Data Quality Improvement and Entity Alignment Optimization for Constructing Large-Scale Knowledge Graphs

Keaton Sullivan[a,b], Fiona Browne[a], Huiru Zheng[b] and Haiying Wang[b]

[a]Datactics, 1 Lanyon Quay, Belfast, Northern Ireland, UK
[b]School of Computing, University of Ulster, Newtownabbey, Northern Ireland, UK

Abstract

Poor data quality can have an impact on the accuracy and analysis of knowledge graphs. Remedying this involves maximizing the data quality of sources used in constructing knowledge graphs and aligning entities. By improving the underlying data quality, knowledge graphs and their analysis are subsequently improved. In this paper we propose and implement a parallelizable data quality driven pipeline. We compare the proposed approach against one utilizing common pre-processing actions. This involves the measurement of entities validated against an external comprehensive dataset. A higher percentage reduces the need for complex algorithms that scale with a polynomial degree. We then show how the validated entities resulting from the pipeline produces high quality nodes and relationships that can be modelled as a realistic knowledge graph.

Keywords

Data quality, Parallelisation, Entity alignment, Textual similarity, Open data

1. Introduction

A global trend of increased publicly accessible datasets has been observed in [1] and attributed to commitments to governmental transparency initiatives such as the 2011 Open Government Partnership and the 2013 G8 Open Data Charter [2]. In addition to the volume, the diversity of data has grown exponentially [3]. The decentralised nature of datasets requires heterogeneous solutions to integrating and representing how different sources of data relate to each other. Knowledge graphs have become an effective solution of modeling these relationships as they provide a more realistic representation of integrated heterogeneous datasets by standardising data to an ontology that is composed of entities and how they relate to each other [4]. The versatility in modeling data from the heterogeneous data sources has found knowledge graphs being used anywhere data sources need to be integrated to support decision making processes - from the original use in creating a semantic web of the internet [5] to developing publicly accessable knowledge graphs of governmental data product offerings [6].

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland
keaton.sullivan@datactics.com (K. Sullivan); sullivan-k@ulster.ac.uk (K. Sullivan); fiona.browne@datactics.com (F. Browne); h.zheng@ulster.ac.uk (H. Zheng); hy.wang@ulster.ac.uk (H. Wang)

The construction of knowledge graphs has been a well researched topic, and an architecture [4] has been established - however the performance of implementing these in real-world large datasets has been identified as a challenge in many studies [4] [7] [8]. Algorithms developed in academia provide very accurate results, however the time complexity of their operations are rarely considered [9]. As a consequence, their practical applications have been found to be limited due to their polynomial time complexity not scaling with realistic datasets, despite their high accuracy.

The ontology and the knowledge base are the two fundamental components to knowledge graph construction.

Datasets undergo a process called knowledge extraction to transform unstructured, semi-structured and external structured data into a knowledge base of structured information. This process transforms the data into entities, their relationships and attributes that describe them. However, being from heterogeneous sources, it contains multiple issues that prevent it from modeling the real scenario, such as: (1) attributes may be spilt between multiple extracted entities (2) there may be duplicated entities (3) there may be subsets of attributes that can be considered independent entities.

Because of this, the extracted knowledge base undergoes a process called knowledge fusion which ultimately constructs the ontology of the knowledge graph and populates it with the knowledge base that is a more realistic approximation of the scenario - this process is often iterative as improvements may only be identified after investigating the completed ontology.

The process of resolving these issues in knowledge fusion before ontology construction is called entity alignment - and the algorithms that carries this out are a focus of academic study. Before these algorithms can be applied, data pre-processing is required to standardise the diverse representations of data and the inconsistencies of how they are recorded. However, academic datasets typically select datasets with limited data quality issues and therefore requires very little data preprocessing actions to be carried out.

The time complexity issue is primarily the product of attempting to apply algorithms with a polynomial time complexity to all entity-relationship pairs - and as such is primarily a limitation in the entity alignment stage which involves those comparisons [9].

Motivated by these challenges facing the practical application of knowledge graphs and with suggestions for solutions provided by [9], [7] and [8], this paper describes how the introduction of a data quality improvement pipe line that extends the existing data pre-processing stage of knowledge fusion can significantly reduce the time complexity challenges that would typically limit the practical implementations of academic entity alignment algorithms. This paper makes the following contributions:

1. development of a knowledge graph with practical applications from Open Governmental Data (OGD) datasets with significant data quality issues that would limit academic algorithms
2. proposed data quality improvement pipeline for practical implementations of entity alignment algorithms
3. demonstrate the time complexity reduction of compared method to existing algorithms and a typical pre-processing approach

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 discusses considered datasets. Section 4 describes the approach to knowledge graph construction and includes the proposed data quality improvement pipeline. Section 5 discusses the results of implementing the proposed pipeline. Section 6 discusses future work. Section 7 concludes the paper.

2. Related Work

In this section we highlight key research in the areas of open data, data quality and the construction of knowledge graphs from these data.

2.1. Data Quality in Datasets

Data quality has been the subject of a number of studies which attempt to define data quality dimensions. However, there is no clear consensus on these dimensions with and across domains despite the research focus.

An attempt to standardize the heterogeneous nature of data quality was conducted in [10]. This research produced a model ontology of data quality characteristics, dimensions and domains providing first steps at standardizing data quality in literature. This research found that data quality assessment was too premature of a field for the ontology to be viable in every approach. It was clear that while objective measures could be made generic, there would always be subjective or business specific measures that required expert assistance in defining.

The approach to data quality evaluation has been investigated. This includes data quality rules executed against data with calculations performed to capture percentage of data that pass such rules. These are straightforward to calculate [11]. However [10] proposes that the fitness for purpose should define beneficial metrics then design rules as complex as required to address them. This complexity applies limits on the number of rules, but is regarded as more beneficial for the users of the data due to being specific. Rules are assigned to a relevant data quality dimension and individual rule results are aggregated to provide a broad overview in dashboards for business users.

Ultimately, data quality is defined as fitness for purpose. Generic objective measures can be carried out [12], however an expert is required for subjective and business specific measures. To provide structure, a framework similar to [6], based on the characteristics described in [10], is a comprehensive approach to measuring data quality.

Knowledge of purpose, descriptive metadata, familiarity of data, comprehensive profiling focusing on regular distribution and outlier values are some of the methods used to inform definitions of data quality rules. The fewer present when describing data quality rules, the less fit-for-use defined rules rules would be.

2.2. Knowledge Graphs and Entity Alignment

Construction of knowledge graphs to model entities and their relations is growing in popularity across diverse domains such as finance, medicine and social sciences. Key challenges in generating high quality robust knowledge graphs are in the areas of data quality defined above

and in knowledge fusion where datasets are integrated together. Zhao et al. [4] conducts a systematic review of knowledge graph construction and identified a typical architecture which is separated into knowledge extraction and knowledge fusion. Entity alignment algorithms are an important topic within knowledge graph construction as identified in [4] and [9] whereby entities from different datasets are matched and presented in a graph structure. We refer to this process as entity alignment. Single alignment algorithms for small-scale datasets have been making significant advancements in the field such as realising knowledge graph representation learning with neural network based models [13], [14]. Alignment in large-scale networks have been facing significant challenges due to the polynomial time complexity of alignment algorithms which have to overcome data quality issues such as consistency. As a consequence, the single algorithm approach for large scale datasets are often reduced to a combination of simple matching strategies [8]. These produce multiple similarity scores which are combined to judge which entities are duplicates. High confidence matched entities often go through further alignment to produce a similarly score vector which is time consuming and not as important as lower confidence matches.

Studies [8], [7] routinely call for the development of algorithms capable of parallelism to address the time complexity. For instance, using a multi threaded approach to increase speed. However, this approach is not feasible with single algorithms as there is little concurrent activity to partition. In addition, as all entities are compared together, there would be an immense amount of inter-partition communication - both of this issues need to be addressed to facilitate parallelism.

In addition, the same studies state that if the data has significant data quality issues, then no algorithm can be used. Surprisingly, this issue is discussed as critical to an algorithms success but the data pre-processing which aims to reduce it is often not described in detail or addresses the minimum consistency and integrity data quality issues.

To address such limitations, we propose a data quality improvement pipeline in this paper. This pipeline aims to reduce the volume of considered entity pairs so that matching using complex alignment algorithms are only performed only on the least confident entity pairs. Furthermore, we focus in detail on the data quality aspect of this pipeline to improve the matching downstream.

3. Datasets

In order to demonstrate the impact of a proposed data quality improvement pipeline on a knowledge base, open governmental datasets were selected as they fulfilled the following requirements: (1) available to the general public, (2) used for practical commercial decision making, (3) no existing knowledge graph is accessible, (4) has measurable data quality issues to improve and (5) heterogeneous entities and relationships between datasets.

The OGD from the UK was selected as the author could act as the expert in decision making due to familiarity with data and the UK has a high ranking on the global open data index. Common datasets available in OGD initiatives include budget allocation, national statistics, environmental data, weather forecasts, location and company registers. Company register was selected due to its common use in finical sectors. The data collection for this register is self

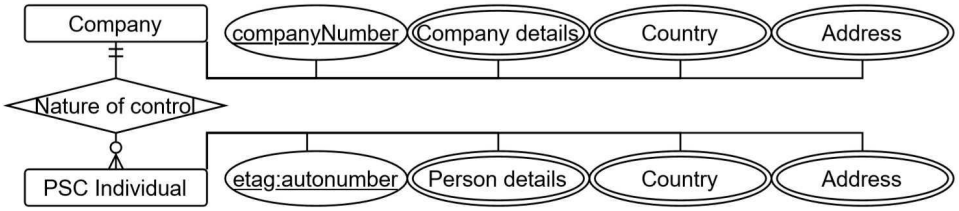

Figure 1: ERD of data sources with unrealistic 1:n cardinality

reported from users resulting in data quality issues. Multiple register datasets are available containing linking information. The specific OGD company register datasets selected in this study are the company data product and their persons of significant control (PSC) from March 2021.

The ontology represented by the data sources is summarised in Figure 1. Issues of why this is unrealistic is discussed in section 4.

Meta data for both companies data and PSC was minimal and only available from unstructured sources. The use of comprehensive profiling allowed the author to understand the content of the fields. A summary of these data along with data quality actions are detailed below along with additional corpora used in the data improvement pipeline for the entity alignment process.

3.1. Company Data Dataset

This set includes the basic name, registered address, operating status and classification of activities of all active UK companies. Updated monthly, available in CSV format and comprising of around 5 million records that have 60 fields. The majority of data quality issues related to companies house can be attributed to the lack of validation on addresses - particularly the lack of retrospectively correcting identified inaccuracies. We select all companies when building the knowledge graph - see Table 1 for summary of cleansing actions taken during pre-processing to ensure data quality was maximized before entity alignment.

3.2. Persons of Significant Control Dataset

This set includes information on the nature of control that individuals, companies and legal entities have over companies. It has user entered address and name information and is updated daily. The data is available in JSON format and comprises of more that 8 million records that have 36 properties. It includes the reference of the company and includes a list of 66 possible natures of control over a company which act as the direct relationship between PSC and companies dataset. PSC can be individuals with name and address information, companies who have address and registration information and legal entities who can have name, registration and address information. While discussed later, all fields except company number (excluding corporate number), kind (determines type of entity) and nature of control have data quality issues associated with them. It also includes companies that are exempt and reasons for non-compliance. We select only individuals as it comprises the majority of the dataset - see Table

Table 1
Table of Companies House company data cleansing and standardisation actions.

Column	Cleansing / Standardisation Process
Company Name	Extracts special character information along with standardising key values such as LTD, Limited.
	Standardise name based on rules from Companies House company search.
Country	List of UN Geoscheme countries was expanded to include UK regions. This was run against the country column to validate all the distinct countries in this column.
	Countries without PSC transparency were labelled as such through comparison to set of countries on the secrecy index dataset.
	As country of origin always contains a value, if country of registration was empty we inputted country of origin here.
Postal Town	When post town was not populated but postcode was, we inputted the post town value.
All fields	For dates – we ensure consistency by parsing all dates into a single format DD/MM/YYYY.
	Standard uppercase and punctuation.
	Cleansing and substitution of values to single values e.g. too, two, to.
	Replace often abbreviated words with their abbreviations.
	Extract characters and commas that could be used in SQL injections such as |, “”,”, .

2 for summary of cleansing actions taken during pre-processing to ensure data quality was maximized before entity alignment.

4. Knowledge Graph Construction

Knowledge graphs construction can be classified as bottom-up when structures within data define the ontology - typically used in iterative implementation with minimal lead time to expanding functionality. Or top-down when well defined domain ontology and schema are considered first then the knowledge base is populated - typically used when interoperability of knowledge graphs is required.

A bottom-up approach was selected as the knowledge graph implemented heterogeneous data sets incrementally - being unable to define the resulting ontology prevented a top-down approach. The main limitation from this is limited interoperability with other knowledge graphs.

The construction of bottom-up knowledge graphs can been observed to follow a common architecture. This has been reviewed extensively by [4].

4.1. Knowledge Extraction

Heterogeneous datasets by their nature are provided in a variety of formats - knowledge extraction describes the approaches to identifying diverse entities, their relations and attributes within semi-structured and unstructured sources and transforming them into a more uniform format.

Table 2
Table of Companies House Persons of Significant Control cleansing and standardisation actions.

Column	Cleansing / Standardisation Process
Name	Titles were not expected in forename and surname yet were present. Applied clean titles rule to removed titles such as "Ms, Dr" from these columns.
	Performed profiling using string length, unexpected characters and numerals to identify outliers.
	Used a regular expression of common English words "to, the, and, at, a" against the name columns to profile non-name words to cleanse.
Country	Performed profiling and standardisation on all country columns.
	The three country columns were appended.
	Profiled the combined list to identify unique values (6.8K). This indicates data quality issues in this column. Shown and discussed in Figure 5
Postal town	Populated postal town based on associated postcode in corpus.
Postcode	Validating the structure of the postcode against valid postcodes from UK, US, India
All rows	For dates – we ensure consistency by parsing all dates into a single format DD/MM/YYYY.
	Standard uppercase and punctuation.
	Cleansing and substitution of values to single values e.g. too, two, to. Replace often abbreviated words with their abbreviations.
	Extract characters and commas that could be used in SQL injections such as |, "";", .

The diversity of entities and relationships dictates the available solutions. Natural Language Processing and Machine Learning methods are practical when handling diverse datasets, however they require significant investment in manually annotating training data to produce a model with acceptable accuracy. Rule-based methods are practical and provide high accuracy if all entities and relationships can be identified by matching with a set of manually developed predicates - it requires comprehensive implementations and a fixed structure which is only feasible in well defined datasets.

Ultimately, so long as the pipeline can differentiate the types of entities then either approach can be used, but a rules-based method was employed as the identified datasets came from semi-structured sources with clearly defined columns - making it easy to identify their type from the structure.

Implementation Figure 2: Entities in JSON were extracted using a series of rules that mapped all attributes to a csv row per entity - flattening the JSON to be in the same format as companies data. Each property had a corresponding column which had the heading of the fully qualified path of the property so no information would be lost. As natures of control was a limited list it was encoded in order to make it simplistic to identify the relationships.

Profiling was carried out to understand the content of each column. Investigating the structure of both datasets to identify candidate entities and the attributes involved in their conditional functional dependencies [7]. In general, the smallest subset of attributes that could be shared between entities were considered separate entities. For example the forename, surname, birth month and birth year was the smallest subset that could describe a person as their address could be considered a separate entity that people and companies share. In addition, if

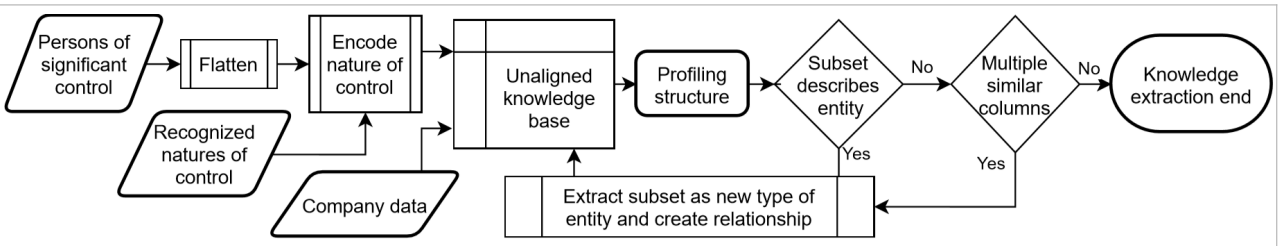

Figure 2: Process of knowledge extraction

the record could have multiple columns describing the same values then it would be best to be extracted as an entity and to create a relationship, such as 'country of origin' and 'registered country' in company data as well as 'country of residence','country registered' and 'country' in PSC.

4.2. Data Quality Improvement Pipeline

The focus of algorithms developed for entity alignment is the prediction accuracy and not the data itself so literature selects datasets with reasonably good data quality as they require minimal preprocessing actions. Even with reasonably good quality data, single algorithm approaches to entity alignment are too complex to be practical for large datasets - so a combination of simple matching strategies is typically used in the entity alignment algorithm. However, even the most advanced combination approaches cannot create knowledge graphs from sources with significant data quality issues [9], which is common place in open data. Therefore to have a practical implementation of knowledge graph construction from open data sources, data quality improvement is necessary regardless of algorithm.

The data quality pipeline encompasses the data preprocessing and entity alignment processes within knowledge fusion stage in the architecture of knowledge graph construction.

4.2.1. Data pre-processing

To address the challenge of parallelism, the pipeline partitions entities into concurrent specialised cleansing and alignment algorithms which align a certain type of entity via a series of matches.

A requirement of parallelism is that there needs to be no inter-partition communication. To satisfy this, entities were profiled then extracted to include any attributes involved in their conditional functional dependencies [7], as this resulted in disjointed partitions that contained all the attributes that would be involved in entity alignment (excluding those in external datasets).

Understanding conditional dependencies within the data goes beyond structural information. An expert or sufficient metadata is required to understand which values result in different entities or relationships being represented. Without this foresight, entities in a partition may have attributes in another - requiring inter-partition communication to complete tasks and therefore not be concurrent.

Comprehensive data cleansing was carried out within the pipeline to maximize data quality and has been summarised in Table 1 and Table 2.

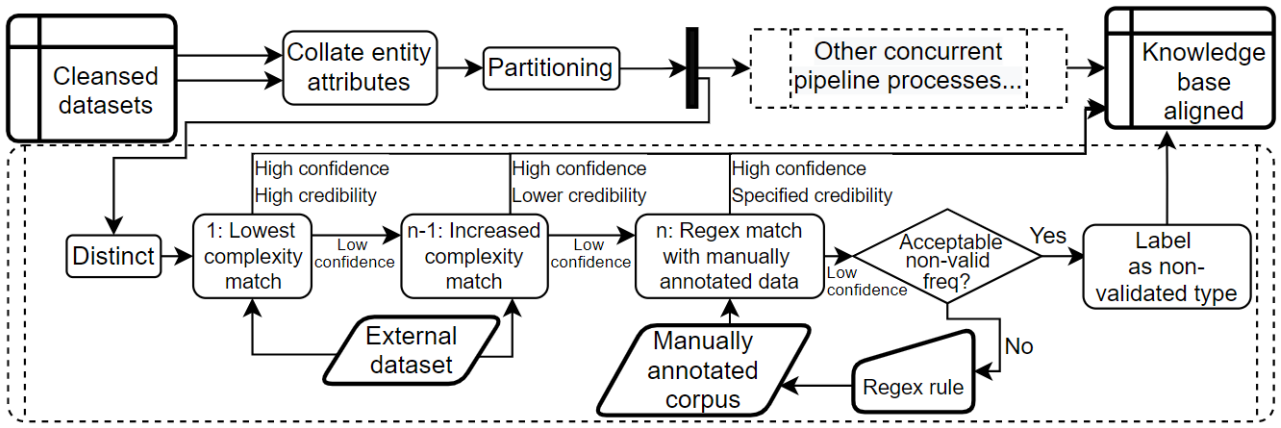

Figure 3: Generic entity alignment algorithm within pipeline

4.2.2. Entity alignment

The aim of entity alignment is to produce a set of unique entities that represent the real scenario. Duplication is expected from knowledge extraction but this leads to an unrealistic model in which relationships of a single real entity are spread across multiple duplicates that may not have any direct connections and therefore any network analysis would be inherently inaccurate.

For example, in the considered dataset this is taken to the extreme as multiple companies can be controlled by a single person and multiple people can control a single company in reality, however the data only represents the latter (shown in Figure 1) as no connections exist between people - network analysis only produces meaning results when the network represents reality, so would be ineffective. Entity alignment in this example takes all duplicated entities that had a single relationship and produces a set of unique entities with all relationships of their duplicates - representing the real many-to-many relationship.

According to [4], entity alignment in bottom-up approaches consists of textual similarity functions (like lexical similarity) producing values used in pairwise alignment algorithms (e.g. Levenshtein) then structural similarity functions being applied on collective alignment algorithms [8]. This process often relies on external datasets to validate entities. Due to the PSC dataset not containing enough distinguishing attributes to rely on the structure - this paper focus on textual similarity like those employed in [8] and demonstrated to be practical for OGD in [6].

The calculation of similarity metrics is used to indicate how likely two entities are to referring to the same entity. Multiple matching strategies may be applied within a single entity alignment algorithm and many exist for specific applications [9]. The resulting values are multidimensional so similarity combination is the process of evaluating all similarity scores and returning a single score. Alignment judgement is interpreting that score to specify which matched entities refer to a single entity then typically determine the winning attributes.

Character-based lexical similarity metrics: Lexical similarity calculates how similar a considered string is when compared to a defined corpus - the corpus is usually constructed from an external dataset of valid entities. Character-based lexical similarity measures how many

subtractions, updates and additions it takes for two strings to be equal.

An example is Levenshtein-ratio (occasionally known as fuzzy string matching), which considers all edit operations between two strings as the same weight [15]:

$$\mathrm{lev}_{a,b}(i,j) \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j))} \end{cases} & \text{otherwise.} \end{cases} \quad (1)$$

"where $1_(ai \neq bj)$ is the indicator function and it is equal to 0 if $a_i = b_j$, equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first i characters in a and the first j characters in b" [15]

The constructed knowledge graph utilizes a series of algorithms that returns lexical similarity - multiple algorithms where applied in series when the corpus that represented the allowable set of entity property values, didn't return satisfactory results.

Content-based category similarity metric: With content-based category metrics, similarity depends on the attributes shared by nodes - so similarity is calculated by "quantitatively evaluating the common information content of two categories" [8], matching-distance between compared content is used to evaluate the likelihood that comparisons refer to the same entity. However it assumes attributes are equally important, when in reality their importance actually depends on context. So improvements consider weights in the calculation.

The matching-distance similarity measure requires a selection of attributes that adequately differentiates the entity for others - attributes with without uniquely identifiable information need to consider many more attributes - considering additional attributes is adding another dimension in algorithms that have a polynomial time complexity so lead to severe performance issues. The comparison of time complexity is discussed in the results and the impact the data quality improvement pipe line provides, but common algorithms are defined below:

Simple matching coefficient compares how similar attributes are. It considered all attributes equally as important. It considers mutual absence in the nominator and denominator - which may not be realistic when comparing two subsets [15]:

$$SMC = \frac{M_{00} + M_{11}}{M_{01} + M_{10} + M_{00} + M_{11}} \quad (2)$$

Jaccard index is similar however it excludes mutual absences of both sets - so is effective when comparing subsets [15]:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{|X \cap Y|}{|X \cup Y|} \quad (3)$$

Overlap coefficient is the overlap of both sets [15]:

$$overlap(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \quad (4)$$

Similarity combination and alignment judgement: The process of similarity combination takes the multidimensional results of those metrics and ultimately produces a single similarity score.

Alignment judgement is determining how those multidimensional results should be evaluated, producing a predicate that determines if entities are referring to the same entity, The typical approach to this is to unilaterally apply algorithms to all entities as this allows all entities to be matched to very high accuracy, however this results algorithms having the worst case of polynomial time complexity.

The chosen algorithms consider a subset of attributes that is the minimal amount to distinguish attributes. When attributes have high confidence in distinguishing entities (like company number), the algorithm requires few attributes - however entities that have only poor confidence in distinguishing attributes requires a larger number of attributes to be considered in algorithms for a similar result. Increasing attributes involves an exponential increase in time complexity of algorithms. Poor confidence attributes mainly exist because they contain non-unique values or that data quality issues have degraded confidence in the data. The implementation minimized the polynomial function in a number of ways:

1. Increasing the data quality before applying these algorithms boosted confidence in attributes so that entities could be matched in earlier algorithms (reducing size of entity pairs being considered), many new entities had enough confidence to match and entities without distinguishing attributes required fewer attributes.
2. Rather than consider all entity-pairs at once, the data quality pipeline selects only entity-pairs that would derive benefit from each algorithm. Rather than consider all metrics at once, the entity-pairs are filtered through a series of algorithms, where as if there is a confident match (e.g. exact country match) then it is bypasses other similarity algorithms. Entities with confidence lower than the matching threshold underwent another algorithm to align as may as possible with the reduced set of unaligned entities. The time complexity of algorithms increased as the number of considered entities decreased - resulting in the algorithms with the highest polynomial function being applied on a minimal set of entity-pairs.

4.2.3. Ontology construction:

Bottom-up approaches like the one implemented have the data drive the formal definition of the ontology. Aligned entities and relationships are considered the knowledge base and are investigated to provide an ontology which would define the knowledge graph.

RDF and graph database are the two most widely implemented approaches to storage of knowledge graphs. Graph database was chosen as it performs faster when running queries and constructing graphs of a pre-defined structure. Neo4j is the most commonly used application for visualizing graph databases so was chosen.

The knowledge base was formatted to specify the ontology of entities and relationships according to neo4j syntax then the knowledge graph was constructed within minutes using their bulk import tool. The knowledge graph was visualized.

As expected, some PSC had such poor DQ that they couldn't be repaired and thus were missing relationships - but we discuss the improvement in results. Address and countries were shared by PSC and companies so were extracted to cluster entities around them. The data quality pipeline repaired data quality issues, derived new attributes and validated entities. The new ontology can be seen represented Figure 4.

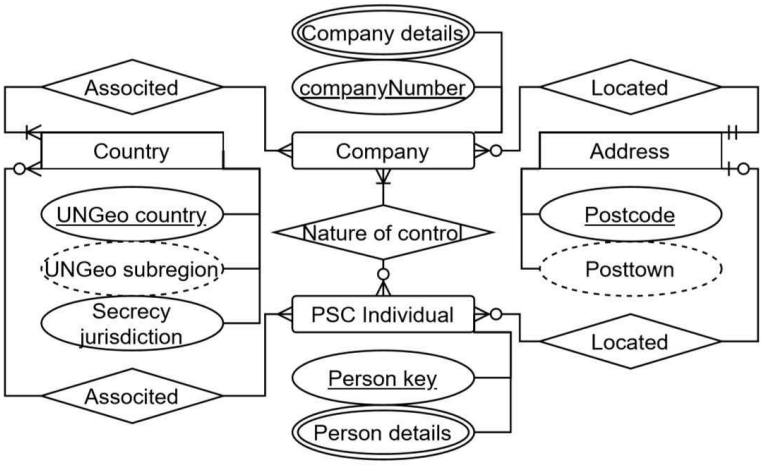

Figure 4: ERD of resulting ontology

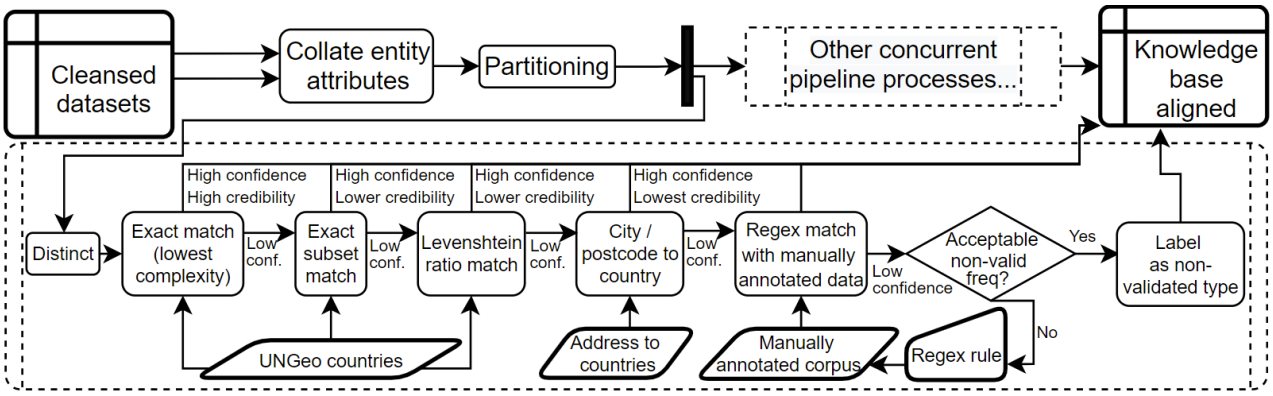

Figure 5: A specific implementation of the entity alignment within the pipeline for the countries partition.

5. Results and Discussion

In this section we are going to highlight the knowledge extraction, developed ontology then the results of implementing the data quality pipeline within knowledge fusion. Those benefits focus on improvements that facilitate knowledge graph construction from realistic and large-scale data sources with significant data quality issues and reducing the time complexity limitations identified as challenges from studies.

5.1. Knowledge Extraction

The rules-based implementation of knowledge extraction shown here Figure 2 was sufficient to extract a knowledge base of unaligned entities. The profiling to understand the structure and conditional functional dependencies was effective in identifying then extracting disjoint sets of entities by type as no inter-partition communication was required during entity alignment - facilitating concurrency of entity alignment algorithms.

5.2. Ontology

We developed an ontology to represent the control asserted by individual PSC.

The scenario represented in the original datasets looks like Figure 1. The lack of a many-to-many relationship between PSC and company meant it was unrealistic as the fundamental many-to-many control structure wasn't represented. Subsets of attributes like address were shared between entities but couldn't be clustered on, so entities at the same address were as closely related as entities in another country.

Whereas a realistic ontology Figure 4 produced from the data quality improvement pipeline Figure 5 has the following ontology. It models the missing1 many-to-many relationships, clusters entities that have similar attributes, validates country and address details and repairs dirty data that would typically degrade the performance of the entity alignment algorithms.

5.3. Boosting Data Quality to Reduce the Reliance on Complex Entity Alignment Algorithms

Time complexity in knowledge graph construction is centred around entity alignment. The presence of data quality issues that would degrade accuracy can be overcome with more complex entity alignment algorithms, however it is a polynomial cost function - addressing the data quality issues is a linear cost function.

We show how the cleansing, validation and repairing in data preprocessing (shown here 1) that addresses data quality issues in attributes that poorly distinguish entities can be sufficient to create a realistic network from the simplest match - where a complex algorithm would otherwise be required.

When there is no significantly distinguishing attributes such as phone number, then it isn't practical for large datasets. We show how an implementation of the data quality improvement pipeline on forename, surname, birth month and birth year can implement cleansing to simplify the required matches.

We select a reduced amount of attributes, profile attributes using string length, unexpected characters and numerals to identify outliers for data quality issues and use those findings to create a manually annotated corpus used to cleanse data. While the pseudonymised individual details in PSC prevents validating individuals that was possible with countries, the cleansing standardised name for consistency and removed irrelevant titles to improve accuracy. The example of this impact is cleansing the forename 'The Executors Of The Estate Of Geoffrey' to 'GEOFFREY' – however while cleansing was applied to all, 1,633 of the records had additional cleansing applied that would have required a more complicated match. We show how this cleansing is practical by implementing the simplest exact match and it creating a realistic structure.

If pre-processing was limited to those found in literature like then the scoring functions of entity alignment algorithm for syntactic matching would implement multiple matching strategies similar in complexity to 1 which is based on edit distance between two strings. As described in [9], even with these complex algorithms the poor data quality would result in missing matches that are significantly different like in the 'GEOFFREY' example, so while a complex algorithm would provide more matches than no cleansing, it would still miss matches with data

quality issues. A common approach in literature is to increase the amount of attributes to dilute the impact of data quality issues, but this results in an exponential increase in complexity.

We find that our data quality pipeline improves that data quality sufficiently so that a reduced amount of attributes need to be considered and the algorithm can be simplified and produce comparable results. The pseudonymised nature of PSC entities means they can't be distinguished further by textual similarity so incorporating structural similarity could be implemented in the future to distinguish individuals with the same details but disjoint egonets.

5.4. Reducing Time Complexity within Entity Alignment

In addition to boosting data quality to reduce the reliance on entity alignment algorithms – we demonstrate how a data quality pipeline can implement a series of increasingly complex matches with a minimal amount of entity-pairs – and how it provides the opportunity to prioritise manual labelling when validating against an expected list. We demonstrate this through aligning the 17,292,145 country values from 4 columns – of which 7,724 are distinct values to an external dataset of recognised countries under the UN Geoscheme [16], implementing ISO country codes [17].

We identify data quality issues originating from (1) column combination, (2) misspellings, (3) aggregated countries, (4) countries that no longer exist, (5) entering values that aren't countries and (6) values that have no logical reason to be entered like '12 July 2012'. All of these are considered valid by companies house but not fit for purpose – the data quality pipe line provided the opportunity to validate countries against a recognised list of countries and to implement multiple effective matching algorithms. While validating knowledge graph construction against rules derived from external dataset has been undertaken before in [18], the data quality pipeline provides the opportunity reduce the entity pairs applied in the alignment algorithms that makes if scale better.

Extract all company columns and partition from rest of the dataset – this allows each process to be carried out in parallel with other cleansing and alignments. Cleansing to standardise values was carried out to apply cleansing rules then profiling was carried out to identify data quality issues. As a limited list was expected, entities were made distinct to reduce the entity pairs. It then undertook a series of matches utilizing the external dataset as shown in Figure 5. The important aspects are that it begins with the simplest and most comprehensive matching then then subsequent matches are more complex. Notice how matches with high confidence results bypass additional similarity metrics – this results in the entity being aligned with the simplest match possible – contrasted to approaches like [8] and [19] in literature which calculate multiple similarity scores for all entity then combines the score into a single value to judge if it should be aligned. While their match considering all similarity scores will result in more accurate matches, the polynomial time complexity makes it impractical for large datasets – this papers implementation of the data quality pipeline provides significant dimensionality reduction to make large datasets feasible.

Non-validated entities can be prioritised based on frequency and therefore the number of entities to manually annotate can be prioritised. As experts are involved in manually annotating training datasets in entity alignment algorithms that utilize ML, this can be a reduced amount of work as only the low-confidence subset requires annotation in the pipeline. The

pipeline including the prioritised annotation resulted in successfully validating over 99% of the country entities - as only 12,000+ couldn't be repaired even with manual annotation. Principles for designing combination of matching strategies algorithm:

1. entity's expected to be from a limited set should be compared against externally recognised reference list
2. comparisons should begin with the lowest time complexity matches like exact matching
3. subsequent algorithms should be increasingly complex with minimal entity pairs
4. matching historic and aggregated data to valid entities may reduce accuracy which conceal malicious intentions, match/label with most likely value that retains suspicion
5. complex algorithms with a low return should be implemented if they can be selectively applied, such as extracting country from postcode only if it matches a postcode regex
6. values that don't match an external set of recognised values should be retained, but labeled as non-validated
7. manual annotation may be required for non-validated data, so integrate profiling to prioritise effort on most common values

We compare this to a typical approach shown in table 3. Each row corresponds to an additional stage that may be applied in pre-processing. A higher percentage of validated entities corresponds to requiring a less complex alignment strategy. Stages: (a) Initial values; (b) Distinct to reduce duplicates; (c) Case insensitive; (d) Standardize punctuation; (e) Remove punctuation.

The external dataset used for validation is the UN Geoscheme countries, it notes 249 countries.As UK is considered a single entity in this dataset, if used without any pre-processing it would result in only validating 3.421% of the dataset before alignment, therefore to provide a more realistic use case, the values of the relevantly clean company data were all mapped to the dataset. 129 matched exactly, 23 required modification to match, 29 were required to be inserted (mostly countries or variations of UK and Irish countries and some dissolved countries) and 97 had no matches so remained as the same value as in the initial dataset. In total, 278 countries were identified in this approach.

Within 3, "$\cap \subseteq$" is used as shorthand to describe the number of distinct validated entities, which can be compared to the "Count" of entities after Stage (b). "$\cap \sum$" refers to the percentage of total validated entities this validates in the selected column(s). The shaded cells

The typical approach achieves below 62% in validating all entities - this requires a more complex entity alignment algorithm to align entities with the same accuracy compared to our approach.

5.5. Derived Attributes

We additionally show that the data quality pipeline can be used to improve the completeness of the knowledge graph by imputing values from the validated values within the graph. While it has been carried out previously – incorporating it into partitions allows for actions to completed concurrently. For example we use postcode and city to extract post town and nation for the dataset – we include additional country information like its region and if it can be trusted. All of this information provides tangible benefit to network analysis. For example, the report

Table 3
Table decomposing the number of distinct validated entities and resulting percentage of countries validated in a typical approach to pre-processing.

	CH PSC									CH CD			CH		
	id.country_registered			country_of_residence			address.country			RegAddress.Country			All		
	Count	∩ ⊆	∩ ∑	Count	∩ ⊆	∩ ∑	Count	∩ ⊆	∩ ∑	Count	∩ ⊆	∩ ∑	Count	∩ ⊆	∩ ∑
(a)	505714	149	47.783%	7084230	204	62.378%	6245555	200	58.063%	3456646	153	62.997%	17292145	217	60.517%
(b)	2934	149	47.783%	4882	204	62.378%	1361	200	58.063%	192	153	62.997%	7724	217	60.517%
(c)	2934	157	73.909%	4882	222	62.609%	1361	212	58.181%	192	153	62.997%	7543	233	61.418%
(d)	2884	174	73.928%	4724	311	62.615%	1322	239	58.208%	192	153	62.997%	7304	351	61.431%
(e)	2539	212	74.052%	4586	327	62.618%	1148	340	58.233%	192	154	63.006%	6691	483	61.446%

'The companies we keep' identifies that individuals who originate from countries high on the financial secrecy index, may not be relied on. They find that certain countries like Ukraine protect details of PSC making it appealing to money laundering. By creating a knowledge graph that is enriched by a data quality pipeline, we can conduct better network analysis by considering missing links and can cluster entities on more shared attributes. The drawback is internal data may need to be duplicated if used by multiple partitions to avoid dependencies, however as our example using countries uses an external list, it wasn't required.

5.6. Limitations

The following are some of the major limitations in the implementation of the pipeline.

Partitions require disjoint datasets to avoid inter-partition communication, which is only possible if data sources have a well-defined structure – this means that it isn't practical for unstructured data sources and data duplication may be required if attributes are shared between entities but cant be extracted as an entity. The series of entity alignment algorithms excluding high confidence matches means that entity judgment is dependant on order of operations and that the most important accurate matches have the lowest time complexity.

Entities that didn't match were excluded, when they should be retained and specified as non-validated. Also original values should be retained to comply with regulations of transparency.

Partitions are based on entity type and not size so parallel processing would never be balanced. This effect is reduced the greater the number of types of entities.

The entire pipeline requires a great deal of expert knowledge to implement. Particularly the cleansing, which becomes less feasible the more entities there are.

The bottom-up approach isn't suitable for interoperability with other knowledge graphs and it hasn't been considered for top-down approaches. Implementing the ontology following a comprehensive schema like one provided by schema.org [20] would allow the same approach but make it easier to integrate with other anthologies.

The datasets were semi-structured and conditional functional dependencies were easy to identify in knowledge extraction due to sample size - it is unlikely that disjoint sets of entities could be extracted from unstructured data sources using the pipeline. NLP approaches could be a solution but the sporadic extraction of attributes would degrade algorithms that relay on complete data.

6. Future Work

Knowledge graph construction is an iterative process, and therefore, additional work can expand it's functionality. The constructed knowledge graph wasn't compared with knowledge graphs using different approaches. Differences in network analysis should be investigated to determine how effective the pipeline is at creating the graph from the same data.

The PSC dataset was limited to individuals which was the majority of entities, however, including other types would provide a more representative network and be more heterogeneous in nature – providing a more accurate comparison against other algorithms.

The current approach doesn't discuss selecting the winning attributes of aligned entities unless they are validated – realistically when referring to an individual entity there should be a golden record that describes it the most likely attributes.

7. Conclusion

As open data use becomes increasingly pervasive in supporting decision making, construction of the knowledge graphs that models and facilitates the network analysis increases in scale and data quality issues.

As described in [21], algorithms that focus on developing overly complicated models to dilute data quality issues or develop solutions with pristine data, degrade their usefulness in real-world scenarios. Academic papers on entity alignment are observed as following this as they propose algorithms with a polynomial time complexity which aren't feasible to create on large-scale open data.

We show how a use-case built from a typical approach to pre-processing in entity alignment 3 can validate entities against external data sources to reduce the need for complex alignment algorithms and that it can be substantially improved with a dedicated pipeline such as what we propose.

The implementation of the data quality pipeline is a practical approach which attempts to overcome identified challenges of large-scale knowledge graph construction from dirty data but requires a great deal of expert knowledge to design initially. Maximizing data quality requires significant development time but it has more of an effect in terms of accurately reflecting entities than increasing the complexity of alignment algorithms.

References

[1] J. Attard, F. Orlandi, S. Scerri, S. Auer, A systematic review of open government data initiatives, Government Information Quarterly 32 (2015) 399–418. doi:10.1016/j.giq.2015.07.006.

[2] R. Matheus, M. Janssen, A systematic literature study to unravel transparency enabled by open government data: The window theory, Public Performance & Management Review 43 (2020) 503–534. doi:10.1080/15309576.2019.1691025.

[3] Company Register, Open knowledge - company register 13 percent open, 2021. URL: https://index.okfn.org/dataset/companies/, accessed: 2021-06-25.

[4] Z. Zhao, S.-K. Han, I.-M. So, Architecture of knowledge graph construction techniques, in: International Journal of Pure and Applied Mathematics, 2018, pp. 1869–1883.
[5] W. Hu, J. Chen, Y. Qu, A self-training approach for resolving object coreference on the semantic web, in: Proceedings of the 20th International Conference on World Wide Web, WWW '11, 2011, p. 87–96. doi:10.1145/1963405.1963421.
[6] J.-w. Lee, J. Park, An approach to constructing a knowledge graph based on korean open-government data, Applied Sciences 9 (2019). doi:10.3390/app9194095.
[7] W. Fan, Data quality: From theory to practice, SIGMOD Rec. 44 (2015) 7–18. doi:10.1145/2854006.2854008.
[8] K. Sun, Y. Zhu, J. Song, Progress and challenges on entity alignment of geographic knowledge bases, ISPRS International Journal of Geo-Information 8 (2019) 77. doi:10.3390/ijgi8020077.
[9] K. Zeng, C. Li, L. Hou, J. Li, L. Feng, A comprehensive survey of entity alignment for knowledge graphs, AI Open 2 (2021) 1–13. doi:10.1016/j.aiopen.2021.02.002.
[10] C. Fürber, M. Hepp, Towards a vocabulary for data quality management in semantic web architectures, in: Proceedings of the 1st International Workshop on Linked Web Data Management, ACM, New York, USA, 2011, p. 1–8. doi:10.1145/1966901.1966903.
[11] D. C. Corrales, A. Ledezma, J. C. Corrales, From theory to practice: A data quality framework for classification tasks, Symmetry 10 (2018). doi:10.3390/sym10070248.
[12] H. Chen, G. Cao, J. Chen, J. Ding, A practical framework for evaluating the quality of knowledge graph, in: Knowledge Graph and Semantic Computing, Springer Singapore, 2019, pp. 111–122. doi:10.1007/978-981-15-1956-7_10.
[13] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, D. Q. Phung, A novel embedding model for knowledge base completion based on convolutional neural network, CoRR (2017). doi:10.18653/v1/N18-2053.
[14] D. Q. Nguyen, T. Vu, T. D. Nguyen, D. Q. Nguyen, D. Q. Phung, A capsule network-based embedding model for knowledge graph completion and search personalization, CoRR (2018). doi:10.18653/v1/N19-1226.
[15] M. K. Vijaymeena, K. Kavitha, A survey on similarity measures in text mining, Machine Learning and Applications: An International Journal 3 (2016) 19–28. doi:10.5121/mlaij.2016.3103.
[16] UN Department of Economic and Social Affairs, Countries or areas / geographical regions, 2020. URL: https://unstats.un.org/unsd/methodology/m49/, accessed: 2021-06-25.
[17] International Organization for Standardization, Iso 3166 - country codes, 2020. URL: www.iso.org/iso-3166-country-codes.html, accessed: 2021-06-25.
[18] S. Guo, Q. Wang, L. Wang, B. Wang, L. Guo, Knowledge graph embedding with iterative guidance from soft rules, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: www.arxiv.org/abs/1711.11231.
[19] Z. Wang, J. Li, J. Tang, Boosting cross-lingual knowledge linking via concept annotation, in: IJCAI, 2013, pp. 2733–2739.
[20] Schema.org, Corporation, 2021. URL: www.schema.org/Corporation, accessed: 2021-6-25.
[21] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. K. Paritosh, Lora, Aroyo, Everyone wants to do the model work not the data work: Data cascades in high-stakes ai, in: 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–15.

PaperID : paper15

Title :
Data Quality Improvement and Entity Alignment Optimization for Constructing Large-Scale Knowledge Graphs

Authors :
Keaton Sullivan, Fiona Browne, Huiru Zheng, Haiying Wang

Pages in Proceedings :
68 - 85

created 04 Dezember 2022 @ 23:31