

RICERCA INDUSTRIALE APPLICATA ALL'ANALISI DI DATI IN AMBITO  
SOCIOECONOMICO: IL PROGETTO SINSE

Alfredo GARRO<sup>1</sup>, Rocco PICARELLI<sup>2</sup>, Andrea PUGLIESE<sup>3</sup>,  
Andrea TAGARELLI<sup>4</sup>, Andrea TUNDIS<sup>5</sup>

**SOMMARIO**

Il contributo presenta i risultati di un'esperienza di ricerca industriale applicata all'estrazione e all'analisi di dati in ambito socioeconomico. L'esperienza è stata condotta all'interno del progetto "SINSE", che ha visto coinvolti l'impresa Contesti S.r.l. e il Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica dell'Università della Calabria. L'attività di ricerca ha riguardato la definizione, applicazione ed estensione di modelli e tecniche orientati a diversi aspetti del processo di estrazione ed analisi dei dati di interesse. In particolare, sono state affrontate problematiche relative all'estrazione flessibile di dati da pagine Web e da social network, applicando tecniche di estrazione a partire da insiemi di seed e parole chiave (nel caso dell'estrazione da pagine Web) o hashtag (nel caso dell'estrazione da social network). Inoltre, sono state applicate metodologie avanzate di estrazione di conoscenza da dati socio-economici, che includono i processi classici di knowledge discovery in databases integrati con nuovi approcci all'analisi frequenziale delle parole, al sentiment analysis sui testi ed al ranking dei nodi di un social network. Infine, sono state applicate tecniche di business performance monitoring agli scenari di interesse individuati definendo servizi semanticamente configurabili per definire cockpit personalizzati e specifici per le diverse tipologie di utente.

---

<sup>1</sup> Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, via P. Bucci, Rende (CS), e-mail: alfredo.garro@unical.it.

<sup>2</sup> Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, via P. Bucci, Rende (CS), e-mail: r.picarelli@dimes.unical.it.

<sup>3</sup> Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, via P. Bucci, Rende (CS), e-mail: andrea.pugliese@unical.it (corresponding author).

<sup>4</sup> Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, via P. Bucci, Rende (CS), e-mail: andrea.tagarelli@unical.it.

<sup>5</sup> Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, via P. Bucci, Rende (CS), e-mail: andrea.tundis@dimes.unical.it.

## 1. INTRODUZIONE

Il progetto di ricerca industriale “SINSE – Sistema di supporto alle decisioni in ambito socioeconomico” ha riguardato lo sviluppo di un sistema informativo avanzato di business intelligence in grado di supportare le attività di analisi e valutazione dei sistemi socio-economici, con particolare riferimento alla costruzione di indici/indicatori sintetici di valutazione e alla realizzazione di confronti settoriali e territoriali, al fine di disporre di un maggior numero di alternative e di informazioni da valutare e processare. Il sistema utilizza modelli e tecniche avanzate di analisi multidimensionale dei dati, estrazione della conoscenza ed interrogazione flessibile di informazioni dalle fonti informative di interesse.

Il requisito principale del sistema sviluppato è stato quello di “accompagnare” sia il progettista con competenze tecniche specifiche che l’operatore più generalista nel rappresentare al meglio uno scenario, sia per gli aspetti economici che per quelli sociali, fornendo un ventaglio di possibilità e informazioni grazie alle quali il decisore possa effettuare le proprie scelte in maniera più consapevole. Altri importanti requisiti sono elencati di seguito.

- Razionalizzazione, integrazione e valorizzazione dei flussi informativi raccolti da fonti diverse.
- Chiarezza, accessibilità, tempestività ed esaustività delle informazioni gestite.
- Rappresentazione interattiva dello scenario e dei risultati.
- Gestione dei processi ed elaborazione dei risultati mediante operazioni interattive e procedure standardizzate.
- Utilizzo di dati descrittivi del territorio.
- Struttura distribuita, che ottimizza la gestione dei dati e rende possibile la fruizione delle informazioni, anche tra postazioni distanti.
- Adattabilità al mutare delle esigenze conoscitive.
- Rappresentazione delle informazioni in accordo al modello multidimensionale, per supportare analisi avanzate basate sulla definizione e valutazione di indicatori di risultato e di processo nonché analisi di scenario.
- Utilizzo di software *open source*.
- Tutela della riservatezza per ciascuno dei soggetti le cui informazioni sono presenti nel sistema.

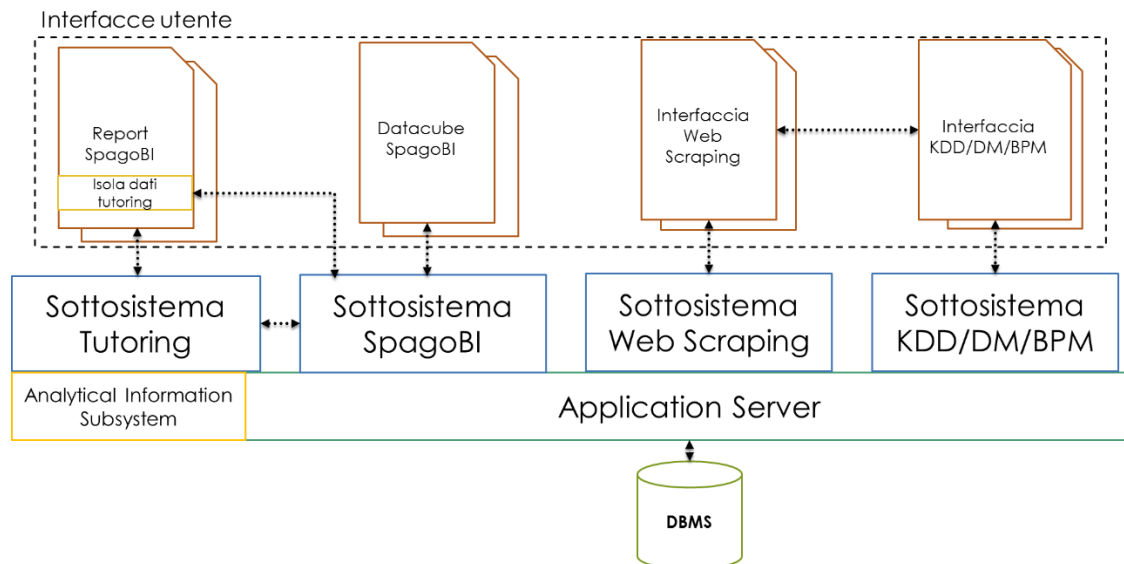
La parte strettamente relativa alle attività di ricerca avanzata ha riguardato la definizione, applicazione ed estensione di modelli e tecniche orientati a diversi aspetti del processo di estrazione ed analisi dei dati di interesse. In particolare, sono state affrontate problematiche relative (i) all’estrazione flessibile di dati da pagine Web e da social network (*Web scraping*), applicando tecniche di estrazione a partire da insiemi di *seed*, parole chiave o *hashtag*; (ii) all’analisi orientata all’estrazione di conoscenza da dati socio-economici, che include i processi classici di *knowledge discovery in databases* integrati con nuovi approcci all’analisi frequenziale delle parole, al *sentiment analysis* sui testi e al *ranking* dei nodi di un *social network*; (iii) al *business performance monitoring* applicato agli scenari di interesse individuati.

Il prosieguo del presente contributo è organizzato come segue. La Sezione 2 descrive l’architettura logica complessiva del sistema e richiama le caratteristiche di base dei sottosistemi “SpagoBI”, basato su prodotti commerciali *open source* di *business intelligence*, e “Tutoring” per il supporto al decisore nella fruizione dei risultati delle analisi e nell’individuazione di nuove modalità o dimensioni di analisi. La Sezione 3 descrive la struttura del sottosistema “Web scraping” per l’estrazione di dati dal Web e le nuove tecniche sviluppate per la progettazione e implementazione del sottosistema stesso. Nella Sezione 4 sono descritte le nuove tecniche sviluppate per il *knowledge discovery in databases*, il *data (text) mining* e il *business performance monitoring* – in questo caso, le tecniche sono implementate in un unico sottosistema denominato “KDD/DM/BPM”. Infine, la Sezione 5 delinea brevemente le conclusioni.

## 2. ARCHITETTURA LOGICA DEL SISTEMA

L'architettura logica complessiva del sistema SINSE è riportata in Figura 1.

Figura 1 – Architettura logica del sistema SINSE



L'architettura è composta dai moduli elencati di seguito.

- **Sistema di gestione di basi di dati (DBMS).** Questo modulo contiene il DBMS relazionale *MySQL* (MySQL, 2015), completamente *open source* e già utilizzato in sistemi informatici di enormi dimensioni (ad esempio *LinkedIn*, *Facebook* e *Wikipedia*). Tutti i dati estratti ed elaborati dal sistema sono memorizzati in questo modulo. I restanti moduli sono progettati e realizzati in modo da poter supportare in modo trasparente qualunque DBMS che soddisfi requisiti di base in termini di raggiungibilità via rete e robustezza.
- **Application Server.** Questo modulo contiene l'application server *Apache Tomcat* (Tomcat, 2015) che da un lato fa da interfaccia tra i moduli soprastanti e il DBMS, dall'altro esegue la maggior parte della logica applicativa del sistema, scritta in linguaggio Java.
- **Sottosistema SpagoBI.** Questo modulo fornisce strumenti e funzionalità di business intelligence coordinando a sua volta specifici motori di esecuzione.
- **Sottosistema Tutoring.** Questo modulo è deputato alla rappresentazione interattiva dello scenario di analisi e dei risultati ad esso associati; permette, inoltre, l'elaborazione interattiva dei risultati anche attraverso suggerimenti al decisore in relazione a possibili percorsi di analisi.
- **Sottosistema Web Scraping.** Questo modulo si occupa dell'estrazione di dati da pagine Web e da social network, applicando tecniche avanzate di estrazione a partire da insiemi di "seed" e parole chiave (nel caso dell'estrazione da pagine Web) o hashtag (nel caso dell'estrazione da social network).
- **Sottosistema KDD/DM/BPM.** In questo modulo sono applicate metodologie avanzate di analisi per l'estrazione di conoscenza da dati socio-economici, che includono i processi classici di "knowledge discovery in databases" integrati con nuovi approcci all'analisi frequenziale delle parole, al "sentiment analysis" sui testi ed al "ranking" dei nodi di un social network. Infine, sono state applicate tecniche di "business performance monitoring" agli scenari di interesse individuati definendo servizi semanticamente configurabili per generare "cockpit" personalizzati e specifici per le diverse tipologie di utente.

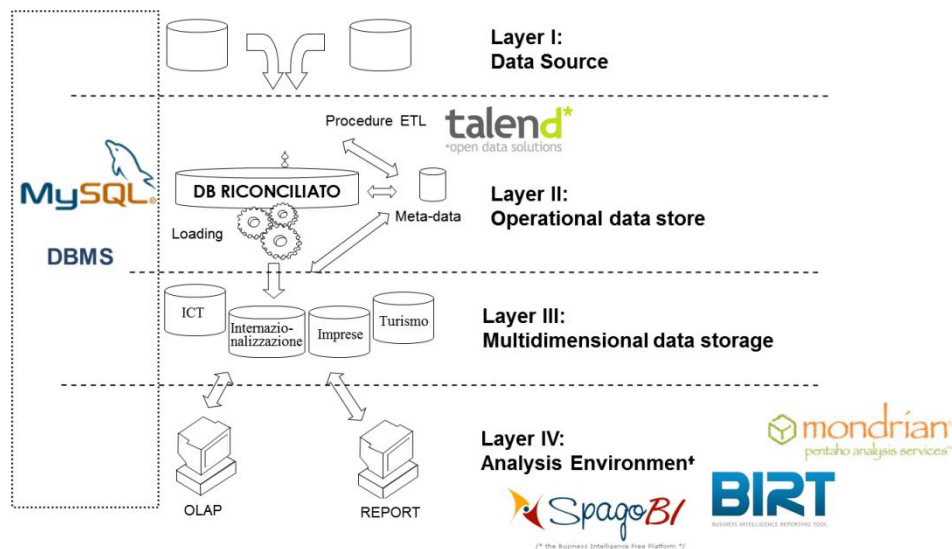
- **Interfacce utente.** Questi moduli gestiscono l'interazione con l'utente, mediante interfacce grafiche Web.

Le rimanenti sottosezioni richiamano le caratteristiche dei sottosistemi SpagoBI e Tutoring, mentre le successive sezioni descrivono in maggiore dettaglio i sottosistemi Web Scraping e KDD/DM/BPM.

## 2.1 Sottosistema SpagoBI

La Figura 2 riporta l'architettura del sottosistema SpagoBI. Tale sottosistema fornisce un'ampia gamma di strumenti e un insieme di funzionalità per amministrare la piattaforma, gli utenti e gli oggetti di business intelligence. Questi oggetti (report, *cubi* di analisi multidimensionale, ecc.) sono gestiti da SpagoBI mediante specifici motori di esecuzione (SpagoBI, 2015). Grazie alla natura "Web-portal" della piattaforma, è possibile accedere direttamente via Web agli strumenti di lavoro e pubblicare nuovi documenti analitici utilizzando apposite procedure guidate.

Figura 2 – Architettura del sottosistema SpagoBI



L'architettura del sottosistema SpagoBI è organizzata in moduli. In questo modo un progetto di analisi, in base alle proprie esigenze, può prevedere l'impiego di tutti o solo di alcuni dei moduli disponibili, permettendo altresì la realizzazione di eventuali estensioni.

Più in dettaglio, il sottosistema SpagoBI è organizzato sui quattro livelli logici descritti di seguito, che a loro volta operano sul DBMS.

- **Layer I - Data source:** rappresenta il livello più basso (in termini di astrazione) del sottosistema, in cui sono memorizzati i dati sorgenti, provenienti da fonti potenzialmente eterogenee e, quindi, tipicamente eterogenei nella forma e nel contenuto.
- **Layer II - Operational data store:** è il livello in cui i dati sorgenti di interesse, identificati e collezionati nel *Layer I*, sono sottoposti ad elaborazione impiegando specifiche procedure ETL (*Extraction, Transformation, Loading*) (Golfarelli e Rizzi, 2006) rispettivamente per l'estrazione, trasformazione e caricamento di tali dati in un database *riconciliato*, al fine di ottenere una visione globale, integrata, consistente, corretta e dettagliata dei dati di interesse indipendentemente dalle sorgenti iniziali, e con una semantica ben definita, basata su un modello comune in accordo a specifici *metadati*. Le procedure che alimentano tale livello sono realizzate in *Talend*, uno strumento software open source per la definizione ed esecuzione di procedure ETL (Talend, 2015).
- **Layer III - Multidimensional data storage:** rappresenta il livello in cui i dati a livello riconciliato sono organizzati e memorizzati in accordo al modello multidimensionale, in base a specifici *fatti*

di interesse, *misure* che ne consentono la valutazione qualitativa e quantitativa e *dimensioni* di analisi.

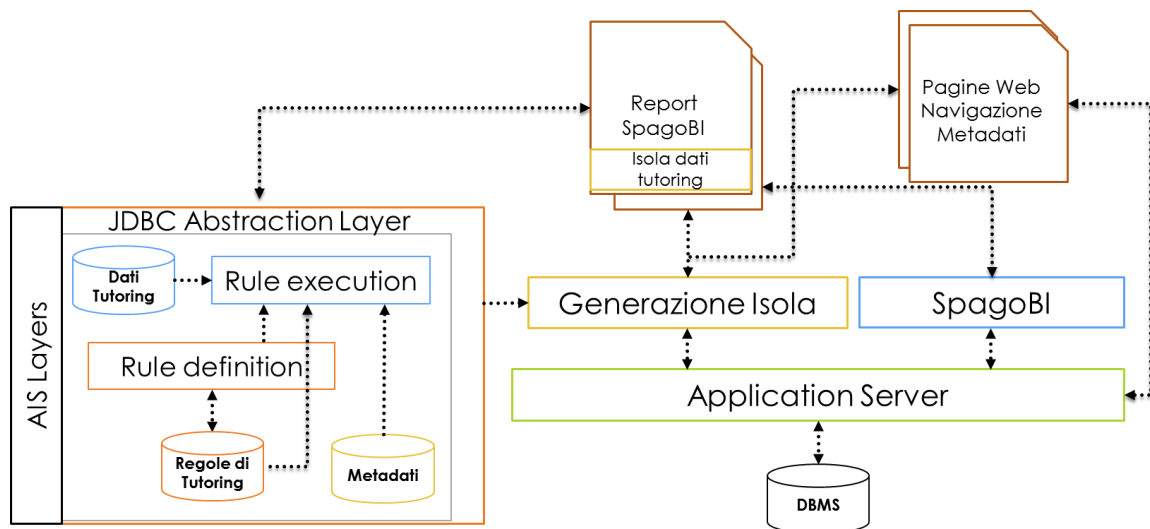
- *Layer IV - Analysis Environment*: è il livello più alto dell'architettura, che consente l'analisi multidimensionale attraverso motori OLAP, i quali a loro volta permettono all'utilizzatore un elevato grado di libertà e flessibilità nell'analizzare i dati e la loro evoluzione nel tempo, nonché la presentazione dei dati mediante strumenti di reportistica avanzata. Tale livello è basato su tecnologie innovative e strumenti avanzati di business intelligence quali ad esempio *Mondrian* e *BIRT* (Mondrian, 2015).

Nell'ambito del progetto SINSE sono stati identificati e costruiti diversi *data mart*. Ad esempio, con riferimento al data mart "Internazionalizzazione", che contiene dati di performance import-export delle imprese italiane, dopo aver individuato le sorgenti dati di interesse memorizzate nei rispettivi database e avendone analizzato i contenuti, è stato definito il livello dei dati riconciliati e le procedure ETL per la sua alimentazione. Attraverso il database riconciliato è stato possibile raccogliere in un unico schema i concetti necessari e di interesse inerenti al settore di business in considerazione, eliminando conflitti strutturali, conflitti semantici sui concetti e conflitti sui nomi, presenti tra le basi di dati sorgenti. A partire dal database riconciliato sono stati poi derivati e progettati diversi fatti di interesse. Tutti i data mart costruiti possono essere integrati a formare un unico *Data Warehouse* per il sistema SINSE.

## 2.2 Sottosistema Tutoring

L'architettura del sottosistema Tutoring è riportata in Figura 3.

Figura 3 – Architettura del sottosistema Tutoring

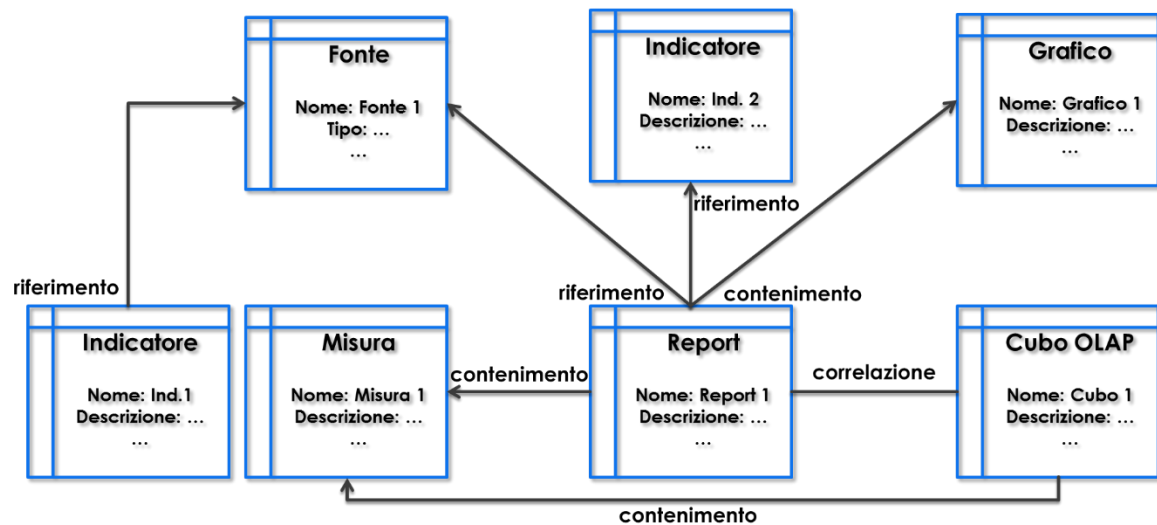


Questo sottosistema ha l'obiettivo di rappresentare in modo interattivo lo scenario di analisi e i risultati ottenuti, anche fornendo suggerimenti al decisore in relazione a possibili ulteriori percorsi di analisi. A tal fine, da un lato vengono create pagine Web per la navigazione dei metadati relativi al processo di estrazione delle informazioni dalle sorgenti esterne, e dall'altro vengono aggiunte "isole" (cioè apposite zone contenenti informazioni e suggerimenti) ai report creati da SpagoBI.

La costruzione delle isole di tutoring avviene attraverso un processo a tre fasi, descritto di seguito.

- Nella prima fase, viene effettuata una descrizione "statica" della reportistica disponibile nel sistema. A tal fine, si utilizza un semplice ma espressivo modello concettuale "a grafo", in cui i nodi sono oggetti di reportistica e gli archi esprimono relazioni (con semantica specifica) tra gli oggetti. Un esempio di modellazione è riportato in Figura 4.

Figura 4 – Esempio di modellazione concettuale dei dati di reportistica per il tutoring



- Nella seconda fase, vengono definite *regole di tutoring*, che guidano il sistema nella costruzione delle informazioni da inserire nell'isola di tutoring. In particolare, è possibile esprimere regole *incondizionate* della forma *evento* → *informazione da aggiungere* e regole *condizionate* della forma *evento [condizione]* → *informazione da aggiungere*. Ad esempio, attraverso regole incondizionate è possibile stabilire che (i) se l'utente sta visualizzando un report *R*, allora è opportuno aggiungere una descrizione del report e delle fonti dei dati rappresentati in *R*, oppure (ii) se l'utente sta visualizzando un oggetto che fa riferimento ad una fonte *F*, allora è opportuno aggiungere dei link ai metadati relativi alla fonte *F*. Attraverso regole condizionate è invece possibile, ad esempio, stabilire che se l'utente sta visualizzando un report *R* ed inoltre una delle fonti a cui fa riferimento *R* ha un periodo di aggiornamento maggiore di 3 mesi, allora è opportuno aggiungere una nota che evidenzia tale periodo.
- Nella terza fase, vengono eseguite le regole di tutoring e generate le isole.

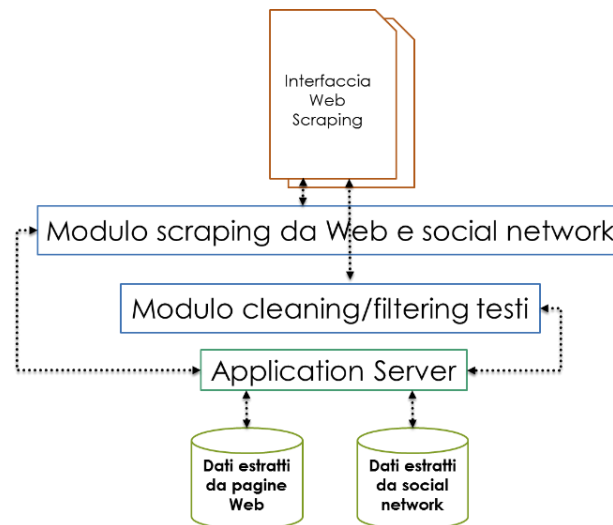
L'architettura dell'*Analytical Information Subsystem*, modulo che contiene metadati, descrizione della reportistica, regole di tutoring e moduli per la definizione e l'esecuzione delle regole, è stata progettata in modo da consentire, attraverso uno *strato JDBC* per l'astrazione dalla rappresentazione fisica dei dati, l'indipendenza logica tra le strutture necessarie al modulo e gli specifici schemi di dati a cui il sottosistema accede durante le sue operazioni. Mediante la stessa tecnologia è possibile gestire la costruzione e l'aggiunta delle isole anche a pagine Web generali, indipendentemente dai report SpagoBI.

### 3. SOTTOSISTEMA WEB SCRAPING

L'architettura del sottosistema Web Scraping è riportata in Figura 5. Il sottosistema è deputato all'estrazione di dati testuali da pagine Web e da social network, alla pulizia/filtraggio dei dati estratti e alla gestione dell'intero processo mediante interfaccia Web. Le funzionalità dei componenti del sottosistema sono descritte di seguito.

- Mediante l'interfaccia Web, è possibile impostare i parametri di input dei processi di estrazione e pulizia/filtraggio, oltre che avviare e controllare l'andamento del processo di estrazione. È possibile, inoltre, esportare i risultati dell'estrazione nei formati più comuni.

Figura 5 – Architettura del sottosistema Web Scraping



- Il modulo che si occupa dello scraping vero e proprio supporta (i) l'estrazione di pagine Web a partire da un dato insieme di *seed*, cioè di indirizzi Web da cui far partire l'esplorazione (che avviene seguendo tutti i link ipertestuali presenti nelle pagine); (ii) la definizione di una “distanza massima”, in termini di numero di link da seguire a partire dai seed; (iii) la gestione di diversi gruppi di parole chiave (la cui assenza dalle pagine estratte “elimina” le pagine stesse dal processo) che vengono combinati mediante congiunzione logica, generando risultati diversi per ognuna delle combinazioni; (iv) la possibilità di limitare l'estrazione alle sole pagine presenti nello stesso dominio Internet del seed; (v) l'estrazione da seed che richiamano motori di ricerca interni specifici per il sito su cui si sta operando; (vi) l'utilizzo di *hashtag* e parole chiave specifiche per il caso dell'estrazione da social network.
- Il modulo di pulitura/filtraggio supporta invece la specifica delle parti da estrarre dalle pagine HTML (intera pagina, parti *body* e *meta*, oppure solo parte *body*) e l'esclusione di parole che compaiono, nei testi estratti, ad una distanza dalle parole chiave maggiore rispetto ad una soglia data.

Sono state inoltre incluse nel sottosistema nuove tecniche orientate a rendere più flessibile e precisa l'estrazione da dati “ad albero”, cioè aventi una struttura gerarchica interna (come tutte le pagine HTML e XML, tipiche dell'ambito Web e *Semantic Web*). Le tecniche sviluppate offrono agli utenti la possibilità di esprimere interrogazioni (*query* nel seguito) in cui si specificano le sottostrutture da ricercare, mediante un'estensione del linguaggio *XPath* (XPath, 2015) progettata *ad hoc*. L'utente può assegnare un punteggio (*score*) a diverse componenti delle sottostrutture e quindi specificare il numero massimo di sottostrutture da estrarre. In base alle preferenze così specificate dall'utente, il sistema si occupa di calcolare in modo efficiente e restituire le “migliori” sottostrutture in base agli score ottenuti.

Si consideri ad esempio la query XPath riportata in Figura 6.

Figura 6 – Esempio di query XPath approssimata con negazione

```

//book{0.1}
  [//authors[/author[contains('Silberschatz')]]]{0.4}
  [NOT[/authors[/author[contains('Galvin')]]]]{0.5}

```

La query sostanzialmente richiede di estrarre i sotto-alberi che descrivono libri (la cui “radice”, elemento principale in cui sono “innestati” gli altri, ha infatti etichetta “book”). La lista degli autori dei libri estratti,



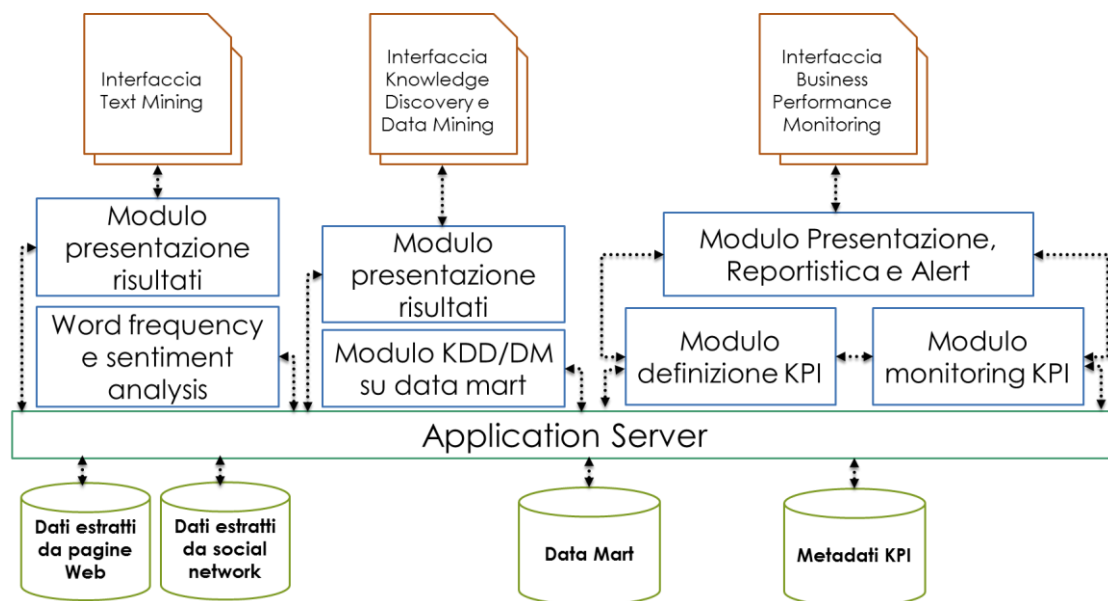
innestata mediante elementi “authors” e “author” al di sotto dell’elemento “book”, deve contenere “Silberschatz” e *non* “Galvin”. Gli score assegnati sono 0.1 per l’etichetta dell’elemento estratto (lo score ottenuto dall’elemento estratto può ad esempio variare nei casi in cui si utilizzino sinonimi dell’etichetta), 0.4 per la presenza di “Silberschatz” e 0.5 per l’assenza di “Galvin”. Per semplicità, dalla figura sono omessi gli score assegnati alle componenti “interne”, quali gli elementi “authors” e “author” – la tecnica supporta comunque l’assegnazione di score a tutte le componenti.

La tecnica ha permesso di risolvere adeguatamente le complessità derivanti innanzitutto dalla presenza dell’operazione di negazione logica (operatore NOT), che influenza sia il processo di approssimazione che la definizione della semantica delle condizioni negate. L’algoritmo di estrazione dei  $k$  elementi (con  $k$  parametro impostabile dall’utente) con score più elevato ha mostrato elevatissime prestazioni, completando il processo in tempi generalmente dell’ordine delle decine di secondi anche su documenti di dimensione dell’ordine delle centinaia di MB. Per maggiori dettagli si veda (Fazzinga *et al.*, 2014).

#### 4. SOTTOSISTEMA KDD/DM/BPM

L’architettura del sottosistema KDD/DM/BPM è riportata in Figura 7. Le successive sottosezioni descrivono in dettaglio le funzioni svolte dal sottosistema KDD/DM/BPM e le tecniche sviluppate e utilizzate all’interno del sottosistema stesso.

Figura 7 – Architettura del sottosistema KDD/DM/BPM



##### 4.1 Knowledge discovery e data mining

Il sottosistema KDD/DM/BPM include tecniche classiche di knowledge discovery e data mining (in particolare di *clustering* o analisi dei gruppi) che individuano relazioni non immediatamente visibili all’utente e organizzano opportunamente le informazioni generando nuova conoscenza. Questo tipo di analisi consente di raggruppare una vasta e variegata popolazione di unità in cluster distinti in cui tutti i componenti condividono caratteristiche simili (Han *et al.*, 2011), (Tan *et al.*, 2015). Sono pertanto indicate, ad esempio, per dividere in segmenti il mercato dei consumatori, o identificare categorie per l’organizzazione dei prodotti.

Il sottosistema si interfaccia ai data mart presenti nel sistema. Le analisi svolte finora hanno riguardato in particolare il data mart “Internazionalizzazione”. I dati disponibili sono collegati al luogo geografico (area, regione, provincia) e alle categorie merceologiche ATECO. L’analisi condotta finora è stata finalizzata al



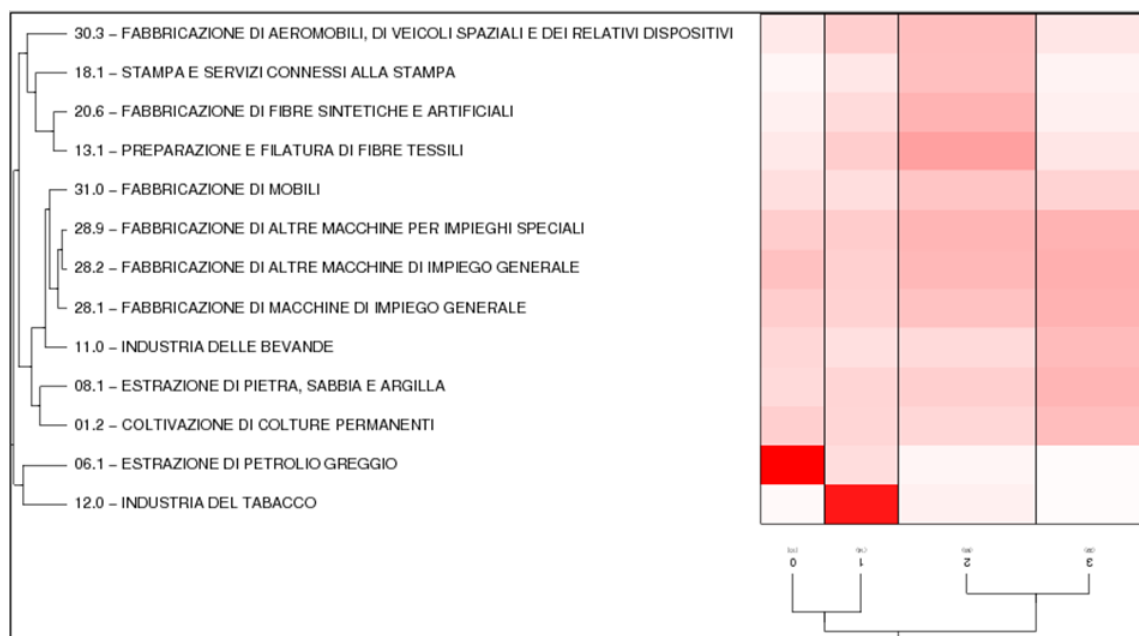
raggruppamento delle province italiane in base ai valori di import-export registrati rispetto alle varie categorie merceologiche.

Il processo di clustering è suddiviso in due macro-fasi. Nella prima fase, vengono analizzati i dati del data mart e generati gli input necessari al motore di clustering. Nella seconda fase, viene invocato il motore di clustering, basato su (CLUTO, 2015), che genera risultati numerici e grafici, a loro volta analizzabili e valutabili dagli esperti di dominio.

Il sottosistema gestisce una molteplicità di parametri di input. Nello studio condotto sono stati impostati come variabili sia le varie classificazioni delle categorie ATECO (gruppi, divisioni, sezioni) che i dati di import e export separatamente. Per l'elaborazione sono stati opportunamente variati i parametri del motore di clustering, in particolare il numero totale di cluster da generare.

La Figura 8 mostra un esempio di output grafico generato dal sistema, dove le categorie merceologiche sono relazionate ai cluster delle province; tonalità più scure corrispondono ad una maggiore caratterizzazione di una o più categorie merceologiche nel rappresentare un cluster di province.

*Figura 8 – Esempio di clustering delle province italiane secondo una rappresentazione basata su categorie merceologiche*



#### 4.2 Word frequency analysis e sentiment analysis

Nel sottosistema KDD/DM/BPM vengono effettuate analisi anche sulle frequenze delle parole e sulle opinioni o stati emotivi degli autori dei contenuti estratti tramite il sottosistema Web Scraping.

Per quanto riguarda le analisi frequenziali, il sistema costruisce, per ognuno dei seed utilizzati nella fase di scraping, le strutture elencate di seguito.

- Una tabella contenente il numero di occorrenze di ognuna delle parole chiave all'interno delle pagine Web estratte a partire dal seed.
- Un grafico a istogrammi contenente le  $k$  parole più frequenti all'interno delle pagine (con  $k$  parametro impostabile dall'utente) in ordine decrescente di frequenza.
- Una *word cloud* contenente  $nwc$  parole ognuna con almeno  $omin$  occorrenze (con  $nwc$  e  $omin$  parametri impostabili dall'utente). Il paradigma visuale delle word cloud, in cui la dimensione dei caratteri è proporzionale alla popolarità dei termini rappresentati, risulta di grande utilità e

immediatezza al fine di caratterizzare la popolarità dei temi discussi. Un esempio di word cloud è riportato in Figura 9.

*Figura 9 – Esempio di word cloud*



Il sistema calcola le informazioni di cui sopra anche rispetto all'*intero* insieme dei seed utilizzati durante lo scraping. Per tutte le analisi, il sistema supporta inoltre (i) l'applicazione di tecniche di *stemming* (estrazione delle radici dei termini) e (ii) la gestione di *stopword* (parole che vengono sempre eliminate dall'analisi, quali articoli, preposizioni ecc.)

In relazione all'analisi del *sentiment*, applicata attualmente ai dati estratti dal social network *Twitter* (Twitter, 2015), il sistema procede secondo le due fasi descritte di seguito.

- In una prima fase, il sistema associa ad ognuno degli elementi estratti un “grado di positività” (come numero razionale compreso tra 0 e 1) dell’opinione o stato emotivo dell’autore, utilizzando una combinazione di tecniche di analisi delle opinioni (OpinionFinder, 2015), (R-Sentiment, 2015), (VADER, 2015).
- In una seconda fase, il sistema supporta l’esecuzione di query *OLAP* (Han *et al.*, 2011) che utilizzano come dimensioni di analisi le posizioni spazio-temporali degli autori o degli elementi estratti e come misura il grado di positività associato. E’ così possibile, ad esempio, ottenere l’andamento medio del grado di positività delle opinioni rispetto ad un certo argomento lungo un dato intervallo di tempo e in relazione ad una specifica area geografica, oppure il grado di positività medio al variare dell’area geografica (con diversi livelli di granularità), ecc.

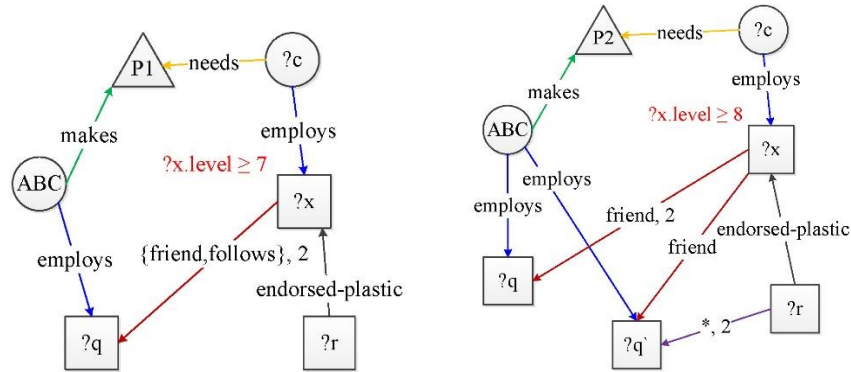
### 4.3 Ranking dei nodi nei social network

Nel sottosistema di cui alla sezione precedente sono state inoltre incluse nuove tecniche per il ranking dei nodi nei social network. Assegnare uno score ai nodi attraverso il ranking, in particolare ai nodi che rappresentano persone partecipanti ai network, è in molti ambiti essenziale al fine di dare maggiore “evidenza”, nelle analisi, ai partecipanti più “vicini” dal punto di vista dell’interesse verso la specifica tematica.

Le tecniche sviluppate comprendono (i) un linguaggio grafico per esprimere *pattern query*, cioè sotto-strutture da ricercare all'interno del social network, completate da un meccanismo per assegnare gli score alle parti del social network che rispecchiano le strutture cercate; (ii) meccanismi di indicizzazione dei dati e calcolo degli score che, dato un insieme di pattern query, permettono di assegnare gli score ai nodi del social network in tempi estremamente brevi.

Si considerino ad esempio le pattern query riportate in Figura 10, basate sul social network *LinkedIn* (LinkedIn, 2015).

Figura 10 – Esempi di pattern query su LinkedIn



Attraverso le query, l'impresa "ABC", che produce due tipi di prodotti in materiale plastico ("P1" e "P2") esprime il fatto che i partecipanti al social network (rappresentati dalla variabile  $?x$ ) sono pesati in maniera diversa a seconda del potenziale interesse verso i prodotti dell'impresa stessa. In particolare, attraverso la prima query (a sinistra nella figura) vengono espressi i seguenti vincoli.

- La persona  $?x$  è individuata come potenziale interessata se (i) lavora presso un'impresa  $?c$  che necessita del prodotto "P1" ed inoltre (ii) è connessa con almeno un dipendente  $?q$  dell'impresa "ABC" attraverso link di tipo "friend" o "follows" (fino a 2 link).
- La persona  $?x$  è stata indicata da almeno un'altra persona  $?r$  come esperta di materiali plastici.
- La persona  $?x$  lavora nell'impresa  $?c$  ad un "livello" ( $?x.level$ ) almeno pari a 7.

La seconda query esprime invece i seguenti vincoli.

- La persona  $?x$  è individuata come potenziale interessata se (i) lavora presso un'impresa  $?c$  che necessita del prodotto "P2" ed inoltre (ii) è connessa con almeno due dipendenti  $?q$  e  $?q'$  dell'impresa "ABC" attraverso link di tipo "friend" (fino a 2 link per  $?q$  ed esattamente un link per  $?q'$ ).
- La persona  $?x$  è stata indicata da almeno un'altra persona  $?r$  come esperto di materiali plastici.
- La persona  $?r$  è anche connessa con  $?q'$ , trovandosi al più a 2 link di distanza da essa (indipendentemente dal tipo dei link).
- La persona  $?x$  lavora nell'impresa  $?c$  ad un "livello" almeno pari a 8.

Il linguaggio permette inoltre di specificare le modalità di calcolo degli score. Ad esempio, lo score della persona  $?x$ , quando essa soddisfa i requisiti presenti nella prima query, potrebbe essere definito come  $(?x.level - 7) + ?q.level / 5$ , mentre nel caso in cui soddisfi i requisiti della seconda (che è maggiormente restrittiva), il suo score potrebbe essere  $(?x.level - 8) + (?q.level + ?q'.level) / 5$ . Lo score complessivo della persona potrebbe essere definito come somma degli score ottenuti per ognuna delle query.

Le tecniche sviluppate hanno permesso di ottenere performance estremamente elevate. Nelle sperimentazioni effettuate è stato possibile estrarre i  $k$  nodi (con  $k$  parametro impostabile dall'utente) aventi score più elevato, in un social network contenente più di 6 milioni di nodi e più di 15 milioni di link, in tempi generalmente inferiori ai 10 secondi. Un articolo che descrive le tecniche è attualmente sottoposto a valutazione per la pubblicazione in una rivista internazionale (Park *et al.*, 2015).

#### 4.4 Business performance monitoring

In letteratura non vi è consenso generale sulla definizione di sistema di Business Performance Monitoring (BPM). Ogni definizione offre una prospettiva diversa su tale concetto e non ci sono definizioni che

concordano sulle specifiche caratteristiche. Tuttavia si può affermare che per BPM si intende il costante e proattivo monitoraggio dei processi di business al fine identificare i problemi di prestazioni prima che influiscano negativamente sul business aziendale. Ad esempio, in accordo a (Havey, 2015), il BPM è il processo essenziale che aiuta a comprendere e valutare la “bontà” del business in termini di raggiungimento degli obiettivi aziendali. Esso fornisce gli strumenti per ottenere una visione degli indici di rendimento del business allo scopo di permettere decisioni aziendali tempestive e mirate. D’altra parte, in accordo a (BPM, 2015) il BPM si riferisce all’aggregazione, all’analisi e alla presentazione delle informazioni in tempo reale sulle attività interne alle organizzazioni coinvolgendo anche partner e clienti. In generale, l’obiettivo del BPM è di fornire informazioni real-time sullo stato e sui risultati delle operazioni aziendali, processi e transazioni. Questo permette ad un’impresa di disporre di un miglior supporto alle decisioni da prendere, di identificare immediatamente le aree che presentano dei problemi e scoprire nuove opportunità di business. E’ quindi possibile: (i) individuare i colli di bottiglia dei processi in tempo reale e adottare efficaci misure correttive; (ii) ottimizzare la fornitura di servizi al fine di garantire che i *service level agreement* siano soddisfatti; (iii) identificare i rischi emergenti e intervenire tempestivamente; (iv) approfittare delle “windows of opportunity” prima che esse terminino.

Al fine di quantificare il soddisfacimento degli obiettivi rispetto al loro grado di raggiungimento, i *Key Performance Indicators* (KPI), possono risultare utili strumenti di valutazione. Essi nascono per stimare lo stato attuale del business e supportare la creazione di un piano di azione per raggiungere i risultati desiderati. Per definire dei KPI, è fondamentale individuare che cosa deve essere misurato, come dev’essere misurato e la provenienza dei dati. Infatti, la possibilità di accedere ai dati dalle diverse sorgenti disponibili è fondamentale per il successo del monitoraggio delle prestazioni aziendali. Al fine di supportare tali meccanismi nell’ambito del progetto SINSE, sono stati raccolti i requisiti architetturali di seguito riportati e le relative funzionalità richieste al sottosistema BPM.

- Strumenti per la definizione di specifici indicatori di performance volti al monitoraggio e alla valutazione di parametri legati al business aziendale.
- Strumenti per la modifica e l’adattamento di KPI, considerando le esigenze e capacità evolutive aziendali legate a possibili fattori esterni.
- Capacità di definire su specifici KPI opportune viste e funzioni di monitoraggio.
- Supporto al processo di definizione di soglie di riferimento utili all’analisi.
- Funzionalità di notifica/alert, definite su specifici parametri di interesse e relativi KPI.
- Funzionalità di filtraggio dei dati in base agli indicatori di interesse selezionati.
- Supporto (semi-)automatico alla generazione della reportistica dei dati aziendali in base agli indicatori desiderati.
- Strumenti di visualizzazione dei dati in base agli indicatori definiti.
- Funzionalità in grado di fornire prospettive multiple dei dati a differenti livelli di dettaglio, eventualmente estese da supporto grafico.
- Funzionalità per la manipolazione delle viste in accordo agli indicatori e ai dati considerati.

Partendo all’architettura in Figura 7, ed in accordo ai requisiti sopra indicati, è descritto, di seguito, il sottosistema BPM. Esso è organizzato modularmente, quindi flessibile e facilmente estendibile, ed in accordo all’architettura generale, suddiviso in quattro livelli: (i) il livello utente o interfaccia; (ii) il livello *core*, che specializza l’architettura; (iii) il livello Application Server; (iv) il livello sorgenti dei dati. I componenti fondamentali di questa architettura di BPM sono elencati di seguito.

- Interfaccia di business performance monitoring, il cui ruolo è quello di mostrare i risultati di una specifica attività di BPM fornendo specifiche viste, in base ai parametri di configurazione accessibili lato utente.
- Modulo di presentazione, reportistica e alert, che si occupa di elaborare le richieste provenienti dall’interfaccia di BPM a cui è direttamente collegato.
- Modulo per la definizione di KPI, che permette la definizione di indici di performance personalizzati.

- Modulo di monitoring KPI, che si occupa di valutare dinamicamente gli indici di performance definiti.

## 5. CONCLUSIONI

Il contributo ha presentato i risultati di un'esperienza di ricerca industriale applicata all'estrazione e all'analisi di dati in ambito socioeconomico, condotta in collaborazione tra l'impresa Contesti S.r.l. e il Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica (DIMES) dell'Università della Calabria. L'attività di ricerca più avanzata ha riguardato problematiche relative (i) all'estrazione flessibile di dati da pagine Web e da social network e (ii) a nuove metodologie di analisi per l'estrazione di conoscenza che includono processi classici integrati con nuovi approcci all'analisi frequenziale delle parole, al "sentiment analysis" sui testi, al "ranking" dei nodi di un social network e al "business performance monitoring" sugli scenari di interesse individuati. I risultati ottenuti sono molto incoraggianti e il sistema sviluppato può da un lato soddisfare nuove esigenze di analisi dei dati e dall'altro supportare una molteplicità di ambiti e scenari applicativi.

## 6. BIBLIOGRAFIA

- BPM (2015), *Theoris - Business Performance Monitoring*. <http://www.theoris.com/business-performance-monitoring>.
- CLUTO (2015), *CLUTO - Software for clustering high-dimensional datasets*. <http://glaros.dtc.umn.edu/gkhome/views/cluto>.
- Fazzinga B., Flesca S., Pugliese A. (2014), Top-k Approximate Answers to XPath Queries with Negation, *IEEE Transactions on Knowledge and Data Engineering*, 26, 10: 2561-2673.
- Golfarelli M., Rizzi S. (2006), *Data Warehouse. Teoria e pratica della progettazione*. McGraw-Hill.
- Han J., Kamber M., Pei J. (2011), *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Havey M., (2005), *Essential Business Processing Modeling*. O'Reilly Media.
- LinkedIn (2015), *LinkedIn professional community*. <http://linkedin.com>.
- Mondrian (2015), *Mondrian OLAP engine*. <http://sourceforge.net/projects/mondrian>.
- MySQL (2015), *MySQL – The world's most popular open source database*. <http://www.mysql.com>.
- OpinionFinder (2015), *OpinionFinder system*. <http://mpqa.cs.pitt.edu/opinionfinder>.
- Park N., Parisi F., Pugliese A., Subrahmanian V.S. (2015), Top-k User-Defined Vertex Scoring Queries in Edge-Labeled Graph Databases, sottoposto a valutazione per la pubblicazione in *IEEE Transactions on Knowledge and Data Engineering*.
- R-Sentiment (2015), *Sentiment analysis with the "sentiment" R package*. <https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment>.
- SpagoBI (2015), *SpagoWorld*. <http://www.spagoworld.org>.
- Talend (2015), *Talend Product Suite*. <https://www.talend.com>.
- Tan P.-N., Steinbach M., Kumar V. (2015), *Introduction to Data Mining*. Addison-Wesley.
- Tomcat (2015), *Apache Tomcat project*. <http://tomcat.apache.org>.
- Twitter (2015), *Twitter*. <http://twitter.com>.
- VADER (2015), *Valence Aware Dictionary and sEntiment Reasoner*. <https://pypi.python.org/pypi/vaderSentiment>.
- XPath (2015), *XML Path Language*. <http://www.w3.org/TR/xpath>.

## ABSTRACT

This paper presents the results of an industrial research experience in the context of socio-economic data extraction and analysis. The experience has been conducted with the support of the “SINSE” grant, which involved the “Contesti” company and the Department of Informatics, Modeling, Electronics and Systems Engineering (DIMES) of the University of Calabria, Italy. The research regarded the definition, application, and extension of models and techniques targeted at different aspects of the data extraction and analysis processes. In particular, we tackled flexible data extraction from Web pages and from social networks by applying seed- and keyword-based (for Web data) and hashtag-based (for social networks) extraction techniques. Moreover, we applied advanced methodologies for knowledge discovery from socio-economic data, which include classical approaches together with new ones for word frequency and sentiment analysis, and for ranking the nodes of social networks. Finally, we applied business performance monitoring techniques to specific scenarios, by defining semantically reconfigurable services to generate personalized cockpits for the various kinds of users.