Dottorato di ricerca in Informatica
XXVIII ciclo

Progetto di ricerca

Dottorando: Dott. Giuseppe Rizzo
Tutor: Prof. Nicola Fanizzi

Cordinatore:
Prof. Donato Malerba

Firma del dottorando:_____

Firma del tutor _____

1 Titolo della ricerca

Tecniche di apprendimento automatico per l'estrazione di conoscenza da Linked Data

2 Area nella quale si inquadra la ricerca

Apprendimento automatico, Web Semantico e Linked Data

3 Obiettivi della ricerca

La ricerca si propone di definire una metodologia, basata sull'utilizzo di algoritmi di apprendimento automatico per l'estrazione di conoscenza dai Linked Data (LD), i quali rappresentano il risultato dell'integrazione di dati provenienti da fonti eterogenee, pubblicati nel formato Resource Description Framework (RDF), utilizzato per rappresentare metadati strutturati [6]. Per esso, gli attuali linguaggi di interrogazione, come SPARQL, non permettono di effettuare inferenza. La ricerca si propone di indagare le problematiche connesse all'ottenimento di nuova conoscenza nel Semantic Web (come la tipica assunzione di mondo aperto che rende l'informazione incompleta e implica l'impossibilità di derivare la verità di determinati assiomi nelle basi di conoscenza [3]), con specifico riferimento ai LD (i quali rappresentano la visione più attuale in merito alla condivisione e al riuso dei dati pubblicati). In particolare si cercherà, inoltre di indagare, sui limiti degli attuali approcci di apprendimento induttivo adottati per inferire sulle basi di conoscenza del Semantic Web, allo scopo di determinare la misura in cui essi siano applicabili anche ai LD. Un ulteriore obiettivo sarà quello di proporre nuovi metodi e di migliorare quelli esistenti e sottoporli a valutazione sperimentale.

4 Motivazioni della ricerca

Nel corso degli ultimi anni l'interesse nel fare evolvere il Web verso il cosiddetto Web Semantico (Semantic Web) è cresciuto, allo scopo di migliorare l'elaborazione automatica delle informazioni disponibili, come per i risultati restituiti dai motori di ricerca. Secondo la visione tradizionale, il Web è una collezione di documenti collegati tra di loro. La semantica dei contenuti e degli eventuali collegamenti è comprensibile esclusivamente dagli esseri umani, nonostante le risorse siano accessibili ed elaborabili dalle macchine. Trasformando il Web in Semantic Web, l'obiettivo che si intende perseguire è rendere l'informazione comprensibile

dalle macchine, riutilizzando l'infrastruttura già esistente e favorendo così l'interoperabilità a livello semantico [5], vale a dire far sì che le informazioni scambiate siano interpretate nello stesso modo da tutte le applicazioni che le utilizzano. La diffusione dei LD rappresenta un modo per garantire questa interoperabilità, rendendo possibile il collegamento di dataset. Inoltre, il linguaggio SPARQL, utilizzato per interrogare dataset in RDF non permette di effettuare inferenza su LD. Perciò è necessario creare una sovrastruttura che sia in grado di affrontare diverse problematiche [26] riconducibili all'enorme quantità di dati esistente e in continua evoluzione, causa di inconsistenza degli stessi e contenente una quantità di informazione non esplicitamente rappresentata che non può essere catturata dalla semantica di RDFS e OWL singolarmente.

A causa di questa serie di problemi e dell'assunzione del mondo aperto del Semantic Web (OWA), si rendono necessarie forme alternative di inferenza, come quelle basate sull'apprendimento induttivo [11], per costruire modelli (di tipo predittivo e non) a partire dai dati tolleranti al rumore e scalabili. Tuttavia lo stato dell'arte del learning su LD non considera totalmente la conoscenza presente, esplicitamente o implicitamente, nelle ontologie alle quali i LD fanno riferimento.

5 Stato dell'arte

In [11], gli autori motivano l'esigenza di utilizzare l'apprendimento induttivo nell'ambito del Semantic Web per sopperire alla mancanza di informazione dovuta all'OWA. Il ragionamento deduttivo, infatti, affinchè possa consentire di trarre conclusioni necessita di una serie di premesse che si assume siano corrette e complete. A causa di ciò il trattamento dell'incertezza non è possibile attraverso il ragionamento deduttivo. Inoltre, in un contesto come quello del Semantic Web, dove il numero di premesse è elevato, l'inferenza deduttiva non appare scalabile da un punto di vista computazionale. Al contrario, l'apprendimento induttivo consente di costruire, a partire dai dati (anche rumorosi o inconsistenti), un modello generale plausibile in grado di spiegarli. La costruzione di tale modello, è possibile anche quando la base di conoscenza non è fuzzy o probabilistica, pemettendo di estrarre nuova conoscenza (non esatta). La letteratura in merito all'apprendimento sul Semantic Web e sui LD, ha proposto diversi approcci induttivi finalizzati alla costruzione di modelli, sia di tipo descrittivo sia di tipo predittivo. Nel primo caso un approccio possible è quello di apprendere un'ontologia a partire dai LD e scoprire grazie all'algoritmo impiegato nuove relazioni. Un esempio è costituito da [7, 20, 32]. In esso, l'approccio utilizzato consiste nel determinare regole di associazione per combinare le informazioni fornite da una pluralità di fonti e rendere possibili nuove inferenze. L'algoritmo Apriori [2] rappresenta uno degli algoritmi

più noti per l'estrazione di regole di associazione ed è quello adoperato nei lavori citati a causa dell' ipotesi alla base, ossia l'assunzione che un itemset frequente di n elementi, possa essere determinato ricorsivamente sulla base degli itemset frequenti di $n-1$ elementi. Nell'articolo [7], l'approccio proposto muove i passi per la scoperta di regole di associazione a partire da due distinte fonti di informazione (assumendo che entrambe condividano un insieme di attributi): un'ontologia e un database relazionale. Attraverso una serie di euristiche, il metodo descritto procede a creare una fonte di informazione risultante dalla integrazione di quelle originali, prima di determinare gli itemset frequenti e successivamente le regole di associazione in funzione dei livelli di supporto e confidenza stabiliti. Un approccio analogo è quello descritto in [32]. Anche qui attraverso una serie di euristiche il grafo è mappato su una serie di assiomi. Per esempio il costrutto `rdf:type` associato a una risorsa permette di determinare la classe di appartenenza della stessa. Una volta realizzato questo mapping, è possibile determinare la tabella delle transizioni. In essa, ogni riga/transizione corrisponde a un individuo (o coppia di essi), mentre i concetti e le properties sono le colonne di questa tabella. D'altra parte per ciò che concerne la creazione di modelli predittivi, nell'ambito dell'apprendimento sui LD [16] ha proposto un approccio semisupervisionato per l'apprendimento di modelli predittivi dai dati estratti dal LD Cloud [17], per superare le difficoltà relative all'etichettare completamente un dataset. In questo lavoro sono stati applicati diversi algoritmi di learning noti in letteratura e un approccio semisupervisionato chiamato *self-training*, il quale consiste nell'utilizzare le predizioni effettuate per arricchire l'insieme dei dati etichettati. Inoltre, nell'ambito dell'apprendimento dai LD, gli approcci di apprendimento statistico-relazionale (SRL) hanno assunto un'importanza crescente e sono stati proposti in lavori come [19,27,31]. Tra gli approcci di SRL occorre citare i modelli grafico-relazionali, che estendono alle teorie del prim'ordine i modelli grafici. I modelli grafico-relazionali [23] comprendono i cosiddetti *Probabilistic Graphic Model*. Gli archi del modello rappresentano la relazione esistente tra i nodi. In questa tipologia di modelli esistono due tipi di incertezza strutturale: l'incertezza di riferimento e l'incertezza di esistenza, legate all'ipotesi che la relazione sia nota a priori. Questo tipo di modelli sono di tipo grafico orientato. Invece, tra i modelli legati a grafi non orientati è possibile menzionare le reti logiche Markoviane (MLN). Una MLN rappresenta un insieme di formule F_i con associato un peso w_i. Per ogni possibile statement/atomo ground si introduce un nodo binario, dato un insieme di costanti. Lo stato di ogni nodo è pari a 1 se l'atomo ground è vero, 0 diversamente. In più MLN contiene una feature per ogni possibile grounding della formula F_i. Il valore di una feature è 1 se è la formula è vera, 0 altrimenti. I pesi utilizzati da una MLN esprimono il livello di confidenza. Dopo che è stato appreso un insieme di reti di Markov, il learning in una MLN consiste proprio nello stimare i valori di questi pesi Un altro

modello adottato è l'*infinite hidden relational model*(IHRM), legato all'utilizzo di variabili latenti, cioè variabili/attributi non noti le quali rappresentano variabili genitori per quelle osservabili. Un vantaggio derivante dall'IHRM consiste nella non necessità di apprendere la struttura della rete, poichè essa è stabilita sulla base della struttura del grafo del Semantic Web [27]. Nei LD sono stati applicati approcci di machine learning anche per trattare il problema dell'instance matching, vale a dire verificare che due URI siano associati alla stessa risorsa. Questo perchè l'equivalenza espressa mediante il costrutto `owl:SameAs` è frequentemente utilizzata per la creazione di link tra dataset. L'approccio descritto in [28] definisce un classificatore in grado di attribuire a una coppia di individui due possibili valori, +1 e -1 per esprimere il fatto che due istanze siano uguali oppure no. Per fare questo gli autori definiscono una misura di distanza basata su un insieme di feature le quali rappresentano informazioni di tipo testuale. Sempre per la creazione automatica di link per i LD, in [29] è stato applicato un approccio di *link prediction*, attraverso la definizione di un framework predittivo che gli autori hanno applicato a un grafo tripartito, cioè costituito da tre insiemi di nodi e due di archi. L'approccio proposto si basa sull'applicazione di metodi di *graph summarization* che permettono la definizione di super-nodi e super-archi, cioè nodi e archi che rappresentano rispettivamente gruppi di nodi e insieme di archi tra nodi. L'utilizzo di questa tipologia di metodi permette di catturare la semantica non solo dei singoli nodi ma anche degli archi che collegano i nodi stessi. Il grafo compatto è successivamente dato in input a una funzione di predizione che restituisce un rank dei link indotti. Il task di link prediction, inoltre, è stato considerato anche in [25], per validare due funzioni kernel proposte dagli autori e basate su grafi. L'interesse verso i metodi kernel è legato alla possibilità di disaccoppiare la rappresentazione dei dati e il particolare task di apprendimento. Gli autori propongono due approcci basati su grafi: uno basato sul grafo di intersezione e un altro basato sull'albero di intersezione (costruiti a partire da un grafo di *neighborhood*, attraverso una visita in ampiezza fino a una profondità k). Le funzione kernel è determinata sulla base di caratteristiche come il numero di *walk*, *path* oppure il numero di *full subtree* nel caso degli alberi di intersezione. Questi metodi, appaiono simili a quanto proposto in [12], sebbene gli autori propongano un kernel basato sul conteggio dei percorsi all'interno di un albero costruito dal grafo RDF e da uno specifico vertice.

Come è stato anticipato nella sezione 4, lo stato dell'arte mira a valutare una serie di algoritmi di learning, senza che la conoscenza rappresentata nell'ontologia sia utilizzata e integrata in qualche maniera. Gli approcci di link prediction così come i kernel descritti precedentemente non considerano la semantica dei dati. Emblematico a tal proposito è il lavoro di [22], in cui è proposto un approccio che mira a predire un link in un grafo multirelazionale. Più esattamente questo approccio mira a stabilire la likelihood di uno statement RDF in funzione degli altri

statement sfruttando i LD come un dominio di valutazione del metodo proposto. Diversamente, in [21], gli autori hanno integrato approcci alternativi per la predizione di relazioni su LD. L'aspetto interessante di questo lavoro è inerente l'aspetto dell'integrazione di forme di ragionamento differente, come quello induttivo (vale a dire le predizioni effettuate con algoritmi di machine learning) e deduttivo(la base di conoscenza). Sebbene il dominio di applicazione di questo progetto di ricerca sia quello dei LD e non delle ABox di una singola base di conoscenza, l'idea di integrare, seppure in modo differente, la conoscenza disponibile nell'ontologia nel learning è descritta in lavori come [8,10,14,15] nei quali sono stati proposti diversi algoritmi di tipo instance-based (il k-nearest neighbor), in [8, 10, 14], oppure basati su propotipi (le Reduced Coloumbus Energy Network) descritte in [15]. Alla base degli approcci instance-based, c'è l'idea secondo la quale individui simili della base di conoscenza hanno un comportamento simile rispetto a un insieme di feature, rappresentato dall'insieme di concetti atomici presenti nell'ontologia considerata, e che prende il nome di *committee* o *contesto*. Il comportamento di un individuo rispetto al contesto può essere determinato proprio grazie a un reasoner. Quindi è stata definita una famiglia di pseudomisure di distanza ispirata alla famiglia di misure Minkowski [9], e applicate congiuntamente con il k-nearest neighbor [8, 10] allo scopo di stabilire se un individuo fosse o meno istanza di un concetto (class-membership) e successivamente anche per inferire induttivamente asserzioni del tipo $R(a, b)$ (filler sui ruoli), data una coppia di individui (a, b) e un ruolo *R*, e per predire il valore (categorico) per un datatype property dato un individuo (filler su datatype properties) [14].

6 Approccio al problema

L'approccio al problema necessita di considerare una serie di aspetti come:

1. l'individuazione delle tecnologie e dei metodi per la raccolta dei dati dal LD Cloud e la creazione di dataset.

2. la individuazione di metodi più opportuni di apprendimento induttivo proponendone sia di nuovi sia di esistenti ma con l'apporto di miglioramenti che andranno a costituire una metodologia. Gli algoritmi di apprendimento proposti dovranno considerare l'integrazione della conoscenza rappresentata (esplicitamente o implicitamente) nell'ontologia.

3. la valutazione sperimentale degli approcci proposti

Da un punto di vista applicativo i metodi proposti che si riveleranno sperimentalmente più promettenti, andranno a costituire una suite potenzialmente impiegabile

per la realizzazione di applicazioni mashup, cioè applicazioni che utilizzano i LD provenienti da più organizzazioni o istituzioni.

7 Ricadute applicative

In generale, il tema dei LD è diventato nel corso degli ultimi anni piuttosto importante. Infatti, una direttiva del Parlamento Europeo e del Consiglio (Direttiva 2003/98/CE) sensibilizza gli enti pubblici rispetto alla condivisione dei propri dati, al fine di utilizzare questa potenzialità sia per completare e migliorare l'offerta di servizio pubblico, sia per contribuire alla creazione di nuove opportunità produttive.Molti enti nazionali, come ad esempio il Governo Italiano (`http://www.dati.gov.it`) e quelli della Spagna (`http://www.datos.gob.es`) e dell'Inghilterra (`http://www.data.gov.uk`), hanno dotato i propri portali web di sezioni dedicate all'esposizione pubblica dei LD. Inoltre, secondo l'analisi a cura della Commissione europea (il cosiddetto rapporto Vickery), il valore di mercato del riuso dell'informazione del settore pubblico è stimato intorno ai 140 miliardi di euro all'anno nell'Unione Europea [4]. Attraverso la disponibilità di LD governativi e dei metodi che saranno sviluppati per l'apprendimento di modelli, potrebbe essere possibile effettuare diagnosi, come per esempio predizioni, relative a quali possono essere strade congestionate di un determinata città e più in generale realizzare tutte quelle applicazioni in grado di gestire informazioni urbane secondo la visione delle *Smart Cities* [24].

Ulteriori ricadute applicative potrebbero riguardare l'impiego dei metodi in ambienti di tipo context-aware allo scopo di supportare persone anziane affette da demenza [30]. Le tecnologie semantiche offrirebbero, quindi, la possibilità di rendere gli ambienti intelligenti consentendo il miglioramento della qualità della vita. Inoltre, nell'ambito della Social Network Analysis è possibile, mediante l'impiego di algoritmi di apprendimento automatico, estrarre nuova conoscenza non considerando soltanto i collegamenti (rappresentati mediante grafi RDF) ma anche la loro semantica [1]. I metodi proposti potrebbero essere applicati anche nel contesto dei Virtual Research Environment (VRE), ambienti collaborativi pensati per supportare gruppi di ricerca multidisciplinare, per realizzare servizi di ragionamento sulle politiche di restrizione dei contenuti condivisi dagli utenti [13]. Integrando opportunamente i modelli (ottenuti con i metodi) proposti, a partire dai LD nei motori di ricerca dovrebbe essere possibile migliorare la qualità dei motori di ricerca, dando la possibilità di ritrovare ulteriori informazioni in merito ai risultati restituiti all'utente [18]. Altre ricadute applicative riguardano la possibilità di integrazione dei metodi di apprendimento proposti in SPARQL consentendo per esempio di aggregare i risultati delle query e arricchendo in questo modo il linguaggio stesso.

Infine la possibilità di apprendere dai LD, consentendo di ottenere informazioni non banali, permetterebbe ai *knowledge engineer* di determinare automaticamente nuovi link e arricchire il LD Cloud [11].

8 Fasi del progetto

Il progetto sarà articolato nelle seguenti fasi:

1. **Studio dello stato dell'arte** (Primo anno). Questa fase si articolerà nelle seguenti attività:
   (a) *Studio dello stato dell'arte in merito ai Linked Data.* In questa attività verranno approfonditi i temi del Semantic Web e dei LD(in particolare legati alla caratteristiche di questo tipo di dati) e quelli inerenti i metodi e le tecnologie applicati per la pubblicazione dei dati e la loro acquisizione
   (b) *Studio dello stato dell'arte e degli algoritmi di learning che è possibile applicare nel campo dei LD*, prendendo in esame anche gli approcci più recenti come l'apprendimento semisupervisionato, lo SRL, ecc.;
   (c) *Studio approfondito delle modalità di integrazione della conoscenza rappresentata negli approcci di learning*. In questa attività verranno presi in esame gli aspetti di incertezza legati ai LD e verranno studiati gli approcci integrabili nel task dell'apprendimento
2. **Sintesi di soluzioni innovative** (Secondo anno)
   (a) *Sintesi di algoritmi innovativi e di varianti di soluzioni esistenti.*Questa attività prevederà di volta in volta, anche la definizione di setting ottimali per le sperimentazioni
   (b) *Pubblicazione delle soluzioni e partecipazione a workshop, conferenze, ecc*.
   (c) *Frequentazione di scuole estive relative al Semantic Web, LD e machine learning*
3. **Definizione della metodologia** (tra Secondo e Terzo anno). Essa si concretizzerà in una sola attività, cioè:
   (a) *selezione dei metodi più promettenti per il ragionamento induttivo con il trattamento dell'incertezza relativa ai LD, per la definizione di un framework*

4. **Stesura della tesi di dottorato** (Terzo anno)

E' previsto, tra il secondo e terzo anno, un periodo di almeno tre mesi di studio presso una struttura di ricerca europea specializzata nell'ambito del Semantic Web e dei LD.

Dal punto di vista delle risorse necessarie a portare a termine il progetto, sarà necessario procedere all'ottenimento dei dataset. A tal proposito potranno essere utilizzati dati impiegati in altri lavori e eventualmente disponibili online. In alternativa potranno essere creati di nuovi, opportunamente estratti dal LD cloud.

Per ciò che concerne i risultati miliari attesi durante gli anni di dottorato, essi consistono nella realizzazione delle pubblicazioni che verranno presentate nei vari workshop, conferenze, riviste inerenti il Semantic Web e i LD.

9 Valutazione dei risultati

Dallo studio della letteratura proposta, molte delle tecniche adottate valutano i risultati degli esperimenti mediante misure classiche del machine learning, come precisione, richiamo e *f-measure* (sia nel caso di modelli predittivi sia nel caso delle regole di associazione come descritto in [32]). Invece per poter valutare i metodi di *link prediction* è possibile utilizzare misure come la *Normalized Discounting Cumulative Gain* allo scopo di valutare i rank predetti [27]. Adottare queste misure permetterebbe di effettuare un primo confronto con i metodi che costituiscono lo stato dell'arte. Tuttavia dai lavori come [8, 10, 15] deriva l'idea di trattare nelle misure assunzioni come quella dell'OWA definendo misure in grado di esprimere completezza e correttezza del metodo adottato. Le misure proposte nei lavori precedentemente citati sono le seguenti: *match*, *commision*, *omission*, *induction*. Chiaramente le misure adottate per la validazione dipenderanno dallo specifico task considerato (predittivo o non) e non è esclusa la possibilità di definirne di nuove. Inoltre potrebbe essere necessario e opportuno condurre *proof-of-concept* allo scopo di comprendere se i metodi proposti hanno le potenzialità per essere applicati anche in applicazioni reali.

10 Eventuali referenti esterni al Dipartimento

Achim Rettinger

Institute of Applied Informatics and Formal Description Methods (AIFB)
Karlshuer Institute für Technologie
Phone: 0721 608 46592
Email: rettinger@kit.edu

Johanna Völker

Research Group Data and Web Science
University of Mannheim
Phone: +49 621 181 2661
E-mail: johanna@informatik.uni-mannheim.de

Riferimenti bibliografici

[1] F. Abel, C. Hauff, G.-J. Houben, and K. Tao, "Leveraging user modeling on the social web with linked data," in *Web Engineering - 12th International Conference, ICWE 2012, Berlin, Germany, July 23-27, 2012. Proceedings*, ser. Lecture Notes in Computer Science, M. Brambilla *et al.*, Eds., vol. 7387. Springer, 2012, pp. 378–385.

[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.

[3] F. Baader and others., Eds., *The description logic handbook: theory, implementation, and applications*. New York, NY, USA: Cambridge University Press, 2003.

[4] M. Barbera and F. D. Donato. (2012) Le implicazioni economiche degli open data in italia. http://www.linkedopendata.it/implicazioni-economiche-open-data.

[5] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, May 2001.

[6] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[7] C. d'Amato, V. Bryl, and L. Serafini, "Semantic knowledge discovery from heterogeneous data sources," in *Knowledge Engineering and Knowledge Management*, ser. Lecture Notes in Computer Science, A. Teije *et al.*, Eds. Springer Berlin Heidelberg, 2012, vol. 7603, pp. 26–31.

[8] C. d'Amato, N. Fanizzi, and F. Esposito, "Analogical reasoning in description logics," in *Uncertainty Reasoning for the Semantic Web I*, ser. Lecture Notes in Computer Science, P. Costa *et al.*, Eds. Springer Berlin Heidelberg, 2008, vol. 5327, pp. 330–347.

[9] C. D'Amato, N. Fanizzi, and F. Esposito, "Induction of optimal semantic semi-distances for clausal knowledge bases," in *Proceedings of the 17th international conference on Inductive logic programming*, ser. ILP'07. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 29–38.

[10] C. d'Amato, N. Fanizzi, and F. Esposito, "Query answering and ontology population: An inductive approach," in *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, ser. ESWC'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 288–302.

[11] ——, "Inductive learning for the semantic web: What does it buy?" *Semant. web*, vol. 1, no. 1,2, pp. 53–59, apr 2010.

[12] G. K. D. de Vries and S. de Rooij, "A fast and simple graph kernel for rdf," 2013, submitted.

[13] P. Edwards, E. Pignotti, A. Eckhardt, K. Ponnamperuma, C. Mellish, and T. Bouttaz, "ourspaces – design and deployment of a semantic virtual research environment," in *Proceedings of the 11th international conference on The Semantic Web - Volume Part II*, ser. Lectures Notes on Computer Science, Cudré-Mauroux *et al.*, Eds. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 50–65.

[14] N. Fanizzi, C. d'Amato, and F. Esposito, "Evidential nearest-neighbors classification for inductive abox reasoning," in *URSW*, F. Bobillo *et al.*, Eds., 2009, pp. 27–38.

[15] N. Fanizzi, C. D'Amato, and F. Esposito, "Reduce: A reduced coulomb energy network method for approximate classification," in *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ser. ESWC 2009 Heraklion. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 323–337.

[16] N. Fanizzi, C. d'Amato, and F. Esposito, "Mining linked open data through semi-supervised learning methods based on self-training," in *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing*, ser. ICSC '12, 2012, pp. 277–284.

[17] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, ser. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[18] I.-C. Hsu, H.-Y. Lin, L. J. Yang, and D.-C. Huang, "Using linked data for intelligent information retrieval," in *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, 2012, pp. 2172–2177.

[19] Y. Huang, V. Tresp, M. Nickel, A. Rettinger, and H. Kriegel, "A scalable approach for statistical learning in semantic graphs," *Semantic Web*, vol. 1, pp. 1–5, 2012.

[20] Q. Ji, Z. Gao, and Z. Huang, "Reasoning with noisy semantic data," in *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ser. ESWC'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 497–502.

[21] X. Jiang, Y. Huang, M. Nickel, and V. Tresp, "Combining information extraction, deductive reasoning and machine learning for relation prediction," in *Proceedings of the 9th international conference on The Semantic Web: research and applications*, ser. ESWC'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 164–178.

[22] X. Jiang, V. Tresp, Y. Huang, and M. Nickel, "Link prediction in multi-relational graphs using additive models," in *SeRSy*, ser. CEUR Workshop Proceedings, M. de Gemmis *et al.*, Eds. CEUR-WS.org, 2012, pp. 1–12.

[23] H. Khosravi and B. Bina, "A survey on statistical relational learning," in *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence*, ser. AI'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 256–268.

[24] F. Lécué, A. Schumann, and M. L. Sbodio, in *Proceedings of the 11th international conference on The Semantic Web - Volume Part II*, ser. Lectures Notes on Computer Science, Cudré-Mauroux *et al.*, Eds., pp. 114–130.

[25] U. Lösch, S. Bloehdorn, and A. Rettinger, "Graph kernels for rdf data," in *Proceedings of the 9th international conference on The Semantic Web: research and applications*, ser. ESWC'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 134–148.

[26] A. Polleres, A. Hogan, R. Delbru, and J. Umbrich, "RDFS & OWL reasoning for linked data," in *Reasoning Web 2013*, ser. Lecture Notes in Computer Science (LNCS), S. Rudolph and H. Stuckenschmidt, Eds. Mannheim, Germany: Springer, July 2013, to appear.

[27] A. Rettinger, U. Lösch, V. Tresp, C. d'Amato, and N. Fanizzi, "Mining the semantic web - statistical learning for next generation knowledge bases," *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 613–662, 2012.

[28] S. Rong, X. Niu, E. W. Xiang, H. Wang, Q. Yang, and Y. Yu, "A machine learning approach for instance matching based on similarity metrics," in *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ser. ISWC'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 460–475.

[29] A. Thor, P. Anderson, L. Raschid, S. Navlakha, B. Saha, S. Khuller, and X.-N. Zhang, "Link prediction for annotation graphs using graph summarization," in *The Semantic Web – ISWC 2011*, ser. Lecture Notes in Computer Science, L. Aroyo *et al.*, Eds. Springer Berlin Heidelberg, 2011, vol. 7031.

[30] T. Tiberghien, M. Mokhtari, H. Aloulou, and J. Biswas, "Semantic reasoning in context-aware assistive environments to support ageing with dementia," in *Proceedings of the 11th international conference on The Semantic Web - Volume Part II*, ser. Lectures Notes on Computer Science. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 212–227.

[31] V. Tresp, M. Bundschus, A. Rettinger, and Y. Huang, "Uncertainty reasoning for the semantic web i," P. C. Costa *et al.*, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. Towards Machine Learning on the Semantic Web, pp. 282–314.

[32] J. Völker and M. Niepert, "Statistical schema induction," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, G. Antoniou *et al.*, Eds. Springer Berlin Heidelberg, 2011, vol. 6643, pp. 124–138.