



Sii-Mobility

Supporto di Interoperabilità Integrato per i Servizi al Cittadino e alla Pubblica Amministrazione

Trasporti e Mobilità Terrestre, SCN_00112

Deliverable ID: DE2-7a

**Titolo: Stato dell'arte degli algoritmi di Data Analytic
nel dominio esterno di Sii-Mobility**

Data corrente	
Versione (solo il responsabile puo' cambiare versione)	0.5
Stato (draft, final)	Finale
Livello di accesso (solo consorzio, pubblico)	
WP	WP2
Natura (report, report e software, report e HW..)	Report
Data di consegna attesa	
Data di consegna effettiva	
Referente primario, coordinatore del documento	Paolo Nesi
Contributor	Paolo Nesi, Daniele Cenni, Angelo Difino, Laura Cocone, Pierfrancesco Bellini, Gianni Pantaleo
Coordinatore responsabile del progetto	Paolo Nesi, UNIFI, paolo.nesi@unifi.it

Sommario

1	Executive Summary	3
2	Indicizzazione ed accesso ai dati	4
2.1	Query testuali e faceted	4
2.2	Query e reasoning spaziali e temporali	6
2.3	Indicizzazione e versioning di RDF Store.....	10
3	Calcolo di matrici origine destinazione e traiettorie	12
3.1	Flussi e traiettorie	12
3.2	Acquisizione di dati di flusso da Wi-Fi.....	12
3.3	Acquisizione di dati di flusso da Spire e sensori.....	13
3.4	Calcolo di matrici OD	18
3.5	Clusterizzazione di traiettorie.....	18
4	Calcolo di Suggerimenti e stimoli.....	19
4.1	Produzione di raccomandazioni	19
4.2	Calcolo del comportamento utente.....	21
4.3	Produzione di regole di assistenza e ingaggio.....	22
5	Calcolo di Sentiment Analysis su dati testuali.....	23
5.1	Distribuzioni e NLP.....	23
5.2	Calcolo della sentiment analysis	24
6	Bibliografia	26
7	Acronimi	35

1 Executive Summary

Questo deliverable descrive lo stato dell'arte e sulla ricerca nel campo dei data analitc. Questo viene sviluppato nell'ottica del contesto di Sii-Mobility ma escludendo tematiche che sono trattate in altri deliverable come deliverable e algoritmi (come da specifica):

- DE2.2a – stato dell'arte e sulla ricerca effettuata nel campo del vehicle routing, M6
 - A06: Algoritmi di instradamento (veicoli e persone), Attività: 2.2.5
- DE2.4a – stato dell'arte nei problemi di percorso ottimo su grafi multimodali, multi-obiettivo, M6
 - A05: Algoritmi di ottimizzazione (percorsi con più fermate, cambi, etc.), Attività: 2.2.4
 - A07: Algoritmi per la produzione di percorsi per flotte merci, Attività: 2.2.5
- DE2.9a - stato dell'arte e sulla ricerca effettuata nel campo del supporto alle decisioni, M6
 - Identificazione di correlazioni inattese
 - Identificazione di condizioni critiche, dashboard, soluzioni di firing
- DE4.1 – analisi dello stato dell'arte e definizione dei sistemi di acquisizione dati e ingestion process, M7
 - A01: Algoritmi e strumenti di Riconciliazione, Attività: 2.2.1
 - A02: Algoritmi e strumenti di Arricchimento, Attività: 2.2.1
 - A08: Algoritmi di aggregazione / riconciliazione, Attività: 4.1.7

In questo deliverable lo stato dell'arte relativo ad algoritmi identificati in fase di specifica:

- A03: Algoritmi di ottimizzazione per la produzione suggerimenti per il parcheggio, Attività: 2.2.1
- A04a: Algoritmi e strumenti per raccolta e computo flussi di persone, Attività: 2.2.1
- A04b: Algoritmi e strumenti per raccolta e computo flussi di mezzi, Attività: 2.2.1

In questo scenario le aree di Data analytic prese in considerazione sono quelle relative a soluzioni e applicazioni per lo sviluppo di:

- Indicizzazione ed accesso ai dati
 - Query testuali e faceted
 - Query e reasoning spaziali e temporali
 - Indicizzazione e versioning di RDF store
- Calcolo di matrici origine destinazione e traiettorie
 - Acquisizione di dati di flusso da Wi-Fi
 - Acquisizione di dati da mobile, utente come sensore
 - Acquisizione di dati di flusso da Spire e sensori
 - Calcolo di matrici OD
 - Clusterizzazione di traiettorie
- Calcolo di Suggerimenti e stimoli, per guida connessa ed assistenza
 - Suggerimenti per similarità
 - Produzione di raccomandazioni
 - Calcolo del comportamento e del profilo utente
 - Produzione di regole di assistenza e ingaggio
 - Produzione di modelli e regole di rewarding
- Calcolo di Sentiment Analysis su dati testuali
 - Distribuzioni e NLP
 - Calcolo della sentiment analysis

L'obiettivo di questo documento non è quello di presentare uno stato dell'arte completo ed esaustivo. Ma quello di presentare lo stato dell'arte attuale, mettendo in evidenza gli aspetti recenti della ricerca e dello sviluppo nei vari settori in modo da servire come punto di partenza per lo sviluppo delle ricerche specifiche.

2 Indicizzazione ed accesso ai dati

Per sistema di indicizzazione si intende un metodo per creare un indice specifico che permetta il recupero con prestazioni accettabili delle informazioni, anche se queste vengono richieste ordinate o integrate (composte) con altri campi. Il recupero (retrieval) delle informazioni può pertanto essere più o meno performante in base alla presenza di indici specifici, e alla loro capacità di produrre risultati in tempi brevi su grandi moli di dati. I sistemi di indicizzazione tipicamente sfruttano le informazioni che sono presenti nei database per creare degli indici aggiuntivi che possono fornire risposte a query complesse in tempi più brevi di quanto potrebbero fare i database stessi.

Lo stato dell'arte dei sistemi di indicizzazione dati per quanto riguarda il contesto di Sii-Mobility deve riferirsi all'indicizzazione delle seguenti informazioni riguardo a:

- **punti di interesse geolocalizzati**, pertanto indici che tengono in grande considerazione le coordinate GPS, le località geografiche, etc.
- **elementi stradali con forme diverse e classificazione semantica, uno SKOS per esempio**. Per esempio segmenti di strada in cui sono collocati elementi stradali e non stradali in grande quantità, punti di interesse ma con classificazione semantica, etc.
- **elementi testuali per permettere la ricerca full text** o per keyword su una vasta gamma di elementi diversi. Dai nomi delle strade, a quelli dei ristoranti, ai nomi delle stazioni delle metro, alle descrizioni testuali dei punti di interesse.
- **Sensori ed elementi di misura**. In questo caso i sensori sono tipicamente classificati e geolocalizzati, ma anche collocati temporalmente. Questi dati possono essere ricercati per identificativo, per posizione, per tipo di misura, per classificazione semantica, per istante, o semplicemente per avere il loro ultimo valore, per farne una valutazione statistica (valor massimo, medio, mediano, varianza, etc.). Le misure sui sensori possono essere riferiti alla temperatura, pressione, inquinanti, livelli dell'acqua, valori sismici, numero puri, etc.

Le informazioni in un sistema di controllo della Smart City risiedono tipicamente in database di grosse dimensioni, spesso noSQL database. Questi non vengono interrogati tramite linguaggi SQL standard ma tramite modelli specifici. Fanno parte di questi noSQL database: Hbase, Mongo, Cassandra, etc. e anche i database a grafo, come per esempio vari RDF store: GraphDB, Virtuoso, StarDog, Jena, Oracle graph db, etc. Questi ultimi vengono interrogati tramite linguaggi specifici per la formulazione di query semantiche come SPARQL. Anche queste soluzioni più avanzate hanno bisogno di particolari accorgimenti e algoritmi per poter permettere di erogare e fornire ricerche efficienti, e per questo necessitano di indici specifici ed efficienti.

2.1 Query testuali e faceted

La rapida crescita delle risorse digitali sul Web ha aperto nuove sfide nello sviluppo di soluzioni di information retrieval efficienti e robuste. Una vasta gamma di contenuti, con diversi formati e tipi di metadati, costituisce un insieme eterogeneo di risorse difficili da affrontare. Un esempio rilevante è fornito da risorse multimediali trasversali, che spesso includono un ricco set di metadati e tecnica mista, affrontando i problemi collegati alla costruzione di indici. In genere, vi è la necessità di: strumenti per l'estrazione dei metadati; schemi, regole e strumenti di mappatura dei metadati; e per i metadati multilingue in certi casi anche la traduzione dei contenuti e la loro certificazione. Le soluzioni per il recupero delle informazioni (IR, information retrieval) sono tenute a fornire risposte

coerenti rispetto a errori di battitura o inflessioni, e devono essere abbastanza efficienti, mentre producono l'ordinamento di enormi elenchi di risultati. Per ottenere una certa raffinatezza nella risposta e nelle ricerche si utilizzano soluzioni faceted (di sfaccettatura), soprattutto nei grandi archivi multimediali. Gli algoritmi di analisi dei documenti devono essere abbastanza veloci per produrre indici di grandi volumi di documenti e tenendo in considerazione tutto il testo ed eventuali arricchimenti. Vi sono pertanto diversi tipi di descrittori di contenuto. Indici e archivi devono essere completamente accessibili, senza tempi significativi di inattività (sempre disponibili, in high availability, HA), in caso di guasti o di aggiornamenti importanti della struttura dell'indice, in servizi di produzione (ad esempio, la ridefinizione dello schema dell'indice, indice di corruzione).

I documenti multilingue richiedono la traduzione delle query o dei metadati per il recupero delle informazioni. Il primo approccio riduce l'utilizzo della memoria e ogni documento viene memorizzato solo una volta nell'indice [McCarley, 1999], mentre il secondo produce indici più grandi ed evita problemi di traduzione della query. Infatti, il processo di traduzione automatica di query potrebbe creare problemi di ambiguità parola, polisemia, flessione e omonimia [Abusalah et al., 2005], in particolare nel caso di domande brevi [Hull, 1996]. Per questo possono essere adottate tecniche di disambiguazione ad esempio utilizzando co-occorrenze dei termini accoppiati [Yuan and Yu, 2007], o adottando approcci statistici. Per aumentare l'efficienza del sistema di recupero delle informazioni possono essere adottate soluzioni come: l'espansione della query [Ballesteros and Croft, 1997], la tecnica pseudo relevance feedback [Attar and Fraenkel, 1977], [Ballesteros and Croft, 1998], thesauri come WordNet [Fellbaum-WordNet -1998] o soluzioni di traduzione strutturata [Sperer and Oard, 2000].

Altri approcci possibili per trattare documenti multilingua si riferiscono a Self Organizing Maps (SOM) [Lee and Yang, 2000] o fanno uso di soluzioni di clustering prima del processo di traduzione [Chen et al., 2003]. Un approccio alternativo alla traduzione delle query prevede l'utilizzo di Corpora paragonabili [Picchi and Peters, 1998]. Essi consistono in una raccolta di testi in linguaggio naturale, in cui ogni documento è tradotto in varie lingue; questi corpora paralleli vengono allineati e sono annotati per abbinare ogni frase nel documento sorgente con le rispettive traduzioni. Così, i documenti sono comparabili quando usano lo stesso vocabolario e trattano lo stesso argomento [Manning et al., 2008].

Algoritmi di ranking consistono nel ordinare la lista dei risultati di uscita di una ricerca dal più probabile e coerente (corretto) alla voce meno probabile [Belkin and Croft, 1987]. In generale, la classifica si basa sul posizionamento e la frequenza; i documenti che presentano maggiori ricorrenze del termine ricercato sono posizionati più in alto. Un esempio è l'algoritmo PageRank [Brin and Page, 1998], che determina la pertinenza di una pagina con un'analisi dei link. L'algoritmo di Relevance Feedback si basa sul presupposto che una nuova query sia una versione modificata di quella vecchia, derivato aumentando il peso dei termini in elementi pertinenti, e diminuendo il peso dei termini in elementi non rilevanti. Al fine di superare i limiti dei motori di ricerca basati su parole chiave in database tradizionali, vengono anche utilizzati modelli Fuzzy. In tal caso, i sinonimi come gli errori di battitura sono valutati in termini di somiglianza rispetto alle chiavi effettivamente indicizzate, in modo da fornire risultati più completi.

Esempi rilevanti di applicazione di tecniche Fuzzy includono la ricerca semantica [Lai et al., 2011], ontologie [Akhlaghian et al., 2010], Cloud Computing [Li et al., 2010], immagine analisi del testo [Berkovich and Inayatullah, 2004], espansione della query [Takagi and Tajima, 2001], il clustering [Matsumoto and Hung, 2010], e piattaforme di ricerca popolari come Apache Lucene. Modelli multidimensionali con tassonomie dinamiche (ad esempio, ricerche faceted [Tunkelang, 2009], [Sacco, 2007], [Sacco and Tzitzikas, 2009]) sono molto popolari, soprattutto nei siti di e-commerce, in cui l'utente ha bisogno di un modo per esplorare con facilità i contenuti, e ogni aspetto possono

essere rappresentati con una tassonomia. Il tipo di documento, gli algoritmi di rilevamento e di analisi per l'estrazione dei metadati sono un fattore importante per l'indicizzazione di risorse di testo (ad esempio, semi-strutturato o documenti non strutturati), con l'obiettivo di Natural Language Processing (NLP). Per esempio, alcuni approcci includono metodi di apprendimento automatico [Han et al., 2003], tabella di estrazione dei metadati (ad esempio, dai PDF [Liu et al., 2006]), contesti e thesauri in combinazione con l'analisi dei documenti [Shepherd et al., 2004] DOM basato estrazione di contenuti [Gupta et al., 2003]. In genere, informazioni estratte dai documenti non strutturati possono essere organizzate come entità (cioè, sintagmi nominali) e le relazioni tra di loro, aggettivi, tabelle ed elenchi [Sarawagi, 2008].

La valutazione dell'efficacia, svolge un ruolo determinante nella valutazione di un sistema di recupero delle informazioni, e questo vale anche per quelle geolocalizzate o collocate temporalmente. Quindi, è importante effettuare una dettagliata analisi IR, soprattutto negli archivi multilingue enormi. Il ranking del sistema di recupero delle informazioni comporta quasi sempre l'utilizzo di valutatori umani, e può contribuire a trovare i problemi che impediscono l'erogazione di una soddisfacente e coinvolgente esperienza di ricerca. Una valutazione comparativa dei sistemi di IR, di solito segue il paradigma Cranfield o altri approcci [Kursten and Eibl, 2011]; in questo contesto, l'efficacia di una strategia di recupero è calcolata in funzione del posizionamento (ranking) dei documenti, e ogni metrica è ottenuto facendo la media sulle domande. In genere, la valutazione dell'efficacia inizia con la raccolta delle informazioni necessarie/ricercate/descrittive di una serie di argomenti. In seguito a queste esigenze, viene definita una serie di query, e quindi un elenco di rilevanza e giudizi che mappano le query ai loro corrispondenti documenti pertinenti, i risultati. Dal momento che le persone spesso non sono d'accordo sulla rilevanza di un documento, la raccolta dei giudizi di rilevanza è un compito molto difficile. In molti casi, con un'approssimazione accettabile, la rilevanza viene considerata una variabile binaria, anche se è definita in un intervallo di valori [Spink and Greisdorf, 2001]. Per superare questi limiti, alcuni approcci iniziano la valutazione dell'efficacia del recupero delle informazioni senza giudizi di rilevanza, facendo uso dei giudizi di pseudo rilevanza [Soboroff et al., 2001] [Aslam and Savell, 2003], [Wu and Crestani, 2003]. A questo fine, strategie di ranking sono spesso eseguite confrontando i coefficienti di correlazione (ad esempio, Spearman [Callan et al., 1999], Kendall Tau) con classifiche ufficiali come TREC. L'efficacia dell'IR è valutata calcolando parametri rilevanti quali la precisione (precision), il richiamo (recall), la precisione media, R-precision, F-measure e il normalized discounted cumulative gain (NDCG) [Manning et al., 2008]. Sono anche utilizzate collezioni di prova e serie di valutazioni per sviluppare uno studio comparativo dell'efficacia del recupero delle informazioni (ad esempio, TREC, GOV2, NTCIR e CLEF).

Nel quadro dell'ottimizzazione delle soluzioni di IR, sono utilizzati approcci stocastici per migliorare l'efficacia IR. Con algoritmi genetici [Radwan et al., 2006], [Pathak et al., 2000a] è possibile migliorare l'efficacia tramite la riformulazione di query [Perez-Aguera, 2007], la selezione di query [Cecchini et al., 2008], [Yang et al., 1992]. Altre tecniche fanno uso di algoritmi Fuzzy [Snasel et al., 2009], analisi del contesto locale [Xu and Croft, 2000], clustering [Zhang et al., 2001], e il ranking improvement [Wang and Zhu, 2010].

2.2 Query e reasoning spaziali e temporali

Le soluzioni per Smart City producono grandi quantità di dati che hanno una grande variabilità, varietà, velocità e dimensioni; e quindi complessità. La varietà e la variabilità dei dati può essere dovuta alla presenza di diversi formati [Mulligan e Olsson, 2013], [Kebus et al., 2011], [Balakrishna 2012], [Wellington et al., 2013] e l'interoperabilità tra semantica dei singoli campi e delle diverse serie di dati [Bellini et al., 2014]. I dati statici sono raramente aggiornati, ad esempio

una volta al mese / anno, anziché i dati dinamici che vengono aggiornati: da una volta al giorno fino a ogni minuto per arrivare a dati in tempo reale. La velocità dei dati è legata alla frequenza di aggiornamento di dati per dati dinamici come posizione del bus, persone e loro stato, la posizione di raccoglitori di rifiuti, flussi di mezzi e persone, ecc. La dimensione dello storage per le Smart City cresce nel tempo accumulando nuovi dati ogni giorno, settimana e mese. A livello architettonico, le soluzioni Smart City in genere adottano architetture n-tier [Anthopoulos et al., 2014].

L'utilizzo di RDF store per Smart City è abbastanza recente, dal momento che nella maggior parte dei casi i servizi sono forniti come servizi verticali. Ad esempio, il sistema di trasporto intelligente, ITS, nella città fornisce informazioni per la posizione o meglio il tempo di arrivo di autobus, senza affrontare la posizione dei servizi della città, i flussi delle persone, gli eventi in città, e come questi aspetti e servizi sono posizionati ed associati con le fermate degli autobus. I servizi integrati sono in genere forniti da aggregatori di dati e fornitori di servizi che eseguono l'integrazione dei dati e consentono di sfruttare modelli di dati integrati. Alcuni integratori di dati della città sono a support di certi servizi come bike e car sharing, il sistema di navigazione, le informazioni turistiche, la prenotazione hotel, etc. Tutte queste soluzioni hanno la necessità di integrare le informazioni geo-location con dati in tempo reale e gli eventi in continuo arrivo da informazioni aggiornate come ad esempio: gli eventi, i voti, i flussi di traffico, commenti, etc. [Chourabi et al., 2012], [Wellington et al, 2013], [Khan et al., 2013]. Per queste applicazioni, l'RDF store (graph database) può essere una soluzione per far fronte alla variabilità dei dati, per permettere ragionamenti geo spaziali, nel tempo e in relazione a concetti complessi [Zaheer et al., 2013]. A questo proposito, ontologie a support di informazioni delle Smart City possono essere efficaci (si veda ad esempio <http://smartcity.linkeddata.es/>), prodotto da progetto Read4SmartCity della Commissione Europea. Uno dei modelli ordinati nel progetto di ricerca Read4SmartCity è Km4City ontologia che viene utilizzata in Sii-Mobility ed è anche adottata in progetti internazionali della Commissione Europea come RESOLUTE H2020, REPLICATE H2020.

Per la valutazione di RDF Store, vengono adottati specifici parametri di riferimento. Alcuni di essi possono essere basati su dati veri del mondo reale, mentre altri forniscono soluzioni per generare un insieme di dati di sintesi per la valutazione. Ad esempio, la soluzione LUBM [Guo et al., 2005] utilizza un set di dati di sintesi nel dominio universitario e copre solo la specificazione SPARQL 1.0 [SPARQL]. Al contrario, la soluzione BSBM [Bizer e Schultz, 2009] genera un set di dati di sintesi nel dominio di e-commerce e copre le analisi di business query SPARQL 1.1. Più recentemente, nel progetto Linked Data Consiglio benchmark sono stati proposti due modelli di riferimento e anche la generazione di un insieme di dati di sintesi, uno nel dominio semantico della pubblicazione (LDBC-SP) e l'altro nel dominio sociale network (LDBC-SN). Questi benchmark eseguono un mix di / aggiornamento / eliminazione e di query operazioni di inserimento e non solo il semplice accesso query. Si noti che l'attuale standard di SPARQL non copre le ricerche spaziali e per parole chiave pertanto una query che coinvolge questi aspetti deve essere costruita adottando costrutti specifici dell'RDF store utilizzato. Lo standard GeoSPARQL [GeoGeoSPARQL] è stato sviluppato dalla Open Geospatial Consortium per coprire le ricerche spaziali, anche se non molte soluzioni attualmente supportano questa specifica. Per quanto riguarda il punto di riferimento di geo e memorizza RDF spaziali il punto di riferimento Geographica [Garbis, et al., 2013] è stato proposto utilizzando sia un insieme di dati di sintesi generati e un vero e proprio set di dati. Analizza il supporto e prestazioni per relazioni spaziali avanzati tra entità spaziali complessi (ad esempio, poligoni). In [Duan et al., 2011], data set di riferimento veri e sintetici sono stati confrontati dimostrando che i dati sintetici sono simili a set di dati generati per parametri di riferimento relazionale di database (TPC-H) e fortemente differenti da serie di dati del mondo reale (ad esempio, DBpedia) che sono molto meno strutturati. In [Schmidt et al., 2009], con lo SPARQL

Benchmark (SP2Bench) si ha un quadro di riferimento specifico per la lingua progettata per i costrutti più comuni SPARQL. Esso include anche un generatore di dati e 17 query di riferimento generale, che non sono adatti per valutare gli aspetti Smart City, ma più si concentra sui metadati.

Recentemente SPARQL è stato esteso per interrogare dati in tempo reale provenienti da flussi di dati RDF. Generalmente la query SPARQL viene eseguita regolarmente nel corso degli stream utilizzando una finestra temporale scorrevole. I risultati dell'interrogazione sono forniti come un flusso continuo. Così la query viene registrato sul server e il client è in grado di connettersi per ricevere i risultati. Ci sono alcune implementazioni come C-SPARQL [Barbieri et al., 2009], SparqlStream [Calbimonte et al., 2010], CQELS [Le-Phuoc et al. 2011] e anche i parametri di riferimento sono stati definiti come SRBench [Zhang et al., 2012] utilizzando i dati provenienti da sensori meteo, LSBench [Le-Phuoc et al. 2012], dati provenienti da social network e CityBench [Ali et al., 2015] utilizzando i dati provenienti da sensori intelligenti della città. Questo tipo di parametri di riferimento sono adatti per lo streaming dei dati che le query che eseguono specifiche richieste con un numero limitato di risultati. W3C ha inoltre esaminato RDF benchmark degli store che mettono in evidenza la loro applicabilità per valutare i diversi aspetti degli RDF Store, e la loro applicazione in diversi store.

Nonostante questa ampio stato dell'arte su RDF store e i loro parametri di riferimento, nessuno degli approcci menzionati specificamente messo a punto per valutare i negozi di RDF contro Smart City. Smart City presenta condizioni molto particolari e specifiche che sfruttano i costrutti più impegnativi e nuovi degli RDF store come query geo-spaziali, le query di testo, query di tempo e combinazioni di esse.

Nel fornire servizi ai cittadini di una città intelligente, un RDF store dovrebbe fornire alcune funzionalità che consentono di supporto funzionalità specifiche. In particolare, le seguenti caratteristiche sono riportate in base alla loro rilevanza e la loro classificazione. In modo che i negozi Smart City dovrebbe fornire il supporto per:

- **indicizzazione spaziale:** fornire informazioni vicino ad un dato punto geografico: come posizione GPS. Per esempio, tutti i servizi che sono presenti vicino ad un dato punto. Si dovrebbe anche sostenere funzionalità avanzata geo-spaziale come in grado di gestire geometrie complesse (ad esempio, informazioni lungo un percorso ciclo, tutti elementi in una determinata zona poligonale).
- **indicizzazione full text:** permette l'integrazione di query semantiche con le ricerche di parole chiave basate sul testo che possono essere presenti negli attributi ed elementi di classe, come triple. I soggetti e oggetti delle triple possono contenere un'area di testo pertinente, quali le descrizioni, i nomi delle strade, i nomi di luoghi, etc.
- **quadruple** e non solo triple per associare metadati ai data set. Le triple sono prodotte sulla base di data set provenienti da diverse fonti. Pertanto, è importante tenere traccia sull'origine dati con metadati e licenze associate. Questa funzione è particolarmente utile per risolvere i processi di licensing e autorizzazione durante l'utilizzo dei dati da parte degli utenti finali tramite le API., in base ai loro profili ed autorizzazione. Per esempio: per cittadini, operatori della protezione civile, saranno accessibili dati diversi.
- **inferenza** come definita per RDFS base o con profili OWL2 più avanzati che permettono di inferire nuovi fatti dai dati disponibili. Questo può essere usato per generalizzare / specializzare sulle entità, tramite sinonimie, equivalenze, relazioni transitive, relazioni simmetriche, etc. La deduzione può implicare la materializzazione di triple nella fase di indicizzazione [Bellini et al., 2015].
- **indicizzazione temporale:** molte informazioni e le caratteristiche che sono memorizzate negli RDF store per smart city cambiando nel corso del tempo. Ad esempio, la situazione meteo e previsioni, il flusso di traffico rilevato dai sensori del traffico, la posizione del bus e gli eventi

che accadono in città. Per questo motivo, è molto importante che l'archivio RDF possa sostenere delle ricerche su base temporale per consentire un facile recupero dei dati che fanno riferimento allo stesso momento. La memorizzazione delle informazioni temporali (che potrebbe cambiare in tempo reale) è la fonte principale dell'aumento delle dimensioni del RDF store, chiedendo soluzioni Big Data per smart city per il volume, velocità e varietà, almeno. Una soluzione a questo aspetto è la memorizzazione nell'RDF store solo degli ultimi dati relativi al sensore tramite azioni di upload/update del "last value", senza aggiungere di fatto triple che vanno a creare il dato storico delle misurazioni. Il dato storico ovviamente non va perduto ma deve essere memorizzato altrove, per esempio.

- **grandi volumi di traffico.** La gestione di RDF store bigdata tramite l'accesso di molti utenti che interrogano i dati è molto impegnativa, per questo motivo può essere necessaria una soluzione di clustering. Potrebbe essere un raggruppamento (scaling verticale o scalare up / down) quando lo stesso servizio è duplicato per consentire l'esecuzione di molte query simultanee ma anche per fornire una soluzione di tolleranza ai guasti. Una soluzione diversa può consistere in un scale-out come clustering (orizzontale) quando i dati vengono suddivisi tra i diversi server, visto che il singolo server non può gestire tutti i dati.

Un requisito non funzionale molto rilevante è dovuto al fatto che molte applicazioni Smart City sono acquisite dalle Pubbliche Amministrazioni tale richiesta si concretizza in (i) soluzioni standard per evitare il rischio di vendor lock-in specie per molte nuove tecnologie come gli Store RDF, e per (ii) soluzioni open source per essere compatibile con le leggi nazionali delle pubbliche amministrazioni che tipicamente spingono per avere soluzioni aperte: standard ma anche con il codice sorgente accessibile e condivisibile tra più amministrazioni pubbliche. Inoltre, ci dovrebbe essere una comunità attiva in grado di sostenere l'evoluzione del prodotto.

L'effettiva capacità di svolgere reasoning geospaziale e temporale in una KB sviluppata tramite un RDFstore dipende da vari fattori:

- la struttura dell'ontologia: la presenza di una ontologia a supporto del reasoning spaziale e temporale. In particolare l'ontologia deve predisporre le strutture in modo che tramite le istanze sia effettivamente effettuare delle query SPARQL tali da sfruttare inferenza e ragionamento.
- la realizzazione di query semantiche che possano sfruttare l'inferenza nelle sue forme anche in relazione ai dati effettivamente caricati nell'RDF store.
- Il supporto al reasoning spaziale nell'RDF store, per esempio tramite degli indici specifici ed una gestione particolare di attributi delle classi che potrebbero rappresentare coordinate GPS.

Table. RDF stores features comparison, where:
OS=Open Source, Cm=Commercial, H=Horizontal cluster, V=Vertical cluster

RDF Store	Features											
	SPARQLv.	Inference	3/4-tuple	Spatial search	Full-text search	Storage	Temporal search	Size (million/billion triples)	License	Dev. Language	Cluster support	Active project
Virtuoso 7.2.4 OS	1.1	RDFS+	4	Adv	Y	RDBMS	N	50BT	OS	C	N	Y
Virtuoso 7.2.4 Comm	1.1	RDFS+	4	Adv	Y	RDBMS	N	50BT	Cm	C	H	Y
Graph DB SE 7.0.1	1.1	OWL2RL	4	Bas	Y	custom	N	10BT	Cm	Java	H	Y
Stardog 4	1.1	OWL2	4	Adv	Y	custom	N	10BT	Cm	Java	H	Y

RDF Store	Features											
	SPARQLv.	Inference	3/4-tuple	Spatial search	Full-text search	Storage	Temporal search	Size (million/billion triples)	License	Dev. Language	Cluster support	Active project
Oracle 12c	1.1	RDFS, OWL2	4	Adv	Y	custom	N	1TT	Cm	C/Java	H	Y
Apache Jena-Fuseki	1.1	RDFS OWL-Lite	3	Bas	Y	custom (TDB)	N	1.7BT	OS	Java	N	Y
Apache Jena-Fuseki	1.1	No	4	Bas	Y	custom (TDB)	N	1.7BT	OS	Java	N	Y
Blazegraph 2.1.2	1.1	RDFS+	3	Bas	Y	custom	N	50BT	OS	Java	V&H	Y
Blazegraph 2.1.2	1.1	No	4	Bas	Y	custom	N	50BT	OS	Java	V&H	Y
CumulusRDF	1.1	No	3	No	N	Cassandra 1.2	N	120MT	OS	Java	V	(Y)
Strabon	1.1	No	3	Adv	N	RDBMS	Y	500MT	OS	Java	N	(Y)
4store	1.1	No	4	No	N	custom	N	15BT	OS	C	V	N
h2rdf+	1.0	No	3	No	N	HBase	N	2.7BT	OS	Java	H&V	N

2.3 Indicizzazione e versioning di RDF Store

Le grandi banche dati a grafo stanno subendo un forte spinta dal mercato per la creazione di nuovi servizi per le smart city, le biblioteche digitali, modellazione delle competenze, l'assistenza sanitaria, l'istruzione intelligente, ecc. Questo fatto è dovuto principalmente alla loro capacità di rappresentare conoscenza e quindi sulla creazione di conoscenza, come Knowledge base [Grosan e Abraham, 2011]. I database a grafo possono essere implementati come RDF Store (Resource Description Framework) [Klyne e Carrol 2004], per creare servizi interattivi in cui il ragionamento e le deduzioni possono essere prodotte da un motore inferenziale. Un archivio RDF si fonda sul concetto di tripla che mette in relazione due entità. Ad esempio, Carl conosce Paolo, consiste in un soggetto, un predicato e un oggetto, che a loro volta sono rappresentati con URI. Predicati, come "conosce", possono essere specificati usando il vocabolario in cui sono definite relazioni e le loro relazioni (per esempi di specializzazione).

Nel contesto di una KB e quindi in un RDF store, un vocabolario definisce le caratteristiche comuni di cose appartenenti a classi e le loro relazioni. Un vocabolario, viene chiamato anche ontologia, e viene definita utilizzando formalismi come RDFS (RDF Schema, RDF Vocabulary Description Language) o l'estensione OWL (Ontology Web Language). Recentemente, gli RDF store sono stati anche classificati come NoSQL per la gestione di big data [Bellini et al., 2013a]. Un grande insieme di ontologie e relativi set di dati sono ora accessibili, si veda ad esempio il gran numero di LOD (linked open data) accessibile e connessi tra loro tramite URI [Berners-Lee, 2006], [Bizer e Schultz, 2011]. Gli RDF store possono essere resi accessibili tramite un punto di ingresso per postare query semantiche formalizzate per esempio in SPARQL [Hartig et al., 2009] (SPARQL Protocol e RDF Query Language, definizione ricorsiva). Le soluzioni non banali di RDF Store sono tipicamente basate sfruttando molteplici ontologie, e supportano il caricamento di dati, il testing / validazione dei risultati ottenuti con delle regole semantiche. Ciò significa che essi sono costruiti utilizzando metodologie per la costruzione di ontologie [Noy e McGuinness, 2001], [Lopez 1999], supportando lo sviluppo della KB tramite un ciclo di vita [Bellini et al., 2015].

Gli RDF store possono crescere nel tempo per (i) l'aggiunta di nuove triple, e (ii) per la loro capacità di imparare, se dotati di un inferenziale ragionatore / motore. Vale a dire, sono in grado di produrre nuova conoscenza, che si formalizza in termini di nuove triple. Così, il motore inferenziale associato al RDF store materializza nuove triple durante ragionamento (ad esempio al momento

della indicizzazione o interrogazione). Questo ultimo fatto è la principale motivazione alla base di basse prestazioni in indicizzazione, e per le prestazioni in caso di eliminazione di triple di RDF store in quanto sono coinvolti nella rimozione anche triple materializzate nello store. Mancando (per cancellazione) fatti che sono stati utilizzati per dedurre altri fatti, anche i fatti dedotti vengono a perdere di significato. Queste caratteristiche impatto sulle prestazioni dello store, e quindi, in letteratura, sono stati proposte molte soluzioni per la valutazione di RDF store. Alcuni di questi usano dati reali come da DBpedia, UniProt, WordNet, altri usano dati sintetici generati come LUBM [Guo et al., 2005] (dominio universitario), BSBM [Bizer et al., 2009] (dominio e-commerce), SP2Bench [Schmidt et al., 2009] (dominio biblioteca).

Recentemente, nel progetto Linked Data Benchmark Council LDBC EU sono stati sviluppati due nuovi parametri di riferimento: [Erling et al, 2015] basato sui Social Network, e il secondo su Semantic Publishing. Mentre i benchmarks LUBM e SP2Bench si basano su dati reali, e valutano solo le query eseguite dopo il caricamento di tali dati. I benchmark BSBM e LDBC valutano un mix di carichi di lavoro di inserimento / aggiornamento / cancellare / query con dati sintetici. Quando RDF store sono utilizzati come supporto per una KB, alcuni dei cambiamenti nello store RDF possono essere distruttivi per il modello a grafo, come per esempio i cambiamenti nelle triple che interessano la definizione dei modelli dell'ontologia su cui sono riportate milioni di istanze, casi e dati. Al fine di mantenere le prestazioni accettabili, lo store RDF deve essere ricostruito da zero o da una qualche versione parziale per risparmiare tempo nel rilascio della nuova versione. Così, il ciclo di vita degli RDF store può presentare più cicli in cui lo store RDF è costruito in modo incrementale tramite i parametri progressivi di mediazione tra: (i) il riutilizzo di modelli ontologici, (ii) l'aumento della capacità di fare deduzioni e ragionamenti sulla base di conoscenza, (iii) il mantenimento delle prestazioni di query e di risposta dei dati accettabili, (iv) semplificando la progettazione dei servizi di front-end, (v) soddisfare l'arrivo di ulteriori dati e modelli e / o correzioni, ecc Un modello del ciclo di vita di condiviso e standard per costruire KB non è ancora disponibile e molti ricercatori hanno cercato di incorporare KB fasi di sviluppo in alcuni modelli del ciclo di vita del software convenzionali [Batarseh, Gonzalez, 2013]. In generale, lo sviluppo di sistemi KB è un processo multifasico che procede iterativamente, utilizzando una strategia di prototipazione evolutiva. Un certo numero di modelli del ciclo di vita sono stati proposti specificamente per i sistemi KB da [Milette 2012] e anche in [Bellini et al., 2015].

Nel modello del ciclo di vita, un cambiamento nell'ontologia può generare la revisione e rigenerazione di una vasta quantità di triple RDF. Il problema della gestione delle versioni dell'ontologia come affrontato in [Klein et al., 2002] e [Noy e Musen 2004] può essere applicato facilmente se l'ontologia non è utilizzato come base per la creazione di un RDF store big data, di una KB. In [Volkel et al, 2005] il controllo delle versioni di CDR KB è stato affrontato in modo simile alle soluzioni CVS utilizzando i comandi come: commit, update, branch, merge, and diff. Le differenze sono calcolate a livello semantico sui file di triple. [Zegins et al., 2007] ha presentato una soluzione per la gestione di versioni RDF ipotizzando la possibilità di stimare il delta tra due modelli RDF eseguendo una serie di aggiunge e cancella da un modello per trasformarlo nell'altro. A livello di database, gli aspetti chiave di un sistema per la gestione delle versioni in RDF store sono lo spazio di archiviazione e il tempo per creare una nuova versione [Tzitzikas et al., 2008]. Pertanto, possibili approcci potrebbero essere quella di memorizzare: (a) ogni versione come triple negozio indipendente [Klein et al, 2002], [Noy e Musen 2004], [Volkel et al, 2005]; (B) i delta in termini di triple tra due versioni consecutive e implementare un processo per applicare tale delta [Zegins et al., 2007].

3 Calcolo di matrici origine destinazione e traiettorie

3.1 Flussi e traiettorie

L'acquisizione dei dati di flusso [M.G. Messina et al. 2011] è effettuata tramite l'utilizzo di tecnologia GPS, oppure tramite altri sensori [Traffic 2013]. Un esempio di caso di studio, avente ad oggetto un'area metropolitana, per il monitoraggio dei flussi di traffico, è presentato in [A. Costanzo 2013]. Studi per l'analisi delle interazioni causali e di pattern anomali in dati di traffico sono presentati rispettivamente in [W. Liu et al. 2011] e [L. X. Pang 2013, L. X. Pang 2013]. L'analisi della mobilità può avere applicazioni diverse, come per esempio la pianificazione di servizi ad essa connessi.

Per esempio in [A. Gemma 2011] viene presentato un sistema di monitoraggio e previsione della mobilità veicolare per l'integrazione tra la rete dell'illuminazione pubblica e la rete della mobilità. In [X. Xiao 2010] è proposta una metodologia di stima di similarità fra gli utenti, utilizzando le loro tracce GPS. Questo approccio modella semanticamente (SLH) le traiettorie dell'utente (e.g., shopping mall -> ristoranti -> cinema), e poi misura la similarità fra SLH con un algoritmo di travel match (MTM).

L'analisi delle traiettorie consente anche di stimare i consumi di carburante e le emissioni degli autoveicoli [J. Shang et al. 2014]. In [Y. Zheng et al. 2010] è descritto un esempio di social networking service per l'analisi delle traiettorie, luoghi e persone, e per scoprire le correlazioni fra utenti e luoghi, in termini di traiettorie GPS. Altri studi di similarità fra gli utenti che utilizzano le tracce GPS sono descritti in [X. Xiao et al. 2012, Q. Li et al. 2008].

Inoltre sono stati sviluppati sistemi per la raccomandazione di amici a partire dallo studio delle tracce [Y. Zheng et al. 2011]. Lo studio delle traiettorie è utilizzato anche per costruire le traiettorie più popolari a partire dalle singole tracce [L.W. Wei et al. 2012], oppure per categorizzare i luoghi [J. Yuan 2012]. Considerando l'esperienza di viaggio degli utenti e la sequenzialità dei luoghi visitati, è possibile calcolare la correlazione delle traiettorie GPS [Y. Zheng and X. Xie 2010, Y. Zheng et al. 2009].

La modalità di trasporto può essere inferita anche con approcci supervised learning [Y. Zheng et al. 2008]. Lo studio delle traiettorie è utilizzato per fornire raccomandazioni di viaggio, come per esempio in [Y. Zheng 2011, Y. Zheng et al. 2009, H. Yoon et al. 2010]. Inoltre sono stati studiati algoritmi per il map-matching globale di traiettorie a basso campionamento, come per esempio in [Y. Lou 2009]. In modo complementare è stato studiato il problema della ricerca delle traiettorie a partire dai luoghi, dove la query è rappresentata da una piccola parte dei luoghi, senza un ordine specifico, mentre l'obiettivo è quello di trovare le *k* Best-Connected Trajectories (*k*-BCT), in modo tale da connettere nel modo migliore le località [Z. Chen et al. 2010]. Applicazioni alle social network sono state studiate utilizzando la posizione degli utenti [Y. Zheng 2012]. Un altro studio per la raccomandazione del percorso più efficiente a partire dalle tracce GPS è descritto in [J. Yuan et al. 2013, J. Yuan et al. 2013].

3.2 Acquisizione di dati di flusso da Wi-Fi

L'acquisizione dei dati di flusso relativi agli utenti in città è un problema importante, nell'ottica di un'efficiente gestione del traffico urbano e della fornitura di servizi ai cittadini. A questo riguardo è utile sviluppare strumenti e tecniche per ottimizzare lo sfruttamento dei vari servizi e misurare in modo efficiente i flussi di traffico urbano.

In particolare l'ottimizzazione dei servizi per i cittadini è una delle sfide più impegnative nell'ambito dell'urban computing [Y. Zheng et al. 2014] e più in generale delle Smart City [Smart cities 2007]. I servizi ai cittadini fanno generalmente riferimento alla mobilità, alla governance,

all'energia, agli eventi culturali, al commercio, all'ambiente. Fra questi, la mobilità è considerata un bene (commodity); quindi l'analisi della mobilità dei cittadini è uno degli aspetti più rilevanti del concetto di Smart City. Secondo [Giffinger, R. et al. 2007], la Smart Mobility [Smart Mobility 2012] è uno dei fattori chiave nel contesto delle moderne Smart City [European Parliament 2014], inclusa l'accessibilità locale e internazionale, la disponibilità di infrastrutture ICT, e di sistemi di trasporto innovativi e sicuri [Caragliu A. 2009].

Nel contesto della mobilità, l'analisi dei flussi di traffico è un importante prerequisito per lo smistamento del traffico, che costituisce la parte centrale del cosiddetto Intelligent Transportation Systems (ITS), per la gestione del trasporto pubblico. L'analisi dei dati di flusso è comunemente usata per facilitare la gestione dei trasporti, per regolare l'accesso controllato alle città, per lo Smart Parking, per la sorveglianza del traffico, per fornire informazioni circa le condizioni delle strade, o per monitorare e controllare le condizioni ambientali, come per esempio le emissioni nocive (e.g., CO₂, PM₁₀, ozono). La Commissione Europea indica, fra i temi principali che dovrebbero essere considerati con speciale attenzione nel contesto del processo CARS 2020, l'implementazione e la promozione dell'ITS [European Commission 2014].

Alcune delle tecniche adottate per il monitoraggio e la gestione del traffico possono essere utilizzate per l'analisi dei flussi delle persone. È importante conoscere il movimento delle persone con una certa precisione, e rilevare dove e come le persone attraversano la città e utilizzano i suoi servizi, per esempio i vari mezzi di trasporto (e.g., auto, bicicletta, taxi, car sharing, bus, tram). Tipicamente, gli operatori di TLC non sono capaci di fornire questo tipo di informazione: conoscono il numero di persone connesse ad ogni cella telefonica in un determinato istante di tempo, ma non hanno informazioni su come gli utenti si muovono nella città.

A questo proposito sono necessari strumenti specifici per strumenti mobile per il tracking delle persone, ovviamente previa autorizzazione da parte degli utenti.

Recentemente, sono state introdotte tecniche che fanno uso di Bluetooth, Real Time Location System (RTLS) e Wi-Fi access point [Danalet, A. 2014, Patil, P. H. 2015]. In alcuni casi la posizione di un veicolo può essere monitorata a partire dalla sua posizione (tramite apparati GPS installati sul veicolo), oppure con l'ausilio di navigatori per smartphone (e.g., Google Maps, TomTom, Waze). RFID è inadatto a causa del suo corto raggio d'azione. Allo stesso modo Bluetooth non è molto utile perché richiederebbe l'installazione di specifiche stazioni di rilevamento. Al contrario, un sistema costituito da access point Wi-Fi, seppure meno preciso dei sistemi GPS, può essere molto conveniente per la stima dei flussi delle persone. In questo contesto gli access point Wi-Fi sono considerati come dispositivi di rilevazione della presenza; data l'alta diffusione dei sistemi mobile, tutti provvisti di funzionalità Wi-Fi, questo sistema si rivela essere molto economico perché non richiede l'installazione di device specifici sulle auto. Va inoltre considerato il basso costo per l'installazione degli access point Wi-Fi, e la gratuità dell'accesso alle reti cittadine; in molti casi le città sono già provviste di un certo numero di access point, gestiti dalle varie municipalità, che coprono più o meno uniformemente le aree cittadine. Si pone quindi il problema del posizionamento efficiente di questi access point, al fine di ottenere una copertura omogenea ed efficiente delle città.

3.3 Acquisizione di dati di flusso da Spire e sensori

Lo sviluppo di Sistemi di Trasporto Intelligente (ITS – *Intelligent Transport Systems*) richiede dati di flusso di alta qualità in tempo reale. L'utilizzo di sensori posti sulla via di transito (es. a *loop* induttivo) per la raccolta dati è necessario, ma non sufficiente, data la loro limitata copertura e i loro alti costi di adozione e manutenzione.

Negli ultimi anni sono state sviluppate diverse sorgenti di dati alternative, come ad esempio il metodo basato sulla collezione di dati geolocalizzati, di velocità, direzione di viaggio e informazioni temporali provenienti dalle unità di bordo dei veicoli guidati (FCD - *Floating Car*

Data). Questo metodo risulta redditizio e compensa le limitazioni che derivano da rilevatori fissi. Anche se l'idea di raccogliere dati provenienti da dispositivi situati all'interno del veicolo (telefoni cellulare o GPS) non è nuova, il mercato FCD sta crescendo solo ora a livello mondiale, portando una vasta gamma di applicazioni e benefici. Questi dati non migliorano solamente la gestione del traffico, ma aiutano anche a soddisfare la crescente domanda degli automobilisti che sono disposti a pagare un servizio per avere accesso a informazioni in tempo reale. Per soddisfare la loro richiesta di informazioni sulla congestione del traffico, però c'è bisogno di ottenere dati di flusso accurati, affidabili, puntuali e completi.

Le tecnologie convenzionali comprendono quei metodi che misurano dati di flusso attraverso rilevatori ubicati sul ciglio della strada. Esse possono essere suddivise in due categorie: metodi intrusivi e non-intrusivi.

I metodi intrusivi consistono in un registratore di dati e in un sensore posto sulla o nella strada. Vengono impiegati da molti anni e i più importanti sono:

- **Tubi pneumatici [Messina et al., 2011], [Leduc 2008]:** sono tra i primi e più semplici dispositivi di conteggio del traffico, dove il sensore è costituito da un tubo di gomma, steso sulla superficie della pavimentazione stradale in direzione ortogonale al senso di marcia e collegato ad un apparecchio contatore disposto al margine della strada. Questo sistema rileva i veicoli dall'onda di pressione generata dal passaggio del veicolo sul tubo di gomma, che attiva un interruttore a membrana, che funge da rilevatore, posto su un'estremità del tubo. Il segnale creato viene quindi registrato e processato dal contatore, che effettua il conteggio dei veicoli facendo corrispondere ad ogni coppia di impulsi un'unità di vettura equivalente. Questo dispositivo di conteggio da una parte ha il vantaggio dell'economicità e della semplicità di installazione e rimozione, ma presenta significative limitazioni in termini di copertura stradale, oltre che un'elevata imprecisione di conteggio nel caso di flussi elevati (con errori anche superiori al 20%) o flussi a basse velocità.
- **Sensori piezoelettrici [Messina et al., 2011], [Leduc 2008]:** i sensori sono generalmente posizionati in un solco all'interno della pavimentazione stradale monitorata. Il principio è quello di sfruttare le proprietà piezoelettriche dei cristalli di quarzo che, sottoposti ad una sollecitazione dinamica, si elettrizzano. La deformazione meccanica del materiale piezoelettrico crea delle distribuzioni non omogenee di elettricità, portando ad una differenza di potenziale sulle facce sollecitate. L'estensione e la frequenza del segnale è direttamente proporzionale al grado di deformazione. Questo sistema può anche essere usato per misurare peso e velocità dei veicoli in marcia.
- **Spire ad induzione magnetica [Messina et al., 2011], [Leduc 2008]:** sono la tecnologia più diffusa per l'acquisizione di dati di flusso. Il sensore è costituito da uno o più avvolgimenti di conduttore elettrico installati nella pavimentazione, alimentati da un generatore di corrente, che permette la creazione di un campo magnetico. Questi sono poi collegati ad un dispositivo di conteggio, posizionato al margine della strada. Il sensore ha una breve aspettativa di vita, dato che può essere danneggiato da veicoli pesanti, ma non risente di negative condizioni meteorologiche. Questo metodo è stato ampiamente utilizzato in Europa negli ultimi decenni, anche se la loro adozione e manutenzione prevede costi piuttosto alti.

Metodi non-intrusivi sono invece basati su tecnologie a controllo remoto, con sensori collocati a lato o al di sopra della strada. In questo caso, il conteggio manuale è il metodo più utilizzato, anche se esistono nuove tecnologie in via di sviluppo.

- **Conteggio manuale [Leduc 2008] [Ministero]:** questo metodo si basa sull'acquisizione dei dati di traffico da parte di un operatore umano, che rileva dati che non possono essere ottenuti in maniera efficiente da dispositivi di conteggio automatici, come ad esempio il

tasso di occupazione dei veicoli, i pedoni e la classificazione dei veicoli. Le attrezzature più utilizzate includono schede di censimento e apparecchi contacolpi. Tuttavia l'osservatore presenta dei limiti naturali, che non possono garantire la correttezza e l'esattezza delle informazioni rilevate, come la non istantanea velocità di registrazione dati, o l'affaticamento, che determina una minore attendibilità all'aumentare del tempo di rilevamento.

- **Sensori a raggi infrarossi passivi e attivi [Leduc 2008] [Ministero]:** in questo caso la presenza e velocità (solo da parte dei sensori attivi) dei veicoli sono rilevati da un apparecchio ricevitore in grado di rilevare l'energia delle radiazioni infrarosse emesse dalla pavimentazione stradale o dalla superficie dei veicoli attraversanti la sua zona di influenza. Queste due tipologie di sensori presentano il vantaggio di non causare disturbo alla circolazione stradale durante la loro installazione, mentre si verifica un degradamento delle loro prestazioni in condizioni di pioggia o neve, ed hanno un limitato raggio di copertura stradale.
- **Sensori magnetodinamici [AIPCR]:** i sensori magnetodinamici rilevano il passaggio di veicoli stradali basandosi sull'analisi della variazione del campo magnetico terrestre indotta dall'interferenza della loro massa metallica. Possono essere posizionati all'interno della pavimentazione stradale o sulla superficie del piano viabile. Oltre al conteggio del numero di veicoli sono anche in grado di fornire i tempi di occupazione, i distanziamenti temporali e di classificare i veicoli in funzione della velocità e della lunghezza. Presentano comunque qualche imprecisione in presenza di code.
- **Sensori a microonde (ad effetto Doppler) [Leduc 2008] [AIPCR]:** installati fuori dalla sede viaria ed orientati verso la corrente veicolare oggetto di rilevamento, questi sensori possono rilevare veicoli in movimento e la loro velocità. Il radar è infatti in grado di registrare dati di conteggio, velocità e fare una semplice classificazione dei veicoli, inoltre presenta il vantaggio di non essere influenzato dalle condizioni atmosferiche, ma risulta più costosa rispetto ad altri rilevatori tradizionali installati sulla superficie stradale, anche se nel lungo termine possono dimostrarsi più economici grazie a minori costi di manutenzione.
- **Sensori acustici (ad ultrasuoni e acustici passivi) [Leduc 2008] [AIPCR]:** questi dispositivi emettono onde sonore, riconoscendo i veicoli dal calcolo del tempo impiegato dall'onda emessa per lasciare la sorgente, rimbalzare sulla superficie riflettente e ritornare sulla sorgente stessa. I sensori ad ultrasuoni sono posti sopra la superficie stradale (es. sotto un cavalcavia), mentre i sensori acustici passivi possono essere posizionati sia sopra che lateralmente alle corsie da monitorare. Entrambi i sensori sono in grado di rilevare il flusso, il tasso di occupazione delle corsie e classificare i veicoli in funzione dell'altezza i primi e della lunghezza i secondi, ma possono anche essere influenzati da cattive condizioni atmosferiche (es. basse temperature e neve). Infine i sensori acustici passivi sono anche in grado di misurare la velocità veicolare.
- **Trattamento di immagini video:** i sistemi di rilevamento con immagini video si basano sull'impiego di telecamere che acquisiscono scene di traffico di un tronco stradale. Il trattamento automatico delle scene di traffico sono in grado di fornire non solo i parametri fondamentali del traffico (es. portata, velocità, tasso di occupazione, densità), ma consentono anche di rilevare anomalie di deflusso, come veicoli fermi, incidenti, ingorghi, ecc. Questo avviene tramite differenti tecniche video, come i sistemi *tripline* e *tracking*. D'altro canto richiedono ingenti capacità di calcolo e sono sensibili a riflessi, ombre, e condizioni climatiche avverse, come ad esempio pioggia, nebbia e neve.

Infine, un'altra categoria è invece quella dei sistemi di **identificazione automatica** (AIS o AVI - *Automotive Vehicle Identification*). In questo caso, da una postazione fissa, i veicoli sono

identificati tramite tag elettronici (es. targhe), che vengono letti al passaggio del veicolo davanti al sensore [Leduc 2008].

La Tabella 1 presenta una panoramica delle varie tecnologie sopra elencate e dei dati che si possono acquisire.

Tabella 1: Dati acquisiti dalle tecnologie convenzionali che utilizzano spire e sensori, Source: [Leduc 2008]

Metodi Acquisizione Dati	Dati di flusso	Tasso di occupazione	Categoria veicoli	Velocità	Tempo di viaggio	Info O/D	Rilevamento incidenti
Tubi pneumatici	Si	Indiretto (derivato da numero di assali)	Parziale (da numero di assali)	Con due rilevatori, non accurata	No (non accurato)	No (non accurato)	No (non accurato)
Sensori piezoelettrici	Si	Si	Si (con peso assali)	Con due rilevatori	No (stime con algoritmi)	Non utilizzato	Non utilizzato
Spire a induzione magnetica	Si	Si	Se velocità disponibile	Con due spire consecutive	No (stime con algoritmi)	No, a meno di algoritmi specifici (richiedere molti sensori)	Si (a meno di algoritmi specifici)
Sensori infrarossi	Si	Si	No	Si, ma solo per sensori attivi	No	No, a meno di algoritmi specifici (richiedere molti sensori)	No
Sensori magneto-dinamici	Si	Si	Si, con due rilevatori	Si, con due rilevatori	No	No, a meno di algoritmi specifici	No
Sensori microonde	Si	Si	Si	Si	No, a meno di algoritmi specifici	No, a meno di algoritmi specifici (richiedere molti sensori)	Si, tracciando il veicolo in movimento
Sensori acustici	Si	Si	Si	Si, ma solo sensori passivi	No	No, a meno di algoritmi specifici	No
Immagini video	Si	Si	Si	Si	No (stime con algoritmi)	Non utilizzato	Non utilizzato

Immagini video con riconoscimento targa	Si	Si	Si	Si	Si (tracciando n. di targa)	Si (tracciando n. di targa)	Si, se combinato con algoritmi di rilevazione incidenti
Tag	Si (se ne sono dotati abbastanza veicoli)	Si (se ne sono dotati abbastanza veicoli)	Si, se l'info è contenuta nel tag	Si	Si	Si (entrata/uscita)	Si

Tra le tecnologie emergenti è importante il cosiddetto *Floating Car Data (FCD)*, un metodo che determina l'andamento del traffico in tempo reale basandosi sulla localizzazione dei veicoli l'unità di bordo o dispositivo mobile sull'intera rete stradale. In pratica, questo significa che ogni veicolo, equipaggiato con un telefono cellulare attivo o GPS, può agire da sensore nella rete stradale. I dati di geolocalizzazione del veicolo, velocità e direzione di viaggio vengono mandati anonimamente ad un'unità centrale di elaborazione. Una volta raccolte ed estratte, le informazioni utili (ad esempio informazioni sullo stato del traffico e su strade alternative) possono essere ridistribuite ai conducenti.

I *Floating car* sono una fonte di dati di alta qualità alternativa e complementare alle attuali tecnologie. Aiutano a migliorare la sicurezza, l'efficienza e l'affidabilità del sistema di trasporto e stanno diventando una tecnologia chiave per lo sviluppo dei nuovi sistemi di trasporto intelligente.

Generalmente i FCD possono essere raccolti in due modi [Leduc 2008] [ministero]: con rilevamento basato su dispositivi cellulari e rilevamento mediante tecnologie di tipo satellitare.

Il metodo di rilevamento basato su celle consente di produrre stime del traffico ovunque sia presente una copertura cellulare, utilizzando i dispositivi di telefonia mobile attivi all'interno dei veicoli guidati. Ciò ha il vantaggio di non richiedere infrastrutture aggiuntive, come sensori, telecamere o ricevitori GPS montati su veicoli ed ha un impatto nullo sui terminali d'utente. Richiede però che l'infrastruttura della rete cellulare sia adeguata al fine di raccogliere informazioni aggiuntive, operazione che prevede significativi investimenti da parte dei gestori della rete stessa [5T]. Dal momento che oggi la maggior parte dei veicoli è dotata di almeno un telefono cellulare, essi diventano sorgenti anonime di traffico. La localizzazione del dispositivo, che è costantemente trasmesso alla rete, viene effettuata tramite un processo chiamato triangolazione, che permette di ricavare la posizione approssimata basandosi sui segnali che questo scambia con i ripetitori (es. reti CDMA¹, GSM², UMTS³ e GPRS⁴), che si servono delle aree di territori e che identificano le celle [Palara 2013]. Elaborando i segnali ricevuti e come questi variano nel tempo è possibile stimare la posizione, direzione e velocità del dispositivo, e quindi del veicolo. La precisione di queste reti aumenta nelle aree metropolitane, dove la maggiore richiesta di traffico comporta un maggior numero di ripetitori distribuiti nello spazio. Anche se la precisione di localizzazione è generalmente bassa (intorno a 300m), questa debolezza è parzialmente compensata dalla grande disponibilità di dispositivi. Inoltre, una precisione di dati maggiore può essere ottenuta con la tecnologia UMTS (3G) [Palara 2013].

Il rilevamento mediante tecnologie di tipo satellitare può avvenire con dispositivi dotati di un ricevitore GPS, come navigatori o smart phones, dai quali è possibile ricavare la posizione e

¹ Code Division Multiple Access, soprattutto utilizzato in Nord America

² Soprattutto utilizzato in Europa, Cina e America Latina

³ Universal Mobile Telecommunications Systems (anche chiamato 3GSM)

⁴ General Packer Radio Service (anche detto GSM++ o GSM2+)

comunicarla tramite comunicazioni radio o rete mobile. In particolare, lo sviluppo della telematica applicata al trasporto stradale sta consentendo l'installazione sui veicoli di terminali di bordo dotati di sistemi di localizzazione satellitare, collegati via etere a centrali operative remote [Ministero]. Questo sistema permette di garantire da una parte servizi di assistenza e protezione ai conducenti, e dall'altra servizi di informazione (infomobilità). I nuclei telematici a bordo dei veicoli si basano su un sistema di telecomunicazioni (GSM – *Global System for Mobile communications*), un sistema di identificazione o localizzazione satellitare (GPS – *Global Positioning System*) e un elaboratore, per la gestione delle unità periferiche associate [Ministero]. In questo caso, la precisione di localizzazione del veicolo è relativamente alta, generalmente meno di 30m [Leduc 2008]. Anche se questo tipo di sistemi sta diventando sempre più conveniente, solo un limitato numero di veicoli ne è al momento dotato, e tipicamente sono servizi di gestione flotte (es. Taxi). In generale, dati di traffico provenienti da veicoli privati o camion sono più adatti per autostrade o zone rurali, mentre flotte di taxi sono particolarmente utili per il traffico urbano, dato il maggior numero di veicoli e i loro sistemi di comunicazione di cui sono dotati.

3.4 Calcolo di matrici OD

Per la stima dei flussi delle persone è di grande utilità il calcolo della cosiddetta matrice Origine/Destinazione (Origin Destination Matrix). Una matrice OD per i flussi di traffico, generalmente presenta su entrambi gli assi le zone della città, che possono essere quartieri storici oppure zone predefinite (e.g., quadrati o rettangoli). I singoli elementi della matrice rappresentano il numero di persone che passano da una zona all'altra (riga-colonna), in un determinato momento o finestra temporale (e.g., giorno, mese). La stima della matrice OD a partire dai dati di traffico è utile per la predizione e la gestione dei flussi, in particolare per pianificare percorsi ottimizzati (e.g., più corti ed efficienti), e per la fornitura di servizi di info-traffico sui dispositivi mobile, per esempio tramite il cosiddetto Advanced Traffic Management Systems (ATMS).

Le matrici OD possono essere tempo-dipendenti, e quindi necessitano di una stima in tempo reale, o almeno la valutazione dei valori nel corso del giorno o della settimana. D'altra parte, le matrici OD sono sensibili alle condizioni del traffico e quindi possono essere utilizzate come descrittori della mobilità. In pratica, si è interessati a misurare il flusso tipico fra diverse zone della città. Nel contesto del flusso di traffico, alcuni metodi fanno uso di tecniche per la stima di parametri (e.g., Maximum Likelihood, Generalized Least Squares, Bayesian inference). Altri metodi basati sul conteggio del traffico includono Combined Distribution and Assignment (CDA) [Cascetta and Postorino 2001], Bi-level Programming [Doblas and Benitez, 2005], algoritmi genetici [Kim et al. 2001], Heuristic Bi-level Programming [Lundgren and Peterson, 2008], Path Flow Estimation (PFE) [Nie et al. 2005], o reti neurali [Gong 1998]. In [Ashok and Ben-Akiva 2000] gli autori utilizzano una tecnica basata sul filtro di Kalman per aggiornare la matrice OD. La stima offline tempo-dipendente utilizza le serie temporali dei conteggi di traffico. Tipicamente, la costruzione di una matrice OD per la mobilità richiede l'installazione di device per il conteggio dei singoli veicoli (ed eventualmente la misura della loro velocità). In [Caggiani et al. 2012], viene presentato uno stimatore OD, tramite l'utilizzo di un nuovo algoritmo di simulazione. L'algoritmo proposto utilizza un modello di caricamento dinamico della rete di tipo mesoscopico in associazione con Bee Colony Optimization (BCO). Un altro approccio per la stima di matrici OD fa uso di informazioni GPS e di una IC card [Zhang and Chen 2011].

3.5 Clusterizzazione di traiettorie

Il clustering delle traiettorie degli utenti costituisce un problema importante nell'analisi dei flussi di traffico acquisiti con tecniche di data mining [Y. Zheng 2015]. Interpretare i pattern di movimento in forma aggregata è infatti un requisito fondamentale per analizzare le criticità a livello infrastrutturale eventualmente presenti nella città. Le tecniche di trajectory clustering trovano molte applicazioni in quei contesti dove marcata è l'esigenza di implementare sistemi efficienti per la

pianificazione del traffico e l'analisi dei movimenti. Le tecniche sviluppate includono per esempio T-DBSCAN, che utilizza un clustering spazio-temporale [J. Chen et al. 2014] con segmentazione delle traiettorie [W. Chen et al., 2014]. In [F. Biljecki, et al. 2013, Y. Zheng et al. 2010, Y. Zheng et al. 2008] gli autori presentano metodi di classificazione dei mezzi di trasporto utilizzando le traiettorie. Altre tecniche prevedono l'utilizzo di misure di distanza Dijkstra based Dynamic Time Warping (DTW) [D. Kumar et al. 2015] In [S. Kisilevich et al. 2010] gli autori presentano alcune applicazioni di questa tecnica. Le tecniche di clustering spazio-temporale 3D prevedono la modellazione delle traiettorie come unione di più segmenti 3D. In [C. Panagiotakis, et al. 2009] gli autori utilizzano un metodo di voting che è applicato ad ogni segmento della traiettoria, per il calcolo di un descrittore della traiettoria locale. Dall'analisi dei descrittori possono essere calcolati i percorsi più rappresentativi per la traiettoria. Un esempio di clustering incrementale è presentato in [Z. Li et al. 2010]. Un esempio di algoritmo per la semplificazione di traiettorie è presentato in [Y. Chen et al. 2009].

In [M. C. Gonzalez et al. 2008] gli autori analizzano 100,000 tracce anonimizzate di dispositivi mobile, collezionate durante un periodo di 6 mesi e dimostrano che esse hanno un alto grado di regolarità spazio-temporale, in contrasto con le traiettorie previste dai modelli random walk. In particolare ogni individuo è caratterizzato da una caratteristica distanza di viaggio percorsa, indipendente dal tempo, e da una significativa probabilità di ritornare in alcuni luoghi maggiormente frequentati. Alcune tecniche studiano l'angolo formato dalle traiettorie con i meridiani, e dal cambiamento di questo valore, insieme alla velocità ed all'accelerazione, estraggono i luoghi più significativi dalle traiettorie GPS [T. Bhattacharya et al. 2012]. In [G. Schoier and G. Borruso 2011] viene presentata una tecnica per la ricerca di cluster lungo le traiettorie, in modo da comprendere la struttura della rete stradale in termini di spazi densi. Fra gli algoritmi che si basano sull'analisi della struttura stradale si vedano ad esempio NEAT [B. Han et al. 2012, B. Han et al. 2013] e TraCLUS [J.G. Lee et al. 2008, J.G. Lee et al. 2007, J.G. Lee and 2008]; quest'ultimo genera una gerarchia di feature tramite il partizionamento delle traiettorie (partition-and-group framework). In particolare le traiettorie sono partizionate come insiemi di segmenti e poi i segmenti simili sono raggruppati in cluster.

4 Calcolo di Suggerimenti e stimoli

4.1 Produzione di raccomandazioni

La costruzione di sistema di raccomandazione usualmente include l'utilizzo di tecniche di Machine Learning e statistiche, tecniche di modellazione di sistemi, tecniche di ottimizzazione, analisi del click rate e più in generale del comportamento utente. Spesso è utile fare uso di tecniche di linguaggio naturale (NLP) per l'interpretazione delle preferenze dell'utente.

I sistemi di raccomandazione possono essere suddivisi in 4 gruppi:

- Generici, ogni utente riceve le stesse raccomandazioni;
- Demografici, ogni utente nella stessa categoria riceve le stesse raccomandazioni;
- Contestuali, le raccomandazioni dipendono soltanto dall'attività corrente;
- Persistenti, le raccomandazioni dipendono da interessi di lungo termine.

Una classificazione dei sistemi di raccomandazione può essere fatta anche tenendo conto delle tecnologie impiegate per la loro implementazione:

- Content-based systems, per esempio se un utente Netflix ha visto molti film western, allora il sistema raccomanda film dello stesso genere;

- Collaborative filtering systems, raccomandano contenuti basati su misure di similarità fra gli utenti e i contenuti stessi. I contenuti raccomandati all'utente sono quelli preferiti da utenti a lui/lei simili.

L'obiettivo di questi sistemi è di fornire la giusta raccomandazione a un utente in un dato contesto

- Ricerca degli utenti con gli stessi pattern di giudizi espressi rispetto all'utente;
- Utilizzo dei rating degli utenti del punto precedente per calcolare una predizione per l'utente corrente.

Per rating si intende un valore numerico assegnato dall'utente a un contenuto. I rating possono essere impliciti, espliciti o ibridi. I rating impliciti sono basati sull'interazione e il tempo (e.g., acquisti, click, visita pagine web). Tali metriche vengono utilizzate per generare un rating implicito. I rating espliciti possono essere numerici, possono esprimere il rapporto fra due contenuti, oppure possono contenere un'informazione semantica (e.g., tag e label). I rating ibridi sono un misto fra i precedenti approcci, prevedendo una scala numerica esplicita e la misura delle attività implicite dell'utente. Un sistema di raccomandazione deve essere efficiente nel fornire buone raccomandazioni. Le raccomandazioni devono contenere informazioni chiare e complete per l'utente in modo da permettergli di valutare il contenuto.

La produzione di raccomandazioni è solitamente ottenuta mettendo in relazione gli utenti con i contenuti. Nel caso di utilizzo di collaborative filtering (uno degli approcci più utilizzati) per esempio si utilizzano le tecniche:

- neighbourhood e latent factor models, utilizzano le relazioni fra i contenuti o fra gli utenti;
- item-item, modellizzano la preferenza di un utente per un contenuto sulla base di rating di contenuti simili da parte dello stesso utente. Modelli latent-factor, come Single Value Decomposition (SVD) mappano sia gli utenti che i contenuti in un unico spazio.

Una delle tecniche spesso utilizzata in questo contesto è quella della matrix factorization [Y. Koren et al. 2009].

Ad esempio Amazon utilizza un approccio di questo tipo:

- costruzione di una matrice item-item, con la determinazione delle relazioni fra coppie di contenuti;
- utilizzando la matrice e i dati sull'utente corrente, si deducono le sue preferenze.

Solitamente tali matrici sono sparse, perché non tutti gli utenti forniscono un rating per ogni contenuto. L'obiettivo di un sistema di raccomandazione è quello di predire gli spazi vuoti della matrice di fattorizzazione. Il popolamento della matrice di fattorizzazione può avvenire chiedendo agli utenti stessi di fornire valutazioni sui contenuti, oppure spiandone i comportamenti. Se per esempio un utente ha visto un film o ne ha letto la recensione, molto probabilmente sta esprimendo una preferenza o almeno un interesse per quel film.

Numerose tecniche di collaborative filtering sono state proposte per risolvere il problema della fornitura di raccomandazioni pertinenti, si veda ad esempio [X. Amatriain 2012, M. D. Ekstrand et al. 2011, A. Galland 2012, J. A. Konstan and John Riedl 2012, F. Ricci and P.B. Kantor 2011]. Per esempio in [H. Kim et al. 2010] gli autori propongono un metodo che fa uso di collaborative filtering per la fornitura di raccomandazioni di qualità. Il sistema fa uso di user-created tags che offre vantaggi significativi, sia per l'incremento della qualità delle raccomandazioni (nel caso di dati sparsi), che per la gestione degli utenti nuovi arrivati. In [G. Linden, et al. 2003] gli autori forniscono una comparazione fra la tecnica di collaborative filtering tradizionale, modelli a cluster, metodi search-based e il loro algoritmo, che chiamano item-to-item collaborative filtering. A

differenza dei sistemi tradizionali, il loro sistema scala indipendentemente dal numero di utenti e il numero di contenuti, producendo raccomandazioni di alta qualità in tempo reale. In [P. Melville et al. 2002] gli autori presentano un framework per combinare contenuti e interazione, utilizzando un predittore basato sui contenuti per arricchire i dati utente esistenti, e fornire suggerimenti personalizzati attraverso il collaborative filtering. Questo approccio ha prestazioni migliori di un predittore puro content-based, di un collaborative filtering tradizionale, e di un approccio naive hybrid. Alcuni approcci fanno uso di dati GPS relativi alla posizione degli utenti. Per esempio in [V. Zheng 2010] viene mostrato come sia possibile predire luoghi di interesse per gli utenti, sulla base della loro storia GPS. Gli autori applicano una tecnica di collective matrix factorization per scoprire luoghi e attività di interesse, per poi utilizzarle per raccomandare agli utenti dove possono recarsi se vogliono compiere determinate attività, e cosa possono fare se si recano in determinati luoghi. Un'altra tecnica item-based è presentata in [M. Deshpande and G. Karypis 2004]; essa prima determina le similarità fra i vari contenuti e poi li usa per determinare l'insieme dei contenuti da raccomandare. Gli autori dimostrano che gli algoritmi item-based sono fino a due ordini di grandezza più veloci dei tradizionali sistemi basati su user-neighborhood. Gli algoritmi item-based inoltre forniscono raccomandazioni qualitativamente migliori. Un altro esempio di approccio collaborative filtering al problema delle raccomandazioni è fornito in [J. Herlocker et al. 2000], dove viene presentato un modello esplicativo basato sul modello concettuale del processo di raccomandazione. Gli autori dimostrano quali sono le migliori componenti di tale modello. Tecniche per migliorare le prestazioni dei sistemi collaborative filtering sono descritte per esempio in [Herlocker et al. 1999].

4.2 Calcolo del comportamento utente

Il continuo aumento della disponibilità di tecnologie atte all'acquisizione di informazioni di posizione (GPS, reti GSM, ...) e di ambiente (temperatura, battito cardiaco, ...) permette oggi di monitorare su grande scala i movimenti delle persone, arricchendo dati spaziotemporalmente con precise informazioni di contesto. Questi dati storici abilitano nuovi scenari d'uso, tra i quali la predizione degli interessi e lo studio delle abitudini delle persone che accettano di essere monitorate, correlando l'utenza e i loro comportamenti con i luoghi che essi frequentano.

In letteratura, già prima dell'avvento dei terminali mobili, si trovano diversi studi che cercavano di modellizzare il comportamento utente in base al suo movimento all'interno di un ambiente, a quei tempi, chiuso. Maggiormente degna di nota è la tesi [Orwant, 1993], nella quale viene descritto Doppelgänger, un software di modellizzazione utente che impara a prevedere come un utente si sposterà all'interno di un numero fisso di posizioni di un edificio. Orwant, per monitorare l'utente, usava tessere magnetiche di ingresso stanza, unix login e un insieme di attività schedulate sui terminali dell'edificio. Nel sistema Doppelgänger, la possibile posizione nella quale un utente si poteva trovare era insita e definita a priori nel sistema e scelta fra un insieme finito e prefissato di luoghi.

Dall'avvento dei sistemi GPS è stato possibile liberarsi dal numero limitato di luoghi/posizioni, determinandoli volta per volta, in modo dinamico, analizzando il movimento dell'utenza in spazi aperti. Successivamente aggregando e filtrando questi dati è possibile inferire i diversi punti di interessi e infine prevedere gli spostamenti degli utenti da un luogo ad un altro. Nello studio [Ashbrook and Starner, 2003] viene presentata una soluzione ai diversi problemi di aggregazione e interpretazione di questi dati.

La sfida odierna si sposta ai nostri giorni nell'analizzare gli storici degli utenti in ordine di grandezza molto più alta rispetto al passato, elaborando in modo appropriato questi tabulati e aggregandoli efficientemente per poi essere utilizzati da un predittore real-time [Esnaashari and Meybodi, 2010].

Le tecnologie maggiormente utilizzate per modellizzare l'utente sono prettamente legate ad algoritmi di machine learning in quanto la sfida è quella di "rendere l'elaboratore capace di interpretare e reagire rispetto a un fenomeno particolarmente complesso: il comportamento utente". Problemi di User Modelling possono essere semplificati distribuendo parte dell'informazione su diversi peer di ragionamento, per poi aggregare in un secondo momento suggerimenti e predizioni. Problematiche di confidenza, reputazione, sincronizzazione e consistenza devono essere tenuti in particolare considerazione [Specht et al., 2009].

Particolarmente degni di interesse sono i modelli basati su Reti Markoviane e Baysiane, nelle quali il comportamento dell'utente viene modellizzato con un insieme finito di stati e di possibili transizioni pesate fra di essi [Ashbrook and Starner, 2002].

Infine, tecnologie basate su Reti Neurali possono contribuire a supporto di scenari non deterministici e nei quali l'analisi punto-punto dello storico dell'utente non sia possibile o prevedibile [Vintan et al, 2004].

4.3 Produzione di regole di assistenza e ingaggio

La grande mole di dati disponibile all'analisi sopra descritta viene spesso aggregata per poter essere più facilmente gestita e, in questo nuovo contesto, l'utente perde il suo classico ruolo di semplice consumatore, ma diventa esso stesso, tramite il suo terminale d'uso, colonna fondamentale del sistema che eroga il servizio. L'utente deve perciò essere fidelizzato all'uso del sistema e assistito nel momento del bisogno, possibilmente anche prevedendo quando questo bisogno si presenterà (il servizio deve andare incontro alle esigenze dell'utente).

In letteratura nuovi aspetti sociologici si sono affiancati ad argomenti prettamente tecnologici, facendo fiorire una nuova gamma interdisciplinare di studi: il comportamento dell'utente nell'uso e in risposta di un sistema informatico, il comportamento dell'utente in mobilità all'interno del suo territori sono solo alcuni dei nuovi argomenti oggetto di attenzione e formalizzazione, soprattutto in un contesto di completa connessione sensoriale [Salima and Usman, 2015].

Lo scopo del monitoraggio su grande scala in contesto cittadino viene realizzato al fine di alzare la qualità della vita dei cittadini e di proteggerli da eventuali casi eccezionali (si parte da una semplice risoluzione di congestione del traffico fino ad arrivare a pianificare un'attività di ripristino a fronte di eventi catastrofici quali alluvioni, terremoti, ...). I sistemi di ingaggio sia da parte di enti pubblici che privati sono visti come elementi critici nel creare una fidelizzazione, uno sviluppo sociale e una coesione sociale nelle comunità locali [Davies and Simon, 2013] e risultano essere molto meno costosi rispetto forzare la cittadinanza a trovare i servizi che altrimenti vengono inviati direttamente a bisogno, eventualmente cercando di innescare comportamenti virtuosi. Incoraggiando attivamente la partecipazione alla cittadinanza, il sentimento di appartenenza rende il cittadino orgoglioso del posto in cui vive e lo responsabilizza nell'uso dello spazio che condivide con gli altri [Jacobson, 2014].

Diversi progetti europei, tra i quali [s3c] cercano di formalizzare questi sistemi di ingaggio, analizzando strumenti di fidelizzazione a lungo e medio termine, e orientando la ricerca a ruoli dell'utente quale consumatore, cliente e cittadino.

L'Osservatorio Nazionale Smartcity [ONS], tramite la produzione del suo "vademecum per la città intelligente", parte dall'analisi delle diverse esperienze delle città italiane partecipanti e dei diversi contributi metodologici sottoposti (urbanistica, design, sociologia, informatica, consulenza e project management, ecc.) per poter ricavare una mappatura piuttosto articolata delle diverse tecniche che una città può effettivamente adottare per coinvolgere cittadini, esperti e stakeholder nel percorso di pianificazione della Smart City.

Le tecnologie disponibili per definire le regole di ingaggio e i sistemi per l'esecuzione automatica di tali regole sono molteplici. Anche la stessa rappresentazione risponde a differenti formati standard, rendendo i diversi strumenti particolarmente frammentati e molto orientati al caso d'uso nel quale

vengono calati e al contesto di interesse (misure GPS nel caso di scenari di mobilità, condizioni meteo in scenari di resilienza, sensori di misura in scenari di efficienza energetica).

Il cappello tecnologico può essere definito quale Sistemi ad Eventi-Condizione-Azione, nei quali il dominio applicativo viene rappresentato con una serie di dati osservati (eventi, fatti) che abilitano l'esecuzione di un insieme di regole definite a priori e composte da azioni che vengono scatenate a fronte dell'avverarsi di particolari condizioni. Tra i più diffusi si citano gli algoritmi RETE, che permettono di scomporre le regole di ingaggio in sotto-regole con l'obiettivo di ritornare risposte in simil real-time anche a fronte di un grande numero di fatti.

Tra i principali standard di rappresentazione delle regole si citano: [RIF], del gruppo W3C, che propone di formalizzare regole di interoperabilità fra i diversi sistemi a regole, soprattutto a motori orientati al Web (piuttosto che formalizzare uno strumento valido per tutti i sistemi); [OWL2RF] che permette la realizzazione di algoritmi di ragionamento usando regole che estendono tecnologie atte ai database, ma che lavorano direttamente su triple RDF; [RULEML] che rappresenta un formato di regole realizzate dall'omonimo strumento e che considera scenari di Delibera, Reazione e Consumo (orientate al mondo giuridico).

Per poter valutare un motore a regole è necessario analizzare quali strumenti fornisca agli utenti e agli sviluppatori, per poter definire le regole e per poter provare il sistema a regole, le performance degli algoritmi di esecuzioni di regole, il workflow per l'editing delle regole, l'integrazione con software già esistente e il supporto alla community. Saranno da prendere in considerazione: Apache jena, jboss drools, jess, ORACLE Engagement Engine e IBM JRULES.

5 Calcolo di Sentiment Analysis su dati testuali

5.1 Distribuzioni e NLP

La rapida crescita delle risorse digitali sul Web ha aperto nuove sfide nello sviluppo di soluzioni per l'estrazione automatica di informazioni. In particolare, la sempre crescente produzione in rete di dati testuali non strutturati in linguaggio naturale rende auspicabile un ausilio da parte di strumenti software per la comprensione automatica del testo. A tale scopo, vengono tipicamente applicate tecniche di Natural Language Processing (NLP), Semantic Computing, modelli statistici, machine learning e soluzioni di clustering. In particolare, i modelli NLP più utilizzati in letteratura sfruttano algoritmi di parsing grammaticali e sintattici, come ad esempio il Part-of-Speech (POS) tagging per l'estrazione di termini sintatticamente annotati a partire da dati testuali non strutturati.

Successivamente, per migliorare le capacità di comprensione automatica del testo, vengono generalmente applicate tecniche di riconoscimento e disambiguazione (in casi di parole polisemiche o differenti contesti), come Named Entity Recognition (NER) [Nadeau, and Sekine, 2007] e Word Sense Disambiguation (WSD) [Navigli, 2009], [Pal and Saha, 2015]. Altri metodi basati su Machine Learning trattano il problema dell'estrazione automatica di keyword come un problema di supervised learning, utilizzando specifici dataset di training, algoritmi naive Bayes [Zavrel et al., 2000], least square support vector machines (LS-SVM) [Wu et al., 2007] ecc. Altri approcci confidano nell'uso di risorse esterne come liste annotate, gazetteer, blacklist per filtrare eventuali keyword indesiderate [Liu et al., 2009a], selezionando solo alcuni tag come potenziali keyword [Liu et al., 2009b]. In aggiunta, vengono utilizzati metodi che stimano la rappresentazione della frase estraendo un grafo orientato (parsing sintattico) che modella e descrive le relazioni e le dipendenze tra i componenti del periodo analizzato, avvalendosi anche di funzioni di ranking per stimare la rilevanza dei nodi, e quindi l'importanza delle keyword ad essi associate [Hasan and Ng, 2014]. Infine, metodi statistici (in particolare tecniche di clustering) sono utilizzati per la stima e l'estrazione del contesto, ad esempio in strumenti di topic-extraction e text summarization [Ginevra et al., 2009], [Liu et al., 2010]. Data la sempre crescente mole di dati da analizzare, sono sempre più

utilizzate le architetture parallele e distribuite per l'implementazione di modelli e algoritmi di NLP in ambito Big Data [Nesi et al., 2015].

5.2 Calcolo della sentiment analysis

Con il termine Sentiment Analysis si intende, in letteratura, l'insieme di tecniche e metodi atti ad individuare l'orientamento di opinioni espresse riguardo una specifica entità, argomento ecc., estratto in maniera automatica da dati testuali (in particolare testo non strutturato). La Sentiment Analysis si basa tipicamente su tecniche e soluzioni di text mining, NLP, metodi statistici e di intelligenza artificiale.

Comunemente vengono presentati diversi approcci di classificazione dei task inerenti alla SA, sulla base della differenziazione del tipo di task, o relativamente ai differenti livelli in cui viene eseguita la classificazione e/o il processing. Nel primo caso, si differenziano i seguenti tipi di task:

- Stima della polarità SO (Subjective or Objective): questo task si occupa di identificare se un testo è soggettivo (se cioè esprime un'opinione o uno stato d'animo) o oggettivo (se cioè ha una natura fattuale).
- Stima della polarità PN (Positive or Negative) nel testo: questo task rappresenta un tipo di classificazione binaria, attraverso la quale si cerca di stimare se un testo soggettivo esprime, nel suo complesso, un parere positivo o negativo su un certo argomento (che può essere anch'esso stimato in maniera automatica con specifici algoritmi NLP di content extraction o summarization, oppure definito a priori).
- Stima della polarità PNO (Positive – Negative – Objective): simile al task descritto precedentemente; in questo caso però la classificazione prevede tre classi, aggiungendo alle polarità positiva e negativa quella neutra.
- Stima dell'intensità della polarità PN: in questo task si cerca di stimare, oltre alla polarità positiva o negativa, anche la sua intensità (per esempio Weakly Positive, Mildly Positive or Strongly Positive ecc. o mediante altre scale di valori), attraverso punteggi o coefficienti opportunamente definiti ed implementati.

Nel caso di classificazione basata sul livello di elaborazione, si possono distinguere le seguenti categorie: Word Level, Phrase Level, Sentence Level e Document Level. A livello di singola parola (Word Level), che è attualmente ancora il più usato in letteratura, l'input testuale viene spesso rappresentato come una *bag-of-words* [Paltoglou and Thelwall, 2013] da cui si estraggono keywords (in maniera generalmente decontestualizza) che vengono poi annotate semanticamente (mediante risorse esterne già annotate, ad esempio dizionari semantici come Wordnet [Miller et al., 1990]) o attraverso altre risorse esterne appositamente progettate per fornire un'interpretazione della polarità (PN-PNO, ovvero positiva, negativa e/o neutra) delle keyword estratte (come ad esempio i dataset SentiWordnet [Esuli and Sebastiani, 2006] e Wordnet Affect [Valitutti, 2004]). La mancanza di informazioni contestuali rende tuttavia le stime basate su questo approccio non molto precise, come evidenziato in [Pang et al., 2002] e [Turney, 2002].

A livello di frase (Phrase Level) sono di solito utilizzate tecniche di pattern matching, e recentemente si è affermato l'approccio che prevede l'analisi degli *N*-grams. Gli *N*-grams sono simili alle *bag-of-words*, con la differenza che, in questo caso, il testo è splittato non in parole singole ma in array di parole di uguale lunghezza, definite dal parametro *N* [Pang and Lee, 2008]. Il valore di *N* dipende dalla natura del testo in ingresso e dal tipo di problema da gestire. Comunemente vengono utilizzati 2-grams, 3-grams e 4-grams. In questo modo, a discapito di un maggior costo computazionale, si possono gestire le dipendenze tra gli elementi della frase, come qualificatori o modificatori, negazioni, doppie negazioni, uso di avverbi ecc. il Sentence Level è simile al Phrase Level, differendo solo per il fatto che una Sentence viene definita come un'entità che può avere opinioni o sentimenti differenti al suo interno, mentre una Phrase generalmente viene

intesa come un'entità che esprime un singolo stato (sentimenti e/o opinioni coerenti tra di loro). A livello di documento (Document Level), l'approccio più comunemente utilizzato è quello di trovare la polarità di sentiment delle singole frasi e combinarle insieme per stimare la polarità del documento (impiegando specifici schemi o modelli per assegnare pesi alle polarità delle varie frasi) [Psomakelis et al., 2014].

Un tentativo di avanzare lo stato dell'arte nel campo della SA, trovando soluzioni anche diverse rispetto all'esclusivo utilizzo di metodi e tecniche di NLP (considerati non sufficienti per comprendere in maniera automatica le numerose regole implicite ed irregolarità proprie di ogni lingua), è stato proposto con l'affermarsi del Sentic Computing. Il Sentic Computing è un approccio multi-disciplinare alla SA che sfrutta sia le tecnologie dell'informazione, sia principi delle scienze sociali per migliorare il riconoscimento e l'interpretazione delle espressioni di sentimento e opinione. In pratica, il Sentic Computing unisce tecniche di Semantic Computing e Artificial Intelligence per aggiungere reasoning e inferenza ai risultati ottenuti tramite l'elaborazione del testo tramite soluzioni di NLP, nonché paradigmi ed elementi di sociologia ed etica per modellare ed interpretare dinamiche e relazioni social (ad esempio l'influenza all'interno di una social network, una community ecc.) [Poria et al., 2014].

Basandosi su approcci di Semantic Computing, e quindi facendo riferimento a risorse semantiche esterne e di terze parti, un limite di questo tipo di approccio è il fatto di essere dipendente dalla ricchezza e dalla completezza della knowledge base di riferimento.

6 Bibliografia

- [5T] 5T S.r.l., Infoblu S.p.A, CSST S.p.A., Tema.mobility, «Specifiche interfaccia sistema di acquisizione dati Floating Car Data (FCD),» Provincia di Firenze, Firenze.
- [A. Costanzo 2013] A. Costanzo, Using GPS data to monitor road traffic flows in a metropolitan area: methodology and case study, 2013
- [A. Galland 2012] A. Galland. Recommender systems. Technical report, INRIA-Saclay, 2012.
- [Abusalah et al., 2005] Abusalah M, Tait J, Oakes MP (2005) Literature review of cross language information retrieval. In: WEC (2)'05, pp 175--177
- [AIPCR] Associazione mondiale della strada - AIPCR, «Monitoraggio delle caratteristiche e delle azioni del traffico veicolare per il progetto e la manutenzione delle pavimentazioni stradali,» Firenze University Press, Firenze, Ottobre 2006.
- [Akhlaghian et al., 2010] Akhlaghian F, Arzanian B, Moradi P (2010) A personalized search engine using ontology-based fuzzy concept networks. In: Proceedings of the 2010 International Conference on Data Storage and Data Engineering, IEEE Computer Society, Washington, DC, USA, DSDE '10, pp 137--141, doi{10.1109/DSDE.2010.30}
- [Ali et al., 2015] Muhammad Intizar Ali, Feng Gao, Alessandra Mileo, "CityBench: A Configurable Benchmark to Evaluate RSP Engines Using Smart City Datasets", 14th Int. Semantic Web Conference, Bethlehem, PA, USA, October 11-15, LNCS, 2015
- [Anthopoulos et al., 2014] Anthopoulos, L.; Fitsilis, P., "Exploring architectural and organizational features in smart cities," in Advanced Communication Technology (ICACT), 2014 16th International Conference on , vol., no., pp.190-195, 16-19 Feb. 2014, doi: 10.1109/ICACT.2014.6778947
- [Ashbrook and Starner, 2002]: Ashbrook, D., and Starner, T. "Learning Significant Locations and Predicting User Movement with GPS". Proceedings at Sixth International Symposium on Wearable Computers, doi:10.1109/iswc.2002.1167224.
- [Ashbrook and Starner, 2003]: Ashbrook, D. and Starner, T. "Using GPS to Learn Significant Locations and Predict Movement Across Multiple User". Personal and Ubiquitous Computing archive, Volume 7 Issue 5, October 2003, pp 275-286, doi:10.1007/s00779-003-0240-0.
- [Ashok and Ben-Akiva 2000] Ashok, K. and Ben-Akiva, M. E. (2000) Alternative approaches for real-time estimation and prediction of time dependent origin-destination flows, Transportation Science 34: 21-36
- [Aslam and Savell, 2003] Aslam JA, Savell R (2003) On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, SIGIR '03, pp 361--362, doi{10.1145/860435.860501}.
- [Attar and Fraenkel, 1977] Attar R, Fraenkel AS (1977) Local feedback in full-text retrieval systems. J ACM 24(3):397--417
- [B. Han et al. 2012] B. Han, L. Liu, E. Omiecinski, NEAT: Road Network Aware Trajectory Clustering, 2012
- [B. Han et al. 2013] B. Han, L. Liu, E. Omiecinski, Road-Network Aware Trajectory Clustering: Integrating Locality, Flow, and Density, 2013
- [Balakrishna, 2012] Balakrishna C., "Enabling Technologies for Smart City Services and Applications," Next Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference on , vol., no., pp.223,227, 12-14 Sept. 2012
- [Ballesteros and Croft, 1997] Ballesteros L, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. SIGIR Forum 31(SI):84--91
- [Ballesteros and Croft, 1998] Ballesteros L, Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, SIGIR '98, pp 64--71
- [Barbieri et al., 2009] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus. "C-sparql: Sparql for continuous querying". In Proc. of WWW, pages 1061–1062. ACM, 2009.
- [Batarseh, Gonzalez, 2013] Batarseh, Feras A., and Avelino J. Gonzalez. "Incremental lifecycle validation of knowledge-based systems through CommonKADS." IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol.43, n.3, 2013, pp.643-654.

- [Belkin and Croft, 1987] Belkin NJ, Croft WB (1987) Retrieval techniques. In: Williams ME (ed) Annual review of information science and technology, vol. 22, Elsevier Science Inc., New York, NY, USA, pp 109--145,
- [Bellini and Nesi, 2013] Bellini P, Nesi P (2013) A linked open data service for performing arts. In: ECLAP, 2nd International Conference on Information Technologies for Performing Arts, Media Access and Entertainment, Springer Verlag LNCS
- [Bellini et al., 2013a] P. Bellini, M. Di Claudio, P. Nesi, N. Rauch, "Tassonomy and Review of Big Data Solutions Navigation", as Chapter 2 in "Big Data Computing", Ed. Rajendra Akerkar, Western Norway Research Institute, Norway, Chapman and Hall/CRC press, ISBN 978-1-46-657837-1, 2013
- [Bellini et al., 2014] P. Bellini, M. Benigni, R. Billero, P. Nesi, N. Rauch, "Km4City Ontology Building vs Data Harvesting and Cleaning for Smart-city Services", International Journal of Visual Language and Computing, Elsevier, 2014
- [Bellini et al., 2015] P. Bellini, I. Bruno, P. Nesi, N. Rauch, "Graph Databases Methodology and Tool Supporting Index/Store Versioning", publication on JVLIC, Journal of Visual Languages and Computing, Elsevier, 2015 doi:10.1016/j.jvlc.2015.10.018 <http://www.sciencedirect.com/science/article/pii/S1045926X15000750>
- [Berkovich and Inayatullah, 2004] Berkovich S, Inayatullah M (2004) A fuzzy find matching tool for image text analysis. In: Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on, pp 101 -- 105, doi{10.1109/AIPR.2004.2}
- [Berners-Lee, 2006] T. Berners-Lee, "Linked Data", <http://www.w3.org/DesignIssues/LinkedData.html>, 2006.
- [Bizer and Schultz, 2009] C. Bizer, A. Schultz. "The Berlin SPARQL Benchmark". International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, Pages 1-24, 2009
- [Bizer et al., 2011] Bizer, C., Jentzsch, A., Cyganiak, R.: State of the LOD cloud. <http://lod-cloud.net/state/> Retrieved July 5, 2014.
- [Brin and Page, 1998] Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1-7):107--117
- [C. Panagiotakis, et al. 2009] C. Panagiotakis, N. Pelekis, I. Kopanakis, Trajectory Voting and Classification based on Spatiotemporal Similarity in Moving Object Databases, 2009
- [Caggiani et al. 2012] L. Caggiani, M. Dell'Orco, M. Marinelli, M. Ottomanelli, A metaheuristic dynamic traffic assignment model for O-D matrix estimation using aggregate data, 2012
- [Calbimonte et al., 2010] J.-P. Calbimonte, O. Corcho, and A. J. G. Gray. "Enabling ontology-based access to streaming data sources" In ISWC, pp. 96–111, 2010
- [Callan et al., 1999] Callan J, Connell M, Du A (1999) Automatic discovery of language models for text databases. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of data, ACM, New York, NY, USA, SIGMOD '99, pp 479--490, doi{10.1145/304182.304224}
- [Caragliu A. 2009] Caragliu A., Del Bo C., Nijkamp P., Smart cities in Europe, 3rd Central European Conference in Regional Science CERS, 2009
- [Cascetta and Postorino 2001] Cascetta, E. and Postorino, M. N. (2001) Fixed point approaches to the estimation of O/D matrices using traffic counts on congested networks, Transportation Science 35: 134-147
- [Cecchini et al., 2008] Cecchini RL, Lorenzetti CM, Maguitman AG, Brignole NB (2008) Using genetic algorithms to evolve a population of topical queries. Inf Process Manage 44(6):1863--1878
- [Chen et al. 2003] Chen HH, Kuo JJ, Su TC (2003) Clustering and visualization in a multi-lingual multi-document summarization system. In: Proceedings of the 25th European conference on IR research, Springer-Verlag, Berlin, Heidelberg, ECIR'03, pp 266--280
- [Chourabi, et al., 2012] Chourabi, H.; Taewoo Nam; Walker, S.; Gil-Garcia, J.R.; Mellouli, S.; Nahon, Karine; Pardo, T.A.; Scholl, Hans Jochen, "Understanding Smart Cities: An Integrative Framework," System Science (HICSS), 2012 45th Hawaii International Conference on , vol., no., pp.2289,2297, 4-7 Jan. 2012
- [D. Agarwal 2011] Deepak Agarwal, Offline components: Collaborative filtering in cold-start situations. Technical report, YAHOO, 2011.
- [D. Kumar et al. 2015] D. Kumar, S. Rajasegarar, M. Palaniswami, X. Wang, C. Leckie, A Scalable Framework for Clustering Vehicle Trajectories in a Dense Road Network, 2015
- [Danalet, A. 2014] Danalet, Antonin, Michel Bierlaire, and Bilal Farooq. "Estimating Pedestrian

- Destinations using Traces from WiFi Infrastructures.” Pedestrian and Evacuation Dynamics 2012. Springer International Publishing, 2014. 1341-1352.
- [Davies and Simon, 2013]: Davies, A and Simon, J, “Engaging Citizens in Social Innovation”. Published by youngfoundation.org, 2013.
- [Doblas and Benitez, 2005] Doblas, J. and Benitez, F. G. (2005) An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix, Transportation Research, Part B: Methodological 39: 565-591
- [Duan et al., 2011] S. Duan, A. Kementsietsidis, K. Srinivas, and O. Udrea. “Apples and oranges: a comparison of RDF benchmarks and real RDF datasets”. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11). ACM, New York, NY, USA, 145-156, 2011
- [Erling et al., 2015] O. Erling, A. Averbuch, J.L. LarribaPey, Hassan Chafi, Andrey Gubichev, Arnau Prat, Minh-Duc Pham, Peter Boncz, The LDBC Social Network Benchmark: Interactive Workload. Proceedings of SIGMOD 2015, Melbourne.
- [Esnaashari and Meybodi, 2010]: Esnaashari and Meybodi. “Data aggregation in sensor networks using learning automata”. Wireless Networks 16(3):687-699, April 2010, doi: 10.1007/s11276-009-0162-5.
- [Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F., “SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining”, In Proc. of the International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
- [European Commission 2014] Report on the state of play of the outcome of the work of the High Level Group, European Commission, October 2014
- [European Parliament 2014] Mapping Smart Cities in the EU, DIRECTORATE GENERAL FOR INTERNAL POLICIES. POLICY DEPARTMENT A: ECONOMIC AND SCIENTIFIC POLICY, European Parliament 2014
- [F. Biljecki, et al. 2013] F. Biljecki, H. Ledoux, P. van Oosterom, Transportation mode-based segmentation and classification of movement trajectories, 2013
- [F. Ricci and P.B. Kantor 2011] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor. Recommender Systems Handbook. Springer, 2011.
- [Fellbaum-WordNet-1998] Fellbaum C (ed) (1998) WordNet An Electronic Lexical Database. The MIT Press, Cambridge, MA ; London
- [G. Linden, et al. 2003] Greg Linden, Brent Smith, and Jeremy York, Recommendations Item-to-Item Collaborative Filtering, 2003
- [G. Schoier and G. Borruso 2011] G. Schoier, G. Borruso, Individual Movements and Geographical Data Mining. Clustering Algorithms for Highlighting Hotspots in Personal Navigation Routes, 2011
- [Gamallo et al., 2002] Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G., and Delima, V., “Mapping syntactic dependencies onto semantic relations”, in Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, 2002.
- [Garbis, et al., 2013] G. Garbis, K. Kyzirakos, M. Koubarakis. “Geographica: A Benchmark for Geospatial RDF Stores”. In the 12th International Semantic Web Conference (ISWC 2013). Sydney, Australia, October 21-25, 2013
- [GeoGeoSPARQL] Open Geospatial Consortium, “GeoGeoSPARQL - A Geographic Query Language for RDF Data”, Sept. 10 2012 <http://www.opengeospatial.org/standards/geosparql>
- [Ginevra et al., 2009] Grineva, M., Grinev M. and Lizorkin, D., “Extracting key terms from noisy and multitheme documents”, in Proc. of the 18th Int. Conf. on World Wide Web, pp. 661-670, 2009.
- [Gong 1998] Gong, Z. (1998) Estimating the urban o-d matrix: a neural network approach, European Journal of Operational Research 106: 108-115
- [Grosan and Abraham, 2011] Grosan, C., and A. Abraham. Intelligent Systems: A Modern Approach, Springer-Verlag, Berlin, 2011.
- [Guo et al., 2005] Y. Guo, Z. Pan, and J. Heflin. “Lubm: A benchmark for owl knowledge base systems”. J. Web Semantics, 3(2-3):158-182, 2005.
- [Gupta et al., 2003] Gupta S, Kaiser G, Neistadt D, Grimm P (2003) Dom-based content extraction of html documents. In: Proceedings of the 12th international conference on World Wide Web, ACM, New York, NY, USA, WWW '03, pp 207--214, doi{10.1145/775152.775182}
- [H. Kim et al. 2010] Heung-Nam Kim, Ae-Ttie Ji, Inay Ha, Geun-Sik Jo, Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation, 2010

- [H. Yoon et al. 2010] H. Yoon, Y. Zheng, X. Xie, W. Woo, Smart Itinerary Recommendation based on User-Generated GPS Trajectories, 2010
- [Han et al., 2003] Han H, Giles CL, Manavoglu E, Zha H, Zhang Z, Fox EA (2003) Automatic document metadata extraction using support vector machines. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, IEEE Computer Society, Washington, DC, USA, JCDL '03, pp 37—48.
- [Hartig et al., 2009] O. Hartig, C. Bizer, J.-C. Freytag. 2009. Executing SPARQL Queries over the Web of Linked Data. In Proc. of ISWC '09, Springer, pp.293-309.
- [Hasan and Ng, 2014] Hasan, K. S. and Ng, V., “Automatic Keyphrase Extraction: A Survey of the State of the Art”, in Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1262-1273, 2014.
- [Herlocker et al. 1999] Herlocker, J.; Konstan, J.; Borchers, A.; and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering
- [Hull, 1996] Hull DA, Grefenstette G (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, SIGIR '96, pp 49--57
- [J. A. Konstan and John Riedl 2012] Joseph A. Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, April 2012.
- [J. Chen et al. 2014] J. Chen, T. Hu, P. Zhang, W. Shi, TRAJECTORY CLUSTERING FOR PEOPLE’S MOVEMENT PATTERN BASED ON CROWD SOURING DATA, 2014
- [J. Herlocker et al. 2000] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl, Explaining Collaborative Filtering Recommendations, 2000
- [J. Shang et al. 2014] J. Shang, Y. Zheng, W. Tong, E. Chang, Y. Yu, Inferring Gas Consumption and Pollution Emissions of Vehicles throughout a City, 2014
- [J. Yuan 2012] J. Yuan, Y. Zheng, X. Xie, Discovering Regions of Different Functions in a City Using Human Mobility and POIs, 2012
- [J. Yuan et al. 2013] J. Yuan Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, Y. Huang, T-Drive: Driving Directions Based on Taxi Trajectories, 2013
- [J. Yuan et al. 2013] J. Yuan, Y. Zheng, X. Xie, G. Sun, T-Drive: Enhancing Driving Directions with Taxi Drivers’ Intelligence, 2013
- [J.G. Lee and 2008] J.G. Lee, J. Han, X. Li, Trajectory Outlier Detection: A Partition-and-Detect Framework, 2008[Giffinger, R. et al. 2007] Giffinger, Rudolf; Christian Fertner; Hans Kramar; Robert Kalasek; Nataa Pichler-Milanovic; Evert Meijers, Smart cities Ranking of European medium-sized cities, 2007 Available at: <http://www.smartcities.eu/>
- [J.G. Lee et al. 2007], Trajectory Clustering: A Partition-and-Group Framework, 2007
- [J.G. Lee et al. 2008] J.G. Lee, J. Han, X. Li, H. Gonzalez, TraClass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering, 2008
- [Jacobson, 2014]: Bob Jacobson. “Getting citizens involved: how cities can use smart technology and social media for improvement” published on <http://www.marshalls.co.uk>, 2014.
- [Kebus et al., 2011] Kehua Su; Jie Li; Hongbo Fu, "Smart city and the applications," *Electronics, Communications and Control (ICECC)*, 2011 International Conference on , vol., no., pp.1028,1031, 9-11 Sept. 2011.
- [Khan et al., 2013] Zaheer Khan, Ashiq Anjum, and Saad Liaquat Kiani. 2013. Cloud Based Big Data Analytics for Smart Future Cities. In Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing (UCC '13). IEEE Computer Society, Washington, DC, USA, 381-386. DOI=<http://dx.doi.org/10.1109/UCC.2013.77>
- [Kim et al. 2001] Kim, H., Beak, S. and Lim, Y. (2001) Origin-destination matrices estimated with a genetic algorithm from link traffic counts, *Transportation Research Record* 1771, Transportation Research Board, Washington, D.C.:156-163
- [Klein et al., 2002] M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov. “Ontology versioning and change detection on the web”. In *Procs of the 13th European Conf. on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 197–212. Springer, 2002.

- [Klyne and Carrol, 2004] G. Klyne, J. Carroll, “Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation”, 2004
- [Kurstien and Eibl, 2011] Kursten J, Eibl M (2011) A large-scale system evaluation on component-level. In: *Advances in Information Retrieval*, vol 6611, Springer Berlin / Heidelberg, pp 679--682
- [L. X. Pang 2013] L. X. Pang, S. Chawlaa, W. Liu, Y. Zheng, *On Detection of Emerging Anomalous Traffic Patterns Using GPS Data*, 2013
- [L.W. Wei et al. 2012] L.W. Wei, Y. Zheng, W.C. Peng, *Constructing Popular Routes from Uncertain Trajectories*, 2012
- [Lai et al., 2011] Lai LF, Wu CC, Lin PY, Huang LT (2011) Developing a fuzzy search engine based on fuzzy ontology and semantic search. In: *Fuzzy Systems (FUZZ)*, 2011 IEEE International Conference on, pp 2684 --2689, doi{10.1109/FUZZY.2011.6007378}
- [Langville and Meyer, 2012] Langville, A. N. and Meyer, C. D., “Google's PageRank and Beyond: The Science of Search Engine Rankings”, Princeton University Press, 2012.
- [Leduc 2008] G. Leduc, «Road Traffic Data: Collection Methods and Applications,» Institute for Prospective Technological Studies - Joint Research Centre, Luxembourg, 2008.
- [Lee and Yang, 2000] Lee CH, Yang HC (2000) Towards multilingual information discovery through a som based text mining approach. In: *PRICAI Workshop on Text and Web Mining*, pp 80--87
- [Le-Phuoc et al., 2011] D. Le-Phuoc, M. Dao-Tran, J. X. Parreira, and M. Hauswirth. “A native and adaptive approach for unified processing of linked streams and linked data”, In *ISWC*, pp. 370–388, 2011.
- [Le-Phuoc et al., 2012] Le-Phuoc, D., Dao-Tran, M., Pham, M. “Linked Stream Data Processing Engines: Facts and Figures” In *International Semantic Web Conference (ISWC 2012)*. Volume 1380, Boston, USA, Springer (2012) 300–312
- [Li et al., 2010] Li J, Wang Q, Wang C, Cao N, Ren K, Lou W (2010) Fuzzy keyword search over encrypted data in cloud computing. In: *INFOCOM, 2010 Proceedings IEEE*, pp 1--5, doi{10.1109/INFCOM.2010.5462196}.
- [Liu et al., 2006] Liu Y, Mitra P, Giles CL, Bai K (2006) Automatic extraction of table metadata from digital documents. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, New York, NY, USA, JCDL '06, pp 339--340, doi{10.1145/1141753.1141835}
- [Liu et al., 2009a] Z. Liu, P. Li, Y. Zheng and M. Sun, “Clustering to find exemplar terms for keyphrase extraction”, in *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, pp. 257–266, 2009.
- [Liu et al., 2009b] Liu, F., Pennell, D., and Liu, Y., “Unsupervised approaches for automatic keyword extraction using meeting transcripts”, in *Proc. of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 620-628. 2009.
- [Liu et al., 2010] Liu, Z., Huang, W., Zheng, Y. and Sun, M., “Automatic keyphrase extraction via topic decomposition”, in *Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing*, pp. 366-376, 2010.
- [Lopez, 1999] M. Fernandez Lopez, “Overview of Methodologies for Building Ontologies”, in: *IJCAI99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, Stockholm, 1999.
- [Lundgren and Peterson, 2008] Lundgren, J. T. and Peterson, A. (2008) A heuristic for the bilevel origin-destination matrix estimation problem, *Transportation Research, Part B: Methodological* 42: 339-354
- [M. C. Gonzalez et al. 2008] M. C. Gonzalez, C. A. Hidalgo, A.L. Barabasi, Understanding individual human mobility patterns, *Nature* Vol 453, 5 June 2008
- [M. D. Ekstrand et al. 2011] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, February 2011.
- [M. Deshpande and G. Karypis 2004] Mukund Deshpande, George Karypis, *Item-Based Top-N Recommendation Algorithms*, 2004
- [M.G. Messina et al. 2011] M.G. Messina, G. Valenti, F. Carapellucci G. Fusco, C. Colombaroni, [A. Gemma 2011] A. Gemma, G. Ciccarelli, S. Lo Sardo, *Sistema di monitoraggio e previsione della mobilità veicolare per l'integrazione tra la rete della illuminazione pubblica e la rete della mobilità*, Report Ricerca di Sistema Elettrico Accordo di Programma Ministero dello Sviluppo Economico – ENEA 2011

- [Manning et al., 2008] Manning CD, Raghavan P, Schtze H (2008), Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA
- [Matsumoto and Hung, 2010] Matsumoto T, Hung E (2010) Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation. In: Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, pp 1 --8, doi{10.1109/FUZZY.2010.5584771}.
- [McCarley, 1999] McCarley JS (1999) Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '99, pp 208--214
- [Messina et al., 2011] Messina, M. G., et al. "Sistema di monitoraggio e previsione della mobilità veicolare per l'integrazione tra la rete della illuminazione pubblica e la rete della mobilità." Ricerca_sviluppo/documenti/ricerca-di-sistema-elettrico/smart-city/rds-330. pdf (2011).
- [Milette 2012] L. Milette, Improving the Knowledge-Based Expert System Lifecycle, UNF report, 2012.
- [Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J., "Introduction to Wordnet: an On-Line Lexical Database", International Journal of Lexicography, Vol. 3(4), pp. 235-244, 1990.
- [Ministero] Ministero dei Lavori Pubblici (Ispettorato Generale per la Circolazione e la Sicurezza Stradale), CISUT di Reggio Calabria, «Sistemi di monitoraggio del traffico - Linee guida per la progettazione,» spettorato Generale per la Circolazione e la Sicurezza Stradale.
- [Mulligan and Olsson, 2013] Mulligan, C.E.A.; Olsson, M., "Architectural implications of smart city business models: an evolutionary perspective," Communications Magazine, IEEE , vol.51, no.6, pp.80,85, June 2013
- [Nadeau, and Sekine, 2007] Nadeau, D. and Sekine, S., "A survey of named entity recognition and classification", Journal of Lingvisticae Investigationes, vol. 30(1), pp. 3-26, 2007.
- [Navigli, 2009] Navigli, R., "Word sense disambiguation: A survey", ACM Comput. Surv., Vol. 41, ACM, New York, NY, USA, pp. 10:1–10:69, 2009.
- [Nesi et al., 2015] P. Nesi, G. Pantaleo, G. Sanesi, "A Hadoop Based Platform for Natural Language Processing of Web Pages and Documents", in Journal of Visual Languages & Computing, Vol. 31 (Special Issue on DMS2015), pp. 130-138, 2015.
- [Nie et al. 2005] Nie, Y., Zhang, H. M. and Recker, W. W. (2005) Inferring origin-destination trip matrices with a decoupled GLS path flow estimator, Transportation Research, Part B: Methodological 39: 497-518
- [Noy and McGuinness, 2001] Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." Technical Report SMI-2001-0880, Standford Medical Informatics. 2001.
- [Noy and Musen, 2004] N. F. Noy and M. A. Musen. "Ontology versioning in an ontology management framework". IEEE Intelligent Systems, 19(4):6–13, 2004.
- [ONS]: <http://osservatoriosmartcity.it>
- [Orwant, 1993]: Orwant, Jon. "Doppelgänger goes to school: Machine learning for user modelling". Thesis at Massachusetts Institute of Technology, Program in Media Arts & Sciences, 1993.
- [OWLRF]: W3C Consortium, "OWL 2 Web Ontology Language Profiles (Second Edition)", W3C Recommendation 11 December 2012, <https://www.w3.org/TR/owl2-profiles>
- [P. Melville et al. 2002] Prem Melville and Raymond J. Mooney and Ramadass Nagarajan, Content-Boosted Collaborative Filtering for Improved Recommendations, 2002
- [Pal and Saha, 2015] Pal, A.R. and Saha, D., "Word Sense Disambiguation: A Survey", International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3. 2015.
- [Palara, 2013] M. Palara, «Una rassegna di applicazioni per il monitoraggio veicolare,» Alma Mater Studiorum - Università di Bologna, Bologna, 2013.
- [Paltoglou and Thelwall, 2013] Paltoglou, G. and Thelwall, M., "More than Bag-of-Words: Sentence based Document Representation for Sentiment Analysis", in Proc. of International Conference on Recent Advances in Natural Language Processing, pp. 546–552, Hissar, Bulgaria, 7-13 September 2013.
- [Pang and Lee, 2008] B. Pang, L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, Vol.2,(1–2), pp. 1–135, 2008.

- [Pang et al., 2002] B. Pang, L. Lee, S. Vaithyanathan, “Thumbs up? Sentiment Classification Using Machine Learning Techniques”, in Proc. Of the ACL-02 Conference on Empirical methods in Natural Language Processing (EMNLP '02), Vol.10, pp 79-86, Philadelphia, Pennsylvania, 2002.
- [Pathak et al., 2000a] Pathak P, Gordon M, Fan W (2000) Effective information retrieval using genetic algorithms based matching function adaptation. In: in Proceedings of the 33rd Hawaii International Conference on System Science.
- [Patil, P. H. 2015] Patil, P. H., and A. A. Kokil. ”WiFiPi-Tracking at mass events.” Pervasive Computing (ICPC), 2015 International Conference on. IEEE, 2015.
- [Perez-Aguera, 2007] Perez-Aguera JR (2007) Using genetic algorithms for query reformulation. In: Proceedings of the 1st BCS IRSG Conference on Future Directions in Information Access, British Computer Society, Swinton, UK, UK, FDIA'07, pp 15--15
- [Picchi and Peters, 1998] Picchi E, Peters C (1998) Cross-language information retrieval: A system for comparable corpus querying. In: Grefenstette G (ed) Cross-Language Information Retrieval, The Springer International Series on Information Retrieval, vol~2, Springer US, pp 81—92.
- [Poria et al., 2014] S. Poria, “Sentic Patterns: Dependency-Based Rules for Concept-Level Sentiment Analysis”, Knowledge-Based Systems, Vol. 69, pp. 45-63, 2014.
- [Psomakelis et al., 2014] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, T. Varvarigou, “Comparing methods for Twitter Sentiment Analysis”, in Proc. Of the Int. Conf. on Knowledge Discovery and Information Retrieval, January 2014
- [Q. Li et al. 2008] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W.Y. Ma, Mining User Similarity Based on Location History, 2008
- [Radwan et al., 2006] Radwan AAA, Latef BAA, Mgeid A, Ali A, Sadek OA (2006) Using genetic algorithm to improve information retrieval systems
- [RIF]: W3C Consortium, “RIF Overview (Second Edition)”, W3C Working Group Note 5 February 2013, <https://www.w3.org/TR/rif-overview>
- [RULEML]: <http://wiki.ruleml.org>
- [S. Kisilevich et al. 2010] Slava Kisilevich, Florian Mansmann, Mirco Nanni, Salvatore Rinzivillo, Spatio-Temporal Clustering: a Survey, 2010
- [s3c]: <http://www.s3c-project.eu>
- [Sacco and Tzitzikas, 2009] Sacco GM, Tzitzikas Y (2009) Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience, 1st edn. Springer Publishing Company, Incorporated
- [Sacco, 2007] Sacco G (2007) Research results in dynamic taxonomy and faceted search systems. In: Database and Expert Systems Applications, 2007. DEXA '07. 18th International Workshop on, pp 201--206, doi{10.1109/DEXA.2007.75}.
- [Salima and Usman, 2015]: Flora Salima and Usman Haque. “Urban computing in the wild: A survey on large scale participation and citizen engagement with ubiquitous computing, cyber physical systems, and Internet of Things”. International Journal of Human-Computer Studies, Volume 81, September 2015, Pages 31–48.
- [Sarawagi, 2008] Sarawagi S (2008) Information extraction. Found Trends databases 1(3):261--377, doi 10.1561/1900000003.
- [Schmidt et al., 2009] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel. “Sp2bench: A sparql performance benchmark”. pages 222–233, Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on. IEEE, 2009.
- [Shepherd et al., 2004] Shepherd M, Watters C, Young J (2004) Context thesaurus for the extraction of metadata from medical research papers. In: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 6 - Volume 6, IEEE Computer Society, Washington, DC, USA, HICSS '04.
- [Smart cities 2007] Smart cities, Ranking of European medium-sized cities, www.smart-cities.eu, Centre of Regional Science, Vienna University of Technology, October 2007
- [Smart Mobility 2012] Smart Mobility for a 21st Century America Strategies for Maximizing Technology to Minimize Congestion, Reduce Emissions and Increase Efficiency, a White Paper by Transportation for America, ITS America, the Association for Commuter Transportation and the University of Michigan’s SMART Initiative, 2012
- [Snasel et al., 2009] Snasel V, Abraham A, Owais S, Plato? J, Kromer P (2009) Optimizing information retrieval using evolutionary algorithms and fuzzy inference system. In: Abraham A, Hassanien AE,

- Carvalho A (eds) Foundations of Computational Intelligence Volume 4, Studies in Computational Intelligence, vol 204, Springer Berlin Heidelberg, pp 299--324
- [Soboroff et al., 2001] Soboroff I, Nicholas C, Cahan P (2001) Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, SIGIR '01, pp 66--73, doi 10.1145/383952.383961.
- [SPARQL] W3C Consortium, "SPARQL 1.1 Query Language", W3C Recommendation, 21 March 2013, <http://www.w3.org/TR/sparql11-query/>
- [Specht et al., 2009]: Specht, Lorenz, Zimmermann. "Towards a Framework for Distributed User Modelling for Ubiquitous Computing". Advances in Ubiquitous User Modelling: Revised Selected Papers, 2009.
- [Sperer and Oard, 2000] Sperer R, Oard DW (2000) Structured translation for cross-language information retrieval. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, SIGIR '00, pp 120--127
- [Spink and Greisdorf, 2001] Spink A, Greisdorf H (2001) Regions and levels: Measuring and mapping users' relevance judgments. Journal of the American Society for Information Science and Technology 52(2):161--173
- [T. Bhattacharya et al. 2012] T. Bhattacharya, L. Kulik, J. Bailey, Extracting Significant Places from Mobile User GPS Trajectories: A Bearing Change Based Approach, 2012
- [Takagi and Tajima, 2001] Takagi T, Tajima M (2001) Query expansion using conceptual fuzzy sets for search engine. In: Fuzzy Systems, 2001. The 10th IEEE International Conference on, vol. 3, pp 1303--1308, doi 10.1109/FUZZ.2001.1008898.
- [Traffic 2013] Traffic Monitoring Guide, U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information, 2013
- [Tunkelang, 2009] Tunkelang D (2009) Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan and Claypool Publishers.
- [Turney, 2002] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", in Proc. Of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), pp. 417-424, Philadelphia, Pennsylvania, 2002.
- [Tzitzikas et al., 2008] Tzitzikas, Yannis; Theoharis, Yannis; Andreou, Dimitris, On Storage Policies for Semantic Web Repositories That Support Versioning, pp.705-719, LNCS 5021 The Semantic Web: Research and Applications, Springer, 2008
- [V. Zheng 2010] Vincent W. Zheng, Yu Zheng, Xing Xie, Qiang Yang, Collaborative Location and Activity Recommendations with GPS History Data, 2010
- [Vintan et al., 2004]: Vintan, Gellert, Petzold, Ungerer. "Person Movement Prediction Using Neural Network". Workshop on Modeling and Retrieval of Context 2004, MRC 2004, located at the 27th German Conference on Artificial Intelligence, Ulm, Germany, September 2004.
- [Vlaitutti, 2004] Valitutti, R., "Wordnet Affect: An affective extension of wordnet", In 4th International Conference on Language Resources and Evaluation,
- [Volkel et al., 2005] M. Volkel, W. Winkler, Y. Sure, S. R. Kruk, and M. Synak. "SemVersion: A Versioning System for RDF and Ontologies". In Procs. of the 2nd European Semantic Web Conf., ESWC'05., Heraklion, Crete, May 29 June 1 2005.
- [W. Chen et al., 2014] W. Chen, M. H. Ji and J. M. Wang, T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation
- [W. Liu et al. 2011] W. Liu, Y. Zheng, S. Chawla, J. Yuan, X. Xie, Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams, 2011
- [Wang and Zhu, 2010] Wang J, Zhu J (2010) On statistical analysis and optimization of information retrieval effectiveness metrics. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '10, pp 226--233
- [Welington et al., 2013] Welington M. da Silva, Alexandre Alvaro, Gustavo H. R. P. Tomas, Ricardo A. Afonso, Kelvin L. Dias, and Vinicius C. Garcia. 2013. Smart cities software architectures: a survey. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13). ACM, New York, NY, USA, 1722-1727. DOI=<http://dx.doi.org/10.1145/2480362.2480688>
- [Wu and Crestani, 2003] Wu S, Crestani F (2003) Methods for ranking information retrieval systems without relevance judgments. In: Proceedings of the 2003 ACM symposium on Applied computing, ACM, New York, NY, USA, SAC'03, pp 811--816, doi 10.1145/952532.952693.

- [Wu et al., 2007] C. Wu, M. Marches, J. Jiang, A. Ivanyukovich and Y. Liang, “Machine Learning-Based Keywords Extraction for Scientific Literature”, *Journal of Universal Computer Science*, Vol. 13(10), pp. 1471-1483, 2007.
- [X. Amatriain 2012] Xavier Amatriain. Building large-scale real-world recommender systems - recsys2012 tutorial. Technical report, Netflix, 2012.
- [X. Xiao 2010] X. Xiao, Y. Zheng, Q. Luo, X. Xie, Finding Similar Users Using Category-Based Location History, 2010
- [X. Xiao et al. 2012] X. Xiao, Y. Zheng, Q. Luo and X. Xie, Inferring social ties between users with human location history, *Journal of Ambient Intelligence and Humanized Computing* 2012
- [Xu and Croft, 2000] Xu J, Croft WB (2000) Improving the effectiveness of information retrieval with local context analysis. *ACM Trans Inf Syst* 18(1):79--112
- [Y. Chen et al. 2009] Y. Chen, K. Jiang, Y. Zheng, C. Li, N. Yu, Trajectory Simplification Method for Location-Based Social Networking Services, 2009
- [Y. Koren et al. 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.
- [Y. Lou 2009] Y. Lou, C. Zhang, Y. Zheng, Map-Matching for Low-Sampling-Rate GPS Trajectories, 2009
- [Y. Zheng 2011] Y. Zheng, X. Xie, Learning Travel Recommendations from User-Generated GPS Traces, 2011
- [Y. Zheng 2012] Y. Zheng, Tutorial on Location-Based Social Networks
- [Y. Zheng 2015] Y. Zheng, Trajectory Data Mining: An Overview
- [Y. Zheng and X. Xie 2010] Y. Zheng, X. Xie, Learning Location Correlation from GPS Trajectories, 2010
- [Y. Zheng et al. 2008] Y. Zheng, L. Liu, L. Wang, X. Xie, Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web, 2008
- [Y. Zheng et al. 2008] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.Y. Ma, Understanding Mobility Based on GPS Data, 2008
- [Y. Zheng et al. 2009] Y. Zheng, L. Zhang, X. Xie, W.Y. Ma, Mining Correlation between Locations Using Human Location History
- [Y. Zheng et al. 2009] Y. Zheng, L. Zhang, X. Xie, W.Y. Ma, Mining Interesting Locations and Travel Sequences from GPS Trajectories, 2009
- [Y. Zheng et al. 2010] Y. Zheng, X. Xie, W.Y. Ma, GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory, 2010
- [Y. Zheng et al. 2010] Y. Zheng, Y. Chen, Q. Li, X. Xie, W.Y. Ma, Understanding Transportation Modes Based on GPS Data for Web Applications, 2010
- [Y. Zheng et al. 2011] Y. Zheng, L. Zhang, X. Xie, W.Y. Ma, Recommending Friends and Locations Based on Individual Location History, 2011
- [Y. Zheng et al. 2014] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban Computing: Concepts, Methodologies, and Applications
- [Yang et al., 1992] Yang JJ, Korfhage R, Rasmussen EM (1992) Query improvement in information retrieval using genetic algorithms - a report on the experiments of the trec project. In: TREC, pp 31--58
- [Yuan and Yu, 2007] Yuan SA, Yu SN (2007) A new method for cross-language information retrieval by summing weights of graphs. In: Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 02, IEEE Computer Society, Washington, DC, USA, FSKD'07, pp 326--330
- [Z. Chen et al. 2010] Z. Chen, H. T. Shen, X. Zhou, Y. Zheng, X. Xie, Searching Trajectories by Locations – An Efficiency Study, 2010
- [Z. Li et al. 2010] Z. Li, J. G. Lee, X. Li, J. Han, Incremental Clustering for Trajectories, 2010
- [Zaheer et al., 2013] Zaheer Khan, Ashiq Anjum, and Saad Liaquat Kiani. 2013. Cloud Based Big Data Analytics for Smart Future Cities. In Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing (UCC '13). IEEE Computer Society, Washington, DC, USA, 381-386. DOI=<http://dx.doi.org/10.1109/UCC.2013.77>
- [Zavrel et al., 2000] Zavrel, J, Degroeve, S., Anne Kool, Daelemans, W. and Jokinen, K., “Diverse Classifiers for NLP Disambiguation Tasks Comparison, Optimization, Combination, and Evolution, in Proc. Of the 2nd CEvoLE Workshop “Learning to Behave”, 2000.

- [Zegins et al., 2007] D. Zeginis, Y. Tzitzikas, and V. Christophides. “On the Foundations of Computing Deltas Between RDF Models”. In Proc of the 6th Intern. Semantic Web Conf., ISWC/ASWC’07, pages 637–651, Busan, Korea, November 2007.
- [Zhang and Chen 2011] R. Zhang, X. Wu, S. Chen, Public Transportation Trip OD Matrix inference Using IC Card Data and GPS Information, 2011
- [Zhang et al., 2001] Zhang J, Gao J, Zhou M, Wang J (2001) Improving the effectiveness of information retrieval with clustering and fusion. Computational Linguistics and Chinese Language Processing 6(1):1--18
- [Zhang et al., 2012] Y. Zhang, M.-D. Pham, O. Corcho and J.-P. Calbimonte. “SRBench: A Streaming RDF/SPARQL Benchmark” In Proc. of the 11th International Semantic Web Conference ISWC 2012. Boston, USA, Nov 2012.
- [Bellini et al., 2012] Bellini P, Cenni D, Nesi P (2012) On the effectiveness and optimization of information retrieval for cross media content. In: KDIR, 4th International Conference on Knowledge Discovery and Information Retrieval, SciTePress, pp 344--347
- [Bellini et al., 2013] Bellini P, Bruno I, Cenni D, Nesi P, Paolucci M, Serena M (2013) A new generation digital content service for cultural heritage institutions. In: Nesi P, Santucci R (eds) Information Technologies for Performing Arts, Media Access, and Entertainment, Lecture Notes in Computer Science, vol 7990, Springer Berlin Heidelberg, pp 26--38, doi{10.1007/978-3-642-40050-6_3}.
- [Bellini et al., 2013b] P. Bellini, M. Benigni, R. Billero, P. Nesi and N. Rauch, "Km4City Ontology Bulding vs Data Harvesting and Cleaning for Smart-city Services", International Journal of Visual Language and Computing, Elsevier, <http://dx.doi.org/10.1016/j.jvlc.2014.10.023>, 2013
- [Bereta et al., 2013] K. Bereta, P. Smeros and M. Koubarakis. Representing and Querying the Valid Time of Triples for Linked Geospatial Data. In the 10th Extended Semantic Web Conference (ESWC 2013). Montpellier, France. May 26-30, 2013.
- [Erling et al., 2009] O. Erling and I. Mikhailov. “Virtuoso: RDF Support in a Native RDBMS”. In Semantic Web Information Management, pages 501-519. Springer, 2009.
- [Gómez-Pérez, 2004] Gómez-Pérez, A. Ontology Evaluation. Handbook on Ontologies. S. Staab and R. Studer Editors. Springer. International Handbooks on Information Systems. Pp: 251 – 274. 2004.
- [Isele, Bizer, 2013] R. Isele, C. Bizer. “Active learning of expressive linkage rules using genetic programming”. Web Semantics: Science, Services and Agents on the World Wide Web 23 (2013): pp.2-15
- [Kyzirakos et al., 2012] K. Kyzirakos, M. Karpathiotakis and M. Koubarakis. “Strabon: A Semantic Geospatial DBMS”. In the 11th International Semantic Web Conference (ISWC 2012), Boston, USA, 11-15 November 2012.
- [Ladwig et al., 2011] G. Ladwig, A. Harth, "CumulusRDF: Linked data management on nested key-value stores", 7th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2011), 2011
- [Ngomo, 2011] Ngomo, A. C. N., & Auer, S. Limes-a time-efficient approach for large-scale link discovery on the web of data. integration, 15, 3. (2011).
- [Papailiou et al., 2013] N. Papailiou, I. Konstantinou, D. Tsoumakos, P. Karras, N. Koziris, “H2RDF+: High-performance distributed joins over large-scale RDF graphs”, Big Data, 2013 IEEE International Conference on , pp.255,263, 6-9 Oct. 2013
- [Rector et al., 2004] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R.; Wang, H., Wroe, C. "Owl pizzas: Practical experience of teaching owl-dl: Common errors and common patterns". In Proc. of EKAW 2004, pp: 63 – 81. Springer. 2004.

7 Acronimi

- API: Application Program Interface
- AVL: Automatic vehicle location
- AVM: Automatic Vehicle Monitoring
- BDaaS: Big Data as a Service
- CAP principle: Consistency Availability Partition Tolerance principle
- CBB: Content Based Billing
- CBB: Content Based Billing

- CEN: European Committee for Standardization
- DBMS: database management system
- FCD: Floating Cellular Data
- GPRS: General packet radio service
- GPS: Global positioning System
- GSM: Global System for Mobile
- ICT: Information and Communication Technologies
- ITS: Intelligent Transport Systems
- LCD: liquid-crystal display
- LOD: linked open data
- MC: Mobile Collector
- MMS: Multimedia Messaging Service
- NLP: Natural Language Processing
- NoSQL: no SQL database
- OD: open data
- OD: Open Data
- OGC: Open Geospatial Consortium
- OWL: Web Ontology Language
- PA: Pubblica Amministrazione
- PMI: Piccola e Media Impresa
- PMS: Private Mobile Systems
- POS: part-of-speech
- RDF: Resource Description Framework
- RFID: Radio Frequency IDentification o Identificazione a radio frequenza
- RTTI: Real-time Travel & Traffic Information
- SDI: Spatial Data Infrastructures
- SII: sistema di interoperabilità integrato
- SIMONE: progetto Simone
- SMS: Short Message Service
- SN: social networking, oppure sensor network
- SOA: Service Oriented Architecture
- SOAP: Simple Object Access Protocol
- SSAMM: Agenzia per la Mobilità Metropolitana strumenti di supporto, TOSCANA
- TPEG: Transport Protocol Experts Group
- TPL: gestore trasporto pubblico locale
- UML: Unified Modeling Language
- UMTS: Universal Mobile Telecommunications System
- UTC: Urban Traffic Control
- UUDI: Universal Description Discovery and Integration
- V2I: Vehicle-to-Infrastructure
- V2V: vehicle-to-vehicle
- VMS: Variable Message Sign
- VWSN: Vehicular Wireless Sensor Networks
- W3C: World Wide Web Consortium
- WSD: Word Sense Disambiguation
- WSDL: Web Services Description Language
- WSN: Wireless Sensor Networks
- XMI: XML Metadata Interchange standard di OMG
- XML: Extensible Markup Language
- ZTL: Zona a Traffico Limitato