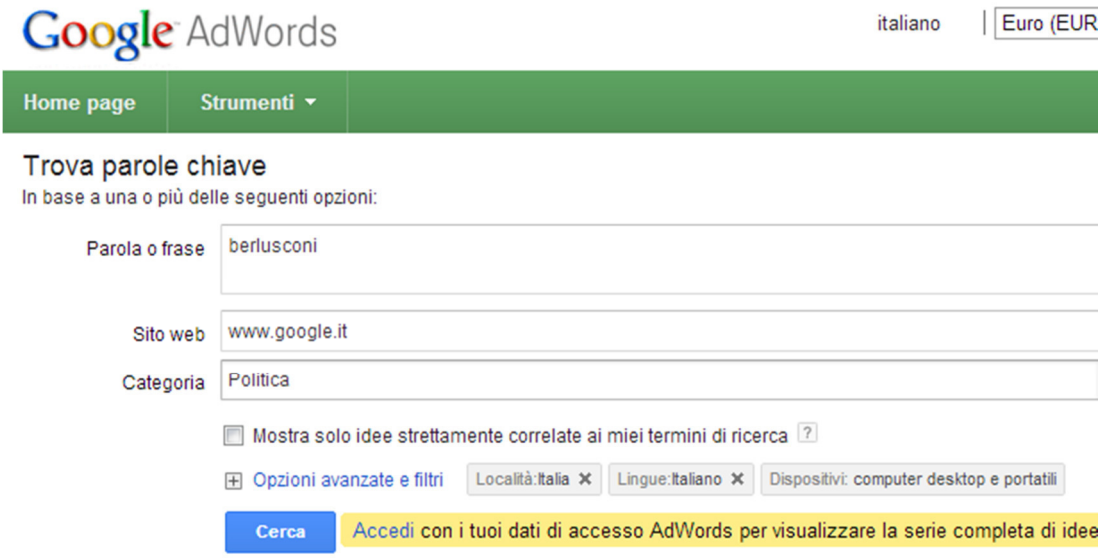

Fig 4.2: Trovare le keyword con Google AdWords

Utilizzando ad esempio come chiave di ricerca "berlusconi" (Fig 4.2), AdWords restituisce in output una lista di idee di possibili parole chiave affiliate a quella utilizzata per eseguire la ricerca, con associata ad ognuna di esse una lista di informazioni relative al volume di ricerche che la coinvolgono (Fig 4.3).

[6] *https://adwords.google.it/*
[7] *http://www.google.com/trends/?hl=it*

Parola chiave	Concorrenza	Ricerche mensili globali	Ricerche mensili locali
berlusconi	Bassa	823.000	550.000

Salva tutto | Idee per le parole chiave (100) | 1 - 50 di 100

Parola chiave	Concorrenza	Ricerche mensili globali	Ricerche mensili locali
governo berlusconi	Bassa	8.100	8.100
berlusconi news	Bassa	5.400	4.400
berlusconi wikipedia	Bassa	6.600	2.900
villa berlusconi arcore	Bassa	1.300	1.000
arcore villa berlusconi	Bassa	1.300	1.000

Fig 4.3: Output della ricerca eseguita con Google AdWords

Espandendo uno ritenuto interessante tra gli output riportati (es: "governo berlusconi"), ci viene data la possibilità di passare a Google Trends (Fig 4.4), e visualizzare delle statistiche più puntuali relative a quel termine.

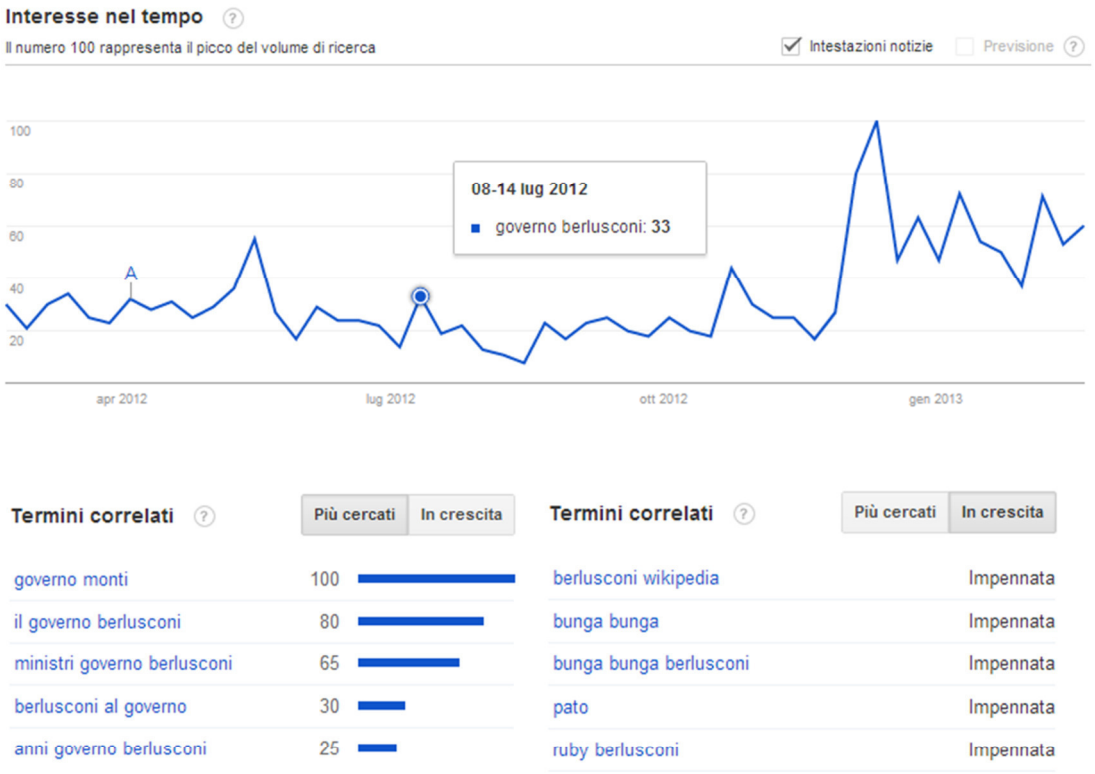

Fig 4.4: Output di Google Trends

Partendo dalla lista dei temi di politica scelti inizialmente e seguendo questo iter di ricerca degli argomenti interessanti, ci siamo creati una prima lista ibrida di 79 termini tra topic e loro alias, ognuno di essi facente capo ad uno o più temi (ad esempio, il topic “Gelmini” appartiene sia al tema “istruzione” che al tema “politici e partiti”).

4.1.2 Source Selection

Una tra le primissime parti ad essere azionata nel funzionamento di un sistema del genere è ovviamente quella relativa al recupero delle informazioni che dovranno essere elaborate. A differenza di un sistema tradizionale di Business Intelligence che recupera i dati da un data warehouse aziendale, questo sistema di Social BI andrà a lavorare su dati recuperati direttamente dal web. Ciò implica una profonda dicotomia tra le due tipologie di sistemi, determinata principalmente su due caratteristiche del dato: il “luogo” in cui il dato si trova e la “forma” del dato stesso. Mentre nel caso di un sistema di BI tradizionale si può assumere, senza paura di sbagliare, che i dati si trovino all’interno di una base di dati ben precisa questo non vale assolutamente per il nostro sistema di SBI il quale viene alimentato da informazioni che possono potenzialmente risiedere sull’hard disk di un computer che accede alla Rete da un qualsiasi angolo del mondo. Per quanto riguarda invece il formato dei dati con i quali si trovano a lavorare i due sistemi, nel caso di BI è certo che un dato, prima di essere memorizzato nel DW aziendale, venga normalizzato e predisposto per essere impiegato in un processo di elaborazione, ciò purtroppo non vale nel caso di SBI. Il sistema infatti, attingendo dalle fonti più disparate, si trova a dover manipolare dati totalmente destrutturati i quali hanno come unico denominatore comune, nel nostro caso applicativo specifico, l’essere testi scritti.

La classificazione delle fonti è funzionale alla possibilità di trattare in modo differente le informazioni a seconda del canale che le ha generate ma può anche servire, molto più semplicemente, a mantenere controllato il volume di informazione

significativa generato da ogni fonte in modo da poter dedurre indicazioni utili come per esempio quale fonte è più attenta e/o sensibile al tema della politica, quale si dimostra più attendibile ed altre informazioni. Ovviamente la classificazione non poteva consistere in un elenco esaustivo dei siti e dei portali web, sia per la proibitiva numerosità dei suddetti ma soprattutto perché il web è un mondo in continuo fermento e le fonti sia avvicendano con una frequenza molto elevata. È stata così redatta una classificazione delle sorgenti di dati, questa prevede 2 macro-categorie:

- **Standard** vengono considerate standard tutte quelle fonti i cui contenuti sono generati da una persona qualificata, come un giornalista, e che quindi presumibilmente non contengono errori ortografici e/o di altri tipi; un esempio di fonte standard può essere considerato il sito internet di un quotidiano di informazione

Fig 4.5: Standard vs. Social

- **Social** vengono considerate social tutte quelle fonti i cui contenuti sono generati da utenti comuni; rientrano dunque in questa categoria fonti come i social network, o i blog di utenti comuni

Selezionare le fonti significa sostanzialmente verificare da un insieme di sorgenti informative (finito nel nostro case study) concordate tra sviluppatori ed esperti del dominio, quelle che sono più *in-target* per il nostro vertical (una fonte è tanto più in-target quando più è grande al suo interno la quantità di contenuti appartenenti al

dominio di interesse), e successivamente sceglierle come contenitore di notizie da cui pescare informazioni attraverso il processo di crawling. Come strumento di ricerca abbiamo utilizzato Google, che è forse ad oggi il miglior strumento in circolazione per il reperimento di qualsiasi dato. Quello che abbiamo fatto è stato creare delle semplici query basate su keyword politiche molto popolari al momento della source selection, e lanciarle sul motore di ricerca. In base alla mole di risultati restituiti per ogni sorgente citata nella query, abbiamo scelto le fonti da utilizzare (vedi esempio in Figura 4.6).

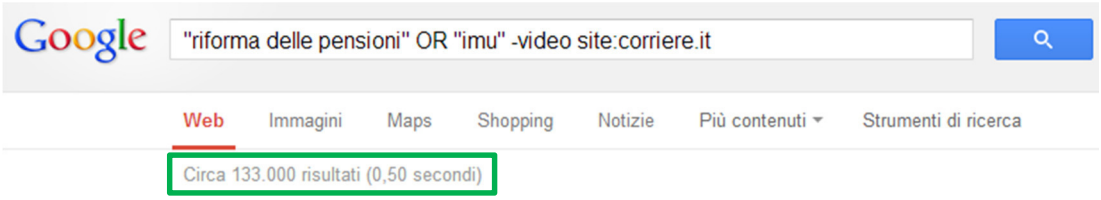

Fig 4.6: Esempio query di "source selection"

La Figura 4.6 ci mostra che la query `riforma delle pensioni OR imu -video site:corriere.it` lanciata su Google restituisce 133.000 risultati: questo significa che il motore di ricerca ha riscontrato che in 133.000 occasioni, sul sito *corriere.it*, compare "riforma delle pensioni" oppure "imu". In accordo con il team di lavoro, sono state ideate le seguenti tre query per stilare la classifica delle fonti

- `"intercettazioni napolitano" -video site:sitoesempio.it`
- `"riforma delle pensioni" OR disoccupazione OR imu OR intercettazioni OR "[riforma | legge] elettorale" -video site:sitoesempio.it`
- `"Elezioni politiche 2013" -video site:sitoesempio.it`

I risultati restituiti da Google per quel che riguarda le fonti standard sono riassunti in Tabella 4.1

query	rank	source	results
1	1	ilgiornale.it	36000
	2	repubblica.it	25800
	3	ilfattoquotidiano.it	4100
	4	ilsole24ore.com	6240
	5	ansa.it	2890
	6	corriere.it	2400
	7	rai.it	2050
	8	sky.it	380
2	1	corriere.it	154000
	2	ansa.it	127000
	3	ilsole24ore.com	79100
	4	ilgiornale.it	72200
	5	repubblica.it	58400
	6	rai.it	33700
	7	sky.it	19700
	8	ilfattoquotidiano.it	14800
3	1	ilsole24ore.com	1440
	2	repubblica.it	1320
	3	ilgiornale.it	865
	4	corriere.it	660
	5	ansa.it	563
	6	rai.it	514
	7	sky.it	81
	8	ilfattoquotidiano.it	14

Tab 4.1: Classifica per query delle fonti Standard

Attraverso una mera somma algebrica delle occorrenze rilevate per ciascuna query di analisi, si è stilata la classifica finale per le fonti standard, e si sono scelte le prime cinque classificate

rank	source	results
1	corriere.it	157060
2	ansa.it	130450
3	ilgiornale.it	109060
4	ilsole24ore.com	86780
5	repubblica.it	85520
6	polisblog	32498

Tab 4.2: Classifica finale fonti Standard

In Tabella 4.2 vediamo sei voci anziché cinque, in quanto è stata aggiunta tra le fonti standard anche il blog di politica "Polisblog", che a differenza delle altre fonti non è il sito di un noto quotidiano, ma può comunque considerarsi una fonte standard di interesse sull'informazione politica.

Come fonti social invece si è scelto di utilizzare i due social network più popolati da sempre e di sempre: Facebook e Twitter.

4.2 Definizione delle condizioni di ricerca

Definire le condizioni di ricerca significa implicitamente completare due processi, uno propedeutico all'altro:

- identificare le keyword su cui incentrare la ricerca dei documenti
- parametrizzare il crawler, in funzione di tali keyword, sulle sorgenti definite in fase di source selection

4.2.1 Identificazione delle keyword

l'identificazione delle keyword è un processo che consiste nell'astrazione, dai topic e dagli alias emersi al passo precedente, dei termini ritenuti dagli esperti come "più significativi" per il reperimento di informazioni ed opinioni di interesse. Bisogna

però prestare attenzione a non usare keyword troppo generiche che rischiano di fare recuperare al crawler una quantità enorme di risultati *off-target*. Si pensi per esempio all'utilizzo della parola "leader": sicuramente il suo impiego determinerebbe il recupero di risultati interessanti per il nostro vertical politico, ma implicherebbe anche il recupero di moltissime informazioni di scarso interesse solo perché contenenti locuzioni dialettali all'interno delle quali viene impiegata la parola "leader": come "Balotelli è subito divenuto leader dell'attacco milanista" (oltre al non trascurabile fatto che la stessa parola con lo stesso significato esiste in inglese, il che significa l'entrata nel sistema di moltissimi dati in una lingua non gestita dal sistema stesso, dati quindi inutili). D'altro canto l'utilizzo di keyword troppo specifiche rischierebbe di indirizzare involontariamente l'attività di ricerca portando così alla perdita di risultati d'interesse. Il rischio derivante dall'utilizzo di parole chiave ambigue sorge invece in un numero di casi non individuabili secondo schemi fissi, queste situazioni devono essere quindi previste o gestite appena individuate. A valle di queste considerazioni, emerge come sia importante non sottovalutare questa fase, e procedere quindi con cautela nello scegliere le parole chiave.

4.2.2 Set up del crawler

Visto che ogni fonte standard da noi scelta presenta sul proprio sito la sezione "Feed RSS", abbiamo deciso di sfruttare questa comodità scegliendo di creare all'interno del crawler tutte fonti documentali di tipo RSS, in modo tale da andare a popolare le fonti attraverso le news pubblicate nella sezione feed del relativo sito.

Fig 4.7: Sezione RSS sul sito di ANSA.it

Il passo seguente alla definizione della fonte di tipo Feed RSS, è la sua parametrizzazione puntuale.

Fig 4.8: Parametrizzazione flusso RSS

Come mostra la Figura 4.8, la parametrizzazione del flusso RSS richiede il riempimento/selezione dei seguenti campi:

- **Flusso RSS** qui dentro andiamo ad inserire il flusso RSS desiderato (es: *http://feeds.ilsole24ore.com/c/32276/f/566665/index.rss*)
- **Contenuto** scegliamo il tipo di contenuto che desideriamo estrarre. Abbiamo due opzioni:
  - **Titolo e abstract RSS** ci consente di estrarre soltanto il titolo e una breve descrizione dell'articolo. Scegliendo questa opzione, non è necessario specificare una regola di segmentazione (vedi in seguito in *regole di*

segmentazione) ed è sufficiente scegliere come tipo di documenti "NOSEZIONI" (vedi in seguito in *tipo documenti*)

- **URL referenziata (la nostra scelta)** scegliamo questa opzione per estrarre tutti i documenti referenziati da URL nella pagina, disponibili per un determinato flusso RSS in base a una regola di segmentazione selezionabile scegliendo l'opzione "CLASSEDIV" (vedi in *tipo documenti*)

- **Filtri su titolo e abstract** indipendentemente dal tipo di contenuto selezionato, è possibile filtrare i documenti estratti in base a parole chiave. Il filtro ha effetto sui titoli e sugli abstract. Per applicare un filtro, è necessario digitare nel campo "Filtro su titolo e abstract" le parole chiave desiderate. Se ad esempio desideriamo estrarre tutti i documenti politici che includono notizie sul "PdL", andiamo a digitarlo all'interno di questo campo. Nel caso invece non volessimo alcuna notizia riguardante il PdL, dovremmo digitare "^PdL". È qui che abbiamo inserito in massa le keyword scelte a valle della fase di Topic Discovery
- **Tipo documenti** qui si sceglie la tipologia di documento che si desidera estrarre
- **Regole di segmentazione** Per determinare la regola di segmentazione più appropriata, è necessario aprire la pagina web desiderata e visualizzare il codice HTML come illustreremo nell'esempio seguente.

Con riferimento alla Figura 4.9, si supponga di voler escludere dalla segmentazione le parti contrassegnate dai numeri 2 e 3, evidenziate rispettivamente con i colori blu e verde, che nel codice HTML sorgente della pagina sono racchiuse tra DIV di classe diversa, ma comunque comprese all'interno del DIV più esterno contrassegnato dal numero 1, che invece vogliamo recuperare.

Fig 4.9: Esempio regole di segmentazione

La sezione 1 (colore rosso) che vogliamo includere nell'estrazione e che costituisce il testo vero e proprio della notizia è racchiusa tra tag DIV di classe `content article`, mentre la sezione 2 (blu) e la sezione 3 (verde) che vogliamo escludere dall'estrazione, sono racchiuse tra tag DIV rispettivamente di classe `news-photo` e `tools-bar`.

```
<html xmls=http://www.w3.org/1999/xhtml xml:lang="it">
    ...
    <div class="content article">
        ...
        <div class="news-photo">
        ...
        </div>
        ...
        <div class="tools-bar">
        ...
        </div>
        ...
    </div>
    ...
</html>
```

Una volta identificate all'interno del sorgente HTML le classi DIV di interesse e quelle invece che si intende scartare, bisogna andarle ad inserire all'interno delle regole di segmentazione. Quindi, dopo aver scelto "CLASSEDIV" come *tipo di documento* ed "URL referenziata" come *contenuto*, ci si comporterà come segue:

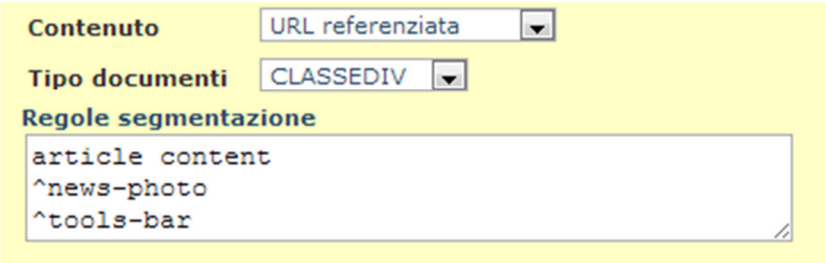

Fig 4.10: Fill in del campo "regole segmentazione"

In questo modo, dal documento andremo ad estrarre il contenuto del tag DIV con classe `article content`, ma dai esso verranno scartati i contenuti dei tag DIV con classi `news-photo` e `tools-bar`.

Si potrebbe erroneamente pensare che una volta settato il crawler per una particolare fonte, non sia più necessario mettervi le mani. Questo non è affatto vero, ed è buona norma andare a controllare ad intervalli regolari che tutto proceda regolarmente, in quanto capita di sovente che i siti web modifichino la struttura delle proprie pagine. In questi casi, le regole di segmentazione definite precedentemente non lavorerebbero più in modo adeguato, andando a cercare di recuperare il contenuto di tag che magari ora non esistono più, o che contengono informazioni diverse dalla tipologia di quelle memorizzate in precedenza. Questa azione di controllo e messa a punto del crawler prende il nome di *tuning*.

4.3 Creazione della conoscenza di dominio

Una peculiarità di SyNTHEMA è quella di disporre già di una tassonomia di termini standard della lingua italiana all'interno della propria ontologia di dominio (Fig

4.11), con associata una polarizzazione per cosi dire "non verticalizzata" dei concetti (non per forza un concetto deve esser polarizzato in modo negativo o positivo: può anche aver associato un sentimento neutrale).

Termine	POS	Sent	Alias	Senso	Lingua	Frq	Relazioni	Azioni
vittima	N					381		
molto	G					374	2 relazioni	
piccolo	G					368	7 relazioni	
poco	G					363	10 relazioni	
vero	G					346	7 relazioni	
bellezza	N					342		
alto	G					341	35 relazioni	
riuscire	V					315	8 relazioni	
buono	G					306	28 relazioni	
sospetto	N					286	2 relazioni	

Fig 4.11: Porzione della tassonomia di base nella KB di SyNTHEMA

Ogni volta che il motore semantico analizza un testo, esso indicizza i termini presenti al suo interno e va ad inserirli, se non sono già definiti, all'interno della base di conoscenza come tipo "standard" con senso comune e caratteristiche definite automaticamente dal motore. Nel caso il termine sia già definito, l'unico dato che va ad arricchirsi è quello relativo alla "frequenza" fino a quel momento riscontrata del termine.

La creazione della conoscenza di dominio, e quindi la verticalizzazione, avviene tra due analisi distinte: il motore semantico produce, al termine di un'analisi, della conoscenza grezza, che va verticalizzata in vista dell'analisi successiva. Agendo in questo modo, la volta successiva che il motore andrà a lavorare su dei documenti testuali, li analizzerà in modo più orientato al nostro vertical politico rispetto all'analisi eseguita all'iterazione precedente. Verticalizzare la conoscenza, significa andare ad aggiungere e/o modificare relazioni e termini presenti nella KB di SyNTHEMA in funzione del dominio su cui stiamo lavorando: sia che essi siano

memorizzati sin dal principio perché presenti nella versione standard della tassonomia, o che essi siano stati aggiunti alla KB a valle di un'analisi del motore. Più lo strumento lavora, migliori sono i risultati che si ottengono di volta in volta. Chiaramente alle prime iterazioni, le migliorie e modifiche da apportare saranno ovviamente in misura molto superiore rispetto a quelle successive, perché agli inizi lo strumento non è orientato al dominio di interesse. Ciò che ci si aspetta, è che la mole di lavoro di verticalizzazione diminuisca con il passare del tempo.

4.4 Acquisizione ed analisi dei dati

La parte di acquisizione ed analisi dei dati ingloba al suo interno due fasi molto importanti: la raccolta delle clip dal web attraverso il processo di crawling (acquisizione), e la loro analisi da parte del motore semantico. Questo è anche il momento in cui si vanno a popolare del tabelle del database descritto nel paragrafo 4.3, attraverso l'ausilio di TOS Data Integration. Andremo ora ad illustrare brevemente in che modo e con quale ordine vengono eseguire le analisi dal motore e popolate le tabelle.

4.4.1 Trasferimento dei documenti dal repository del crawler al database

Una volta che il processo di crawling si conclude, i documenti recuperati dalle fonti documentali definite nel crawler vengono tutti quanti indicizzati e salvati all'interno del repository di iSyN SC (Fig 4.12).

Fig 4.12: Repository del crawler di SyNTHEMA

È necessario fare una piccola premessa: come abbiamo più volte detto, un sistema di Social BI si serve di un crawler per il recupero delle informazioni da analizzare. Il *provider* del servizio di crawling nella maggior parte dei casi è diverso dal motore semantico: il primo servizio si fruisce as-a-service il secondo si può anche internalizzare. Questo significa che solitamente il motore semantico non ha accesso ai contenuti del repository del crawler. Volendo creare una soluzione di Social BI universalmente corretta, abbiamo creato con l'ausilio di Talend una procedura avente il compito di prelevare le clip recuperate dal crawler e collocate sul proprio repository, e salvarle all'interno di una base dati accessibile dal motore semantico.

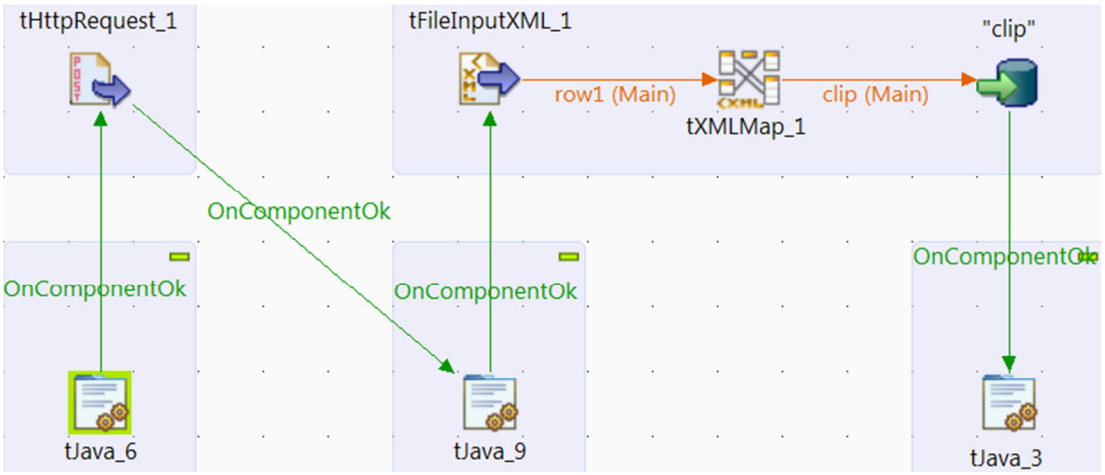

Fig 4.13: Trasferimento clip dal repository del crawler al tabella clip del database

Nel nostro caso specifico non era necessario realizzare questo passaggio, in quanto il motore semantico ha accesso al repository del crawler. Ma dovendo e volendo realizzare una versione universalmente corretta, in cui il servizio di crawling può benissimo non essere fornito dallo stesso provider del motore semantico, abbiamo implementato il *job* mostrato in Figura 4.13 attraverso l'utilizzo del modulo Data Integration di TOS che popola la tabella "clip" del database con i documenti prelevati dal crawler. Il vantaggio più grande derivante dalla realizzazione di questo step è legato al fatto che se in un futuro si volesse cambiare il partner di crawling, non vi sarebbe alcun problema, in quanto il motore semantico lavora in modo

scollegato dal crawler andando a prelevare i documenti dal database, e non dal suo repository.

4.4.2 Autoanello di analisi

Eseguito il trasferimento sul database delle clip raccolte dal crawler, ha inizio quello che nella parte di architettura funzionale del sistema di Social BI (capitolo 2.7) abbiamo definito "autoanello di analisi".

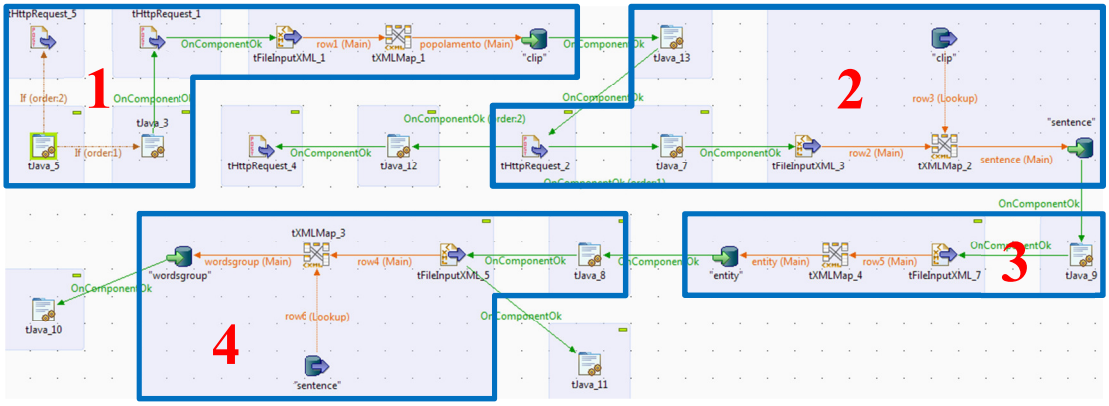

Fig 4.14: Job Talend relativo all'autoanello di analisi

Facendo riferimento a quanto indicato in Figura 4.14, gli step di analisi eseguiti dal motore semantico sono i seguenti:

1. Le clip "grezze" salvate sul database vengono prelevate e date in pasto al motore semantico, che le analizza, e le riscrive (non sovrascrive: le clip "non lavorate" rimangono salvate) sul database nella tabella "clip" andando a valorizzare anche il campo "sentiment" relativo alla polarizzazione della clip con un valore numerico negativo, positivo, o neutro a seconda del risultato dell'analisi.
2. Il motore semantico spezza in frasi (seguendo regole proprie del motore) ogni clip analizzata, ed associa ad ogni frase una polarizzazione basandosi sulle informazioni presenti nella base di conoscenza. Terminata la lavorazione, le

sentence vengono salvate nella corrispondente tabella del database con il proprio sentiment associato.

3. In questa fase avviene l'estrapolazione dalle clip di tutte le entità rilevate. Ogni entità poi viene salvata nella tabella "entity" del database se e solo se non era già stata salvata in un ciclo di analisi temporalmente anteriore a quello attuale. Oltre al lemma della entità, nel record della tabella viene valorizzato anche l'attributo POS corrispondente.
4. L'ultimo step è quello che si occupa di valorizzare la tabella "wordsgroup" all'interno del database. Il motore estrapola da ogni frase presente in ogni clip tutti i gruppi di parole individuati, e va a valorizzare un record per ognuno di essi scrivendo di ognuno: polarizzazione, sentiment, prima parola del gruppo, seconda parola del gruppo, tipo di relazione che lega le parole.

4.5 Diagnosi dei risultati

Terminata l'analisi del motore, è tempo di diagnosi. Attraverso l'utilizzo di opportune query SQL si interrogano le tabelle del database per scoprire quali e quanti dati sono entrati, e con quale qualità. Un'analisi di questo tipo, ci mette in condizione di individuare le modifiche da apportare alla base di conoscenza, quali chiavi di ricerca aggiungere o togliere, quali topic/temi/alias aggiungere. Modifiche della KB a parte, gli altri interventi sono tutti attuabili fianco a fianco con il cliente, che in quanto esperto del dominio di interesse, saprà e dovrà consigliare quali modifiche attuare tra quelle presentategli, e quali no.

Capitolo 5

Analisi dell'efficacia

Al fine di fornire una valutazione quantitativa dell'efficacia del sistema di Social BI si riportano i risultati di un insieme di test svolti sul dominio di ascolto della politica italiana. L'attività di testing si è concentrata sul tema della correttezza delle operazioni di opinion mining che risulta essere la funzionalità più complessa da automatizzare e che necessita di una costante attività di perfezionamento e controllo a seguito della rapida evoluzione del linguaggio e dei contenuti presenti nel web. Oltre a fornire una valutazione "statica" dell'efficacia i test hanno inoltre l'obiettivo di fornire al lettore una indicazione circa la variazione delle performance a seguito delle attività di verticalizzazione.

5.1 I benchmark e l'organizzazione dei test

La Tabella 5.1 sintetizza la composizione dei dataset utilizzati. Il primo dei due (*Dataset Standard*) si compone di testi recuperati dalle cosiddette fonti "qualificate" ovvero tutti quei siti solitamente riconducibili a un quotidiano, ad una rivista o

comunque a siti web che hanno nella pubblicazione di notizie il loro *core-business*. L'attingere informazioni da queste fonti garantisce l'estrazione di testi corretti dal punto di vista sintattico e scritti utilizzando un linguaggio il più possibile coerente con le regole della grammatica italiana, il che costituisce sicuramente un aspetto importante quando si valutano i risultati dell'analisi effettuata da un motore semantico. Il secondo dataset utilizzato (*Dataset Social*) è stato invece costruito a partire da dati estratti da fonti "social", nella fattispecie i social network Facebook e Twitter. I testi recuperabili hanno la peculiarità di essere scritti nella grande maggioranza dei casi utilizzando il cosiddetto dialetto del web: ovvero una scrittura piena di termini dialettali, neologismi e forme sintattiche poco aderenti alla formulazione canonica prevista dalla grammatica italiana. Questa caratteristica non può essere trascurata in fase di valutazione dell'analisi in quanto potrebbe renderla difficoltosa influenzandola negativamente. Si sottolinea che, al fine di separare i fattori di complessità crawler – motore semantico, i dataset sono composti da testi selezionati manualmente tra quelli restituiti dal crawler e non dal risultato grezzo del crawler stesso. Ciò significa che si sono scelti frasi complete da punto a punto assumendo quindi che il crawler sia stato in grado di eliminare eventuali tag presenti nel testo e di segmentare correttamente un documento in più frasi. I due dataset sono stati etichettati da un gruppo di 5 esperti classificandoli, oltre che in "positivi", "negativi" o "neutri" in "facili" e "difficili" sulla base di quanto difficoltosa sia risultata l'analisi linguistica e la polarizzazione manuale degli stessi. Oltre alla loro composizione in termini puramente quantitativi, la Tabella 5.1 riporta anche i dati relativi alla concordanza media relativa ai dataset nel loro complesso e a ogni sottocategoria. Per concordanza media si intende la media di coincidenza del giudizio dato da ogni esperto rispetto al giudizio degli altri in fase di polarizzazione dei testi che costituiscono il dataset. Questa informazione permette di misurare l'effettiva complessità del dataset in quanto determina il tasso di errore medio di un utente umano rispetto all'opinione espressa da un oracolo teorico che fornisca sempre la risposta corretta. La risposta dell'oracolo è definita come l'opinione preponderante (*Majority Group*) tra le valutazioni degli esperti. Si noti che la somma

di tutti i testi facili e difficili non coincide esattamente con la somma di quelli positivi, negativi e neutri poiché per questi ultimi non sempre è stato possibile identificare una polarizzazione prevalente.

L'analisi del dato di concordanza è in linea con gli studi accademici che riportano che un valore medio di *inter-tagger agreement* dell' 80% (Gliozzo & Strapparava, 2009). Questo risultato aiuta a settare le aspettative sul grado di precisione di uno strumento automatico di sentiment analysis visto che è improbabile che questo superi le capacità di un utente umano.

		Positivi	Negativi	Neutri	Totale	AVG concord.
Dataset Standard	Facile	174	255	222	656	92,10%
	Difficile	82	122	99	310	85,10%
	Totale	256	377	321	966	89,20%
Dataset Social	Facile	83	119	37	244	93,88%
	Difficile	43	148	47	239	82,30%
	Totale	126	267	84	483	88,54%

Tab 5.1: Composizione dei dataset usati per il testing del sistema e concordanza media inter-tagger

I dataset sono stati sottoposti a più cicli di valutazione a seguito del processo di verticalizzazione descritto in Figura 4.1. in Tabella 5.2 sono sintetizzate le operazioni svolte ad ogni iterazione e la relativa durata. Le modifiche, ad ogni ciclo di verticalizzazione, sono state mirate verso quelle clip ove veniva riscontrata una totale inversione di polarizzazione del sentiment associato dal motore semantico rispetto a quello riscontrato dal Majority Group: questo perché si è pensato, in linea teorica, che gli errori più gravi (e quindi più meritevoli di esser soggetti ad una correzione) si condensassero in quella casistica dove uomo e macchina dicono l'esatto opposto.

Iterazione	Descrizione	Durata G/U
R1-R2	• Clip corrette: **41** • Termini aggiunti: **1** • Termini modificati: **39** • Relazioni aggiunte: **46** • Relazioni modificate: **20**	9
R2-R3	• Clip corrette: **30** • Termini aggiunti: **0** • Termini modificati: **45** • Relazioni aggiunte: **39** • Relazioni modificate: **17**	7
R3-R4	• Clip corrette: **38** • Termini aggiunti: **7** • Termini modificati: **40** • Relazioni aggiunte: **55** • Relazioni modificate: **23**	4

Tab 5.2: Descrizione dei cicli di verticalizzazione per SyNTHEMA

Nel passaggio da una release a quella successiva (ogni iterazione in Tabella 5.2), indichiamo con:

- **Clip corrette**, numero di clip su cui sono state eseguite delle modifiche (aggiunta/aggiornamento/cancellazione) a relazioni e/o termini
- **Termini aggiunti**, quantità di termini inseriti ex novo nella base di conoscenza
- **Termini modificati**, quantità di termini su cui è stata eseguita una modifica (che può riguardare POS, sentiment, tipo di concetto)
- **Relazioni aggiunte**, quantità di relazioni inserite ex novo nella base di conoscenza
- **Relazioni modificate**, quantità di relazioni su cui è stata eseguita una modifica (che può riguardare sentiment, tipo di relazione)

"Durata G/U" ci dice i giorni-uomo di lavoro spesi per portare a termine ciascuna iterazione.

5.2 Risultati dei test e considerazioni

Prima di passare all'analisi dei risultati si sottolinea che la percentuale di concordanza base di polarizzazione, quella che si potrebbe ottenere tirando un dado a tre facce (positivo-negativo-neutro), è del 33% non del 50% come si potrebbe essere erroneamente portati a pensare.

La Figura 5.1 mostra la curva di miglioramento delle prestazioni del software a fronte di un'attività di verticalizzazione delle risorse linguistiche specifiche (a destra fonti standard, a sinistra fondi social). Come era lecito aspettarsi, l'entità del miglioramento ottenuto con ogni release si è constatato essere più consistente durante i primi cicli di verticalizzazione poiché sono questi a introdurre le modifiche maggiormente significative volte a risolvere le problematiche che più incidono sulla qualità dell'elaborazione. Un chiaro esempio di quanto appena detto si può osservare nell'incremento esistente tra le release 1 e 2, in particolare il riferimento alla spezzata

rossa del grafico di destra in Figura 5.1, che verte sulle clip del data set Standard etichettate come "difficili".

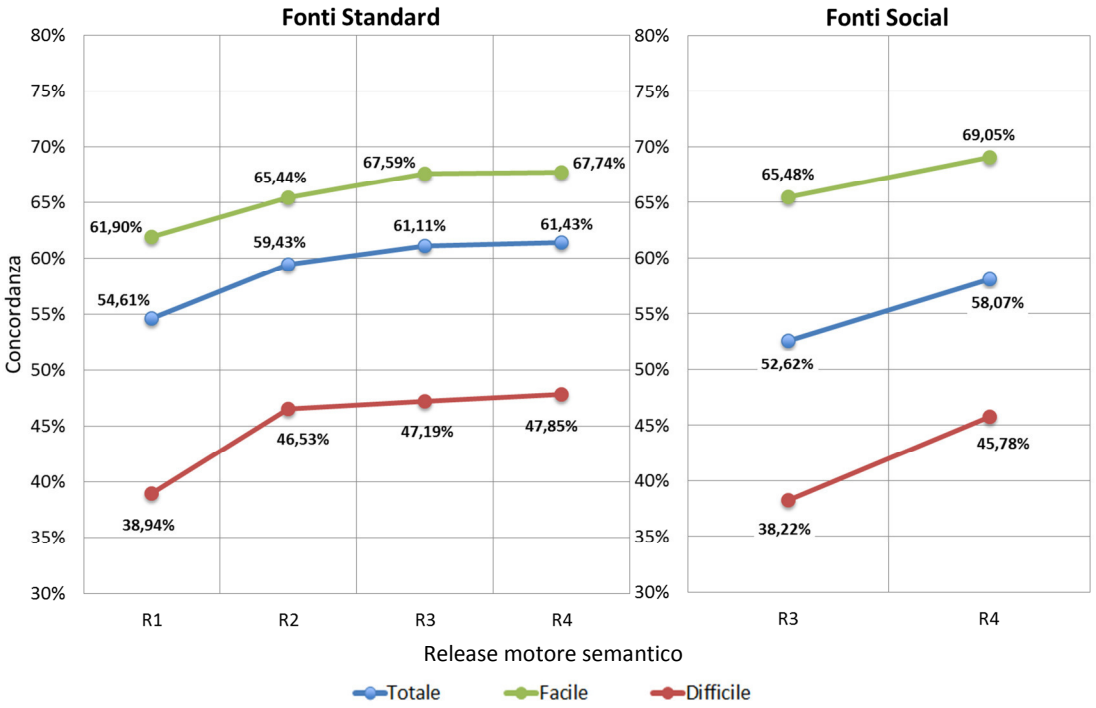

Fig 5.1: Curva di cambiamento della concordanza (easy/hard/totale) rispetto al rilascio di successive versioni delle risorse linguistiche

la curva misurata per SyNTHEMA si dimostra regolare e presenta un miglioramento complessivo dell'analisi contenuto: si parla infatti di un incremento del 6,8% della concordanza media per le fonti standard e 5,45% per le fonti social, derivanti da una correzione di circa 110 clip (70 standard e 40 social) erroneamente polarizzate. La contenuta entità di questo incremento trovano giustificazione nel tipo di situazione da cui è partito il motore, dovuto all'approccio analitico utilizzato. SyNTHEMA grazie all'approccio linguistico registra, indipendentemente dal dominio specifico di utilizzo, risultati fin da subito paragonabili a quelle che saranno le performances finali dello strumento; se avessimo avuto a che fare con un motore semantico basato su di un approccio statistico, avremmo ottenuto inizialmente prestazioni molto peggiori per poi avere un miglioramento più consistente rispetto a quello riscontrato con SyNTHEMA. Il fatto che il processo di verticalizzazione attuato attraverso l'utilizzo di SyNTHEMA abbia portato a miglioramenti contenuti non deve trarre in inganno facendo erroneamente pensare ad un insuccesso. Bisogna tenere a mente tre

elementi che in un progetto reale assumono una rilevanza maggiore rispetto a quella acquisita nello sviluppo di un dimostratore:

1. L'attività di testing è stata condotta su un numero, per quanto ampio, finito di testi, e sulla base degli errori riscontrati in fase di analisi i motori semantici sono stati modificati. Noi che abbiamo curato la verticalizzazione delle risorse linguistiche alla base dell'attività di opinion mining, siamo dell'opinione che per ottenere gli stessi risultati riportati in Figura 5.1 durante un progetto reale, in cui il volume di dati in ingresso e la velocità di generazione degli stessi sono significativamente maggiori, la mole di lavoro ed il tempo necessario siano sicuramente più elevati. Inoltre, a differenza di SyNTHEMA, un motore semantico basato su approccio statistisco necessiterebbe di un processo di verticalizzazione estremamente più corposo a causa della sua natura che ha bisogno, per "imparare" un determinato costrutto o locuzione, di averla incontrata almeno una volta in fase di elaborazione del testo. SyNTHEMA non incorre invece in questo problema che riesce, almeno in parte, a superare grazie ad un'attività di analisi morfo-sintattica dei dati, pagando però questo pesante sfruttamento dell'analisi logica e grammaticale del testo con una minor elasticità in fase di corretta polarizzazione dei costrutti non standard del linguaggio naturale.
2. I risultati della verticalizzazione, soprattutto per quanto riguarda l'approccio statistico, sono tanto migliori quanto più chi effettua il tuning dei motori è in grado di prefigurarsi, e quindi trasferire nei motori, uno scenario complessivo di tutti quelli che sono i pattern di estrazione testuale da impiegare. Ovviamente questa operazione risulta più semplice se l'insieme dei dati utilizzati per il testing è finito e statico.
3. Infine bisogna sempre tenere presente che più le prestazioni di uno strumento migliorano rispetto ad un determinato insieme di dati più si va incontro al rischio di *overfitting*, ovvero creare un mezzo la cui verticalizzazione dia ottimi risultati sul dataset specifico ma non goda assolutamente di generalità

portando ad un considerevole calo delle prestazioni in tutte le altre analisi. Questo rischio specifico è maggiore per i motori che utilizzano un approccio statistico, esiste tuttavia anche per i motori linguistici nel caso si facciano scelte troppo stringenti sulla polarizzazione dei termini specifici del dominio di ascolto: per esempio "crisi" è un vocabolo che deve sempre essere considerato negativo a priori? Verrebbe naturale dire di si ma come comportarsi poi davanti ad espressioni, tutt'altro che rare, come "gestire la crisi", "affrontare i passaggi della crisi", "superare la crisi"?

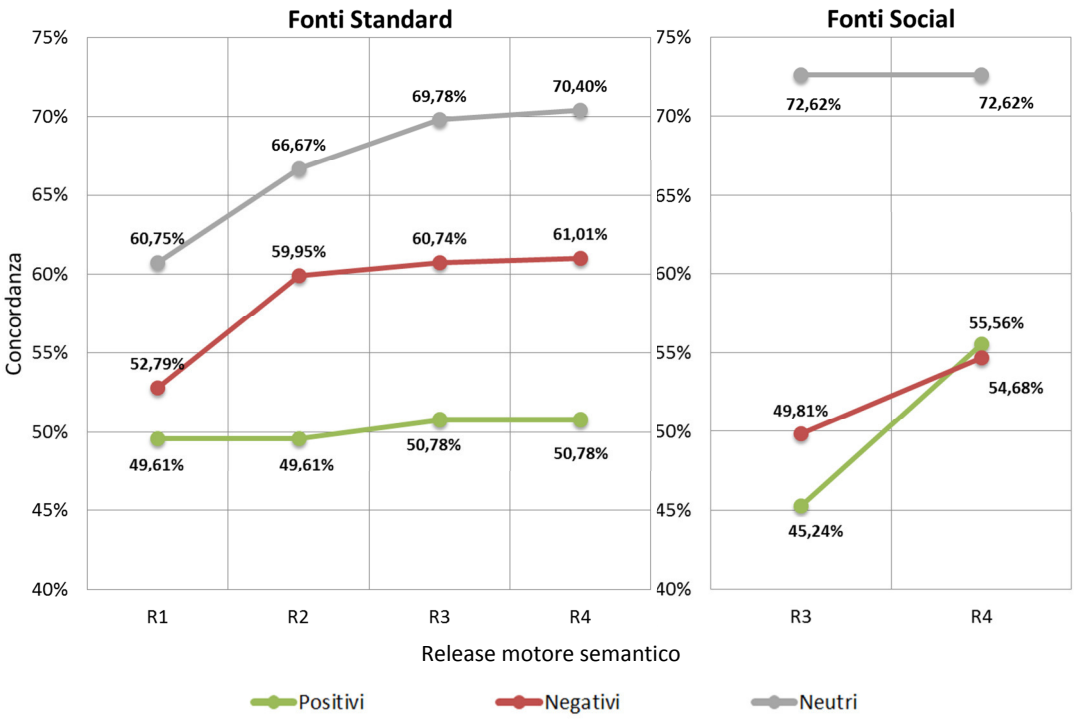

Fig 5.2: Curva di cambiamento della concordanza (pos/neg/neu) rispetto al rilascio di successive versioni delle risorse linguistiche

In Figura 5.2 è invece riportato un confronto sul cambiamento delle percentuali di concordanza dei singoli rilasci di release focalizzandosi non sulla difficoltà dei testi analizzati, come fatto precedentemente, ma sulle tre diverse classi di polarizzazione: positivo, negativo e neutro. Si può osservare dai grafici che sia per quanto riguarda le fonti standard sia per quelle social il motore presenta prestazioni di individuazione

dell'assenza di opinione esplicita (polarizzazione neutra) decisamente superiori rispetto a quelle di estrazione del sentiment positivo o negativo. Questo risultato non è da trascurare in quanto studi riportano come dei documenti estratti dalla rete circa il 70% risulti, all'analisi di più esperti distinti, sostanzialmente neutro (Fitz-Gibbon, 2011).

Interessante è anche notare come, mentre l'incremento che il motore ha avuto nella capacità di individuare correttamente le opinioni negative sia sostanzialmente prossimo al 6-8% in entrambe le tipologie di fonti, la capacità di polarizzare correttamente le frasi positive ha goduto di una crescita significativa soprattutto in ambito social dove l'incremento si avvicina al 10% contro il molto più contenuto 2% rilevato su testi standard. Anche questo risultato merita un'attenzione particolare poiché, come si può notare dai dati riportati in Tabella 5.1, i testi polarizzati positivamente estratti dal web sono molti meno rispetto a quelli negativi, per questo non riuscire ad interpretarli correttamente potrebbe falsare fortemente le conclusioni traibili osservando i risultati dell'analisi.

		neutro		positivi		negativi		tot	
		R3	R4	R3	R4	R3	R4	R3	R4
TW	|MJ,SyN|	46	47	23	26	65	69	134	142
	%(MJ,SyN)	83,64%	85,45%	46,94%	53,06%	44,83%	47,59%	53,82%	57,03%
	Δ(MJ,SyN)	1,82%		6,12%		2,76%		3,21%	
FB	|MJ,SyN|	15	14	34	44	68	77	117	135
	%(MJ,SyN)	51,72%	48,28%	44,16%	57,14%	55,74%	63,11%	51,32%	59,21%
	Δ(MJ,SyN)	-3,45%		12,99%		7,38%		7,89%	
TW+FB	|MJ,SyN|	61	61	57	70	133	146	251	277
	%(MJ,SyN)	72,62%	72,62%	45,24%	55,56%	49,81%	54,68%	52,62%	58,07%
	Δ(MJ,SyN)	0,00%		10,32%		4,87%		5,45%	

Tab 5.3: Cambiamento della concordanza tra R3 ed R4 per le fonti Social

Un'ultima situazione che ci si attendeva di trovare, e che l'analisi ha confermato, è stata quella relativa alle considerazioni sull'incremento della concordanza post-verticalizzazione relativamente alle fonti Social. La Tabella 5.3 riporta: per ogni social network, come varia:

- la cardinalità della concordanza tra SyNTHEMA e il Majority Group (indicata con $|MJ,SyN|$)
- la percentuale della concordanza tra SyNTHEMA e il Majority Group (indicata con $\%(MJ,SyN)$)
- il delta della percentuale di concordanza tra SyNTHEMA e il Majority Group (indicato con $\Delta(MJ,SyN)$)

da una release a quella successiva. Questo dato viene mostrato sia per ogni categoria di polarizzazione (neutro, positivo, negativo), sia in visione totale. ed anche Dalla Tabella 5.3 si evince come, nel passaggio dalla release 3 alla release 4, sia più marcato il miglioramento relativo ai giudizi delle clip prelevate da Facebook (+7,89%) rispetto a quello delle clip relative a Twitter (+3,21%): più del doppio. Questa differenza è legata al fatto che, per quanto in entrambi i social network i testi possano essere mal strutturati e comprensivi di errori ortografici, in Twitter vi è una grossa abbondanza di parole legate tra loro senza spazi e precedute dal carattere *hashtag* '#'. In queste situazioni il motore semantico non riesce a comprendere la parola o la sequenza di parole legate tra loro, e o le esclude dall'analisi, o le include attribuendogli un significato errato.

5.2.1 Testing di un caso reale

Prima di concludere definitivamente il progetto di tesi, abbiamo voluto testare l'efficienza del lavoro di verticalizzazione svolto facendo eseguire al sistema di Social BI un ciclo realistico di lavorazione: utilizzando come release del motore semantico l'ultima prodotta in fase di verticalizzazione, abbiamo fatto analizzare al motore semantico quanto recuperato dalle fonti documentali (senza raffinare quanto prelevato dal crawler), ed estratto dal database un campione post-analisi di 300 frasi lunghe all'incirca quando le clip utilizzate per i benchmark utilizzati nei test pregressi. Perché il campione fosse il più rappresentativo possibile, le 300 sentence prelevate dal database sono state scelte secondo il seguente criterio:

- 50 ultrapositive (sentiment ≥ 2)
- 50 positive (sentiment = 1)
- 100 neutre (sentiment = 0)
- 50 negative (sentiment = -1)
- 50 ultranegative (sentiment ≤ -1)

Dove con "sentiment" intendiamo chiaramente quello associato da SyNTHEMA, e non da noi. Successivamente, come per gli altri test, un gruppo di esperti ha associato ad ogni record di questo campione il proprio parere di polarizzazione e difficoltà, ma non solo. In questo caso è stata raccolta una ulteriore informazione relativa alla qualità della segmentazione del testo eseguita dal motore semantico: mentre nei test precedenti le clip venivano spezzate manualmente in porzioni di testo di senso compiuto, questa volta il lavoro è stato fatto eseguire in maniera totalmente automatica al sistema di Social BI. Nel test si parlerà di "segmentato OK" quando, secondo gli esperti, il motore ha spezzato la frase in modo linguisticamente corretto; si indicherà invece con "segmentato KO" il caso inverso.

SENTENCE TEST (R4)

TOTALE				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	100	139	68	48,92%
positivi	100	49	36	73,47%
negativi	100	108	63	58,33%
	300	296	167	56,42%

EASY				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	47	60	37	61,67%
positivi	45	31	25	80,65%
negativi	44	44	33	75,00%
	136	135	95	70,37%

HARD				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	53	79	31	39,24%
positivi	55	18	11	61,11%
negativi	56	64	30	46,88%
	164	161	72	44,72%

DS STANDARD (R4)

TOTALE				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	476	321	226	70,40%
positivi	199	256	130	50,78%
negativi	291	377	230	61,01%
	966	954	586	61,43%

EASY				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	316	222	168	75,68%
positivi	136	174	97	55,75%
negativi	204	255	176	69,02%
	656	651	441	67,74%

HARD				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	160	99	58	58,59%
positivi	63	82	33	40,24%
negativi	87	122	54	44,26%
	310	303	145	47,85%

Tab 5.4: Vecchio data set standard e nuovo test set a confronto

Per comprendere quanto contenuto nella Tabella 5.4, diamo un'istruzione su come interpretare i dati (il caso evidenziato in giallo).

- SyNTHEMA ha etichettato come neutro 100 sentence su 300
- Il Majority Group ha etichettato come neutro 139 clip su 296 (perché in 4 casi, non ho una maggioranza sulla valutazione del sentiment da parte degli esperti)
- Delle 139 sentence etichettate come neutro dal Majority Group, 68 sono etichettate come neutro anche da SyNTHEMA (nel 48,92% dei casi, SyNTHEMA e il Majority Group hanno espresso la stessa polarizzazione riguardo alla stessa sentence: neutro in questo caso)

La release del motore semantico che ha svolto questo test è stata la quarta (R4), per cui i risultati verranno rapportati (Tabella 5.4) con i quelli relativi al data set Standard ottenuti anch'essi con la release 4 di SyNTHEMA. Prima di effettuare qualsiasi considerazione, bisogna fare una premessa: le statistiche relative al data set Standard si basano su di una mole di 966 campioni, che è più del triplo rispetto alle 300 sentence appena raccolte. Per poter effettuare delle considerazioni proporzionate, è stato necessario svolgere una normalizzazione delle percentuali di concordanza, come illustrato in Tabella 5.5.

	DS std	SyN,MJ	sentence	SyN,MJ
easy	656	67,74%	136	70,37%
hard	310	47,85%	164	44,72%
	966		300	

(67,74 × 70,37) + (47,85 × 44,72) = 69,07%

Tab 5.5: Normalizzazione della concordanza

69,07% è da considerarsi, anziché 56,42%, il risultato di concordanza totale raggiunto con il motore semantico sul test effettuato con le 300 sentence.

A monte dei risultati dei test, la domanda principale che ci si era posti è stata: i risultati saranno affetti da overfitting oppure no? In altre parole, le percentuali di concordanza sul nuovo test saranno molto peggiori rispetto a quelle riscontrate per il data set di training in virtù del fatto che le modifiche alla KB sono state effettuate in funzione degli errori presenti nel data set Standard? Osservando la Tabella 5.4, riscontriamo con molto piacere che la concordanza tra l'uomo ed il motore non peggiora! Sul totale (non facendo distinzione tra easy ed hard), abbiamo un miglioramento della concordanza di +7,64%, sulle clip "easy" c'è una differenza in positivo di 2,63%, e sulle clip "hard" un discostamento (in negativo) di 3,13%.

Il secondo quesito che ci siamo posti in questo test è stato: quanto la segmentazione corretta/errata della frase (eseguita da SyNTHEMA) incide sulla capacità del motore di polarizzare con successo un testo? La Tabella 5.6 ci dice che delle 300 sentence, il gruppo di esperti ha stabilito che 215 sono state segmentate in modo corretto dal motore semantico, mentre 85 in modo errato. Il resto dei dati vanno letti seguendo lo stesso criterio illustrato per la Tabella 5.5.

SEGMENTATO KO				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	29	41	20	48,78%
positivi	36	12	12	100,00%
negativi	20	32	14	43,75%
	85	85	46	54,12%

SEGMENTATO OK				
POLARIZ	SyN	MJ	SyN,MJ	
neutro	71	98	48	48,98%
positivi	64	37	24	64,86%
negativi	80	76	49	64,47%
	215	211	121	57,35%

Tab 5.6: Prestazioni relative alla segmentazione

Come possiamo osservare in Tabella 5.6, la polarizzazione avviene in maniera più concorde all'uomo nel caso di una segmentazione corretta (come volevasi dimostrare), ma il risultato non si discosta di tanto rispetto a quanto ottenuto con una segmentazione errata: il delta è di 3,23%. Da questo fatto possiamo desumere che evidentemente, la presenza di termini polarizzati all'interno di una sentence riesce comunque ha dare una caratterizzazione abbastanza corretta nonostante la frase non sia spezzata in modo propriamente idoneo.

Conclusioni

Con l'esplosione dei *Social Media* e la conseguente moltiplicazione degli *User Generated Content*, la quantità di informazioni potenzialmente interessanti di natura destrutturata sul Web è cresciuta a dismisura. Si è incominciata dunque a diffondere nel mondo IT l'esigenza di avere a disposizione un nuovo tipo di tecnologie che permettessero di gestire ed analizzare in automatico un insieme esteso di testi non strutturati, estrarre da questi le informazioni più rilevanti e classificarli sulla base dell'argomento trattato, risultando in questo modo di estremo aiuto ne processo decisionale dell'impresa.

Con il presente lavoro di tesi si è sviluppato un sistema di Social Business Intelligence in grado di rispondere alle principali esigenze a cui la Social BI deve far fronte. La soluzione offerta si è dimostrata all'altezza delle aspettative, svolgendo correttamente le funzionalità e i compiti progettati: dal crawling delle informazioni, passando per l'analisi dei testi, opinion mining, verticalizzazione delle risorse linguistiche di iSyN SC. Inoltre, Talend Open Studio ha dato prova di essere un

validissimo strumento per l'integrazione dei dati: la sua GUI di semplice fattura, assieme alla grandissima varietà di operazioni rese disponibili e all'ampia documentazione presente sul Web, hanno sicuramente diminuito la complessità del lavoro di implementazione del sistema di SBI.

Tra gli sviluppi futuri del sistema realizzato vi è senza dubbio la parte legata al dashboarding ed al reporting dei risultati. Il mercato offre diverse soluzioni per gestire questa ultima parte dell'architettura funzionale. Quella che ci è sembrata più adatta tra le opzioni presenti è stata la via offerta da *SpagoBI*[8]: l'unica suite di business intelligence completamente open source, completa e flessibile. Offre soluzioni e temi innovativi, non solo i soliti strumenti di business intelligence, quali reporting, grafici, analisi multidimensionali e data mining, ma soluzioni originali per i nuovi domini della business (KPI, cruscotti interattivi, ecc). La sua disponibilità aperta ed il suo sviluppo di livello industriale sono accompagnati da un set completo di servizi professionali di supporto.

Un aspetto importante in vista di cambiamenti futuri è stata la scelta di sviluppare il sistema di Social Business Intelligence con un approccio modulare: la parte legata al recupero dei documenti è totalmente scissa da quella di analisi. Questo ci permette, in futuro, di poter scegliere un qualsiasi nuovo o diverso strumento di crawling tra quelli offerti dal mercato, ed integrarlo nel sistema con uno sforzo molto minore rispetto a quello che sarebbe stato necessario eseguire se avessimo sviluppato una soluzione connessa e quindi meno manutenibile.

[8] *http://www.spagoworld.org/xwiki/bin/view/SpagoBI/?language=it*

Bibliografia

Big Data. (n.d.). In *Wikipedia*. Retrieved February 21, 2013, from http://it.wikipedia.org/wiki/Big_data.

Bolasco, S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. In *Quaderni di statistica* (Liguori Ed., 7, pp. 17-53).

Contenuto generato dagli utenti. (n.d.). In *Wikipedia*. Retrieved February 20, 2013, from http://it.wikipedia.org/wiki/Contenuto_generato_dagli_utenti.

Digital Marketing. (n.d.). In *Wikipedia*. Retrieved February 23, 2013, from http://en.wikipedia.org/wiki/Digital_marketing.

Enterprise Service Bus. (n.d.). In *Wikipedia*. Retrieved February 25, 2013, from http://it.wikipedia.org/wiki/Enterprise_Service_Bus.

Grimes, S. (June 7, 2011). Text Analytics and Sentiment Analysis. In *BeyeNETWORK*. Retrieved February 20, 2013, from http://www.b-eye-network.com/view/15276.

Bibliografia

Grimes, S. (January 23, 2012). Why Sentiment Analysis Doesn't Depend on Text Analytics. In *All Analytics*. Retrieved February 24, 2013, from http://www.allanalytics.com/author.asp?section_id=1408&doc_id=238072.

Hinchcliffe, D. (February 1, 2010). Exploring Why Social Business Will Drive 21st Century Enterprises. In *Enterprise Irregulars*. Retrieved February 15, 2013, from http://www.enterpriseirregulars.com/11731/exploring-why-social-business-will-drive-21st-century-enterprises/.

Hinchcliffe, D. (January 11, 2012). What's Coming Up in Social Business, CoIT, Open APIs, and More. In *Dion Hinchcliffe*. Retrieved February 17, 2013, from http://dionhinchcliffe.com/2012/01/11/whats-coming-up-in-social-business-coit-open-apis-and-more/.

Hinchcliffe, D. (January 7, 2013). Sizing up social business for 2012. In *ZDNet*. Retrieved February 21, 2013, from http://www.zdnet.com/sizing-up-social-business-for-2012-7000009426/.

Ironìa. (n.d.). In *Dizionario Italiano*. Retrieved February 24, 2013, from http://www.dizionario-italiano.it/definizione-lemma.php?definizione=ironia&lemma=I0ABBC00.

Knowledge Base. (n.d.). In *Wikipedia*. Retrieved February 25, 2013, from http://en.wikipedia.org/wiki/Knowledgebase.

Lavenia, G. (2007). "Introduzione alle nuove dipendenze on-line" in M. Marcucci e M. Boscaro. In *Manuale di Psicologia delle Dipendenze Patologiche*. Urbino, IT: Mediateca delle Marche.

Liddy, E.D. (2003). Natural Language Processing. In *Encyclopedia of Library and Information Science*. 2nd Ed. NY.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool.

Malloy, T. (October 29, 2012). Revolutionizing Digital Marketing with Big Data. Invited speaker at *CIKM 2012*. Retrieved November 6, 2012, from http://www.cikm2012.org/doc/cikm2012_Malloy.pdf.

MySQL. (n.d.). In *Wikipedia*. Retrieved February 22, 2013, from http://it.wikipedia.org/wiki/MySQL.

Ontologia (Informatica). (n.d.). In *Wikipedia*. Retrieved February 25, 2013, from http://it.wikipedia.org/wiki/Ontologia_(informatica).

Pearson, M. (January 7, 2013). Social Media Can Play a Role in Business Process Management. In *Harvard Business Review*. Retrieved February 22, 2013, from http://blogs.hbr.org/cs/2013/01/social_media_can_play_a_role.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+harvardbusiness+%28HBR.org%29.

Phneah, E. (March 13, 2012). 'Ultraviolet data' necessary for biz insights. In *ZDNet*. Retrieved February 20, 2013, from http://www.zdnet.com/ultraviolet-data-necessary-for-biz-insights-2062304171/.

Plummer, D. C., & Gartner, Inc. (Firm). (2006). *Gartner's top predictions for IT organizations and users, 2007 and beyond*. Stamford, CT: Gartner, Inc.

Bibliografia

Robin. (December 4, 2009). PARTS-OF-SPEECH TAGGING. In *NATURAL LANGUAGE PROCESSING*. Retrieved February 25, 2013, from http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html.

Savitz, E. (August 13, 2012). 7 Ways To Get Actionable Insights From Social Data. In *Forbes*. Retrieved February 22, 2013, from http://www.forbes.com/sites/ciocentral/2012/08/13/7-ways-to-get-actionable-insights-from-social-data/.

Souza, J. (August 30, 2012). How To Gain Business Intelligence With Social Media. In *Social Media Marketing University*. Retrieved February 16, 2013, from http://socialmediamarketinguniversity.com/gain-business-intelligence-social-media/.

SyNTHEMA, (May, 2011). *iSyN Semantic Center*. Pisa, Italy (via G. Malasoma 24 56121 Pisa (PI)).

Talend, Inc. (2006). Talend Open Studio. In *Talend*. Retrieved February 20, 2013, from http://www.talend.com/products/talend-open-studio.

Talend, Inc. (2006). Data Integration. In *Talend*. Retrieved February 20, 2013, from http://www.talend.com/products/data-integration.