Discovering the topics of a data source: a statistical approach*

Sonia Bergamaschi[1], Davide Ferrari[2], Francesco Guerra[1], and Giovanni Simonini[1]

[1] Università di Modena e Reggio Emilia, Italy
sonia.bergamaschi@unimore.it
francesco.guerra@unimore.it
giovanni.simonini@unimore.it
[2] The University of Melbourne, Australia
dferrari@unimelb.edu.au

Abstract. In this paper, we present a preliminary approach for automatically discovering the topics of a structured data source with respect to a reference ontology. Our technique relies on a signature, i.e., a weighted graph that summarizes the content of a source. Graph-based approaches have been already used in the literature for similar purposes. In these proposals, the weights are typically assigned using traditional information-theoretical quantities such as entropy and mutual information. Here, we propose a novel data-driven technique based on composite likelihood to estimate the weights and other main features of the graphs, making the resulting approach less sensitive to overfitting. By means of a comparison of signatures, we can easily discover the topic of a target data source with respect to a reference ontology. This task is provided by a matching algorithm that retrieves the elements common to both the graphs. To illustrate our approach, we discuss a preliminary evaluation in the form of running example.

1 Introduction

Data-intensive applications (e.g., e-commerce applications, digital libraries, ...), which rely on the information stored in private databases, are now common over the Internet. The data behind the application is in general not accessible by external applications and represents the so-called deep web. The value derived from the re-use of this kind of data has been considered of paramount importance for both research and business activities.

For this reason, the research community has put a lot of effort in the last years for searching information in the deep web [13] and extracting knowledge from it [10]. Three key factors have recently affected this well-known architecture for web applications:

1. **The Semantic Web vision is now reality.** Research outcomes have provided standards, techniques and tools enabling the Web to move from a "Web of Documents", where the data are typically optimized for the direct human-consumption, to a "Web

* The authors would like to acknowledge the networking support by the COST Action IC1302 (www.keystone-cost.eu)

of Data", where the data is structured thus making it more efficiently and effectively usable by software applications. Moreover, data from different sources can be linked with each other, thus fostering the interoperability of the information. This is the Semantic Web vision, which is now becoming reality.

2. **A large amount of structured data is available in the web.** Public Sector and Enterprises have started to consider the web as the primary place for publishing their data with structured and open formats. The EU Commission, for example, is promoting the publication and the reuse of public sector information as open data, so that it can be publicly accessible by other Institutions and Enterprises[1].
3. **Applications for Data Analytics are now handy.** Big data has become a hot trend topic. Data science is now a common term and denotes techniques and applications for extracting knowledge from data [6]. Several software packages, tools, and case studies are now available for managing, extracting, transforming and analyzing large amount of data.

These elements have radically changed the ways for accessing structured data: several data sources are "emerged" from the deep web and are available as open data. A direct user consumption of structured data (not mediated by any application) is now possible. Nevertheless, in this scenario, a new problem arises: how to find the data sources satisfying specific information needs. The usual paradigm for accessing information in the web is based on search engines as the entry points for users looking for information. Unfortunately, search engines cannot be an actual solution since they are not conceived for indexing structured data. Consequently, source retrieval can be a critical task. Some solutions have been proposed to deal with this problem: web portals (see for example the European Union Open Data Portal[2]) can support this task. In portals, the data sources are in general indexed on the basis of some metadata (e.g., title, author, content description, ...) manually provided by the data source owner. This is a tedious and error prone work that can generate biased results if the metadata have not been accurately selected. An automatic data-driven approach for extracting metadata from a target source can help managing this issue.

In this paper, we introduce a preliminary proposal for automatically discovering the topics of a target data source with respect to a reference ontology. Our approach relies on three key elements: a reference ontology, an algorithm for computing the "signature" of a data source, and a graph matching algorithm. We conceive the *reference ontology* as a vocabulary of concepts and related properties describing a real world domain. The *signature* is a concise weighted graph-based representation of the data source topics which is independent of the specific vocabulary adopted in the source (i.e. labels used to describe schema elements, and domains associated to the attributes). In particular, nodes represent concepts and attributes. Edges model three kinds of relationships: relationships between attributes belonging to the same concepts, relationships between concepts and the respective attributes and relationships between attributes and concepts. Entropy is used to weight nodes, thus giving an account of their importance in terms of information power. Mutual information is used to provide weights associated to edges,

[1] Digital Agenda for Europe, http://ec.europa.eu/digital-agenda/, Pillar I, Action 3
[2] https://open-data.europa.eu/

thus measuring the correlation between the involved nodes. We claim that such a signature can be used as a semantic identifier of a domain, i.e. two sources representing the same subject have a similar signature independently of the actual attribute domains adopted. The technique for extracting the signature is a critical task and represents the main contribution of the paper. Finally, a *graph matching* algorithm is used for comparing the signatures of the reference ontology and the target data source. The goal is the identification of possible matches which correspond to concepts in the ontologies described in the source.

Without loss of generality, in the following, we will focus on RDF data sources. RDF is becoming a standard way for publishing structured data on the web and several sources are available[3]. Moreover, working with RDF allows us to use DBpedia as reference ontology[4]. DBpedia is a large knowledge base derived from Wikipedia, which currently describes 4.0 million of "things" with 470 million of "facts". Thus we can easily evaluate our approach in different domains with different data sets. We will experiment two ways for computing graph weights: one based on the classical computation of entropy and mutual information, the second based on composite likelihood to estimate those values.

The main advantage derived by the estimation of the weights is to reduce the sensitivity to a specific type of estimation error related to underestimation of the probability of rare labels combinations. The classic mutual information is known to be very sensitive to regions corresponding to small probabilities; thus, when label combinations are rare, assigning graph weights based on mutual information is expected to produce unstable results. This motivates the introduction of a composite divergence measure based on a linear combination of divergences. Estimation is based on a composite likelihood methodology, a well-known approach for complex models that has proved useful in statistics and machine learning; see [12] for an exhaustive overview. Our preliminary empirical results suggest increased reliability of the new approach based on out-of-sample performance on real data.

Finally, in this preliminary proposal, we do not investigate any advanced technique for graph matching. We adopt the distance-measure proposed in [9] for evaluating signature matches. Summarizing, the main contributions of this paper are: 1) a model for defining signatures representing topics of RDF sources based on schema information, entropy and mutual information; 2) a technique for computing the estimation of the signature weights; and 3) a preliminary evaluation of our proposal by means of a running example.

The rest of the paper is organized as follow: Section 2 introduces the problem, Section 3 describes our proposal for estimating the weights and in Section 4 a running example provides the reader an intuition of our approach. Section 5 describes some related work and finally in Section 6 we sketch out some conclusion and future work.

[3] See for example http://linkeddata.org/data-sets for a list of possible data sets.
[4] http://dbpedia.org/

2 Problem statement

Let us consider RDF sources with a RDFS schema as a *Knowledge Base*. We model the schema information as a *total dependency graph* where each node represents either a concept or a property of the knowledge base, and edges can represent: 1) E_k^{PP}, relationships between properties related to the same concept k (i.e. there is an edge between two nodes representing properties if the properties have the same concept as domain), 2) E_k^{CP}, relationships between concepts and properties (i.e., there is an edge between a node representing a property and a node of representing concept indicated as domain), and 3) E_k^{PC} relationships between properties and concepts (i.e., there is an edge between a node representing a property and a node of representing concept indicated as range, if any). For completeness, in the rest of the section we provide a formal definition of the signatures[5].

Definition (Knowledge base) *Let L be the set of literals, U the set of URIs. A knowledge base is a set of triplets $KB \subset (U \times U \times (U \cup L))$. We use $R = \{r \in U \mid \exists (s,p,o) \in KB : (r = s \vee r = o)\}$ to represent the set of resources, $P = \{p \mid \exists s,o : (s,p,o) \in KB\}$ to represent the set of properties, and $C = \{c \mid \exists s : (s\ rdf:type\ c) \in KB\}$ to represent the set of concepts.*

Definition (Properties of a concept) *Given a concept $k \in C$, the set of properties of k, P_k is defined as $P_k = \{p \mid \exists r_1, r_2 \in R, p \in P : (r_1, p, r_2) \in KB \wedge (r_1\ rdf:type\ k) \in KB\} \cup \{p \mid \exists r_1 \in R, p \in P, l \in L : (r_1, p, l) \in KB \wedge (r_1\ rdf:type\ k) \in KB\}$.*

The properties of a concept need to be better qualified for the definition of the total dependency graph. In particular, we define

Definition (Set of Property-to-Property (P-P) relationships) *Given a concept $k \in C$, we define its set of Property-to-Property (P-P) relationships E_k^{PP} as the set of relationships existing between properties having the concept k as domain. More formally, $E_k^{PP} \equiv (P_k \times P_k)$.*

Definition (Set of Concept-to-Property (C-P) relationships) *Given a concept $k \in C$, we define its set of Concept-to-Property (C-P) relationships E_k^{CP} as $E_k^{CP} = \{(k, p_j) \mid p_j \in P_k\}$. This is the set of relationships holding between a concept k and its properties P_k.*

Definition (Set of Property-to-Concept (P-C) relationships) *Given an object property, its range concept is defined as $C_p^{range} = \{k \mid \exists o \in R, \exists s : (s,p,o) \in KB \wedge (o\ rdf:type\ k) \in KB\}$. We define the set of Property-to-Concept (P-C) relationships of a concept $k \in C$, as $E_k^{PC} = \{(p_j, c) \mid p_j \in P_k \wedge c \in C_{p_j}^{range}\}$. This is the set of relations between properties and their target concepts.*

The Total Dependency Graph summarizes all these kinds of semantics in a unique graph as follow.

[5] We extend the notation used in [16]. In a similar way, we do not model blank nodes to keep the presentation clear.

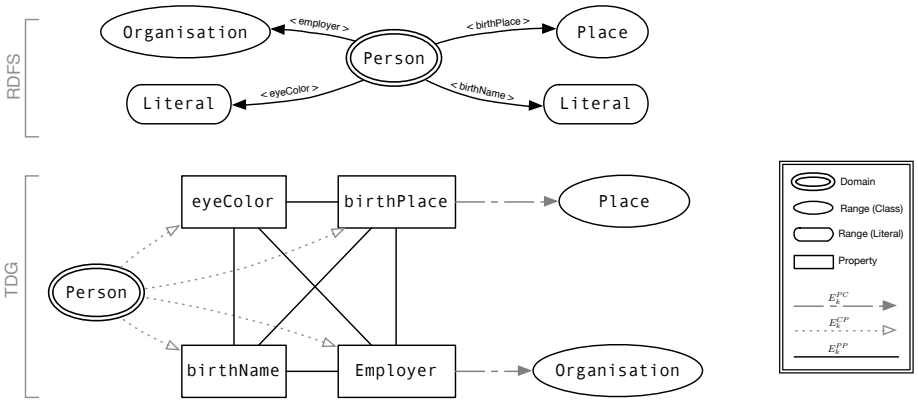

Fig. 1. A simple RDF schema and its TDG.

Definition (Total Dependency Graph - TDG) *A Total Dependency Graph - TDG is the quintuple* $TDG = (C, P, E^{PP}, E^{CP}, E^{PC})$ *where:* C *and* P *are set of nodes, representing concepts, and properties respectively, and* E^{PP}, E^{CP}, E^{PC}*) are sets of edges denoting:*

- $E^{PP} = \bigcup_{k \in C} E_k^{PP}$*, the union of the sets of the P-P relationships built for each concept in the knowledge base;*
- $E^{CP} = \bigcup_{k \in C} E_k^{CP}$*, the union of the sets of C-P relationships built for each concept in the knowledge base;*
- $E^{PC} = \bigcup_{k \in C} E_k^{PC}$*, the union of the sets of P-C relationships built for each concept in the knowledge base.*

Fragments of a Total Dependency Graph concerning only one node associated to a concept are simply called *Dependency Graphs*.

Definition (Dependency Graph - DG) *The Dependency Graph of a (reference) concept* $k \in C$ *is defined as* $DG_k = (P_k, E_k^{PP})$.

Example 1. In Figure 1 a small RDF Schema and its corresponding TDG are shown. Classes and properties of the RDF schema are transformed into nodes in the TDG. Edges between nodes representing properties (E^{PP}), connecting the reference class with its attributes (E^{CP}), and connecting properties with external classes (E^{PC}) are shown.

The *signature* of a structured source is represented as a TDG with weigthed nodes and E^{PP} and E^{CP} edges. E^{PC} are not weighted since they represent possible connections between classes in the data source.

3 Composite likelihood estimation of signatures

A characterization of signatures is carried out by determining a weighted graph that summarizes the topics of a source. The weights in such graphs can be assigned using two basic information-theoretical quantities: entropy and mutual information [9]. After defining such quantities, we discuss their estimation based on data samples by composite likelihood techniques.

Definition (Entropy) *Let X be a random variable representing an attribute with alphabets $\mathcal{X}$ and probability mass function $p(x|\theta)$, with unknown parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. The entropy $H(X)$ is defined by $H(X) = -E_X \log p(X|\theta)$, where $E(\cdot)$ denotes expectation with respect to $p(x|\theta)$.*

Note that the above definition does not involve realized values for data instances, thus making the signature independent of the class represented. In particular, entropy describes the uncertainty of values in an attribute. Thus, one problem is estimation of $H(X)$ from available data instances by means of some appropriate approximation of $p(x|\theta)$. If n samples of X are available, then an estimate $\hat{\theta}$ can be obtained by some statistical estimation method, such as maximum likelihood estimation, so that $H(X)$ could be estimated by using $\theta = \hat{\theta}$ in the definition above. To measure the information shared by two attributes at the time we introduce the concept of mutual information.

Definition (Mutual Information) *Let X and Y be two random variables representing attributes with alphabets $\mathcal{X}$ and $\mathcal{Y}$ with joint mass function $p(x, y|\theta^{XY})$ and marginal mass functions $p(x|\theta^X)$ and $p(y|\theta^Y)$. The mutual information of X and Y is:*

$$I(X;Y) = E_{XY}\left[\log \frac{p(X,Y|\theta^{XY})}{p(X|\theta^X)p(Y|\theta^{XY})}\right] = H(X) + H(Y) - H(X,Y) \qquad (1)$$

where $H(X)$ and $H(Y)$ are marginal entropies for X and Y and $H(X,Y)$ is the entropy for the pair (X,Y).

Firstly, note that $I(\cdot;\cdot)$ measures different levels of association (or shared information) between pairs of nodes. If the association is strong, then the estimated joint frequency $p(x, y|\theta^{XY})$ is large compared to the estimated frequency of separate nodes, $p(x|\theta^X)$ and $p(y|\theta^Y)$. Secondly, note that similarly to entropy, also the mutual information needs to be estimated from data instances. To estimate $I(X;Y)$ we need to obtain parameter estimates $\hat{\theta}^X$, $\hat{\theta}^Y$, and $\hat{\theta}^{XY}$. In our approach, entropy and mutual information are computed by means of the cardinality of the URI for nodes representing concepts and the cardinality of the range for nodes representing properties.

We remark that in the proposed TDG, a node can assume different roles (i.e., it can be the reference concept in a DG, the range value of several properties according to the source schema). This means that, depending on the role considered, a concept can assume different cardinalities: it may vary from its maximum value, when a concept node is considered as alone in a DG, to a number of other possible values, one for each property it is involved. As a consequence, its entropy and the mutual information of its edges may assume different values.

The method for estimating $H(X)$, $H(Y)$ and $H(X,Y)$ from data samples is crucial to obtain representative signature. A suitable method should be able to prevent over-fitting. The estimated signature does not have to perfectly replicate a specific data source, but rather provide us with a synthetic representation of a reference ontology which, in turn, should describe a more complex real world. Over-fitting is important in the presence of very large alphabets for the attributes under exam, with only a few observed instances. The elements of the alphabets with very low frequency typically inflate the overall noise thus deteriorating the quality of the available information.

Another issue related to the high dimensionality of the problem is computing. The high number of instances usually collected in the RDF knowledge bases available online makes the calculation of the actual values expensive from the computational point of view. For example, the class Person (one of the 529 classes which form a subsumption hierarchy) of the DBpedia Ontology (version 3.9) contains 832.000 instances and has 101 properties. This means that the cardinality of the set E^{PP} built considering only the class Person is 5,050. To address the above issues, we propose an approach for an approximate computation of entropy and mutual information.

3.1 Parameter estimation

Let Y be an attribute of a binary alphabet $y = (y_1, \ldots, y_q) \in \{0,1\}^q$. If the discrete alphabet is not binary we convert it into a binary alphabet. The full dependency of q labels can be represented by the joint distribution $p(y|\theta)$, $\theta \in \Theta$. We consider a composite likelihood function constructed from marginal models, $p(y_i|\mathrm{x}, \theta_i)$ $(i = 1, \ldots, q)$, and pairwise models, $p(y_i, y_j|\theta_{ij})$ $(1 \leq i < j \leq q)$. Here $\{\theta_i\}$ and $\{\theta_{ij}\}$ are two sets of parameters vectors for univariate and bivariate models. The marginal and pairwise densities are combined to form the composite model

$$p(y|\theta, w) = \prod_{i=1}^{q} p(y_i|\theta_i)^{w_i} \prod_{i<j} p(y_i, y_j|\theta_{ij})^{w_{ij}}, \tag{2}$$

where $w = (w_{ow}^T, w_{pw}^T)^T$ is a vector including nonnegative elements $w_{ow} = \{w_i : i = 1, \ldots, q\}$ and $w_{pw} = \{w_{ij} : 1 \leq i < j \leq q\}$. These are importance parameters determining the contribution of the marginal and pairwise in the composite likelihood function. Estimation of the joint distribution of two attributes (X, Y) will be analogous.

Given a set of N training samples $D = \{y^n\}_{n=1}^N$, we compute the maximum composite likelihood estimator (MCLE), $\hat{\theta}$, defined as the maximizer of the log-composite likelihood function

$$\ell(\theta, w) \equiv \sum_{n=1}^{N} \log p(y^n|\theta, w) = \prod_{n=1}^{N} L(\theta, w|y^n). \tag{3}$$

Since parameters in different sub-likelihood components are independent, the MCLE may be computed by maximizing separately marginal and pairwise log-likelihood func-

tions:

$$\hat{\theta}_i = \underset{\theta_i}{\text{argmax}} \sum_{n=1}^{N} \log p(y_i^n|\theta_i), \;\; 1 \leq i \leq q, \tag{4}$$

$$\hat{\theta}_{ij} = \underset{\theta_{ij}}{\text{argmax}} \sum_{n=1}^{N} \log p(y_i^n, y_j^n|\theta_{ij}), \;\; 1 \leq i < j \leq q. \tag{5}$$

Therefore, the problem of maximizing (3) is divided into $q + q(q-1)/2$ separate optimization tasks involving the estimation of q binary classifiers and $q(q-1)/2$ 4-class classifiers. Although the computational complexity of the above task is manageable, the policy of keeping all available likelihood components is not well justified in terms of efficiency relative to MLE, since inclusion of redundant factors can deteriorate dramatically the variance of the corresponding composite likelihood estimator [5]. A better strategy would be to choose a subset of likelihood components which are maximally informative, and drop noisy or redundant components to the maximum extent.

3.2 Weights selection

The estimated one- and pair-wise models, $f(y_i^n|\hat{\theta}_i)$ $(1 \leq i \leq q)$ and $f(y_i^n, y_j^n|\hat{\theta}_{ij})$ $(1 \leq i < j \leq q)$, are combined by composite likelihood decomposition in (2), according to the vector, w. The importance parameter w_j is selected to be small when, for a value of θ that is appropriate for the majority of the data subsets, the likelihood function for the jth data subset is relatively large. To this end, we use the importance scheme often used for model combining (see [4] and references therein). For a given $\alpha > 0$, we compute

$$\hat{w}_i \propto \exp\left\{\alpha \sum_{n=1}^{N} \log p(y_i^n|\hat{\theta}_i)\right\}, \;\; 1 \leq i \leq q, \tag{6}$$

$$\hat{w}_{ij} \propto \exp\left\{\alpha \sum_{n=1}^{N} \log p(y_i^n, y_j^n|\hat{\theta}_{ij})\right\}, \;\; 1 \leq i < j \leq q. \tag{7}$$

where $\{\hat{\theta}_i\}$ and $\{\hat{\theta}_{ij}\}$ are maximum likelihood estimates computed in the previous section. The method is a type of regularization approach that favors simpler likelihoods by producing weights tending to zero as α increases. For sufficiently large α, incompatible sub-models are down-weighted to the maximum extent, thus resulting in sparse composite likelihood objects.

3.3 Estimation of mutual information

The mutual information defined in (1) is estimated by replacing the distributions $p(x, y)$, $p(y)$ and $p(x)$ by empirical counterparts estimated by the composite likelihood approach described above. Particularly, we propose to approximate the entropy of

attributes X and Y by the fitted composite likelihood functions

$$\hat{H}(X) = \ell(\hat{\theta}^X, \hat{w}^X) = \sum_{n=1}^{N} \log p_X(x^n|\hat{\theta}^X, \hat{w}^X) \tag{8}$$

$$\hat{H}(Y) = \ell(\hat{\theta}^Y, \hat{w}^Y) = \sum_{n=1}^{N} \log p_Y(y^n|\hat{\theta}^Y, \hat{w}^Y) \tag{9}$$

where p_X, p_Y are composite models for X and Y defined as in (2) and $\{\hat{\theta}^{(X)}, \hat{w}^{(X)}\}$ and $\{\hat{\theta}^{(Y)}, \hat{w}^{(Y)}\}$ the corresponding sets of parameter estimates. Similarly, the entropy of the variable pair (X, Y) is approximated by

$$\hat{H}(X,Y) = \ell(\hat{\theta}^{XY}, \hat{w}^{XY}) = \sum_{n=1}^{N} \log p_{XY}(x^n, y^n|\hat{\theta}^{XY}, \hat{w}^{XY}) \tag{10}$$

where p_{XY} is the composite model for the variable pair (X, Y) and $\{\hat{\theta}^{(YX)}, \hat{w}^{(YX)}\}$ denote parameter estimates for the joint model. As a measure of mutual information, we propose to use the following empirical approximation of the mutual information I based on the fitted likelihood functions:

$$\widehat{I}(X;Y) = \sum_{n=1}^{N} \log \left\{ \frac{p_{XY}(x^n, y^n|\hat{\theta}^{XY}, \hat{w}^{XY})}{p_X(x^n|\hat{\theta}^X, \hat{w}^X) p_Y(y^n|\hat{\theta}^Y, \hat{w}^Y)} \right\} = \hat{H}(X,Y) - \hat{H}(X) - \hat{H}(Y). \tag{11}$$

Note that to avoid taking expectation over alphabets $\mathcal{X}$ and $\mathcal{Y}$ as in (1), the expectation is replaced by summation over observations (x^n, y^n), $1 \leq n \leq N$, in (11). Hence, the summands in the right hand side of (11) correspond to the fitted composite likelihood function as defined in (3). This approach reduces considerably the computational burden compared to the exact approach.

4 Motivating Example

The DBpedia ontology (version 3.9) conceptualizes the real world through a hierarchy structure made of 610 classes (see the DBpedia website for more details[6]). Each class comprises a rich set of datatype and object properties (e.g., the class Person includes more than 3k properties), and a large number of instances is provided for most of the classes (e.g., there are more than 760k instances belonging to the class Person in the English version, more than 300k belonging to the class Work).

The goal of this preliminary evaluation is to show that signatures can effectively represent topics. For reaching this purpose, we performed three experiments and we tested if: 1) Casual partitions of the instances of the same concept provide similar signatures; 2) The signatures of a concept and the one of some superset concept are close; 3) The signatures of two not related concepts are different. We started our evaluation

[6] http://wiki.dbpedia.org/Datasets/DatasetStatistics

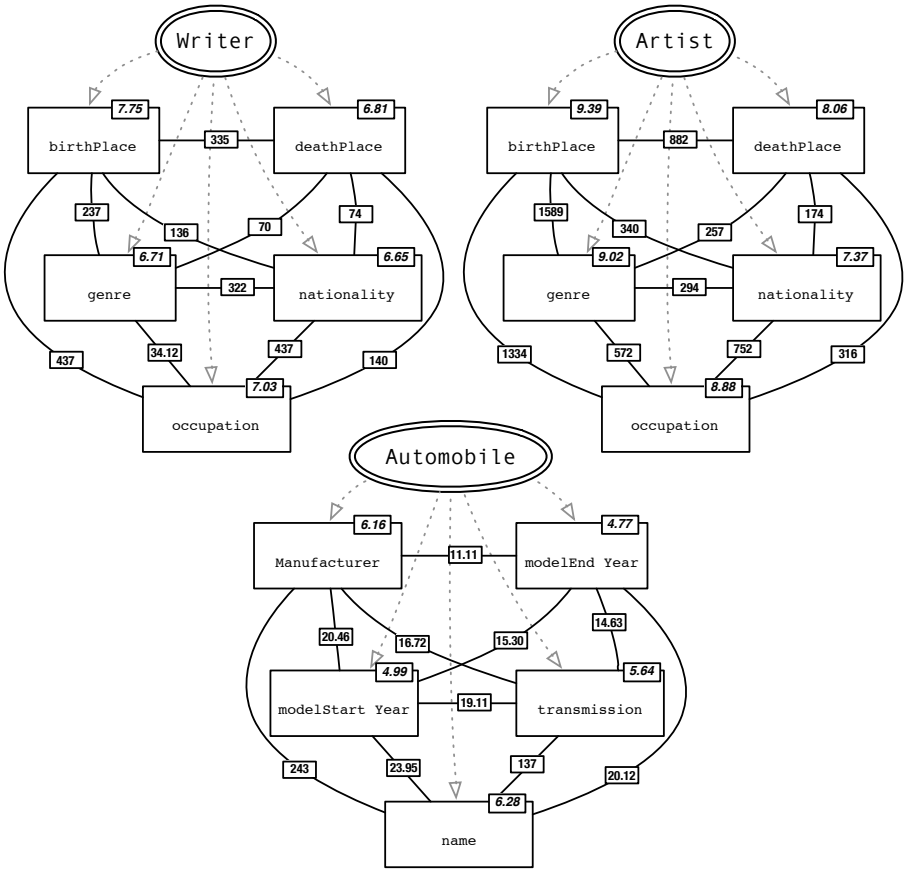

Fig. 2. Three TDGs from DBpedia classes. The values in the boxes are estimated with the proposed technique.

by selecting three classes from DBpedia (*Writer, Artist, Automobile*) and building their TDGs as shown in Figure 2. The first signature represents a fragment of the DBpedia Writer class, including only five representative properties for simplicity. The second TDG describes the Artist class, i.e. a superclass of Writer. Note that both classes share the same properties. Finally, the third TDG represents five properties of the Automobile class. In Figure 2, we show also the weights for representing the actual and the estimated values (in the boxes) of Entropy (on the nodes) and Mutual Information (on the edges).

Since we are interested to evaluate the specific contribution provided by entropy and mutual information alone, we performed separate evaluations, by considering firstly only the nodes (thus measuring the contribution of the entropy) and secondly the edges (thus measuring the contribution of the mutual information). We adopted a Euclidean distance-based metric as, in [9], defined as follows. Let A and B be two equal size

dependency graphs and a_i, b_j the entropy of the node i and j in graph A and B, respectively. Let m be an index that maps a node in graph A into the matching node in graph B (i.e., m(node in A) = matching node in B). The distance metric based on entropy for graph A and B is:

$D = \sqrt{\sum_i (a_i - b_{m(i)})^2}$

An analogous distance measure can be easily defined by considering the mutual information instead of entropy. The result of our experiments is shown in Table 1, where Rows 1-3 compare signatures obtained by random equal-size partitions of the instances of the concepts Writer, Artist and Automobile (actually, the result shown is the mean of the distance measures obtained evaluating 10 random partition). Rows 4-5 show the distances between the signature of concept Writer and its superset Artist (with correct and random matches between the properties). Rows 6-7 show the distances between the previous concepts (Writer and Artist) and the concept Automobile. The columns of the Table represent the types of distances between the graphs computed: we considered nodes and edges with the standard and the estimated measures for entropy and mutual information.

#	Comparison	Distance (H - Nodes)	Distance (MI - Edges)	Distance (Est. H - Nodes)	Distance [log] (Est. MI - Edges)
1	Artist - Artist	0.016	0.327	1.178	6.010
2	Writer - Writer	0.008	0.348	0.292	5.777
3	Automobile - Automobile	0.036	0.845	0.452	4.461
4	Artist - Writer (best matches)	0.205	16.038	26.550	8.879
5	Artist - Writer (random matches)	2.025	20.811	750.570	11.344
6	Artist - Automobile (best matches)	1.825	58.664	903.417	11.169
7	Writer - Automobile (best matches)	1.727	54.719	197.669	11.406

Table 1. Evaluation of the TDGs.

A qualitative evaluation of the preliminary results shows that all the techniques can detect signatures representing similar or different concepts. Our estimated values produce more polarized values, thus making the understanding of diverse classes easier. Moreover, as in [9], our experiments show that the entropy alone provides a good account of the similarities between the classes. Nevertheless, since we considered only few properties, we found some results not strictly consistent with the data (e.g., the value of the distance between Artist and Writer based on mutual information is higher than the one we were expecting, since the classes represent similar world concepts). Finally, the evaluation would definitely provide better results by considering a distance relying on all the weights (nodes and edges).

4.1 Preliminary discussion

The evaluation shown in the previous section permits us to draw some preliminary conclusions, which will constitute the basis of our future work.

1. The technique proposed and, in particular, the signature based on approximate weights is promising: it can effectively support the process of identifying the topics of a data source.
2. Signatures can also be experimented coupled with other techniques for detecting similarities between graphs. In particular, we think to obtain better results with matching approaches based on the source schema, like for instance names of classes/properties comparisons. In this way, techniques, relying on different kinds of information, can complement each other.
3. The definition of an effective and efficient algorithm for comparing signatures is a critical task. The graph matching algorithm should be able to work with: (1) graphs of different sizes, making possible to match graphs and subset of graphs; (2) many-to.one, one-to-many and many-to-many concepts mappings. Our signature extraction method is applicable for both (1) and (2). The example showed only signatures representing a single concept in both the reference ontology and in the target data sources. This because for now the matching algorithm (based on the Euclidean distance) only works for one-to-one concept match; but it is only a metter of matching algorithm and goes beyond the scope of this paper, i.e. proving that a signature based method can be exploited to discover topics of a data source.

5 Related work

To provide users with techniques and tools for automatically understanding the topics of a data source is a hot and challenging issue. The problem is well known in the IR Community, where it is applied to unstructured documents with important outcomes [3]. In the context of structured data sources, the proposed techniques face the issue following three main perspectives: providing summaries, exploiting reference ontologies and supporting users with visual tools.

Summary-based approaches aim to identify and extract a small subset of the information which is representative of the entire contents of the data source. In [14] and [15] two approaches dealing with relational databases and graphs, respectively, have been proposed. Both the approaches compute the closeness between data structures and the importance of the data taking into account entropy and mutual information. In [2], the goal is to summarize an attribute domain. A mix of techniques is applied for clustering the attribute values and identifying in each cluster a single representative value. Ontology-based approaches try to match content and data structures into some reference ontology. Summarized attributes can support the keyword search task as depicted in [1]. The research community in the Semantic Web is studying for fifteen years this process and several algorithms based on heuristic, syntactic and semantic rules have been proposed [7]. Finally, in the data science field, several code libraries and tools have been proposed for extracting visual summaries from the content of a data source (see for example Tableau[7] or Gephi[8]).

Our approach mixes some features from both the summary and the ontology-based approaches. The idea of creating a datasource signature starts from [9] where a depen-

[7] http://www.tableausoftware.com/
[8] http://www.gephi.org/

dency graph is built for supporting schema matching in a data integration approach. In this paper we adapted the approach for RDF sources and we extended the technique with the introduction of different kinds of edges connecting nodes.

Other approaches have applied entropy and mutual information to RDFS graphs (see for example [8]). Nevertheless, in this paper we adopted a novel technique for estimating the mutual information based on composite likelihood.

Finally, it is important to observe that Sindice.com [11], an RDF search engine, could be considered as a possible solution of the problem on hand. Nevertheless, Sindice focuses on finding triples containing particular keywords and not discovering data sources topics.

6 Conclusion and future work

In this paper we presented our preliminary proposal for providing users with an insight of a target data source topic. The approach relies on a reference ontology, a technique for generating signatures and an algorithm for graph matching. The preliminary results show that our proposal can really support the user in this task.

Future work will be devoted to three main tasks. Firstly, we will develop and implement a graph matching algorithm able to effectively match signatures from different data sources. Secondly we will perform a deep evaluation of the proposed approach in different domains and with data sources having different features in terms of numbers of attributes and instances. Thirdly, we will improve the technique for estimating entropy and mutual information to weighting the graph. In particular, we will experiment other statistical measures for evaluating the correlation of the values in order to obtain more effective signatures.

References

1. Sonia Bergamaschi, Elton Domnori, Francesco Guerra, Mirko Orsini, Raquel Trillo-Lado, and Yannis Velegrakis. Keymantic: Semantic keyword-based searching in data integration systems. *PVLDB*, 3(2):1637–1640, 2010.
2. Sonia Bergamaschi, Claudio Sartori, Francesco Guerra, and Mirko Orsini. Extracting relevant attribute values for improved search. *IEEE Internet Computing*, 11(5):26–35, 2007.
3. David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.
4. Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.
5. D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.
6. Vasant Dhar. Data science and prediction. *Commun. ACM*, 56(12):64–73, 2013.
7. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.
8. Lushan Han, Tim Finin, and Anupam Joshi. Schema-free structured querying of dbpedia data. In Xue wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *CIKM*, pages 2090–2093. ACM, 2012.
9. Jaewoo Kang and Jeffrey F. Naughton. On schema matching with opaque column names and data values. In Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, editors, *SIGMOD Conference*, pages 205–216. ACM, 2003.

10. Jayant Madhavan, Loredana Afanasiev, Lyublena Antova, and Alon Y. Halevy. Harnessing the deep web: Present and future. In *CIDR*. www.cidrdb.org, 2009.
11. Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *IJMSO*, 3(1):37–52, 2008.
12. Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.
13. Alex Wright. Searching the deep web. *Communications of ACM*, 51(10):14–15, 2008.
14. Xiaoyan Yang, Cecilia M. Procopiuc, and Divesh Srivastava. Summarizing relational databases. *PVLDB*, 2(1):634–645, 2009.
15. Xiaoyan Yang, Cecilia M. Procopiuc, and Divesh Srivastava. Summary graphs for relational database schemas. *PVLDB*, 4(11):899–910, 2011.
16. Gideon Zenz, Xuan Zhou, Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. From keywords to semantic queries-incremental query construction on the semantic web. *Journal of Web Semantics*, 7(3):166–176, 2009.