Knowledge Harvesting from Text and Web Sources

Fabian Suchanek and Gerhard Weikum
Max Planck Institute for Informatics
66123 Saarbruecken, Germany
E-Mail: suchanek@mpi-inf.mpg.de, weikum@mpi-inf.mpg.de

***Abstract*—The proliferation of knowledge-sharing communities such as Wikipedia and the progress in scalable information extraction from Web and text sources has enabled the automatic construction of very large knowledge bases. Recent endeavors of this kind include academic research projects such as DBpedia, KnowItAll, Probase, ReadTheWeb, and YAGO, as well as industrial ones such as Freebase and Trueknowledge. These projects provide automatically constructed knowledge bases of facts about named entities, their semantic classes, and their mutual relationships. Such world knowledge in turn enables cognitive applications and knowledge-centric services like disambiguating natural-language text, deep question answering, and semantic search for entities and relations in Web and enterprise data. Prominent examples of how knowledge bases can be harnessed include the Google Knowledge Graph and the IBM Watson question answering system. This tutorial presents state-of-the-art methods, recent advances, research opportunities, and open challenges along this avenue of knowledge harvesting and its applications.**

I. Motivation

Knowledge harvesting from Web and text sources has become a major research avenue in the last five years. It is the core methodology for the automatic construction of large knowledge bases [1], [2], going beyond manually compiled knowledge collections like Cyc [13], WordNet [9], and a variety of ontologies [17]. Salient projects with publicly available resources include KnowItAll [7], [4], [8], ConceptNet (MIT) [16], DBpedia [3], Freebase [5], NELL [6], WikiTaxonomy [15], and YAGO [18], [11] (our own project at the Max Planck Institute for Informatics). Commercial interest has been strongly growing, with evidence by projects like the Google Knowledge Graph, the EntityCube (Renlifang) project at Microsoft Research [14], and the use of public knowledge bases for type coercion in IBM's Watson project [12].

These knowledge bases contain many millions of entities, organized in hundreds to hundred thousands of semantic classes, and hundred millions of relational facts between entities. All this is typically represented in the form of RDF-style subject-predicate-object (SPO) triples. Moreover, knowledge resources can be semantically interlinked via owl:sameAs triples at the entity level, contributing to the Web of Linked Open Data (LOD) [10]. For example, a knowledge base may contain the following triples:

```
(Ennio_Morricone type composer)
(Elvis_Presley type singer)
(composer subclassOf musician)
(composer subclassOf musician)
(Ennio_Morricone bornIn Rome)
(Elvis_Presley buriedIn Graceland)
(Ennio_Morricone wonPrize Academy_Award)
(Elvis_Presley wonPrize Grammy)
(Maestro_Morricone sameAs Ennio_Morricone)
(Elvis_Presley hasName "The King")
```

Knowledge bases are a key asset for many kinds of intelligent applications, including question answering, reasoning tasks, semantic search over web contents and social media, contents analytics, and more.

Large knowledge bases are typically built by mining and distilling information from sources like Wikipedia which offer high-quality semi-structured elements (infoboxes, categories, tables, lists), but many projects also tap into extracting knowledge from arbitrary Web pages and natural-language texts. Despite great advances in these regards, there are still many challenges regarding the scale of the methodology and the scope and depth of the harvested knowledge:

- covering more entities beyond Wikipedia and discovering newly emerging entities,
- increasing the number of facts about entities and extracting more interesting relationship types in an open manner,
- capturing the temporal scope of relational facts,
- tapping into multilingual inputs such as Wikipedia editions in many different languages,
- extending fact-oriented knowledge bases with commonsense knowledge and (soft) rules,
- detecting and disambiguating entity mentions in natural language text, and
- large-scale sameAs linkage across many knowledge and data sources.

This 90-minute tutorial will give an overview on knowledge harvesting and will discuss hot topics in this field, pointing out research opportunities and open challenges. As the relevant literature is widely dispersed across different communities, we also venture into the neighboring venues of Web Mining, Artificial Intelligence, Natural Language Processing, Semantic Web, and Data Management. The presentation will be structured according to the following sections and subsections. The first part covers the realm of knowledge harvesting. The second part covers knowledge linking.

References

[1] AKBC 2010: First Int. Workshop on Automated Knowledge Base Construction, Grenoble, 2010, http://akbc.xrce.xerox.com/

[2] AKBC-WEKEX 2012: The Knowledge Extraction Workshop at NAACL-HLT, Montreal, 2012, http://akbcwekex2012.wordpress.com/
[3] S. Auer, C. Bizer, et al.: DBpedia: A Nucleus for a Web of Open Data. ISWC 2007
[4] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. IJCAI 2007
[5] K.D. Bollacker et al.: Freebase: a collaboratively created graph database for structuring human knowledge. SIGMOD 2008
[6] A. Carlson et al.: Toward an Architecture for Never-Ending Language Learning. AAAI 2010
[7] O. Etzioni et al.: Unsupervised named-entity extraction from the Web: An experimental study. Artif. Intell. 165(1), 2005
[8] A. Fader. S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction, EMNLP 2011
[9] C. Fellbaum, G. Miller (Eds.): WordNet: An Electronic Lexical Database, MIT Press, 1998
[10] T. Heath, C. Bizer: Linked Data: Evolving the Web into a Global Data Space, 2011
[11] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum: YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Artif. Intell. 2012
[12] IBM Journal of Research and Development 56(3/4), Special Issue on "This is Watson", 2012
[13] D.B. Lenat: CYC: A Large-Scale Investment in Knowledge Infrastructure. CACM 38(11), 1995
[14] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, W.Y. Ma: Web Object Retrieval. WWW 2007
[15] S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007
[16] R. Speer, C. Havasi: Representing General Relational Knowledge in ConceptNet 5, LREC 2012
[17] S. Staab, R. Studer: Handbook on Ontologies, 2nd Edition, Springer, Morgan & C. 2009
[18] F.M. Suchanek, G. Kasneci, G. Weikum: Yago: a core of semantic knowledge. WWW 2007

II. Knowledge Harvesting

A. Harvesting Entities and Classes

Every entity in a knowledge base (such as `Elvis_Presley`) belongs to one or multiple classes (such as `singer`). These classes are organized into a taxonomy, where more special classes (such as `singer`) are subsumed by more general classes (such as `person`). WordNet [3] already contains a large number of classes. Wikipedia, in contrast, contains a large number of entities. By intelligently mapping Wikipedia categories to WordNet, projects like Yago [9] and WikiTaxonomy [7] have managed to build very large taxonomies.

Alternative work has been pursuing the goal of populating classes ab initio, that is, without resorting to Wikipedia-style sources. Set-expansion methods, typically bootstrapped with a few seed instances, exploit special patterns in natural-language sentences or Web tables (e.g., [1], [2], [4], [5], [6], [11]). The results are usually smaller and noisier than the above knowledge bases. However, for capturing class instances that cannot be found in Wikipedia, Web-based methods are indispensable. Harvesting long-tail entities (e.g., electronics products, or less notable musicians, scientists, etc.) keeps being a demanding research issue.

References

[1] E. Alfonseca, M. Pasca, E. Robledo-Arnuncio: Acquisition of instance attributes via labeled and related instances. SIGIR 2010
[2] B.B. Dalvi, W.W. Cohen, J. Callan: WebSets: extracting sets of entities from the web using unsupervised information extraction. WSDM 2012
[3] C. Fellbaum, G. Miller (Eds.): WordNet: An Electronic Lexical Database, MIT Press, 1998
[4] M.A Hearst: Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992
[5] Z. Kozareva, E.H. Hovy: A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web. EMNLP 2010
[6] M. Pasca: Ranking Class Labels Using Query Sessions. ACL 2011
[7] S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007
[8] S.P. Ponzetto, M. Strube: Taxonomy induction based on a collaboratively built knowledge repository. Artif. Intell. 175(9-10), 2011
[9] F.M. Suchanek, G. Kasneci, G. Weikum: Yago: a core of semantic knowledge. WWW 2007
[10] P.P. Talukdar, F. Pereira: Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL 2010
[11] R. Wang, W.W. Cohen: Language-independent set expansion of named entities using the web. ICDM 2007
[12] F. Wu, D.S. Weld: Automatically refining the wikipedia infobox ontology. WWW 2008
[13] W. Wu, H. Li, H. Wang, K.Q. Zhu: Probase: a probabilistic taxonomy for text understanding. SIGMOD 2012

B. Harvesting Relational Facts

For factual knowledge about entities, most work has focused on instances of binary relations, disregarding higher-arity cases. Examples are:

```
(Ennio_Morricone composed Ecstasy_of_Gold)
(Elvis_Presley sang In_the_Ghetto).
```

Gathering and cleaning such facts involves finding pairs of entities, in text or semi-structured tables, and inferring which relationships hold between them. To this end, methods from pattern matching (e.g., regular expressions), computational linguistics (e.g., dependency parsing), statistical learning (e.g., factor graphs and MLN's), and logical consistency reasoning (e.g., weighted MaxSat or ILP solvers) are combined in many interesting ways.

Terminological diversity in the ways how relations are referred to (e.g., `bornIn` versus `birthplace`) need to be reconciled automatically, but this issue becomes easier with wider adoption of standardized vocabularies like `schema.org` or (class-specific) infobox templates in Wikipedia. Hot research also addresses the scalability challenge, robustness to noise, and the ability to tap into the long tail of facts while minimizing the amount of human supervision. [8], [22], [25] survey these methods; further references on original work are given below.

References

[1] E. Agichtein, L. Gravano: Snowball: extracting relations from large plain-text collections. ACM DL 2000
[2] S. Auer, C. Bizer, et al.: DBpedia: A Nucleus for a Web of Open Data. ISWC 2007
[3] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. IJCAI 2007
[4] S. Brin: Extracting Patterns and Relations from the World Wide Web. WebDB 1998
[5] M.J. Cafarella: Extracting and Querying a Comprehensive Web Database. CIDR 2009
[6] A. Carlson et al.: Toward an Architecture for Never-Ending Language Learning. AAAI 2010
[7] P. Cimiano, J. Völker: Text2Onto. NLDB 2005: 227-238
[8] A. Doan, L. Gravano, R. Ramakrishnan, S. Vaithyanathan (Eds.): ACM SIGMOD Record 37(4), Special issue on managing information extraction, 2008

[9] P. Domingos, D. Lowd: Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool 2009
[10] O. Etzioni et al.: Unsupervised named-entity extraction from the Web: An experimental study. Artif. Intell. 165(1), 2005
[11] G.Gottlob et al.: The Lixto Data Extraction Project - Back and Forth between Theory and Practice. PODS 2004
[12] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum: YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Artif. Intell. 2012
[13] R. Hoffmann, C. Zhang, D.S. Weld: Learning 5000 Relational Extractors. ACL 2010
[14] N. Kushmerick, D.S. Weld, R.B. Doorenbos: Wrapper Induction for Information Extraction. IJCAI 1997
[15] Y. Li, F. Reiss, L. Chiticariu: SystemT: A Declarative Information Extraction System. ACL 2011
[16] A. Machanavajjhala et al.: Collective extraction from heterogeneous web lists. WSDM 2011
[17] B. Marthi, B. Milch, S. Russell, First-Order Probabilistic Models for Information Extraction. IJCAI 2003
[18] N. Nakashole et al.: Scalable knowledge harvesting with high precision and high recall. WSDM 2011
[19] F. Niu, C. Zhang, C. Re, J. Shavlik: DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference, VLDS Workshop 2012
[20] M. Palmer, D. Gildea, N. Xue: Semantic Role Labeling Morgan & Claypool 2010
[21] S. Riedel, L. Yao, A. McCallum: Modeling relations and their mentions without labeled text. ECML 2010
[22] S. Sarawagi: Information Extraction. Foundations & Trends in Databases 1(3), 2008.
[23] F.M. Suchanek et al.: SOFIE: a self-organizing framework for information extraction. WWW 2009
[24] P. Venetis, A. Halevy, J. Madhavan, et al.: Recovering Semantics of Tables on the Web, PVLDB 2011
[25] G. Weikum, M. Theobald: From information to knowledge: harvesting entities and relationships from web sources. PODS 2010
[26] J. Zhu et al: StatSnowball: a statistical approach to extracting entity relationships. WWW 2009

C. Open-Domain Extraction

No knowledge base can ever be fully complete. New, so far uncovered entities become important or come into existence. This is considered only very partially in Wikipedia-centric knowledge harvesting. Moreover, many knowledge bases focus on a prespecified set of relations, often oriented towards frequent or particularly clean properties in Wikipedia infoboxes. The number of relation types in DBpedia, Freebase, NELL, and YAGO ranges from about a hundred to several thousands, thus missing out on many interesting relationships.

Open information extraction (IE) [1] aims to close this gap, by aggressively tapping into noun phrases as entity candidates and verbal phrases as prototypic patterns for relations. Example "factoids" or "statements" from this approach could be:

```
("Elvis" "alive and seen in" "Tibet")
("Tarantino" "picked music by" "the maestro")
```

While increasing recall this way, the result tends to be noisy and degrades precision. Thus, this is an active research area of great importance.

References

[1] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. IJCAI 2007
[2] A. Fader. S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction, EMNLP 2011
[3] C. Wang, J. Fan, A. Kalyanpur, D. Gondek: Relation Extraction with Relation Topics. EMNLP 2011
[4] T. Mohamed, E.R. Hruschka, T.M. Mitchell: Discovering Relations between Noun Categories. EMNLP 2011
[5] N. Nakashole, G. Weikum, F. Suchanek: PATTY: A Taxonomy of Relational Patterns with Semantic Types. EMNLP 2012
[6] M. Nickel, V. Tresp, H.-P. Kriegel: Factorizing YAGO: scalable machine learning for linked data. WWW 2012
[7] L. Yao, S. Riedel, A. McCallum: Unsupervised Relation Discovery with Sense Disambiguation. ACL 2012

D. Harvesting Temporal, Multilingual, Visual, and Commonsense Knowledge

Relationships like presidents of countries, CEOs of companies, and even spouses change from time to time. Thus, a rich knowledge base should be aware of the timespans during which certain facts hold and of the timepoints at which certain events happen. For example, we would like to capture:

```
id1:(Elvis_Presley marriedTo
Priscilla_Presley)
id2:(id1 validDuring [1967,1973])
id3:(Ennio_Morricone wonPrize Academy_Award)
id4:(id3 happenedOn 25-February-2007)
```

where SPO-triples can be reified (via identifiers) in other triples about temporal properties. This calls for a temporal dimension [17] in the process of knowledge harvesting – a recently tackled and widely open research challenge [4], [6], [13], [14], [19], [20].

There is also a multilingual dimension in knowledge harvesting: aiming to capture names and surface expressions for entities, classes, and general concepts from many different languages and cultural contexts [1], [2], [8], [9], [12]. Visual knowledge like images for entities and classes, and their properties (e.g., typical shapes, sizes, geometric features) are another direction to pursue [3], [10], [16]. Here, ImageNet is the most prominent project that populates ten thousands of WordNet classes with photos [3].

Finally, a dimension that complements factual knowledge is commonsense knowledge: properties and rules that every child knows but are hard to acquire by a computer. Machines should have formal representations of statements such as:

```
(pasta hasTexture al_dente)
(steak hasTexture tender)
∀x:(x type composer) ⇒∃y:(x playsInstrument y)
∀x,y:(x type deadPeople) ⇒ ¬(x sightedAt y)
```

References

[1] G. de Melo, G. Weikum: Towards a universal wordnet by learning from combined evidence. CIKM 2009
[2] G. de Melo, G. Weikum: MENTA: inducing multilingual taxonomies from wikipedia. CIKM 2010
[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. CVPR 2009
[4] G. Garrido, A. Penas, B. Cabaleiro, A. Rodrigo: Temporally Anchored Relation Extraction, ACL 2012
[5] N. Lao, T.M. Mitchell, W.W. Cohen: Random Walk Inference and Learning in A Large Scale Knowledge Base. EMNLP 2011
[6] X. Ling, D.S. Weld: Temporal Information Extraction. AAAI 2010
[7] C. Matuszek et al.: Searching for Common Sense: Populating Cyc from the Web. AAAI 2005
[8] V. Nastase et al.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. LREC 2010

[9] R. Navigli, S. Ponzetto: BabelNet: Building a Very Large Multilingual Semantic Network. ACL 2010
[10] M. Rohrbach et al.: What helps where - and why? Semantic relatedness for knowledge transfer. CVPR 2010
[11] R.Speer, C.Havasi, H.Surana: Using Verbosity: Common Sense Data from Games with a Purpose. FLAIRS 2010
[12] V.I. Spitkovsky, A.X. Chang: A Cross-Lingual Dictionary for English Wikipedia Concepts. LREC 2012
[13] P.P. Talukdar, D.T. Wijaya, T. Mitchell: Coupled temporal scoping of relational facts. WSDM 2012
[14] P.P. Talukdar, D. Wijaya, T. Mitchell: Acquiring Temporal Constraints between Relations. CIKM 2012
[15] N. Tandon, G. de Melo, G. Weikum: Deriving a Web-Scale Common Sense Fact Database. AAAI 2011
[16] B. Taneva, M. Kacimi, G. Weikum: Gathering and ranking photos of named entities with high precision, high recall, and diversity. WSDM 2010
[17] M. Verhagen, I. Mani, R. Sauri, J. Littman, R. Knippen, S. B. Jang, A. Rumshisky, J. Phillips, J. Pustejovsky. Automating temporal annotation with tarsqi. In ACL, 2005
[18] J. Völker, P. Hitzler, P. Cimiano: Acquisition of OWL DL Axioms from Lexical Resources. ESWC 2007
[19] Y. Wang, L. Qu, B. Yang, M. Spaniol, G. Weikum: Harvesting Facts from Textual Web Sources by Constrained Label Propagation. CIKM 2011
[20] Y. Wang, M. Dylla, M. Spaniol, G. Weikum: Coupling Label Propagation and Constraints for Temporal Fact Extraction. ACL 2012

III. Knowledge Linking

A. Named-Entity Disambiguation

When extracting knowledge from text or tables, entities are first seen only in surface form: by names (e.g., "Elvis") or phrases (e.g., "the late rock and roll idol"). Such entity mentions are often highly ambiguous; mapping them to canonicalized entities registered in a knowledge base is known as the task of named-entity disambiguation, NED for short. State-of-the-art NED methods [13], [9], [7] combine context similarity between the surroundings of a mention and salient phrases associated with an entity, with coherence measures for two or more entities co-occurring together. Although these principles are well understood, NED remains an active research area towards improving robustness, scalability, and coverage.

References

[1] R.C. Bunescu, M. Pasca: Using Encyclopedic Knowledge for Named Entity Disambiguation. EACL 2006
[2] S. Cucerzan: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP 2007
[3] M. Dredze et al.: Entity Disambiguation for Knowledge Base Population. COLING 2010
[4] P. Ferragina, U. Scaiella: TAGME: on-the-fly annotation of short text fragments. CIKM 2010
[5] J.R. Finkel, T. Grenager, C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. ACL 2005
[6] X. Han, L. Sun, J. Zhao: Collective entity linking in web text: a graph-based method. SIGIR 2011
[7] J. Hoffart, M. A. Yosef, I. Bordino, et al.: Robust Disambiguation of Named Entities in Text. EMNLP 2011
[8] J. Hoffart et al.: KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. CIKM 2012
[9] S. Kulkarni et al.: Collective annotation of Wikipedia entities in web text. KDD 2009
[10] G.Limaye et al: Annotating and Searching Web Tables Using Entities, Types and Relationships. PVLDB 2010
[11] T. Lin et al.: No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. EMNLP 2012
[12] X. Ling, D.S. Weld: Fine-Grained Entity Recognition. AAAI 2012
[13] D.N. Milne, I.H. Witten: Learning to link with wikipedia. CIKM 2008
[14] R. Navigli: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 2009
[15] A. Rahman, V. Ng: Coreference Resolution with World Knowledge. ACL 2011
[16] L. Ratinov et al.: Local and Global Algorithms for Disambiguation to Wikipedia. ACL 2011
[17] S. Singh, A. Subramanya, F.C.N. Pereira, A. McCallum: Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models. ACL 2011

B. Entity Linkage

Even when entities are explicitly marked in structured or semi-structured data (e.g., RDF triples), the problem arises to tell whether two entities are the same or not. This is a variant of the classical record-linkage problem (aka. entity matching, entity resolution, entity de-duplication), with the additional requirement to map also relations and schema information. For knowledge bases and Linked Open Data [5], it is of particular interest because of the need for generating and maintaining owl:sameAs information across knowledge resources. Surveys on record-linkage methods are given by [3], [7], [10]. Referencs on recent work, often using statistical learning and graph algorithms, are given below.

References

[1] I.Bhattacharya, L. Getoor: Collective Entity Resolution in Relational Data. TKDD 1(1), 2007
[2] C. Böhm et al.: LINDA: Distributed Web-of-Data-Scale Entity Matching. CIKM 2012
[3] A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios: Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng. 19(1): 1-16 (2007)
[4] R. Hall, C.A. Sutton, A. McCallum: Unsupervised deduplication using cross-field dependencies. KDD 2008
[5] T. Heath, C. Bizer: Linked Data: Evolving the Web into a Global Data Space. Morgan&Claypool, 2011
[6] A.Hogan et al.: Scalable and distributed methods for entity matching. J. Web Sem. 10, 2012
[7] H. Köpcke et al.: Evaluation of entity resolution approaches on real-world match problems. PVLDB 2010
[8] J.Li, J.Tang, Y.Li, Q.Luo: RiMOM: A dynamic multistrategy ontology alignment framework. TKDE 21(8), 2009
[9] S. Melnik, H. Garcia-Molina, E. Rahm: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. ICDE 2002
[10] F. Naumann, M. Herschel: An Introduction to Duplicate Detection. Morgan&Claypool, 2010
[11] T. Nguyen et al.: Multilingual Schema Matching for Wikipedia Infoboxes. PVLDB 2012
[12] V. Rastogi, N. Dalvi, M. Garofalakis: Large-Scale Collective Entity Matching. PVLDB 2011
[13] P. Singla, P. Domingos: Entity Resolution with Markov Logic. ICDM 2006
[14] F. Suchanek et al.: PARIS: Probabilistic Alignment of Relations, Instances, and Schema. PVLDB 2012
[15] J. Wang, T. Kraska, M. Franklin, J. Feng: CrowdER: Crowdsourcing Entity Resolution. PVLDB 2012
[16] Z. Wang, J. Li, Z. Wang, J. Tang: Cross-lingual knowledge linking across wiki knowledge bases. WWW 2012
[17] S.E. Whang, H. Garcia-Molina: Joint Entity Resolution. ICDE 2012