Finding Synonymous Attributes in Evolving Wikipedia Infoboxes

Paolo Sottovia[1], Matteo Paganelli[2], Francesco Guerra[2], and Yannis Velegrakis[3]

[1] University of Trento
paolo.sottovia@unitn.it
[2] Università di Modena e Reggio Emilia
firstname.lastname@unimore.it
[3] Utrecht University
i.velegrakis@uu.nl

Abstract. Wikipedia Infoboxes are semi-structured data structures organized in an attribute-value fashion. Policies establish for each type of entity represented in Wikipedia the attribute names that the Infobox should contain in the form of a template. However, these requirements change over time and often users choose not to strictly obey them. As a result, it is hard to treat in an integrated way the history of the Wikipedia pages, making it difficult to analyze the temporal evolution of Wikipedia entities through their Infobox and impossible to perform direct comparison of entities of the same type. To address this challenge, we propose an approach to deal with the misalignment of the attribute names and identify clusters of synonymous Infobox attributes. Elements in the same cluster are considered as a temporal evolution of the same attribute. To identify the clusters we use two different distance metrics. The first is the co-occurrence degree that is treated as a negative distance, and the second is the co-occurrence of similar values in the attributes that are treated as a positive evidence of synonymy. We formalize the problem as a correlation clustering problem over a weighted graph constructed with attributes as nodes and positive and negative evidence as edges. We solve it with a linear programming model that shows a good approximation. Our experiments over a collection of Infoboxes of the last 13 years shows the potential of our approach.

1 Introduction

Wikipedia, with its more than 5.8 million entries[4] is one of the largest human curated sources of knowledge. A Wikipedia entry provides information about some real world entity, of a specific type, like an event, a person, an organization, a product, etc. It consists of two parts: the unstructured part, which is free text, and the structured part, that is known as the *infobox* and is a set of attribute-value pairs. These pairs describe the main characteristics of the entity that the

[4] `https://en.wikipedia.org/wiki/Wikipedia:Statistics` updated on 24 March 2019

entry describes. The importance of the Infoboxes is significant. They may contain information that is also found in the text of the of Wikipedia entry, yet, they are highly more structured. This means that the semantics of the information they contain is much easier to interpret, queried, analyzed, combined and explored in general. The attributes (i.e., the names of the attributes) to be present in an Infobox of an entity depend on the type of the entity and are dictated by the Wikipedia policies.

The Wikipedia entries are highly dynamic data. Real world entities evolve over time, and so does the knowledge that we have about them. This real world evolution is reflected into the Wikipedia entries. Users are continuously updating the Wikipedia entries in order to always contain in the best possible way the knowledge we have about an entity. This means that by studying the evolution of the Wikipedia pages, it is possible to understand the evolution of the entities through time. To do so, a fundamental task is to be able to identify and link, across different versions in time of the same Wikipedia page, the parts that model the same kind of information. This is typically done using the schema information, i.e., the attribute names.

Unfortunately, the evolution of the Wikipedia entries is not only on the content but also on the attribute names, making the required linking a challenging task. Attribute names are often changed to more accurately or completely represent the semantics of the attribute values in the Infobox. As a result, different attribute names in different versions of the same Wikipedia entry may be used to represent the same semantic information, and the same attribute name in two different versions may be used to model semantically different pieces of information.

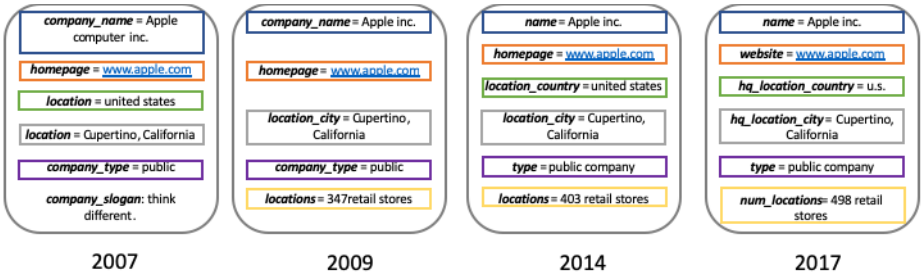


Fig. 1: Evolution of schema and values of the entity Apple Inc.

As an illustration of that situation, consider Figure 1 that contains four Infoboxes from different versions in time of the entity "Apple Inc.". Note the attribute that describes the location of the company. Initially (in 2007), the attribute name *location* was used to specify the country and other geographical data. In 2009 it disappeared. In 2014 two new attributes were introduced to indicate the location: the *location_country* and *location_city*. Finally, in 2017, the

attributes were renamed to *hq_location_country* and *hq_location_city* to more accurately specify that the location is the location of the headquarters.

The aforementioned example is also indicating another situation. The fact the same can happen to the attribute values. For instance, it can be noticed that the value indicating the country USA was originally "united states" and later changed to "u.s.".

In this work we deal with the problem of attribute name alignment in Wikipedia pages across time. We want to analyse and identify sets of attribute names that, across the evolution history of the pages of a specific type in Wikipedia, have been used to represent the same semantic concept. At the same time, we want to identify cases in which the same name has been used in attributes that model semantically different information.

Attribute alignment is a well-known problem that has been studied extensively in the past, mainly in the case of schema matching and in ontology alignment. The straightforward approach would be to look for synonym words in attribute names. This is the approach that has been followed in the area of Natural Language Processing. The techniques that have been developed there can be classified in two main categories. The first are those that exploit dictionaries, i.e., being based on the semantics of the attribute names. They are however, limited. Their limitation lies mainly in the number of synonyms that can be encoded in a knowledge base. Furthermore, they are context-independence, i.e., are not able to differentiate synonyms according to the context in which they are used. Another approach that has been studied in the context of identifying correspondences in Web form schema matching [12] is one that exploits correlations. The idea behind these techniques is that the search for synonyms is implemented by discovering attributes that correlate negatively or that do not co-occur. Unfortunately, this concept cannot be directly applied in the context of the Wikipedia Infobox attributes. This, because the Wikipedia content is so rich that using only co-occurrence information results in high false positives and also false negatives.

In an effort to overcome the limitations of the co-occurrence approach we have developed an extension of it that exploits the attribute values. In particular, the occurrence of same values between two different attributes in Infoboxes of different versions of the same type is treated as a positive indication that the two attribute names model the same real world concept. Furthermore, we treat co-occurrences as a negative indication. In particular, high degree of correlation (co-occurrence) in Infoboxes between two different attributes is an indication that these two attribute are referring to different concepts, thus, the high correlation is a negative indication. We turn the above two indicators into two different metrics, and create a network of attribute names where the values of the different metrics are used as a distance function. Then we apply clustering [10] to identify those sets that form mutually close names. Such sets are those we consider as semantically related.

Our contributions are specifically as follows: (i) We provide a novel approach to the problem of attribute name alignment in Wikipedia Infoboxes that exploits

co-occurrence information as a negative evidence and common attribute values as a positive evidence; (ii) We turn the evidences into metrics and treat the problem as a clustering problem, providing an efficient implementation of it that is based on lineal programming that provides a good approximation; (iii) We apply the technique on the set of Wikipedia entries of 13 years and report our findings on how effective such approach is indeed.

The remainder of the paper is structured as follows. Section 2 defines formally the problem with which we deal and Section 3 introduces our approach. Section 4 provides an extensive evaluation of our technique and reports our findings. Related works are presented in Section 5 alongside details on how our approach differs from these works.

2 Problem statement

The paper deals with entities described in Wikipedia articles. We assume that an article describes only an entity, identified by an identifier *id* (e.g., the page title). The type T specifies the subject of the entity, e.g, an event, person, organization, product, etc. Wikipedia articles consist of two components: an unstructured textual component and a list of attribute-value pairs called *infobox*.

Definition 1 (Infobox and infobox schema). *We define the entity infobox* $I = \{\langle a_1, v_1\rangle, ..., \langle a_n, v_n\rangle\}$*, where* $\langle a_i, v_i\rangle$ *are attribute-value pairs. We denote with* S_I *the infobox schema, that is the set of attributes included within it, and with* V_I *its values.*

For each type of entity T*, Wikipedia policies specify a template for the infobox schema, i.e. the list of attributes that should describe that type of entity.*

The data shown in an infobox may change over time. This happens mainly for two reasons: 1) the referred entity changes, i.e, the infobox values change and/or 2) Wikipedia releases new policies defining the infobox schema associated to a type of entity.

We define I_t the infobox at time t and E_t the entity at time t it is describing.

Definition 2 (Entity). *An entity at time* t*, denoted as* E_t*, is a triple* $\langle id, T, I_t\rangle$ *where id is the entity identifier,* T *is the entity type and* I_t *the associated infobox.*

The set of all the changes occurred to the entity can be collected from all the infoboxes and constitutes the *entity evolution*.

Definition 3 (Entity evolution). *Assuming the existence of a set of times values* $\mathcal{T}$ *that correspond to all possible times instances* t_i*, we define entity evolution* E *as the triple* $\langle id, T, I_\mathcal{T}\rangle$ *where id is the entity identifier,* T *is the entity type and* $I_\mathcal{T}$ *the set of all infoboxes* I_{t_i} *describing the entity over time. We identify with* $S_{I_\mathcal{T}}$ *the schema of* $I_\mathcal{T}$*, that is the union of the schemas of all the infoboxes contained within it. Similarly, we define* $V_{I_\mathcal{T}}$ *as the set of values in all the included infoboxes.*

The problem we want to address is to find, for each entity type, lists of synonymous attributes, i.e., attributes that are used over the time to describe the same property of an entity. The set of attributes used in the infoboxes of a specific entity type is called *set of entity type attributes*.

Definition 4 (Set of entity type attributes). *For each entity type T, A^T is the set of entity type attributes and includes all attributes used in at least an infobox schema at any time for describing an entity of type T.*

Synonymous attributes are clusters of entity type attributes that describe the same real-world entity property.

Problem 1 (Finding synonymous attributes)**.** *Given a set of entity type attributes A^T, we want to find a disjoint partitioning of A^T, denoted as $S = \{S_1, ..., S_m\}$, where the attributes $a_j \in S_i$ are used to describe the same real-world entity property.*

Furthermore, in the following, we denote:

- $V_I(a)$ as the set of values assumed over time by the attribute a within all the infoboxes $I_\mathcal{T}$ associated with an entity;
- $\Delta t_I(a)$ as the time interval (i.e., a list, even if not contiguous, of time instants) in which attribute a is *valid*, i.e. it appears in some infoboxes $I_\mathcal{T}$ associated with an entity;
- $\mathcal{I}$ as the set of the infoboxes $I_\mathcal{T}$ collected over time for a collection of entities.

3 The approach

In this section, we present our proposal for finding synonymous attributes in Wikipedia entities having the same type. For each pair of attributes, two measures are computed, assessing the extent in which the attribute represent (and do not represent) the same entity property, respectively. In this way, they provide a positive and a negative evidence of the synonymy. The measures are presented in Section 3.1. Then Section 3.2 shows how to use the knowledge provided by these measures to generate clusters of synonymous attributes. For this purpose, we reduce our problem to the one addressed by correlation clustering [4], where data points are partitioned into groups based on their similarity. A linear-programming approach has been adapted for this purpose. The work has been inspired from [12], where a similar technique has been adopted in the context of web search engines.

3.1 Positive and negative evidence for synonymy

We can model the synonymy relationship between attributes by analyzing their co-occurrences in the same infobox. In this perspective, we assume that synonymous attributes cannot appear simultaneously in the same infobox otherwise there would be information redundancy. In other words, we leverage the co-occurrence of two attributes in the same infobox as a negative evidence for their synonymy.

Example 1. Consider for example the attributes *name* and *type* which definitely describe different aspects of an entity. They are very common attributes: in a random sample of 60,760 infoboxes describing companies collected over the last 13 years, they appeared, together or separately, in 74.61% of the cases (i.e. for describing 45,332 entities). Within these entities the attributes coexist in the same infobox in 99.89% of the cases (i.e., 45,281 times), and they do not co-occur 51 times. According to our idea, they cannot be considered as synonyms. Conversely, the *name* and *company_name* attributes, that instead can be used to describe the same characteristic of an entity, show an inverse co-occurrence pattern: in 0.08% of the cases are present simultaneously in the same infobox and in 99.92% of the cases do not co-exist.

More formally, given two attributes a_i, a_j belonging to infoboxes describing the same entity type, Equation 1 provides a measure of their "negative" co-occurrence.

$$NegCoocc(a_i, a_j) = \frac{|\{I_{\mathcal{T}} \in \mathcal{I} | \Delta t_I(a_i) \cap \Delta t_I(a_j) \neq \emptyset\}|}{|\{I_{\mathcal{T}} \in \mathcal{I} | a_i, a_j \in S_{I_{\mathcal{T}}}\}|} \quad (1)$$

To compute this measure, all infoboxes in the collection are evaluated. For each of them, the presence of both input attributes (i.e., a_i and a_j) in a time frame is verified. This check is carried out by identifying whether there are overlaps in their validity time interval (i.e., $\Delta t(a_i)$ and $\Delta t(a_j)$ respectively). The number of entities for which there is overlap, normalized by the overall number of entities in $\mathcal{I}$, provides the correlation value that we consider as negative evidence for their synonymy.

As the experimental evaluation shows, the adoption of this measure only is not enough to accurately identify the synonymous attributes.

Example 2. The highest values obtained by the application of Equation 1 to the attribute *company_logo*, are with the attributes *logo*, *name* and *type*. The results, in our collection are respectively 0.9988, 0.965 and 0.94. Obviously, only the first pair of attributes are synonyms. The other pairs are attributes representing very different information. After a careful analysis of the temporal evolution of the infobox schemas we noticed that the attributes with the "company" prefix have been introduced with an old Wikipedia policy to identify all attributes describing "company" type entities. Today, this policy is no longer adopted, in favor of more concise and direct attributes (such as *type* and *name* instead of *company_type* and *company_name* respectively). However, a delay in the application of the new policy produces misalignments in the infoboxes and make Equation 1 not enough accurate.

To produce more accurate results, we introduce a measure for positive synonymy evidence. In particular, we measure the values shared between attributes as the indication that they are really synonyms. We analyze the different value representations of the attributes throughout the entire history of Wikipedia and we calculate their fraction of overlap through the Jaccard similarity. We do not

deliberately consider other string similarity techniques to have a more general approach, which does not rely on specific domain knowledge. In more detail, for each pair of attributes we select the values that generate the maximum fraction of overlap within the data collection. Equation 2 provides the formulation of the measure we adopt.

$$PosOverlap(a_i, a_j) = \frac{\sum_{I_\mathcal{T} \in \mathcal{I} | a_i, a_j \in S_{I_\mathcal{T}}} max\Big(\sum_{v_i \in V_I(a_i), v_j \in V_I(a_j)} Jaccard(v_i, v_j) \Big)}{|\{I_\mathcal{T} \in \mathcal{I} | a_i, a_j \in S_{I_\mathcal{T}} w\}|} \tag{2}$$

where, with reference to the notation introduced in Section 2, $V_I(a_i)$ and $V_I(a_j)$ represent respectively all values assumed over time by the attributes a_i and a_j for all infoboxes in $I_\mathcal{T}$ where they are valid.

Example 3. Let us consider the application of Equation 2 with the same input as in Example 2. We obtain the following results: $PosOverlap(company\_logo, logo) = 0.8899$; $(company\_logo, name) = 0.003$; and $(company\_logo, type) = 0.007$. We can observe that the high value computed for the pair (*company_logo*, *logo*) confirms the previous evidence of synonymy. The very low values for the other pairs do not confirm the evidence of synonymy resulting from Equation 1.

3.2 Holistic approach for synonym discovery

The measures of synonymy between pairs of attributes are used to compute clusters of synonymous attributes which constitute the result of our work. Our idea is to model the synonymy relations between the attributes by means of a graph and to apply a clustering algorithm over the graph to extract groups of synonymous attributes.

Given some positive and negative evidence for attributes synonymy, we model attributes and their synonymy relationship as an *attribute-synonymy graph*, where the nodes correspond to the attributes and the edges to the synonymy relations between the attributes. The edges are labeled according to whether the measure associated with them should be interpreted as positive or negative evidence of synonymy.

Definition 5 (*attribute-synonymy graph*). *An* attribute-synonymy graph *is a graph $G = (V, E)$ with vertices representing the attributes of the infoboxes we want to analyze. The edges associate to each pair of vertices provide a measure of their synonymy through a weight $w_{i,j} \geq 0$. Let $L_{i,j}$ be the label associated to each edge (i, j). L can assume the value $+$ or $-$ according to whether the edge is representing the measure of the negative or the positive evidence for their synonymy expressed by Equation 1 and Equation 2, respectively. Let E^+ be the set of edges identified by a label of value $+$: $E^+ = \{(i, j) | L_{i,j} = +\}$, and, analogously, E^- (i.e., $E^- = \{(i, j) | L_{i,j} = -\}$) the set of edges identified by a label of value $-$. A representation of this graph is provided in Figure 2, where solid edges indicate edges with positive weights and edges with crosses the negative ones.*

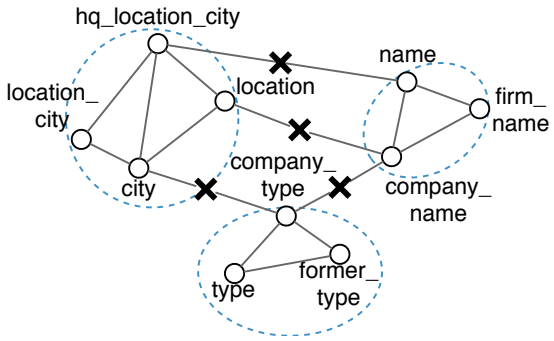


Fig. 2: Attribute-synonym graph for "company" type entities

Our goal now is to apply a clustering strategy that partitions the nodes of the *attribute-synonymy graph* so that each attribute is associated with a single cluster with its synonyms (see the dashed blue circles of Figure 2). To obtain this result, we adopt a correlation clustering algorithm [4] which provides a method for clustering data points into the optimum number of clusters based on their similarity without specifying that number in advance. In our implementation, the aim is to identify the partitioning of the infobox attributes that best respects the positive and negative evidence of synonymy provided as input.

Problem 2 (discovery of synonymous infobox attributes)**.** *Given an* attribute-synonymy graph $G = (V, E)$*, we want to find a disjoint partitioning of* V*, denoted as* $\mathcal{S} = \{S_1, ..., S_m\}$*, that agrees as much as possible with the labels* L *associated to the edges* E *of the* attribute-synonymy graph*. More precisely, we want a clustering that maximizes the weight of agreements: the weight of* $+$ *edges within clusters plus the weight of* $-$ *edges between clusters.*

The resolution of this problem exploits a heuristic procedure already proposed in the literature [10] for solving the correlation clustering problem. This technique is divided into two steps. First a linear programming approach is used to provide an approximate solution to the problem. The results produced by this model are fractional values that correspond to scores of synonymy between attributes. In a second step a technique called region-growing is applied to group attributes with a high synonymy level within the same cluster and remove the attributes that describe different information about the referred entity.

Linear-programming approach In the first phase of the approach the following linear model has to be solved.

$$\begin{aligned} &\text{minimize} \quad \sum_{(i.j)\in E^-} w_{i,j}(1 - x_{i,j}) + \sum_{(i.j)\in E^+} w_{i,j}x_{i,j} \\ &\text{subject to} \quad x_{i,j} \in [0,1], \qquad x_{i,j} + x_{j,k} \geq x_{i,k}, \qquad x_{i,j} = x_{j,i}. \end{aligned}$$

The goal of this model is to identify a valid assignment of the variable $x_{i,j}$ that minimizes the sum of the negative edges included in a cluster and maximizes the sum of positive edges. Intuitively, this variable provides an indication of the

collocation of the nodes in the clusters (i.e., it assumes, in the borderline cases, the value 0 when two attributes are included in the same cluster and 1 in the opposite case). An assignment of $x_{i,j}$ is considered valid if $x_{i,j} \in [0, 1]$ and $x_{i,j}$ satisfies the triangular inequality. This motivates the inclusion of the constraints in the problem formulation. The adaptation of this linear model to our problem requires the addition of a further constraint, which requires that the negative weights (i.e., $w_{i,j}$ in the first sum) are defined according to Equation 1, and the positive ones (i.e., $w_{i,j}$ in the second sum) according to Equation 2.

Region growing Once a first approximated solution to the problem is obtained, we apply the region growing technique. Its objective is to convert the approximate cluster membership indication of the attributes provided by this first solution, into an exact distribution of the attributes in the different clusters. More precisely, this technique is used to convert the fractional solution x in an integral solution which identifies if two attributes belong to the same cluster. Since this technique represents a classical clustering strategy, below we provide only an insight into its operation. More details instead can be found in [10]. The intuition behind this technique is to construct, in an iterative way and starting from randomly selected seed nodes, some balls (i.e., groups of graph nodes) modifying, step by step, their coverage radius on the graph. The growth of these balls is determined by the weights associated with the graph edges: a ball will continue to grow as the sum of the positive weights included inside the subgraph identified by the ball is advantageous. On the contrary, the ball will stop growing, causing the creation of a new ball (or a new cluster), when its growth would incorporate dissimilar nodes compared to those already included in the cluster. The arrangement of these balls within the graph determines its final partitioning.

4 Experimental evaluation

In this section, we firstly provide a description of the dataset used for the experimental evaluation (Section 4.1) and then we qualitatively (Section 4.2) and quantitatively (Section 4.3) evaluate the effectiveness of the approach. Finally, a case study is presented (Section 4.4) to show how Wikipedia synonymous attributes can be used in a real scenario.

4.1 Dataset description

The dataset used in the experimental evaluation is a collection of infoboxes of entities having type associated to the "concept of company" (i.e., we consider entities having type company, organization, dot-com company, etc.). This collection includes, for each entity, its complete history between August 2004 to August 2017, i.e. all updates in the infobox schemas that have been introduced by Wikipedia users. The result is $60,760$ entities and around $1,861,252$ changes.

The number of attributes used in the infoboxes varies: it is not fixed per entity type and in the time. Table 1 provides some statistics about attributes

	avg	**std**	**max**	**min**
# attribute per entity	12.84	5.64	253	1
# value per entity	25.35	23.67	503	1

Table 1: Number of attributes and values per entity

and values. The average number of attributes and values per entity are 12.84 and 25.35 respectively. Moreover, the maximum number of attributes associated to an entity, in the considered period of time, is equal to 253, and the maximum number of different values is 503.

In Tables 2a and 2b the top 10 most frequent attributes and values are reported. The attribute "name" is the most used: it appears in 97% of the collected entities, while "company_name", appearing in 47% of the entities, is the 10th most used attribute. Concerning the values, "united states", "privately held companies" and "public companies" appear respectively in 31.78%, 31.63% and 23.47% of the entities and are the most frequently values used in the collection.

attr	**freq**	**freq (%)**
name	59017	97.13%
industry	51845	85.33%
foundation	49033	80.70%
homepage	47076	77.48%
type	46015	75.73%
logo	40102	66.00%
key_people	36176	59.54%
products	33388	54.95%
location	32490	53.47%
company_name	28565	47.01%

(a) Top 10 most frequent attributes

value	**freq**	**freq (%)**
united states	19310	31.78%
privately held company	19221	31.63%
public company	14259	23.47%
united states dollar	9692	15.95%
private	7983	13.14%
subsidiary	7144	11.76%
united kingdom	5986	9.85%
worldwide	5973	9.83%
yes	5289	8.70%
chief executive officer	4793	7.89%

(b) Top 10 most frequent values

Table 2: Frequencies of attributes and values

Tables 3 provides an insight on the evolution of the attributes and values in the considered period of time. In particular, Table 3a shows the attributes whose values were most frequently subject to change, and Table 3b the top 10 entities affected by the greatest number of changes over time. Note that 10% of all the infobox updates involves the "key_people" attribute, and the most modified entity is "Eurosport".

A more detailed analysis of the evolution is shown in Figure 3, where the top 5 most updated types of entity are analyzed. Table 3a shows the number of entities collected per type and the total number of changes. Figure 3b plots some statistics about the number of updates per entity. Although the total number of updates in the "company" category is the highest, a particularly high number of average updates has been applied to entities belonging to the "television" type.

attr	freq	freq (%)
key_people	162465	10.07%
products	112440	6.97%
location	94936	5.89%
foundation	87828	5.45%
industry	86707	5.38%
homepage	84182	5.22%
name	77894	4.83%
logo	62653	3.88%
type	61655	3.82%
revenue	59206	3.67%

(a) Top 10 most changed attributes

entity title	freq	freq (%)
Eurosport	1924	0.10%
National Geographic (TV channel)	708	0.04%
Canada	672	0.04%
Apple Inc.	594	0.03%
Nintendo	580	0.03%
HBO	538	0.03%
Cuba	527	0.03%
Animax Asia	526	0.03%
General Motors	525	0.03%
Amazon.com	509	0.03%

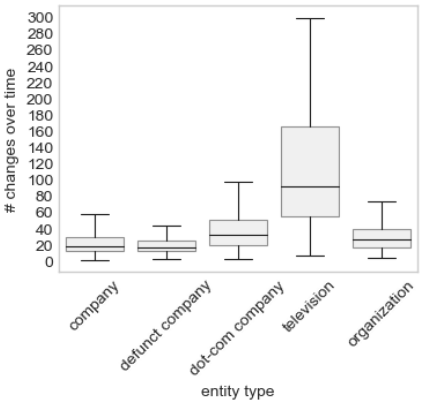
(b) Top 10 most changed entities

Table 3: Updates in Wikipedia entries

The other categories of entities, on the other hand, present an average number of updates which is approximately the same (i.e., the range varies between 20 and 40 updates).

entity type	# entities	# total changes
company	57,553	1,494,245
defunct company	664	15,304
dot-com company	127	6,340
television	40	4,953
organization	155	4,819

(a) Entities and updates per entity type



(b) Statistics on the updates

Fig. 3: Top 5 most updated entity types

4.2 Qualitative evaluation of the effectiveness

In this section we qualitatively evaluate the effectiveness of our approach by analyzing a sample of its results. Table 4 shows 10 clusters of synonymous attributes generated by our approach. We can observe that our approach is able to identify

interesting and non-trivial synonymy relations between attributes. For example, it is able to find the correspondences between attributes like "established" and "founded" or "predecessor" and "former_name" which would not be identifiable by a string similarity technique. Furthermore, we match attributes expressed in different languages, such as "employees" with "mitarbeiterzahl" and "city" with "sitz". Analyzing these results, we can observe the various textual forms used over time by the Wikipedia community to indicate the same characteristic of an entity. This variety of forms presumably derives from the adoption of different schema guidelines/policies imposed by Wikipedia[5]. The attributes "name", "company_name" and "type", "company_type" are examples of this situation. The inclusion of the prefix "company" has been introduced by a policy to make more explicit the type of entity described by the attributes.

cluster
num_staff, employees, number of employees, num_employees, numemployees, mitarbeiterzahl
established, opened, formation, founded_date, start_year, date_founded, foundation, gründungsdatum, introduced, founded
logo, non-profit_logo, network_logo, company_logo, firm_logo
web, url, website
operating_profit, ebitda, operating income, operating_income
creator, founder(s), founder, founders
predecessor, former_names, former_name, predecessors
company_type, type, unternehmensform, former type, former_type, company type, non-profit_type
headquarters, headquaters, hq_city, location_city, city, sitz, residence, hq_location_city, location, place, hq_location
agency_name, network_name, group_name, name, non-profit_name, company_name, firm_name, company name

Table 4: Example of synonymous attributes produced by our approach

4.3 Quantitative evaluation of the effectiveness

The effectiveness of the proposed approach is assessed in quantitative terms. The main goal of this analysis is to empirically demonstrate that both the measures contribute in identifying synonymous attributes. To perform this evaluation, firstly a ground truth has been manually created. We have exploited public attribute mappings directly provided by the Wikipedia Template pages to obtain a first minimal set of attribute matches. This basic information has been then extended with new manually inserted attribute correspondences. The generated ground truth includes about $2,000$ attributes clustered in 454 groups of synonyms. Once an exact set of attribute correspondences has been generated, we

[5] Note a cleaning procedure has been applied to the input infoboxes to remove the "noise" generated by human mistakes.

evaluated our approach on a sample of the entire collection of Wikipedia infoboxes. In more detail, we tested our approach on an *attribute-synonymy graph*, generated starting from the input dataset, consisting of $6,854$ attributes and $52,707$ synonymy relations. The results provided by our approach were finally compared with the ground truth.

To provide a measure of the quality of the clusters generated by our approach with respect the ground truth, we adopted four measures: precision, recall, f1 score and rand index. We calculate precision as $\frac{\#\ true\ synonyms\ in\ cluster}{\#\ total\ attributes\ in\ estimated\ cluster}$, and recall as $\frac{\#\ true\ synonyms\ in\ cluster}{\#\ total\ attributes\ in\ real\ cluster}$. The f1 score is a combination of precision and recall defined as $2 * \frac{precision*recall}{precision+recall}$. Finally, the rand index [14] was used to evaluate the similarity between the clustering solution produced by our approach and that provided by the ground truth.

The experiment aims, in particular, to evaluate the contribution of each measure in obtaining the final result. To force such behavior, a linear combination of the two measures has been introduced. Its formulation is proposed in Equation 3, where the α parameter is used to weight the contribution of the measures.

$$SynonymyScore(a_i, a_j) = \alpha * PosOverlap(a_i, a_j) + (1-\alpha) NegCoocc(a_i, a_j) \quad (3)$$

The results of this experimentation are given in Table 5. The results show that, with reference to the company entity type, linear combinations that assign more importance to the positive evidence of synonymy produce better results. Table 5 shows only α ranging from 0.6 to 1, however with lower α results follow a similar trend: precision decreases and instead recall increases. The best configuration is $\alpha = 0.8$, that obtains the highest values in all evaluation measures. We observe that the configuration with $\alpha = 1$, where there is no contribution from the negative evidence of synonymy, is the one that obtain the highest precision level. Nevertheless, in that configuration, the recall, rand and f1 score levels decrease considerably.

α parameter (positive contribution)	precision	recall	f1 score	rand index
0.6	0.375	0.764	0.372	0.146
0.7	0.817	0.759	0.757	0.886
0.8	0.797	0.767	0.760	0.947
0.9	0.791	0.754	0.752	0.942
1	0.831	0.668	0.723	0.858

Table 5: Effectiveness evaluation with different positive contributions

4.4 Case study

In this section we provide a small case study to show that synonymous attributes can support the extraction of high quality and accurate information from Wikipedia. Table 6 introduces 5 information needs a user would like to satisfy

against the collection of infoboxes described in Section 4.1. Each information need has been transformed into 2 structured queries: one with the original filtering condition formulated by the user and the second where the attributes have been substituted with a number of disjunctive clauses, each one expressing the same information need but by using synonymous attributes. Table 6 shows the number of entities retrieved when both the queries are executed and the number of results it is expected to be retrieved. Last two columns show the same information in percentage. We observe that synonymous attributes largely support the retrieval of all results. The maximum improvement is obtained with the last query (i.e., `type="public company", num_employees>10,000, year=2010`) where the application of synonymous attributes allows us to retrieve all results, instead of 19.15% of them, as we obtain with the original formulation.

query	# query results				
	original clauses	using synonymous attributes	ground truth	original clauses (%)	using synonymous attributes (%)
location_city="tokyo"	268	293	298	89.93	98.32
founded<1900	681	705	705	96.59	100
location="USA", year=2014	381	531	531	71.75	100
location="united states", net_income>1 billion	371	429	429	86.48	100
type="public company", num_employees>10,000, year=2010	221	1154	1154	19.15	100

Table 6: Case study

5 Related Work

Importance and usage of Wikipedia infoboxes. Wikipedia infoboxes have been used in a large number of research projects. The most significant works include techniques for building structured knowledge bases [18, 3], for analyzing the evolution of specific kinds of entities [13] and for applying structured queries on the Wikipedia content [2]. Although several works support the formulation of structured queries, no previous effort has considered the evolutionary nature of Wikipedia. All previous approaches consider only a static snapshot of the infoboxes as input.

Schema matching. Schema matching is one of the most studied topics in the database community. Books [5, 11] and surveys [16, 6] introduce the existing approaches in the literature. According to the categorization proposed by [16], schema matching approaches can be classified into schema-only matchers and instance-based matchers. Our proposal follows a hybrid approach since we in-

corporate holistic correspondence refinement that belongs to the category of collective matching approaches [6].

Our approach implements a strategy similar to the one proposed by He et al. [12] that consider as negative evidence the co-occurence of attribute names in the same schema. Their approach focus on the discovery of synonyms supporting a web search engine and is based on the combined use of web tables and query logs. The intuition is that users who are looking for the same results provide different synonyms as query terms on a search engine. This is used as positive evidence for attribute synonymy. Instead, attributes within the same web tables are not likely to be synonyms, thus providing a negative evidence. Our approach adapts that idea to work with Wikipedia infoboxes by extracting positive and negative evidence of synonymy analyzing co-occurrences of attributes and values.

Schema matching and alignment over infoboxes. Schema matching techniques have been applied against Wikipedia infoboxes in the context of finding correspondence between schemas in different languages. Adar et al. [1] propose a framework called Ziggurat that creates a supervised classifier based on features that are learned from a set of positive and negative examples extracted from data with heuristics. Bouma et al. [9] match attributes based on the equality of their values. Two values are equal if they have the same cross-language link or exactly the same literals. A different approach [17] exploits value similarity over infobox templates, where first an entity matching process is done, then templates are matched to obtain inter-language mappings between templates and finally attribute matching is done by means of similarity metrics. These approaches rely on similarity metrics that are sensitive to the syntax of the underline languages: they work well if the compared languages have the same root. To overcome this limitation [15] exploits different evidences for similarity and combines them in a systematic manner.

Exploration of schema and value changes. In the context of data exploration, a recent line of research focuses on the exploration of changes over time. [7] is vision paper that introduces innovative concepts related to understanding changes that happen in the data over time. Furthermore, [8] designs a set of primitives supporting the exploration over schema and data of evolving datasets.

6 Conclusion

In this paper, we introduced an approach that automatically defines clusters of synonymous temporal-evolving infobox attributes. The approach is mainly based on two kinds of knowledge: a negative evidence of synonymy provided by co-occurrences of the attributes in the same infobox in a given time instance, and a positive evidence of synonymy generated by co-occurrences of similar values for the attributes in different time instances. We formalized this issue as a correlation clustering problem over a weighted graph and we used a linear programming model to solve it. Our experiments, over the last 13 years infoboxes history, shows that our approach is effective in discovering synonymous attributes.

References

[1] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual wikipedia. In *Proceedings of WSDM*, pages 94–103, 2009.

[2] P. Agarwal and J. Strötgen. Tiwiki: Searching wikipedia with temporal constraints. In *Proceedings of WWW*, pages 1595–1600, 2017.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC*, pages 722–735, 2007.

[4] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[5] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. *Schema Matching and Mapping*. Data-Centric Systems and Applications. Springer, 2011.

[6] P. A. Bernstein, J. Madhavan, and E. Rahm. Generic schema matching, ten years later. *PVLDB*, 4(11):695–701, 2011.

[7] T. Bleifuß, L. Bornemann, T. Johnson, D. V. Kalashnikov, F. Naumann, and D. Srivastava. Exploring change - A new dimension of data analytics. *PVLDB*, 12(2):85–98, 2018.

[8] T. Bleifuß, L. Bornemann, D. V. Kalashnikov, F. Naumann, and D. Srivastava. Dbchex: Interactive exploration of data and schema change. In *Proceedings of CIDR*, 2019.

[9] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual alignment and completion of wikipedia templates. In *Proceedings of the Workshop on Cross Lingual Information Access*, pages 21–29. Association for Computational Linguistics, 2009.

[10] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.*, 361(2-3):172–187, 2006.

[11] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, 2007.

[12] Y. He, K. Chakrabarti, T. Cheng, and T. Tylenda. Automatic discovery of attribute synonyms using query logs and table corpora. In *Proceedings of WWW*, pages 1429–1439, 2016.

[13] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, 194:28–61, 2013.

[14] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 12 1971.

[15] T. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire. Multilingual schema matching for wikipedia infoboxes. *PVLDB*, 5(2):133–144, 2011.

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

[17] D. Rinser, D. Lange, and F. Naumann. Cross-lingual entity matching and infobox alignment in wikipedia. *Inf. Syst.*, 38(6):887–907, 2013.

[18] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A large ontology from wikipedia and wordnet. *J. Web Semant.*, 6(3):203–217, 2008.