On the Quest for Changing Knowledge

Marco Brambilla, Stefano Ceri, Florian Daniel, Emanuele Della Valle
Data Science Lab, DEIB, Politecnico di Milano. 20133 Milano, Italy
{fisrtname.lastname}@polimi.it

ABSTRACT

For centuries, science (in German "Wissenschaft") has aimed to create ("schaffen") new knowledge ("Wissen") from the observation of physical phenomena, their modeling and empirical validation. With this vision paper, we propose to do so by observing not the physical, but the virtual world, namely the Web with its ever growing stream of data materialized in the form of social network chattering, content produced on demand by crowds of people, messages exchanged among interlinked devices in the Internet of Things, and similar. The knowledge we may find there can be dispersed, informal, contradicting and ephemeral today, while already tomorrow it may be commonly accepted. The challenge is capturing knowledge that is new, has not been formalized yet (e.g., in existing knowledge bases), and is buried inside a big, moving target (the stream of online data). The purpose is to provide data-driven innovation scenarios with the necessary food (up-to-date knowledge) and to do so timely.

CCS Concepts

•Information systems → Data management systems; Web searching and information discovery;

1. INTRODUCTION

In their recent book *Creating Innovation Leaders* [2], Banerjee and Ceri collect a set of contributions that show that creating innovation and innovation leaders first of all requires to exploit appropriate and timely knowledge. With this paper we aim to provide a bottom-up approach to build operational knowledge, a.k.a., domain knowledge, and to capture how it evolves over time, so as to enable innovation leaders to take truly informed decisions. The distinction between data, information, knowledge and wisdom is subtle. The distinction is best expressed by the Data-Information-Knowledge-Wisdom (DIKW) Hierarchy [1]. *Data* are the raw symbols and characters used to represent facts, such as a plain number or a sequence of characters. Data is what machines process and can be communicated, for example through networks like the Internet. *Information* is an interpretation of data that gives meaning and context to the data. For example, the number may be a temperature reading, and the sequence of characters may be a line of text. *Knowledge* in turn puts information into context and expresses patterns. For example, temperature is measured with a thermometer, and text can be written in different languages. In other words, knowledge expresses *entities*, *properties* and *relations* between entities and properties as well as respective instances. *Wisdom* is the understanding of general principles that can range from hard scientific principles to soft ethical/moral principles.

While this hierarchy clarifies well the differences between its elements, it however represents only a *static* picture. What we are interested in is instead the *dynamic* picture, that is, understanding when data becomes information and when information turns into knowledge – all this from a computational point of view. We do not want to talk about wisdom here, as achieving wisdom may require ingredients, for example a soul [3], that are generally out of the reach of computing machinery. Inside this dynamic picture, we are then interested in identifying that knowledge that is of *changing* nature, that is, knowledge that emerges (slowly or suddenly) from the public discourse or that, on the contrary, becomes obsolete at some point in time. The location where the public discourse we want to observe happens is the Web with its social networking sites, blogs, news sites, publication repositories, etc. The challenge lies in the dynamics of the knowledge and in capturing entities, properties and relationships that may be volatile, ephemeral and hidden in the stream of data flowing through the Web in "Internet time."

In general, the process of ontological knowledge discovery [4] tends to focus on the most popular items, those which are mostly quoted or referenced, and is less effective in discovering less popular items, those belonging to the long tail (e.g., the portion of the entity's distribution having a large number of occurrences far from the "head" or central part of the distribution itself). Even the largest knowledge bases are largely incomplete for what concerns low-frequency data. It turns out, however, that knowing the long tail has a strong relevance, e.g., in e-commerce or search[1]. While high-frequency entities include well established brands, low-frequency data typically include *emerging* brands, those that have a small impact today but may have a high one tomorrow. The early discovery of low-frequency data and their ontological properties is thus a very inter-

[1]The commercial success of Amazon and Google is due to their ability to discover goods or pages in the long tail.

DDI'16, May 22-25 2016, Hannover, Germany

ISBN 978-1-4503-4360-2/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2911187.2914582

esting problem, with economic and practical implications in the innovation process.

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. This process has involved huge efforts but had a huge payoff: DBpedia is now the crystallization point of linked data, while Google and Facebook saw the business value of this idea and have hugely invested in continuous and manual integration of databases for the development of knowledge graphs. However, social content has fueled the new discipline of Social Media Analytics [5], concerned with analyzing real world phenomena using social media.

Given these premises, our research more precisely focuses on the problem of **discovering emerging knowledge** belonging to the long tail, by extracting the low-frequency entities and relationships, with their attributes, from social content, thereby enriching existing domain knowledge. We do so by using the methods for crawling social content and for entity recognition which are well established within social media analytics; our notion of ontology is broad, and includes classic cases, such as DBpedia or PubMed, but also any authoritative source of knowledge, such as the NY Stock Exchange Listings, or software projects in Github, or locations available in Open Street Map. These sources are used to define the ontological content of high-frequency entities.

We approach this problem with general, domain independent methods, but also with a well defined focus. We do not attempt at building full knowledge graphs, but rather we build small graphs, called *enriched domain graphs*, where the emphasis is on a given domain, and the enrichment is concerned with emerging concepts extracted from the long tail. Examples are: discovering emerging fashion designers (their identity / trends / brands)[2]; or discovering bloggers or narrative writers; or scouting emerging startups or products while they are becoming popular. Domain knowledge is of course very useful in order to extract the relevant facts about the domain, e.g., high-frequency entities or relationships (thus, we know about Gucci or Prada) or structures from existing knowledge graphs (thus, we know that data about fashion designers can be linked to hubs such as fairs or magazines). We use such domain knowledge as the driver to select and organize relevant social content.

The method takes advantage of initial knowledge, that we call *seeds* and is typically provided by domain experts, to scout relevant *candidates* for the various kinds of emerging knowledge, extracted from social content, and ranked according to a variety of mechanisms, from syntactic to semantic ones, from information retrieval to machine learning, possibly helped by crowdsourcing; the first elements in the ranking are new concepts (e.g., entities or relationships), that can be validated by domain experts or, when confidence is sufficient, entered in the enriched domain graph.

We also plan to use social content to approach the dual problem of **detecting obsolete knowledge**, i.e., of knowledge that may have appeared at a given time but has not been confirmed as it has lost social confirmation. Examples in the medical domain include therapeutic options or theories about diseases which are very popular for a limited amount of time but then they either are ignored or confuted. In this case, we start from domain graphs, i.e., restrictions of knowledge graphs to specific domains, and we solve the dual problems of finding obsolete entities, relationships or attributes, and of discovering that certain types of the domain graph have lost relevance.

[2]This problem is particularly relevant in Milano with its well-known fashion industry; it has been presented to us by the Fashion Design research group within Politecnico.

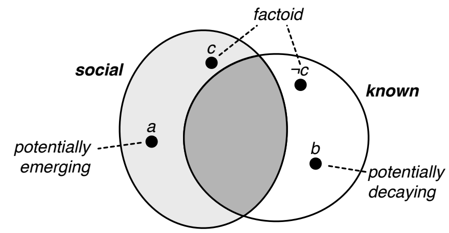

Figure 1: Relation between social knowledge, commonly known and accepted knowledge and factoids.

As an intellectual exercise, we are also interested in **detecting and confuting factoids**, i.e., studying the correctness of the domain graph. Specifically, one can search for factoids, i.e., assumptions or speculations that have been reported and repeated so often that they have become commonly accepted "facts," even though they lack any validity or truth. For instance, the belief that the Great Wall of China is visible from the moon is a factoid, as doing so would require a 17,000 times better eye resolution than we actually have (https://en.wikipedia.org/wiki/Great_Wall_of_China).

2. PROBLEM STATEMENT

Given a domain-independent *knowledge graph*, we consider its restrictions to given domains as a *domain graph*; the main objective of this research is to produce an *enriched domain graph*, where enrichment means adding or annotating concepts, such as entities, attributes or types. We also use the term *social knowledge* to refer to all that knowledge that is part of the public discourse (the Web) but not necessarily part of the above formalized knowledge. Thus, we distinguish a hierarchy of problems of growing complexity. In the first class, we consider problems where the structure of the domain graph is given, and social knowledge is used only for its enrichment by identifying *emerging* knowledge:

- Finding new entities of the enriched domain graph with known type (e.g., `@ABC` is an emerging fashion brand).
- Finding new relationships of the domain knowledge graph (e.g., `@Vogue` published an article about `@ABC`).
- Describing the attributes of such entities and relationships (e.g., `@ABD` has a website `http://abc.com`).

This problems is relevant and unsolved and will be our first focus. If we restrict to problems with invariant domain graphs, two additional problems are relevant:

- Studying the *obsolescence of the domain graph*, i.e., inferring from social knowledge that certain facts of the domain graphs are no longer true.
- Studying the *correctness of the domain graph.* More specifically, we can detect and confute factoids.

Figure 1 illustrates the relationship between social knowledge and commonly accepted domain knowledge. If we restrict knowledge to triples (as in RDF) and we use conventional interpretations for sets and implications, we can represent the problems discussed above as follows:

$$emerge(a) \Leftrightarrow social(a), \neg known(a)$$
$$decay(b) \Leftrightarrow known(b), \neg social(b)$$
$$factoid(c) \Leftrightarrow social(c), known(\neg c)$$

A more complex problem is concerned with the *evolution of the domain graph.* One may discover that the domain graph itself is changing, e.g., discover new types or relationships in the enriched domain graph; for example, one may discover that sports cars are relevant to fashion brands because of new commercial agreements connecting the fashion and sports car markets. A dual problem is discovering that certain types or relationships no longer deserve to be in the domain graph; for example, sports cars may no longer be related to cigarette brands because of the drop of commercial agreements that related them earlier.

Assumptions. In what follows, we exemplarily refer to DBpedia, which is publicly available through its open API, as the generic source of ontological high-frequency knowledge. We use Twitter as social content source, 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"). Next, we enumerate some assumptions used for framing our model. (1) DBpedia *types* are used to partition the existing ontological knowledge; they are organized within a type hierarchy; types which have no descendants are denoted as the (most) *concrete types.* (2) Entities that can be extracted from DBpedia are considered as *high-frequency entities*; each of them is associated to possibly many DBpedia types, including one of them which is their most concrete type. (3) Social content associated with a given Twitter account is analyzed by an Entity Recognition procedure based on DBpedia types; entities identified within a Tweet that refer to any DBpedia type are considered high-frequency entity, while the other entities are considered as *low frequency.*

3. APPROACH OUTLINE

3.1 Finding New Knowledge

We formalize the problem as follows. Given:

- $S = \{s_1, \ldots, s_n\}$, a set of seeds (i.e., low frequency entities of a given type E) provided by a domain expert for a given domain;
- $T = \{t_1, \ldots, t_m\}$, a set of types, also chosen by the expert as relevant for the given domain.

We use T for defining the features of the seeds and then use them to solve the following problems:

1. Discover other low-frequency entities $C = \{c_1, \ldots, c_k\}$ of type E, denoted as *candidates*, which will be identified as *emerging low-frequency entities.*
2. Find relationships between each $c_i \in C$ and high-frequency entities of types T.

Figure 2 shows some of the ingredients of our problem. DBpedia contains both types and high-frequency instances, which are related among them and to the types to which they belong. One of the types, called central type, is the most relevant type of a given domain (e.g., the type of fashion designers); other relevant types of the domain connect to it. The collection of relevant types and of their instances form the domain graph. Then, some of the low-frequency entities of the central type act as seeds; problem (i) is concerned with finding good candidates, i.e., other low-frequency entities that are instances of the central type; problem (ii) is concerned with finding relationships between seeds, candidates, and high-frequency instances.

We assume that we know the social media handles of seeds; we also assume that, given a selected set of features, the entities of a given type are very similar among each other and noticeably different from the entities of any other type; thus, a simple approach based on vector feature similarity can solve the problem; such simple approach can be refined in many ways in order to improve precision.

3.2 Finding Obsolete Knowledge

We formalize the problem as follows. Given:

- f, a fact, i.e., a high-frequency entity of a given type E, belonging to a domain indicated by a domain expert;
- $T = \{t_1, \ldots, t_m\}$, a set of types, also chosen by the expert as relevant for the given domain;
- $R = \{r_1, \ldots, r_n\}$, the set of high-frequency relationships of f inside the knowledge base.

We use again T for defining the features of the fact and then use them to solve the following problems:

1. Discover low-frequency entities $FB = \{fb_1, \ldots, fb_l\}$ of type E, denoted as *fact backers*, whose presence below a given threshold tf_{min} (including complete absence) will be interpreted as a *decay of the fact* f.
2. For each of the relationships $r \in R$, discover similar low-frequency relationships $RB = \{rb_1, \ldots, rb_k\}$, denoted as *relationship backers*, whose presence below a given threshold tr_{min} (including complete absence) will be interpreted as a *decay of the relationship* r.

Pruning a domain graph does therefore not ask for new low-frequency seeds, as the problem is no longer to identify a set of similar low-frequency entities to build a new high-frequency entity, but rather to confirm or not the presence of a sufficient number of low-frequency entities to confirm the validity of an already given high-frequency entity.

The challenge of deciding if a given fact f or relationship r decays is understanding which are the minimum threshold values tf_{min} and tr_{min} that allow one to take a decision. Also, the very choice of the fact or relationship to study is a problem on its own, as in general there may be different "types" of knowledge that may not lead to the presence of respective low-frequency entities on the Web. In this respect, we take a very pragmatic view on the problem and ask domain experts to define and delimit a given domain of interest, containing facts with own relationships to study.

3.3 Unmasking Factoids

We formalize the problem as follows. Given:

- f^*, a presumed factoid, i.e., an entity of a given type E or a relationship among given entities part of the social knowledge, indicated by a user;

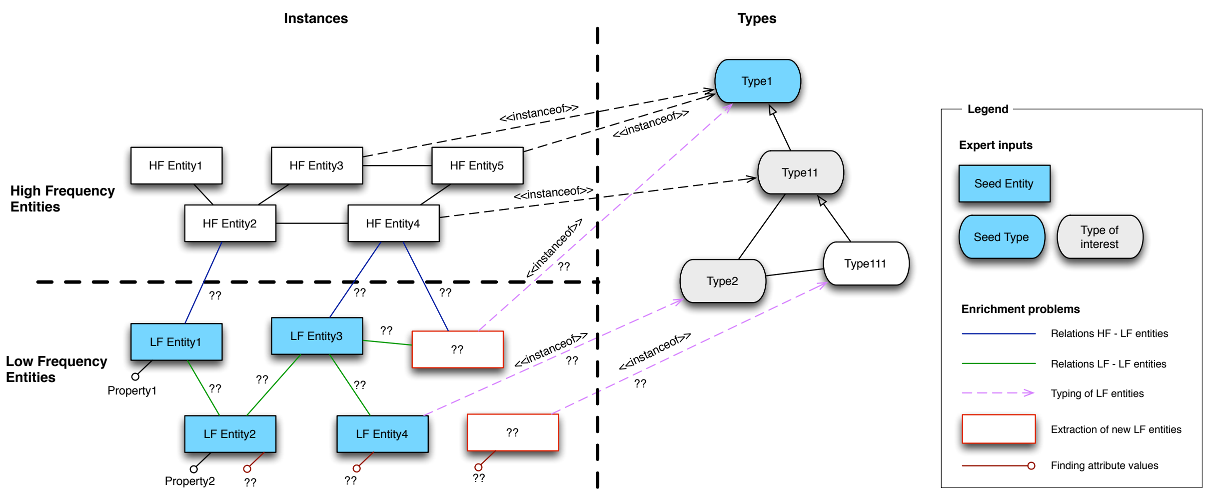

Figure 2: The role of types, instances, high-frequency instances (common knowledge) and low-frequency instances (social knowledge) in the problem setting.

– $T = \{t_1, \ldots, t_m\}$, a set of types, also chosen by the user as relevant for the domain of the presumed factoid.

We use again T for defining the features of the presumed factoid and then use them to solve the following problems:

1. Discover high-frequency entities or relationships $OO = \{oo_1, \ldots, oo_k\}$ that negate the factoid f^*, denoted as *factoid opposers*, whose presence (even only one) will be interpreted as *invalidation of the factoid*. The absence of opposers does not allow to take a decision.

2. For an invalidated factoid, discover low-frequency entities $OB = \{ob_1, \ldots, ob_n\}$ of type E or type relationship, denoted as *factoid backers*, whose presence above a given threshold to_{min} will be interpreted as *social acceptance of the factoid*. The number of factoid backers serves as a measure of the social acceptance.

Unmasking factoids thus requires an additional ingredient compared to finding new or obsolete knowledge, i.e., the capability to understand and match positive and negative assertions about a given entity or relationship. The respective interplay of social and commonly accepted knowledge is, for instance, well exemplified by the conviction about the visibility of the Great Wall of China from the moon (the factoid, the positive assertion) contradicted by Wikipedia (the negative assertion contained in a knowledge base); another example is the claim that autism is correlated with vaccines. Again, automatically checking all low-frequency entities would be computationally unfeasible, and we fall back to a user providing a presumed factoid as input. An interesting aspect of factoids is understanding how widely they are accepted by the crowd of social network users.

4. PRELIMINARY RESULTS

In order to validate our vision and demonstrate the feasibility of the approach, we ran a first batch of experiments regarding the first idea presented in this paper: finding new knowledge. Specifically, we extracted knowledge from Twitter in three very different domains:

- **Fashion**: we considered the problem of identifying emerging fashion brands, which are not yet globally recognized and thus are not present in the knowledge graph. Domain experts provided us with 200 emerging brands in the Italian market, and we discovered others.

- **Literature**: we considered the problem of identifying non-famous writers, starting from a set of writers engaged in a literature event in Australia.

- **Live events**: we considered the domain of the Universal Exposition (EXPO 2015) that took place in Milan last year, and we constructed knowledge about the exhibition pavilions, given a limited set of known ones.

Fashion is characterized by a very high concentration of the domain in a few brands, most of which are known; on the opposite, literature is a quite open domain where authors can be considered widespread; and live events typically count a very small number of entities of interest (e.g., pavilions in Expo were around 100) and have a short duration.

The analysis for knowledge extraction we applied to all the domains can be summarized as follows. Starting from the set of seeds provided for each domain:

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

2. 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 at least two seeds (which leads to sets counting hundreds of thousands of candidates);

3. selection of the top candidates (in the order of thousands of instances) according to a ranking calculated through a variant of a td–idf measure;

4. definition of a vector representation for each seed and top candidate;

5. representation of seeds as one single seed prototype (calculated as the centroid of one single seed cluster), assuming that seeds are somehow homogeneous;

6. ranking of the candidates based on the distance from the seed prototype;

7. selection of the entities (of the same type of the seed) to be added to the knowledge graph, assuming that the candidates that are closer to the seed prototype are most promising.

Notice that our contribution is not focused on improving entity extraction, NLP, or machine learning algorithms. We rely on existing techniques for the low-level content analysis. We focus instead on identifying and selecting, among a large number of extracted candidates, the ones that are most likely represeting instances of the expert types of interest. This selection is done by looking into what these entities talk about and by comparing this with already identified good candidates or seeds.

In order to maximize the quality of the selected set of candidates, we defined a parameterization of the pipeline, which generated about 4,400 strategies. These strategies consisted of:

- 44 alternative feature vector configurations for describing the entities, including 12 basic feature selection alternatives and 32 aggregated strategies that pair the 12 basic alternatives in different ways and with different weights α;

- 9 different values for α, in the range [0.1-0.9], used for weighting the two contributions in the combined strategies above;

- 5 levels of recall [0.00, 0.25, 0.50, 0.75] for the entity extraction algorithms applied to the posts;

- 3 distance measures (cosine, correlation and euclidean) used for assessing the distance of each candidate from the seed prototype.

While the baseline strategy was a purely syntactical one (basically calculating text similarity among the tweet streams of the candidates), our alternative strategies considered different ways to build the feature vector of a candidate starting from the semantic analysis of its tweet stream. For instance, the semantic strategies build the feature vectors as: any identified concept instance (i.e., belonging to any type in the KB); only instances belonging to "leaf" types in the KB (assuming a taxonomy of types); only instances belonging to a set of types relevant to the domain (provided by the domain expert); considering the frequency of the types only, instead of the frequency of the instances; and so on.

All the strategies were executed on the three domains, and their results were assessed against the respective ground truth. At this point, in order to identify the best strategies, we applied a greedy pruning algorithm that iteratively dropped the parameter value that caused the least impact on the number of high quality results. This was repeated until we noticed that removing another parameter value would drop more good results than bad ones. Since our aim was to identify the top candidate entities, we compared the performance of the strategies considering precision@10 (precision of top-10 candidates).

Figure 3 shows the different steps of the pruning algorithm. In each step we report the parameter values that were dropped, the number of parameter combinations available, and the average cross-domain probability of randomly choosing a parameter combination whose precision@10 is greater than the best purely syntactic strategy.

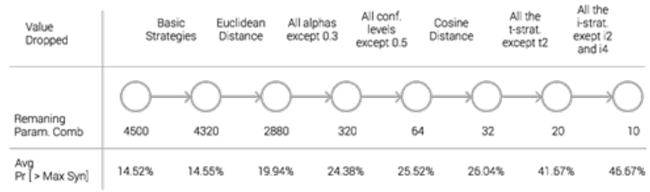

Figure 3: Steps of the pruning algorithm.

At the end of the pruning, we were able to select 10 strategies. Those strategies granted precision@10 equal to 40% in the fashion domain (the hardest one, as we started from a very complete set of seeds already, and therefore finding new elements resulted extremely challenging), 50% in the event domain, and 70% in the writers domain.

For the fashion domain, we also performed an experiment on Instagram as a source of knowledge, where images were tagged through a deep learning approach for extracting the concepts captured by the photos. In that case, the parameterization was simpler (basically, only 3 different strategies were available) but the results were slightly better for the best strategy (60% precision@10). This is due to the richness of the photographic medium for the specific domain of fashion, where photos are crucial in the communication strategies of brands.

5. CONCLUSION

In today's information society, innovation is driven by the quality and availability of suitable information and, hence, knowledge. Failing to keep pace with the dynamics of knowledge undermines innovation and means lagging behind in the competition. Our vision is to help everybody, especially domain experts (the potential innovators), to be aware of how knowledge evolves, laying the foundation for data-driven innovation. Our preliminary results provide substantive evidence that there indeed is considerable knowledge that is not yet captured by consolidated KBs, such as DBpedia, but that we are able to capture. At the current, preliminary stage, our approach has proven to be flexible across domains and reasonably scalable. We are currently refining our strategies for knowledge extraction, expanding to other data sources, and covering new domains. Next, we will study the spatio-temporal dynamics of knowledge.

6. REFERENCES

[1] R. Ackoff. From data to wisdom. *Journal of Applied Systems Analysis*, 16:3–9, 1989.

[2] B. Banerjee and S. Ceri, editors. *Creating Innovation Leaders: A Global Perspective*. Springer, 2016.

[3] G. Bellinger, D. Castro, and A. Mills. Data, Information, Knowledge, and Wisdom. http://www.systems-thinking.org/dikw/dikw.htm, 2004.

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

[5] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger. Social media analytics. *Business & Information Systems Engineering*, 6(2):89–96, 2014.