Extracting Emerging Knowledge from Social Media

Marco Brambilla Stefano Ceri Emanuele Della Valle
Riccardo Volonterio Felix Xavier Acero Salazar
Politecnico di Milano. Dipartimento di Elettronica, Informazione e Bioingegneria
Piazza Leonardo da Vinci, 32 – 20133 Milano, Italy
{firstname.lastname}@polimi.it

ABSTRACT

Massive data integration technologies have been recently used to produce very large ontologies. However, knowledge in the world continuously evolves, and ontologies are largely incomplete for what concerns low-frequency data, belonging to the so-called long tail. Socially produced content is an excellent source for discovering emerging knowledge: it is huge, and immediately reflects the relevant changes which hide emerging entities. Thus, we propose a method for discovering emerging entities by extracting them from social content.

Once instrumented by experts through very simple initialization, the method is capable of finding emerging entities; we use a purely syntactic method as a baseline, and we propose several semantics-based variants. 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 then ranks the candidates by using their distance from the centroid of seeds, returning the top candidates as result. The method can be continuously or periodically iterated, using the results as new seeds. We validate our method by applying it to a set of diverse domain-specific application scenarios, spanning fashion, literature, and exhibitions.

Keywords

Knowledge Extraction; Web Science; Social Media Analysis; Content Analysis; Emerging Knowledge; Entity Typing; Big Data; Online Communities; Web Data Mining; Evolving Knowledge.

1. INTRODUCTION

Massive technologies have been used recently to produce very large ontologies: Google's Knowledge Graph, produced and maintained by integrating data from high-quality structured sources, is used for industrial-strength semantic query interpretation [28]; DBpedia is continuously evolving by extracting data from Wikipedia infoboxes [14], and it contained as of October 2016[1] 4.58 million things, out of which 4.22 million are classified in a consistent ontology.

However, knowledge in the world continuously evolves, at a pace that cannot be traced by Wikipedia. In particular, the process of ontological knowledge discovery tends to focus on the most popular items, those which are mostly quoted or referenced, and is less effective in discovering less popular items, belonging to the so-called *long tail*, i.e. the portion of the entity's distribution having fewer occurrences, far from the central part of the distribution itself. Even the largest ontologies are incomplete for what concerns low-frequency data [17]. It turns out, however, that knowing the long tail may have a strong relevance, e.g. in e-commerce [7]; high frequency entities are generally well known, but low-frequency entities include *emerging entities*, that have a small impact today but may have a high impact tomorrow. Therefore, the early discovery of low-frequency entities and establishing their ontological properties is a very interesting problem, with economic and practical implications; this is the main goal of our research.

For discovering emerging knowledge, there is today a new and extremely powerful source: socially produced content. Social networks are huge (1.8 billion of users) and fresh (their content is produced while events occur). Thus, one can conjecture that somewhere, within such a massive content, any low-frequency, emerging entity has left some traces. The problem is that such traces are unclassified, dispersed, disorganized, uncertain, partial, possibly incorrect. Therefore, deriving information about low-frequency emerging entities from social content is extremely difficult.

The research community has not considered social content yet in building ontological knowledge; DBpedia, Yago, the Knowledge Graphs in Google and Facebook derive from structured or semi-structured curated data [14, 25, 28, 30]. However, social content has fueled the new discipline of Social Media Analytics, concerned with analyzing real world phenomena using social media; this in turn has created a wealth of useful technologies and tools, including new entity recognition and linking techniques which apply special-purpose NLP techniques to social content [29].

1.1 Problem Definition

We consider a portion of knowledge graph which describes a specific domain of interest, called *domain graph*; we will

[1]Data from `https://en.wikipedia.org/wiki/DBpedia`, English version, accessed on February, 15th , 2017.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052697

next consider the domains of fashion, writing, and exhibitions. The main objective of this research is to find *emerging entities for a core entity type*; these entities are not included in the domain graph but they are present in the social content, hence we can think of them as being included in the long tail, but also in the process of emerging from it. This problem was suggested to us by experts of fashion design (which is distinctive of Milano in the world), who are very interested in early discovery of emerging fashion designers.

Our method requires minimum involvement of domain experts. We ask them to produce a small number of low-frequency entities of the type of interest, together with their social handlers (account IDs). We regard these low frequency entities as **seeds**, and our method consists of a workflow of tasks for searching within the social content other entities which are similar to those seeds. Similarity is defined by standard feature comparison, where features include the mentioning of hashtags, user handles, terms and the co-occurrence with other entities; in this way, the determination of emerging entities can be conducted to a classic information retrieval problem. Feature extraction is facilitated by the availability of several instruments from social media analytics, such as social network crawlers (based on their public APIs) and entity recognition tools.

Although the above problem is defined in general terms, we next focus on specific choices of knowledge base and of social content:

- We refer to DBpedia as generic source of ontological high-frequency knowledge. DBpedia is publicly available through its open API; DBpedia *types* are used to partition the existing ontological knowledge, organized within a type hierarchy; types which have no descendants are denoted as the (most) *concrete types*. Entities that can be referred to any DBpedia type are considered as *high-frequency*, while the other entities are considered as *low frequency*.
- 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 restrict to tweets produced after a given time threshold; this allows us to focus on *recent history* (hence, to precisely define what we mean by emerging).

While providing seeds is the minimum requirement for the method, we may also ask domain experts to indicate a small number of other domain types whose entities are typically related to the core entity type, thereby providing a very simple model of the semantics of the domain of interest. In this way, we move from pure syntactic co-occurrence to models which use semantics; we show that such limited semantic injection may significantly improve the quality of the method.

1.2 Research Questions

The research questions tackled in this paper are:

- R1: Are we able to extract low-frequency domain-specific emerging entities from social media content?
- R2: Can we use domain knowledge i.e. can we design semantically-enriched methods for improving over syntactical extractions?
- R3: Can we find good methods across domains, without requiring domain-specific tuning?

The main result of this paper is a class of methods for finding low frequency entities of a given type; the approach requires a simple initialization by domain experts and achieves good precision in determining a small number of emerging entities; precision is preferred to recall, because the method can be continuously or periodically iterated upon the entities resulting from one method application in order to derive new emerging entities. We discuss several variants of the base method, and present experiments that select the class of best performing methods across all variants; we consider Twitter as representative social networking platform and we select three very different domains of application; we show that the method has good precision on all the domains and is rather robust to small variations of the parameters.

1.3 Method Outline, Paper Organization and Limitations of our Approach

The paper is organized as follows. *Section 2* presents the selection of candidate entities starting from expert-defined seeds. *Section 3* presents methods for ranking the candidates according to precision. The first class of methods is *purely syntactic*, and consists of extracting feature vectors for each candidate based on co-occurrence with handles or hashtags appearing in the tweets, and then ranking candidates based on their distance from the seed's centroid. The second class of methods is *semantic* as it uses matches to DBPedia types, and specifically distinguishes between: *all* DBPedia types, *concrete* (i.e. most specific) types, and *expert* types, i.e. suggestions by experts of the most relevant DBpedia types in the context of the specific domain.

Section 4 introduces three domains of interest: fashion designers, fiction writers, and expo pavilions; *Section 5* presents our evaluation; we first select the best syntactic strategy, that we use as baseline, then the evaluation of semantic methods against the baseline. Our findings indicate that one class of methods has higher precision at K than the syntactic method in all the three domains, for most precisions with $K \leq 25$, and is robust to small variations of parameters. We also exemplify the seeds which are found in the fashion domain, that indeed are valuable for our experts.

Our work has the following limitations: (1) using expert-provided seeds may bias the result. Indeed, the method requires experts to provide good initial input, (2) using tweets may be most suitable for given domains, as the majority of twitter accounts represents persons, organisations or locations, (3) our method applies to concepts which are already present in DBPedia; it cannot be used for extracting instances of completely new concepts.

2. SELECTION OF CANDIDATES

The first part of the method is concerned with the selection of suitable candidates, i.e. of potential low-frequency entities which are similar to the seeds. This part of the method uses simple heuristics and rankings, with an emphasis on recall rather than precision. We formalize the problem as follows.

- Given $E = \{e_1, e_2, \ldots, e_n\}$, a set of seeds (i.e., low frequency entities of a given type) provided by the domain expert in the form of Twitter handles;
- Find $C = \{c_i | c_i \in C \wedge F(c_i) > \beta\}$, a set of candidates, where c_i also denotes social network (Twitter) handles.

The condition $F(c_i) > \beta$ indicates that we can rank the candidates according to their features and then set a threshold on such ranking (as next described), in order to reduce the computational complexity of their subsequent comparative analysis, which requires downloading the whole social content associated to each candidate.

Since our whole approach is based on similarity of the candidates to the seeds, we need the seeds to be reasonably homogeneous. Therefore, we initially run a principal component analysis over the seeds and then a k-means clustering with $k = 2$ to check homogeneity. For each seed s_i, we compute a feature vector describing its characteristics; as vector dimensions, we use the set $S = \{s_1, s_2, s_3, \dots\}$ of *spots*, i.e. the hashtags, mentioned handles, and any other relevant token, that co-occur with each seed, extracted from the content posted by the twitter accounts of the seeds.

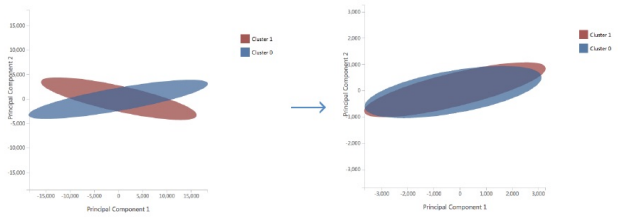


Figure 1: Evaluation of seeds' homogeneity through principal component analysis and *k*-means clustering.

Using the whole set of seeds may result in quite diverse clusters, basically due to outliers (Figure 1 left displaying two clusters with respect to the two principal components). In such case, we use the Coefficient Variation method to find a more homogenous set of seeds. We remove the outliers, i.e., the elements that are two standard deviations away from the mean, and thus we obtain a more homogeneous set of seeds (Figure 1 right). We then consider such set of seeds and we calculate the centroid e_c, which will be used as a global representative of the seeds in the comparison with the candidate emerging entities (see next section).

Given the list of homogeneous seeds, we restrict our interest to the social content posted by those seeds; we compute then the set of **candidates** $C = \{c_i | c_i \in S \wedge F(c_i) > \beta\}$, as the set of the twitter handles appearing in such social content; the intuition behind this choice is that emerging concepts of given type typically relate to each other, hence by looking to the content produced by the seeds there is a good probability of finding concepts which are similar to them.

In order to limit their number, we restrict the analysis to a subset selected by using a ranking function F and a threshold β. The ranking function is based on the following intuition: candidates that occur in relation with many seeds are more interesting than candidates that occur frequently, but with only a few seeds. The ranking function, defined according to the above objectives, assumes the form of a typical information retrieval function[2]:

$$\frac{df * ttf}{(N - df + 1)}$$

where:

- df = Number of seeds with which a candidate co-occurs with;
- ttf = Total number of times a candidate occurs in the analyzed content;
- N = Number of seeds.

3. RANKING OF CANDIDATES

The second part of the method is concerned with the ranking of candidates; the top candidates in the ranking are returned as result and typically confirmed by the expert by accessing their social media information, e.g. through search or by using twitter accounts.

We start with retrieving the social content of all candidates. Note that in a typical setting there are many more candidates than seeds, and that for each candidate we make access to social media through their APIs; in general, the size of social content retrieved at this stage is larger than the size of content which is used for candidate selection. We use estimates of the workload to set the thresholds in the filtering of candidates, described in the previous section, which however is domain-specific, as it depends from the social content size which is produced in the domain.

The social content's text is interpreted, using different tools (see Section 3.2), in order to give values to a *vector of features*, and then each candidate's feature vector is compared to the vector of the seeds' centroid e_c, and candidates are ranked by their distance to the centroid. As distance measure we used *cosine distance*, which gave better precisions than other options (we considered also *Euclidean distance* and *Pearson correlation*).

The main aspect of the method is the choice of considered features yielding to the best ranking; we considered several alternatives, discussed in the next two subsections.

3.1 Syntactic Methods

Syntactic methods apply simple text analysis approaches in the definition of the feature vectors, by selecting appropriate text spots and their frequencies. We focus on the parts of Twitter texts that users explicitly identify as relevant elements, like handles (i.e., user IDs, identified through the prefix @) and hashtags (identified by the prefix #). Based on simple testing on precision, we focus on the following two alternatives:

- *Strategy Syn1* considers the frequencies of all the handles;
- *Strategy Syn2* considers the frequencies of all the handles and hashtags.

Their comparison is presented in Section 5.1[3].

3.2 Semantic Methods

Domain-independent knowledge can be extracted from a general knowledge base (such as DBPedia); the focus of our

[2]Note that, despite some similarity with the *tf-idf* function, our ranking function is different in terms of format and meaning.

[3]Considering also arbitrary spots in text is needlessly noisy as it frequently includes tokens which carry no meaning, yielding to low precision; we excluded such alternative.

research is determining if the use of such knowledge for defining variables may increase the precision of candidate selection over syntactic methods. To enhance our options, we ask a domain expert to define a set of *DBpedia types relevant to the domain* $T = \{t_1, t_2, \ldots, t_m\}$; they can be used to give more relevance to given features, related to T either at the entity or at the type level. For instance, *fashion designers* are related to *brands*, *photographers*, *magazines*. These are generally few (order of 5-10) and very easy to define through an interaction with the expert. The use of domain-specific types for characterizing some of the variables can increase the precision of candidate selection, as shown next.

An important ingredient of the method is the matching of texts to semantic entities. We use a commercial software called Dandelion[4] which matches a text to either instances or types of DBpedia; the matching uses a parameter ranging from 0 to 1 for setting different levels of recall. In our experiments, we set the recall parameter from 0.15 (low recall and high quality matches) to 1 (high recall).

3.2.1 Basic Semantic Strategies

The first category of features considers *entities* (e.g. `Prada`); of course, matching to entities may generate a large number of features. For this reason, we match just **handles and hashtags to entities (HE)**, and we further separate arbitrary DBpedia entities from concrete and expert entities, yielding to the following three classes of variables:

- *AHE (All Handle and hashtag Entities)*: variables are associated to all the hashtags and handles that correspond to an entity in the knowledge base, i.e., are identifiable as high frequency entities;
- *CHE (Concrete Handle and hashtag Entities)*: variables are associated to the hashtags and handles that correspond to an entity in the knowledge base, whose type is a concrete type in the type hierarchy of DBpedia, i.e. the most specialized matching type in the DBpedia type hierarchy.;
- *EHE (Expert Handle and hashtag Entities)*: variables are associated just to hashtags and handles that correspond to one of the semantic types T provided by a domain expert.

The second category considers *types* (e.g. `city`, `magazine`) and focuses on matches of **handles and hashtags to types (HT)**. Variables are built as follows:

- *AHT (All Handle and hashtag Types)*: variables are associated to the counts of the handles and hashtags that match any type in the knowledge base. The association holds along type hierarchies: for instance, if we recognize that an handle is of type `city` and in turn `city` is a subtype of `settlement` and `place`, then the handle is also associated with a variable for types `settlement` and `place`.
- *CHT (Concrete Handle and hashtag Types)*: variables are associated to the counts of the handles and hashtags that match a concrete (i.e. most specialized) type in the knowledge base. In the above example, the handle is associated only with a variable for the `city` type.

[4] `https://dandelion.eu/`

Table 1: Summary of the nine semantic strategies.

	HE (hashtag and handle entities)	HT (hashtag and handle types)	ST (spot types)
A (all types)	AHE	AHT	AST
C (concrete types)	CHE	CHT	CST
E (expert types)	EHE	EHT	EST

- *EHT(Expert Handle and hashtag Types)*: variables are associated just to the counts of the handles and hashtags that match an expert type.

The third category of features considers again types but is not restricted to hashtags and handles, as it includes any entity (or *spot*) in the text that can be matched to a type; thus the method accounts for any **text Spot Type (ST)**. Variables are built as follows:

- *AST (All Spot Types)*: any spot that corresponds to an existing entity in the knowledge base is identifiable as high frequency entity and contributes to variables associated with this type and with all its ancestor types, as discussed in the AHT case;
- *CST (Concrete Spot Types)*: any spot that corresponds to an existing entity in the knowledge base is identifiable as high frequency entity and contributes to the relevance of its type (but not of all its ancestor types).
- *EST (Expert Spot Types)*: any spot that corresponds to one of the types T selected by the domain expert, and thus is identifiable as a low frequency entity, contributes to the relevance of its type.

In this way, we have defined nine basic semantic strategies for building our feature vectors, as summarized in Table 1. Each strategy can be configured with different levels of recall of BDpedia matches.

3.2.2 Mixed Semantic Strategies

Mixed strategies combine one instance base (*HE*) and one type based (*HT* and *ST*) strategy; the corresponding feature vectors include instance-based features and type-based features. Thus, 18 mixed strategies are obtained, denoted as $S_1^\frown S_2$ (for instance, the strategy that mixes *AHE* and *CST* is denoted as $AHE^\frown CST$). We weight the features by using a parameter $\alpha \in [0..1]$, so that the features which use entities (HE) are weighted by the multiplier α and features generated by types (HT or ST) are weighted by $(1–\alpha)$. In summary, by setting α and the recall parameters we obtain 55 variants, leading to a total of 990 semantic strategy variants, consisting of:

- 18 alternative feature vector configurations for describing the candidates;
- 11 different values for α, in the range [0..1], used for weighting the two contributions in the combinations;
- 5 levels of recall [0.15, 0.25, 0.50, 0.75, 1] for the entity extraction algorithms applied to the posts.

4. DOMAINS

In order to validate our approach, we considered three very different domains:

- **Fashion designers**: we considered the problem of identifying emerging fashion designers, which are not yet globally recognized and thus are not present in DBpedia. Domain experts, who are sector leaders [5], provided us with 200 emerging brands in the Italian market, that we took as seeds. We crawled about 250,000 tweets and analyzed 237,000 of them (the ones written in one of the supported languages, i.e., English or Italian). Out of this analysis, we extracted 1.1 million high frequency entities, which were categorized against the 740 DBpedia types.

- **Fiction writers**: we considered the problem of identifying non-famous writers, starting from a set of about one hundred writers engaged in a literature event in Australia[6]. Based on the event program, we defined a set of 22 seeds; we then crawled 14,590 tweets, and from there we extracted several thousand handles; after ranking, we selected the top 200 candidates and crawled their twitters, summing up to 101,200 tweets.

- **Live events**: we considered the domain of the Universal Exposition (EXPO 2015) that took place in Milan in 2015, and we constructed knowledge about the exhibition pavilions, using 15 official accounts as seeds. We crawled 24,000 tweets of the seeds, produced in the period from January to May 2015; in this content, we identified 23,000 high frequency entities and 5000 Twitter handles mentioned by the seeds; we selected 200 of them as candidates, and then crawled more than 500,000 tweets published by the candidates.

The scenarios cover different situations, which are all interesting for our research. Fashion is characterized by a very high concentration of the domain in a few brands, most of which are known; the scenario was also characterized by availability of a good corpus of seeds, which however made candidate selection more difficult (availability of many seeds left us with few candidates to be discovered); on the opposite, writing of fiction is a quite open domain where authors can be considered very widespread; and live events typically count a very small number of entities of interest (e.g., pavilions in Expo are only 140) and have a short duration.

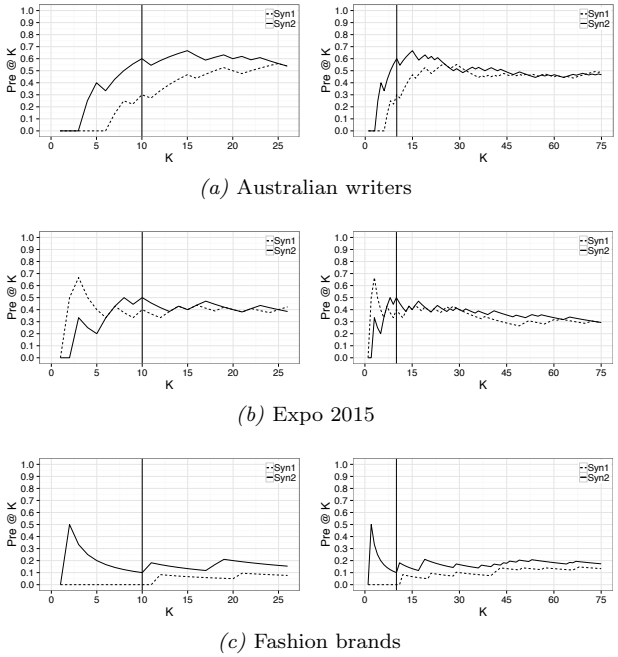
According to our method, starting from the set of seeds provided for each domain, we proceeded with the following phases:

1. elimination from the seeds of outliers according to principal component analysis, and computation of the centroid of the filtered seeds;

2. collection of all the posts of each such seed on the social network of interest;

[5]Data were made available from research related to an emerging fashion designers exhibit which took place in Milano: `http://www.vogue.it/vogue-talents/contest-opportunities/2015/11/20/il-nuovo-vocabolario-della-moda-italiana-triennale/`.

[6]We used the information from the artist section of the 2015 Web site for the Melburne Emerging Writers festival, which is held annually; see the 2016 edition at: `http://www.emergingwritersfestival.org.au/`.



(a) Australian writers

(b) Expo 2015

(c) Fashion brands

Figure 2: Comparison of precision at K of syntactic strategies (up to $K = 25$ on the left, up to $K = 75$ on the right), where *Syn1* is represented by a dashed line, *Syn 2* by a continuous line, and the vertical bars indicate $K = 10$.

3. definition of the set of candidate new entities as all the user handles (i.e., user IDs in the social network) that are mentioned by the seeds (which leads to sets counting hundreds of thousands of candidates);

4. selection of the top candidates (in the order of thousands of instances) according to the ranking function;

5. definition of a vector representation for each top candidate;

6. ranking of the candidates based on the distance from the seed centroid, and production of the result based upon the ranking.

5. EVALUATION

The choice of evaluation measure is a relevant decision in the process. Indeed, depending on the usage scenario, one can choose among a large set of options for measuring the quality of the method: *precision@K* (where also the value of K is relevant), ROC AuC (area under the curve in a true vs. false positive curve), and so on. Given that our usage scenario is to propose a small set of results to an expert for confirmation, we focus the evaluation on *precision@K* as preferred measure, with small K (we use *K=10* and we present scatter plot diagrams for $K \leq 25$), because most expert are interested in shortlists of few good quality results rather than perusing longer lists of low quality results. For the evaluation, we asked design experts to classify all top-ranking candidates, and we manually evaluated all candidate writers and pavilions (with $K \leq 25$).

5.1 Syntactic Methods

Fig. 2 compares the methods *Syn1* and *Syn2*, at work with different scales of precision (up to *K=25* on the left, up to *K=75* on the right). By selecting *Precision@10*, we see that *Syn2* produces higher quality results. More specifically, *Syn1* is competitive for very small K in the case of Expo, but *Syn2* results superior to *Syn1* in Fashion and Australian Writers with any $K \leq 25$. We hence select *Syn2* as the best syntactic strategy – and our baseline for semantic methods. Note that with increasing value of K, the two methods extract more overlapping candidates, leading to similar precision.

5.2 Semantic Methods

Semantic analysis consists of the comparative evaluation of 990 methods over 3 domains; we recall that this number is obtained by starting from all combinations of 3 entity-based and 6 type-based methods, while parameters include 11 values for the α coefficient and 5 values for precision. Also for semantic analysis, we use *Precision@10* as quality criterion. Fig. 3 provides a summarization of the evaluation; it shows how many strategies perform worse, the same and better than the respective syntactical ones in the three domains. By observing the summary (Fig. 3 (d)) we see that semantic methods have potential of moving the overall quality much beyond the benchmarks in the three domains, however we also see that many semantic methods actually perform worse than syntactic methods, hence the selection among them has to be careful.

The first step of analysis is a pruning of the six type-based strategies; after exhaustive analysis we note that 70% of the 165 methods which use all spot types (AST), combined with a choice of α smaller than 0.7, give better or same precision (62% strictly better precision) with respect to the syntactic baseline, whereas with all other types and settings of α this comparison falls below 70%; hence, we decide to focus on AST and disregard the other five type-based cases (AHT, CHT, EHT, CST, EST). The intuition behind such superiority is that type-based features are few dimensions (they are counts over semantic types), hence all types are preferred to expert and concrete types (which are too few for similarity analysis) and similarly all spots are preferred to just handles and hashtags.

We next consider, Precision@K of the combinations of entity-based methods with the AST method in the three domains; boxplots in Figs. 4, 5, and 6 illustrate the difference of *Precision recision@K*, with $K \leq 25$, for the three instance-based strategies (AHE, CHE, EHE) with the various alternative parameter settings. We note that precision is generally higher than the best syntactic method; however, by looking at *precision@10*, we also note that some methods with given parameter settings produce lower precision than the baseline.

Finally, in Fig. 7 we focus on the 39 semantic methods that always perform better than the best syntactic one in all the three domains using *precision@10*; boxplots illustrate the difference of *precision@K*, with $K \leq 25$, in the 39 methods. We note that 35 cases out of 39 methods include AST and 25 of them further include EHE. Based on these considerations, we select as best semantic method the **combination** $EHE^\frown AST$; this combination uses the expert types for classifying the handles and hashtags occurring in the tweets, and uses all the semantic types for aggregating the spots occurring in the tweet.

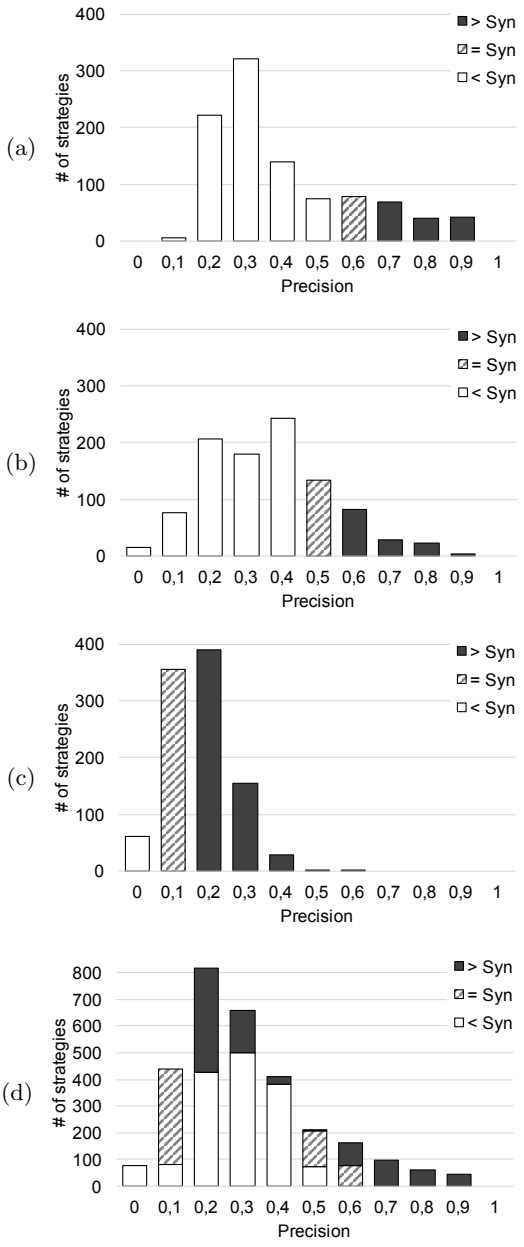


Figure 3: Distribution of the number of strategies based on the precision they achieve, for Australian writers (a), Expo 2015 (b), Fashion (c), and a summary cross-scenario (d). The bar-charts highlight how many strategies perform worse, the same and better than the respective syntactic ones.

Table 2 shows the influence of parameter setting over quality with the combination $EHE^\frown AST$ (we consider 55 possible settings over 3 domains). The best parameter setting is obtained with α*=0.7* and *recall=1*. The table shows that *alpha* is rather stable (any value for alpha between 0 and 0.8 yields to a method that improves over the best syntactic method in more than 60% of the times); *recall* to be used

Table 2: Percentage of strategies that perform better than the syntactic strategy depending on: (a) the weight α assigned to instance vs. type level features; and (b) on the recall of the semantic entity extraction.

(a)

	α										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
EHE⌒AST	60%	60%	60%	60%	73%	73%	73%	87%	67%	40%	7%

(b)

	Recall				
	1.00	**0.75**	**0.50**	**0.25**	**0.15**
EHE⌒AST	82%	67%	48%	55%	48%

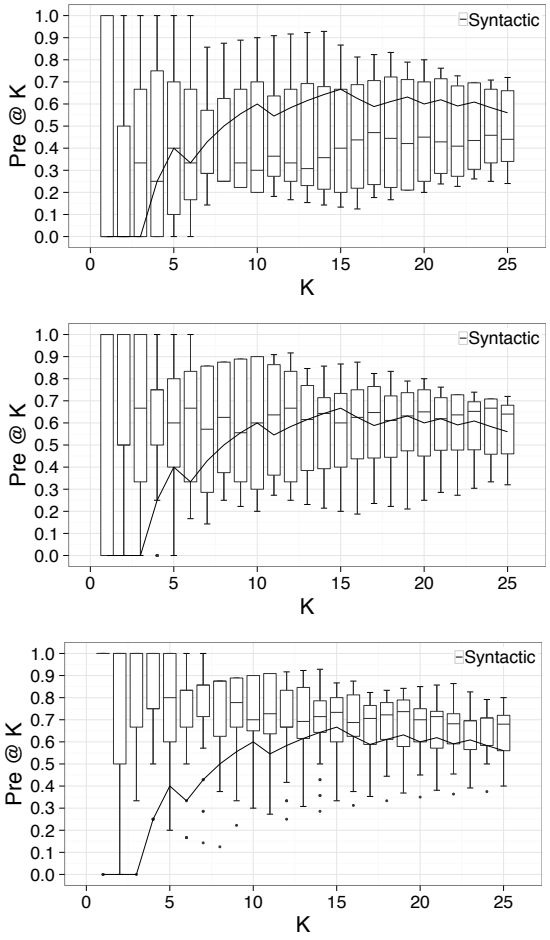


Figure 4: Australian writers scenario: Pre@K for combined strategies of AST with all the 3 instance-based strategies (AHE, CHE and EHE respectively).

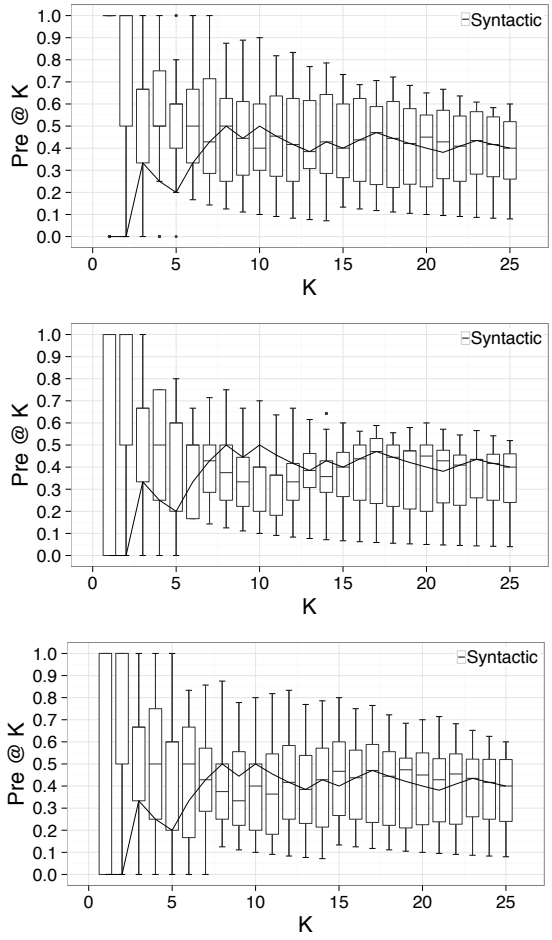


Figure 5: Expo 2015 scenario: Pre@X for combined strategies of AST with all the 3 instance-based strategies (AHE, CHE and EHE respectively).

for term matching with DBpedia is less stable, and the acceptable values are 1.0 and 0.75.

Finally, Table 3 shows the actual fashion designer candidates extracted by the best method, i.e. $EHE^{\frown}AST$ with *alpha=0.7* and *recall=1*. Note that *precision@10* is 0.6, i.e. 6 of the extracted handles are confirmed as emerging fashion designers by experts, whereas *precision@10* of the benchmark is only 0.1. This result was very well received by experts, who will periodically use the method for updating their catalog, by using confirmed candidates as seeds. Table 4 shows the 13 confirmed candidates extracted by the method (with N=25) in the Expo scenario; one can observe

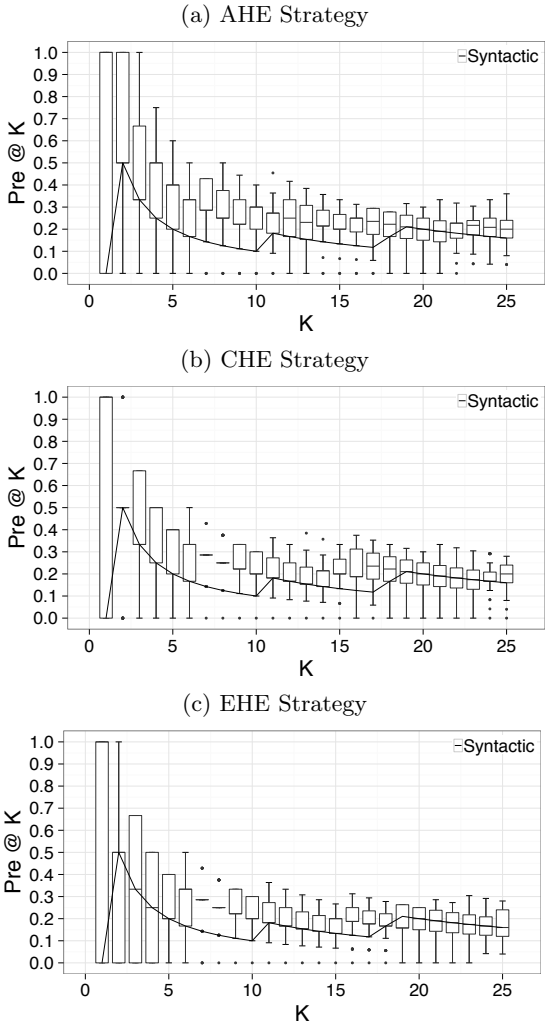


Figure 6: Fashion designers scenario: Pre@X for combined strategies of AST with all the 3 instanc-based strategies (AHE, CHE and EHE respectively).

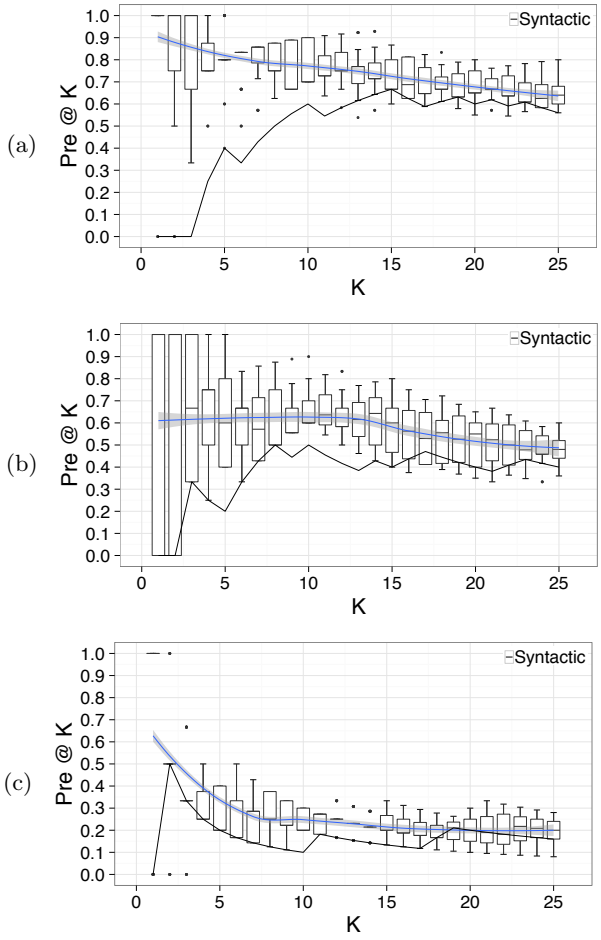


Figure 7: Pre@X of the 39 semantic strategies that always perform better than the syntactic one in all the scenarios for Pre@10, for Australian writers (a), Expo 2015 (b), and Fashion (c), compared to the syntactic baseline.

that pavilion handles are diverse and hard to find (e.g., many of them do not include either of the words *pavilion* or *expo*.)

6. RELATED WORK

This paper presents a method to harvest the collective intelligence of the Social Web in developing a collective knowledge system – a human-computer systems in which machines enable the collection and harvesting of large amounts of human-generated knowledge [12].

This research area was vastly studied in recent years. P. Mika pioneered this area in [19], by identifying broader and narrower terms using social network analysis methods such as centrality and other measures like the clustering coefficient. On the same line, P. Heymann and H. Garcia-Molina propose to measure tag generality using betweeness centrality [13]. P. Schmitz [26], instead, used a statistical approach to model tag subsumption. The work [23] proposed a bottom-up, semi-automatic service registration process exploting existing knowledge bases and text processing techniques for semantic annotation of data services, while [18] focused instead on publishing of arbitrary data sources as semantic streams.

However, our interest is on the *circle of knowledge life* proposed in [27], where emerging knowledge is extracted from the Social Web using known facts captured in a knowledge graph. To this end, the work presented in this paper uses various methods of Knowledge Extraction (see Section 3.2 and footnote 4). Interested readers can refer to [8] for supervised knowledge extraction approaches, to [4] and [11] for semi-supervised ones, to [21] for "distantly" supervised ones; and to [22] and [24] for entity resolution, through record linkage and instance matching, respectively.

As pointed out in [32], [9] and [10], the grand challenge in automating the discovery of emerging knowledge is to find entities, relationships and attributes that are not mainstream, but belong to niches in the long tail [7]. Our approach is grounded in a key aspect of social networks: entities are related when they have similar characteristics. This is called *homophily* [22, 15] and it can be used to explain the

Table 3: Ranking of the top-10 candidates extracted in the Fashion Scenario.

Twitter Handle	Score	Pre@X
eleonoracarisi	0.722917563	1.00
themenissue	0.718428825	0.50
albertomoni	0.712592288	0.33
lindatol001	0.671432151	0.25
iodonnalive	0.662630480	0.20
danielbeckerman	0.656952183	0.33
andreaincontri	0.655022364	0.43
antoniomarras	0.648776969	0.50
miraduma	0.648021181	0.56

Table 4: Confirmed Expo handlers extracted from top-25 candidates in the Expo 2015 Scenario.

serbiaexpo2015	euexpo2015	colombiaexpo
maltaexpo2015	expo2015rd	moroccoexpo2015
azerpavilion	chinapavilion15	un_expo2015
msia_pavilion	friulza	israelexpo2015
worldexpo_md		

scale-free nature[7] of social networks [3]. See [1] for one of the earliest work that exploits homophily to predict relationships between individuals. Interested readers are referred to [11] and [16] to deepen their understanding of the state of the art and challenges in link prediction. In our approach, the use of seeds guides the recursive supervised process that identifies homophily patterns and thus constructs the domain graph. This technique has been exploited for the first time in [6] and then in [2].

In the state of the art we found two works that also proposed to use twitter for ontology enrichment. P. Monachesi and T. Markus in [20] proposed an ontology enrichment pipeline that can automatically enrich a domain ontology using: data extracted by a crawler from social media applications, similarity measures, the DBpedia knowledge base, a disambiguation algorithm and several heuristics. The pipeline is very similar to the one proposed in this paper. Differently from their work, we further investigate the idea of exploiting similarity measurements (see the various semantic strategies proposed in Section 3.2.1) and, in particular, we introduce the idea of exploiting the types provided by a domain expert.

C. Wagner and M. Strohmaier [31] investigated a network-theoretic model called *tweetonomy* in order to study emerging semantics. Complementary to our work, they investigated how the selection of the tweets (which they call Social Awareness Streams) can lead to different results. Incorporating their studies in our research is part of our future work.

7. CONCLUSIONS

Results of this study show that our research questions can be positively answered: *social media include low-frequency emerging knowledge that can be extracted by using domain-independent semantic methods.* Our results suffer from the limitations of DBpedia, which offers few types and therefore limited specificity in distinguishing text tokens; on the other hand, we could delegate to a commercial tool the task of matching tokens to DBpedia types. This work is a first step of a larger project on knowledge discovery, whose general framework is illustrated in [5]. Many directions of research remain to be explored, including:

- Enriching feature vectors with other aspects concerning with popularity (e.g. followers, like counts) and content (e.g. sentiment). This extension would alter the ranking in a way that favors most popular candidates.
- Combining multiple social sources and their multimedia content, like of Instagram, Pinterest, and others. This extension applies to domains whose entities are present also through images (e.g., it applies to fashion designers and to Expo pavilions, but not to writers); also with images it is possible to use commercial tools for tag extraction[8] . Preliminary results show that the combined use of Twitter and Instagram over domains well supported by images improves the precision; of course, seeds must be associated both to Instagram and Twitter accounts.
- Using alternative methods with domain-specific tuning. This extension requires higher expert investment in the definition of gold truth, but may also lead to precision improvements.

8. ACKNOWLEDGMENTS

We wish to thank the Fashion In Process Lab[9] at Politecnico di Milano, and especially Paola Bertola, Chiara Colombi and Federica Vacca, who supported us in the definition of the domain-specific knowledge related to the fashion use case.

9. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007.

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

[4] R. Brachman and H. Levesque. *Knowledge Representation and Reasoning (The Morgan Kaufmann Series in Artificial Intelligence)*. Morgan Kaufmann, May 2004.

[5] 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.

[6] S. Brin. Extracting patterns and relations from the world wide web. In *WebDB*, volume 1590 of *Lecture Notes in Computer Science*, pages 172–183. Springer, 1998.

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

[8] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language*

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

[8] For instance, `http://www.clarifai.com`

[9] `http://www.fashioninprocess.com/`

Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal, 2004.

[9] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *KDD*, pages 601–610. ACM, 2014.

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

[11] L. Getoor. Link mining and link discovery. In *Encyclopedia of Machine Learning*, pages 606–609. Springer, 2010.

[12] 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.

[13] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. 2006.

[14] 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.

[15] D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[16] L. Lu and T. Zhou. Link prediction in complex networks: A survey. *CoRR*, abs/1010.0725, 2010.

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

[18] A. Mauri, J. Calbimonte, D. Dell'Aglio, M. Balduini, M. Brambilla, E. D. Valle, and K. Aberer. Triplewave: Spreading RDF streams on the web. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, volume 9982 of *Lecture Notes in Computer Science*, pages 140–149. Springer, 2016.

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

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

[21] H. B. Newcombe and J. M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11):563–566, 1962.

[22] M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[23] S. Quarteroni, M. Brambilla, and S. Ceri. A bottom-up, knowledge-aware approach to integrating and querying web data services. *TWEB*, 7(4):19:1–19:33, 2013.

[24] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB J.*, 10(4):334–350, 2001.

[25] 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 *International Semantic Web Conference (2)*, volume 9982 of *Lecture Notes in Computer Science*, pages 177–185, 2016.

[26] P. Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, volume 50, 2006.

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

[28] 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.

[29] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger. Social media analytics - an interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6(2):89–96, 2014.

[30] 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.

[31] 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.

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