Iterative Knowledge Extraction from Social Networks

Marco Brambilla, Stefano Ceri, Florian Daniel,
Marco Di Giovanni, Andrea Mauri, Giorgia Ramponi
Politecnico di Milano. Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)
Milan, Italy
[firstname].[lastname]@polimi.it

ABSTRACT

Knowledge in the world continuously evolves, and ontologies are largely incomplete, especially regarding data belonging to the so-called long tail. We propose a method for discovering emerging knowledge by extracting it from social content. Once initialized by domain experts, the method is capable of finding emerging entities by means of a mixed syntactic-semantic method. The method uses seeds, i.e. prototypes of emerging entities provided by experts, for generating candidates; then, it associates candidates to feature vectors built by using terms occurring in their social content and ranks the candidates by using their distance from the centroid of seeds, returning the top candidates. Our method can run continuously or with periodic iterations, using the results as new seeds. In this paper we address the following research questions: (1) How does reconstructed domain knowledge evolve if the candidates of one extraction are recursively used as seeds? (2) How does the reconstructed domain knowledge spread geographically? (3) Can the method be used to inspect the past, present, and future of knowledge? (4) Can the method be used to find emerging knowledge?

CCS CONCEPTS

• **Information systems → Web searching and information discovery**; **Web mining**; **Document representation**;

KEYWORDS

Social media analysis, knowledge extraction, domain model, Twitter

ACM Reference Format:
Marco Brambilla, Stefano Ceri, Florian Daniel, Marco Di Giovanni, Andrea Mauri, Giorgia Ramponi. 2018. Iterative Knowledge Extraction from Social Networks. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3184558.3191578

1 INTRODUCTION

Massive technologies have been used recently to produce very large ontologies: DBpedia, YAGO, the Knowledge Graphs in Google and Facebook derive from structured or semi-structured curated data [7, 11, 13, 14]. However, knowledge in the world continuously evolves: new entities continuously emerge, and existing ones change their properties or become obsolete. In many cases, these new entities belong to the so-called *long tail*, i.e. the portion of the entity's distribution having fewer occurrences [8], and as such they are not included within knowledge graphs.

For discovering knowledge and its evolution, we can take advantage of an extremely powerful and massive source: the content produced on social media. One can conjecture that somewhere, within such a massive content, any entity (and its evolution) has left some traces. The problem is that such traces are unclassified, dispersed, disorganized, uncertain, partial, possibly incorrect. Therefore, deriving information about entities from social content is extremely difficult.

In this paper we propose an iterative knowledge extraction method for discovering knowledge by extracting it from social content. The method is the evolution and extension of our research presented at WWW 2017 [3] and is defined in the context of a broader vision on knowledge discovery, whose general framework is illustrated in [2]. We use **Twitter** as social content source; Twitter can be accessed via its public APIs, which extract tweets related to a given hashtag or Twitter account. We refer to **DBpedia** as generic source of ontological knowledge; DBpedia is publicly available through its open API. DBpedia *types* are used to partition the existing ontological knowledge, organized within a type hierarchy.

The domain of interest is described by a selection of DBpedia types. This selection is performed by domain experts and typically includes few (from 5 to 10) types. We find entities within such domain, by extracting them from the social content. The expert must also provide **seeds**, i.e. prototypes of the interesting entities, simply described by providing their twitter accounts. A small set of seeds is sufficient: we normally use 10 to 20 seeds.

Once initialized by domain experts, the method is capable of finding entities by means of a mix of syntactic and semantic techniques. Our method collects information from the seed's tweets and generates **candidates**, i.e. other twitter accounts which are mentioned within the extracted tweets; then, it associates each candidate to a **feature vector**, built by using terms occurring in their social content, giving more relevance to terms which match the types selected by the expert; then it associates each candidate to a **score**, equivalent to the distance of each candidate from the centroid of the seeds; finally, it returns the top candidates, listed in decreasing score order. Once the candidates are generated, they can be forwarded to a **crowd of evaluators** that can assess the correctness of the extraction. Furthermore, the user can select a subset of candidates and reuse them as new seeds in a new execution of the knowledge extraction process.

In this paper we study how the method captures *evolving knowledge*; this is a crucial aspect, as the method can be repeatedly applied in an iterative manner over the social content to capture new trends

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191578

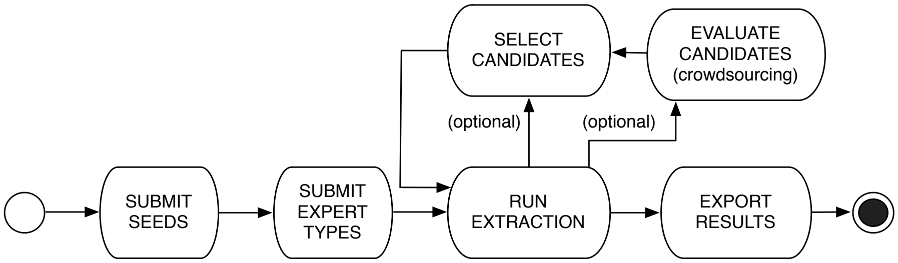


Figure 1: Iterative knowledge extraction process.

or to track knowledge spreading and evolution. We answer the following questions:

RQ1: *How does reconstructed domain knowledge evolve if the candidates of one extraction are iteratively used as seeds for the next extraction?*
RQ2: *How does the reconstructed domain knowledge spread geographically?*
RQ3: *Can the method be used to inspect the past, present, and future of knowledge?*
RQ4: *Can the method be used to find emerging knowledge?*

The paper is organized as follows: Section 2 describes our iterative knowledge extraction process; Section 3 presents the seven domains of interest over which we experiment the approach; Section 4 describes the four usage scenarios that respond to the above research questions and show the method at work on eacho of them; Section 5 describes our implementation; Section 6 discusses the related work; and Section 7 concludes.

2 ITERATIVE EXTRACTION PROCESS

The **knowledge extraction process** we propose is reported in Fig. 1:

(1) *The user submits a set of seeds* in input as samples of concepts to search for. These seeds consist of Twitter handles (usernames);
(2) *The user submits a set of expert types* as descriptors of the domain of interest.
(3) *The extraction of new candidates* is then launched and proceeds as follows:
   (a) Elimination from the seeds of outliers according to principal component analysis and computation of the centroid of the filtered seeds;
   (b) Collection of all the posts of each seed;
   (c) Definition of the set of candidate new entities as all the user handles that are mentioned by the seeds (which may lead to several thousand candidates);
   (d) Filter of candidates based on *tf-df* similarity [3], which allows one to reduce the space of analysis of the candidates to a limited set of relevant ones;
   (e) Collection of all the posts of each such candidate;
   (f) Computation of the feature vector representing each candidate;
   (g) Rank of the candidates based on the vectorial distance from the seed centroid and production of the result based upon the ranking.
(4) Once the candidates are retrieved and ranked, the user can:
   (a) *Export* them (in CSV format for human consumption or data analysis purpose or in RDF format for further integration in existing semantic knowledge bases);
   (b) Forward them to *domain experts or a generic crowd for result evaluation* purposes (validation);
   (c) Use them (or a subset of them) as new seeds and *iterate the whole pipeline.*

3 EXAMPLE DOMAINS OF INTEREST

We applied our method on different domains and usage scenarios, so as to demonstrate its generality:

- **Fashion designers**: the research team of the Fashion In Process Lab[1] (especially Paola Bertola, Chiara Colombi and Federica Vacca) was among the inspirators of this work, as it brought to us the problem of identifying emerging fashion designers. In the original experiment, the domain experts started with 200 emerging Italian brands as seeds.
- **Finance influencers**: a team of economics and statistics researchers at University of Pavia executed experiments on the extraction of influencers in finance. In this case the team selected as seeds 120 bloggers and journalists in the finance sector.
- **Fiction writers**: We considered some fiction authors engaged in the Melbourne Emerging Writers Festival[2] by picking 20 seeds from the participants to the event.
- **Craft breweries**: We considered as seeds a set of 20 well-known US craft breweries, all present in DBpedia.was to understand if it is possible to identify new craft breweries before they are widely acknowledged by consumers.
- **Chess players**: We used a list of 20 top chess players and their accounts.[3]
- **Jazz players**: We used a list of 10 top jazz players and their accounts.[4]
- **Fashion models**: We used a list of 20 fashion top models known in fashion, by extracting the top 20.

[1] http://www.fashioninprocess.com/
[2] http://www.emergingwritersfestival.org.au
[3] https://www.reddit.com/r/chess/comments/32t5ov/list_of_top_chess_player_journalist_twitter/
[4] http://oneworkingmusician.com/10-jazz-musicians-you-should-follow-on-twitter/

Table 1: Precision @10 and @20 of iterative knowledge extraction experiments using candidates produced in one run as seeds of a consecutive run; #results are the overall identified candidates.

	RUN #1				**RUN #2**				**RUN #3**			
SCENARIO	#seeds	#results	Pre@10	Pre@20	#seeds	#results	Pre@10	Pre@20	#seeds	#results	Pre@10	Pre@20
Fashion Designers	20	407	0.3	0.3	6	282	0.1	0.25	5	295	0.2	0.15
Fiction Writers	20	426	0.7	0.55	11	435	0.4	0.5	10	439	0.3	0.55
Chess Players	20	418	0.7	0.5	10	389	0.7	0.6	12	424	0.6	0.6
Finance	20	777	0.5	0.3	6	432	0.3	0.45	9	514	0.4	0.45
Craft Breweries	20	972	0.1	0.25	5	240	0.1	0.1	2	128	0.4	0.3
Jazz Players	20	428	0.8	0.8	15	431	0.8	0.8	16	426	0.9	0.85
Fashion Models	20	413	0.1	0.2	4	138	0.1	0.2	4	211	0.4	0.35
Talk Shows	20	423	0.5	0.45	9	440	0.3	0.45	9	437	0.4	0.35

- **Talk shows**: We used a list of 20 official Twitter accounts of popular TV talk shows.

These scenarios cover different information needs and domains. For instance, fashion design is characterized by a very high concentration of the domain in few brands only, most of which well known; on the opposite, fiction writers is an open domain where authors can be considered widespread; finance is a well established domain with renowned influencers; and craft beer is experiencing a tremendous growth with new craft breweries emerging almost daily.

4 EXTRACTION SCENARIOS

At the purpose of responding to the four research question presented in Section 1, we describe four possible usage scenarios for our method, and we report the findings obtained by experimenting with the scenarios on the seven domains discussed above.

4.1 Iterative Knowledge Extraction

The first usage scenario we want propose is *iterative* knowledge extraction, where successful candidates of one extraction are used as seeds for a subsequent extraction. We identified 20 seeds per domain, and ran 3 iterations of the method for each of them.

Table 1 shows the precision@10 and precision@20 obtained for each extraction for the eight domains; the #seeds in the second and third run correspond to the candidates of the respectively first and second run that were considered correct among the top 20 candidates (#results is the number of all candidates identified in a given run). Correctness was assessed against a manually tagged ground truth built through crowdsourcing; each run was executed twice. Every run after the first takes the good candidates of the previous run as seeds.

Within a given domain, consecutive runs tend to produce similar precision, independently of the number of seeds and results. It seems that certain domains are most suited to the method, such as chess or jazz players, most likely because the twitter accounts of these entities are focused on (if not limited to) their respective domains, whereas the method is less effective in other domains, such as breweries or fashion models. This latter result may be due to tweets that are less focused and contain generic topics, making similarity search less effective, but also to the presence of entities with high similarity but different ontological types (e.g., beer lovers/distributors or fashion bloggers). If initial entities are chosen from a specific subdomain (e.g., writers in Melbourne), iterations progressively extract entities from a wider semantic and geographic domain (e.g., from outside Australia).

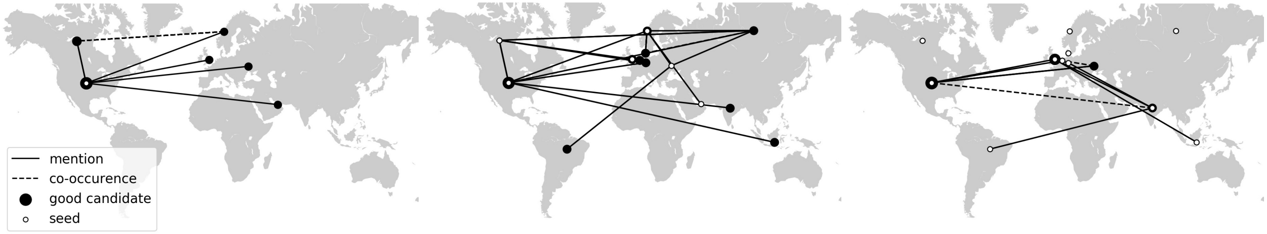
If we consider all runs as independent (considering neither the domain nor the order of execution), we find a correlation of 0.65 between the number of seeds and that of results (at the edge of significance) and one of 0.91 between precision at 10 and precision at 20. If we analyze the domains individually, pair-wise t-tests among the three runs neither identify any significant difference ($\alpha = 0.05$) between precision at 10 nor between precision at 20. The method thus works well even after several iterations, as precision remains rather stable (it decreases in certain domains, but it also increases in others); hence *a recursive application of knowledge extraction methods finds an increasing number of domain entities* **(RQ1)**. Especially when precision is high, one can find a good number of correct emerging entities from within the list of top-20 candidates.

4.2 Geographical Spreading of Knowledge

In order to study the geographical spreading of knowledge, we applied a similar iterative knowledge extraction approach as in Section 4.1 to one selected domain: chess players from the US. We decided to focus on this sub-domain (of all chess players) to study if our method can find entities from other geographical regions and, if yes, how fast the knowledge graph expands.

The experiment lasted three runs. For the first run of the experiment we took 7 seeds and a set of expert types. The next two re-runs were performed selecting the correct candidates from the top 20 results of the respective previous run. The actual localization of candidates was performed manually, either using the declared Twitter user location or, if that was missing, by searching other social resources and matching entities. After careful study of the location field as used by different Twitter accounts, we set the granularity of the locations to the level of individual countries.

The result is an instance-based graph of *mentions* from a seed to a candidate and *co-occurences* of two candidates, i.e., tweets by one of the seeds that mentioned the two candidates together, mapped to physical locations. The result is illustrated in Figure 2. The first knowledge extraction produced 12 good candidates from different countries and continents, reaching Europe and the Middle-East. The first iteration found 15 good candidates, adding new data points



(a) First run (b) Second re-run (c) Third re-run

Figure 2: Geographical dispersion of the knowledge graph in response to iterative knowledge extractions (US chess players).

also to South America and Asia. The second iteration produced 10 good candidates, while some seeds did not find any valid candidate.

We conclude that *discovered knowledge which is iteratively found spans large geographical areas very fast* **(RQ2)**. The finding is somewhat surprising, but can likely be explained with the open nature of Twitter, e.g., compared to Facebook (where we would expect a slower spreading).

With this experiment we proved that even starting from a small set of seeds from a single country, our approach is able to find good candidates from different states, countries and continents.

4.3 Capturing of Knowledge Evolution

The third usage scenario we propose is the study of how knowledge evolves over time. While the previous two scenarios are instances of an iterative knowledge extraction process, with selected candidates being used as seeds, here we propose a *periodic* extraction process, with knowledge extracted at periodic time intervals. For convenience, we fix the interval to three months, starting from September 2017 and looking back until January 2016. At each period, we consider all tweets since the beginning of 2016 up to the last month of the period being studied, constructing smaller reference data sets as we go back in time. It is important to note that to go back in time all cut-offs are computed from one cumulative download of tweets performed in the end of September 2017 using one set of seeds.

Table 2 reports the numbers of candidates extracted for the four domains studied so far. The four domains have a different evolution over time, with *Finance* being the youngest domain (our seeds started tweeting only in 2017). For the other three domains, one can observe that the *Fashion* domain growing slower than both *Chess* and *Australian Writers*. Looking at the table, it is also important to note that the rate at which knowledge increases is fast, that is, the knowledge we extract today is significantly bigger then the one we would have extracted only 3 months ago. Projected into the future, this solicits a continuous knowledge extraction instead of a periodic or random extraction. In conclusion, *knowledge can be extracted from social data at arbitrary points of time in the past and it is possible to trace how knowledge will evolve in the future*, thanks to the possibility to extract knowledge continuously **(RQ3)**.

4.4 Identification of Emerging Knowledge

For the analysis of how much knowledge reconstructed from social content can be considered as *emerging* (low-frequency entities not yet included in generic ontologies with high-frequency knowledge), we refer to *Wikipedia* as generic source of knowledge. We performed the analysis over four domains (*Fashion designers*, *Finance*, *Chess Players* and *Australian Writers*) we took the candidates produced with *one* iteration of the method and calculated the percentage of correctly identified candidates. To assess this aspect, we proceeded by counting how many candidates have a Wikipedia page, i.e., had already been captured formally.

Table 3 plots the obtained results. These are very domain dependent, likely due to the different social context behind the domains. For instance, in the *Fashion Designers* domain, the method produced an unexpected 100% of emerging designers. Instead, the domain that produced the lowest number of emerging entities is Australian writers (36%). Despite these fluctuations across domains and the fact that the reported results may not grant statistical representativeness, it is however important to note that in all cases *knowledge extracted from social content using the described method includes some relevant emerging knowledge that can be added to ontologies* **(RQ4)**.

Table 2: Looking back in time in the four domains: periodic knowledge extractions over a period of 21 months.

Time interval	Chess	Finance	Writers	Fashion
2016/01 - 2017/09	545	151	780	153
2016/01 - 2017/06	310	52	329	123
2016/01 - 2017/03	210	45	237	103
2016/01 - 2016/12	146	0	177	95
2016/01 - 2016/09	78	0	94	79
2016/01 - 2016/06	43	0	45	61
2016/01 - 2016/03	10	0	25	27

Table 3: Emerging knowledge compared to Wikipedia among the correctly identified candidates.

Domain	Emerging entities
Fashion Designers	100%
Finance	77%
Chess Player	42%
Australian Writer	36%

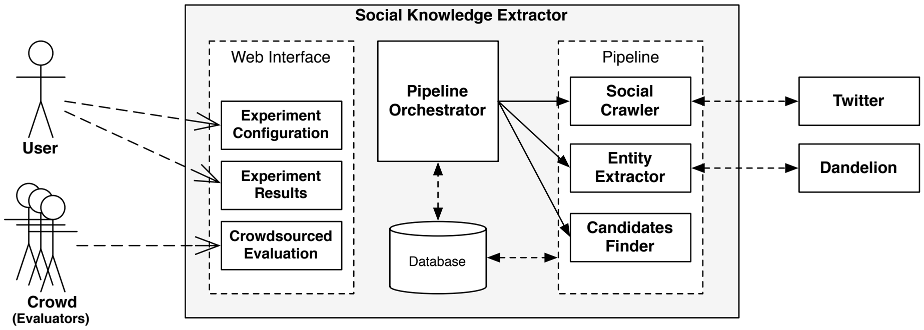


Figure 3: Architecture of the tool implementing our approach.

5 IMPLEMENTATION

The proposed approach has been implemented as a Python application. With respect to our original research [3], which extensively investigated more than 900 alternative extraction strategies, in this work we propose a light-weight tool, which only applies one strategy (the best one in [3]). While the quality of the results slightly decreases, the tool performance is quite good, in terms of tweets download, DBpedia matching, and score computation; the performance is adequate for exploring many domains on daily basis. The tool can run continuously or with periodic iterations, using the results as new seeds.

Fig. 3 represents the high-level view of the **system architecture**. The *Web Interface* allows users to interact with the system: it supports the phases of experiment definition and results visualization by the expert, as well as the validation of the results by the crowd. The *Pipeline Orchestrator* manages the execution of the process and is responsible of coordinating the components that perform each step of the analysis. The involved components are the following:

- *Social Crawler*: this component receives in input a list of Twitter handles (i.e., user identifiers) and uses the Twitter API [5] to crawl their tweets. It is used for retrieving the posts of both the seeds and the candidates.
- *Entity Extractor*: this component receive in input the text of the tweet and uses the Dandelion API[6] to find entities mentioned in the text. Dandelion is a commercial software which matches a text to either instances or types of DBpedia.
- *Candidate Finder*: this component is responsible of ranking the candidates using the information retrieved by the other components. In particular it creates the feature vectors and computes the similarity score of each candidate.

The data involved in the process is persisted in a MongoDB[7] database, which stores, for every user, the track of all his experiments, in terms of seeds, candidates, and evaluations.
The tool is available on GitHub under the Apache 2.0 open source license.[8]

6 RELATED WORK

This paper presents a method and tool to harvest the collective intelligence of the Social Web in developing a collective knowledge system [6]. P. Mika pioneered this area in [9], by identifying broader and narrower terms using social network analysis methods such as centrality and other measures like the clustering coefficient. Our interest is on the *circle of knowledge life* proposed in [12], where emerging knowledge is extracted from the Social Web using known facts captured in a knowledge graph. Our approach is grounded in *homophily*, a key aspect of social networks: entities are related when they have similar characteristics [1]. Homophily can be used to explain the scale-free nature[9] of social networks; in our approach, the seeds guide the process that identifies homophily patterns and thus constructs the domain graph. As pointed out in [16] and [5], the grand challenge in automating the discovery of emerging knowledge is to find entities, relationships and attributes not mainstream, belonging to niches in the long tail [4].

We found two works that also proposed to use Twitter for ontology enrichment. P. Monachesi and T. Markus in [10] proposed an ontology enrichment pipeline that can automatically enrich a domain ontology using data extracted by a crawler from social media applications. C. Wagner and M. Strohmaier [15] investigated a network-theoretic model called *tweetonomy* to study emerging semantics. Complementary to our work, they investigated how the selection of tweets (so-called Social Awareness Streams) can lead to different results. Incorporating their work is part of our future work.

7 CONCLUSIONS

In this paper, we explored a method for discovering knowledge from social media. The method consists of an iterative approach, in which the entities produced by one application of the method are used as seeds for the next application; we considered eight different domains. We specifically described the geographic and temporal spreading of entities extracted by the method. Finally, we measured the number of emerging entities found by the method; we regard as emerging those entities which are not present in Wikipedia.

We show that the method succeeds in achieving a high precision after several iterations in many domains, and particularly in the domains of chess and jazz players; we observe that in such domains

[5] https://dev.twitter.com/rest/public
[6] https://dandelion.eu/
[7] https://www.mongodb.com
[8] https://github.com/DataSciencePolimi/social-knowledge-extractor

[9] I.e., the vertex connectivity follows a power-law distribution.

the terms used in social communications are the most domain-specific. Future work includes the semi-automatic building of a richer domain model, by studying other twitter features (such as verbs and the bag of words which appear in tweet texts). This work is part of a general effort for building automatic knowledge discovery systems on top of socially provided content.

Acknowledgement. This work was partially supported by the ERC Advanced Grant 693174, Data-Diven Genomic Computing.

REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] M. Brambilla, S. Ceri, F. Daniel, and E. Della Valle. On the quest for changing knowledge. In *DDI@WebSci*, pages 3:1–3:5. ACM, 2016.

[3] M. Brambilla, S. Ceri, E. Della Valle, R. Volonterio, and F. X. Acero Salazar. Extracting emerging knowledge from social media. In *WWW 2017*, pages 795–804, 2017.

[4] A. Chris. *The long tail: Why the future of business is selling less of more.* New York: Hyperion, 2006.

[5] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction: The second generation. In *IJCAI*, pages 3–10. IJCAI/AAAI, 2011.

[6] T. Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web semantics: science, services and agents on the World Wide Web*, 6(1):4–13, 2008.

[7] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[8] A. Maedche. *Ontology learning for the semantic web*, volume 665. Springer Science & Business Media, 2012.

[9] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *International semantic web conference*, pages 522–536. Springer, 2005.

[10] P. Monachesi and T. Markus. Using social media for ontology enrichment. In *Extended Semantic Web Conference*, pages 166–180. Springer, 2010.

[11] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *ISWC 2016*, pages 177–185, 2016.

[12] A. Sheth, C. Thomas, and P. Mehra. Continuous semantics to analyze real-time data. *IEEE Internet Computing*, 14(6):84, 2010.

[13] A. Singhal. Introducing the knowledge graph: things, not strings. Available online at http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html, 2012.

[14] E. Sun and V. Iyer. Under the hood: The entities graph. Available online at https://www.facebook.com/notes/facebook-engineering/under-the-hood-the-entities-graph/10151490531588920/, 2013.

[15] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the 3rd International Semantic Search Workshop*, page 6. ACM, 2010.

[16] G. Weikum and M. Theobald. From information to knowledge: harvesting entities and relationships from web sources. In *PODS*, pages 65–76. ACM, 2010.