Knowledge Graphs in Support of Credit Risk Assessment

(research in progress)

Abstract

In comparison with domain ontologies, knowledge graphs are less complex to build. They remove the burden of specifying boundaries for the domain and reduce completeness and consistency requirements. They have been successful in facilitating knowledge reuse and maintenance. Adding knowledge continuously, in small localised chunks, is easier than the holistic engineering required for ontologies. In this paper, we exploit this to use knowledge graphs in combination with ontologies for *transfer learning* in machine learning. Through the use of knowledge graphs, data is extracted and transformed from one domain to another where data is lacking. This synthesized data is then used to support machine learning overcoming the lack of data. This approach is illustrated to support transfer learning in lending risk assessment. The approach provides a template for supporting data driven innovation as a finance company explores new markets and designs new products.

Keywords: Innovation in finance, Ontology alignment, knowledge graph, lending, transfer learning**.**

1 Introduction

With continued increasing competition from Fintech and the current COVD-19 pandemic-triggered recession, financial institutions have been pushed to innovate by introducing new lending products, target new customer segments, or looking at existing customers in a different lens. Applying traditional risk assessment using historical lending data is often not sufficient to truly understand the customers to assess credit risk in a rapidly changing environment. Relying on historical data alone can result in limited or unaffordable credit for some individuals and small businesses. Transfer learning can potentially reduce this limitation, by leveraging knowledge from related domains, with sufficient outcome data (Suryanto et al 2019). Transfer learning from related domains is a potential solution to augment this lack of information and improve financial inclusion. For instance, transferring knowledge from credit card/debt consolidation loans to more risky small business loans or from utility bill payments to loan repayments could potentially deliver a more accurate scoring model. In this paper, we propose an approach to support transfer learning by using *ontology alignment* across domains to adapt data from existing domains to domains where data is lacking. Ontology alignment (Dragisic et al 2016) between two domains facilitates the mapping of data by identifying higher order relationships between the corresponding concepts in the two domains. For example, the “loan limit” concept in credit card domain has a different operationalisation in “small business loans”. With appropriate mapping across other features, it would be possible to produce a mapping function between the two domains but the mapping will need to be expressed using high order concepts. With appropriate mapping, data use in credit card risk assessment can be processed to generate reusable data in “small business loans”. Synthesis of mapping between domains often requires intermediary bridging domains. To ensure knowledge bridges are available, a knowledge graph (KG) based architecture is proposed to support mappings when required. The architecture integrates a knowledge graph with a financial data lake to enable an easier formulation of the ontology mappings across related domains, to support transfer learning. Knowledge graph technology has received increasing industry interest due to simple maintenance and traceability. A number of publicly available general knowledge graphs have become recently available e.g. Yago, NELL and DBPedia.

The proposed architecture advocates a smaller customised knowledge graph that accumulates organisational know-how without imposing the engineering burden of a formal knowledge structure. This architecture also enables the financial institutions to explain the credit assessment logic, which is a requirement for the ML adoption in banks. Concepts in a KG are sparsely connected to enable complete reasoning to enable data mapping across two domains reliably. Our approach resolves this by combining KGs with a richer description of specific domains using ontologies. The rest of the paper is organised as follows: Section 2 presents the background and related work that supports the proposed approach. Section 3 discusses the proposed approach and the KG-based architecture. Section 4 presents an exemplar of data mapping between two related lending areas illustrating how the approach can produce data from one data rich domain to another data poor domain. Finally, Section 5 concludes with a discussion of future possibilities.

2 Related Work and Background

An ontology is a formal and reusable knowledge structure that pertains to a specific domain of expertise. An ontology consists of a set of concepts that describe the domain and their relationships. In addition to knowledge reuse, once available, an ontology can provide system interoperability, problem solving methods reuse and readability (Beydoun et al 2020). Capitalizing on ontologies holds a promise to provide solutions that improve the transparency and traceability in artificial intelligence. Their use in combination with machine learning can support accountability requirement in many applications. This is particularly true in financial decision making. In fact, this is a regulatory requirement in many jurisdictions. As an interoperability mechanism, ontology alignment is the process of mapping concepts/relationships from one source ontology to another target ontology (Dragisic et al 2016). It is akin to language translation but rooted in formal symbols and logical relationships. Predominant formalism of current technology is Description Logic. In practice, tagging concepts and relationships in one ontology using terms from the other ontology. Once this alignment is established, data in the domain from the source ontology can be retagged with terms from the target ontology e.g. (Alruqimi 2019). This operation is of particular interest to our proposed approach to support transfer learning in finance and will later be illustrated.

The challenge in reusing ontologies, whether for data mapping or to enhance readability, is having appropriate ontologies at hand. An effective ontology needs to be complete and consistent. This requires deep domain expertise. An ontology gets developed with reusability in mind. This ideally takes place in the form of retrieving an ontology from an existing set of ontologies (a repository). The retrieval uses a 'synset' as a key to retrieve the most relevant ontology. Several cross-ontology similarity finding methods have been described in the last decade which, for the most part, make use of one or more techniques in combinations (Beydoun et al 2014). Often they propose matching some significant subset of the terms found within the two ontologies. The simplest means for assembling the term similarity techniques into cross-ontology similarity assessors is to assemble the two ontologies into a merged single ontology, inside which the earlier term-to-term tests may be applied. The assembly of such a unified ontology is a non-trivial task (AlMubaid et al, 2009). This approach can be computationally expensive when making numerous cross-ontology comparisons for the purposes of retrieving the best match from an ontology repository. A related approach is to make use of some large-scale and highly descriptive third ontology, such as WordNet. This approach offers the advantage of not needing to construct numerous merged ontologies. It typically makes use of feature-based comparison techniques which requires that the ontologies under review have sufficient descriptive features, concepts or attributes. However, it may not always suit scenarios in which relatively rapid or light-weight ontology creation and comparison is sought. This approach has been refined in recent years to enable a 'large ontology' to become easier to maintain. These refinements include simplifying relationships between concepts and storing instances (data) with known links to concepts. Any unknown links can later be discovered and added. Thus, the knowledge structure grows without any revision requirement. This approach has become quite popular in recent years and spawned into what is currently known as *Knowledge Graphs*. For our purpose, a knowledge graph is essentially built as a large ontology with simplified relationships between concepts where instances of concepts are also stored with the concepts. This can simply take the form of higher order features of the instances (data), or data tags. Most importantly, in a KG knowledge is constantly added as it becomes available, without been constrained by the semantic boundaries of a domain. This removes the burden of completeness and consistency, and enables easy maintenance and construction of KGs. However, this also makes KG's less reliable when accuracy and completeness are required. To have the best of both worlds, we combine KGs and domain ontologies.

In the proposed approach, instead of using a multitude of ontologies, we propose the use of a knowledge graph to act as rich metadata layer above all learning data, *a data lake*. Access to this data lake, during transfer learning, is mediated with ontologies. The focus of this paper is to illustrate the practicality of the proposal by highlighting the semantic mapping requirements and how these requirements can be resolved through ontology mappings.

3 KG and Ontology Mapping Based Approach

Data is a valuable asset of many businesses offering intelligent decision support services. E.g. lending decisions, land use decision, etc.. Data builds up as decisioning service providers build their customer base. Whilst the use of data is restricted and bound by confidentiality agreements, the learning models are often not. Hence, there is scope of transfer learning. It provides an opportunity to transfer models between domain without violating standing agreements. In addition, in many domains e.g. lending, it enables identifying overlooked market opportunities and scope for additional social responsibilities (e.g.

lending to disadvantaged communities where borrowers would be able to repay). This is where our approach is most compelling transferring learning to new markets where data is yet to exist or where only limited data is available.

With an appropriate ontology alignment (see Figure 1), suitable data to support learning is generated from existing data. Semantic relations are defined between ontologies and are applied to existing data. This transforms data from the source ontology to instances of a target ontology (Martins and Silva 2009). The relations can be identified through analysis and creating new tags for concepts describing old data, and subsequently used to transform the old data. A challenge is identifying a suitable set of initial data to execute the alignment. This is where the innovation in our proposal is compelling. By using a single data lake, supported by a KG, the appropriate data is automatically selected through the source ontology. In other words, the selection of the data is a two-step process:

- The ontology identifies suitable concepts in the KG.
- The concepts from the KG filter the required data.

The approach is illustrated in Figure 1 below. The architecture shown in Figure 1 requires business processes to maintain three elements as new domains are encountered:

1. Ontologies need to be created for each domain supported by the service provider.
2. The KG needs to be expanded as required as a result of the new ontology
3. Links between the data and the KG need to be maintained.

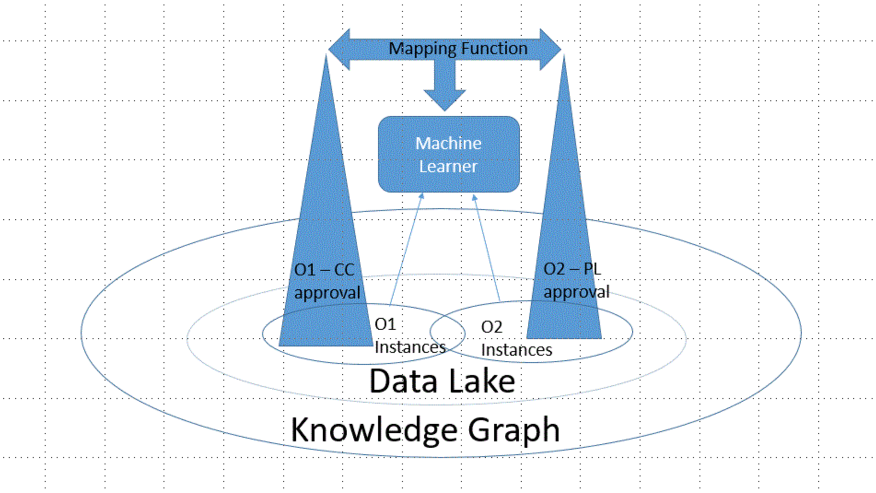

Figure 1. *KG-based Data Management Architecture*

Step 3 is possible only where data exists. If data does not exist for the new domain (say O2), then an ontology mapping is created. Data is created through an ontology mapping (say O1) that targets the domain where data is missing. This enables a transfer operation from one domain (O1) to another domain (O2). The mapping between two ontologies will depend on the differences between the two domains. If the domains are closely related and the two ontologies share most of the concepts, the mapping could be achieved through concept-to-concept translation. If there is a high degree of concepts misalignments, external ontologies to support the mapping would be needed. Metamodels can be used to achieve concept alignment. In many applications, there are extant metamodels that exist e.g. in lending, a standard metamodel Lixi exists and this can support mapping between two domains. In higher degree of misalignment, additional external ontologies may be needed. This is further elaborated with examples in the next section. The focus of this *Research-in-progress* paper is to illustrate how the above architecture can be operationalised through ontology mappings. For the purpose of this paper, sourcing the ontologies is assumed at this stage to rely on the available finance expertise, rather than reuse. In our illustration in the next section, reuse is confined to sourcing additional ontologies (or metamodels), to support ontology mappings.

4 Illustration: *Transfer Learning in Lending Assessment*

We illustrate the approach in credit risk. In addition to authentication of the applicant, the applicant is assessed for the likelihood that they are able and willing to pay by the period of the proposed loan. In this section, we present three different lending domains and illustrate how ontology alignment will enable generating data from one lending domain to another. The three domains are the following: *Payday Lending, Instalment Based Lending* and *Merchant Lending*. In all these three domains, the purpose of the assessment is to ensure that a client is capable of repaying before the loan is approved. All three loan types are unsecured i.e. they are not covered by any securities that can be repossessed if customer defaults. They can be lucrative to a lender but they are also risky. The loan products vary in terms of the period, the amount, the risk and the repayable amount. The third product type, *Merchant Lending,* is more different than the other two in that the borrower is a small business and the process requires revenue information of the borrower (a merchant) rather than personal income (as for the first two). The first two differ in period and amount. The overall assessment of the risk for each is different. For each domain, we review the data requirements for each risk assessment and present an ontology snippet. Ontology mappings across the domains would then enable data to be mapped from one to another.

4.1 Lending domains descriptions

Details of each lending domain is detailed in what follows.

PayDay loans: These are short term loans for cash strapped clients. They are riskier with a higher return. The loan amount is usually small and is typically less than the amount of immediate pay period. The pay period may vary from 1 week to a month. The interest rate is typically high, perhaps as high as 15% for a month. But the loan is also small, typically $500. Full repayment is expected at the next pay date. The repayment would be the principal and a fixed fee which include processing fees and the interest incurred over the short period of the loan. Loans of this type are risky and a default rate of 5-10% is not unusual.

Instalment loans: These are longer term loans than *payday loans*. A typical period is six months to 3 years. The loan amount is also larger. The amount is usually less than 40% of the income for the period of the loan, for example, if the expected income within 24 months is $24000, then max loan amount is $9600. The repayment is broken into instalments rather than full payment required in *payday loans*. The instalment repayment is aligned with the pay period, e.g fortnightly or monthly. The repayment date is usually after the first salary pay date. The interest is much lower compared to payday loan, typically 10% to 20% annually. The default rate is also lower (less than 5%).

Merchant lending is a new type of loan for which data does not yet exist. It stipulates a long term relationship between a lender and a business owner. The assessment is based on the revenue of the business rather than the net income. A lender's risk is offset by being able to sell services to support the transactions of the business and at the same time gain visibility of the business performance. Loan amount is assessed against card (credit/debit) payment received. Hence, the approval process can be expedited and the lender's visibility of the business also enables them to offer flexibility in the repayment. For example if the average daily revenue (received through card payments) is $1000, the loan repayment is set at 10% of the actual revenue, i.e. $100. These loans can also offer flexibility in the repayment period according to the performance of the business. For instance, during a pandemic period (COVID19 for instance), the period can be stretched.

4.2 Ontology Mapping and KG usage Outline

The ontologies for the first two loan types, PayDay and Instalment loans, are quite similar in the concepts used (these are shown in Table 1). This makes the synthesis of the ontology mapping easier. Concepts constraints and attributes do differ. The mapping will require taking those differences into account. For example, for PayDay Loans the maximum loan is $1500 or 40% of the pay amount (the smaller of the two). The DTI (debt to income ratio) for PayDay Loans is loan amount/monthly income whereas for Instalment Loans it is loan amount/yearly income. The risk grade for all these three products is a probability of default function. It is shown here as follows:

Risk Grade = E, if PD (Attributes of applicant, Attributes of Loan) >= 0.2
= D, if 0.2> PD >= 0.1
= C, if 0.1> PD >= 0.05
= B, if 0.05> PD >= 0.01
= A, if 0.01> PD >= 0

The calculation of the risk grade is a function that depends on the attributes of the applicant and loans. Lenders rely mainly on modelling, e.g. logistic regression, machine learning, for credit scoring. The mapping of risk grades between the three domains requires mapping between attributes of the applicants and the loans. The mapping will be based on the respective ontologies. This mapping can also make use of higher order functions where the input to the mapping from one domain to another, requires the risk function itself as input.

Concept	**Instalment Loans examples**		**PayDay Loans examples**	
	Example 1	*Example 2*	*Example 1*	*Example 2*
Loan Amount	*12000*	*15000*	*1500*	*1000*
Debt to Income Ratio (DTI)	*0.2*	*0.25*	*0.5*	*0.33*
Annual Income	*60000*	*60000*	*36000*	*36000*
Income Frequency	*Monthly*	*Fortnightly*	*Monthly*	*Monthly*
Income Type	*Permanent full time*	*Disability Benefit*	*Casual*	*Part time*
Repayment Amount	*1300*	*721*	*1725*	*1150*
Repayment Frequency	*Monthly*	*Fortnightly*	*1*	*1*
Interest	*30% per year*	*25% per year*	*35% month*	*30% month*
Fee	*0*	*0*	*100*	*100*
Loan Term	*12 months*	*12 months*	*30 days*	*30 days*
Start Date of Loan	*15/06/2020*	*25/05/2020*	*15/07/2020*	*25/07/2020*
Date of first repayment	*15/07/2020*	*08/06/ 2020*	*14/08/2020*	*24/08/ 2020*
Risk Grade	*B*	*A*	*D*	*C*
Job Type	*Labourer*	*Unemployed*	*Labourer*	*Professional*
Years at current job	*1*	*5*	*1*	*2*

Table 1. Concepts and examples within *PayDay Loan* and Instalment Loan domains

The knowledge graph can be used to support the quality of mapping. In some cases, additional knowledge can be used to provide additional insights. Risk depends on the applicant and what they do for living. In other words, risk profile of certain roles may differ even though they may have similar income. This role of the KG will become essential to deal with completely new domains that are substantially different. For instance, the Merchant Lending domain is quite different from the above two domains. The mapping between the concepts involved requires access to additional external knowledge. The role of the knowledge graph is more prominent in this case. For instance, to support the mapping between revenue and income, an external ontology describing various business attributes including their profit margins is required. For example, $600K revenue in a restaurant running at a profit margin of 20% is similar to net income of $ 120 K/yr. Whereas for an antic store business running at 50% profit margin, the same revenue is similar to a net income of $300k/yr. With access to such an external ontology, LRR can then be mapped.

Concept	**Example 1**	**Example 2**
Loan Amount	*60000*	*100000*
Loan Revenue Ratio (LRR)	*0.1*	*0.1*
Annual Revenue	*600000*	*1000000*
Frequency of Revenue Test	*Daily*	*Weekly*
Business Category	*Restaurant*	*Bar*
Repayment Amount	*164*	*274*
Repayment Frequency	*Daily*	*Daily*
Interest	*6000 (12%)*	*10000 (12%)*
Fee	*7200*	*12000*
Repayment period	*1 years*	*1 years*
Start Date of Loan	*1/02/2017*	*1/07/2019*
Date of first repayment	*2/02/2017*	*2/07/2019*
Risk Grade	*B*	*A*

Table 2. Concepts and examples within *Merchant Lending domain*

A strength of the above approach in generating artificial data, is that various policy settings can also be explored. For example, the mapping function can have additional dynamic parameters to adjust risk. Merchant lending data conversion to ‘payday lending’ can be made to produce more negative than positive learning instances.

5 Discussion and future work

In this paper, we have presented an approach to integrate the use of ontologies and knowledge graphs to support transfer learning. It is important to highlight that the approach has a wider applicability to support organisational innovation. Digitising operations of an organisation yields the required knowledge graph. The beauty of the approach is that operational knowledge is utilised with expert knowledge (in the ontologies) to support long term innovation. Within the banking sector, known for its conservative outlook, innovation pace can be enhanced with an approach that creates reliable artificial data for new product scenarios.

Our approach is based on a three layered architecture: data, knowledge graphs and ontologies. We illustrated how the architecture enables ontology alignment between different lending domains to generate data from a data rich domain to a data poor domain i.e. to support transfer learning. When a lender expands into new market segments, a new credit risk model is required to assess the credit risk of loan applications. The current approach is based on expert rules, where the credit risk expert builds business rules based on data and available derived data, combined with the expert's experience and knowledge. Lenders initially used an expert model to gather sufficient labelled data, to build a supervised learning model.

Supporting transfer learning is only one specific benefit of combining the use of ontologies and knowledge graphs. From a machine learning perspective, it also supports addressing the challenge of providing readability and traceability of AI-based Information Systems. For instance within the lending industry, the approach presents the reasoning and trace from data to the features, to serve as the missing link between transparency and explainability. We currently can explain how the features work within a model, but couldn't answer the question why we use these features. Knowledge graph will help us to answer the latter. The approach can also provide a different viewpoint and presentation/interpretation of data for different stakeholders to extract insights, e.g. virtual CFO dashboard for SME (from business owner viewpoint), account health check (from banker viewpoint), and portfolio dashboard (from credit analyst viewpoint).

The approach still requires synthesis of complementary processes to support the KG development and maintenance. It also requires automation of the ontology mappings. We plan to use WWW language offerings to create a working prototype.

6 References

Al-Mubaid, H., and Nguyen, H. A. (2009). Measuring semantic similarity between biomedical concepts within multiple ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* 39(4), 389-398.

Beydoun, G., Low, G., Garcia-Sanchez, F., Valencia-Garcia, R & Martinez R. (2014) Identification of ontologies to support information systems development. *Information Systems* 46,45-60.

Dragisic, Z., Ivanoa, V., Lambrix, P., Faria, D., Jimenez-Ruiz, E., Pesquita, C. (2016). User Validation in Ontology Alignment. In: Groth P. et al. (eds) The Semantic Web – ISWC 2016. ISWC 2016. Lecture Notes in Computer Science, vol 9981. Springer, Cham. https://doi.org/10.1007/978-3-319-46523-4_13.

Martins, H., Silva, N. (2009). A User-driven and a semantic-basd ontology evolution approach, ICEIS 2009 - Proceedings of the 11th International Conference on Enterprise Information Systems, Volume DISI, Milan, Italy, May 6-10, 2009.

Suryanto H., Guan C., Voumard A., Beydoun G. (2020) Transfer Learning in Credit Risk. In: Brefeld U., Fromont E., Hotho A., Knobbe A., Maathuis M., Robardet C. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science, vol 11908. Springer, Cham. https://doi.org/10.1007/978-3-030-46133-1_29

Beydoun, G., Hoffmann, H., Valencia Garcia, R., Shen, J., Gill, A. (2020).Towards an assessment framework of reuse: a knowledge-level analysis approach, Complex & Intelligent Systems (2020), Springer, 6:87–95.

Alruqimi M., Aknin N., Al-Hadhrami T., James-Taylor A. (2019) Towards Semantic Interoperability for IoT: Combining Social Tagging Data and Wikipedia to Generate a Domain-Specific Ontology. In: Saeed F., Gazem N., Mohammed F., Busalim A. (eds) Recent Trends in Data Science and Soft Computing. IRICT 2018. Advances in Intelligent Systems and Computing, vol 843. Springer, Cham.