Departamento de Ciencias de la Computación
e Inteligencia Artificial

Knowledge Discovery in Multi-relational Graphs

Advisor

This dissertation has been submitted by
Pedro Almagro Blanco
in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy
by the University of Seville

D. Fernando Sancho Caparrini

Seville, April 2017.

Contents

Chapter 1

Introduction

Man's curiosity to understand the environment in which he lives has allowed him to advance in his way of dealing with the unknowns and day to day problems. This primary need to understand the behavior of the phenomena that surround him leads him to create increasingly efficient methods used to discover and explain the workings of what he can perceive through his senses. For example, the enunciation and understanding of universal laws of physics allows him to emulate hypotheses about the past conditions of phenomena, to reason about their relation to the present that he perceives, and to predict their future evolution, going ahead in the decisions about events that have not yet occurred. Without possessing a perfect system of inference, it is certainly the capacity that differentiates him the most from the rest of living beings with which he shares his existence.

However, he must repeatedly redo and expand the discovery mechanisms he uses, adding new tools that allow him to continue to expand the frontiers of knowledge. Among past tools, the Scientific Method, and the scientific paradigms that are established by its application, stand out as great referents, being that they are characterized by demanding a rational justification of principles that needed to be proved and by rejecting absolute affirmations, assuming that any of them is susceptible to being refuted. The oldest scientific paradigms are the empirical (validate a hypothesis through experimental repetition) and the theoretical ones (to prove something through logical derivations). Later on, immersed in the computer age, a new scientific paradigm emerged that has come to be called a computational paradigm (to validate something through computational simulations that complement experimental observations). In recent years, and as a consequence of the use of computer systems for mass data processing, a new paradigm of data-based inference has emerged, which is giving rise to an emerging discipline called Data Science which, despite not being new, only in the last few decades has an effort been made to endow it with solid theoretical foundations and viable implementations, which has placed it in the spotlight of many economic, social and research interests.

Among the phenomena that can be approached with these types of mechanisms, those in which both the description of the elements and the interactions between them are important are especially interesting, especially when there may be different types of interactions or when we have to consider the internal properties that characterize them. The data structures that allow to express this type of phenomena usually are oriented to the description of its elements, to the relations between them, or have a hybrid orientation that allows a similar level of expressiveness for both.

Most methods in Data Science are oriented to work with structures that naturally express the properties of the elements, but usually have certain limitations to express their interactions in a natural way. For example, vectors or tables, with which the most widely used machine learning algorithms usually work, are well suited for describing elements, but not for expressing the complex relationships between them. Classic databases provide mechanisms for expressing complex relationships between elements, but the available query and modification tasks are not optimized to work on features derived from these relationships.

The mathematical structure that seems to better express the interactions between elements of a system is the graph, and recently extensions of these have started to appear, such as property graphs, that allow to express in a natural way all the components that make up a system: its elements and the relationships between them. Based on this conceptual structure graph databases were developed, which allow to store the information in such a way that both the storage process and the subsequent query process are efficient.

1.1 Property Graphs

As previously mentioned, mathematical objects usually used to model system's elements tend to have some linearity in their structure and have good computational implementations (vectors, registers, tables, ...). For those cases where it has been necessary to have less linear structuring tools, ad-hoc implementations have been created that have covered the desired requirements (XML, RDF, ...). In search of a common mathematical object that models this type of data structures, property graphs stands out as a concept that provides a good balance between correctly expressing the description of the elements of the system and that of their relationships. This structure, which has not been formally defined until a few years ago, does not yet have a robust theory to support it, although, as we shall see, a correct definition can make it contain different types of abstract structures that support databases.

Historically, uni-relational graphs (those in which all relationships are of the same type) formed the basis of classical mathematical graph theory. Multi-relational graphs (those in which different relationships may have different types) are closer to the intuitive description of a system. property graphs allow each element (node or relation) to have an indeterminate number of associated heterogeneous properties. With these premises, this structure is able to contain practically any other relational structure and it can be adapted to the level of detail required for the system according to the needs of later analysis.

For these reasons, in this work we will use this type of mathematical objects as a base structure, both in the previous phase of modeling and storage of information and in the later phase of analysis and inference about it.

1.2 Machine Learning on Graphs

Once we have selected a suitable mathematical structure to describe the systems that interest us, we need tools that allow us to make inferences automatically about them, usually to be able to predict them. Traditionally, the various scientific paradigms have made use of tools based on logical, geometric, algebraic, analytical, etc., and more recently they make extensive use of algorithmic tools.

In the same way as in the scientific method, the enunciation of hypotheses after the observation of phenomena allows extracting general laws from experience, the use of computational techniques can be useful to obtain general laws from examples analyzed with algorithmic tools. This process can be done at various levels. We can make use of computational tools as a means to aid the process of inference of the researcher, or we can try to create an algorithm capable to make the complete inference and enunciate laws itself, automatically. In the latter case, we say that we are dealing with a machine learning process (discussed more fully in Chapter 2).

Originally, most of these techniques had been devised to generalize from a series of isolated examples, elements described through a series of predetermined properties, usually expressed in the form of registers or tables. This way of structuring the examples coincides with the first type of structures presented in the previous section, those that allow to correctly express the properties of the elements that make up a system, but not their relations. This limits learning ability as it does not explicitly consider an important part. If in the phenomenon under study (and which is intended to learn from) the interactions seem to be determining in the understanding of the same, we must choose a mathematical structure that reflects them correctly. However, the efforts in the development of machine learning seem to have left these considerations behind.

Designing algorithms that learn from structured data in the form of property allows to learn in a more natural way both on the basis of the properties of the elements of a system and the relationships between them. In addition, although some information is always lost from the real phenomenon, the flexibility of graphs allows a greater adaptation without making too many transformations that would break from reality.

If we allow machine learning algorithms to work with graphs as the natural structure from which to learn, they can manipulate relations explicitly in the same way as they do with the properties of the elements. This type of learning has come to be called relational learning (multi-relational in case there are several types of relationships between data) and luckily, despite not being a focus of attention, has made great strides and has been an active research area for many years. It is usual in literature to find relational learning divided into three blocks: (1) Statistical Relational Learning (SRL), within which developments like the Markov logical networks would be included, that uses a coding of multi-relational graphs making use of probabilistic models; (2) Path Ranking Methods, which explicitly explore the relationship space through random paths; And (3) Immersion Based Models, which obtain a vector representation of the graph through matrix / tensor factorization, Bayesian clustering or neural networks. In addition to these three blocks, we can include the algorithms that perform the discovery of relational patterns by refining a hypothesis through a series of steps, in this last block we could include algorithms like Top-Down Induction of Logical Decision Trees, Multi-Relational Decision Tree Learning or Graph Based Induction Decision Tree, which we will detail in Chapter 4 of this report. As can be observed, advances have been made in relational learning using decision trees, neural networks and probabilistic models, among others, but there is still a long way to go, and there are other algorithms that, even though they have demonstrated great potential in learning from non relational data, have not yet been exploited from this perspective, as is the case of Random Forest.

Usually, according to various criteria, machine learning models are classified in Supervised vs. Unsupervised (by type of learning carried out), Regression, Calsification or Ranking (by the type of expected output), etc. In addition, and with respect to the interpretability of the results by a human, we can classify the models in those that are able to offer an explanation that accompanies and justifies the result they provide (white box models) and those that sacrifice such justification for better efficiency (black box models).

With regard to the white box methods, one of the most representative models is the decision tree, which results in a succession of tests that explain the prediction of each of the examples. With respect to the black box ones, one of the most representative models are artificial neural networks, because although they have proved to be very efficient in classification and regression tasks, they present great difficulties in offering a justification interpretable by a human user.

1.3 Thesis Goals

Given the small number of methodologies that perform relational machine learning, the main objective of this research has been to provide new methods to carry it out, as well as to optimize some of the existing ones. In order to carry out this task, and without naming objectives related to bibliographical or comparative revisions between models and implementations, a series of concrete objectives to be covered are proposed:

1. Define flexible and powerful structures that allow phenomena modeling based on the elements that compose them and the relations established between them. Such structures must be able to express naturally complex properties (continuous or categorical values, vectors, matrices, dictionaries, graphs,...) of the elements, as well as heterogeneous relationships between them that in turn may possess the same level of complex properties. In addition, such structures must allow to model phenomena in which the relationships between the elements do not only occur between pairs, but also between any number of them.

2. **Define tools to build, manipulate and measure such structures**. However powerful and flexible a structure is, it will be of little use if you do not have the right tools to manipulate and study it. These tools should be efficient in their implementation and should cover construction and consulting tasks.

3. **Develop new black box relational machine learning algorithms**. In tasks related to automatic classification and regression black box models can be used, since the goal is not to obtain explanatory models, this characteristic can be sacrificed for better efficiency.

4. **Develop new white box relational machine learning algorithms**. When an explanation about the operation of the systems being analyzed is needed we will look for white box models.

5. **Improve query, analysis and repairing tools for databases**. Some of the queries in databases are computationally expensive, preventing us

from proper analysis in some information systems. In addition, graph databases lack methods that allow to normalize or to repair the data automatically or under the supervision of a human. It is interesting to develop tools that carry out this type of tasks increasing efficiency and offering a new layer of query and standardization that allows to cure the data for a more optimal storage and recovery.

All marked objectives must be developed on a solid formal basis usually based on information theory, learning theory, artificial neural network theory or graph theory. This basis will allow the results obtained to be sufficiently formal so that the contributions made can be easily evaluated. It is also sought that the developed abstract models be easily implemented on real machines to be able to verify experimentally its operation and to offer to the scientific community useful solutions of this type in a short space of time.

1.4 Contributions

The work carried out has meant an incursion into the formalization of graphs and relational machine learning and, as reflected in this report, has allowed to unify and condense different perspectives in these areas. In addition, it has allowed the development of new techniques to carry out this type of tasks using more general formalizations as well as making use of new learning methods that are able to work with property graphs as basic structure from which to learn.

We describe below the contributions that can be found in this work:

1. **Generalized Graph**, simple mathematical structure that generalizes virtually all classical definitions of graph, from uni-relational graph to property hypergraph. In this paper, concepts belonging to graph theory are redefined from this new perspective and a solid, simple and flexible basis is provided to support systems in which both the description of the elements and that of the relations are important.

2. **Extension of the measures** already defined for uni-relational graphs to the generalized graph concept. Because this structure generalizes most of the range of definitions for graphs, this extension allows to make measurements on the different types of graphs existing.

3. **Property Query Graph**, query tool for generalized graphs that allows to evaluate structures based on their content and the elements with which it is related. Query languages such as Cypher and other query tools such

as Selection Graphs are specific cases of this tool. PQG are expressed through a generalized graph, allowing to work with the same structure for both the information source and the query.

4. **PQG-ID3**, relational machine learning algorithm that allows to discover patterns in enriched structures of data, and to construct decision trees to classify subgraphs from a set of classified examples. The relational patterns extracted by this algorithm are expressed through a generalized graph, allowing its easy interpretation by any human / machine.

5. **Methodology for the embedding of generalized graphs in vector spaces**, maintaining the original semantics and allowing the discovery of new information, retrieval of missing information, automatic classification of data and providing improvements in other tasks such as long distance queries.

6. **Implementations** of the tools developed throughout the work, and detailed in the Implementation Appendix.

In addition to the points indicated, no less significant are the following contributions: a first step towards a tool for the normalization of graph structured data through analysis of vector structures obtained from embedding; A first family of refinements that allow automatic manipulation of complex predicates on graphs. In addition, throughout this report, you can find other minor conceptual and technical contributions that have not been named in this section.

1.5 Thesis Structure

The content of this report is described below, describing the various chapters of which it consists and that pretend to cover the research goals marked previously.

In Chapter 2, **Fundamentals**, we introduce the fundamental theories that serve as a transversal axis to all the work done, we present the common structures for the different chapters of this report. We present a framework for graphs that unifies definitions as a uni-relational graph or graph with properties through the concept of generalized graph, redefining concepts belonging to Graph Theory. In addition, an introduction to machine learning is presented and the models that will be used throughout the memory are briefly presented.

In Chapter 3, **Property Graph Pattern Matching**, a study about the evaluation and extraction of information in relational structures is made, presenting a review of the technologies and foundations that make this task possible. In addition, we present the Property Query Graphs (PQG), our proposal to evaluate structures immersed in a property graph, and which works as predicates on subgraphs, so that they are ideal as a basis for discovery tasks. The chapter concludes by showing some concrete examples of PQG.

In Chapter 4, textbf Decision trees for property graphs, we face the problem of automatically constructing subgraph classifying trees in property graphs. It begins by reviewing the operation of decision tree induction algorithms, from those that learn object (described through a set of properties) classifier trees, to those who learn to classify records from a relational database taking into account their relationships. The central part of the chapter presents the PQG-ID3 algorithm, our proposal to build multi-relational decision trees based on PQG. We conclude by showing a collection of examples of trees constructed using this tool and analyzing the resulting semantic patterns.

Chapter 5, **Semantic Embeddings of Property Graphs**, makes a different approach to multi-relational learning, this time making use of neural networks to learn a vectorial encodings of property graphs. This encoding allows to make use of the usual machine learning methods designed to work naturally with objects described vectorially. A review is made of the learning methods that make use of neural networks, and we analyze methods that have been used to obtain encodings of other types of structures. We present a methodology that makes use of neural encoders to carry out property graph embeddings in vector spaces, experiments are carried out on real data to verify that the obtained projection allows to capture properties present in the original graph. In addition an empirical demonstration is carried out, proving that the proposed immersion methodology allows to successfully perform machine classification tasks, information extraction, missing information retrieval and long-distance queries.

Finally, in the chapter Conclusions and Future Work, we present the conclusions obtained from the research process structured in accordance with the chapters of this report, as well as the conclusions obtained globally across the work. In this last chapter we also present possible future lines of work, made possible by this research and that are related to the formalization of graphs, relational machine learning, graph pattern matching and discovery procedures in databases.

Chapter 2

Conclusions and Future Work

As discussed in various sections of this report, and despite its potential, relational machine learning has been in the background in relation to the more standard machine learning, which makes use of non relational information, usually in form of tables and other regular structures. After the work done, we can identify certain reasons that have led to this situation.

On the one hand, we have detected that the scientific community has some inertia in its research, which leads it to prioritize the optimization and modification of existing algorithms over the creation of new ones, or the use of new data structures from which to learn. Certain scientific publications get impact by overcoming well-founded earlier results and this may be a reason why entering a new method is not as fruitful (academically speaking) for the researcher as continuing a methodology that has already proven to be well founded and useful.The creation of completly new models usually does not find the adequate platforms for the presentation of its work or simply requires a greater validation effort.

On the other hand, the most commonly used information systems, in which we store most of the studied phenomena, make use of schemas and systems based on relational databases. As many studies have shown, these classical databases do not show an optimal performance when working with complex relationships, which is one more reason for the impediment of developing methodologies oriented to work well with elements and their relations.

In addition, the greater expressive richness of the more complex information imposes difficulty in making new algorithms and provides, at least in the first approximations, results less striking than the more refined and more traditional methods.

Finally, there is a Rich gets richer process, the more conventional methods are better known by the scientific community and, therefore, are knowm by more researchers, which causes that the majority of them to work withs non relational machine learning methods.

We would like to emphasize that to reason about the formal structures used to store the information concerning the systems we analyze is essential if we want to carry out this type of tasks in an optimal way. Researchers often transform data obtained from a system into one of regular formats, usually vectors or tables, missing facets of important relational information whose expression is not natural in these formats. When analyzing a system through techniques derived from Machine Learning can become as important the data structure used to express the information as the algorithm used. After this research, we consider that there is no proportional effort in the area between optimizing the structures from which to learn and the chosen learning methods. In this paper we have explored the learning capacity from structured data in the form of property graphs.

With respect learning from property graphs, it is possible to emphasize that there are several lines of work that transform the original data towards other structures that algorithms are able to handle of more natural way (because they were created to work specifically with such structures). This is the case of graph embeddings in vector spaces. In our view, these are valid approximations that should continue to be investigated, but other options should be considered, such as working directly with the graph structure, which has been one of the lines of research followed in this work and which has proved to be valid.

At times, efforts to work with data are focused on obtaining automatic predictions that quantitatively improve previous results, usually measured through benchmarks. However, there are methods related to prediction that can provide results (quantitative and qualitative) that are not easily measurable through this type of techniques. In addition, there are other interesting tasks, not related to prediction, that can be carried out with data and that have not received the attention they deserve. For example, analyzes related to the semantic purity of a data set can be interesting to evaluate the structure of the data set and to detect inconsistencies such as overlap between data types or redundancy in the data or schema, as well as improving the efficiency of queries on these datasets.

Another related task, is the discovery of information through white box models. If we analyze the trend in learning, relational or not, we can observe that in recent years many efforts have been invested in black box methods, with explanatory methods remaining in the background. Explanatory algorithms like MRDTL emerged at the end of the 1990s and it seems that their development has stalled in recent years. Undoubtedly, black box models are showing shocking, unexpected results, through their greatest exponent, Deep Learning, but whose interpretation is too diffuse to be understood by a human (although there are many efforts to create tools that could narrow this gap). We consider that this fact can be dangerous, black box methods allow us to predict the evolution of systems to a certain extent,but prevents humans from realizing real learning about how the system works, simply providing a tool that predicts it but does not add additional knowledge to the researcher. In this way they become useful tools for Engineering (which justifies the work done on them), but not for Science. If we want to know the phenomena that surround us and advance in the understanding of our surroundings, we must find ways to understand it. Therefore, we consider that a superior effort should be made in the study and development of white box machine learning methods.

We will now take a more detailed look at some conclusions that can be derived from the different approaches in this work.

2.1 PQG and its use for PQG-ID3

Chapter 3 addressed the goal of obtaining a tool to evaluate subgraphs in property graphs that can be used in discovery procedures in relational information. To achieve this goal, several requirements had to be fulfilled. On the one hand, it was necessary to have a grammar to express the queries in a way close to the structures on which it is going to work. And thanks to the expressive capacity of property graphs, we have presented a query tool that can be expressed naturally by means of a property graph. In addition, it was necessary to provide queries that when used as logical predicates on graphs, they behaved consistently and robustly. In addition, it was necessary, since we will also use them to generate machine learning methods, that the queries could be modified in a controlled way by means of atomic operators that translated the topological control into a logical control. In this sense, a first family of refinements (a refinement acts as a query partition) has been introduced that allow an ordered collection of queries to be constructed from an initial query (which may be empty).

Any relational data structure can be viewed as a graph and any query can be viewed as pattern matching, thus, most query languages in databases can be viewed as Graph Pattern Matching (perhaps primitive) tools in property graphs. In Chapter 3 we have also analyzed some of the existing Graph Pattern Matching tools as well as the feasibility to be used in automatic procedures. One of the tools analyzed, Selection Graph, allows to evaluate registers in relational databases using acyclic patterns that can be refined through basic operations, allowing to obtain complementary patterns in each case. It does not require an exact projection of the pattern representing the selection graph on the subgraph to be evaluated, but rather the fulfillment of a series of predicates expressed through said pattern. It should be noted that if a projection is required when carrying out the verification of a pattern, the task of evaluating the non-existence of certain elements is complicated. Specifically, the selection graphs evaluate the existence / non existence of paths that are incident to the registry under evaluation (they are only able to evaluate individual records). It is verified if a conjunction of predicates on paths that depart from the analyzed registry is fulfilled, which can be seen as the evaluation of the existence of a tree rooted in the node that represents the registry under evaluation.

Property Query Graph, the tool presented in Chapter 3, extends the concept of Selection Graph allowing the evaluation of general subgraphs, beyond a single node, using predicates through the definition of a language on the elements of the graph and allowing cyclical patterns. As it becomes a requirement not to use a projection for the verification of a pattern, these objectives have been achieved by extending the form of evaluation, which can be seen as the evaluation of a tree rooted by each node present in the pattern. Although each node of a PQG evaluates the existence of a node that fulfills the conditions imposed by its predicate and the edges in which it participates, by allowing the edges to be identified with paths in the graph (Regular Pattern Matching) there is the evaluation of one tree per node, not a single star. It is through the intersections that occur between the various trees and the constraints imposed on the nodes as the evaluation of cyclical patterns in PQGs is allowed.

Like the selection graphs, the PQG can be modified and constructed from refinements, but unlike the simple case of selection graphs, refinements are usually not binary, since their application can modify more of one predicate in the pattern, resulting in sets of size 2^k (where k is the number of modified predicates). As shown in Chapter 4, this is not a problem when it comes to building learning models, such as decision trees, since they do not have to be binary. Through the definition of certain operations of simplification and equivalence, the refinements shown can be simplified giving rise to simple tools that allow to express complex queries in graphs.

In general, refinements result in partitions of the structures they evaluate, making them ideal tools for white-box procedures. After carrying out a first (but fully functional) proof-of-concept implementation, it has been experimentally demonstrated that PQGs are viable under mild conditions and meet the stated objectives.

An explicit use of these capabilities is shown in Chapter 4, with the presentation of the algorithm PQG-ID3, which makes use of the Property Query Graphs as test tools for the construction of a decision tree following the foundations of ID3 algorithm. In the results of the experiments carried out, it is shown that PQG-ID3 is able to extract interesting patterns that can be used in complex learning tasks. PQG contained in the leaves can be considered as new attributes discovered by the algorithm. In this way, in addition to constructing a classifier tree, the algorithm is able to discover patterns that characterize different structures in the graph (Graph Pattern Mining) and that can be used as attributes of the structures that classify in later tasks (Feature Extraction).

MRDTL algorithm can be seen as a particular case of the algorithm PQG-ID3 in which only PQG with tree form are allowed (since they use selection graphs) and where it learns only from structures formed by a single node. In this sense, PQG-ID3 is a leap forward in a line of work started years ago and considered open since then. As a curiosity, we have to say that the work done on PQG-ID3 was done in a completely independent way, and it was only when writing this memory that we could relate it to the selection graphs and the MRDTL algorithm.

As we saw throughout the chapter, the main problem presented by multi-relational decision tree construction algorithms is that the hypothesis space is extremely large. To solve this problem several solutions can be proposed. On the one hand (and as an extension to the proposal in MRDTL-2), the frequency of occurrence of certain structures can be analyzed in a statistical way with the purpose of reducing the number of possible refinements to be applied in each case and thus to reduce the cost of the search for the best refinement. All of this prior analysis makes use of the various measures introduced in Chapter 2 (and which extend the simpler frequency measures used in the case of MRDTL-2). On the other hand, you can create more complex refinement families (for example, combining the refinement *add edge* with *adding property to an edge* in a single step) to reduce the number of steps to get complexes PQG. If this last option is carried out properly (unifying the refinements according to the frequency of occurrence of structures in the graph), the algorithm can be brought closer to the solution faster. In both cases, an improvement in efficiency is achieved by sacrificing the possibility of covering a wider hypothesis space (but probably offering alternatives in which the impurity reduction is smaller). In this sense, a minimal set of well-constructed refinements has been offered in this paper, but it should be borne in mind that they are not offered with the intention of being optimal for all learning tasks.

The second major problem with the PQG-ID3 agorithm (and inherited by all algorithms inspired by ID3) is the inability to undo the decisions made during the construction of the tree. In such a way that the options of refinement in a determined step of the algorithm depend on the refinements chosen in previous steps. To solve this problem, it is usual to use some backtracking procedure to undo decisions if they have resulted in a bad result or a use a Beam-Search procedure as used in the GBI algorithm that allows you to make several decisions in parallel and finally select the one that has resulted in a better solution.

Consequently, queries on graph based on PQG allow us to obtain powerful and simple tools suitable for automatic construction and to be used in white-box tasks on multi-relational information with controlled complexity, due in part to good properties related to complementarity and containment of queries. In addition, the combination of the PQG decision tree type through an aggregate model, such as Random Forest, can achieve very good results when performing automatic classification (although this will reduce the interpretive capacity of the models obtained).

2.2 Semantic Embedding

The purpose of this last chapter has been to offer the possibility of performing relational machine learning tasks through more traditional algorithms by making an automatic feature selection. In this way, and in addition to the approach presented in the two previous chapters, we try to analyze what options traditional algorithms offer when we want to not lose the enriched structures of relational information.

If there is an element (a subgraph) that is immersed in a database (a property graph) the task of constructing attributes for learning from the relationships that it presents in the global structure can be very complicated. The approximation presented in this chapter consists of constructing a vector representation of each element in the system from a sampling of the information present in the network. In this way, we avoid, on the one hand, the manual work of selecting the attributes to be taken into account and, on the other hand, we obtain a learning algorithm which feeds from a representation of the elements of a graph obtained from global information.

Compared to other machine learning tasks, there are few jobs that have used neural encoders to perform property graph embeddings in vector spaces. Our methodology has sought to use simple architectures to obtain vector representations that maintain the semantic and topological characteristics of the original graph. In addition, it has been demonstrated experimentally that with the obtained embedding one can obtain semantic connections that do not appear explicitly in the original graph (due to incompleteness in the stored data, or to inconsistencies), or even to optimize queries in databases.

We have verified that the geometric characteristics of the structures formed by the nodes and edges in the new vector space can help to assign missing types or properties to the original graph elements (using measures related to distance, linearity, or clustering, among others), or may even help identify new relationships between elements that are not explicitly present in the original graph. This functionality can be very useful in processes that work with large sets of relational data, where incompleteness of data is a common obstacle.

In addition, as has been observed from the evaluation tests, the performance and accuracy of machine learning tasks on these vector representations can provide information about the semantic structure of the data set itself, and not only about the algorithms in use. For example, the confusion of some nodes / edges in classification tasks can give us information about the need to make an adjustment in the data schema to reflect the semantic characteristics correctly. A detailed report on how different types, properties, and clusters overlap and confuse in the resulting embedding would be useful for making decisions related to the standardization of data schemas, something that almost all current analysis proposals lack.

It is evident that the size of the training set and of the selection window positively influence the application capacity of the resulting embedding, but these influences must be studied deeper, since they can throw keys for the automation of the embedding parameters.

In addition, this chapter has explored how vector structures can be used to retrieve information from property graphs, as shown in Entity Retrieval and Typed Paths experiments. Looking for complex structures in the projected space can be simpler than in the original one. In fact, the use of a second layer of learning models after neural encoding can improve the results of various tasks related to the retrieval of information in semantic graphs. The results show that this is a line of research that is worth considering. Although not enough experiments have been carried out on long-distance queries through the representative vectors in the new space, the results obtained show that the query times can be reduced considerably, sacrificing the optimality. This type of queries are very expensive in the databases, and although graph databases have helped to reduce their computational cost they continue to present great problems of efficiency.

Compared to other approaches in the same direction, this paper presents the novelty of working with more general semantic contexts, and not only with random paths, which assume a linearization of the original graph structure. But these are not the only options to carry out property graph embeddings through neural networks. As will be discussed in 2.3, we can get continuous embeddings of property graphs using neural autoencoders, so that the neuronal encoder will learn the identity function for the elements of the graph, allowing the encoding to work without any bias imposed by any function that relates the elements to their context.

With this work we have given an initial framework to perform machine learning tasks from property graphs in which we take into account information from the complete graph to encode each element. This new representation of property graphs allows to work with relational data stored in almost any system of persistence in a vectorial way, taking advantage of the power that the processors and GPUs currently have to work with this type of structures.

It should be noted that during the review of this document new tools based on the *Word2Vec* architectures have been published, they optimize the process of learning latent semantics from natural language [34]. In spite of the probable improvement that these tools would suppose in our methodology, we have decided not to take them into account since they do not imply a change in the fundamental part of our results, although it would possibly improve the associated computational cost.

2.3 Future Work

In this last section we want to show some of the new lines opened by this research. Some are already being studied, while others represent simple ideas that have emerged during the work and have been targeted to be addressed whenever possible. As far as possible, we will try to maintain the natural order that has been followed in this report.

The power obtained through the definition of a language on the elements of a graph such as the one presented in this work has allowed to construct a discovering tool that generalizes and enhances the standard in multi-relational decision tree construction. However, the restriction that we have imposed that such predicates can only evaluate nodes, edges or paths, implies that the global pattern represents a predicate that is finally a conjunction of relatively simple ones. These types of patterns are constructed through some structure in the form of a graph that unifies said predicates, several trees in the case of PQG-ID3. In a PQG, the predicates that compose it have a node perspective (the PQG is constituted by a predicate for each node it owns), but PQG could be constructed to evaluate another structure, for example, PQG in which assigns a predicate for each cycle of length 3 (triangles perspective). This limitation suggests that the concept of pattern, as hitherto conceived, is not general enough to express powerful and flexible predicates about relational data. A recursive definition of pattern, which could lead to a redefinition of the concept of graph through the recursion of structures by levels, could allow to express a larger set of patterns without losing power, flexibility, or the capacity to be constructed through complementary refinements such as those presented.

In PQGs, the sign of nodes and edges have a clear intuitive interpretation: positive elements are must be found in the graph under evaluation, and negatives impose non-existence constraints. Because adding constraints to a non-existence condition resulting in a less restrictive condition, negative elements have not been amenable to being refined through refinements presented. However, there is no reason not to investigate possible ways of refinement through the negative elements. For this, it would be enough to propose this type of predicates based on disjunctions, in this way the situation would be the inverse and the negative restrictions would be susceptible of being refined. Thus, in order to improve the utility of PQGs and the tools derived from them, a way of working should be in generating families of refinements that expand the expressive capacity of PQGs that can be built automatically.

The advances made in multi-relational decision trees through PQG should be used by the family of ensemble methods and in particular by Random Forest. As we have discussed, the interpretative ability of decision trees is diluted when several trees are combined to explain the same result, but they can greatly expand their predictive capacity. Another option derived from the use of methods combined with decision trees using PQG is that the trees obtained have tests in their leaf nodes that evaluate complete semantic patterns, so an option would be combining them in a probabilistic way to give rise to combined patterns that can be interpreted as probabilistic decision tools, opening up an interesting line in white box relational machine learning.

PQG presented in chapter 3 represent predicates on property subgraphs that are capable of evaluating characteristics beyond the structural and semantic properties of the subgraph under evaluation, since they allow to express restrictions in the surroundings of said subgraph (this surroundings can become all the graph in which it is immersed, if the appropriate predicate is used). This feature, in addition to the ones already discussed, make the PQG into descriptors of relational structures (whether they were built automatically or manually designed by experts in the area), erected as suitable candidates to be used as additional attributes in relational learning tasks. In addition, as already mentioned, the complexity in the PQG-ID3 method can be reduced by using statistical analysis to evaluate the frequency of occurrence of different patterns in the graph and in this way reduce the possible refinements available in each step, or combine several refinements into one.

The efficiency improvements in long-distance queries deserve to be evaluated in greater depth and compared with other similar methods. Some results related to the semantic analysis of property graphs have not been carried out in depth and have not been presented in this report although they are expected to be presented in later works. Options such as sampling the context of the edges, perform a embedding of the same and from this infer an embedding for the nodes have not been taken into account and can offer interesting results. With respect to property graph embeddings in vector spaces through neuronal encoders, it must be taken into account that, having inspired us in the architectures corresponding to *Word2Vec*, the function that this encoder tries to learn relates each node with its context and this therefore determines the distribution of the obtained embedding. In this work it has not been considered that this encoder could learn other functions, but we consider that it is a point to take into account since the function that learns the encoder is determinant in the use of the subsequent embedding. For example, if we use the identity function (in this case the network would be an autoencoder) we could get an aseptic immersion, not determined by any previous criteria. This would allow to avoid the problems derived by the definition of the contexts and in other works related to graph embeddings through random walks. Another option could be to use patterns like PQG to perform the encoding. Given a structure, the function to be learned by the coder will relate it to its associated PQG, so the obtained embedding would reflect the semantics associated with the PQG used, and the resulting representation could be optimal if supervised (classification) or non-supervised (clustering) learning is later related to the structure shown in the PQG. Undoubtedly, the possibilities of mixing the expressiveness provided by patterns such as PQG and the efficiency and performance provided by the vectorial representations shown are broad and promising.

During the conception, implementation and experimentation of this work new research lines have been opened that can be considered to analyze the characteristics of the obtained embeddings.

A first consideration is how to construct the training set that is consumed by the neuronal encoder to obtain the vector representation of a property graph. In the experiments carried out the construction of the training set has been totally random, that is, all nodes have the same probability of being sampled, as well as all their properties and neighbors. This may not be the most appropriate way depending on the type of activity to be performed with the resulting embedding. For example, it may be beneficial to construct the training set so that those nodes with a greater semantic richness are more likely to appear in it, which may contribute to regions that are less likely to be considered.

Another line to take into account is to build a neural network that works with the contexts of an element as input (in one-hot format) and learn to return a particular property of the element as an output, i.e. connect a neuronal classifier directly with the encoder, in order to learn the proper encoding and classification simultaneously. Similarly, it would be interesting to think of neuronal encoders that make use of recurrent neural networks to be able to analyze the behavior of dynamic relational information.

It should also be noted that the possibility of working with continuous properties in nodes and edges is open, this feature is not present in the datasets used, but should be considered to expand the capacity of presented methodology. For both PQG and conttinuous embedding there are direct mechanisms to include the presence of continuous properties, it is still a matter of work to begin by testing these more obvious mechanisms and then to measure the extent to which other approaches can be taken into account.

In summary, presented research has opened numerous lines of work in various connected areas. The most evident have been in the formalization of relational structures, formalization of procedures for constructing queries about them, relational machine learning and relational knowledge discovery, feature extraction, representation learning, and analysis and normalization of data. Some of them have been presented here but, undoubtedly, new challenges will arise in the form of ideas from this work. For this reason, we are pleased to present a thesis in which, despite having meticulously addressed the initial objectives, more questions have been opened that roads have closed. Undoubtedly, this profusion of possible ways of continuity shows that the study of relational information systems can become a fruitful line of research worth paying attention to.

Bibliography

[1] Cypher into patterns. http://neo4j.com/docs/stable/cypher-intro-patterns.html.

[2] Cypher introduction. http://neo4j.com/docs/stable/cypher-introduction.html.

[3] D2r server: Accessing databases with sparql and as linked data. http://d2rq.org/d2r-server.

[4] The sparql2xquery framework. http://www.dblab.ntua.gr/~bikakis/SPARQL2XQuery.html.

[5] Sparqlimplementations. https://www.w3.org/wiki/SparqlImplementations.

[6] Xml 1.0 origin and goals. https://www.w3.org/TR/REC-xml/#sec-origin-goals.

[7] Xml: The angle bracket tax. https://blog.codinghorror.com/xml-the-angle-bracket-tax/.

[8] Xpath - retrieving nodes from an xml document. http://sqlmag.com/xml/xpath151retrieving-nodes-xml-document.

[9] *Fast Graph Pattern Matching*, April 2008.

[10] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

[11] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. Connectionist models and their implications: Readings from cognitive science.

chapter A Learning Algorithm for Boltzmann Machines, pages 285–307. Ablex Publishing Corp., Norwood, NJ, USA, 1988.

[12] Boanerges Aleman-Meza, Christian Halaschek-Wiener, Satya Sanket Sahoo, Amit Sheth, and I. Budak Arpinar. *Template Based Semantic Similarity for Security Applications*, pages 621–622. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[13] Faisal Alkhateeb, Jean-Francois Baget, and Jérôme Euzenat. *RDF with regular expressions*. PhD thesis, INRIA, 2007.

[14] Noga Alon and Raphael Yuster. On a hypergraph matching problem. *Graphs and Combinatorics*, 21(4):377–384, 2005.

[15] Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *J. ACM*, 42(4):844–856, July 1995.

[16] Kartik Anand and Ginestra Bianconi. Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E*, 80(4):045102, 2009.

[17] Renzo Angles. A comparison of current graph database models. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering Workshops*, ICDEW '12, pages 171–177, Washington, DC, USA, 2012. IEEE Computer Society.

[18] Renzo Angles, Pablo Barceló, and Gonzalo Ríos. A practical query language for graph dbs. In *7th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)*, 2103.

[19] Renzo Angles, Arnau Prat-Pérez, David Dominguez-Sal, and Josep-Lluis Larriba-Pey. Benchmarking database systems for social network applications. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 15:1–15:7, New York, NY, USA, 2013. ACM.

[20] T.M. Apostol. *Mathematical Analysis*. 1957.

[21] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection, 2009.

[22] Anna Atramentov, Hector Leiva, and Vasant Honavar. *A Multi-relational Decision Tree Learning Algorithm – Implementation and Experiments*, pages 38–56. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[23] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[24] Richard Barber and Austin Gibbons. Automatic annotation in multirelational information networks. 2011.

[25] Pablo Barceló, Leonid Libkin, and Juan L. Reutter. Querying graph patterns. In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '11, pages 199–210, New York, NY, USA, 2011. ACM.

[26] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[27] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[28] George M. Bergman. *An Invitation to General Algebra and Universal Constructions.* Henry Helson, 15 the Crescent, Berkeley CA, 94708, 1998.

[29] Ginestra Bianconi, Anthony C. C. Coolen, and Conrad J. Perez Vicente. Entropies of complex networks with hierarchically constrained topologies. *Physical Review E*, 78(1):016114+, July 2008.

[30] Bahareh Bina, Oliver Schulte, Branden Crawford, Zhensong Qian, and Yi Xiong. Simple decision forests for multi-relational classification. *Decision Support Systems*, 54(3):1269–1279, 2013.

[31] Hannah Blau, Neil Immerman, and David D. Jensen. A visual language for relational knowledge discovery. Technical Report UM-CS-2002-37, Department of Computer Science, University of Massachusetts, Amherst, MA, 2002.

[32] Hendrik Blockeel and Luc De Raedt. Top-down induction of logical decision trees.

[33] Hendrik Blockeel and Luc De Raedt. Top-down induction of first-order logical decision trees. *Artificial Intelligence*, 101(1):285 – 297, 1998.

[34] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[35] D. Bonchev. *Information Theoretic Indices for Characterization of Chemical Structures.* Chemometrics research studies series. Research Studies Press, 1983.

[36] U. S. R. Bondy, J. A.; Murty. *Graph Theory. Graduate Texts in Mathematics.* Springer, 2008.

[37] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc., 2013.

[38] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer, 2005.

[39] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[40] I. Bratko and M. Bohanec. Trading accuracy for simplicity in decision trees, 1994.

[41] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA, 1984.

[42] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[43] Eric Browne. The myth of Self-Describing XML, 2003.

[44] A.E. Bryson. *Applied Optimal Control: Optimization, Estimation and Control.* Halsted Press book'. Taylor & Francis, 1975.

[45] Horst Bunke, Peter Dickinson, Miro Kraetzl, Michel Neuhaus, and Marc Stettler. *Matching of Hypergraphs — Algorithms, Applications, and Experiments*, pages 131–154. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[46] Yang Cao, Wenfei Fan, Jinpeng Huai, and Ruizhe Huang. Making pattern queries bounded in big graphs. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 161–172, 2015.

[47] Steve Cassidy. Generalizing XPath for directed graphs. In *Proceedings for the Extreme Markup Languages conference*, 2003.

[48] Kai-Wei Chang, Scott Wen-tau Yih, Bishan Yang, and Chris Meek. Typed tensor decomposition of knowledge bases for relation extraction. In

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. ACL – Association for Computational Linguistics, October 2014.

[49] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 119–128, New York, NY, USA, 2015. ACM.

[50] Jie Cheng, Christos Hatzis, Hisashi Hayashi, Mark-A. Krogel, Shinichi Morishita, David Page, and Jun Sese. Kdd cup 2001 report. *SIGKDD Explor. Newsl.*, 3(2):47–64, January 2002.

[51] Yu Cheng and Daniel Suthers. Social network analysis—centrality measures. 2011.

[52] W. J. Christmas and C Fl W. J. Christmas. Structural matching in computer vision using probabilistic reasoning, 1995.

[53] E. F. Codd. Relational database: A practical foundation for productivity. *Commun. ACM*, 25(2):109–117, February 1982.

[54] Thayne Coffman, Seth Greenblatt, and Sherry Marcus. Graph-based technologies for intelligence analysis. *Commun. ACM*, 47(3):45–47, March 2004.

[55] A. Colmerauer and P. Roussel. *La naissance de Prolog*. Editions universitaires europeennes EUE, 2014.

[56] Mariano P. Consens and Alberto O. Mendelzon. Graphlog: A visual formalism for real life recursion. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '90, pages 404–416, New York, NY, USA, 1990. ACM.

[57] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *IJPRAI*, 18(3):265–298, 2004.

[58] Diane J. Cook and Lawrence B. Holder. Substructure discovery using minimum description length and background knowledge. *J. Artif. Int. Res.*, 1(1):231–255, February 1994.

[59] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27, September 2006.

[60] Quinlan Quinlan Cs and J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

[61] Balázs Csanád Csáji. Approximation with artificial neural networks. *MSc Thesis, Eötvös Loránd University (ELTE), Budapest, Hungary*, 2001.

[62] Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph Gonzalez, and Matei Zaharia. Graphframes: an integrated API for mixing graph and relational queries. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems, Redwood Shores, CA, USA, June 24 - 24, 2016*, page 2, 2016.

[63] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3:041022, Dec 2013.

[64] H. Decker, L. Lhotská, S. Link, M. Spies, and R.R. Wagner. *Database and Expert Systems Applications: 25th International Conference, DEXA 2014, Munich, Germany, September 1-4, 2014. Proceedings.* Number parte 1 in Lecture Notes in Computer Science. Springer International Publishing, 2014.

[65] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, abs/1606.09375, 2016.

[66] Matthias Dehmer. Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1–2):82 – 94, 2008.

[67] Matthias Dehmer. A novel method for measuring the structural information content of networks. *Cybern. Syst.*, 39(8):825–842, November 2008.

[68] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM.

[69] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes,

N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232. Curran Associates, Inc., 2015.

[70] Sašo Džeroski. Multi-relational data mining: An introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16, July 2003.

[71] Frank Emmert-Streib and Matthias Dehmer. Information theoretic measures of uhg graphs with low computational complexity. *Applied Mathematics and Computation*, 190(2):1783–1794, 7 2007.

[72] Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959 1959.

[73] Floriana Esposito, Donato Malerba, and Giovanni Semeraro. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):476–491, May 1997.

[74] Wenfei Fan. Graph pattern matching revised for social network analysis. In *Proceedings of the 15th International Conference on Database Theory*, ICDT '12, pages 8–21, New York, NY, USA, 2012. ACM.

[75] Wenfei Fan. Graph pattern matching revised for social network analysis. In *Proceedings of the 15th International Conference on Database Theory*, ICDT '12, pages 8–21, New York, NY, USA, 2012. ACM.

[76] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu. Adding regular expressions to graph reachability and pattern queries. In Serge Abiteboul, Klemens Böhm, Christoph Koch, and Kian-Lee Tan, editors, *ICDE*, pages 39–50. IEEE Computer Society, 2011.

[77] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu, and Yunpeng Wu. Graph pattern matching: From intractable to polynomial time. *Proc. VLDB Endow.*, 3(1-2):264–275, September 2010.

[78] Wenfei Fan, Jianzhong Li, Shuai Ma, Hongzhi Wang, and Yinghui Wu. Graph homomorphism revisited for graph matching. *Proc. VLDB Endow.*, 3(1-2):1161–1172, September 2010.

[79] C. Fellbaum. *WordNet: An Electronic Lexical Database.* Language, speech, and communication. MIT Press, 1998.

[80] Philip Fennell. Extremes of xml. In *XML London 2013*, 2013.

[81] Michael J. Fischer and Richard E. Ladner. Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, 18(2):194 – 211, 1979.

[82] Scott Fortin. The graph isomorphism problem. Technical report, 1996.

[83] Steven Fortune, John Hopcroft, and James Wyllie. The directed subgraph homeomorphism problem. *Theoretical Computer Science*, 10(2):111 – 121, 1980.

[84] Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey Holmes, and Ian H. Witten. Using model trees for classification. *Machine Learning*, 32(1):63–76, 1998.

[85] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, March 1977.

[86] Brian Gallagher. Matching structure and semantics: A survey on graph-based pattern matching. *AAAI FS*, 6:45–53, 2006.

[87] Warodom Geamsakul, Takashi Matsuda, Tetsuya Yoshida, Hiroshi Motoda, and Takashi Washio. *Classifier Construction by Graph-Based Induction for Graph-Structured Data*, pages 52–62. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[88] Warodom Geamsakul, Tetsuya Yoshida, Kouzou Ohara, Hiroshi Motoda, Hideto Yokoi, and Katsuhiko Takabayashi. Constructing a decision tree for graph-structured data and its applications. *Fundam. Inf.*, 66(1-2):131–160, November 2004.

[89] S. B. Gelfand, C. S. Ravishankar, and E. J. Delp. An iterative growing and pruning algorithm for classification tree design. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(2):163–174, 1991.

[90] Xavier Glorot, Antoine Bordes, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. *CoRR*, abs/1301.3485, 2013.

[91] Charles F. Goldfarb. *The SGML Handbook*. Oxford University Press, Inc., New York, NY, USA, 1990.

[92] Charles F. Goldfarb. The roots of sgml - a personal recollection. `http://www.sgmlsource.com/history/roots.htm`, 1996.

[93] Charles F Goldfarb. The roots of sgml: A personal recollection. *Technical communication*, 46(1):75–83, 1999.

[94] Joris Graaumans. *Usability of XML Query Languages*. PhD thesis, Proefschrift Universiteit Utrecht.

[95] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016. cite arxiv:1607.00653Comment: In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[96] Andrey Gubichev and Manuel Then. Graph pattern matching: Do we have to reinvent the wheel? In *Proceedings of Workshop on GRAph Data Management Experiences and Systems*, GRADES'14, pages 8:1–8:7, New York, NY, USA, 2014. ACM.

[97] S. Gupta. *Neo4j Essentials*. Community experience distilled. Packt Publishing, 2015.

[98] J. F. Guo H. Zheng and J. Y Wang. A relational data classification algorithm with user guide.

[99] F. Harary. *Graph Theory*. Addison-Wesley, 1994.

[100] Olaf Hartig. Reconciliation of rdf* and property graphs. *CoRR*, abs/1409.3288, 2014.

[101] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009. Autres impressions : 2011 (corr.), 2013 (7e corr.).

[102] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 517–526, New York, NY, USA, 2002. ACM.

[103] Huahai He and Ambuj K. Singh. Graphs-at-a-time: Query language and access methods for graph databases. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 405–418, New York, NY, USA, 2008. ACM.

[104] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society.

[105] Yi He, Jian-chao Han, and Shao-hua Zeng. *Classification Algorithm based on Improved ID3 in Bank Loan Application*, pages 1124–1130. Springer London, London, 2012.

[106] Ivan Herman and M Scott Marshall. Graphxml—an xml-based graph description format. In *International Symposium on Graph Drawing*, pages 52–62. Springer, 2000.

[107] I. N. Herstein. *Topics in Algebra.* Ginn and Company, 1964.

[108] G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[109] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[110] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[111] Paul W Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 2(2):107–124, 1971.

[112] Florian Holzschuher and René Peinl. Performance of graph query languages: Comparison of cypher, gremlin and native access in neo4j. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 195–204, New York, NY, USA, 2013. ACM.

[113] Jiewen Huang, Kartik Venkatraman, and Daniel J. Abadi. Query optimization of distributed pattern matching. In *ICDE*, 2014.

[114] Shan Shan Huang, Todd Jeffrey Green, and Boon Thau Loo. Datalog and emerging applications: An interactive tutorial. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 1213–1216, New York, NY, USA, 2011. ACM.

[115] Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 373–382, New York, NY, USA, 2014. ACM.

[116] Li Ji, Wang Bing-Hong, Wang Wen-Xu, and Zhou Tao. Network entropy based on topology configuration and its computation to random networks. *Chinese Physics Letters*, 25(11):4177, 2008.

[117] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[118] Ian L Kaplan, Ghaleb M Abdulla, S Terry Brugger, and Scott R Kohn. Implementing graph pattern queries on a relational database. *Lammerce Livermore National Laboratory, Tech. Rep*, 2008.

[119] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.

[120] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[121] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2):119–127, 1980.

[122] Yusuf Kavurucu, Pinar Senkul, and Ismail Hakki Toroslu. Confidence-based concept discovery in multi-relational data mining.

[123] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 381–388. AAAI Press, 2006.

[124] Nikhil S. Ketkar, Lawrence B. Holder, and Diane J. Cook. Subdue: Compression-based frequent pattern discovery in graph data. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 71–76, New York, NY, USA, 2005. ACM.

[125] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[126] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *SCIENCE*, 220(4598):671–680, 1983.

[127] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *CoRR*, abs/1309.7233, 2013.

[128] Arno J. Knobbe. Multi-relational data mining. In *Proceedings of the 2005 Conference on Multi-Relational Data Mining*, pages 1–118, Amsterdam, The Netherlands, The Netherlands, 2005. IOS Press.

[129] Arno J. Knobbe, Arno Siebes, Danil Van Der Wallen, and Syllogic B. V. Multi-relational decision tree induction. In *In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre*, pages 378–383. Springer, 1999.

[130] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[131] Ron Kohavi and Ross Quinlan. Decision tree discovery. In *IN HANDBOOK OF DATA MINING AND KNOWLEDGE DISCOVERY*, pages 267–276. University Press, 1999.

[132] Christian Krause, Daniel Johannsen, Radwan Deeb, Kai-Uwe Sattler, David Knacker, and Anton Niadzelka. *An SQL-Based Query Language and Engine for Graph Pattern Matching*, pages 153–169. Springer International Publishing, Cham, 2016.

[133] Mark-A. Krogel and Stefan Wrobel. *Transformation-Based Learning Using Multirelational Aggregation*, pages 142–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[134] S. Kullback. *Information Theory And Statistics.* Dover Pubns, 1997.

[135] P. Laird. Weighing hypotheses: Incremental learning from noisy data. In *Proc. of the 1993 AAAI Spring Symposium on Training Issues in Incremental Learning*, pages 88–95, Stanford, California, 1993.

[136] Rolf Landauer. Computation: A fundamental physical view. *Physica Scripta*, 35(1):88, 1987.

[137] Ni Lao, Tom Mitchell, and William W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 529–539, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[138] Daniel T. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining.* Wiley-Interscience, 2004.

[139] Héctor Ariel Leiva, Shashi Gadia, and Drena Dobbs. Mrdtl: A multi-relational decision tree learning algorithm. In *Proceedings of the 13th International Conference on Inductive Logic Programming (ILP 2003*, pages 38–56. Springer-Verlag, 2002.

[140] Juan Li. Improved multi-relational decision tree classification algorithm.

[141] Gilles Louppe. *Understanding Random Forests: From Theory to Practice.* PhD thesis, University of Liege, Belgium, 10 2014. arXiv:1407.7502.

[142] L. H. Luan and G. L. Ji. Research on the decision tree classification technology, 2004.

[143] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. Capturing topology in graph pattern matching. *Proc. VLDB Endow.*, 5(4):310–321, December 2011.

[144] J Kent Martin and Daniel S Hirschberg. *The time complexity of decision tree induction.* 1995.

[145] Takashi Matsuda, Hiroshi Motoda, Tetsuya Yoshida, and Takashi Washio. *Knowledge Discovery from Structured Data by Beam-Wise Graph-Based Induction*, pages 255–264. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[146] Warren S. McCulloch and Walter Pitts. Neurocomputing: Foundations of research. chapter A Logical Calculus of the Ideas Immanent in Nervous Activity, pages 15–27. MIT Press, Cambridge, MA, USA, 1988.

[147] William J. McGill. Applications of information theory in experimental psychology*. *Transactions of the New York Academy of Sciences*, 19(4 Series II):343–351, 1957.

[148] Caroline R. McNulty, George F.; Shallon. Inherently nonfinitely based finite algebras. *Universal algebra and lattice theory (Puebla, 1982), Lecture Notes in Math., 1004, Berlin, New York: Springer-Verlag, pp. 206–231, doi:10.1007/BFb0063439, MR 716184*, 1983.

[149] Jim Melton and Alan R Simon. *SQL: 1999: understanding relational language components.* Morgan Kaufmann, 2001.

[150] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[151] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[152] Tomas Mikolov, Jiri Kopecky, Lukas Burget, Ondrej Glembek, and Jan Cernocky. Neural network based language models for highly inflective languages. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4725–4728, Washington, DC, USA, 2009. IEEE Computer Society.

[153] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[154] Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013.

[155] R. Milner. *Communication and Concurrency*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[156] R. Milo and et al. Network motifs: simple building blocks of complex networks, 2002.

[157] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[158] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[159] Abbe Mowshowitz and Matthias Dehmer. Entropy and the complexity of graphs revisited. *Entropy*, 14(3):559–570, 2012.

[160] 2Mr. Kushik K Rana Mr. Brijain R Patel. A survey on decision tree algorithm for classification, 2014.

[161] Sreerama K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Min. Knowl. Discov.*, 2(4):345–389, December 1998.

[162] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281, September 2003.

[163] Lorenzo De Nardo, Francesco Ranzato, and Francesco Tapparo. The subgraph similarity problem. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):748–749, 2009.

[164] Phu Chien Nguyen, Kouzou Ohara, Akira Mogi, Hiroshi Motoda, and Takashi Washio. *Constructing Decision Trees for Graph-Structured Data by Chunkingless Graph-Based Induction*, pages 390–399. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[165] Phu Chien Nguyen, Kouzou Ohara, Hiroshi Motoda, and Takashi Washio. *Cl-GBI: A Novel Approach for Extracting Typical Patterns from Graph-Structured Data*, pages 639–649. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[166] Maximilian Nickel and Volker Tresp. A three-way model for collective learning on multi-relational data.

[167] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing yago: Scalable machine learning for linked data. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 271–280, New York, NY, USA, 2012. ACM.

[168] Vincenzo Nicosia, Regino Criado, Miguel Romance, Giovanni Russo, and Vito Latora. Controlling centrality in complex networks. *arXiv preprint arXiv:1109.4521*, 2011.

[169] Caleb C. Noble and Diane J. Cook. Graph-based anomaly detection. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 631–636, New York, NY, USA, 2003. ACM.

[170] Jae Dong Noh and Heiko Rieger. Random walks on complex networks. *Physical review letters*, 92(11):118701, 2004.

[171] A.B.J. Novikoff. On convergence proofs on perceptrons. 12:615–622, 1962.

[172] SH. OATES-WILLIAMS. Graphs und universal algebras. *Lecture Notes Math. 884 (1981), 351-354.*, 1981.

[173] Cristina Olaru and Louis Wehenkel. A complete fuzzy decision tree technique. *Fuzzy Sets Syst.*, 138(2):221–254, September 2003.

[174] Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, 2013.

[175] Alberto Paccanaro and Geoffrey E. Hinton. Learning distributed representations of concepts using linear relational embedding. *IEEE Trans. on Knowl. and Data Eng.*, 13(2):232–244, March 2001.

[176] Neelamadhab Padhy and Rasmita Panigrahi. Multi relational data mining approaches: A data mining technique. *CoRR*, abs/1211.3871, 2012.

[177] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[178] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM.

[179] Gordon Plotkin. Automatic methods of inductive inference. 1972.

[180] Nataliia Pobiedina, Stefan Rümmele, Sebastian Skritek, and Hannes Werthner. *Benchmarking Database Systems for Graph Pattern Matching*, pages 226–241. Springer International Publishing, Cham, 2014.

[181] Reinhard Pöschel. Graph algebras and graph varieties. *algebra universalis*, 27(4):559–577, 1990.

[182] E. Prisner. *Graph Dynamics*. Chapman & Hall/CRC Research Notes in Mathematics Series. Taylor & Francis, 1995.

[183] J. Punin and Mukkai Krishnamoorthy. XGMML (eXtensible Graph Markup and Modeling Language) 1.0 Draft Specification., 2001.

[184] J. R. Quilan. Machine intelligence 11. chapter Decision Trees and Multi-valued Attributes, pages 305–318. Oxford University Press, Inc., New York, NY, USA, 1988.

[185] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

[186] J. R. Quinlan. Simplifying decision trees. *Int. J. Man-Mach. Stud.*, 27(3):221–234, September 1987.

[187] J. R. Quinlan. Learning logical definitions from relations. *MACHINE LEARNING*, 5:239–266, 1990.

[188] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[189] Luc De Raedt and Sašo Džeroski. First-order jk-clausal theories are pac-learnable. *Artificial Intelligence*, 70(1):375 – 392, 1994.

[190] Ronald C. Read and Derek G. Corneil. The graph isomorphism disease. *J. Graph Theory*, 1(4):339–363, 1977.

[191] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[192] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[193] Juan L. Reutter. *Graph Patterns: Structure, Query Answering and Applications in Schema Mappings and Formal Language Theory.* PhD thesis, The school where the thesis was written, Laboratory for Foundations of Computer Science School of Informatics University of Edinburgh, 2013.

[194] Pedro Ribeiro and Fernando Silva. G-tries: An efficient data structure for discovering network motifs. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1559–1566, New York, NY, USA, 2010. ACM.

[195] Matthew Richardson and Pedro Domingos. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136, February 2006.

[196] Carlos R. Rivero and Hasan M. Jamil. Anatomy of graph matching based on an xquery and RDF implementation. *CoRR*, abs/1311.2342, 2013.

[197] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases.* O'Reilly Media, Inc., 2013.

[198] Marko A. Rodriguez. A multi-relational network to support the scholarly communication process. *CoRR*, abs/cs/0601121, 2006.

[199] Marko A. Rodriguez. The gremlin graph traversal machine and language. *CoRR*, abs/1508.03843, 2015.

[200] Marko A. Rodriguez and Peter Neubauer. The graph traversal pattern. *CoRR*, abs/1004.1001, 2010.

[201] Marko A. Rodriguez and Peter Neubauer. A path algebra for multi-relational graphs. *CoRR*, abs/1011.0390, 2010.

[202] Marko A. Rodriguez and Joshua Shinavier. Exposing multi-relational networks to single-relational network analysis algorithms. *CoRR*, abs/0806.2274, 2008.

[203] L. Rokach and O. Maimon. Top-down induction of decision trees classifiers - a survey. *Trans. Sys. Man Cyber Part C*, 35(4):476–487, November 2005.

[204] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[205] F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Report: Cornell Aeronautical Laboratory. Cornell Aeronautical Laboratory, 1957.

[206] Mohamed Rouane-Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. Relational concept analysis: mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, 67(1):81–108, 2013.

[207] S. Ruggieri. Efficient c4.5. *IEEE Trans. on Knowl. and Data Eng.*, 14(2):438–444, March 2002.

[208] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988.

[209] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[210] S. Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology, 1991.

[211] Marcos Salganicoff, Lyle H. Ungar, and Ruzena Bajcsy. Active learning for vision-based robot grasping. *Machine Learning*, 23(2):251–278, 1996.

[212] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.

[213] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

[214] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.

[215] Toby Segaran, Colin Evans, Jamie Taylor, Segaran Toby, Evans Colin, and Taylor Jamie. *Programming the Semantic Web.* O'Reilly Media, Inc., 1st edition, 2009.

[216] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[217] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[218] L. Shapiro and R. Haralick. Structural descriptions and inexact matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:504–519, 1981.

[219] Dennis Shasha, Jason T. L. Wang, and Rosalba Giugno. Algorithmics and applications of tree and graph searching. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 39–52, New York, NY, USA, 2002. ACM.

[220] Jitesh Shetty and Jafar Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 74–81, New York, NY, USA, 2005. ACM.

[221] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1986.

[222] Ajit P. Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 650–658, New York, NY, USA, 2008. ACM.

[223] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc., 2013.

[224] Ilya Sutskever, Joshua B. Tenenbaum, and Ruslan R Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1821–1828. Curran Associates, Inc., 2009.

[225] Anand Takale. *Constructing Predictive Models to Assess the Importance of Variables in Epidemiological Data Using A Genetic Algorithm System employing Decision Trees.* PhD thesis, UNIVERSITY OF MINNESOTA, 2004.

[226] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, New York, NY, USA, 2015. ACM.

[227] Yuanyuan Tian and Jignesh M. Patel. Tale: A tool for approximate large graph matching. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 963–972, Washington, DC, USA, 2008. IEEE Computer Society.

[228] Hanghang Tong, Christos Faloutsos, Brian Gallagher, and Tina Eliassi-Rad. Fast best-effort pattern matching in large attributed graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 737–746, New York, NY, USA, 2007. ACM.

[229] Godfried T. Toussaint. Bibliography on estimation of misclassification. *IEEE Trans. Information Theory*, 20(4):472–479, 1974.

[230] Vaibhav Tripathy. A comparative study of multi-relational decision tree learning algorithm.

[231] Robert P. Trueblood and John N. Lovett, Jr. *Data Mining and Statistical Analysis Using SQL.* Apress, Berkely, CA, USA, 2001.

[232] J. R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42, January 1976.

[233] S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(5):695–703, September 1988.

[234] Paul E. Utgoff. Incremental induction of decision trees. *Mach. Learn.*, 4(2):161–186, November 1989.

[235] Anneleen Van Assche. A Random Forest Approach to Relational Learning.

[236] Oskar van Rest, Sungpack Hong, Jinha Kim, Xuming Meng, and Hassan Chafi. Pgql: a property graph query language. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems*, page 7. ACM, 2016.

[237] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database: A data provenance perspective. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, pages 42:1–42:6, New York, NY, USA, 2010. ACM.

[238] Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1063–1064, New York, NY, USA, 2012. ACM.

[239] Radim Řehůřek. *Scalability of Semantic Analysis in Natural Language Processing*. PhD thesis, Masaryk University, May 2011.

[240] Hongzhi Wang and Jianzhong Li. Gxquery: Extending xquery for querying graph-structured XML data. *CIT*, 19(2):83–91, 2011.

[241] Huazheng Wang, Bin Gao, Jiang Bian, Fei Tian, and Tie-Yan Liu. Solving verbal comprehension questions in IQ test by knowledge-powered word embedding. *CoRR*, abs/1505.07909, 2015.

[242] Zhiguang Wang and Tim Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. 2015.

[243] Takashi Washio and Hiroshi Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1):59–68, July 2003.

[244] Scott White and Padhraic Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 266–275, New York, NY, USA, 2003. ACM.

[245] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *Trans. Evol. Comp*, 1(1):67–82, April 1997.

[246] David Wood, Markus Lanthaler, and Richard Cyganiak. RDF 1.1 concepts and abstract syntax, February 2014.

[247] Peter T. Wood. Query languages for graph databases. *SIGMOD Rec.*, 41(1):50–60, April 2012.

[248] Huayu Wu, Tok Wang Ling, Gillian Dobbie, Zhifeng Bao, and Liang Xu. Reducing graph matching to tree matching for XML queries with ID references. In *Database and Expert Systems Applications, 21th International Conference,DEXA 2010, Bilbao, Spain, August 30 - September 3, 2010, Proceedings, Part II*, pages 391–406, 2010.

[249] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *CoRR*, abs/1412.6575, 2014.

[250] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Learning multi-relational semantics using neural-embedding models. *CoRR*, abs/1411.4072, 2014.

[251] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Crossmine: Efficient classification across multiple database relations. In *Proceedings of the 2004 European Conference on Constraint-Based Mining and Inductive Databases*, pages 172–195, Berlin, Heidelberg, 2005. Springer-Verlag.

[252] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. *CrossMine: Efficient Classification Across Multiple Database Relations*, pages 172–195. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[253] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Efficient classification across multiple database relations: A crossmine approach. *IEEE Trans. on Knowl. and Data Eng.*, 18(6):770–783, June 2006.

[254] Kenichi Yoshida, Hiroshi Motoda, and Nitin Indurkhya. Graph-based induction as a unified learning framework. *Applied Intelligence*, 4(3):297–316, 1994.

[255] R. Zass and A. Shashua. Probabilistic graph and hypergraph matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, June 2008.

[256] Wei Zhang. Multi-relational data mining based on higher-order inductive logic programming. *2013 Fourth Global Congress on Intelligent Systems*, 2:453–458, 2009.

[257] H. Zheng. Research on the relational data classification algorithm based on background knowledges.

[258] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms.* CRC press, 2012.

[259] Lei Zou, Lei Chen, and M. Tamer Özsu. Distance-join: Pattern match query in a large graph database. *Proc. VLDB Endow.*, 2(1):886–897, August 2009.