Dottorato di ricerca in Informatica
XXVIII ciclo

Progetto di ricerca

Dottorando: Dott. Pasqua Fabiana Lanotte
Tutor: Prof. Michelangelo Ceci

Cordinatore:
Prof. Donato Malerba

Firma del dottorando:_____

Firma del tutor _____

1 Titolo della ricerca

Tecniche di Data Mining per l'estrazione di dati strutturati dal Web

2 Area nella quale si inquadra la ricerca

Data Mining, Web Mining

3 Obiettivi della ricerca

Nonostante il Web sia considerato come la più grande libreria di "documenti ipertestuali", interconnessi e condivisi (pagine Web), esso contiene una significativa quantità di "dati strutturati" di qualsiasi tipo, dai prodotti alla finanza, ai dati pubblici (open data), ai dati scientifici (es. Connectivity Map, DBLP), agli annunci immobiliari, news e tanto altro. Questi dati, rappresentanti entità, oggetti o concetti, condividono una struttura comune con i dati tipicamente contenuti nei database. Tuttavia i motori di ricerca attuali (es. Google) indicizzano le pagine Web basandosi quasi completamente sui contenuti non strutturati delle pagine stesse, e sfruttando solo i dati strutturati di particolari pagine Web (es. Wikipedia). Inoltre a causa dell'eterogeneità del Web, sebbene le basi di dati siano correntemente usate per generare i contenuti di alcuni siti, gli schemi sottostanti, anche se usati per rappresentare gli stessi tipi di oggetti, sono spesso tra loro inconsistenti. Quindi l'utilizzo combinato di strumenti capaci di: 1) estrarre i "dati strutturati" presenti nel Web e 2) popolare e/o integrare basi di dati con le informazioni estratte, consentirebbe di riconciliare la natura non strutturata del Web con i modelli strutturati tipici delle basi di dati. In questo modo sarebbe ad esempio possibile integrare le informazioni strutturate provenienti da un database bibliografico relativo a pubblicazioni scientifiche, docenti e conferenze (es. DBLP) con informazioni provenienti dai contenuti strutturati delle pagine Web dei docenti. In tale ottica, l'obiettivo della ricerca è quello di sintetizzare e implementare algoritmi, computazionalmente efficienti, capaci di:

1. estrarre "dati strutturati" nel Web (in generale presenti sotto forma di liste e data record [7]);
2. scoprire pagine Web Entità [16];
3. associare i dati scoperti a record nei database;

4. apprendere modelli di classificazione di pagine Web Entità;

In questo modo sarà ad esempio possibile costruire motori di ricerca in grado di restituire dati strutturati in risposta a query testuali ed integrare la natura parzialmente strutturata del Web con le basi di dati. I risultati restituiti rappresenterebbero punti di arrivo per i bisogni informativi degli utenti, piuttosto che punti di partenza per nuove ricerche.

4 Motivazioni della ricerca

L'estrazione e l'integrazione automatica di informazioni strutturate e non strutturate è un obiettivo che coinvolge diverse discipline di ricerca (Intelligenza artificiale, Data Base, Data Mining, Information Retrieval, NLP, comunità Web, etc.). Le cause di una simile popolarità sono da ricercarsi nell'impatto potenziale che il raggiungimento di un simile obiettivo potrebbe avere nel modo in cui l'informazione è gestita e ricercata. Tale integrazione potrebbe ad esempio permettere l'utilizzo di query non strutturate (es. interrogazioni Google) per restituire contenuti non strutturati in forma strutturata, o l'utilizzo di query strutturate (es. interrogazioni SQL) per restituire contenuti non strutturati (es. pagine Web). Attualmente i più popolari motori di ricerca effettuano ranking delle pagine Web basandosi su due assunzioni: 1) le pagine Web contengono inter-domain link (es. la pagina Web di un docente appartenente al dominio Uniba contiene link a pagine Web del dominio Wikipedia), 2) tutte le pagine sono rappresentabili tramite insiemi di keyword (secondo modelli VSM o LSI). Sistemi siffatti funzionano bene per pagine statiche, contenenti dati non strutturati ed editate da esseri umani, che hanno interesse a creare link tra pagine Web semanticamente correlate ed appartenenti a domini differenti. Tuttavia nel Web esiste una grande quantità di pagine Web che:

1. sono generate automaticamente a partire da basi di dati (Hidden Web Database);

2. non contengono link inter-domain (es. le pagine web dei prodotti di Amazon non contengono link alle pagine web degli autori);

3. contengono dati strutturati e/o rappresentano delle entità ben distinte (docenti, libri, case in vendita,etc.)

Queste pagine non sono attualmente indicizzate dai vari motori di ricerca, perchè 1) non contengono link inter-domain (sono penalizzate nel ranking

dei risultati di ricerca) 2) le entità ed i dati strutturati presenti non sono rappresentabili tramite un insieme di keyword. Per superare questi limiti gli attuali approcci utilizzano wrapper, costruiti in modo più o meno ad hoc, per estrarre dati strutturati, relativi a particolari entità, dalle pagine web e regole per integrare i dati estratti, al fine di migliorare i risultati di ricerca. Soluzioni di questo tipo risultano computazionalmente costose e soprattutto non automatizzabili, in quanto richiedono un importante apporto da parte dell'umano. Inoltre in questi casi una qualsiasi modifica alla struttura di un sito web renderebbe i wrapper addestrati inutilizzabili in quanto obsoleti. Realizzare degli strumenti generali capaci si analizzare in modo automatico ed efficiente il Web, estraendo ed integrando i dati strutturati, consentirebbe la fruizione di questi ai motori di ricerca, garantendo un miglioramento del ranking delle pagine Web.

5 Stato dell'arte

Di seguito verranno riportati brevemente gli approcci esistenti in letteratura relativi all'estrazione di dati strutturati nel web, scoperta di pagine Entità, classificazione e associazione delle pagine scoperte ai database.

5.1 Estrazione di dati strutturati da pagine Web

Negli ultimi anni sono stati proposti in letteratura diversi approcci capaci di estrarre dati strutturati e semi-strutturati dalle pagine Web. In recenti lavori Cafarella et al. [2] hanno mostrato come le tabelle HTML rappresentano sorgenti ricche di dati strutturati. I loro risultati mostrano l'esistenza nel Web oltre di 150 milioni di tabelle HTML contenenti dati relazionali. In particolare in [3] gli autori propongono WebTables, un approccio per estrarre dati relazionali dal Web espressi attraverso tag HTML. A partire da un insieme di pagine Web, WebTables ottiene un database relazionale attraverso la realizzazione di 2 fasi: 1) estrazione di tabelle HTML contenenti dati relazionali, 2) integrazione dei dati estratti in tabelle relazionali. In entrambe le fasi il sistema utilizza tecniche di apprendimento supervisionato e regole scritte da esperti per selezionare, tra tutte le tabelle presenti in una pagina Web, quelle contenenti dati relazionali e verificare che i dati estratti siano conformi allo schema della base di dati. Questo rappresenta un forte limite in quanto non permette l'estensione di un tale approccio a qualsiasi dominio. I dati estratti possono ad esempio essere utilizzati per motori di ricerca che, data una pagina web contenente una tabella che descrive l'apporto calorico di alcuni alimenti, restituiscono la pagina contenente questa tabella tra i

primi risultati per la query "latte calorie", anche se i termini della query sono spazialmente distanti. Elmeleegy et al. [7] assumono che le pagine Web contengono dati strutturati sotto forma di liste, le quali non rispettano particolari template e i singoli elementi appartenenti alle liste (data record) potrebbero non contenere la stessa quantità di informazioni. Gli autori propongono in questo caso un approccio non supervisionato e indipendente dal dominio per estrarre tabelle relazionali da liste web. Infine Chang et al. [5] propongono Cazoodle, un sistema capace di estrarre dati da centinaia di sorgenti online (es. deep database, forum, etc.) e di integrare i dati ottenuti in modo da generare motori di ricerca verticali. Un motore di ricerca verticale è un sistema capace di effettuare ricerche dedicate rispetto ad un particolare argomento (es. Trivago.com, skyscanner.com). Il sistema non permette di realizzare motori di ricerca generici, ma solo limitati a alcuni domini (appartamenti, vacanze, prodotti elettronici).

5.2 Individuazione di Entity Web Pages e Associazione di Pagine Entità a Database

In letteratura sono stati proposti diversi approcci capaci di estrarre entità nominate da documenti strutturati come pagine Web [13]. In particolare, nel campo del Named Entity Extraction, un'entità è un sintagma nominale, composto da un piccolo numero di token, presente in testo non strutturato. I sistemi esistenti sono costruiti partendo da dataset contenenti entità nominate e definendo caratteristiche sulle entità etichettate e sulle parole circostanti. Questi approcci quindi estraggono soltanto alcune tipologie predefinite di entità, quali ad esempio nomi di Persone, Luoghi, Compagnie e Date. Estendendo l'idea alla base dei modelli così costruiti [15] propone un metodo semi-supervisionato per l'estrazione di Pagine Web Entità all'interno di un sito Web. Una Pagina Entità è una pagina che descrive un'entità nominata (es. pagina web di un professore, di un libro in Amazon, etc.). Partendo da un esempio di pagina Web Entità il sistema sfrutta la struttura ad hyperlink di cui il sito si compone per estrarre tutte le pagine entità dello stesso tipo di quella di esempio. In [14] è proposto Winacs un sistema capace di estrarre entità strutturate dalle pagine Web e costruire un grafo eterogeneo di entità appartenenti al dominio dell'information science, integrando i contenuti presenti nel database DBLP con le entità estratte dal Web.

5.2.1 Classificazione di pagine Web Entità

Nel campo del Named Entity Extraction i modelli di classificazione sono costruiti partendo da dataset e definendo feature sulle entità etichettate e sulle parole circostanti [13]. Alla stessa maniera è possibile creare modelli di classificazione di pagine entità in cui le feature possono essere costruite a partire dalle pagine etichettate e dalle pagine collegate ad esse tramite hyperlink. I vantaggi derivanti da questo approccio sono duplici. In primo luogo a rispetto al testo, dove i dataset di training sono costruiti manualmente, per le entity page questi possono essere definiti utilizzando algoritmi di estrazione semi-supervisionati [16]. Inoltre, da una Entity Page è possibile estrarre un numero maggiore e più informativo di feature rispetto al semplice testo. Queste possono essere costruite dalla struttura e dal contenuto delle pagine entità, degli hyperlink entranti ed uscenti e dalla struttura e dal contenuto delle pagine collegate tramite hyperlink. Per ciò che riguarda il processo di classificazione delle pagine Web esistono in letteratura molteplici approcci capaci di generare modelli predittivi di grande qualità. In [18] gli autori utilizzano Supported Vector Machine (SVM) per classificare pagine web in base a feature estratte dal contenuto delle pagine e dalle url. Liu et al. [10] usano un approccio semi-supervisionato che codifica un insieme di pagine web in un grafo pesato e dopo usano un meccanismo basato su prodotto matriciale per la propagazione delle etichette dai nodi etichettati a quelli non etichettati. I pesi sono ottenuti conisiderando il contenuto delle pagine web (titolo, testo delle ancore, testo in grassetto, etc.) senza usare alcuna informazione sulla struttura delle pagine. Nonostante i numerosi approcci proposti per la classificazione di pagine Web, non esistono ad oggi soluzioni che combinano le assunzioni alla base dei sistemi di Named Entity Extraction e dei sistemi di classificazione di pagine Web, per permettere la classificazione di pagine entità.

6 Approccio al problema

L'approccio che si intende perseguire consiste nella realizzazione di quattro task, quali:

1. Estrazione di web list (record contenenti hyperlink) da pagine web. In questo modo è possibile sfruttare l'ipotesi che che hyperlink raggruppati in liste portano a pagine web semanticamente simili. [6, 8];

2. Scoperta di path paralleli. Formalmente un sito web può essere visto come un grafo orientato in cui i nodi rappresentano le pagine web e

gli archi rappresentano gli hyperlink che collegano le pagine tra loro. L'obiettivo di questo task è dunque quello di effettuare visite nel grafo per individuare path paralleli, ossia percorsi root-link, che appartengono alla stessa web list;

3. Individuazione delle pagine web con path paralleli a quelli dell'entità data come esempio [16, 17];

4. Selezione dei possibili identificativi [15], per tutte le pagine entità.

Queste attività possono poi essere proficuamente utilizzate per l'individuazione di pagine entità. Formalmente, il compito di estrarre pagine web entità, è definito come segue: dati 1) l' url di un sito web e 2) un insieme di url di pagine entità, rappresentati l'insieme di addestramento, basandosi sull'assunzione che le pagine entità dello stesso tipo semantico generalmente condividono inter page path simili all'interno del sito web e DOM path (intra path) simili, l'obiettivo è quello di identificare tutte le pagine entità dello stesso tipo degli esempi positivi forniti in input. In questo modo, dato ad esempio il dominio web www.di.uniba.it e un insieme di pagine web dei professori del dipartimento di informatica (esempi positivi) ed eventualmente un insieme di esempi negativi (es. pagine web dei corsi), è possibile ottenere tutti i professori del dipartimento di Informatica.

7 Ricadute applicative

Di seguito sono analizzate le diverse ricadute applicative della ricerca sopra descritta.

Migliorare la precisione ed il richiamo dei risultati delle ricerche sul web

Gli attuali motori di ricerca sono incapaci di utilizzare i dati contenuti nei Deep Database. L'estrazione delle entità e delle loro proprietà e l'indicizzazione di simili dati può migliorare la qualità delle risposte, rendendoli più robusti ed affidabili.

Question Answering

Le entità estratte nel Web potrebbero essere utilizzate per la realizzazione di sistemi di question answering capaci di fornire risposte soddisfacenti alle interrogazioni degli utenti (es. Chi era il presidente della Repubblica Italiana nel 1956?).

Mashup

Poter aggregare ed integrare automaticamente dati presenti nel web è una

grande sfida di ricerca per la comunità scientifica. Nel campo delle basi di dati questo processo prende il nome di "data integration", mentre nel web viene chiamato "mashup". L'applicazione di tecniche di data integration nel campo delle basi di dati e nel web è completamente differente. Un amministratore di basi di dati può integrare due basi di dati di impiegati con un lungo impiego di tempo ottenendo risultati qualitativamente alti, lo stesso però non si può dire per il web. Ad esempio integrare i libri presenti in Amazon con quelli presenti in AbeBooks è un processo altamente dispendioso e che può portare a risultati mediocri. Ciò è dovuto al fatto che mentre nel campo dei database i dati sono memorizzati con una strutturazione puntuale e definita, nel web i dati sono memorizzati in forma semi-strutturata. Poter estrarre dati strutturati dal web avrebbe quindi una immediata ricaduta applicatica nel campo del data "mashup".

Altri ambiti applicativi

Altre ricadute applicative possono spaziare dal Business and competitive Intelligence [1] alla Homeland Security [11, 12], al crawling di social web plataform [4] alla BioInformatica [9]. Più specificatamente, nel contesto della Homeland Security, risulta indispensabile effettuare l'analisi di news pubblicate su siti web specializzati (quotidiani, agenzie di stampa), o di report pubblicati su siti istituzionali (e.g. Ministero degli Interni, Polizia di Stato e Questure, Arma dei Carabinieri e Procure della Repubblica) può essere utilizzata per migliorare l'attività investigativa e di controllo delle forze dell'ordine. I dati estratti da queste pagine entità possono essere utilizzate per la realizzazione di Crime Map e l'analisi dei trend criminali. Le "Crime Map" sono uno strumento quasi sconosciuto in Italia, ma molto diffuso negli Stati Uniti ed in Canada e permettono di rappresentare geograficamente statistiche sulla criminalità in modo da renderle al tempo stesso accessibili e utili. Il tutto permetterebbe di supportare l'azione investigativa e preventiva del crimine. Un ulteriore contesto applicativo è quello della analisi di dati di natura biomedica che vengono messi a diposizione sia in forma di basi di dati relazionali che su siti web specializzati che contengono sia tabelle che informazioni di natura testuale e che sono legate a particolari entità (geni, patologie, indagini cliniche). Questo permetterebbe di individuare e studiare potenziali correlazioni tra oggetti appartenenti a entità differenti. Infine l'estrazione automatica di pagine entità potrebbe essere utilizzata per confrontare pagine web entità duplicate all'interno di siti web diversi. Questo potrebbe essere sfruttato in diversi domini applicativi quali in ambito giornalistico per individuare news pubblicate su siti differenti (e.g. Corriere.it, lagazzettadelmezzoggiorno.it), ma riguardanti stesso evento o nel campo dell'ecommerce per permettere confronti tra stessi prodotti commer-

ciali venduti su più siti web. Per ciò che riguarda quest'ultimo ambito sono già disponibili applicazioni web (e.g. Segugio.it, skyscanner.com, etc.) capaci di estrarre e confrontare pagine entità con le informazioni associate (e.g. nome prodotto, categoria, prezzo) da più siti web. Questi sistemi realizzano estrazione e confronto di pagine entità simili (libri, alberghi, fotocamere, etc.) attraverso un parsing sintattico dei siti web prescelti. Di conseguenza, in tali applicazioni, è necessario realizzare tanti parser quanti sono i siti web utilizzati per l'estrazione delle pagine entità e ogni modifica alla struttura sintattica dei siti web di origine, e più in particolare alla struttura sintattica di una pagina web entità (DOM), comporta una modifica al corrispondente parser.

8 Fasi del progetto

Primo Anno: Studio del materiale di ricerca di base. In particolare:

1. Studio approfondito del Web Structure mining, Web Content Mining, Web Usage Maning, NLP e Named Entity Extraction.
2. Approfondimento delle tematiche relative all'estrazione di liste all'interno di pagine Web.
3. Studio della letteratura relativa alla scoperta di pattern frequenti.

Secondo Anno: Sintesi e realizzazione di algoritmi nel campo del Web Mining

1. Analisi e sintesi di algoritmi per la scoperta di liste Intra- e Inter-pagina;
2. Sintesi di algoritmi per la scoperta di pattern frequenti;
3. Sintesi di modelli di scoperta per Pagine Web Entità utilizzando tecniche di pattern matching e scoperta di pattern frequenti;
4. Pubbicazione dei risultati conseguiti in riviste e conferenze di carattere internazionale;
5. Stage presso università straniera e confronto con l'attività svolta presso gruppi di ricerca con obiettivi affini.

Terzo Anno : Applicazione al dominio applicativo scelto e sviluppo della tesi di dottorato

1. Stage presso università straniera e confronto con l'attività svolta presso gruppi di ricerca con obiettivi affini;
2. Ottimizzazione del metodo proposto e implementazione di strumenti di supporto al test sul dominio applicativo scelto;
3. Analisi dei risultati sperimentali;
4. Sviluppo della tesi di dottorato.

Attvità	Primo Anno			Secondo Anno			Terzo Anno		
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 1	Trimestre 2	Trimestre 3
Studio approfondito del Web Structure mining, Web Content Mining, Web Usage Maning, NLP e Named Entity Extraction	■	■	■						
Approfondimento delle tematiche relative all'estrazione di liste all'interno di pagine Web	■	■							
Studio della letteratura relativa alla scoperta di pattern frequenti		■	■						
Analisi e sintesi di algoritmi per la scoperta di liste Intra- e Inter-pagina				■	■				
Sintesi di algoritmi per la scoperta di pattern frequenti				■	■				
Sintesi di modelli di scoperta per Pagine Web Entità utilizzando tecniche di pattern matching e scoperta di pattern frequenti					■	■			
Pubbicazione dei risultati conseguiti in riviste e conferenze di carattere internazionale						■			
Stage presso università straniera e confronto con l'attività svolta presso gruppi di ricerca con obiettivi affini					■	■	■	■	
Ottimizzazione del metodo proposto e implementazione di strumenti di supporto al test sul dominio applicativo scelto							■	■	■
Analisi dei risultati sperimentali									■
Sviluppo della tesi di dottorato								■	■

9 Valutazione dei risultati

In letteratura esistono numerose misure utilizzate per la valutazione dei dati strutturati estratti. In particolare le misure più frequentemente utilizzate risultano precisione richiamo e F-measure. Per la valutazione di questa ricerca verranno utilizzate le misure sopra descritte. In particolare queste saranno molto utili per valutare la qualità del processo di estrazione delle liste Web, la scoperta delle pagine web entità, l'associazione delle entità scoperte a record in basi di dati e la valutazione dei modelli per la classificazione delle pagine web entità. Buona parte della valutazione di queste misure verrà fatta costruendo dataset ad hoc e valutando manualmente i risultati dei metodi implementati con quelli attesi.

10 Eventuali referenti esterni al Dipartimento

Tim Weninger
Siebel Center for Computer Science Room 2119
201 N. Goodwin Avenue
Urbana, IL 61801
Phone: +16206647020
Email: weninge1@illinois.edu

References

[1] Robert Baumgartner, Georg Gottlob, and Marcus Herzog. Scalable web data extraction for online market intelligence. *Proc. VLDB Endow.*, 2(2):1512–1523, August 2009.

[2] Michael J. Cafarella, Alon Halevy, and Jayant Madhavan. Structured data on the web. *Commun. ACM*, 54(2):72–79, February 2011.

[3] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549, August 2008.

[4] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Crawling facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 52:1–52:8, New York, NY, USA, 2011. ACM.

[5] Tao Cheng and Kevin Chen-Chuan Chang. Beyond pages: supporting efficient, scalable entity search with dual-inversion index. In *Proceedings of the 13th International Conference on Extending Database Technology*, EDBT '10, pages 15–26, New York, NY, USA, 2010. ACM.

[6] Valter Crescenzi, Paolo Merialdo, and Paolo Missier. Clustering web pages based on their structure. *Data Knowl. Eng.*, 54(3):279–299, 2005.

[7] Hazem Elmeleegy, Jayant Madhavan, and Alon Halevy. Harvesting relational tables from lists on the web. *The VLDB Journal*, 20(2):209–226, April 2011.

[8] Fabio Fumarola, Tim Weninger, Rick Barber, Donato Malerba, and Jiawei Han. Hylien: a hybrid approach to general list extraction on the web. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 35–36, New York, NY, USA, 2011. ACM.

[9] Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman. The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771, October 2011.

[10] Wei Liu, Jun Wang 0006, and Shih-Fu Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624–2638, 2012.

[11] Colleen McCue. Data mining and predictive analytics in public safety and security. *IT Professional*, 8(4):12–18, July 2006.

[12] Shyam Varan Nath. Crime pattern detection using data mining. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology*, WI-IATW '06, pages 41–44, Washington, DC, USA, 2006. IEEE Computer Society.

[13] Sunita Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, March 2008.

[14] Tim Weninger, Marina Danilevsky, Fabio Fumarola, Joshua Hailpern, Jiawei Han, Thomas J. Johnston, Surya Kallumadi, Hyungsul Kim, Zhijin Li, David McCloskey, Yizhou Sun, Nathan E. TeGrotenhuis, Chi Wang, and Xiao Yu. Winacs: construction and analysis of web-based computer science information networks. In *Proceedings of the*

2011 ACM SIGMOD International Conference on Management of data, SIGMOD '11, pages 1255–1258, New York, NY, USA, 2011. ACM.

[15] Tim Weninger, Fabio Fumarola, Jiawei Han, and Donato Malerba. Mapping web pages to database records via link paths. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1637–1640, New York, NY, USA, 2010. ACM.

[16] Tim Weninger, Fabio Fumarola, Cindy Xide Lin, Rick Barber, Jiawei Han, and Donato Malerba. Growing parallel paths for entity-page discovery. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 145–146, New York, NY, USA, 2011. ACM.

[17] Tim Weninger and Jiawei Han. Exploring structure and content on the web: extraction and integration of the semi-structured web. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 779–780, New York, NY, USA, 2013. ACM.

[18] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. Pebl: Web page classification without negative examples. *IEEE Trans. on Knowl. and Data Eng.*, 16(1):70–81, January 2004.