sisinflab: an ensemble of supervised and unsupervised strategies for the NEEL-IT challenge at Evalita 2016

Vittoria Cozza, Wanda La Bruna, Tommaso Di Noia
Polytechnic University of Bari
via Orabona, 4, 70125, Bari, Italy
{vittoria.cozza, wanda.labruna, tommaso.dinoia}@poliba.it

Abstract

English. This work presents the solution adopted by the sisinflab team to solve the task NEEL-IT (Named Entity rEcognition and Linking in Italian Tweets) at the Evalita 2016 challenge. The task consists in the annotation of each named entity mention in a Twitter message written in Italian, among *characters*, *events*, *people*, *locations*, *organizations*, *products* and *things* and the eventual linking when a corresponding entity is found in a knowledge base (e.g. DBpedia). We faced the challenge through an approach that combines unsupervised methods, such as DBpedia Spotlight and word embeddings, and supervised techniques such as a CRF classifier and a Deep learning classifier.

Italiano. *Questo lavoro presenta la soluzione del team sisinflab al task NEEL-IT (Named Entity rEcognition and Linking in Italian Tweets) di Evalita 2016. Il task richiede il riconoscimento e l'annotazione del testo di un messaggio di Twitter in Italiano con entità nominate quali personaggi, eventi, persone, luoghi, organizzazioni, prodotti e cose e eventualmente l'associazione di queste entità con la corrispondente risorsa in una base di conoscenza quale, DBpedia. L'approccio proposto combina metodi non supervisionati quali DBpedia Spotlight e i word embeddings, e tecniche supervisionate basate su due classificatori di tipo CRF e Deep learning.*

1 Introduction

In the interconnected world we live in, the information encoded in Twitter streams represents a valuable source of knowledge to understand events, trends, sentiments as well as user-behaviors. While processing these small text messages a key role is played by the entities which are named within the Tweet. Indeed, whenever we have a clear understanding of the entities involved in a context, a further step can be done by semantically enriching them via side information available, e.g., in the Web. To this aim, pure NER techniques show their limits as they are able to identify the category an entity belongs to but they cannot be used to find further information that can be used to enrich the description of the identified entity and then of the overall Tweet. This is the point where Entity Linking starts to play its role. Dealing with Tweets, as we have very short messages and texts with little context, the challenge of Named Entity Linking is even more tricky as there is a lot of noise and very often text is semantically ambiguous. A number of popular challenges on the matter currently exists, as those included in the SemEval series on the evaluations of computational semantic analysis systems[1] for English, the CLEF initiative[2] that provides a cross-language evaluation forum or Evalita[3] that aims to promote the development of language and speech technologies for the Italian language.

Several state of the art solutions have been proposed for entity extraction and linking to a knowledge base (Shen et al., 2015) and many of them make use of the datasets available as Linked (Open) Data such as DBpedia or Wikidata (Gangemi, 2013). Most of these tools expose the best performances when used with long texts. Anyway, those approaches that perform well on newswire domain do not work as well in a microblog scenario. As analyzed in (Derczynski et al., 2015), conventional tools (i.e., those trained

[1] https://en.wikipedia.org/wiki/SemEval
[2] http://www.clef-initiative.eu/
[3] http://www.evalita.it/

on newswire) perform poorly in this genre, and thus microblog domain adaptation is crucial for good NER. However, when compared to results typically achieved on longer news and blog texts, state-of-the-art tools in microblog NER still reach bad performance. Consequently, there is a significant proportion of missed entity mentions and false positives. In (Derczynski et al., 2015), the authors also show which tools are possible to extend and adapt to Twitter domain, for example DBpedia Spotlight.The advantage of Spotlight is that it allows users to customize the annotation task. In (Derczynski et al., 2015) the authors show Spotlight achieves 31.20% of F1 over a Twitter dataset.

In this paper we present the solution we propose for the NEEL-IT task (Basile et al., 2016b) of Evalita 2016 (Basile et al., 2016a). The task consists of annotating each named entity mention (characters, events, people, locations, organizations, products and things) in an Italian Tweet text, linking it to DBpedia nodes when available or labeling it as NIL entity otherwise. The task consists of three consecutive steps: (1) extraction and typing of entity mentions within a tweet; (2) linking of each textual mention of an entity to an entry in the canonicalized version of DBpedia 2015-10 representing the same "real world" entity, or NIL in case such entry does not exist; (3) clustering of all mentions linked to NIL. In order to evaluate the results the TAC KBP scorer[4] has been adopted. Our team solutions faces the above mentioned challenges by using an ensemble of state of the art approaches.

The remainder of the paper is structured as follows: in Section 2 we introduce our strategy that combines DBpedia Spotlight-based and a machine learning-based solutions, detailed respectively in Section 2.1 and Section 2.2. Section 3 reports and discusses the challenge results.

2 Description of the system

The system proposed for entity boundary and type extraction and linking is an ensemble of two strategies: a DBpedia Spotligth[5]-based solution and a machine learning-based solution, that exploits Stanford CRF[6] and DeepNL[7] classifiers. Before applying both approaches we pre-processed the tweets used in the experiments, by doing: (1) data cleaning consisting of replacing URLs with the keyword `URL` as well emoticons with `EMO`; This has been implemented with ad hoc rules; (2) sentence splitter and tokenizer, implemented by the well known linguistic pipeline available for the Italian language: "openNLP"[8], with its corresponding binary models[9].

2.1 Spotlight-based solution

DBpedia Spotlight is a well known tool for entity linking. It allows a user to automatically annotate mentions of DBpedia resources in unstructured textual documents.

- Spotting: recognizes in a sentence the phrases that may indicate a mention of a DBpedia resource.
- Candidate selection: maps the spotted phrase to resources that are candidate disambiguations for that phrase.
- Disambiguation: uses the context around the spotted phrase to decide for the best choice amongst the candidates.

In our approach we applied DBpedia Spotlight (J. et al., 2013) in order to identify mention boundaries and link them to a DBpedia entity. This process makes possible to identify only those entities having an entry in DBpedia but it does not allow a system to directly identify entity types. According to the challenge guideline we required to identify entities that fall into 7 categories: `Thing`, `Product`, `Person`, `Organization`, `Location`, `Event`, `Character` and their subcategories. In order to perform this extra step, we used the "type detection" module, as shown in Figure 1 which makes use of a `SPARQL` query to extract ontological information from DBpedia. In detail we match the name of returned classes associated to an entity with a list of keywords related to the available taxonomy: Place, Organization (or Organisation), Character, Event, Sport, Disease, Language, Person, Music Group, Software, Service, Film, Television, Album, Newspaper, Electronic Device. There are three possible outcomes: no match, one match, more than one match. In the case we find no match we discard the entity while in case we have more than one match we choose

[4] https://github.com/wikilinks/neleval/wiki/Evaluation

[5] urlhttps://github.com/dbpedia-spotlight/dbpedia-spotlight

[6] http://nlp.stanford.edu/software/CRF-NER.shtml

[7] https://github.com/attardi/deepnl

[8] `https://opennlp.apache.org/index.html`

[9] `https://github.com/aciapetti/opennlp-italian-models/tree/master/models/it`

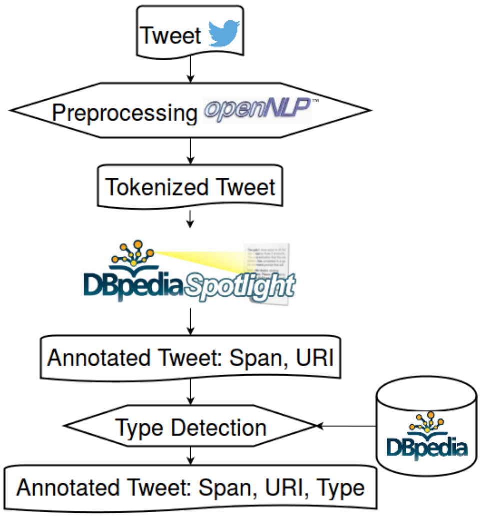

Figure 1: Spotlight based solution

the most specific one, according the NEEL-IT taxonomy provided for the challenge. Once we have an unique match we return the entity along with the new identified type.

Since DBpedia returns entities classified with reference to around 300 categories, we process the annotated resources through the Type Detection Module to discard all those entities not falling in any of the categories of the NEEL-IT taxonomy. Over the test set, after we applied the Ontology-based type detection module, we discarded 16.9% of returned entities. In this way, as shown in Figure 1, we were able to provide an annotation (span, uri, type) as required by the challenge rules.

2.2 Machine learning based solution

As summarized in Figure 2, we propose an ensemble approach that combines unsupervised and supervised techniques by exploiting a large dataset of unannotated tweets, Twita (Basile and Nissim, 2013) and the DBpedia knowledge base. We used a supervised approach for entity name boundary and type identification, that exploits the challenge data. Indeed the challenge organizers provided a training dataset consisted of 1,000 tweets in italian, for a total of 1,450 sentences. The training dataset were annotated with 801 gold annotations. Overall 526 over 801 were entities linked to a unique resource on DBpedia, the other were linked to 255 NIL clusters. We randomly split this training dataset in `new_train` (70%) and `validation` (30%) set. In Table 1 we show the number of mentioned entities classified with reference to their corresponding categories. We then pre-processed the `new_train` and the `validation` sets with the approach shortly described in Section 2 thus obtaining a corpus in IOB2-notation. The annotated corpus was then adopted for training and evaluating two classifiers, Stanford CRF(Finkel et al., 2005) and DeepNL(Attardi, 2015) as shown in Figure 2, in order to detect the span and the type of entity mention in the text.

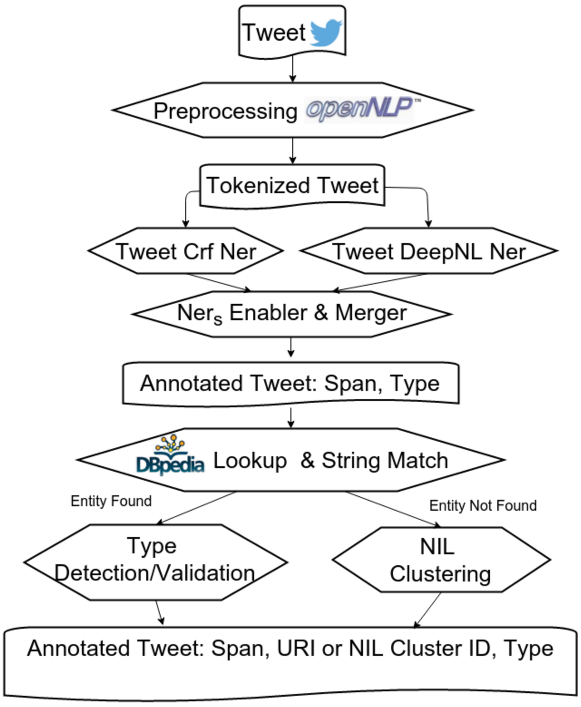

Figure 2: Machine Learning based solution

The module *NERs Enabler & Merger* aims to enabling the usage of one or both classifiers. When them both are enabled there can be a mention overlap in the achieved results. In order to avoid overlaps we exploited regular expressions. In particular, we merged two or more mentions when they are consecutive, and we choose the largest span mention when there is a containment. While with Spotlight we are allowed to find linked entities only, with this approach we can detect both entities that matches well known DBpedia resources and those that have not been identified by Spotlight (NIL). In this case given an entity spot, for entity linking we exploited DBpedia Lookup and string matching between mention spot and the labels associated to DBpedia entities. In this way we were able to find both entities along with their URIs, plus several more NIL entities. At this point, for each retrieved entity we have the span, the type (multiple types if CRF and DeepNL disagree) and the URI (see Figure 2) so we use a type detection/validation module for assigning the correct type to an entity. This module uses ad hoc

	#tweets	Character	Event	Location	Organization	Person	Product	Thing
Training set	1,450	16	15	122	197	323	109	20
New_train set	1,018	6	10	82	142	244	68	12
Validation set	432	10	5	40	55	79	41	8

Table 1: Dataset statistics

rules for combining types obtained from the classifier with CRF, DeepNL classifier if they disagree and from DBpedia entity type, when the entity is not NIL. For all NIL entities, finally we cluster them, as required by the challenge, by simply clustering entities with the same type and surface form. We consider also surface forms that differ in case (lower and upper).

CRF NER. The Stanford Named Entity Recognizer is based on the Conditional Random Fields (CRF) statistical model and uses Gibbs sampling for inference on sequence models(Finkel et al., 2005). This tagger normally works well enough using just the form of tokens as feature. This NER is a widely used machine learning-based method to detect named entities, and is distributed with CRF models for English newswire text. We trained the CRF classifier for Italian tweets with the new_train data annotated with IOB notation, then we evaluate the results across the validation data, results are reported in Table 2. The results provided follow the CoNLL NER evaluation (Sang and Meulder, 2003) format that evaluates the results in term of Precision (**P**) and Recall (**R**). The F-score (**F1**) corresponds to the strong_typed_mention_match in the TAC scorer. A manual error analysis showed that even when mentions are correctly detected, types are wrongly identified. This is due of course to language ambiguity in a sentence. As an example, for a NER it is often hard to disambiguate between a person and an organization, or an event and a products are not. For this reason we applied a further type detection and validation module which allowed to combine, by ad hoc rules, the results obtained by the classifiers and the Spotlight-based approach previously described.

Entity	**P**	**R**	**F1**	**TP**	**FP**	**FN**
LOC	0.6154	0.4000	0.4848	16	10	24
ORG	0.5238	0.2000	0.2895	11	10	44
PER	0.4935	0.4810	0.4872	38	39	41
PRO	0.2857	0.0488	0.0833	2	5	39
Totals	0.5115	0.2839	0.3651	67	64	169

Table 2: CRF NER over the validation set

DeepNL NER. DeepNL is a Python library for Natural Language Processing tasks based on a Deep Learning neural network architecture. The library currently provides tools for performing part-of-speech tagging, Named Entity tagging and Semantic Role Labeling. External knowledge and Named Entity Recognition World knowledge is often incorporated into NER systems using gazetteers: categorized lists of names or common words. The Deep Learning NLP NER exploits suffix and entities dictionaries and it uses word embedding vectors as main feature. The entity dictionary has been created by using the entity mention from the training set, and also the locations mentions provided by SENNA[10]. The suffix dictionary has been extracted as well from the training set with ad hoc scripts. Word embeddings were created using the Bag-of-Words (CBOW) model by (Mikolov et al., 2013) of dimension 300 with a window size of 5. In details we used the software word2vec available from https://code.google.com/archive/p/word2vec/, over a corpus of above 10 million of unlabeled tweets in Italian. In fact, the corpus consists of a collection of the Italian tweets produced in April 2015 extracted from the Twita corpus (Basile and Nissim, 2013) plus the tweets both from dev and test sets provided by the NEEL-IT challenge, all them pre-processed through our data preprocessing module, with a total of 11.403.536 sentences. As shown in Figure 3, we trained a DeepNL classifier for Italian tweets with the new_train data annotated with IOB-2 notation then we evaluate the results across the validation data. Over the validation set we obtained an accuracy of 94.50%. Results are reported in Table 3.

Entity	**P**	**R**	**F1**	**Correct**
EVE	0	0	0	1
LOC	0.5385	0.1750	0.2642	13
ORG	0.4074	0.2	0.2683	27
PER	0.6458	0.3924	0.4882	48
PRO	0.4375	0.1707	0.2456	16
Totals	0.5333	0.2353	0.3265	104

Table 3: DeepNL NER over the validation set

2.3 Linking

For the purpose of accomplish the linking sub task, we investigated if a given spot, identified by the machine learning approach as an entity, has a cor-

[10] http://ronan.collobert.com/senna/

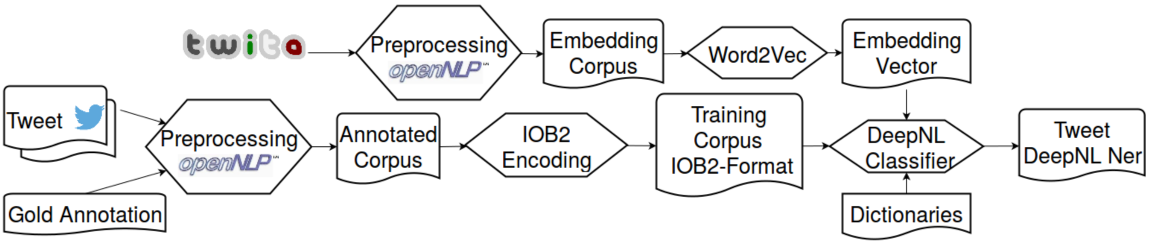

Figure 3: DeepNL: Training phase

responding link in DBpedia. A valid approach to link the names in our datasets to entities in DBpedia is represented by DBpedia Lookup[11] (Bizer et al., 2009) which behaves as follows:

candidate entity generation. A dictionary is created via a Lucene index. It is built starting from the values of the property `rdfs:label` associated to a resource. Very interestingly, the dictionary takes into account also the `Wikipedia:Redirect`[12] links.

candidate entity ranking. Results computed via a lookup in the dictionary are then weighted combining various string similarity metrics and a PageRank-like relevance rankings.

unlinkable mention prediction. The features offered by DBpedia Lookup to filter out resources from the candidate entities are: (i) selection of entities which are instances of a specific class via the `QueryClass` parameter; (ii) selection of the top N entities via the `MaxHits` parameter.

As for the last step we used the Type Detection module introduced above, to select entities belonging only to those classes representative of the interest domain. We implemented other filters to reduce the number of false positives in the final mapping. As an example, we discard the results for the case of Person entity, unless the mention exactly matches the entity name. As a plus, for linking, we also used a dictionary made from the training set, where for a given surface form and a type it returns a correspondent URI, if already available in the labeled data.

Computing canonicalized version. The link results obtained through Spotlight and Lookup or string match, refer to the Italian version of DBpedia. In order to canonicalized version as required by the task, we automatically found the corresponding canonicalized resource link for each Italian resource by means of the `owl:sameAs` property.

As an example the triple `dbpedia:Multiple_endocrine_neoplasia> owl:sameAs <http://it.dbpedia.org/resource/Neoplasia_endocrina_multipla>` maps the Italian version of *Neoplasia_endocrina_multipla* to its canonicalized version. In a few cases we were not able to perform the match.

3 Results and Discussion

In this section we report the results over the gold test set distibuted to the challenge participants, considering first 300 tweets only.

In order to evaluate the task results, the 2016 NEEL-it challenge uses the TAC KBP scorer[13]. TAC KBP scorer evaluates the results according to the following metrics: **mention_ceaf**, **strong_typed_mention_match** and **strong_linked_match**.

The overall score is a weighted average score computed as:

$$\textbf{score} = 0.4 \cdot \textbf{mention\_ceaf} + 0.3 \cdot \textbf{strong\_link\_match} + \\ +0.3 \cdot \textbf{strong\_typed\_mention\_match}$$

Our solution combines approaches presented in Section 2.1 and Section 2.2. For the 3 runs submitted for the challenge, we used the following configurations: **run1** Spotlight with results coming from both CRF and DeepNL classifiers; **run2** without CRF; **run3** without DeepNL.

As for CRF and DeepNL classifiers, we used a model trained with the whole training set provided by the challenge organizers. In order to ensemble the systems output we applied again the NERs Enabler & Merger module, presented in Section 2.2 that aims to return the largest number of entity annotations identified by the different systems without overlap. If one mention has been identified with more then one approach, and they disagree about the type, that returned by the Spotlight approach is chosen. Results for the different runs are shown in Table 4 together with the results of

[11] `https://github.com/dbpedia/lookup`

[12] `https://en.wikipedia.org/wiki/Wikipedia:Redirect`

[13] https://github.com/wikilinks/neleval/wiki/Evaluatio

System	mention_ceaf	strong_typed__mention_match	strong_link_match	final_score
Spotlight-based	0.317	0.276	0.340	0,3121
run1	0.358	0.282	0.38	0.3418
run2	0.34	0.28	0.381	0.3343
run3	0.358	0.286	0.376	0.3418
Best Team	0.561	0.474	0.456	0.5034

Table 4: Challenge results

the best performing team of the challenge. In order to evaluate the contribution of the Spotlight-based approach to the final result, we evaluated the **strong_link_match** considering only the portion of link-annotation due to this approach over the challenge test set, see Table 5. We had a total of 140 links to Italian DBpedia, then following the approach described in Section 2.3 we obtained 120 links, 88 of which were unique. It was not possible to convert into DBpedia canonicalized version 20 links. Final results are summarized in Table 5. Looking at the Spotlight-based solution (row 1), compared with the ensemble solution (row 2) results, we saw a performance improvement. This means that machine learning-based approach allowed to identify and link entities that were not detected by Spotlight thus improving precision results. Moreover, combining the two approaches allowed the system, at the step of merging the overlapping span, for a better identification of entities. This behavior lead sometime to delete correct entities, but also to correctly detect errors produced by the Spotlight-based approach and, more generally, it improved recall results.

System	P	R	F1
Spotlight-based	0.446	0.274	0.340
run1	0.577	0.28	0.380

Table 5: **strong_link_match** over the challenge gold test set (300 tweets)

In the current entity linking literature, mention detection and entity disambiguation are frequently cast as equally important but distinct problems. However, in this task, we find that mention detection often represents a bottleneck. In **mention_ceaf** detection, our submission results show that CRF NER worked slightly better then Deep NER, as already showed in the experiments over the validation set in Section 2.2. Anyway according to experiments in (Derczynski et al., 2015) with a similar dataset and a smaller set of entities, we expected better results from CRF NER. A possible explanation is that errors are due also to the larger number of types to detect as well as to a wrong recombination of overlapping mentions, that has been addressed using simple heuristics.

References

G. Attardi. 2015. Deepnl: a deep learning nlp pipeline. *Workshop on Vector Space Modeling for NLP, NAACL.*

P. Basile and M. Nissim. 2013. Sentiment analysis on italian tweets. In *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.*

P. Basile, A. Caputo, A. L. Gentile, and G. Rizzo. 2016a. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016).* Associazione Italiana di Linguistica Computazionale (AILC).

P. Basile, F. Cutugno, M. Nissim, V. Patti, and R. Sprugnoli. 2016b. EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016).* Associazione Italiana di Linguistica Computazionale (AILC).

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. {DBpedia} - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165.

L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrel, R. Troncy, J. Petrak, and K. Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the 43rd ACL '05.*

A. Gangemi. 2013. A comparison of knowledge extraction tools for the semantic web. In *Proc. of ESWC.*

J., M. Jakob, C. Hokamp, and P. N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proc. of the 9th I-Semantics.*

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *In Advances in Neural Information Processing Systems*, pages 3111–3119.

E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proc. of 7th CONLL*, pages 142–147.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on KDE*, 27(2):443–460.