Habilitation à diriger des recherches

université PARIS-SACLAY

AgroParisTech

Integrating Experts' Knowledge in Machine Learning

Habilitation à diriger des recherches de l'Université Paris-Saclay

Thèse présentée et soutenue à Saclay, le 25 juin 2024, par

Cristina MANFREDOTTI

Composition du jury

Fatiha Saïs Professor, Paris Saclay University	Présidente
Philippe Leray Professor, University of Nantes	Rapporteur
Andrea Passerini Professor, university of Trento	Rapporteur
Nathalie Pernelle Professor, Sorbonne University of Paris Nord	Rapportrice
Thomas Guyet Full researcher, Inria Center of Lyon	Examinateur
Nicolas Maudet Professor, Sorbonne University of Paris 6	Examinateur
Alexis Tsoukias CNRS full researcher, Paris Dauphine University	Examinateur
Antoine Cornuejols Professor, AgroParisTech	Tuteur

université PARIS-SACLAY

AgroParisTech

Titre: Integration de la connaissance experte dans l'apprentissage automatique

Mots clés: Connaissance Experte, Réseaux Bayésiennes, Modelés Probabilistes Relationnels, Ontologies, Graphes de Connaissance, Systèmes de Recommandation, Réseaux de Neurones Récurrents.

Résumé: Dans mes travaux, j'intègre les connaissances des experts dans l'apprentissage et l'inférence pour l'apprentissage automatique appliqué à divers domaines d'application. Au cours des dix dernières années à AgroParisTech, j'ai formalisé les connaissances des experts dans le cadre des ontologies. Ces méthodes ont l'avantage de fournir des solutions faciles à interpréter et à expliquer, grâce à l'interaction et à l'aide des experts. Dans ce manuscrit, je présente deux axes de travail.

Je présente des méthodes pour coupler des modèles probabilistes et des ontologies afin de modéliser l'incertitude et la causalité dans le domaine des sciences de la vie. Dans ce domaine, l'acquisition de données est difficile pour différentes raisons : les expériences sont souvent menées avec peu de répétitions, les matériaux utilisés sont souvent coûteux et les connaissances disponibles sont rarement complètes. En particulier, j'ai travaillé sur le raisonnement autour des processus de transformation. J'ai formalisé les connaissances des experts dans le cadre d'une ontologie. J'ai proposé de modéliser un processus de transformation avec un modèle relationnel probabiliste appris à partir de données enrichies par l'ontologie. Cela permet une approche efficace capable de modéliser l'incertitude.

Le deuxième groupe de travaux porte sur les systèmes de recommandation dans le domaine de la nutrition. Mes contributions incluent des méthodes pour trouver des suggestions de substitutions alimentaires acceptables obtenues par l'analyse des données de consommation afin de proposer un régime alimentaire plus sain, des techniques de recommandation basées sur le contexte de consommation et des méthodes pour fournir des recommandations à un groupe d'utilisateurs, qui doit manger ensemble, avec préférences incertaines. De plus, en accord avec mes travaux sur la formalisation des connaissances des experts dans le cadre des ontologies, dans le projet EXERSYS, je propose de définir une ontologie pour formaliser les connaissances des experts dans le domaine de la nutrition et de l'utiliser pour améliorer la tâche de recommandation dans le but de recommander une séquence de menus en tenant compte des préférences de l'utilisateur, des contraintes nutritionnelles et du contexte de consommation.

université PARIS-SACLAY

AgroParisTech

Title: Integrating Experts' Knowledge in Machine Learning

Keywords: Expert's Knowledge, Bayesian Networks, Probabilistic Relational Models, Ontologies, Knowledge Graphs, Recommender Systems, Recurrent Neural Networks.

Abstract: In my work, I integrate experts' knowledge in learning and inference in machine learning applied to various application domains. In the last ten years at AgroParisTech I formalized the expert's knowledge within the ontology framework. These methods have the advantage of providing solutions that are easy to interpret and explain, thanks to the interaction with and the lead of the expert. In the manuscript, I present two main groups of works.

I present methods to pair probabilistic models and ontologies to model uncertainty in life science domains and reason about causality in these domains. In this field, acquiring data is difficult for different reasons: experiments are often conducted with little repetitions, the materials used are often expensive and the knowledge available is seldom complete. In particular, I worked on reasoning about transformation processes. I formalised the experts' knowledge within the ontology framework. I proposed to model a transformation process with a probabilistic relational model learnt from data enriched by the ontology. This allows for an efficient approach that can model uncertainty in transformation processes.

The second group of works deals with recommender systems in the nutrition domain. My contributions include methods to find suggestions of acceptable food swaps obtained by analyzing consumption data in order to propose an healthier diet, recommendation techniques based on the context of consumption and methods for providing joint recommendations to a group of users, that has to eat together, in the case of uncertain preferences. Moreover, in line with my works on the formalisation of the expert's knowledge within the ontology framework, in the EXERSYS project I propose to define an ontology to formalise the experts' knowledge we can access on the nutrition domain and use it to improve the recommendation task with the purpose of recommending a sequence of menus taking into account user's preferences, nutritional constraints and the context of consumption.

Grazie[1]

09 luglio 2024

Ho cominciato a pensare a cosa scrivere in questa pagina da quando ho saputo che potevo scrivere questo manoscritto (per i miei figli e i miei genitori : *il libro*). Pensavo anche di dire tutto quello che sto per scrivere all'orale, una volta terminata la discussione ma non ne ho avuto il coraggio. Ho ringraziato tutti e poi le lacrime stavano cominciando a spuntare e mi si sarebbe rovinato il trucco ...

Stamattina in radio hanno detto che ringraziare, essere riconoscenti, fa stare bene. Ho realizzato che, durante questi due anni in cui mi sono dedicata -ad intermittenza- alla stesura di questo *libro*, ho sempre, allo stesso tempo, pensato a cosa scrivere in questa pagina e, forse, é proprio per questo che sono stata bene ...

Voglio approfittare di questo momento, l'ultimo prima di pubblicare questo manoscritto, per ringraziare tutte le persone che mi hanno, in un modo o nell'altro, sostenuta. Durante questo percorso mi sono resa conto di quante belle persone incontro ogni giorno.

A cominciare dai componenti della giuria, gli unici che ho ringraziato espressamente il 25 giugno. Voglio ringraziarvi per il tempo che mi avete dedicato -quando di tempo non ne abbiamo mai abbastanza-, per le *reviews* fatte, per la discussione che abbiamo avuto il 25 giugno. Ora posso dirlo, come in molti me lo hanno detto prima del 25 giugno : é stata proprio una bella esperienza! Grazie a voi. Come dicevo dopo la discussione, ho scelto ogniuno di voi per delle ragioni scientifiche (ovviamente!) ma anche "sentimentali". Grazie di aver accettato il mio invito!

Grazie in particolare agli "Chefs"che mi hanno sempre sostenuto. Appena ho comunicato che ero stata autorizzata ad iscrivermi per discuter il mio HDR, Antoine mi ha chiesto chi avrei invitato nella giuria, come a dare per scontato che sarei arrivata alla discussione senza intoppi. Julien mi ha subito detto : " Ovvio che ti sostengo! Il laboratorio pagherà i viaggi della giuria e il ristorante". Alla notizia che Julien aveva firmato il documento in cui si impegnava a pagare i viaggi per i membri della giuria, Christophe ha subito domandato cosa si sarebbe mangiato dopo la discussione ...

Grazie ai membri dell'equipe Ekinocs. Stephane che ci é stato fino alla fine (ed ha persino letto un capitolo del *libro*). Liliana, con i suoi consigli sempre appropriati. Vincent, cosa avrei fatto senza di te? E tutti gli altri che con brevi messaggi o con la loro presenza mi hanno sorretta ed accompagnata. Grazie!

Grazie ai membri di MIA Paris Saclay e ai membri del dipartimento MMIP. Il vostro sostegno é stato tangibile nei messaggi di chat, nella vostra presenza in quella sala troppo calda per essere sopportabile, negli apprezzamenti all'idea di Spritz, idea, perché faceva troppo caldo per berne quanto avrei voluto ne beveste. Il mio ringraziamento va a chi, in un modo o nel l'altro, ha contribuito all'organizzazione della giornata del 25 giugno,

[1] I decided to write the acknowledgments in Italian because I think who will read these are most probably people who do not speak English and do not have easy access to translating tools as most of you have :-)

attraverso consigli, richieste più o meno celate o nella scelta della sala. Grazie!

Grazie anche a Saclay ... E da un po che ci penso. Saclay ci ha obbligato ad essere un po più uniti (nella disavventura). Anche se non tanto presente quanto vorrei, mi sento parte di un bel gruppo. Grazie a voi tutti!

Un grazie particolare ai miei studenti, senza il (più o meno) duro (ma sempre apprezzato) lavoro dei quali questo *libro* non si sarebbe potuto scrivere.

Ed, in fine, un grazie particolare a chi ci é sempre stato, senza mai capirci un gran ché. Che continua a sostenermi dicendo che é l'ultima volta ma poi si rende conto che ce ne sarà un'altra e loro ci saranno. Grazie a chi li ha sorretti da casa, capendoci ancora meno e che voleva le foto ancora prima che ci fosse la discussione. Grazie a chi mi ha guardato da casa e non ha evitato di dire che una camicia potevo anche metterla.

Grazie a voi, piccoli, per la vostra pazienza, per essere venuti il 25 alla "cresima" della mamma, per esservi vestiti bene, per avermi portato i fiori, per aver dimostrato tutta la pazienza, la stima, il sostegno, l'amore che sapete dare. E grazie a te, mio copilota, per aver loro insegnato questo amore, questa stima nei miei confronti e per avermi lasciato andare per prima in questo viaggio che ora ti aspetta.

Grazie!

E qualche volta si crea qualcosa di bellissimo
Anche dalla spazzatura e dagli scarti
Perché l'insieme è maggiore della somma delle parti
Insieme siamo più forti di Hulk
Più veloci di Flash
Più magici di un bibbidi bobbidi bu
Più belli di Belle
Insieme più grandi di Olly e di Benji
In uno stadio quando c'è la ola
E siamo il suono di mille voci
Che diventano una voce sola

cit. Lorenzo baglioni: Insieme

Contents

List of Acronyms

Chapter 1

Introduction

Contents

This manuscript constitutes my *Habilitation à diriger des recherches* (HDR) and, while covering my main research contributions after completing my PhD, it reflects my overall vision about what I have been doing, mainly at AgroParisTech, and what I intend to do in the near future. My work deals with methods for integrating experts' knowledge in learning and inference in machine learning applied to various application domains and different classes of statistical learning approaches, spanning from Probabilistic Relational Models (PRMs) to recommender systems. In this chapter I retrace my research activities that are based on my involvement in several collaborations, projects proposals writing and students supervision.

1.1 Research Interests and Context

My research domain is Artificial Intelligence (AI). Since my PhD, obtained in 2009, I always cared about models that were explicable and could be improved by expert's knowledge. My PhD thesis focused on modeling and inference with probabilistic models mainly for video surveillance problems. During the postdoc years (January 2010-July 2014) I focused on the problem of learning these models.

My appointment as *Maîtresse de conférences* at AgroParisTech (from September 2014) marked a shift towards the conceptualisation of expert's knowledge within the ontology framework. These works allowed me to formalise the importance of considering the expert's knowledge for reasoning in complex domains. Moreover, they provided explainable approaches that can guarantee the interpretability and explicability of the models learnt.

Reasoning with Uncertainty taking into account Expert's Knowledge Uncertainty is an important aspect of AI: it arises from different sources, including noise on measurements and the limited amount of available data. Probability theory presents a consistent framework for quantifying and manipulating uncertainties; the Bayesian probability interpretation provides reliable and general methods for modeling uncertainty, allowing revisions based on new observations. In this context, Bayesian Networks (BNs) [Koller & Friedman 2009] make it possible to represent the structure of complex probability models in a simple way allowing efficient calculations. Handling uncertainty is important for reliable, robust and trustworthy intelligent systems. Returning the quantified uncertainty to the user, we provide him with all the information he needs to take informed decisions.

Numerous applications are based on sequential data (such as object tracking or time series forecasting or recommender systems for sequences as playlists). In these domains, BNs have been extended to Dynamic Bayesian Networks (DBNs) [Murphy 2002] in order to provide a suitable framework to represent uncertainty in sequences.

I have dealt with domains that are richly structured, containing a multitude of entities, defined by a set of various characteristics, related to each other in different ways. To model the structure of the domain, BNs have been extended to Relational Bayesian Networks (RBNs) [Jaeger 1997]. Indeed, many domains require to model the behaviors of multiple agents, to understand their roles, the context and to detect anomalies. Examples of these domains can be surveillance systems (where, *e.g.*, one must identify the activity of multiple agents interacting with each other), sales support systems (discount campaigns for a specific type of customer, product recommendations based on previous purchases that went well with the recommendation made, *etc.*), in bio-informatics (relationships between the genetic profiles of patients and their drug responses), or in understanding human activities (relationships between active components, such as joints in body motion analysis). A key feature of many situations is that interactions (or relationships) are dynamic and can change over time. In my PhD thesis, at this purpose, I defined Relational Dynamic Bayesian Networks (RDBNs) that extend RBNs to model dynamic situations [Manfredotti 2009].

These models, while modeling uncertainty, are well suited to expert's knowledge integration that is often expressed as relations or properties between entities. In my work, I showed that the explicit recognition of the relationships between entities in the model, improves the understanding we have of their behaviors and helps predict future trends.

Reasoning with Expert Knowledge Expressed in Knowledge Bases Taking into account experts' knowledge is of crucial importance when reasoning about complex systems: it can drastically improve the reasoning task of an AI system. This is especially true in situations where data is scarce, or it is difficult or expensive to obtain.

In my work, I propose to associate machine learning and knowledge representation

methods: we enrich, or complete, data with knowledge, obtained from experts and formalised in knowledge bases, before using them for learning and reasoning. Once the model is learnt, our system interacts with the experts to better specify his knowledge and improve the model.

These methods have the advantage of providing solutions that are easy to interpret and explain. While deep neural networks have been dominating the research landscape supplying impressive advances in automated prediction, they suffer from a lack of interpretability; most of the times, their behaviors can only be explained *post-hoc* [Lipton 2018]. This can be an obstacle when trying to rely on them for decision-making, especially in situations when decisions have to be fair, transparent and accountable (that is often the case in life sciences applications). Interpretability assures some guarantees on the obtained results. Thanks to the interaction with and the lead of the expert's knowledge, my work aims at models that assure the interpretability and the values of the results and, thus, are suited to those kinds of applications.

In my work, I propose to interact with the expert whose knowledge is formalised in an ontology. This approach is situated within the *human-in-the-loop* idea [Zanzotto 2019, Mosqueira-Rey *et al.* 2023] , that views humans and AI agents working together to solve problems. In my work, the information coming from experts is not just used as initial data, but interaction with human experts are an important part of the process.

The originality of my work lies in the combination of two philosophies: the machine learning approach that favors the statistical analysis to reason on the data and the ontological approach which is based on experts' knowledge modeling to represent the domain. I see this as a step forward to come back to AI as a whole: after a period in which the different research domains have been progressing on their own, this is a step towards the goal of putting all the progresses made together, for a better understanding of the world around us.

1.2 Research Contributions

1.2.1 Overview

Before coming to AgroParisTech, I studied models and algorithms to solve problems in domains where many relationships between different entities were present. In particular I worked on the problem of simultaneously tracking (following) several interacting objects and activity recognition (mainly in video-surveillance systems) using probabilistic inference methods.

The common point between these different works is the need to recognize and understand the environment and the activities that take place: who are the actors, their roles and their states. When the environment is particularly complex, in particular taking into account the interaction between several distinct entities whose states can be correlated, automated reasoning becomes particularly difficult.

During my PhD thesis, I extended DBNs to explicitly model interactions between agents moving in a scene and developed an inference method to track these agents and recognize online what they are doing. I modeled experts' knowledge as variables of a RDBN for tracking moving interacting objects in videos [Cattelani *et al.* 2014]. In this case, experts' knowledge was mainly about relations or interactions between moving objects. I showed that the explicit representation of the interconnected behaviors of the targets can provide good models to capture the key elements of the activities in the scene and these variables revealed particularly important when dealing with the problem of online activity recognition [Manfredotti *et al.* 2011].

Motivated by the need to have an activity model $a\ priori$, I also dealt with the problem of learning models of complex activities (activities involving interactions between several objects) from data. During my postdoc at the University of Regina, I developed an approach (the LEMAIO framework) that learns a RDBN from data in an unsupervised way [Manfredotti *et al.* 2013]; this approach discovers relations between moving objects from data and encapsulates this information in the activity model learnt.

At AgroParisTech, I have focused my expertise on probabilistic models in the field of life sciences. In particular, I worked on reasoning about transformation processes and food recommender systems.

Life science poses a lot of challenging questions, it presents sources of uncertainty and it is a field where interactions, the context and time play an important role. But, in the era of *big data*, one of the main issues raised by most life science applications is the lack of data. Indeed, in this field, acquiring data is difficult for different reasons: experiments are often conducted with little repetitions (*e.g.* if you overcook a cake, you do not repeat exactly the same recipe), the materials used are often expensive and the knowledge available is seldom complete (*e.g.* making users fill out forms -such as to know their eating habits or what they ate- is annoying and often imprecise).

While these applications are rich of uncertainty, that makes probabilistic models the appropriate methods to use, the scarcity of data makes very difficult to learn models that take into account uncertainty and experts' knowledge (interactions and the context). Based on this observations, at AgroParisTech, I started a collaboration with Juliette Dibie to formalise the experts' knowledge within the ontology framework to reason on transformation processes [Manfredotti *et al.* 2015]. I wrote an AgroParisTech project, obtaining the fundings for three internships on these topics. To extend this work, we asked and got a grant to the ABIES doctoral school. The PhD thesis of Melanie Münch, that I co-supervised with Juliette Dibie and Pierre-Henry Wuillemin, focused on the causal discovery task [Münch *et al.* 2019a]. On related topics, in 2015, I submitted an ANR project to the *Programme Jeunes Chercheuses Jeunes Chercheurs (JCJC)* dealing with this problem extended to the use of transfer learning techniques to learn multiple probabilistic models given an ontology. Project that has not been financed.

In the same period, I initiated a collaboration with Nicolas Darcel, Antoine Cornuéjols and Fabien Delaert of Danone Nutricia Research for a recommender system in the nu-

trition domain. In the PhD thesis of Sema Akkoyunlu, we proposed methods to find swapping suggestions from food consumption data [Akkoyunlu *et al.* 2017]. In this case, experts' knowledge is about the context in which the swap can be proposed and it is detected from data.

In line with my works on the formalisation of the expert's knowledge within the ontology framework, to better formalise expert's knowledge in the nutrition domain, in 2020, I started a collaboration with Fatiha Saïs to define an ontology to formalise the experts' knowledge we can access on the nutrition domain and use it to improve the recommendation task. This collaboration became a real consortium in 2022. I coordinated the writing of the EXERSYS project[1] that proposes to define an ontology to formalise the experts' knowledge we can access on the nutrition domain and use it to improve the recommendation task with the purpose of recommending a sequence of menus taking into account user's preferences, nutritional constraints and the context of consumption. This project has been funded by the DATAIA Institute[2] which granted the economic support for an internship and a PhD thesis. On a related topic, I proposed a thesis subject for the European project submitted to the Marie Sknlodowska - Curie Action, SWAPS. In this thesis, in collaboration with Vincent Guigue and Michael Schumacher from the HES school in Valais, Switzerland, I proposed to implement methods to merge knowledge representation techniques and machine learning to provide sequences of recommendations of menus. The SWAPS project has not been accepted.

My PhD work took into account expert's knowledge formalised into variables, but it did not give a central role to the expert. This is what I did at AgroParisTech: I modeled expert's knowledge and incorporated it into the system, I developed approaches and methods around this knowledge so to be more transparent and explicable.

In the following, I briefly present the two research domains I dealt with at AgroParisTech. My contributions will be developed in the following chapters.

1.2.2 Experts' Knowledge and PRMs

When I first arrived at AgroParisTech I met people interested in knowledge representation and ontology and found out that the structure of those were not too different from that of a PRM[3]. We started collaborating to merge the two frameworks to model uncertainty in transformation processes and, in particular, in stabilisation processes.

Modeling Stabilisation Processes A stabilisation process is a transformation process that aims at freezing or drying an organism to keep it in the current state to be

[1] EXERSYS -an EXplainablE Recommander SYStem for the nutrition domain, combining knowledge graphs, ontologies and machine learning- is a project in collaboration with Nicolas Darcel, Stephane Dervaux, Vincent Guigue, Fatiha Saïs, Paolo Viappiani and me.

[2] https://www.dataia.eu/

[3] PRMs are extensions of BNs similar to RDBNs

used in the future. An example can be the process of drying a yeast to be used afterwards for culinary reasons. The uncertainty that characterizes the most a stabilization process comes from the fact that the treated cells are (and must be) living systems and living systems are difficult to predict and control. The analysis of a stabilization process gives heterogeneous observations that comes from different sources.

To deal with the heterogeneity of data, an ontology for stabilization processes, the ontology PO^2 (*Process and Observation Ontology*), has been modeled by collegues at AgroParisTech [Ibanescu *et al.* 2016]. This ontology collects and standardizes experts' knowledge and information from different sources, acquired from different domains and at different scales of the studied product, but cannot cope with uncertainty. In order to take into account the uncertainty that characterizes a transformation process, we proposed to work on a method to align ontologies to PRMs [Manfredotti *et al.* 2015].

PRMs extend BNs with the concept of class linked in a relational schema. A class in a PRM is a BN over a set of internal attributes and a set of attributes of other classes referenced by reference slots. PRMs are defined at the class level and represent generic probabilistic relationships within classes that will be instantiated for each specific situation. The relational schema of a PRM describes a set of classes, associated with attributes and reference slots. PRMs provide the qualitative high-level description of the structure domain (*i.e.* the relational schema) and the quantitative information of the probability distribution [Torti *et al.* 2010]. In [Manfredotti *et al.* 2015], based on the similarity between ontologies and the relational schema of PRMs, we proposed an approach able to align the two structures.

Learning PRMs from Ontologies In the thesis of Mélanie Münch, that was defended in 2020, we extended the idea above. The objective of this thesis was to guide the learning of probabilistic relations with experts' knowledge in domains described by ontologies. To do this, knowledge bases have been coupled with PRMs with the aim of filling statistical learning with experts' knowledge in order to learn a model as close as possible to reality and to analyze it quantitatively (with probabilistic relationships) and qualitatively (with causal discovery) [Münch *et al.* 2019a].

The combination of the two approaches (the Bayesian and the ontological) makes it possible to improve both reasoning under uncertainty and experts' knowledge. Thanks to the PRM learnt from data and the ontology it was possible to reason on transformation processes taking into account uncertainty [Münch *et al.* 2018a]. At the same time, learning PRMs from data enriched by an ontology is easier than learning the model from data alone [Münch *et al.* 2017] and the obtained model is transparent and explicable because learnt from expert's knowledge.

Transformation of Urban Waste To continue this work, a postdoc has been offered to Melanie Münch that I supervised with Patrice Buche. In this collaboration, the work presented in Melanie Münch's thesis has been applied to model the process

of the transformation of urban waste for the production of packaging material. In this context, urban waste (dry leaves, small pieces of wood, ...) are shredded and then mixed with a polymer that makes the final product waterproof and resistant. The challenge is, then, to find the right compromise between the quality of the pre-treatment of the waste and the quantity of polymer used (which is expensive). We used one of the algorithms presented in Mélanie Münch's thesis coupled with the PO^2 ontology to learn a PRM which was, then, used with state-of-the-art inference algorithms to answer this challenge [Münch *et al.* 2021, Münch *et al.* 2022].

In the production of packaging materials from urban waste, several techniques can be used to save the waste and the polymer used. For each technique, experimental data are collected but they are generally not enough to learn a probabilistic model. Putting all the data from different techniques together to be used to learn a probabilistic model, could be a reasonable solution for statistical learning, but the data from different techniques are, unfortunately, often not comparable and therefore cannot be used as such for training the same probabilistic model. During the two internships I co-supervised in 2018 and 2019, we researched transfer learning methods that can be used at this purpose.

Transfer Knowledge in Ontologies During my postdoc at the University of Paris 6[4], we studied an algorithm to learn a DBN from a similar one [Gonzales *et al.* 2015]. Based on these work, in 2018 I initiated a collaboration with Juliette Dibie and Fatiha Saïs to study graph matching techniques to transfer knowledge between different domains represented by the same ontology. An ontology can be seen as a graph that structures the data. Graph matching techniques can be used to find similarities or discrepancies between data expressed in a graph. Therefore graph matching techniques can be used to find relationships between data represented in an ontology. In the two internships I co-supervised in 2018 and 2019, we studied methods to transfer these information from a PRM mapped from an ontology to another to easy the learning.

If we use the same ontology to represent the experimental data acquired with the different techniques for the production of packaging materials, we could use the methods researched in the internships to make the data comparable and use all of them for learning a PRM. This would, thus, make it possible to model the uncertainty of the different techniques used with a unique model and reason about the problem taking into account the data obtained by the different experiments that have been conducted.

1.2.3 Experts' Knowedge in Recommender Systems

At AgroParisTech, I dealt, as well, with the nutritional domain. I initiated a collaboration with Nicolas Darcel, with whom we thought that computer science could have and important impact in data analysis for recommending health nutritional choices. Thanks

[4]Now Sorbonne University.

to his connection with Danone Nutricia Research, we obtained a grant for a PhD thesis that I co-supervised with Nicolas Darcel and Antoine Cornuéjols.

Most chronic diseases are correlated to unhealthy eating habits [Rep 2003]. Public health agencies have created dietary guidelines targeting the general population in order to push people for healthier eating habits: "eat at least 5 fruits or vegetables per day", "limit your consumption of salt". The compliance to these guidelines by the general public is relatively low, although the awareness about healthy diets is rather good [Ivens 2016]. There are different causes that contribute to this: cultural and personal preferences, difficulty of implementation, availability and price of food items [Webb 2015] and so on.

A better strategy might be to give recommendations specifically chosen for a given consumer or for a group of similar individuals, taking into account their preferences and health. For example, we could design a recommendation engine capable of providing a consumer with a weekly menu with the aim of improving the nutritional quality of his diet, while respecting his eating habits in terms of associations and social context. This goal is different from the goal of commonly used e-commerce recommender systems for different reasons: (1) providing a weekly menu means giving sequences of recommendations and not a recommendation on an item; (2) a meal is a complex item, giving a sequence of recommendations can be similar to recommending a *playlist* for a music recommendation engine, but music is nevertheless a simpler item than meals; (3) in an e-commerce system, purchases are made online, we have, therefore, the history of the purchases made on which we can rely to learn the preferences of the users, in our case, obtaining the history of what the user ate is not easy; finally (4) taking into account the user's eating habits means knowing (or learning from data) his habits in relation to different information that are not explicit and often difficult to know without asking directly to the user.

Automatically Learn Food Contexts A first step towards the development of this tool was taken in the thesis of Sema Akkoyunlu, co-supervised by Antoine Cornuéjols, Nicolas Darcel and myself, where we studied the co-occurrences of different food items in daily food consumption data[5]. We developed a tool that finds the food contexts where an item is most often eaten [Akkoyunlu *et al.* 2017]. Once the dietary contexts of a food item have been discovered and given a wish of a user to eat something, it is possible to give recommendations for substitutions of this food item, which are acceptable because they respect the contexts of the desired food [Vandeputte *et al.* 2023].

Food Choices and Group Recommendations We rarely eat alone and satisfying the preferences of several people together is another challenge of food recommender systems. In 2020 and 2021 I co-supervised two internships on making recommendations to a group of people. On this subject I initiated a collaboration with Nicolas Darcel and

[5]We used the data from the INCA survey (étude Individuelle Nationale des Consommations Alimentaires) https://www.anses.fr/fr/glossaire/1205

Paolo Viappiani. In Maéva Caillat's internship, we studied Bayesian methods for group recommendation and interactive preference elicitation. We compared different elicitation strategies and, in simulations, we improved the performance of group recommendation algorithms compared to the state of the art [Caillat *et al.* 2020]. In Youhan Wang's internship, which I co-supervised with Paolo Viappiani, we continued this study and we proposed an approach capable of scaling up based on the Plackett Luce model.

The thesis of Thomas Dheilly, that started in November 2023 and I am co-supervising with Nicolas Darcel, Patrick Taillandier, Sabrina Tessier and Paolo Viappiani, investigates how social information (*e.g.* what people around the user is eating) influences the nutritional choice of the user. It will verify different hypothesis and social choices models with user cases and simulations.

Food Recommender Systems In recent years a lot of work has been done to define an ontology that groups all the ontologies that describe the terms in the nutrition domain together [Dooley *et al.* 2018, Dooley *et al.* 2021]. The goal of the two-months internship of Ayoub Hammal, I co-supervised with Stephane Dervaux and Fatiha Saïs, was to investigate this and other food ontologies with the purpose of enriching the data available for a better use in the recommender system development. During this internship, we defined a new ontology that is currently being improved by another student.

During the internship of Noémie Jacquet, that I co-supervised with Stephane Dervaux, Vincent Guigue, Fathia Saïs and Paolo Viappiani, and was financed by the EXERSYS project, we proposed to use the distance between food items found by the Word2Vec algorithm [Mikolov *et al.* 2013b] to define classes of consumers and provide recommendations based on these classes. We used, then, the ontology defined in Ayoub Hammal internship to "filter" the recommendation based on some experts' rules. Those allowed to identify, for example, a recommended item that is not gluten-free and, for this reason, incompatible with some diet[6]. Thanks to the ontology, our system is also able to explain the recommendation (because it satisfies all the rules) or the exclusion of some aliments from the recommendation.

Sequential aspects of Food Choices Another challenge raised by a food recommender system is linked to the sequential aspect of food choices. In the second part of Noémie Jacquet's internship we investigated the use of Recurrent Neural Networks (RNNs) [Chung *et al.* 2015] to simultaneously model the sequential structure of meals and individual tastes.

The thesis of Alexandre Combeau, part of the EXERSYS project, that started in October 2023 and is supervised by Vincent Guigue, Fatiha Saïs, Paolo Viappiani and

[6]It is obvious at which extents this is important, but it is worthy to notice that it is not possible to detect that from data alone because for most of the foods registered in a standard collection this (and other) information is not registered.

me, will study the sequential aspect of food choices integrating machine learning methods and ontologies. We aim at using the ontology not only as a tool to "filter" the recommendation provided by the machine learning algorithm but also to integrate the expert's knowledge into the system to provide an "informed" (or already filtered) recommendation.

Consumption Sequence Generation One of the problems we encountered during the analysis of food consumption data is the insufficient length of the consumption sequences at our disposal for machine learning analysis. From the classes found in Noémie Jacquet's internship, we could learn a probabilistic model that could be used to simulate consumption sequences.

Given that the model will have been learned from a small data-set, to allow the simulated sequences to be closer to the truth, the simulation could be done interactively with the user: instead of simulating a complete (long) sequence, we can simulate the next meal, given the meals already recorded, and ask the user what he thinks of it. To better diversify the sequence, one could use sampling techniques like the particle filtering algorithm that I proposed in my thesis. A last aspect that should be considered is the context of consumption (we eat fish on Fridays, we drink white wine with fish, we drink beer with friends ...). To do this, we could use a PRM for the interactive simulation so that the relationships between variables and, therefore, the consumption context will be taken into account.

1.3 Manuscript Plan

In this manuscript I am going to retrace my research activities by providing the motivations behind the different research questions, positioning my works in the context of the discipline, highlighting the significance and the novelty of my work, and outlining future directions.

In the last ten years at AgroParisTech I formalized the expert's knowledge within the ontology framework. This allows for an explicable approach that puts the expert at the center of the process interacting with him for the construction of an explicable model.

I hope to convey, with this HDR thesis, the multitude facets of my research activities and my increased gain of experience and maturity, acquired in the related activities of student supervision, research collaboration and project proposal writing. Indeed, I pursued my research by collaborating with several colleagues and students; in particular, in order to pursue my aim of linking machine learning with knowledge engineering, I established collaborations with researchers from these two sub-areas of AI. My research activities often involve students: I co-supervised bachelors, M2 interns, PhD students and postdocs. In order to support these activities, I wrote several grants and research project proposals; I have dealt with different application domains and different (sometimes very

specific) data-settings.

In future years, I intend to continue working at the interface of machine learning and knowledge representation methods. I intend to develop more the ongoing EXERSYS project to take into account sequences of meals in order to, for example, make recommendations of weakly menus; I also plan to study methods for groups recommendation taking into account user's preferences and the context of consumption. This will be a natural follow up of the thesis of Alexandre Combeau and Thomas Dheilly to obtain a recommender system able to provide suggestions that are understandable by the user and that take into account preferences, past consumptions and the social context of the consumption.

Interaction between machine learning and knowledge representation methods is intrinsically linked with the interaction with the experts and this cannot be disjointed from taking into account the human interface; for this purpose I intend to initiate a collaboration with experts in information visualisation. I intend, as well, to initiate new collaborations at AgroParisTech, at the University of Paris-Saclay and with industries.

I also maintain a particular attachment with the research topics that I have addressed before coming to AgroParisTech concerning tracking, anomaly detection and video analysis. A possible relevant topic is the domain of following the growth of crops from drone's images (the expert knows if a particular speciem influences the growth of another, and this information can be taken into account to better follow the different speciems); with Jean-Marc Gilliot, I submitted a project on this topic at AgroParisTech. A second possible direction concerning tracking is the detection of anomalies in time series taking into account expert's knowledge; a topic that is of interest in real applications.

To present the two major research domains I dealt with in the last ten years, the manuscript is organised as follows.

Chapter 2 mainly presents the works done in relation with the thesis of Melanie Münch, starting from the first idea of mapping ontologies and PRMs to the aligning we did of PO^2 and a PRM.

Chapter 3 merges together all the research I have done on food recommendation. I present the concept of context, the context of a food item as the food items eaten with it (Sema Akkoyunlu's work) and of context of consumption as the companies we eat with (the works done in the two internships). Moreover, I introduce the EXERSYS project and the idea of modeling the context -and every possible definition of it- with an ontology.

Chapter 4 concludes the manuscript, giving some final remarks and perspectives.

Chapter 2

Experts' Knowledge and PRMs

Contents

In this chapter, I present how we proposed to map an ontology representing experts' knowledge about transformation processes to probabilistic relational models (PRMs). Motivated by the necessity of reasoning about transformation experiments and their results, I proposed methods that use data enriched with expert's knowledge formalised within the ontology framework to learn probabilistic models. I show how this approach allows to deal with the problems of (1) modeling a transformation process, (2) reasoning with the uncertainty present on it, (3) causal discovery and (4) parameters control.

These works well correspond to the two themes that characterise my research. Modeling a transformation process with a PRM allows to consider uncertainty in a richly structured domain which complexity can be considered only thanks to expert's knowledge. Moreover, in our works, causal discovery and parameters control are dealt thanks to the interaction with the expert. This underlines the importance of bringing the human in the loop of the process, as discussed in the Introduction. In this chapter, after giving some background in section 2.1, I show different approaches that are motivated by different goals and in which the expert takes different roles.

- In the first approach, presented in section 2.3, we proposed to align ontologies and PRMs to model transformation processes. Experts' knowledge is organised in an ontology that is used to deduce the structure (the relational schema) of a PRM. In this case the role of the expert is that of providing a structured dataset (structured by the ontology that has been constructed thanks to the experts' knowledge) and the expert does not intervene in the process.

- In section 2.4, I present our approach dealing with causal discovery, where the expert has an active role. In fact, the approach we presented requires the expert to organise attributes following a cause-effect relation order and, when a cause-effect relationship is learned, asks him to verify it and, eventually, re-organise the attributes to obtain a different relationship closer to the knowledge he has.

- In section 2.5, I present how we addressed the problem of parameters control, where the expert, at first, is asked to organise the attributes according to temporal and causal constraints; he can, afterwards, modify this ordering if he is not satisfied by the obtained result.

- In section 2.6, I present the POND framework that is a unifying approach that uses ontology axioms and properties to do inference on the domain and, when this is not enough, it uses an approach that aligns an ontology to a PRM to do inference taking into account uncertainty in the domain. Interaction with the expert is required for aligning the ontology and the PRM.

- Finally, in section 2.7, I show possible extensions of these works to the reasoning on other experimental settings that could involve, as well, the expert in the process.

The works I present in this chapter are the results of a collaboration I initiated when I arrived at AgroParisTech with Juliette Dibie, Cedric Baudrit, and Pierre-Henri Wuillemin, the PhD thesis of Mélanie Münch that I co-supervised with Juliette Dibie and Pierre-Henri Wuillemin, her postdoc's work that I supervised with Patrice Buche and different Master internships.

2.1 Background

2.1.1 Related Works

Different works in the literature map ontologies into BNs. In [Devitt *et al.* 2006] and in [Fenz 2012], for instance, two different approaches are presented that build BNs starting from a knowledge base modelled as an ontology. These approaches take advantage of the information provided by the ontology simplifying the BN learning. One of the biggest issues of these approaches is that, while learning a BN, the authors flatten the information coming from the ontology loosing its relational aspect. This is one of the reasons why we used PRMs.

The method proposed in [Truong *et al.* 2005] brings together ontologies and PRMs, merging them in a new model on which different types of reasoning are supported. To implement Bayesian reasoning on this model, a BN is constructed from the unified model. In this way, as in the works above, the reasoning is done on a BN and not on the PRM.

In [Ishak *et al.* 2011] an approach for learning probabilistic graphical models from an ontology is presented. Their approach learns object-oriented BNs by morphing a given ontology. Object-oriented BNs are an extension of BNs using the object-oriented paradigm that determine a set of "interface" nodes which allow the communication between objects but they are less generic than PRMs and, for this reason, less suitable (because less similar) to ontology morphing than the latter. With the aim of maintaining the structural and relational information expressed in the ontology, in [Manfredotti *et al.* 2015], we presented a mapping of an ontology of transformation processes into PRMs.

2.1.2 Transformation Processes

A transformation process is a dynamic process composed of a sequence of operations which allows inputs to be transformed in several different outputs. Cooking recipes and yeast stabilisation processes are examples of transformation processes. A transformation processes relies on data and knowledge coming from heterogeneous sources and presents several interesting characteristics:

- it is *complex*, multiple operations can occur at the same time and are linked together; inputs and outputs are characterized at multiple scales (*i.e.* environment,

population, cellular and molecular) and studied with different types of measurements (*e.g.* physiological, biochemical, genetic);

- data is *scarce*, due to the difficulty to obtain results, this imposes to gather information from various sources;
- it presents problems of *missing data* (*e.g.* when a parameter is not controlled) and *missing values* (*e.g.* when the process' instructions are not precise);
- even with complete information, it is characterized by *uncertainty*, instruments used to take measurements during a process are able to return only an estimation of the quantity observed because their calibration cannot be entirely defined and repeated from an experiment to another and some internal and uncontrollable parameter (from both devices or outside the experiment) can influence the final result.

Reasoning on a transformation process supposes to be able, for instance, to predict future outputs given certain inputs or given that some inputs are missing, to diagnose how to obtain the best output by determining the important inputs, to control the process and to suggest the best sequence of operations. To do that, it is necessary to face two main locks: (1) data and knowledge heterogeneity and (2) uncertainty quantification.

In order to face the first lock, a relevant solution is to use ontologies, as for example in [Fridman Noy 2004] or [Doan *et al.* 2012]. Many works propose solutions to manage uncertainty in ontologies such as adapting the querying process using fuzzy sets [Buche *et al.* 2005], reasoning using a possibilistic and probabilistic description logic reasoner [Qi *et al.* 2010, Lukasiewicz & Straccia 2008], reasoning in fuzzy ontologies [Bobillo *et al.* 2013] or using existing knowledge bases to predict unfilled information [Saïs & Thomopoulos 2014]. Other languages model uncertainty in ontologies. Extensions of the Ontology Web Language (OWL) to model uncertainty in semantic web are, *e.g.*, BayesOWL [Pan *et al.* 2005], OntoBayes [Yang & Calmet 2005] and PROWL [da Costa *et al.* 2008, Carvalho *et al.* 2013]. PROWL provides a method to write ontologies containing probabilistic information, this information can be processed but it cannot be enriched as in the case of learning or updating from new data. BayesOWL and OntoBayes add to the ontology a BN that models the uncertainty on the domain, providing a pair ontology-BN. In [Helsper & van der Gaag 2002] BNs are built to integrate knowledge expressed by experts in an ontology. The BNs built with these approaches cannot summarize the information contained in the ontology because BNs cannot represent relational information for this reason, the two models need to be paired.

We proposed to quantify uncertainty in reasoning with probability theory and in particular within the Bayesian framework. We modelled transformation processes with Probabilistic Relational Models (PRM).

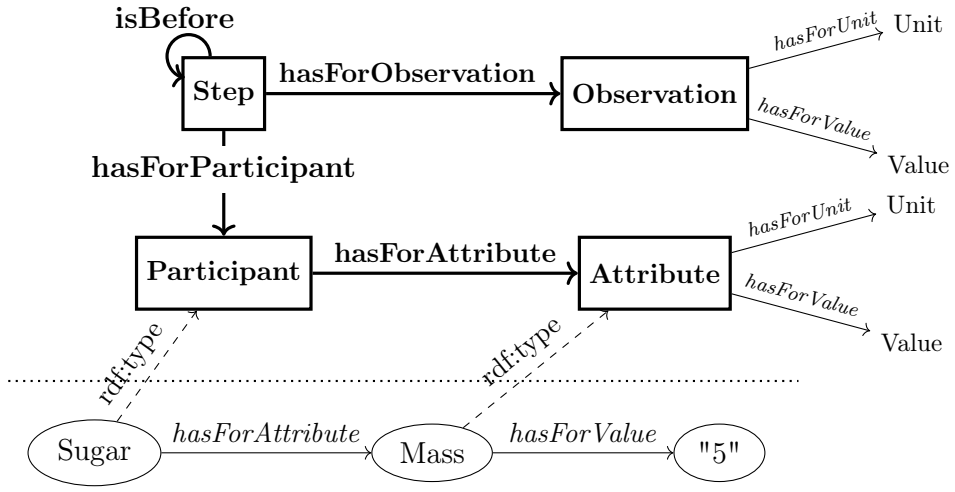


Figure 2.1: Excerpt of a knowledge base about transformation processes

2.1.3 An Ontology for Transformation Processes: PO^2

Ontologies represent the knowledge on a domain with classes, relations between these classes and instances of these classes. They are used as a common and standardized vocabulary for representing a domain. A clear definition of ontologies can be found in [Guarino *et al.* 2009] or in [Staab & Studer 2009].

Data (or observations) at hand (*e.g.*, coming from some experiments) can be collected in a knowledge base that organises data according to the structure defined by an ontology. In the works presented in this manuscript, ontologies are defined by the OWL knowledge representation language[1]. A knowledge base $\mathcal{KB}$ is defined by a couple ($\mathcal{O}$, $\mathcal{F}$) where:

- the ontology $\mathcal{O} = (\mathscr{C}, DP, OP, A)$ is defined in OWL by a set of classes $\mathscr{C}$, a set of *owl:DataTypeProperty* DP in $\mathscr{C} \times T_D$ with T_D being a set of primitive datatypes (*e.g.* integer, string), a set of *owl:ObjectProperty* OP in $\mathscr{C} \times \mathscr{C}$, and a set of axioms A (*e.g.* subsumption, property's domains and ranges).

- the knowledge graph $\mathcal{F}$ is a collection of triples (s, p, o) in the standard Resource Description Framework (RDF)[2], called instances, where s is the subject of the triple, p is a property that belongs to $DP \cup OP$ and o is the object of the triple; for a triple (s, p, o), we note $domain(p) = s$ and $range(p) = o$.

Most of the works presented in the rest of this chapter rely on the PO^2 ontology. The Process and Observation Ontology (PO^2) [Ibanescu *et al.* 2016] has been designed

[1] https://www.w3.org/OWL/

[2] RDF is a standard model for data on the Web. It has features that facilitate data merging and it specifically supports the evolution of schemas over time without requiring all the data to be changed. https://www.w3.org/RDF/

to represent transformation processes. In PO^2, a transformation process is denoted as a sequence of steps (*i.e.* operations), with different participants (*i.e.* inputs) and designed to obtain a specific product (*i.e.* output). Fig. 2.1 gives an excerpt of the PO^2 ontology[3] (on the top) associated with a small example of a knowledge graph (in the bottom). PO^2 is composed of four main classes: the *step* class, that defines the different steps of a transformation process and how they are linked together in time; the *participant* class, that defines the different objects used during a step; the *observation* class, that defines the observations made on the participants and the class *attribute* that characterizes the participants[4]. Using the previous notations, referring to Fig. 2.1, *Step* $\in \mathscr{C}$, *Unit* $\in T_D$, *hasForParticipant* $\in OP$, *hasForValue* $\in DP$.

In this ontology, a sequence of different *steps* linked to each other defines an *itinerary*: each *step* is associated to the one(s) following it according to a chronological order and a dependency relation. A set of itineraries that share the same goal is called *process*. A *step* is defined both by its duration and its participants.

Participants in a *step* can be *mixture*s, *method*s or *device*s. Participants are characterized by inner attributes defined by experimental conditions; a *mixture* is composed of different *products* that represent its composition. *Method*s, *mixture*s and *device*s are subclasses of the ontology class *participant*.

During each *step*, one or more *observation*s can take place to make measurements of one *participant*: they are made using specific *participant*s (independently of the other step's *participant*s) and at a specific *scale*. They have for result a *sensor output* and/or a *computed result*, each of them can have for value a function or a simple measure. A *measure* is characterized by either a quantity and a unit of measure or a symbolic class.

Each *step* is defined as a class to which a set of descriptor classes is linked: participants (*i.e.* devices, mixtures and methods) are classes whose parameters are set *a priori*; observations are classes whose parameters are measured during the step. Therefore there exists for each step a compartmentalization between the different domain's objects. Moreover, the *time relation* linking steps gives information about their relative time (inside the process and with other steps). The instance component of PO^2 allows one to represent different transformation processes by a succession of instances of steps and instances of their associated descriptors.

In this manuscript, I present the results of the algorithms and methods developed on different data collected in different projects that make use of the ontology PO^2. For each project, the ontology PO^2 has been specialised in a *domain ontology*. It is important to notice that, while data were organised by the same (core) ontology PO^2, for each project, we have results on different data-sets and, for this reason, we have different knowledge graphs.

[3] http://agroportal.lirmm.fr/ontologies/PO2

[4] In this manuscript, we refer to the version 1.5 of PO^2.

2.1.4 Probabilistic Relational Models (PRMs)

A BN [Koller & Friedman 2009] is the representation of a joint probability over a set of random variables that uses a Directed Acyclic Graph (DAG) to encode probabilistic relations between variables (Figure 2.2(a)). However, in the case of numerous random variables with repetitive patterns (for instance different steps in the same transformation process), it cannot efficiently represent every probabilistic link.

PRMs extend BNs with the notion of class of relational databases. They extend the BN representation with a relational structure (the *relational schema*) between (potentially repeated) fragments of BN called *classes* [Torti *et al.* 2010]. A class is defined as a DAG over a set of inner attributes and a set of outer attributes from other classes referenced by so-called *reference slots* (Figure 2.2(b)). A *slot chain* is defined as a sequence of reference slots that allows to put in relation attributes of objects that are indirectly related.

The analysis of the BNs in Fig. 2.2(a) reveals two recurrent patterns, that can be translated into two interconnected classes $\mathscr{E}$ and $\mathscr{F}$, as presented in Fig. 2.2(b). In a PRM, the *relational schema* describes a set of classes C, associated with attributes $A(C)$ and reference slots $R(C)$[5].

The probabilistic models are defined at the class level over the set of inner attributes, conditionally to the set of outer attributes and represent generic probabilistic relations inside the classes. This is the *relational model* of the PRM (see Fig. 2.2(c)).

Classes can be *instantiated* for each specific situation (see Fig. 2.2 (d)). A *system* in a PRM provides a probability distribution over a set of instances of a relational schema [Wuillemin & Torti 2012]. PRMs define the high-level, qualitative description of the structure of the domain and the quantitative information given by the probability distribution [Friedman *et al.* 1999].

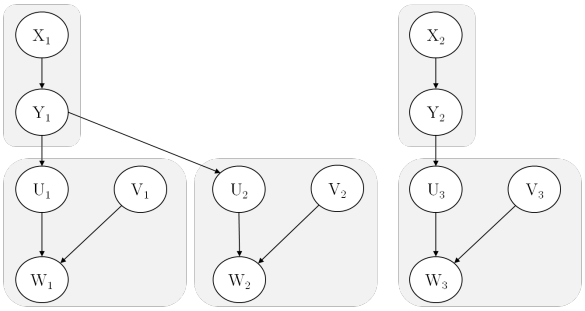
An instantiated system of a PRM is equivalent to a BN. As a consequence, alongside the construction of the PRM, we obtain an Essential Graph (EG) for the PRM.

2.1.5 Essential Graphs (EG)

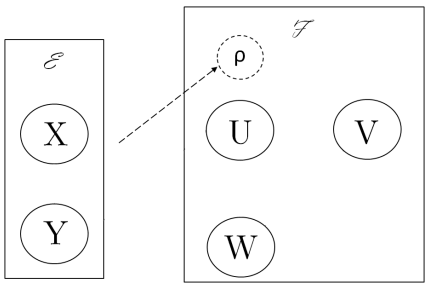
An Essential Graph (EG) [Madigan *et al.* 1996] is a semi-directed graph associated to a BN. They both share the same skeleton (*i.e.*, the set of nodes and not-oriented links between nodes), but the EG's edges' orientation depends on the BN's Markov equivalence class[6]. If an edge's orientation is the same for all the equivalent BNs, it means that its orientation is necessary to keep the underlying probabilistic distribution encoded in the graph: in this case, the edge is also oriented in the EG and it is called

[5] Using the standard object-oriented notation, we will write $C.X$ (respectively $C.\rho$) to refer to a given attribute X (respectively, reference slot ρ) of a class C.

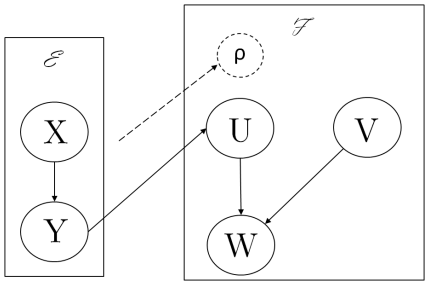
[6] A BN's Markov equivalence class is the set of all BNs that represent the same probabilistic distribution over the same set of random variables. They include the same random variables but can have different edges and edge's orientation.



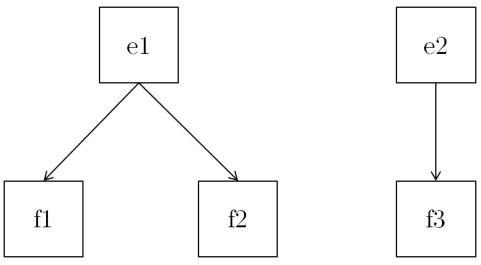
(a) An example of two BNs. The gray areas represent the repetitive patterns, but they are not part of the BN specification.



(b) The **relational schema** of the PRM. It is composed of two connected classes $\mathscr{E}$ and $\mathscr{F}$. ρ is a reference slot in $\mathscr{F}$ which indicates that attributes of $\mathscr{F}$ (U, V, W) can have parents in $\mathscr{E}$ (X, Y).



(c) The PRM *relational model*. Relational links between attributes were added to the relational schema in (b).



(d) A system for the PRM in (c). Instantiation of the classes of the PRM representing the BNs in (a).

Figure 2.2: BNs and PRMs: the analysis of the BN in (a) reveals two recurrent patterns, that can be translated into two interconnected classes $\mathscr{E}$ and $\mathscr{F}$ of a PRM (b) and (c). An equivalent system can, thus, be constructed through the instantiation of twice the class $\mathscr{E}$ and three times the class $\mathscr{F}$ (d).

an *essential arc*. On the contrary, if an edge's orientation is not the same for all the equivalent BNs, it means that its orientation can be both ways without changing the probabilistic distribution, and it is unoriented in the EG. The EG expresses whether the orientation of an arc between two nodes can be reversed without modifying the probabilistic distribution encoded in the graph: whenever the constraint given by an essential arc is violated, conditional independence requirements are changed and the structure of the model itself has to be changed. An example of an EG and two possible interpretations of it are given in Fig. 2.3.

With a BN learned under causal constraints, its EG can give us a new insight. If an arc is oriented, then its orientation has to be kept if we want to conserve all the information we have provided during the learning, this means also the causal information.

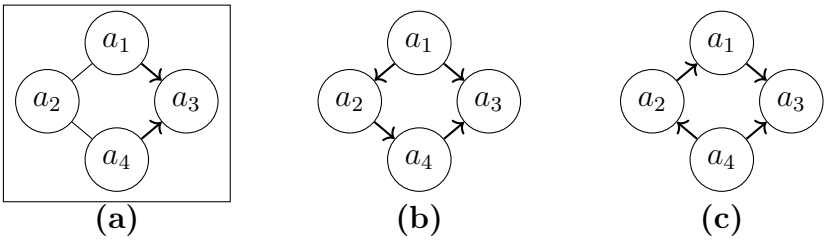


Figure 2.3: Example of an essential graph **(a)** and two BNs **(a)** and **(b)** representing possible interpretations.

2.1.6 Causal Models

Being able to provide explanations about a domain is an hard task that requires the ability to reason about causal knowledge. However, causal discovery from data alone remains a challenging question: previous works have presented the use of interventions, but these require to be able to change certain variables while keeping other constant, which is not always easily doable.

Causal models are DAGs that allow to express causality between their different variables [Pearl 2009]. There are two types of methods for structure learning of causal models from data: independence-based (as for example the PC algorithm described in [Spirtes *et al.* 2000]) and score-based (as for example the Greedy Equivalent Search algorithm described in [Chickering 2003]). Usually independence-based methods give a better outlook on the causality between the attributes by finding the *true* arc orientation, while score-based methods find a structure that maximizes the likelihood considering the data. Finally, other algorithms, such as MIIC [Verny *et al.* 2017], use independence-based algorithms to obtain information considered as partially causal allowing to discover latent variables.

Other works have proposed the use of the EG for learning causal models: for instance [Hauser & Bühlmann 2014], proposes two optimal strategies for suggesting in-

terventions in order to learn causal models with score-based methods and an EG. Integrating knowledge in the learning has also been considered: [Ćutić & Gini 2014] uses ontological causal knowledge to learn a BN and discover new causal relations with its EG; [Ben Messaoud *et al.* 2009] presents a method to iterative causal discovery by integrating knowledge from beforehand designed ontologies to causal BN learning; [Amirkhani *et al.* 2017] proposes two new scores for score-based algorithms using experts knowledge and their reliability and [Besnard *et al.* 2014] presents a tool combining ontological and causal knowledge in order to generate different arguments and counter-arguments in favor of different facts by defining enriched causal knowledge.

In [Münch *et al.* 2018a] and in [Münch *et al.* 2019a] we proposed to learn a PRM based on an ontology and causal constraints defined by an expert. We looked at the EG of the system of the PRM learnt to see if the causal information provided was confirmed to be causal. Our assumption was that, given the PRM learned under causal constraints, the EG of its system should encapsulate causal information as well: in this case, an oriented arc in the EG means that there exist a causal relation between the linked variables and the parent in the link is the cause.

2.1.7 Learning PRMs

The task of learning a PRM is composed of two different parts: *structure selection* and *parameters estimation*. Structure selection can be done in two steps: a first step that organizes the knowledge under an entity-relation pattern using classes and references (this is the *relational schema* learning); and a second step that employs a graphical language to represent the probability distribution in a compact way by exploiting the probabilistic dependencies between the attributes (at this step, the *relational model* is learnt). Due to these multiple steps, the number of free parameters is high and the target model is not unique: selecting one requires to make subjective choices. Moreover, the richness of this tool allows us to represent new and complex systems where data can be scarce or incomplete. This can be another obstacle in learning PRMs.

In [Friedman *et al.* 1999] an algorithm based on an *heuristic search* is proposed to select the legal structure (*i.e.* a structure representing a coherent probability model) with the highest score. The score proposed has a decomposability property that helps to analyze small parts of the structure, easing the search. Other score-based approaches have been equally proposed in [Getoor & Taskar 2007] based on a relational extension of this.

On the contrary of heuristic search, *dependency analysis* tries to discover dependency relations from data itself and then attempts to learn the structure. This constraints guided approach was exploited in [Li & Zhou 2007] that extends to the relational context, or in [Ettouzi *et al.* 2016] that proposes an exact approach to learn PRMs.

In [Manfredotti *et al.* 2015] we proposed to use the knowledge of an ontology to define the relational schema of a PRM, based on this, in [Münch *et al.* 2017], we proposed

a method that learns a PRM from data using the semantic knowledge of an ontology describing these data. Using an ontology helps us to ease the learning in complex domains by integrating the experts' knowledge.

In our works, experts' knowledge is often included in the learning as constraints. Different related works showed that using constraints while learning BNs brings more efficient and accurate results for parameters [de Campos & Ji 2008] or structure learning [De Campos *et al.* 2009]. With experiments, we demonstrated that this is true also for PRMs.

2.2 Mapping Ontology with PRMs

In the following, I present the work presented in [Manfredotti *et al.* 2015], where we described a general approach to deduce relational schema from a given ontology of transformation processes. For this first work, we did not use the PO^2 ontology but we defined a simple ontology for the cooking domain arguing that this is an easily understandable domain that exemplifies the more general and complex domain of transformation processes. The defined ontology is simpler than PO^2 but very similar to it and specific to the cooking domain. In the following, I introduce this ontology that extends the Suggested Upper Merged Ontology[7] (SUMO) [Niles & Pease 2001] and the method, we introduced, to map this to a PRM's relational schema.

We choose the upper level ontology SUMO because it separates physical from abstract entities and gives a definition of object, separated from the definition of process; properties that seamed appropriate for the cooking domain and that we can find in PO^2 as well. The fact that the defined ontology refers to an upper level ontology, guarantees its interoperability because an upper level ontology is general and largely used.

2.2.1 An Ontology for Recipes

In [Despres 2014] an ontology of numeric cooking is presented. We kept four of the classes introduced in this work: ingrédient that we called *product*, matériel called *device* (using the SUMO concept's name), technique de base called *operation* and étapes de réalisation, *realization step*. To these, we added two classes, the class *attribute* already defined in the SUMO ontology, and the class *observation* that records the values assumed by an attribute during the process. Figure 2.4 presents the general schema of relations between these classes that are detailed below.

We defined a recipe as a sequence of *realisation steps*. Each realisation step is composed of one or more *operation(s)* applied either to one or more *product(s)* using one or more *device(s)* or to a *device* in order to change some of its *properties*. The product output(s) of one operation can be the input of another following it in the

[7] http://www.ontologyportal.org/

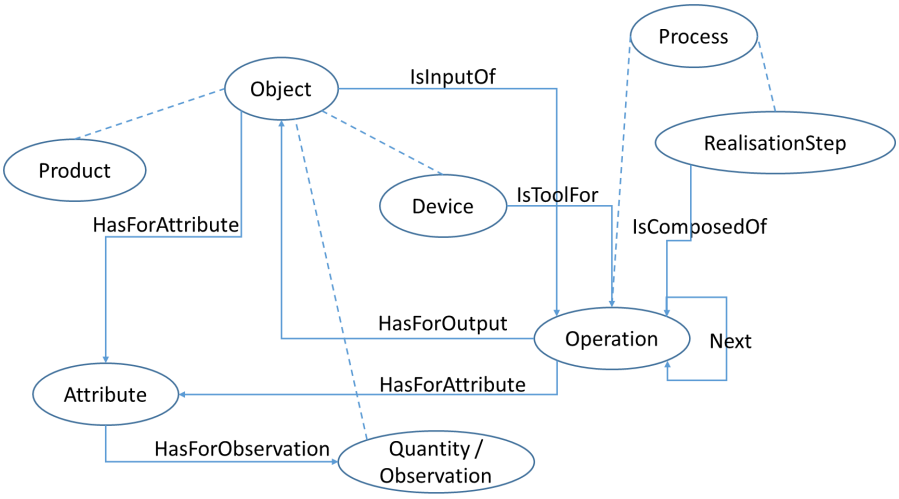


Figure 2.4: The general schema of relations between the classes used to describe the proposed ontology. Subclasses are connected with discontinuous lines.

sequence given by the recipe. In Figure 2.5, we report part of the SUMO ontology highlighting the classes we used and the ones we defined.

In the SUMO ontology, *cooking* is a subclass of the class *process*. We defined two subclasses of the class *cooking*: *operation* and *realization step*. An *operation* can be applied to a *device*. For example, the *operation* of *pre-heating* the oven at a certain temperature has as input the *device oven* and operates changing its state. An *operation* can also be applied to one or more *product*(s). The *device mixer* can be used to *whip eggs*, whipping takes as input eggs and returns eggs with changed properties. The *operation whipping* uses the *device mixer* to modify some of the properties of the object *eggs* given as input. Another example of *operation* applied to one or more *products* is the operation of *mixing flour* and *sugar*. The *device*s *spoon* and *bowl* are used by the *operation mixing*. The *device spoon* is used to *mix* the two *product*s in a *bowl*, to return a *product* that is an *intermediary mixture*.

In the SUMO ontology, *food* and *device* are subclasses of the class *object*. We defined a subclass of *object* that is a superclass of the class *food*. We called it *product*. This can be a *food* or an *intermediary mixture* with its own recipe. For instance, *flour* is an ingredient of a recipe of a cake, it is a *food* and so a *product*; the *mix* made of *flour* and *sugar* ready to be added to *eggs* in the cake baking process is the output of the *mixing* operation; the *cream* to be put on top of a cake is an ingredient of the recipe which can be prepared separately with its own recipe.

The SUMO class *attribute* represents qualities of objects or operations. The *food flour* has *attribute type* which can have value 'whole grain', the *device oven* has *attribute*

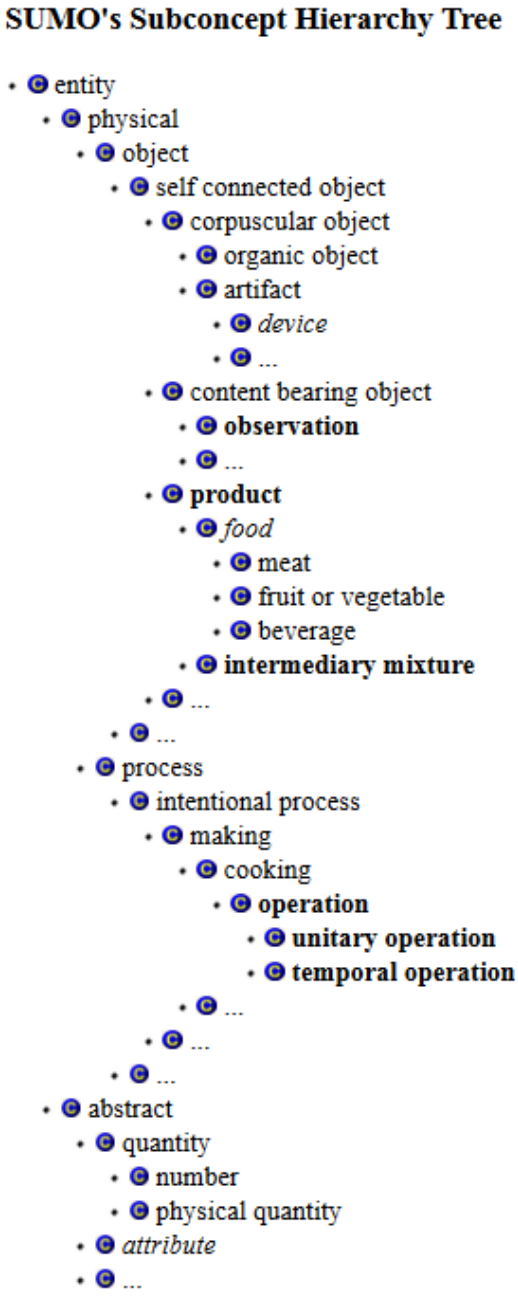


Figure 2.5: Part of the SUMO ontology, highlighting in italic the classes we used and in bold the classes we introduced. We have omitted part of the classes we did not use.

temperature which can have value '280°' and the *operation mix* has *attribute speed* with value 'quick'. To record the values of the *attributes* we defined the class *observation* as a subclass of the *content bearing object* SUMO class[8]. While making a cake, we can observe the *mixture* of *flour* and *sugar* and record its *color* and *temperature* (color and temperature are attributes of the mixture, the observations about them are collected in the class observation). While observing the *mixture* of *butter* and *sugar* we will register also its *granularity*. Observations cannot be modified by the transformation process.

In a recipe, there are operations that have a duration, we called them *temporal operation*s and we differentiated them from *unitary operation*s. Temporal properties can be described by the time ontology[9] of the semantic web proposed in [Hobbs & Pan 2004]. A *temporal operation* is a subclass of the time ontology class *interval*; a *unitary operation* is a subclass of the time ontology class *instant*. *Temporal operations* are *unitary operations* with a duration, for this reason we can represent them as a concatenation (or sequence) of the same *unitary operation*.

The time ontology classes interval and instant are subclasses of the time ontology class *temporal entity*, so the classes temporal operation and unitary operation are subclasses of the time ontology class temporal entity (Figure 2.6). Thus, we would use properties of the time class *temporal entity* to represent temporal relations between operations and so partially ordering the operations of a recipe in *realization steps*.

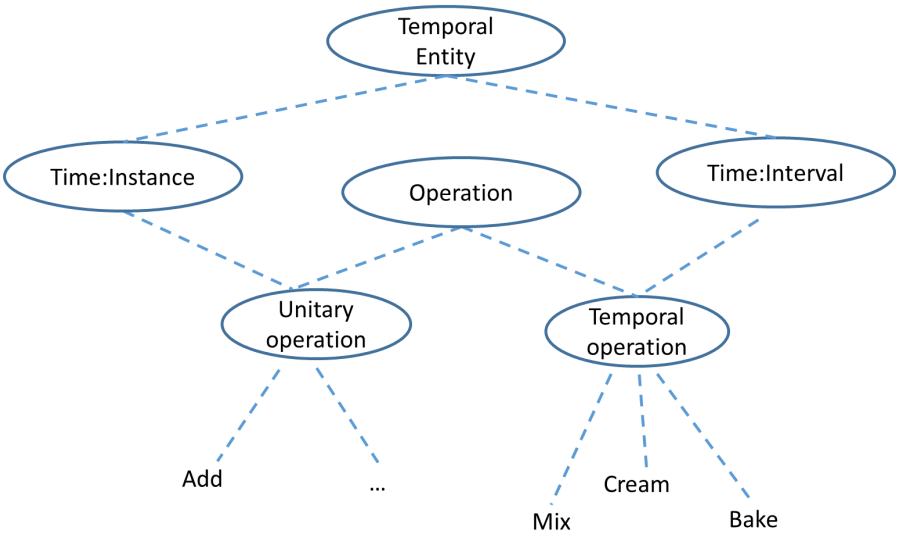


Figure 2.6: Operation's subclass hierarchy tree.

[8] A *content bearing object* is defined as a *self connected object* which expresses information.

[9] http://www.w3.org/TR/owl-time/

A Recipe Example Let's consider the following recipe for the Aunt Lila's cookies[10]:

Aunt Lila's cookies

1/2 lb butter
2 c Nuts ground
2 c All-purposes flour
4 tb Sugar
2 ts Vanilla
to roll Powdered sugar

Preheat oven to 180°C. Cream sugar and butter until light and fluffy. Add vanilla and nuts. Add flour gradually. Roll into small balls. Place on baking sheet. Bake 15 to 20 minutes. Roll baked balls in powdered sugar while still warm.

and its knowledge graph reported in Figure 2.7.

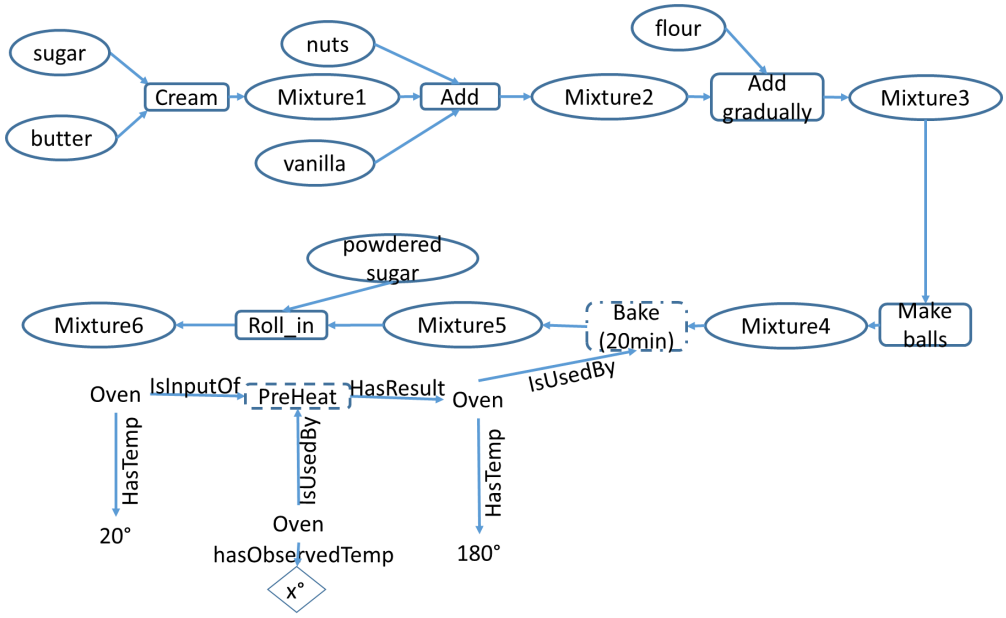


Figure 2.7: The knowledge graph for the Aunt Lila's cookies based on our ontology.

The operation *preheat* the oven is a temporal operation which relates with an observation (the $x°$ in the rhombus in Figure 2.7). Representing the observation of the

[10]The recipe for the Aunt Lila's cookies was first presented in the TAAABLE project (`http://intoweb.loria.fr/taaable3ccc/`) that is now closed.

temperature of the oven during time, could help a decision process on when to put the cookies in the oven, which can be an uncertain information.

2.2.2 Mapping

We proposed a method to deduce the relational schema for a PRM from the ontology introduced in 2.2.1. In the following I describe the mapping, for the ontology's classes: *object*, *unitary* and *temporal operation*, *attribute* and *observation* [Manfredotti *et al.* 2015].

The class *object* and its subclasses *product*, *device* and *observation* (see Figure 2.4) are mapped into (PRMs) classes, called ObjectClass:

Definition 1 *An* **ObjectClass** *in a PRM is a class which attributes are mapped from the properties of the ontology class object.*

For each class *object* in the ontology we have a class in the PRM called ObjectClass: the attributes of the ObjectClass are mapped from the properties of the class *object* in the ontology. In Figure 2.8, the class *input1* with properties *att1* and *att2* is mapped into the ObjectClass Obj.input1 with attributes the variables att1 and att2.

We proposed to map the ontology class *unitary operation* to a specific (PRM) class that we called OperationClass.

Definition 2 *An* **OperationClass** *in a PRM is defined by (1) a DAG over*

- *the reference slots giving access to the attributes of the ObjectClasses mapping the input object(s) and the device object(s) of the operation,*
- *an attribute for each property of the operation and*
- *the attributes mapping the properties of the output object(s) of the operation;*

and (2) a probability distribution over the attributes of the ObjectClasses mapping the results objects of the operation given the values of the attributes of the ObjectClasses mapping the input and the device objects.

Figure 2.8 shows (at the top) the relational schema and (at the bottom) the PRM for two OperationClasses: *operation1* and *operation2*. The output of the first operation is input for the other, so a reference slot (ρ_4) exists between the two OperationClasses. Each ObjectClass mapping the *inputs* and the *device* (Obj.input1, Obj.input2, Obj.Device1, Obj.input3 and Obj.Device2) are referred to by a reference slot in the OperationClass ($\rho 1$, $\rho 2$, $\rho 3$, $\rho 5$ and $\rho 6$). The attributes mapping the *properties* of the *output object* of the *operation* (att4, att5) define a class to which other OperationClasses can refer (see ρ_4 in Figure 2.8)[11].

[11] With respect to the literature on PRMs, we should have represented the attributes representing the properties of the object output of the operation as a class outside the class operation. Here, we represented it inside, to mean that the output is part of the operation itself.

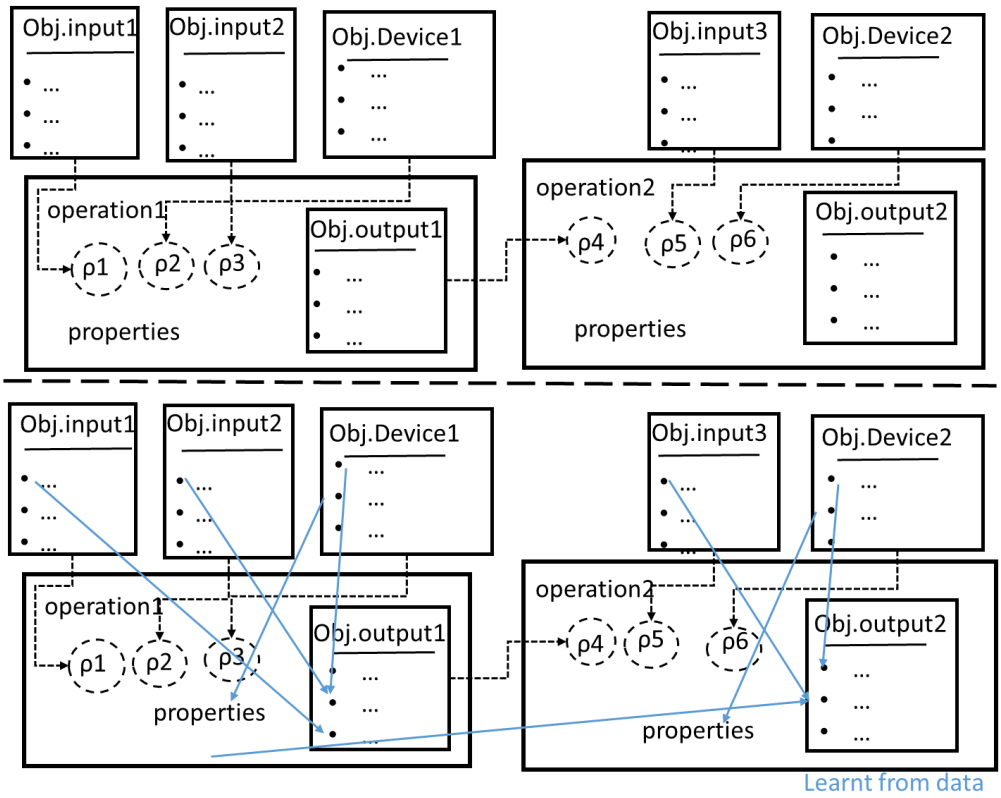


Figure 2.8: (top) The relational schema and (bottom) the PRM for two OperationClasses. A ρ_i in a class represents the reference slot giving access to the attributes of the class it refers to. Each square represents an ObjectClass.

A temporal operation is a concatenation of the same *unitary operation*. Following the standard definition of dynamic BN [Murphy 2002] we can define a PRM mapping a *temporal operation* that we called TemporalOperationClass.

Definition 3 *A* **TemporalOperationClass** *is a pair of OperationClasses with a reference slot among them:*

- *one ($OperationClass_0$) representing the dependencies between variables at the beginning of the operation and*
- *another ($OperationClass_{\rightarrow}$) representing the dependencies from the generic instant of time i to the next instant $i+1$, with a reference slot to itself.*

The second OperationClass ($OperationClass_{\rightarrow}$) refers to itself, creating a (possibly infinite) loop. To avoid the loop to run forever, we fix the number of times this OperationClass can refer to itself. In this way, we ensure the overall model to describe a probability distribution. Figure 2.9 shows the relational schema of the PRM for a TemporalOperationClass and an OperationClass. As before, the output of the *temporal operation* is input for the *unitary operation*, so a reference slot exists between the two (PRM) classes mapping them. The output of the OperationClass $OperationClass_0$ is input of the $OperationClass_{\rightarrow}$. A reference slot exists, also, between $OperationClass_{\rightarrow}$ and itself. The number of time the TemporalOperationClass can refer to itself is fixed (reported in the triangle).

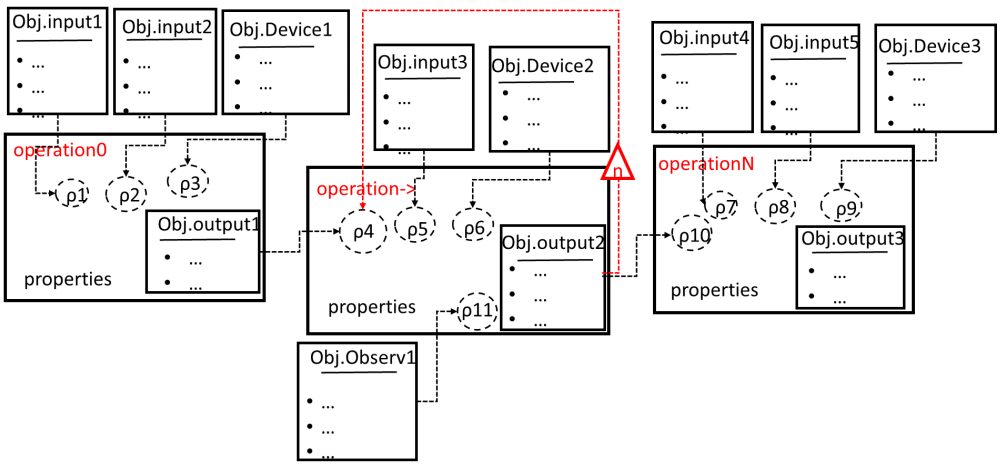


Figure 2.9: The relational schema of the PRM for a TemporalOperationClass linked to an OperationClass.

Our ontology of transformation processes is mapped into the relational schema of a PRM that is a concatenation of classes representing *realisation steps* chained by reference

slots. In our ontology, *attributes* are abstract entities representing properties of *objects* or *processes*. We mapped ontology's *properties*, in the PRM, as attributes of the classes mapping the *objects*. Finally, *observations* are ontology classes that record a particular measurement done over an *object* or *process*. In a PRM, an *observation* is mapped to a class to which an *attribute* can refer to.

A PRM for the Example Reasoning about mapping our ontology for transformation processes in a PRM led us to better define the ontology itself. In a BN, the conditional probability distribution of a node depends upon the number of its parents. Referring to the Aunt Lila's cookies example, the ontology of the operation *add* in Figure 2.7 is the same no matter the number of products we have to add together. For a PRM, instead, changing the number of parents of an attribute changes its conditional probability distribution. Following this observation, we better specified our ontology saying that, when an operation can be done on multiple products, we constraint it to be done on pairs of them. So the operation *add nuts* and *vanilla* is maintained (Figure 2.10) but the operation *add nuts*, *vanilla* and *milk* is replaced by a sequence of two operations *add*: *add nuts* and *vanilla* and then *add milk* and the *mixture of nuts and vanilla*.

Following this modification, the operation *add* became *add2* to specify that the number of its inputs is constrained to 2. In the following, we report the mapping for only three operations. The operation *add2* is mapped in a PRM with three reference slots, two for the inputs of the operation (*nuts* and *vanilla*) and one for the device used by the operation (*bowl*). The PRM defines a class *mixture1* output of the operation. In Figure 2.10 we report the relational schema of this PRM with arrows representing possible dependencies between the attributes of the classes.

The operation *bake* is a temporal operation. It is mapped in a pair of classes: one representing how the operation *bake* starts, the other representing the probability distribution of the process of baking. The OperationClass mapping the operation *bake* reported in Figure 2.11 is equivalent to a PRM consisting of the first class in the pair and 20 copies (if the duration of a time step is equivalent to 1 minute) of the second. Being *mixture4* an output of the *making balls* operation, it is formed by small balls to be put in the oven. The class *mixture4* has as property the *diameter* of the balls that is mapped as an attribute of the PRM ObjectClass mixture4. The *diameter attribute* of *mixture4* influences the consistency of the output of the baking operation *mixture5*, as expressed by the probabilistic dependency that exists between these two attributes.

The operation *add gradually* is a special temporal operation because the ontology does not give us the number of times the probabilistic model has to loop over the second class in the pair before passing to the operation that is next to it (Figure 2.12). To treat this problem we proposed two solutions.

- **Structure uncertainty.** If a probabilistic distribution p on the number of times

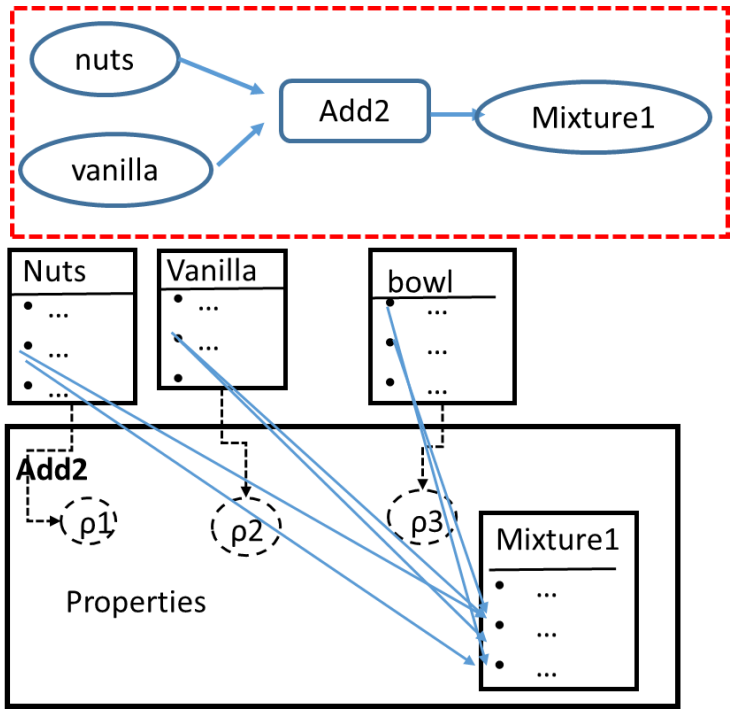


Figure 2.10: The PRM for the operation add2.

the loop has to be done is given, we can make the structure uncertain. We add a parameter θ parent of the operation following the temporal one. The probability of the operation given θ is given by p.

- **Simulation process.** We can define a simulation process on top of the PRM ruled by the conditions underlining the exit of the loop (*e.g.* cook till brown). The exit from the loop will depend on the value of this condition.

Having a PRM for the Aunt Lila's cookies recipe can help reasoning about different questions that are not possible to be answered with an ontology alone. For instance, we could compute the probability of having tasty Aunt Lila's cookies, given the fact that we have/haven't cream well butter and sugar (this is the prediction problem). We could also infer the probability of having done a good job in creaming butter and sugar having observed tasty cookies (inference problem). The defined PRM can be used to suggest a specific sequence of operations to obtain a certain output. For instance, given the butter at a certain temperature, we could suggest the best speed at which to use the mixer to cream it with sugar (process control). Finally, we could use the PRM to simulate experiences under different conditions.

Based on this work, in [Münch *et al.* 2017], we proposed to use the knowledge of an ontology to learn the parameters of a PRM from data. Using an ontology helps us

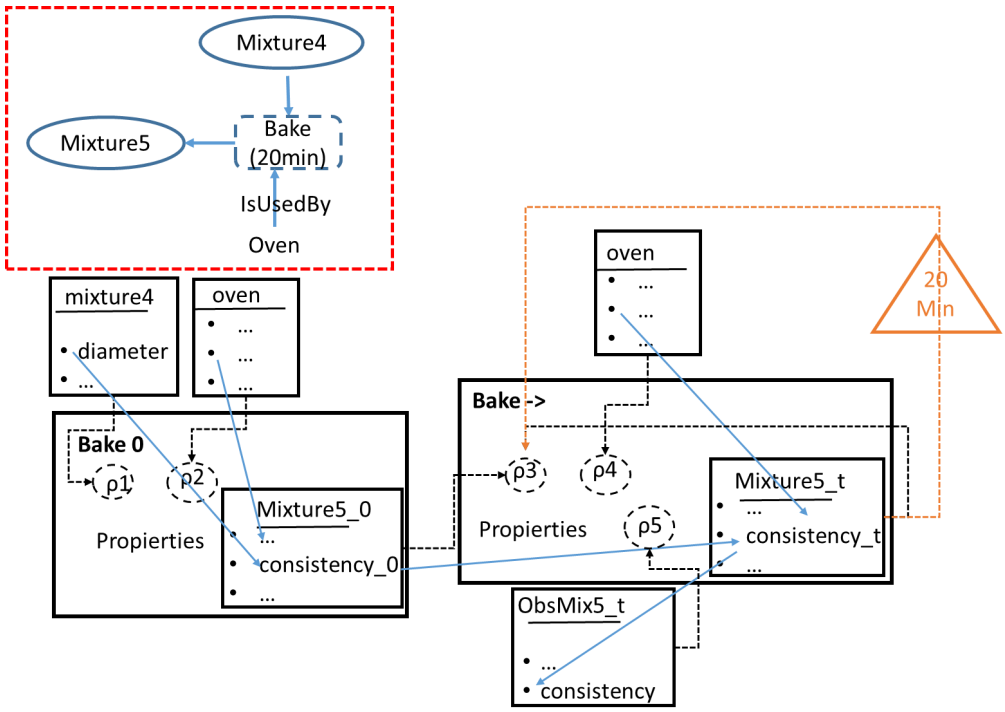


Figure 2.11: The PRM for the operation *bake*.

by integrating the experts' knowledge to ease the learning in complex domains. In the next section, I present this approach.

2.3 Learning a PRM from an Ontology

In [Manfredotti *et al.* 2015] we presented an approach to deduce a relational schema from a given ontology, once the structure of the relational schema is known, learning the relational model of a PRM can be compared to selecting the structure of a BN [Getoor & Taskar 2007]. The main difference is that probabilistic dependences between attributes in the same class have to be identical. At this purpose, the PRM relational schema and the ontology's semantic knowledge give us patterns on which to learn. Following this idea, in [Münch *et al.* 2017], we proposed a method that learns a PRM from data using the semantic knowledge of an ontology describing these data in order to make the learning easier. Based on [Manfredotti *et al.* 2015], we proposed to use the knowledge of an ontology to define the relational schema of a PRM and to learn the relational model of this PRM from data.

This work is part of the thesis work of Mélanie Münch, whose one of the goals was to provide a tool for reasoning on transformation processes. For this reason, to illustrate our approach, instead of the simple ontology for cooking recipes, we proposed to use

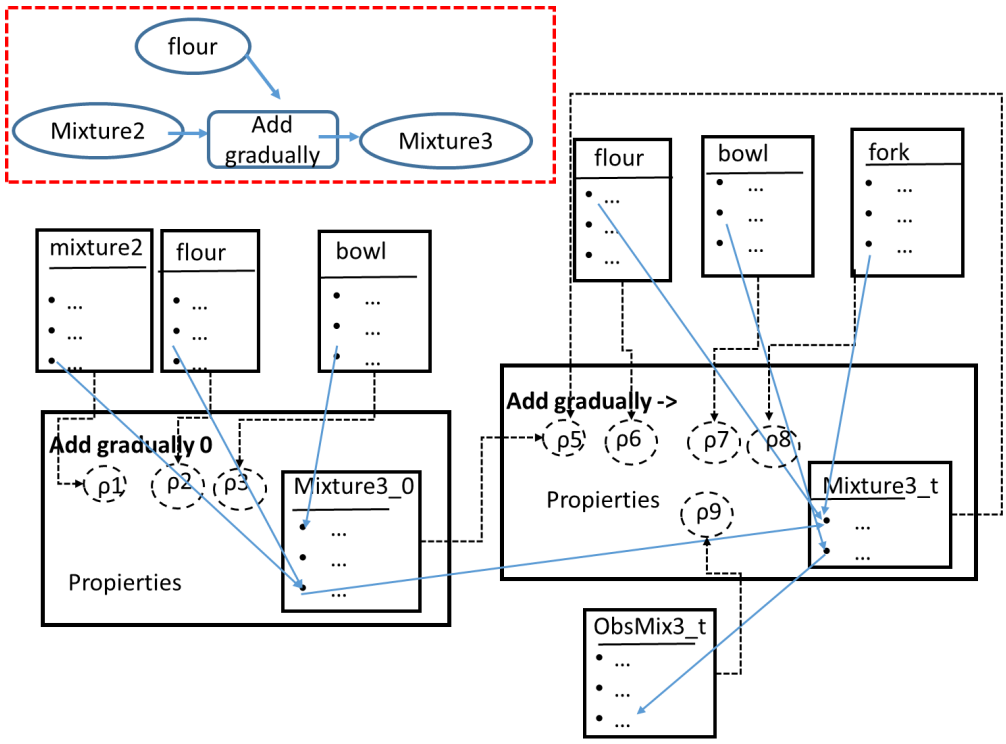


Figure 2.12: The PRM for the operation *add gradually*.

the ontology PO^2 [Ibanescu *et al.* 2016], that is close to the ontology presented above but broader in domains applications and already used in the literature.

In [Münch *et al.* 2017], we introduced an example of the ontology PO^2 specialised in the micro-organisms stabilization process domain denoted by PO^2_{stab}. A micro-organisms stabilisation process can be described as a transformation process. Fig. 2.13(a) gives an excerpt of PO^2_{stab} where there are 3 steps, *Fermentation*, *Culture* and *Stabilization* which are sub-classes of the class *Step* and 2 attributes, *SugarQuantity* and *Temperature* which are sub-classes of the classes *Attribute*. Fig. 2.13(b) shows an instance of PO^2_{stab}. In this example, there are three instanciated steps linked by a linear temporal dependency *Fermentation_1* that is before *Culture_1* that is before *Stabilization_1*. The instance *Fermentation_1* of the class *Fermentation* has for participant *Mixture_1* (an instance of the class *Mixture*) which has for sugar quantity (the instance *SugarQuantity_1* of the class *Attribute*) the value $2g$. Moreover, an observation (the instance *Observation_1* of the class *Observation*) was made on the temperature (the instance *Temperature_1* of the class *Attribute*) of *Mixture_1* which has for value $5°$.

2.3.1 Relational Schema Mapping for PO^2

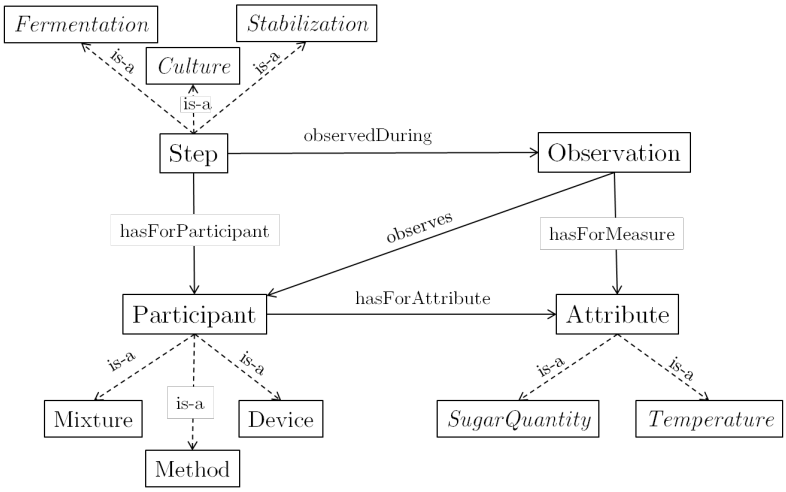
In [Münch *et al.* 2017], motivated by the description of transformation processes given by the ontology PO^2, we proposed a mapping that is slightly different from that described in [Manfredotti *et al.* 2015]. We refer ourselves to the definition of *state* in the theory of control and expert systems that allows to have a complete description of the state of the system over time.

In the theory of control, a system can be described as a succession of *states* through time [Thrun *et al.* 2005]. A state contains a set of every attribute that enables to describe the system. Observations can be made to evaluate these attributes; however, the act of observing is independent of the state itself. These definitions and the semantic representation of transformation processes in PO^2 brought us to define the following temporal dependencies properties.

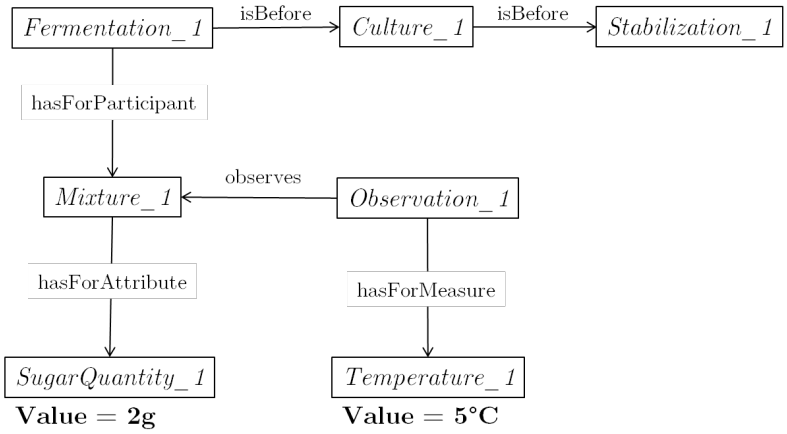
- **Observations can be longer in time than the states they observe.** For instance, some measurement methods in biology are based on time dependent reactions; in this case, the result of observations can be physically obtained even if the step linked to these has ended before;

- **States influence the result of observations, but observations do not influence states' values.** From this property, we can deduce that observations cannot influence other observations.

In the relational schema, we, therefore, proposed to define two classes built from the ontology's classes: the *Participant Class*, $\mathscr{P}$, that groups every *a priori* attribute and the *Observation Class*, $\mathscr{O}$, that groups every measured attribute. At each time step *t*,



(a) Excerpt of the ontology PO^2_{stab}



(b) Knowledge graph about the micro-organisms stabilization transformation process

Figure 2.13: An example of a knowledge base about the micro-organisms stabilization transformation process that uses the PO^2_{stab}

the method instantiates these two classes: $\mathscr{P}_t$ and $\mathscr{O}_t$. We called *Step*, denoted by $\mathscr{S}_t$, the couple $\mathscr{P}_t$ and $\mathscr{O}_t$.

The temporal dependencies properties introduced above can be formalized between the two classes $\mathscr{P}_t$ and $\mathscr{O}_t$ as the following *temporal dependencies constraints*. $\mathscr{P}_t$ can have none or multiple $\mathscr{P}$ parents at time *t-1* (that we called altogether $\mathscr{P}_{t-1}$), but always maximum one child at time *t+1* ($\mathscr{P}_{t+1}$). $\mathscr{O}_t$ only depends on $\mathscr{P}_t$. To each $\mathscr{P}$ class an $\mathscr{O}$ class is linked. Through slot chain, each $\mathscr{P}_T$ class has access to every attribute of $\mathscr{P}_t$ with t<T. Each $\mathscr{O}_t$ has only access to the attributes of $\mathscr{P}_t$.

The relational schema mapped from the PO^2 ontology is represented in Fig. 2.14: the arrows, o→, represent the reference slots. Given two classes $\mathscr{P}_t$ and $\mathscr{P}_{t-1}$, $\mathscr{P}_{t-1}$o→ $\mathscr{P}_t$ means that attributes of $\mathscr{P}_{t-1}$ can be parents of attributes of $\mathscr{P}_t$. According to the temporal dependencies constraints, attributes of $\mathscr{O}_{t-1}$ cannot be parents of attributes of $\mathscr{O}_t$. This is expressed by the absence of a reference chain between $\mathscr{O}_{t-1}$ and $\mathscr{O}_t$.

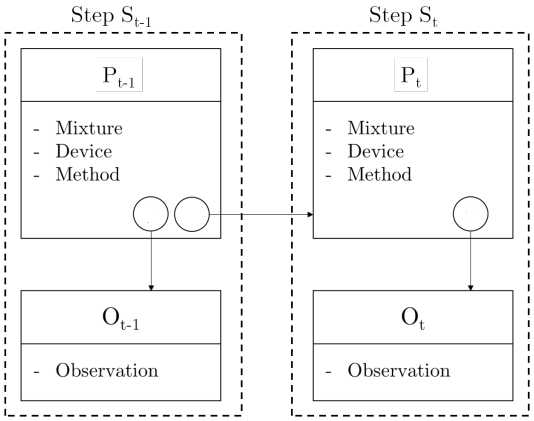


Figure 2.14: Relational Schema mapped from the PO^2 ontology for two steps.

This relational schema has two interesting properties we used in the learning. First, it preserves the *compartmentalization* between the different steps and between the participants in the process and the observations about the process. In the relational schema the attributes of an observation class only depend on the attributes of the participant class it is associated with. This allows us to consider, while learning the relational model, only meaningful attributes, defined by the ontology. In the example of Fig. 2.13, we can deduce that the *SugarQuantity* is an attribute of the mixture in the Participant class and the *Temperature* is an attribute of the Observation class. Moreover, we can deduce that these two attributes are specific to the Fermentation step.

Second, it preserves the integrity of the steps through time: a choice made at time *t* (*i.e.* the value of an attribute of $\mathscr{P}_t$) cannot influence an observation at time *t-1*. This led us to define the *direction learning constraint* used in the learning: if attributes are dependent in the domain ontology, the learnt links between them can only have one

direction. In the example of Fig. 2.13, from the instances of PO^2_{stab} we can deduce that the sugar quantity, an attribute of the mixture, can have an influence on the temperature, an observation attribute of the mixture. Moreover, considering that the fermentation step is before the culture step, the sugar quantity can also have an influence on the values of the attributes associated with the culture step.

The ON2PRM algorithm, presented in [Münch *et al.* 2017], learns PRMs' relational models using its relational schema (mapped from PO^2) and the ontology PO^2.

2.3.2 The ON2PRM Algorithm

Let us consider a knowledge graph K about a transformation process, where each attribute is represented using concepts defined in the ontology. Following the *compartmentalization* property introduced in the relational schema, during the learning from K, we create several sub-databases, each containing data of only one step: only the attributes of the step (*i.e.* attributes from the $\mathscr{P}_t$ and the $\mathscr{O}_t$ classes) and their parents (*i.e.* attributes from the $\mathscr{P}_{t-1}$ class). This ensures to preserve the organization between participants and observations. Afterwards, using the *direction learning constraint*, we force a learning order over the attributes of the same sub-database. This ensures that the temporal order between steps is preserved. However, preserving organization and temporal order does not imply links existence but only that, if a link exists, its orientation is defined by the direction and the organization given. From the instances of PO^2_{stab} (Fig. 2.13(b)), if we consider the fermentation at time t, we can deduce that the attribute quantity of sugar will be part of the participants classes fermentation ($\mathscr{P}_t$) and culture ($\mathscr{P}_{t-1}$), while the attribute temperature will be an observed attribute of the class fermentation ($\mathscr{O}_t$).

We called ON2PRM(M) our algorithm that learns a PRM relational model from an ontology where M is a learning method for BNs that can be used to draw probabilistic dependencies between attributes from a database. For each step (*e.g.* the steps fermentation, culture and stabilization in Fig. 2.13), the ON2PRM(M) algorithm uses M over the attributes (*e.g.* the attributes quantity of sugar and temperature) following the established learning order, to learn a small BN for each identified class of the PRM. This means that if, at time t, the step fermentation is occurring, the ON2PRM(M) algorithm uses M over the values of the attribute quantity of sugar at time t and $t-1$ (for the participants classes $\mathscr{P}_t$ and $\mathscr{P}_{t-1}$) and the values of the attribute temperature at time t (for the observation class $\mathscr{O}_t$) to learn a small BN for $\mathscr{P}_t$, $\mathscr{P}_{t-1}$ and $\mathscr{O}_t$. Once every class has been learnt, the PRM relational model is defined and can be instantiated (see Algorithm 1).

The PRM relational model can be instantiated with the variables in K providing the system of the PRM. In [Münch *et al.* 2017] we used the instantiated PRM to compare the performance of our approach to that of a method that learns a BN directly from data. We demonstrated that, thanks to the use of the semantic knowledge represented in

Input: ontology PO^2 + relational schema + knowledge graph K + learning method M
Result: a PRM relational model
//the **for** loop is justified by the compartmentalization property of the relational schema
//the identification of the steps relies on the classes and classes' hierarchy defined by the classes of PO^2
for *each step at time* t **do**
  //the identification of the attributes relies on the classes and classes' hierarchy defined in PO^2 ;
  identify attributes for $\mathscr{P}_t$;
  identify attributes for $\mathscr{P}_{t-1}$;
  identify attributes for $\mathscr{O}_t$;
  create a **sub-knowledge graph** from K from the identified attributes;
  //the **learning order** is defined from the instance of PO^2 as defined in the direction constraint ;
  define the **learning order** ;
  learn a BN of a PRM class from **sub-knowledge graph** + **learning order** + **method** M;
end
//the PRM relational model is the set of the PRM classes generated above, linked to each other following the PRM relational schema ;
create the PRM relational model ;

Algorithm 1: ON2PRM(M): Learning a PRM using an ontology

the ontology, learning a PRM with an ontology is more efficient than learning it without. We compared the performance of learning with our algorithm, ON2PRM(M), to the performance of learning only with the method M for two different learning methods: the *Greedy Hill Climbing* algorithm with BIC score (that we called $M1$) and the *Local Search with Tabu List* algorithm with BDeu score ($M2$)[12].

The learning was performed on 64 000 databases. We compared the instantiation of the relational models learnt by the ON2PRM algorithm using both learning methods $M1$ and $M2$ (ON2PRM($M1$) and ON2PRM($M2$)), with the BNs learnt by $M1$ and $M2$ alone. With the approach at the state of the art, the learning was done directly from the database. Both compartmentalization and the direction constraint drastically

[12]These are two standard, well known, methods for learning BNs. The description of these and others methods can be found in [Neapolitan 2003]

reduced the number of possibilities the method M must consider in the ON2PRM(M) algorithm.

Thanks to the addition of the semantic knowledge the learning's complexity was reduced and the learnt models were more meaningful than those learnt with a simple direct learning. In our experiments we demonstrated the efficiency of our approach compared to the one without prior knowledge, even in low-complexity processes or with few data. In [Münch *et al.* 2019a] we extended this approach to causal relations discovery. This is what I show in the next section.

2.4 Causal Discovery

Discovering causal relations in a knowledge base is an interesting task as it gives a new way to understand complex domains. In [Münch *et al.* 2019a], we presented a method to combine an ontology with a PRM, in order to help a user to check his assumption on causal relations between data and to discover new relationships. This assumption guides the PRM construction and provides a learning under causal constraints.

In [Münch *et al.* 2017], we showed that using the semantic and structural knowledge contained in a knowledge base, PRM learning can be greatly eased. Moreover, we showed that the learned model is closer to the reality described by the ontology. However, different PRMs can be defined from a same knowledge base. Thus, in order to select one, in [Münch *et al.* 2019a], we considered a causal assumption given by a user (a domain expert) of the form "*Does attribute C have a causal influence over attribute E?*" that he wants to check.

In [Münch *et al.* 2019a] we defined causal constraints as an ordering between the different attributes of a PRM. Following [Münch *et al.* 2017] each class of the PRM can be learnt as a BN. We can learn these BNs under precedence constraints with the K2 algorithm [Cooper & Herskovits 1992], that requires a complete ordering on the variables, or with other algorithms (as for example the one presented in [Parviainen & Koivisto 2013]) which require only a partial order on the variables. As a consequence, a system of instantiated classes linked together is equivalent to a big BN composed of small repeated BNs and thus can be associated to a (big) EG. The intuition expressed in [Münch *et al.* 2019a] is that if a BN is learned under causal constraints, its EG can give us a new insight: if an arc is oriented, then it could represent a causal relation.

2.4.1 Causal Discovery Driven by an Ontology

In [Münch *et al.* 2019a] we presented a four-steps interactive approach that learns a PRM from a domain represented by a knowledge graph (KG) guided by a causal assumption provided by an expert:

1. the user expresses expert's knowledge in a causal form "*The attribute E has a*

causal influence over the attribute C" that he wants to check in a given knowledge base;

2. the attributes of the user's causal assumption are used to define, from the knowledge base, the attributes of two classes in the relational schema (RS), the explaining and the consequence class;

3. the attributes previously defined for each class of the RS are enriched with new attributes from the knowledge base, judged as interesting by the expert for the study of the causal assumption;

4. using the defined RS a PRM is learned, whose analysis will validate the expert's causal assumption and, eventually, uncover new causal relations.

The *expert's causal assumption* $\mathcal{H}$ is of the form: "E_1, ..., E_n *have a causal influence on* C_1, ..., C_p" with E_i an explaining attribute and C_j a consequence attribute. We denoted the sets of the attributes of $\mathcal{H}$ as $A_E^{\mathcal{H}} = \{E_1, ..., E_n\}$ and $A_C^{\mathcal{H}} = \{C_1, ..., C_p\}$, with $A^{\mathcal{H}} = A_E^{\mathcal{H}} \cup A_C^{\mathcal{H}}$.

Using the instantiated transformation process of Fig. 2.15, a user's assumption $\mathcal{H}_f$ over our PO^2 example could be: "*The attributes of* p_1 *and* p_2 *have an influence over* o_4 *and* o_5", with $A_E^{\mathcal{H}} = \{a_1, a_2, a_3\}$ and $A_C^{\mathcal{H}} = \{o_4, o_5\}$ (grayed out in Fig. 2.15). In order to construct a PRM as close as possible to the causal assumption provided by the user, we defined a generic RS composed of two different types of classes, the *explaining* and the *consequence* classes, whose attributes are respectively denoted as explaining and consequence attributes. Distinguishing between them influences the causal discovery: if a relation is found between an explaining and a consequence attribute, the direction of causality is automatically determined from the explaining to the consequence attribute. We defined this as a *causal constraint*.

The class order guides the PRM learning, as we restrained our set of possible structures only to those that respect these causal constraints. Once the RS has been defined, we need to select attributes to fill the classes. Since the probabilistic dependencies are learned using a score-based Bayesian learning method, this depends on statistical evaluation. Thus not all attributes from a knowledge base can be selected: they must fit certain conditions and be useful for the learning.

In a knowledge base $\mathcal{KB}$ we call *useful learning attribute* an attribute a that is not constant and whose set of values is bound. These useful learning attributes correspond in $\mathcal{KB}$ to datatype properties $p \in DP$. In the example of Fig. 2.15, if we consider that all instances of a same attribute have the same unit, then the datatype property *hasForUnit* is not useful.

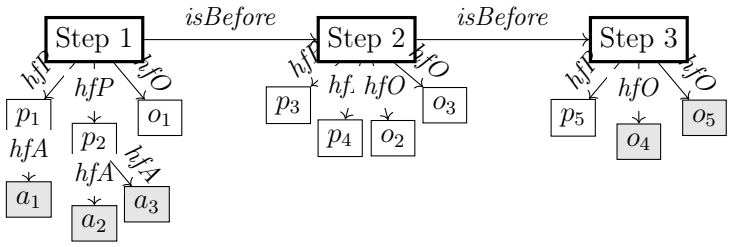


Figure 2.15: Example of a knowledge graph of a transformation process that uses the PO^2 ontology. *hfP: hasForParticipant, hfO: hasForObservation, hfA: hasForAttribute*

2.4.1.1 Assumption's Attributes Identification

In order to identify the attributes of the explaining and the consequence classes of the *RS*, we proposed to build the set $S_{\mathcal{H}}^{\mathcal{KB}}$ of all useful learning datatype properties of $\mathcal{KB}$ corresponding respectively to the explaining and the consequence causal assumption's attributes. To do so our approach starts from each attribute a of $A^{\mathcal{H}}$ of the assumption $\mathcal{H}$, and builds the set S^a of its corresponding datatype properties in $\mathcal{KB}$. First it uses the Jaccard similarity measure to compute for each attribute $a \in A^{\mathcal{H}}$ the similarity between its name and a $\mathcal{KB}$ entity's label. If it is higher than a certain α experimentally fixed in [0,1], it does one of the following:

(i) if the entity is a datatype property, it is added to S^a;

(ii) if the entity is a class, the approach adds to S^a all of its datatype properties;

(iii) if the entity is an object property, the approach gathers its range and domain classes and applies (ii).

Second, for each datatype property added, it checks whether they are useful for the learning and, if not, it deletes them from the set. For all S^a it also verifies that a connected KG could be constructed from their union, to prevent cases where each datatype property has individually enough instantiations but not enough global instances that link them together.

Finally, the user checks each S^a and chooses to exclude those datatype properties he judges inadequate. At the end of this process, for each attribute a, its set S^a is either entirely checked or empty: in this last case, it means that the attribute a is not relevant for $\mathcal{KB}$ and that $\mathcal{H}$ cannot be checked.

In our example, $\mathcal{H}_f$ defines three participants a_1, a_2 and a_3 considered as explaining, and two observations o_4 and o_5 considered as consequence. Only the useful datatype property *hasForValue* is selected. As a consequence $\mathcal{H}$ can be checked.

2.4.1.2 Enriching the Set of PRM Attributes

Most of the time the attributes expressed in $\mathcal{H}$ are not enough to find causal relations between data. This requires to find other useful learning attributes to improve the RS building. The approach presented in [Münch *et al.* 2019a] makes successive iterations on the KG over the properties, starting from the entities found and following the ones to which they are linked if they have enough instances. If it finds a datatype property through a path with enough instances that is useful for the learning and relevant, it adds it to the set of attributes for the analysis. When adding a datatype property, the user has to decide in which class he wants to put it: if he does not know, it is put in the higher explaining class by default.

In our PO^2 example the other participants' values attributes and observations are selected. The separation into steps induces the need for new classes: we wanted to be able to separate for each step explaining and consequence attributes. As a matter of fact, if we consider that each step happens at a distinct moment and that attributes can only be explained by those that happened at the same time or before, then we need to define at least one explaining and one consequence class for each considered time step. Fig. 2.16 (b) presents the RS defined to respect these constraints.

2.4.1.3 PRM Construction

In order for the user to check the model, we proposed *an interactive and iterative method* based on the study of the EG. Considering that the PRM has been learned under causality constraints (given by the expert), we made the assumption that the EG helps to determine causal relations: if an edge is oriented in the EG, then it is said causal assuming that (1) the data we dispose is representative of the reality, (2) all the attributes interesting for the problem are represented and (3) the causal information brought by the user is considered as true. We made two verification: a first one for the inter-classes relations and a second one for the intra-class relations.

The EG *inter-classes* relations are the first to be presented to the expert since they are the ones he had direct control over: if he detects a wrong orientation, it means that the RS has been badly constructed and has to be modified. The *intra-class* relations are then presented. In the case of a *non oriented* relation in the EG, the user can choose to keep the orientation as it is in the learned PRM or inverse it. In the case of an *oriented* relation in the EG, the expert can choose to keep this orientation, or declare it wrong according to his knowledge on the domain. If the expert wants to modify the orientation of the relation, a modification of the RS is required. When *modifying the* RS, two cases are possible: if one of the nodes only needs to change class, then the same class structure is kept in the RS; otherwise new classes need to be introduced.

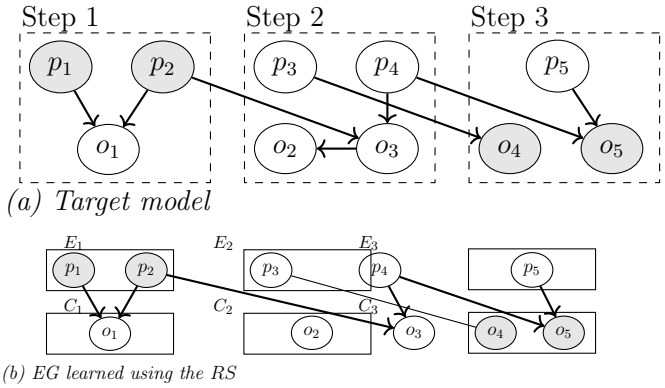


Figure 2.16: Comparison of the model used to generate the database (a) with the learned EG using the RS (b).

2.4.2 Experiments

From the KG of Fig. 2.15 and the probabilistic relations defined in Fig. 2.16 (a), we generated a data-set of 5000 different instances (165,000 RDF triplets) and applied our method. Fig. 2.16 (b) shows the EG learnt. All relations except for one inter-class are oriented, meaning that considering our knowledge base and the constraints brought both by the ontology (*i.e.* time constraint) and the causal assumption, only one result is possible. Using it, we can see that p_1 and p_2 do not explain o_4 and o_5: $\mathcal{H}_f$ is therefore not checked.

In [Münch *et al.* 2019b] we illustrated this method with a part of the DBpedia[13] KG dedicated to writers. The DBpedia database collects and organizes all available information from the Wikipedia[14] encyclopedia. Since it describes 4.58 million things (including persons, places, ...), we have decided to study only a small part of it, on a subject simple enough where we could easily play the role of experts. As a consequence, we have restrained our study to a much smaller KG[15], dedicated to writers. We have selected four classes to represent our domain: Writer, University, Country and Book. The selected KG is presented in Fig. 2.17. Considering all possible Datatype Properties (DPs) for every instance of these classes and also all Object Properties (OPs) between them, we have a data-set of 2,966,073 triples.

We wanted to study the possible influence of a university over a writer's work. Using the data-set and the RS defined by the causal assumption "Does the university have an influence on the books of a writer", it was possible to learn a PRM and study its EG (Fig. 2.19 (a) and (b)).

Despite not being experts of the domain, most of our results appeared to agree with common sense. For instance, it seems logical that a university's ARWU rank and its

[13] https://wiki.dbpedia.org/
[14] https://www.wikipedia.org/
[15] https://bit.ly/2X0eeCw

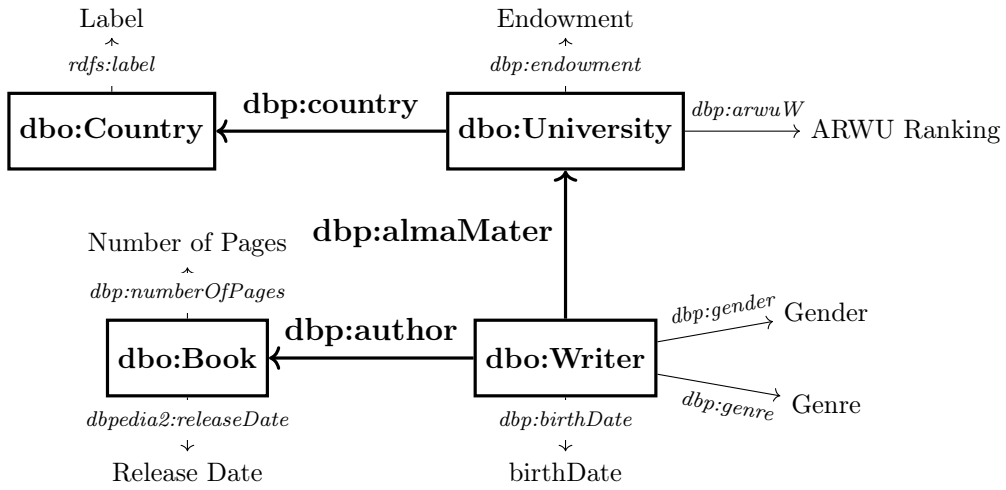


Figure 2.17: Ontology of the used excerpt of DBPedia with the Datatype Properties kept in the final *RS*.

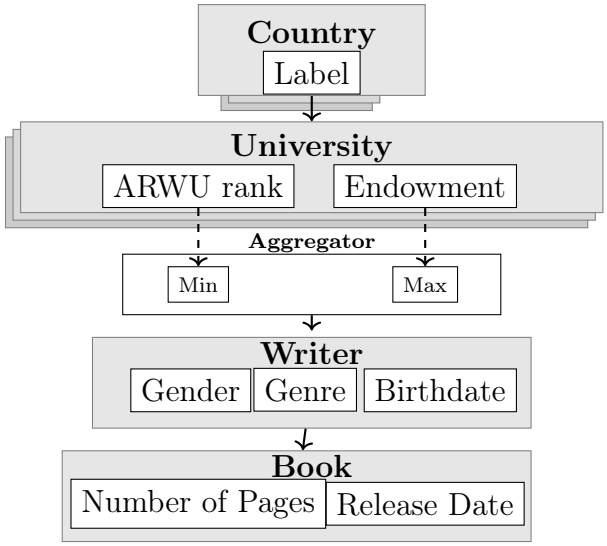


Figure 2.18: Relation Schema defined from ontological and user's knowledge. Since a writer can have multiple universities, we introduced an aggregation between the two classes.

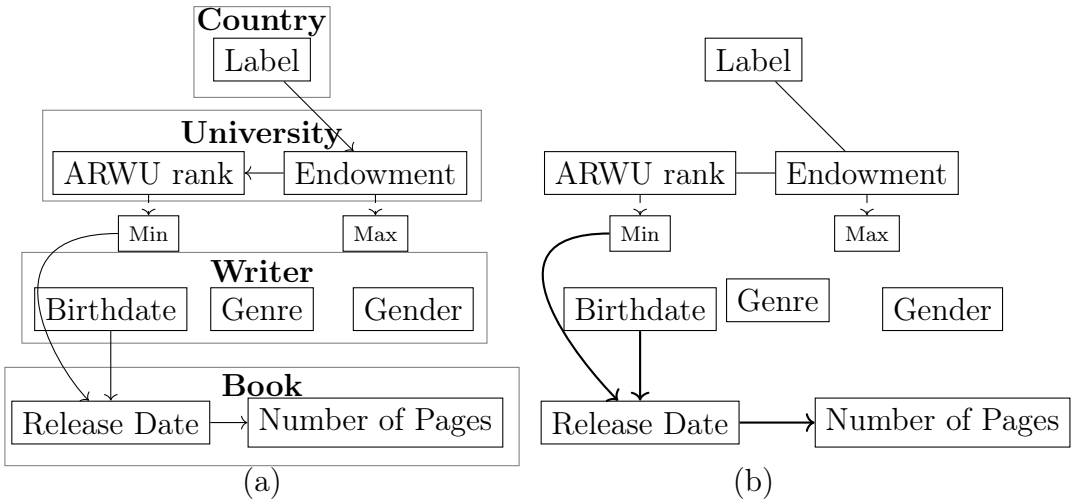


Figure 2.19: (a) PRM learned. Plain arrows indicates probabilistic relations. (b) Associated EG. Plain arrows indicates essential arcs, unoriented ones indicate the edges. Dashed arrows only serve as a visual cue to indicate aggregation.

endowment are correlated. However our KG's representativeness casts doubts on other results. For instance, we found that a book's release date can be explained by both the highest rank of the university its author went to and the author's birth date. Basically, authors born before 1950 tend to publish more before 1980 when they are from a top-tiers school. On another hand, youngest authors tend to publish after 1980, which at first seems logical: writers born after 1980 would hardly be able to publish books prior to their birth date. However, we have no instance in our data-set of books published before 1980 written by persons born after 1950, which explains why we learned this relation. This underlines the importance of a complete and verified KG: if our data-set is representative, then we acknowledge the fact that youngest authors cannot publish before 1980. On another hand, if our data-set is not representative, it means that the learned relation cannot be causal, as we have missing arguments.

The interest of this work is twofold: first, it can help a user validate his hypothesis on a domain; second, it can suggest new experiments to conduct to test new hypothesis. This method is *interactive* (*i.e.* the user can interact with the algorithm to give his inputs and influence the learning) and *generic* (*i.e.* it can be applied on any KG as long as it is relevant for causal discovery). It is also dependant on the quality of the data-set: the data-set has to be checked (*i.e.* no errors) and complete (*i.e.* no missing attributes or incomplete data). In [Münch *et al.* 2018a] we used this method to suggest strong links between plausible control variables and some parameters in the process of the cheese fabrication. Thanks to this method we were also able to propose to conduct some other experiments to understand the whole process.

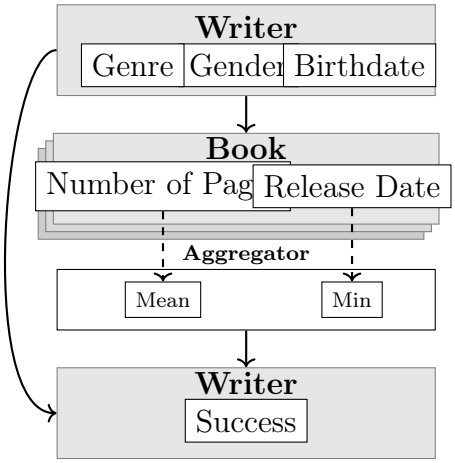


Figure 2.20: Example of a *RS* class split with creation of an aggregation.

2.5 Identifying Control Parameters

Cheese processing is a complex domain involving many different variables. Their combination leads to final products that can differ in quality which can be assessed by different criteria (*i.e.* sensory, nutritional, ...). Parameters that are necessary to explain all these criteria are denoted as *control parameters*. Modeling cheese fabrication processes helps experts to check their assumption on the domain such as finding which control parameters can explain the final products and its properties. In order to help experts assess and check their assumptions, tools and methods are needed to analyze data. These involve various parameters and reasoning over different steps. The approach presented in [Münch *et al.* 2019a] did not consider temporal information that is important to model cheese processing. The approach presented in [Münch *et al.* 2018a] is a more general approach that allows a user to integrate causal and temporal information represented by precedence constraints in order to model a cheese fabrication process and identify the transformation process control parameters[16].

2.5.1 Cheese Processing

This work has been applied on a real application about cheese processing using data from the TrueFood project. The goal of the TrueFood project was to investigate to what extent the impact of some combinations of thermophile lactic bacteria (*i.e.* Streptococcus thermophiluss, Lactobacillus helveticus LH with 2 distinct levels and Lactobacillus delbrueckii LD with 2 distinct levels) on the characteristics of hard cooked cheese is affected by the use of milks with various compositions and by the use of different technological

[16][Münch *et al.* 2019a] was published after [Münch *et al.* 2018a]. In [Münch *et al.* 2018a] we refer to [Münch *et al.* 2018b] that I find well summarised by [Münch *et al.* 2019a], this is why I refer to it.

conditions (such as distinct temperature for the heating of the milk in the vat). The data used are about 24 hard cooked cheese of 10 kg each manufactured during three weeks in January 2008, and made using 100 liters vats. Three kinds of milk, differing in their protein content and their production conditions, were used for the cheese production. During the cheese making, three different temperatures (53°C, 55°C and 57°C) were applied for the milk heating. During this process various parameters were monitored, such as different measures of proteolysis. In particular, the potentially bioactive peptides content of the cheeses were measured at several steps of the cheese ripening. Their sensory properties were also assessed at the end of the ripening step: texture and flavor were evaluated by 11 panelists on a 10 points scale.

The influence of milk heating and of combination of lactic bacteria during cheese manufacture on the formation of peptides has already been observed in the literature [Santiago-López *et al.* 2018]. Moreover the impact of the type of milk used for the cheese manufacture (especially the influence of the cows feeding system) on the organoleptic properties of hard cheeses has been shown in [O'Callaghan *et al.* 2017].

In our experiments, the experts made the assumption that the three factors of variation of the cheese making process (*i.e.* type of milk used for the cheese making, combination of thermophile lactic bacteria added to it and the milk temperature) are the control parameters for the potentially bioactive peptide content of the cheese and its sensory properties. Those attributes are measured at different times during the cheese making process. The aim of our work was to check this assumption.

2.5.2 Integrating Temporality in Causal Discovery

In the following, we consider a knowledge base $\mathcal{KB} = (\mathcal{O}, \mathcal{F})$ and a user's assumption about possible causal relations between data in the form "$E_1, ..., E_n$ *have a causal influence on* $C_1, ..., C_p$". From the assumption and the $\mathcal{KB}$, a database S is created and used for the learning. It is composed of the explaining and consequence attributes as well as of other inferred attributes as presented in [Münch *et al.* 2019a].

Our method gives the user the possibility to check his assumption about possible causal relations between the data of $\mathcal{KB}$. The integration of explaining and consequence attributes helps him express his own knowledge of the domain and guide the learning towards a coherent causal model.

In [Münch *et al.* 2018a], we observed that explaining attributes at one time step can become consequence attributes at the next time step. We denoted by **event** a group of attributes that happen at the same time. When dealing with temporal information, it is possible that the consequence attributes of an event e_t at time t become the explaining attributes of another attribute of another event e_{t+1} at time $t + 1$. Moreover, we can suppose that all the attributes from an event can have an influence over all the attributes of the following events. For this reason, in [Münch *et al.* 2018a], we proposed an extension of this method dealing with both causality and temporality constraints.

We defined a new kind of model that we called the *stack model* that allows every event to have an influence on the attributes of the events that happened after it[17].

2.5.2.1 Stack Model: Determining Precedence Constraints

Using [Münch *et al.* 2019a] where we defined explaining and consequence attributes, we proposed to decompose the precedence constraints into two sub-constraints: the causal constraints and the temporal constraints. *Causal constraints* are information on the relations between attributes of the type "*The attribute A is a possible cause for the attribute B*". *Temporal constraints* are information on the relations between attributes of the type "*The attribute A happens before the attribute B*". These causal and temporal constraints both imply two things: (1) the value of B *can be* explained by A (but it does not have to); (2) B can never explain the value of A.

Causal and temporal constraints are differentiated by their nature: temporality is immediate and objective (*i.e.* the past can influence the future and not the contrary), while causality usually needs expert's knowledge.

- **Temporal constraints.** When possible, the temporal information is provided in the knowledge base through the time ontology[18] that helps anchoring events in time. In some cases it is also possible to introduce temporal information directly from experts. In all cases we suppose that attributes can be attached to a specific event in time and, as a consequence, they contain temporal information.

- **Causal constraints.** Causal information can be brought by experts or by the ontology itself. In certain cases it is also possible to use statistical independence tests such as the χ^2 test in order to guess some possible causal relations [Spirtes *et al.* 2000].

2.5.2.2 Stack Model: Description

The stack model has been built in order to graphically represent the two kinds of precedence constraints. If an attribute is put higher in the stack then it has a precedence constraint on all attributes below it; if two attributes are on the same level then they do not have precedence constraints.

It is also possible to encounter parallel events. In this case, we supposed we had enough information from the knowledge base to differentiate the events, in order to know which attribute corresponds to which event. In this case, we defined paths for each parallel event. Events on the same path have parenthood links: temporal constraints can be established between them. On the contrary, events that do not share parent events

[17] This was also possible in the state-observation model introduced in [Münch *et al.* 2017] but it was not explicit.

[18] https://www.w3.org/TR/owl-time/

are on two separated paths and we supposed they cannot influence one another. As a consequence, there cannot be precedence constraints between them, neither causal nor temporal.

Starting from a user's assumption, the model construction is based on two operations[19].

1. **Defining temporal constraints.** Groups of attributes that happen at the same time are put at the same level. If they are from a same event, they are put in the same stack; if they are from parallel events we create different paths, each with a stack.

2. **Defining causal constraints.** Inside a stack some attributes might have a causal influence over others. In order to express those causal constraints, we sort the attributes such as higher attributes can explain lower attributes and that attributes at the same level share no causal influence between each other.

An example of this construction is given in Fig. 2.21. We considered four events: one at time t_1, two parallel at time t_2 and one at time t_3 (Fig. 2.21 (a)). When constructing the model we first considered only temporal constraints (Fig. 2.21 (b)): we created two paths with on one side a stack with the group of attributes A and on another side two stacks with respectively the group of attributes B and C, the first being above the second. Finally, we created a fourth stack below all the others, including the group of attributes D. we defined temporal constraints between the different stacks: since the group B was not on the same path as A, no temporal constraint was drawn between them. At the end, we defined causal constraints (Fig. 2.21 (c)). In our example, we supposed that the expert distinguished between explaining and consequence attributes in the group A, respectively subgroups A_1 and A_2. In order to lighten the figure, arrows between groups of attributes inside different stacks were not represented: however, if two stacks were linked, it meant that each attribute on the higher stack had a temporal constraint over those on the lower.

2.5.2.3 From Stack Models to PRMs

The final stack model is used to construct a PRM's relational schema, which defines the classes and the attributes of the PRM. Each subgroup of attributes becomes a class, which are linked together with reference slots following the different precedence constraints. For instance in the model in Fig. 2.21 (c), it would lead to five classes and six reference slots.

Once the relational schema is defined, the PRM can be learned using the database S extracted from the knowledge base [Münch *et al.* 2017]. This PRM can then be

[19] For convenience and in order to ease the readability of the presentation we used a *top-down* construction (from temporality to causality). However nothing prevents us to use the opposite *bottom-up* construction.

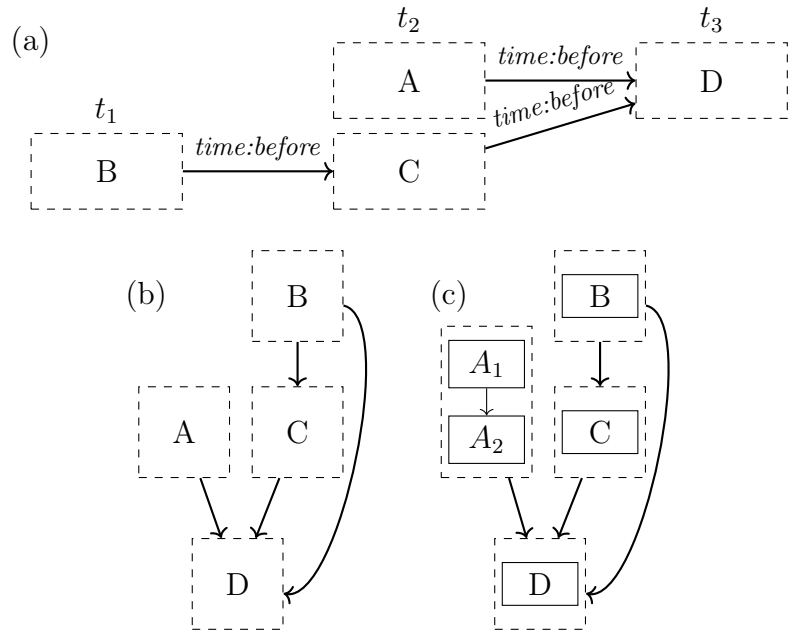


Figure 2.21: (a) Example for a system (knowledge graph) with parallel events. (b) Definition of the temporal constraints. (c) Definition of the causal constraints.

instantiated in order to obtain a BN representing the learned model. It will include causal information as it was learned under causal constraints; however, it is not a complete causal BN considering that the learning of dependencies between attributes inside the same group was dealt like a classical BN. In order to deal with causal information, following [Münch *et al.* 2019a], we used the EG: if an arc is oriented in the EG, it can mean that there is a causal relation.

2.5.3 Experiments

Considering the TrueFood project, the experts would like to model the different relations between the attributes in order to explain the products at the end and infer its characteristics. More particularly they want to check if "*The temperature, ferments and type of milk have a causal influence on the potentially bioactive peptide content of the cheese and its sensory properties*". Following the approach presented in [Münch *et al.* 2019a] temperature, ferments and type of milk are the only explaining attributes of the problem, while the other are consequences. Since those three are fixed at the beginning, they correspond to the **control parameters**. The knowledge graph is composed of data (instances) from three different steps that are part of a cheese fabrication and tasting process: Step in the vat, Ripening and Mastication.

- **Step in the vat:** is described by three processing control parameters (Tempe-

rature, Starters and Type of milk), and two measured (hardening and clotting times).

- **Ripening:** is described by the measured value of five different concentrations in cheese: butyric acid, propionic acid, acetic acid, free amino acids and free amino groups.
- **Mastication:** in this step, a panel of 11 judges has evaluated each cheese sample on 45 different criteria (*e.g.* spice aroma, sugar or fat perception). Those sensory notes can be divided into two categories, cheese texture (10 attributes) and cheese flavor (35 attributes). The scores are ranged from 0 to 10.

The times measured during the step in the vat are a pre-requisite to *study bioactive peptide contents*, even if they do not represent their quantities. On another hand the attributes measured during the ripening and the mastication steps are useful to *evaluate the cheese sensory properties*.

The obtained model is presented in Fig. 2.22, where the different steps are underlined by the dashed squares and $\times i$ denotes the number of attributes of the given type.

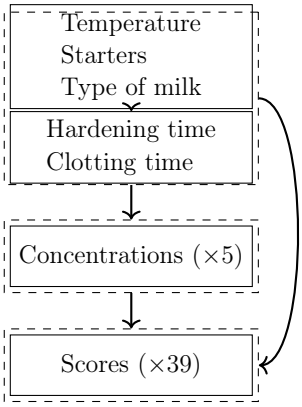


Figure 2.22: Model constructed from the expert assumption.

While analyzing the PRM we focused on the inter-step relations that give a whole new reading of the model. It indeed helped us generate new information about the temporal aspect, in particular discovering if attributes at some steps can explain all the other attributes, or, on the contrary, if a step has no influence on the process. In our case, we would like to see at what extent the control parameters are able to explain (in)directly the other attributes. To do that, we extended our study on inter-step relations, also including the inter-subgroup relations between the control parameters and the two attributes Solidifying time and Clotting time.

The vast majority of the observed inter-steps relations found *confirms the experts assumption*: "*The temperature, ferments and type of milk have a causal influence on the potentially bioactive peptide content of the cheese and its sensory properties*". Some

of them are directly explained, while others are linked to attributes of the same group that are explained by the control parameters. Only three sensory notes are not linked at all to any parameter. Those results and the number of relations found are summarized in Fig. 2.23.

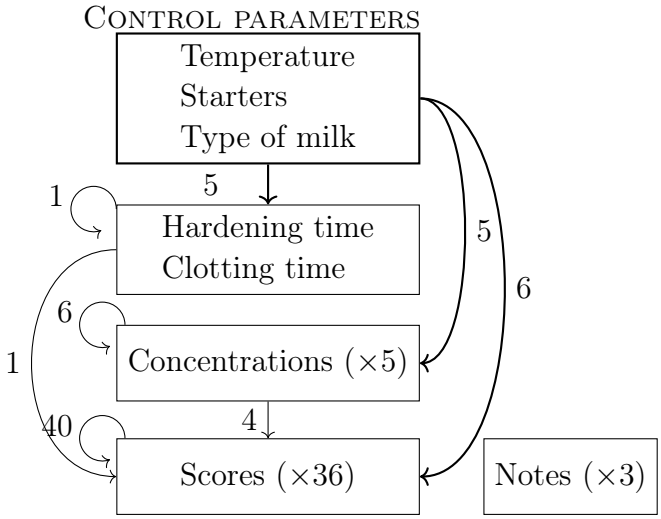


Figure 2.23: Summary of the number of observed inter and intra step relations.

The learned PRM gave us two ways of analysis. First, using the EG of its system, we could check the expert's assumption. Considering that our control parameters were fixed at the beginning of the process, if a relation was found between them and an attribute, then we could conclude that the parameters may control this attribute. Second, once the model was validated by the experts, it could be used to predict results. For instance if we wanted to control the cheese texture scores in order to keep them within a certain range, we could identify the control parameters we had to act on.

The POND workflow [Münch *et al.* 2022] extends these works to a pipeline to support technical itineraries for reverse engineering purposes. We applied this approach to the processing of bio-composites for food packaging.

2.6 POND

POND (Process and observation ONtology Discovery) is a workflow dedicated to answer expert's questions about processes, it addresses two main issues: 1) how to represent the processes inner complexity and 2) how to reason about processes taking into account uncertainty and causality. In [Münch *et al.* 2022], we showed how to use a knowledge base to answer some of the expert's questions concerning the processes, using semantic web languages and technologies. Then, we described how to learn a predictive model, to discover new knowledge and provide explicative models by integrating the semantic model

into a PRM. The result is a complete workflow able to extensively analyse transformation processes through all their granularity levels and answer expert's questions about their domains. An example of this workflow is given on biocomposites manufacturing for food packaging.

The goal of the POND workflow is to propose a way of representing and reasoning on data extracted from different sources about a specific transformation process. Our originality stems from (1) the adaptability of the representation part, that allows the combination of two knowledge sources (the ontology and the expert's inputs); and (2) the scope of the questions that can be answered through this workflow, some being answered by directly querying the data and others by analysing a model, learned for the occasion, that is able to reason with the transformation process complexity.

The functionalities of POND have been defined in the framework of several interdisciplinary projects involving computer scientists, data scientists and biomass processing experts for food and bio-based material production. We presented them from a generic point of view. While the PO^2 ontology allows us to define experts' knowledge by unifying it under common semantic terms, reasoning about this heterogeneity requires to define specific questions that we aim to answer. We denote them as *Expert Queries* (EQs), and separate them into two subsets:

- **Competency Questions** (CQs). In ontology engineering, CQs are natural-language questions that outline the scope of the knowledge represented by an ontology and the applications exploiting it [Grüninger & Fox 1995]. CQs represent functional requirements in the sense that the ontology and the knowledge base to which it belongs to should be able to answer them. Typical CQs addressed by PO^2 are:

  CQ$_1$ Which steps compose a given transformation process?

  CQ$_2$ Which attribute values are associated with each step?

  CQ$_3$ What are the attribute values associated with an input (or output) for a given step of a given transformation process?

  CQ$_4$ What are the changes for an attribute value of an input during a given step?

- **Knowledge Questions** (KQs). Similarly to CQs, these EQs query the knowledge base, but require an analysis of the relations between the variables able to deal with the uncertainty. KQs can be expressed in two different ways:

  KQ$_1$ Does a given attribute have a (causal) relation with another attribute?

  KQ$_2$ How does a change in a given attribute's value (causally) influence the values' distribution of another attribute?

  Differently from CQs, KQs require a two-times analysis: first we have to build a database representing the attributes of the question as variables and then we

have to learn a probabilistic model from the database to answer the question. More generally, CQs rely on specific classes or properties of the ontology; while KQs require BNs and PRMs concepts, such as variables that need to be defined beforehand.

In order to answer these questions, POND implements different functionalities:

F1 The workflow provides a model allowing to express both expert's and ontological knowledge and a tool to structure and store data using this model.

F2 In a collection of experimental data acquired during different projects, the workflow provides a way to extract from the knowledge base, in a semi-automatic way, attributes of interest for the analysis.

F3 The workflow is able to compute a model for reasoning with variables of interest.

F3 is specific to KQs. The three functionalities are provided by the following steps of the POND system, presented in Figure 2.24:

Step 1. **Knowledge Collection**. Expert's knowledge is collected under the form of experimental data or expert's interviews and structured using an ontology. PO^2 is used to annotate experimental data and to store it in a RDF knowledge graph. An EQs set is defined and, depending on its type, it will either be processed in Step 2 (**Knowledge Base Querying**) for CQs or in Step 3 (**PRM Learning**) for KQs.

Step 2. **Knowledge Base Querying**. CQs are expressed as **SPARQL**[20] **queries** and executed against the RDF knowledge graph. A specific Web application, SPO^2Q, has been designed in order to assist users to query the PO^2 RDF database.

Step 3a. **Mapping between PO^2 and PRM**. Answering KQs requires the learning of a PRM; however, in order to integrate the expert's knowledge expressed during the **Knowledge Collection**, a mapping is needed before interrogating the PRM. It is used to automatically translate expert's knowledge into constraints to guide the learning and it is expressed under two forms: first a mapping of the attributes, then the expression of the precedence constraints [Münch *et al.* 2019a].

Step 3b. **PRM Exploitation for Reasoning**. Directly following the **Mapping**, this sub-step consists in the learning of the PRM and in the validation of this by the expert, who can accept or reject the result using tools to criticize the model (as done in [Münch *et al.* 2019b]). If the model is rejected, the expert is invited to reconsider the knowledge integration done during the **Mapping** (step 3a), and a

[20]SPARQL is an RDF query language able to retrieve and manipulate data stored in RDF format.

new iteration begins. If, despite those iterations, the expert cannot validate the model, it means that the expert knowledge defined in the **Knowledge Collection** cannot be used to answer the KQ. In this case, the identified problems (such as the lack of knowledge) are given to the expert as guideline to improve the learning. On the contrary, if the model is validated, we continue to the final step.

Step 4. **Expert Query Analysis**. The results of the SPARQL query formulated in the **Knowledge Base Querying** step or the explicative models validated in the **PRM Exploitation for Reasoning** step are received and analyzed to answer the EQ: for instance, if no answer has been found, we can find out whether this is due to a lack of information or to a problem within one of the involved step.

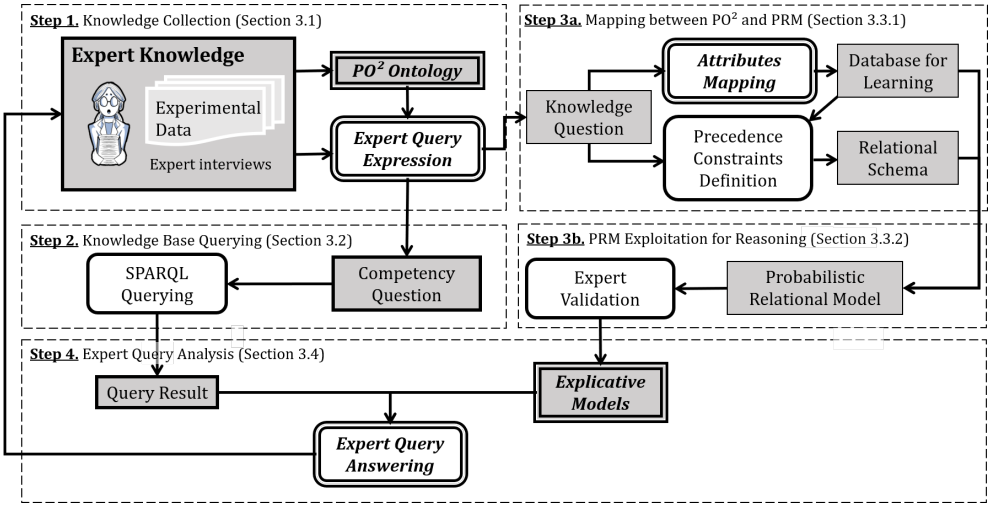


Figure 2.24: POND global overview. White boxes indicate actions that require the expert's intervention, grey boxes indicate a concrete object automatically built from expert's inputs.

Figure 2.24 presents these different steps and the different possible sequences. To be noted, the passage from Step 3b to either 3a or 4 depends on whether the expert has rejected or validated the learned PRM. The POND workflow helps the expert to gain a new overview of the studied domain, for instance by suggesting new experiments or ways to improve the data-set.

2.7 Conclusions and Future Possible Directions

In this chapter I presented how we considered experts' knowledge for learning and reasoning with PRMs: in different works we proposed to map an ontology in the relational schema of a PRM that is, then, learnt from data. We showed how doing that allows

for an easier learning even with a small data-set, for causal discovery and parameters control.

The thesis of Mélanie Münch presented different approaches for this mapping following the purpose of the application at hand. Motivated by the necessity of modeling uncertainty, we started with an ontology of which an expert has some insights and is willing to guide the learning of the PRM. With the participant and observation classes, the expert differentiates the variables whose values he can change from the variables he does not have control on. When we introduced the user's (causal) assumption, we introduced a separation between the explaining attributes and the consequence attributes. With the stack model we introduced temporal constraints on the variables. All of these approaches introduced an ordering over the attributes of the PRM. Learning is done following this ordering and taking into account the relational constraints introduced by the ontology and reveled to be easier than learning the same PRM with no ordering and no relational constraints. Moreover, the user's (causal) assumption could be validated and the expert can interact with the system leading to a more precise model.

Of course, one can think about an approach that puts all the different mappings presented in Mélanie Münch's thesis in a general one that sees the separation between the attributes we can control and the attributes we cannot, the attributes that are cause and the attributes that are consequence following a certain (causal) assumption, all this organised in the stack model. This would lead, probably, to a more general approach that we did not experiment.

What would be a natural next step from Mélanie Münch's thesis is to make a system able to give feedback to the knowledge base to improve the data and, eventually, the ontology itself. We could use the ability of the POND workflow, the assumption verification and the parameter control abilities, to evaluate the quality of potential new data (probable, not probable, impossible) and help the expert finding outliers or to suggest relational constraints that are not present in the ontology so to improve the ontology itself according to new information gathered.

Mélanie Münch's thesis focused on transformation processes and used the PO^2 ontology as it was shown by the POND workflow. While, in my opinion, extending the approaches presented in her thesis to all the ontologies, developing a mapping that suits all kind of knowledge base is not promising, I think that a possible future work could be an approach to transfer the knowledge we have over a knowledge base to another one. The idea could be to use data linking methods to ease the learning of a PRM of a new knowledge base by transfer from an existing "close" one and its PRM. I started thinking about this in 2018 with a M2 internship that I co-supervised with Juliette Dibie and Fatiha Saïs. The idea was to merge data linking techniques in semantic web to find similarities between the two knowledge bases and transfer learning techniques to transfer the knowledge we have over the system of a PRM to the other.

Data linking [Ferrara *et al.* 2013] consists in detecting whether different descriptions refer to the same real-world object. It is mostly based on a calculation of similarity

between data using elementary similarity measures [Cohen *et al.* 2003], but it uses also ontology alignment approaches whose objective is to detect the correspondences between the concepts and the relations of different ontologies.

Transfer learning techniques reuse knowledge already acquired in one domain to improve or accelerate learning in new domains. Works on transfer learning propose to use a common representation space in which the decision functions are close [Glorot *et al.* 2011]. This approach has similarities with reasoning by analogy [Murena & Cornuéjols 2016] and can be used with the perspective of choosing a descriptive model common to data from different domains. In the work I did during my postdoc at Sorbonne University [Gonzales *et al.* 2015], we presented an algorithm to learn a DBN given a similar one, this approach uses a particular learning score that could be used to learn a PRM from a similar one (mapped from the same ontology).

The idea could be to use methods from both these families of techniques to learn a PRM ($PRM2$) modeling a domain of which we have a knowledge base ($KB2$) that is close to another one ($KB1$) that has been mapped into a PRM ($PRM1$) itself. The internship I supervised in 2018, proposed to use data linking techniques and ontology alignment to find correspondences between the two knowledge bases and transfer learning techniques to transfer the (causal) knowledge and the probabilistic relations we have over the first PRM to the new one. Even if we did not have any results, yet, I really think that this is a very promising way that could improve the reasoning ability of the system.

Data linking techniques could also be used to merge data coming from different experiments. Several interdisciplinary projects contributed to the definition of the functionalities of the POND framework, but not all the data could have been used because they did not "match" altogether. We could use data linking techniques to find correspondences between the different data-sets and transfer learning techniques to learn new PRMs for the different experiment settings.

Another possible extension of Mélanie Münch's work could be to deal with domain evolution. There are domains whose settings change over time due to environmental changes or needs of the system. We could think to use transfer learning techniques to adapt the PRM obtained by mapping from an ontology representing the system to new data and to map this to a new ontology that will be the result of the first one evolved with the changes taken into account.

In the next chapter I present our research in recommender systems for the nutrition domain. I will show how some of these ideas can be pertinent (as future work) for this domain as well.

Chapter 3

Food Recommender Systems

Contents

In this chapter, I present the works I have been doing at AgroParisTech with the purpose of developing a recommender system for the nutrition domain. The aim of this set of works is to encourage people towards healthier eating habits.

Most chronic diseases such as diabetes, obesity and cardiovascular conditions are correlated to unhealthy eating habits [Rep 2003]. For this reason, public health agencies have created dietary guidelines targeting the general population in order to push people for healthier eating habits. These are the guidelines we all have seen on the public way, for instance "eat at least 5 fruits or vegetables per day", "limit your consumption

of salt"[1]. Although the awareness about healthy diets is rather good, the compliance to these guidelines by the general public are relatively low [Ivens 2016]. There are different causes that contribute to this: cultural and personal preferences, difficulty of implementation, availability and price of food items [Webb 2015] and so on.

A solution to this problem could be to develop a food related recommender system that would be able to provide suggestions that satisfy both the user (preferences, choices – *e.g.* vegetarian, religious, *etc.* – and allergies) and the nutritional constraints established by the experts. Early studies showed that web-based personalized interventions are more effective than standard public health advice for inducing compliance with healthy eating recommendations [Hageman PA 2014]. Recommender systems are based on the general idea of "suggesting similar items to similar users". They are tools that have become more and more popular for supporting the user in finding personalized suggestions of products, services and information [Adomavicius & Tuzhilin 2010, Delporte *et al.* 2014]. They have been very successful in a variety of domains (*e.g.*, movies, shopping, social networks, job portals) and deployed in a large number of applications.

They can have diverse objectives. For suppliers, the main objectives are to increase the number of items sold, build consumer loyalty and gain a better understanding of what consumers want. For users, the objectives can be finding the best items, influencing, *etc.* It is key for any recommender system to define its objectives upstream, both for the supplier and for the user [Ricci *et al.* 2015].

For a given user, a recommender system predicts his taste or usefulness score for a list of products, enabling them to be ordered. These scores are personalised and each user can have his own recommendations [Delporte *et al.* 2014]).

Recommender systems use three types of data: products, users and relationships between products and users (transactions). The most commonly used recommender systems are generally fairly content-poor. Some techniques incorporate knowledge (ontological descriptions of users or products), constraints and users' social relationships, but they are less frequently used [Ricci *et al.* 2015]. User ratings may be explicit, such as ratings given to products used or consulted, or implicit, such as clicks or online purchases, *i.e.* signals of user behaviour which may indicate preferences.

Recommender systems have been classified into three groups on the basis of the approach used to generate recommendations [Adomavicius & Tuzhilin 2005]: the content-based filtering approach, the Collaborative filtering (CF) approach and the hybrid approach. CF is the classic approach in recommender systems [Sarwar *et al.* 2001]. It exploits users' traces aiming at implicitly modeling the similarities between users according to their tastes. Content-based (CB) approaches exploit users and items descriptions to suggest related items; they are based on textual data or knowledge bases and are able

[1] https://www.mangerbouger.fr/l-essentiel/les-recommandations-sur-l-alimentation-l-activite-physique-et-la-sedentarite

to explain the suggestions made from items descriptions [Ricci *et al.* 2015].

Both approaches have limitations. CB approaches generally do not take into account the quality of the items in the recommendation process [Lops *et al.* 2019], while CF suffer from the cold start problem, *i.e.* the difficulty of dealing with new products and new users [Ricci *et al.* 2015]. Hybrid approaches have the advantage of overcoming these limitations, taking advantage of each system: both preference information from CF and contextual information specific to users or items provided by CB approaches [Delporte *et al.* 2014]).

In the last few years, recommender system architectures based on deep-learning have made it possible to exploit both user traces and item content in order to offer particularly powerful hybrid systems [Dong *et al.* 2017]. Deep learning architectures also allow to pull sequences and produce structured recommendations, such those presented in [Sutskever *et al.* 2014].

In addition to those systems, several approaches that use Knowledge Graphs (KGs) and ontologies have been proposed [Guo *et al.* 2020] to enhance recommendation algorithms [Zhang *et al.* 2016, Wang *et al.* 2019]. Other approaches exploit ontologies to provide a taxonomic classification of items. This allows indexing users by the entities from the ontology and weighting each dimension according to users' preferences. Most of these approaches either use lightweight ontologies or use ontologies as a source for a controlled vocabulary [Sheridan *et al.* 2019]. This data source, generally verified by experts, is characterized by its high quality: it is the ideal support to explain the recommendation in a reliable way [Catherine *et al.* 2017]. We must note that, recently, the capacity of explicability of a recommender system has become its main acceptance factor [Shin 2021], so this is an important aspect to consider when developing a recommender system.

To be helpful to public health, nutrition recommender systems should be able to induce a change in individuals' eating habits. This is challenging, thus food based recommendations should better be easy to follow [Bier *et al.* 2008]. Moreover, nutritionists stress the fact that, in order to make practical food-based recommendations, it is crucial to understand consumer behaviour. One fair assumption is that people are more likely to follow recommendations if these are acceptable from their point of view. We hypothesize that the user acceptance is a prerequisite for the compliance and could be improved by producing user-tailored recommendations that take into account dietary habits. On the long term, our objective is to build a nutrition recommender system taking into account dietary habits in order to encourage people towards healthier alternatives with high compliance.

In this chapter I present the work I have been doing with some students and colleagues towards the development of such a system.

- First I present the state of the art in recommender systems for nutrition and the data at our disposal (section 3.1).

- Second, I present the work related to the thesis of Sema Akkoyunlu (section 3.2) that, first, proposed to use some algorithms to find food items that can be eaten in the same eating context and, for this reason, are substitutable, then proposed to use the concept of context to group users with similar eating habits.
- Third, I present the ongoing EXERSYS project (section 3.3) that aims at taking advantage of the experts' knowledge expressed with a KG to ameliorate the recommendation.
- Fourth, I present some works we have done on considering a group of people eating together and group recommendation methods (section 3.4). The questions we asked are: How does the behaviour of the user change in presence of others? How does it change knowing what others have been doing?
- Finally, in section 3.5, I conclude the chapter with some ideas on future directions.

These works have been funded by different agencies: Danone Nutricia Research funded Sema Akkoyunlu's thesis, the EXERSYS project was funded by DATAIA that provided support for an internship and a PhD Thesis. Given the interdisciplinarity of the subjects, to develop these works, I initiated collaborations with different people, experts in different research domains.

3.1 Background

3.1.1 Recommender Systems for Nutrition

We can define various needs that a recommender system for nutrition must meet in order to ensure user satisfaction: the need to personalise the recommendation, the need to take into account a user's specific nutritional constraints (allergies, preferences, *etc.*), the need to provide a sequence of food items and the need to provide a meal (or a sequence of meals) that is coherent within the items that compose it.

In food related recommender systems, the recommended items can be recipes, food items or menus. Recipe recommender systems take advantage of users' past recipes ratings to propose items that they might like [Freyne & Berkovsky 2010, Harvey *et al.* 2013, Teng *et al.* 2012, Trattner & Elsweiler 2017]. Menu based recommender systems combine meals that users showed preference for with nutritional constraints based on the nutritional requirements for a user [Elsweiler & Harvey 2015]. Food item recommender systems [Massimo *et al.* 2017] are designed to learn users' tastes.

Most of these systems use popular recommendation algorithms often based on matrix factorization techniques which learn an embedding space for representing users and food items simultaneously. However, this representation does not take into account that food items are seldom consumed in isolation and that users' preferences for food items can

change in response to the other food items consumed (*i.e.* the dietary context) and to the context of consumption (*e.g.* eating croissant for breakfast is acceptable, but it is not for lunch). It seems necessary to take into account these aspects for increasing the efficacy of food related recommendation in real-life settings.

Context-aware recommender systems seem, therefore, to be an appropriate approach. However, modelling the context is highly dependent on the domain at hand. It is, thus, necessary to first model eating behaviours and understand how it is impacted by the context. This is what we attempted to do in [Akkoyunlu *et al.* 2017].

In nutritional science, dietary behaviours are modelled using two main types of methods: theoretical and empirical methods [Newby & Tucker 2004]. Theoretical methods use dietary indexes developed by research groups or agencies in order to rank the healthiness of eating behaviours. Indexes are constructed based on the current knowledge in nutrition but can also include current dietary guidelines and recommendations which are usually deduced from empirical research. However, in [Newby & Tucker 2004] it is pointed out that there can be a conflict when there is no scientific consensus about the definition of "healthy behaviour" that results in indexes that measure different definitions of this term. In empirical methods, there is no nutritional *a priori* about eating behaviours, this means that there is no definition about what a healthy behaviour is. Patterns are found with no nutritional *a priori*. In the works I present, we only focused on empirical methods, our goal has been to learn eating behaviours based on consumption data in an unsupervised way and, in a second moment, including expert's knowledge.

In the literature, two methods stand out for discovering eating behaviours: clustering and factor analysis. Cluster analysis aims at discovering groups of behaviours, while factor analysis seeks the most relevant factors to represent the behaviours. Clustering may use factor analysis as a preprocessing step. The K-Means algorithm is often applied to the matrix of consumption of food items directly [Reedy *et al.* 2009] or after dimension reduction using, for example, Principal Component Analysis (PCA) [Thorpe MG 2016] or Non-Negative Matrix Factorization (NMF) [Zetlaoui *et al.* 2011].

To our knowledge, there is no comprehensive review about the methods used for empirically deriving eating patterns [Newby & Tucker 2004]. Each study works on its own data-set and, most of the time, only one method of dimension reduction is applied for deriving eating behaviours. There is no apparent gold standard method, but the existing literature seems to favour the use of PCA.

These methods are reductionist: they only consider food items alone. Nutrition experts argue that this perspective may not be efficient for recommendation purposes: deeper and more complex information are needed [Wendel *et al.* 2013].

Opposed to this point of view, the holistic approach considers the diet as "a dynamic interaction of the parts of their synthesis" [Hoffmann 2003]. In the holistic approach, food items interactions are also used for eating behaviours modelling. In this approach dietary data are considered in a meal-based form. Meal pattern analysis provides more details regarding the way people compose their meals and could provide more insights

for characterising eating behaviours. This approach takes into account the complexity of the diet and aims at overcoming the limitations of the study of food items in isolation.

In [Woolhead *et al.* 2015], a meal based approach for discovering eating behaviours is introduced. Frequent item-sets techniques are used to generate a generic meal classification to derive 63 generic meals across all meal types. For each subject, mean daily intakes of energy percentage contribution of each generic meal type is computed, then PCA is applied to discover eating behaviours. Authors themselves argue that this methodology induces a subjective classification. Besides, relying on frequent item-sets to code meals may overlook infrequent eating patterns at a population level but frequent at an individual level, discarding these patterns as noise. This shows the necessity of an adequate representation of meals.

Developing a food recommender system that takes into account meals and their context, and not only food items, requires to meet two main challenges: (1) finding a proper meal description model in which distances between meals can be computed and (2) discovering an adequate way of aggregating several meals for computing distances between users in order to discover clusters of eating behaviours. This is what we attempted to do in [Akkoyunlu *et al.* 2018].

As a first step in the EXERSYS project we integrated the expert's knowledge to the approach presented in [Akkoyunlu *et al.* 2018] to find meaningful groups of users and provide each user with an informed recommendation that takes into account possible constraints (allergies, believes, preferences, ...) that may concern him. In the EXERSYS project we considered also the dynamics of the consumption. Our aim is not to provide a food item recommender system, not a meal recommender system but a menu recommender system. We are currently researching methods to model the sequence of consumption in all of its complexity.

3.1.2 Food Data

3.1.2.1 Consumption data

Several dietary assessment methods are available to collect data about eating habits or consumption. Those are: food frequency questionnaire (FFQ), 24-hour dietary recall (24HR) and food diaries.

- FFQ are questionnaires on the frequency of consumption of certain food items. They are tailored by research groups with a specific aim in mind. They are easy to implement and cost-effective however, their accuracy is not enough for recommendation purposes.

- 24HR method is an interview on the consumptions of a single day. It requires 30 minutes, it is rather precise but one day of consumption per user is not sufficient to learn the preferences of the user.

- Food diaries are a prospective openended food consumption assessment method where consumers write down all the food items and beverages consumed over a specific time period [Shim JS 2014]. Quite often, the time period goes from 3 to 7 consecutive days. The main advantages are that the whole process can be automatized, it is adapted for recommendation purposes and it provides several days of consumption, in this way, changes in diet can be captured.

The French INCA studies are an example of food diaries, carried out every 7 years: INCA1 (1998-1999), INCA2 (2006-2007) and INCA3 (2014-2015). These studies provide, at a given moment, a snapshot of the food consumption habits of the French population. Combined with monitoring plans and databases on the composition of foods, these data make it possible to know the intake of beneficial substances present in our diet (vitamins, essential fatty acids, *etc.*) as well as exposures, *i.e.* ingested doses of harmful substances likely to be present in foods (heavy metals, pesticide residues, toxins, *etc.*). INCA studies also contribute to assessing the impact of public health measures taken in the food domain.

INCA2[2] is the result of such a survey conducted during 2006-2007. Individual 7 consecutive days food diaries are reported for 2624 adults and 1455 children over several months taking into account possible seasonality in eating habits. In this data-set, a day is composed of three main meals: breakfast, lunch and dinner. The moments in between are denoted as snacking. For the main meals, the location (home, work, school, outdoor) and the companion (family, friends, coworkers, alone) are registered. The 1280 food entries are organized in 44 groups and 110 subgroups of food items. As a first step of this survey, different information about the individuals who participated in the study were collected: demographic and socioeconomic, food choice criteria, food preparation and storage, lifestyle habits, state of health, attitudes and opinions regarding food, consumption of food supplements...

For the estimation of nutritional intakes, each of the 1280 food entries is associated with a nutritional vector from the national database of the Food Quality Information Center (CIQUAL). The CIQUAL[3] composition table provides the contents of lipids, fatty acids, carbohydrates, total sugars and profile of individual sugars, proteins, salt, vitamins and minerals of more than 3185 foods, representative of those consumed in France. Data on nutrients present on the nutritional labeling of processed foods, collected by ODALI (nutritional section of the Food Observatory)[4], are also integrated into this database.

In 2014, the INCA3[5] study has integrated numerous new features and improvements as part of a procedure to be harmonized with other databases at the European level. Improvements are, for example, the inclusion of children under 3 years old, the study

[2] https://www.anses.fr/fr/content/inca-2-les-resultats-dune-grande-etude
[3] https://ciqual.anses.fr
[4] https://odalim.inrae.fr/
[5] https://www.anses.fr/en/content/raw-data-ansess-inca-3-study-now-available

of food consumption from organic farming or personal production, as well as a more precise food description system which makes it possible to refine estimates of nutritional intake and risk assessments on various themes (packaging materials, consumption of raw foods, *etc.*). This methodological change does not make it possible to precisely study changes in food consumption, energy and nutritional intake between the INCA2 and INCA3 studies.

At a different degree of involvement, 7566 adults and 6775 children participated in the INCA3 study. Information on the individuals' food consumption (with the 24HR method) were collected over two or three non-consecutive days (two weekdays and one weekend day) spread out over three weeks minimum. Moreover, a food frequency questionnaire to determine eating habits over a longer period was administered to the participants of the study. In this way, the INCA3 data-set collects data over a longer time period (and not only one week), it collects data about the food consumption and the food habits of a larger number of participants but it cannot represent the sequentiality of the food intake as the INCA2 study does.

To harmonise the study at the European level, each food entry in the INCA3 data-set is associated with a code in the FoodEx2[6] data-set, a standardised system for classifying and describing food. FoodEx2 consists of descriptions of a large number of individual food items aggregated into food groups and broader food categories in a hierarchical parent-child relationship. The current version of the FoodEx2 data-set has seven food hierarchies. The FoodEx2 nomenclature is the standard nomenclature for food data-sets in Europe.

The way the INCA3 data were collected makes the data-set less suitable for our purpose of developing a nutrition recommender system. We aim at providing a recommender systems able to suggest sequences of menus and this is not possible starting from sporadic food consumptions of the users. For this reason, our preliminary works are based on the INCA2 data-set. One of the results of Ayoub Hamal's internship was the association of the INCA2 data with the FoodEx2 nomenclature, this provides a first step towards the unification of the two data-sets for a more complete learning and recommendation.

These data-sets provide clean access to user consumption data enriched with dish compositions and quantitative nutritional data on ingredients: they are a rich and deep resource associated with an almost perfect data quality. These resources will allow to initialize the system and to anchor the explanations associated with the suggestions. To gain in scope and coverage, a menu recommender system must also consider more massive and qualitative sources like online web communities gathering cooking enthusiasts, such as https://www.marmiton.org or the data-set presented in [Achananuparp & Weber 2016]. These data pose a challenge of information extraction and synchronization of sources in terms of ingredients, preparation methods and

[6] https://www.efsa.europa.eu/en/data/data-standardisation

dish names, this is the reason we started working on the INCA2 data-set. It is our intention to integrate those data in futures studies.

3.1.2.2 Ontologies in the food domain

There exist in the literature, several ontologies related to food [Snae & Bruckner 2008, Boulos *et al.* 2015, Haussmann *et al.* 2019a, Tumnark *et al.* 2019, Caracciolo *et al.* 2023]. While these ontologies can be used to model a particular aspect of food items and to recommend a particular item based on what a user likes, they are not suitable to recommend recipes or menus.

Different cooking ontologies exist (see [Min *et al.* 2022] for a survey) that focus on the cooking act. They could model the cooking process, recipes or integrate cooking methods and instruments [Nanba *et al.* 2014, Després 2016, Singh & Deepak 2022]. Those may be used in a recommender system to suggest recipes based on what one prefers but they are not built to model user's consumption to learn users's preferences and suggest sequences of menus, as we want to do.

FoodOn [Dooley *et al.* 2018] is an ontology built to define all parts of animals, plants and fungi which can have a food role, as well as derived food products and the processes used to make them. The purpose that aims the project behind this ontology is to develop a semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes. For these reasons, FoodOn is composed of multiple (continuously growing in number) facets dedicated to terms focusing on a particular food subdomain. The aim of the FoodOn ontology is to put all the food ontologies and process together to have a unified vocabulary. Its goal is to represent the different aspects of the food domain but it was not built for recommendation purposes and, at the moment, it does not include a description of the user.

All these ontologies are not developed specifically for recommendation purposes and, if they are used at this purpose, they are used to select an item based on some stated preferences. Moreover, at the best of our knowledge, the ontologies presented in the literature are not linked to consumption data or sequences of consumption data. One of the aims of the EXERSYS project is to learn "informed" preferences from data structured by an ontology.

During Ayoub Hamal's internship we made a first step towards the modelisation of this ontology: we built an ontology able to model the information represented in the INCA2 study. More precisely, we developed a KG where both users and sequential consumption data are represented. This ontology could be enriched by domain experts' knowledge in the form of additional axioms and rules that are specific to recommendation of sequences of menus. This will allow the recommender system to consider, in a declarative way, the nutrition guidelines and recommendation context (*e.g.* health, ethics, economics) as we started to do in the EXERSYS project.

In the next sections I, first, present the works done in Sema Akkoyunly's thesis, then the EXERSYS project and the works done considering how the companion of the eating act can influence the eating choices.

3.2 The meal as the Context of a Food Item

Food based dietary guidelines are insufficiently followed by consumers. One of the principal explanations of this failure is that they are too general and do not take into account eating habits. Providing personalized dietary recommendations via nutrition recommender system can, hence, help people improve their eating habits. At this scope, understanding eating habits is a keystone in order to build a recommender system that delivers personalized dietary recommendations.

In Sema Akkoyunlu's thesis we made a first step towards this goal[7]. We explored food relationships on real-world data using the INCA2 data-set. We particularly focused on extracting food substitutions (*i.e.* food items that can replace each other) from consumption data. We considered that two food items can be substituted if they are consumed in similar food contexts. In [Akkoyunlu *et al.* 2017], we defined what a food context is and we introduced a measure of substitutability between food items based on consumption data that encodes the food context.

The ultimate goal of Sema Akkoyunlu's work was to build a food item based recommender system able to deliver messages such as "instead of eating x, eat y". In order to extract meaningful relationships between food items, in Sema Akkoyunlu's thesis, we considered contextual information.

Knowing substitutability relationships between items has been proven relevant for recommender systems [McAuley *et al.* 2015, Zheng *et al.* 2009]. Moreover, it has been shown [Adomavicius & Tuzhilin 2010] that context-aware recommender systems produce better recommendations than recommender systems that do not take into account the context.

We specifically investigated food substitutability. To do that, we defined the concept of *dietary context* as the set of food items a food is consumed with and the concept of *food intake context* as the setting of the food consumption. Our intuition is that two food items are substitutable if they are consumed in similar dietary contexts and that substitutability differs according to the food intake context.

[7] Sema's thesis has been co-supervised by Antoine Cornuéjols, Nicolas Darcel and myself. After two years and two publications, Sema Akkoyunlu decided to stop her thesis work. In this section I present the work we have done together, calling it thesis, even if a thesis was never produced.

3.2.1 Substitutability of Food Items

Let X be the set of food items. In [Akkoyunlu *et al.* 2017] we defined a meal as a collection of food items consumed at the same moment: $\{coffee, bread, jam, juice\}$ is a meal. The meal database DB is the set of all meals. We denoted $DB_{breakfast}$ the database of breakfasts and DB_{lunch} the database of lunches. Given a database of meals, we wanted to extract substitutability relationships based on the way people compose their meals. No nutritional information was used during this process. Instead, contextual information was used in order to extract meaningful substitutability relationships.

3.2.1.1 The Concept of Context

It is difficult to universally define the notion of context. In recommender systems, the context is usually defined according to the domain of application. We defined two types of contexts for the nutrition domain: the *dietary context* and the *food intake context*.

- The dietary context of a food item x is the set of food items C with which x is consumed; for instance, in the meal $\{coffee, bread, jam, juice\}$, the dietary context of $\{coffee\}$ is $\{bread, jam, juice\}$. We think that the dietary context is fundamental when seeking substitutes for food items because the way people compose their meals is intrinsically dependent on the relationships between the food items.

- The food intake context is the set of all variables that add information to the meal itself, such as the type of meal (breakfast, lunch, dinner, snack), the location (home, workplace, restaurant), the participants (family, friend, coworkers, alone). This corresponds to the notion of context more often used in context-aware recommender systems [Adomavicius & Tuzhilin 2010].

To the best of our knowledge, only one study tackled the subject of food substitutability based on real-world consumption data [Achananuparp & Weber 2016]. However, in [Achananuparp & Weber 2016] the food intake context information is not taken into account.

In the literature, there exist three paradigms for incorporating context in recommender systems: contextual pre-filtering, contextual post-filtering and contextual modelling [Adomavicius *et al.* 2022]. Contextual pre (or post)-filtering consists in splitting the data-set according to the contextual variables before (or after) applying the recommendation algorithm. Contextual modelling consists in incorporating contextual information in the algorithm.

In our framework, dietary context was used in order to model substitutability whereas the food intake context was used for contextual pre-filtering. Our objective was to investigate substitutability among food items based on the assumption that two food items are highly substitutable if they are (usually) consumed in similar dietary contexts and in the same food intake context.

Investigating all possible dietary contexts of a food item is computationally expensive because the number of possible dietary contexts is exponential in the number of food items and in the length of the dietary context. Instead of investigating all the dietary contexts of a food item, we decided to explore collections of meals that differ only by one item. We defined the dietary context of a meal database, or meal context C, as the intersection of a set of meals S_M such that:

$$len(C) = \max_{x \in S_M} (len(x) - 1) \tag{3.1}$$

We defined the substitutable set S_C associated to a meal context C as the set of food items such that the context C plus one item of S_C can be effectively consumed together. For instance, given the collection of meals

$$\{bread, jam, juice, coffe\}, \{bread, jam, juice, tea\}, \{bread, jam, juice, yagourt\},$$

the substitutable set of the meal context $C = \{bread, jam, juice\}$ might be $S_C = \{coffee, tea, yagourt\}$.

3.2.1.2 Mining Substitutable Items

To efficiently retrieve interesting sets of meal contexts and their substitutable set, we proposed an approach based on graph mining techniques. Let us denote the meal graph $G = (V, E)$ where V is the set of nodes representing meals from the database and E is the set of edges such that two nodes are connected if there is at most one item that changes between them. A meal should appear at least once in the database in order to appear as a node in the graph. Figure 3.1 is a simple illustration of a meal network.

In this way, the nodes of the substitutable set of a meal context are adjacent and they form a maximal clique. In our settings, discovering substitutable sets is similar to mining maximal cliques in a graph. To search for maximal cliques, we used the algorithm of Bron-Kerbosh [Bron & Kerbosch 1973].

We searched for cliques the intersection of the nodes of which defined a eating context. We denoted these cliques as substitutable cliques. However, we might encounter cliques that are uninteresting, as in Figure 3.2, where the intersection of the nodes is $\{A\}$, from this substitutable clique we could not derive a substitutable set. To avoid retrieving uninteresting cliques, we applied Algorithm 2 that filters out substitutable cliques.

Input: a clique : *clique* **Result:** a boolean : b
context = get_Context(*clique*)
lenmax = max(*len*(*x*) **for** *x* **in** *clique*)
if *lenmax - len(context) = 1* **then**
  b = TRUE
else
  b = FALSE
end

Algorithm 2: Finding substitutable clique

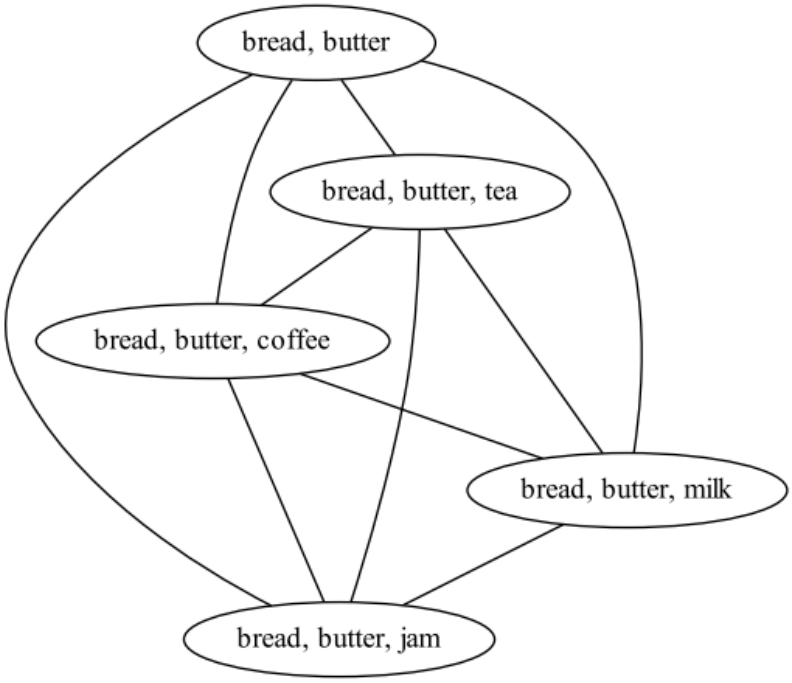


Figure 3.1: Example of a simple meal network.

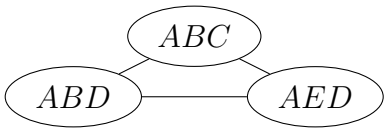


Figure 3.2: Example of an uninteresting clique

For instance, when we apply our algorithm to the example of Figure 3.1, we get that this graph is a substitutable clique. The context is $\{bread, butter\}$ and the substitutable set associated to this context is $\{coffee, tea, milk, jam, nothing\}$. In this particular case, it is possible to substitute an item by nothing because $\{bread, butter\}$ can be consumed as such.

3.2.1.3 Computing a Substitutability Score

We observed that substitutability is not a binary relationship because if two items are very often consumed together, they might be associated and, for this reason, they could be less substitutable. Therefore, we defined a function to quantify the relationship of substitutability that incorporates the associativity as well. We made the hypothesis that two items are highly substitutable if they are consumed in similar dietary contexts. We defined a substitutability score such as:

1. two items are highly substitutable if they are consumed in similar contexts;
2. two items are less substitutable if they are consumed together;
3. substitutability is a symmetrical relationship.

For an item x we called C_x the set of meal contexts in which x is a substitutable item. If the cardinality of C_x, $|C_x|$, is high, then x is substitutable in many meal contexts. For two items x and y, the condition (1) is described by the cardinality of the intersection of C_x and C_y. If $|C_x \cap C_y|$ is big, then x and y are consumed in similar contexts. We denoted $A_{x:y}$ the set of contexts of x where y appears :

$$A_{x:y} = \{C \subseteq C_x | y \in C\} \tag{3.2}$$

The cardinality of $A_{x:y}$ denotes at which degree y is associated to x.

Inspired by the Jaccard index [Jaccard 1912], we proposed the *substitutability score*, $f(x, y)$:

$$f(x, y) = \frac{|C_x \cap C_y|}{|C_x \cup C_y| + |A_{x:y}| + |A_{y:x}|} \tag{3.3}$$

The substitutability score equals 1 when x and y appear in exactly the same contexts and $A_{x:y} = A_{y:x} = 0$. If x and y are never consumed in the same contexts, then, the score equals 0. The higher $|A_{x:y}| + |A_{y:x}|$ is, the higher the association of x and y is and the smaller the score is.

3.2.1.4 Experiments

We conducted some experiments on the INCA2 data-set. In order to capture inter- and intra-groups substitutability relationships, we have chosen to consider the medium level of hierarchy of the data-set, the subgroups separation. Only adults' consumptions were considered. All meals have been collected in a meal database, DB_{meals}, regardless the type of meal, next this database has been splitted according to contextual information. We compared the results of our methodology on three data-sets: $DB_{breakfast,\, lunch}$, $DB_{breakfast}$ and DB_{lunch}.

Applying our algorithm to $DB_{breakfast}$ yielded 2368 contexts. Some of these and their substitutable sets are given in Table 3.1. Our experiments showed coherent results: for example, either bread, rusk or viennoiserie can be consumed for breakfast with coffee, sugar and water.

Table 3.1: Results of context and substitutable set retrieval for breakfasts

Context	Substitutable set
coffee, sugar, water, butter	bread rusk viennoiserie
tea/infusions, donuts	yogurt sugar jam/honey nothing

Table 3.2: Top 3 substitutable items for several items for breakfast and lunch

	Breakfast and Lunch		Breakfast		Lunch	
Food item	Substitute items (ordered by score)	score	Substitute items (ordered by score)	score	Substitute items (ordered by score)	score
Bread	Rusk	0.2234	Rusk	0.3716	Fruits	0.0497
	Viennoiserie	0.1359	Viennoiserie	0.2010	Yoghurt	0.0490
	Cakes	0.0745	Cakes	0.1243	Potatoes	0.0468
Coffee	Tea	0.2799	Tea	0.4219	Sodas	0.065
	Cocoa	0.1729	Chicory	0.2550	Yogurt	0.0642
	Chicory	0.1486	Cocoa	0.2255	Fruits	0.0633
Tea	Coffee	0.2799	Coffee	0.4219	Cakes	0.0536
	Cocoa	0.1721	Chicory	0.1965	Viennoiserie	0.0417
	Chicory	0.1289	Cocoa	0.1462	Coffee	0.0412
Cocoa	Chicory	0.2171	Chicory	0.2211	Cereal Bars	0.25
	Coffee	0.1729	Coffee	0.2077	Preprocessed Vegetables	0.0526
	Tea	0.1289	Tea	0.1965	Hamburger	0.0256
Butter	Margarine	0.2413	Margarine	0.4030	Margarine	0.0602
	Honey/jam	0.0924	Chocolate spread	0.1240	Fruits	0.0431
	Chocolate spread	0.0786	Honey/jam	0.1175	Sauces	0.0431
Milk	Juice	0.1409	Yogurt	0.1815	Doughnut	0.0869
	Yogurt	0.1264	Juice	0.1504	Other milk	0.0666
	Sugar	0.1089	Tap water	0.1361	Milk in powder	0.0625
Wine	Sodas	0.0814	/	/	Sodas	0.0860
	Beer	0.0704			Tap water	0.0755
	Tap water	0.0412			Beer	0.0746
Pizza	Sandwich baguette	0.2429	/	/	Sandwich baguette	0.2810
	Other sandwiches	0.1729			Other sandwiches	0.2177
	Meals w pasta or potatoes	0.1513			Meals w pasta or potatoes	0.1658
Potatoes	Pasta	0.1111	/	/	Pasta	0.1142
	Green beans	0.0922			Green beans	0.0941
	Rice	0.0602			Rice	0.0616

We applied our algorithm to the three data-sets. The results are reported in Table 3.2. We obtained inter-group substitutions such as $\{potatoes \rightarrow greenbeans\}$ but also intra-group substitutions as $\{bread \rightarrow rusk\}$.

The substitutions found are consistent with regards to eating habits. Substitutes of drinks are also drinks: the substitutes of coffee are tea, cocoa and chicory. It is also the case for spreadable food items: the substitutes for butter for breakfast are spreadable items. No semantic information describing how a food item can be eaten is available in the data-set and yet, considering the dietary context helped us retrieving this kind of information.

Substitutions between food items of the same nutritional food groups were found as well. For instance, the substitutes found for potatoes are pasta and rice that all contain starches.

Applying the method to the databases splitted according to the contextual variable "type of meal", we obtained different substitutes and scores. Coffee can be substituted by tea, chicory and cocoa for breakfast whereas for lunch it can be substituted by sodas, yogurt and fruits. This is in line with the observation that food items are consumed differently according to the type of meal and the relationship of substitutability is, therefore, different too.

Differences of scale in scores are observed according to the variable "type of meal". It may be due to the fact that the diversity of food items consumed during lunch and dinner is higher than during breakfast[8].

3.2.2 Grouping Users Based on what They Have Eaten

In [Akkoyunlu *et al.* 2018] we proposed a new approach to model meal representation by applying the Doc2Vec algorithm [Mikolov *et al.* 2013b] in order to learn a meal embedding space. This allows the use of a cosinus similarity adapted to matrices to compute similarities between users and to infer clusters of users. We compared our method to the state of the art methods used in the nutrition science community.

3.2.2.1 State of the Art Methods

In eating behaviour science, researchers work mostly on food items. They transform food consumption data into matrices where the columns correspond to the frequency or the quantity of consumption of food items and the rows to users as shown in Figure 3.3.

Afterwards, they apply PCA or NMF [Lee & Seung 1999]. PCA consists in finding a set of linearly independent variables, called principal components, that capture as much as possible the variance of the data points. NMF is similar to PCA but imposes a non-negativity constraint on the parameters of the model. This is found

[8]More observations about the diversity of food items consumed during breakfast and during lunch and dinner are reported when presenting the internship work of Noémie Jacquet.

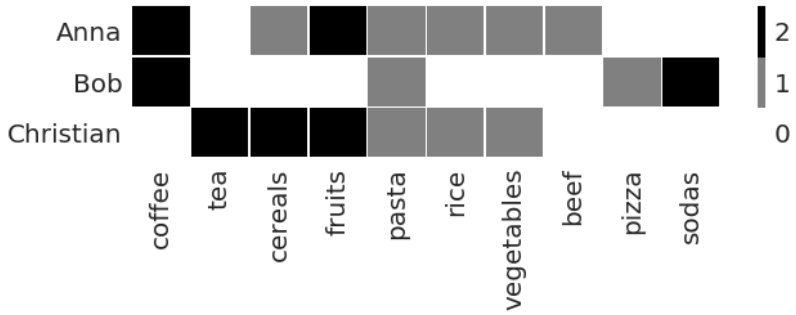


Figure 3.3: Matrix of consumption of the toy example.

useful in many domains such as signal processing and recommender systems, because more amenable to interpretation by experts [Luo *et al.* 2014]. Clusters of eating behaviours are, then, discovered by applying K-Means algorithm on the result of PCA or NMF [Zetlaoui *et al.* 2011]. In order to find the optimal number of clusters the silhouette coefficient is usually used [Rousseeuw 1987]. In [Akkoyunlu *et al.* 2018] we proposed a new approach based on the Doc2Vec algorithm.

3.2.2.2 Applying Doc2Vec to Users

The Doc2Vec algorithm [Mikolov *et al.* 2013b] learns distributed representations of arbitrarily large units of text such as sentences, paragraphs or documents. It has been proposed in two flavours: Distributed Bag Of Words (DBOW) and Distributed Memory version of Paragraph Vector (DMPV). DBOW is simpler than DMPV as it does not take into account the order of the words when learning the embedding space. It is the version that is suited for our task as, in [Akkoyunlu *et al.* 2018], we decided not to take into account the order of the food items in the meals. Besides, empirical evaluations of Doc2Vec showed that DBOW performs better than DMPV [Lau & Baldwin 2016].

The food based approach considers that a user is described by the frequency with which he consumed single food items. In our approach, a user is considered as a document where the food items eaten over a specific amount of time play the role of words.

Figure 3.4 is an illustration of what applying Doc2Vec algorithm on individual eating consumptions means. Individual consumption are fed in the model as documents. The result is an embedding space of users based on their eating consumptions which means that each user is described by a set of coordinates. In Figure 3.4, users are represented as vectors because similarity between users is computed with the cosine similarity, a metric commonly used in document retrieval. We computed the similarity matrix of users and we clustered users according to their similarity.

To cluster the users we used a spectral clustering algorithm, that is a method that exploits similarity measures by considering data points as nodes of a weighted connected

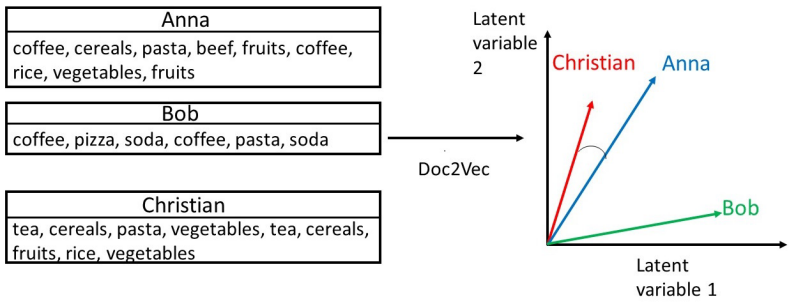


Figure 3.4: Application of Doc2Vec on user consumptions in the food based approach.

graph. Clusters are found by partitioning this graph based on the eigenvectors of the Laplacian matrix derived from the similarity matrix.

Choosing the optimal number of clusters is often a problem for clustering algorithms. There are several heuristics adapted for spectral clustering, we used the eigengap heuristic: the optimal number of clusters k is the number such that the difference between the eigenvalues of the similarity matrices $\lambda_{k+1} - \lambda_k$ is the largest [von Luxburg 2007].

This approach clusters users based on what they have eaten during a period of time. In [Akkoyunlu *et al.* 2018] we compared this approach to the state of the art and to another approach that clusters users represented in an embedding space for meals.

3.2.2.3 A Meal Based Method Using Doc2Vec

Learning an Embedding Space for Meals We observed that applying the Doc2Vec algorithm directly to users is against the philosophy of the holistic approach as it ignores interactions that may exist between food items in a meal. An elegant way of learning such interactions is to use the Doc2Vec algorithm to learn an embedding space of meals. We defined a meal as a combination of food items simultaneously consumed by one user at a single moment of consumption on one day. Meals are lists of food items. In our meal based approach, the objective is to be able to compute similarities between meals in order to compute similarities between users to derive clusters of users.

Indeed, the embedding is learned in such a way that similar meals are closer in the induced space as showed in Figure 3.5. In this embedding, each user is described by a matrix where the rows correspond to the meals he consumed and the columns to the coordinates of the meals in the Doc2Vec induced space.

Computing Distances Between Users Once the meal representation learned, the challenge became to compute a similarity between users. Mathematically speaking, this means to compute a similarity between matrices. We used the cosine ker-

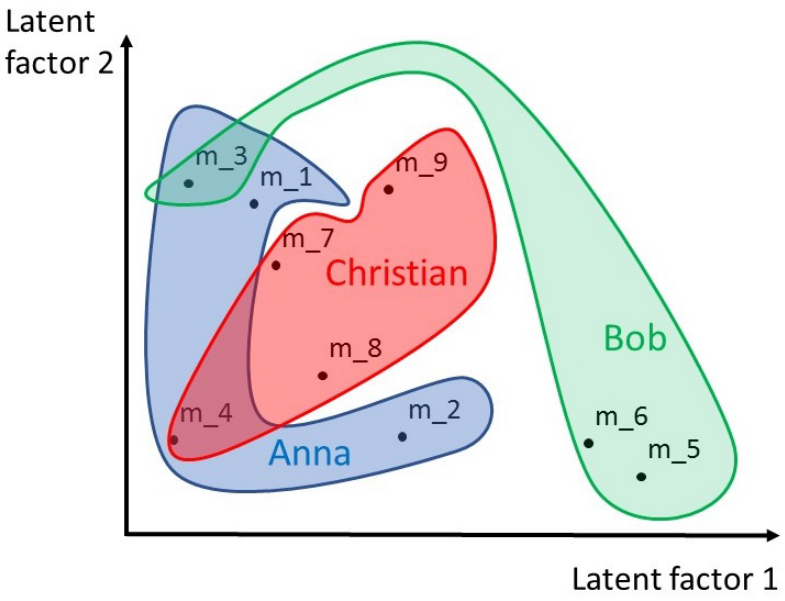


Figure 3.5: Application of Doc2Vec on meals in the meal based approach.

nel [Mijangos *et al.* 2017]:

$$\cos(A, B) = \frac{\langle A, B \rangle}{\|A\|_F \cdot \|B\|_F} \tag{3.4}$$

where A and B are two documents, $\langle , \rangle$ is the Frobenius inner product and $\|x\|_F$ is the Frobenius norm.

Using the Frobenius inner product it is possible to compare the similarity of the sentences to determine the similarity of the documents. Let us denote s_A and s_B the number of sentences in document A and document B respectively.

This formula implies that the cosinus similarity is computed between the first sentences of both documents then the second ones and so on until the $min(s_A, s_B)$-th sentences. If one document is longer than the other, the last sentences of the longer document are not taken into account for the similarity computation. For eating behaviour modelling, this means that two consumers are similar if they eat similar meals at the same moment of the day on the same day. This is a rather strong assumption concerning eating behaviour modelling, we accepted.

3.2.2.4 Experiments

We compared the performance of the Doc2Vec algorithm for clustering users (considering meals or not) to the PCA and NMF based methods. In our experiments we used the INCA2 data-set. We decided to work on the subgroups of the survey because the vocabulary is larger than that used for groups while having enough repetitions unlike

when considering food items. We did not impose the number of clusters to be the same for all the methods as we wanted to see if the number of clusters that each method discovered was different and if the clusters were overlapping or not.

PCA and NMF on Consumption Data The state of the art methods require the selection of two parameters: the number of components C of the reduction of dimensionality method and the number of clusters k. The number of clusters k was determined by using an internal clustering evaluation score: the silhouette score. The optimal number of clusters is found when the silhouette score is maximised. For PCA and NMF, we varied the number of clusters between 2 and 30 and computed the silhouette score. The score was maximised for $k = 9$.

The loadings factors for the PCA and for the NMF gave us an hint about the new representation space of the users. Figure 3.6 shows the loadings factors for the PCA according to food items. For ease of reading only food items whose absolute value of contribution to any factor is superior to 0.005 were displayed. NMF factors are shown in Figure 5. The food items are displayed if their contribution to any factor was superior to 0.3.

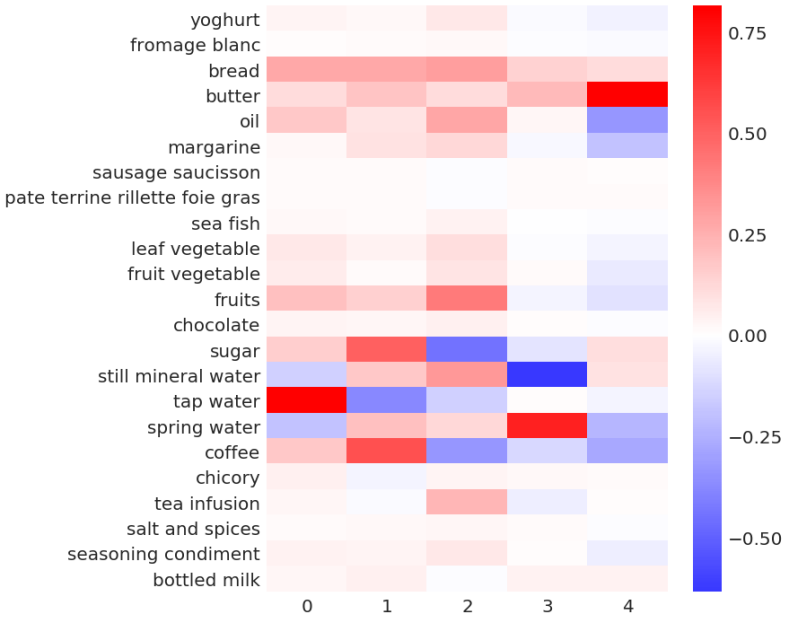


Figure 3.6: Factor loadings of PCA: explaining the new representation space.

Doc2Vec on Users We constituted the corpus of the document aggregating the food item consumption per user, each user constituting a document. We used the Gensim implementation of Doc2Vec in order to learn our model. The corpus contained

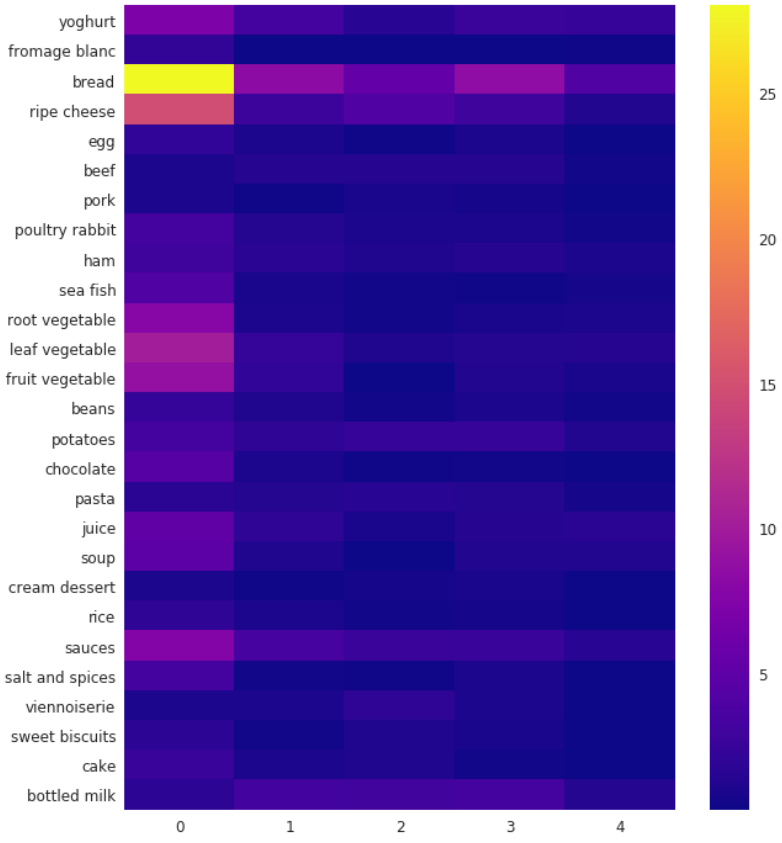


Figure 3.7: Factor loadings of NMF: explaining the new representation space.

Table 3.3: Comparison of clustering results with Adjusted Rand Index

	FOOD BASED			MEAL BASED
	PCA	NMF	Doc2Vec users	Doc2Vec meals
PCA	1	0.93	0.14	0.017
NMF		1	0.13	0.018
Doc2Vec users			1	0.013
Doc2Vec meals				1

2624 documents. After learning the model, we computed the cosinus similarity of users and performed spectral clustering. Using the eigengap heuristic, we found the optimal number of clusters corresponding to 5 clusters of users.

Doc2Vec on Meals We gathered the corpus of meals by aggregating the food items consumed at the same moment of consumption, at the same day, by the same user. The corpus was constituted by 37283 unique meals. A meal embedding was learned using the Gensim Doc2Vec implementation. For each user, we computed the vector of each of his meals, obtaining the user matrices. Applying the cosine kernel, we obtained the similarity matrix between users. We found clusters of users using the spectral clustering technique where the number of clusters was determined by the eigengap heuristic. We found 3 clusters.

Comparison of the Clustering Results We compared the clustering results to determine in which cases a food-based approach was adequate and the contribution of a meal-based approach. In order to compare the agreement between clustering results, we computed the Adjusted Rand Index (ARI) [Rand 1971]. This is a popular measure which consists in computing the agreement between two partitions. It is recommended for cases where the number of clusters is different, which is our case. The ARI takes values in [-1, 1], 1 meaning that both partitions agree, values close to 0 mean that the partitions do not agree.

Table 3.3 shows that no matter the factorization method used before the clustering step, the clustering results are very similar according to the ARI. This means that the choice of the factorization method for clustering users based on their food consumptions is not primordial.

However, as shown in Figures 3.6 and 3.7, the eating behaviours discovered are different. The coefficients of PCA can be interpreted as consumptions when positive and non consumption when negative. For instance, the eating behaviour 0 consists in drinking tap water but not spring or mineral water. We can also extract information such that those who consume coffee do not consume tea and vice versa. On the opposite, the coefficients of NMF are strictly positive, hence the interpretation only concerns

food consumptions. For instance, the eating behaviour 0 consists in eating all types of vegetables. The extracted eating behaviours are different according to the method of reduction of dimensionality used.

We reported, in Figure 3.8, the repartition of users in clusters according to the different methods. From one method to another, the number of clusters as their dimension is very different.

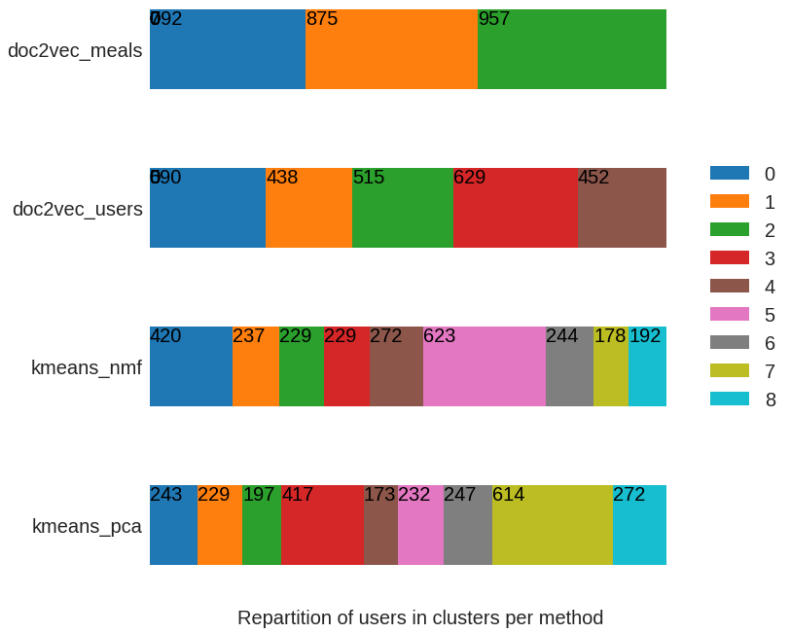


Figure 3.8: Repartition of users in clusters per method.

It is in the meal based approach that the number of clusters is the smallest. This shows that consumers of this data-set described by the way they compose their meals are less diverse as we only find 3 clusters. This result should be interpreted in the light of the assumption made about eating behaviours. We considered that two consumers are similar in the meal based approach if they consume similar meals on the same moment of the day on the same day, a strong assumption on 7-day food diary data. This may lead to more or less low values of similarity overall between users yielding in lesser clusters.

We applied the Doc2Vec algorithm directly to users in order to challenge the state of the art methods in food based approaches as we wanted to see how Natural Language Processing (NLP) method performed on this task. The number of clusters using the Doc2Vec algorithm on users yielded a smaller number of clusters and clustering results are rather different.

3.2.3 Remarks

In [Akkoyunlu *et al.* 2017] we proposed a score of substitutability based on consumption data with the assumption that two items are substitutable if they are consumed in similar contexts. This score can be used in a recommender system together with other scores such as a nutritional score that takes into account the nutritional contribution of the substitution and a user preference score.

In [Vandeputte *et al.* 2023] some experiments were conducted to prove the acceptability and the plausability of this score. In a first experiment a panel of humans was asked to tell if some proposed substitutions were suggested by a human or a machine. Generally the participants well identified if the substitutions have been made by a human or an artificial intelligence. In another experiments the participants were asked to tell what they were going to eat the day after and a coaching system, that uses the substitutability score presented in [Akkoyunlu *et al.* 2017], proposed a substitution that they were asked to accept or not. Of the 162 interaction outcomes between the participants and the coach 74 of them resulted in the acceptance of a recommendation, reflecting an overall average acceptability of 46%.

In [Akkoyunlu *et al.* 2018] we explored user modelling in food consumption for clustering users for recommendation purposes. We proposed a new food-based approach by considering food consumptions as textual data and learned an embedding model with the Doc2Vec algorithm. We observed that the application of Doc2Vec to user food consumption is adequate for user clustering, however it is not adapted for extracting eating behaviours. We argued the importance of having an holistic approach towards nutrition in order to make acceptable recommendations. We proposed a new meal based approach which consisted in learning a meal embedding space and then computing user similarity based on their meals' similarity.

The notion of dietary context defined in [Akkoyunlu *et al.* 2017], was implicitly modelled by the Doc2Vec algorithm in [Akkoyunlu *et al.* 2018]. In the internship work of Noémie Jacquet we continued the investigation of the Doc2Vec algorithm for the analysis of food consumption data and extend it taking into account experts' knowledge.

In [Akkoyunlu *et al.* 2018] we used Doc2Vec in two different ways:

1. we assumed the document to be all the consumptions and phrases to be the single user's consumptions along all the analysed period; in this case the Doc2Vec algorithm automatically found distances between users expressed by their food-consumptions;

2. we assumed each phrase to be a meal of a user, a user is, then, represented as a matrix which rows are the meals he has consumed expressed by their coordinates in the space induced by the Doc2Vec algorithm.

Differently from [Akkoyunlu *et al.* 2018], in Noémie Jacquet's work, the Doc2Vec algorithm was applied on food items and a distance between meals was defined as the

distances between the food items composing them. To respect the holistic approach that attributes importance to the relations between food items in a meal, we introduced experts' knowledge expressed by a KG. This is investigated in the EXERSYS project that financed Noémie Jacquet's internship, Ayoub Hammal's internship and the PhD thesis of Alexandre Combeau that started in October 2023[9].

3.3 The EXERSYS Project

In the past years, several recommender systems have been proposed as a promising solution to facilitate the adoption of an healthy diet [Vandeputte *et al.* 2022, Ge *et al.* 2015]. In food related recommender systems, the recommended objects can be recipes, food items, meals or menus. A menu is a *complex item* composed of different meals which are composed of different dishes, users' preferences for a dish can change in response to the other dishes consumed in the same meal, users' health situation (*e.g.* diabetes, arterial tension, allergies) may add constraints on possible dishes/ingredients to consider in a menu. Hence, recommending menus requires to check if the dishes are compatible and fitting the user preferences and his health constraints. Moreover, a food-related recommender system may consider the sequential aspect of the eating consumption (what we accept to eat today may be related to what we have eaten yesterday), while recommending an item to buy on a website is a one-shot recommendation, recommending a food-related item one needs to consider at which frequency the item should/could be recommended and when it has been consumed by the user. (*e.g.* depending on the user profile, bread can be recommended more often than fish and if fish has been eaten at the previous meal the system should probably not recommend it).

As introduced in [Akkoyunlu *et al.* 2017], a food-related recommendation must also consider the *context of the consumption*: user's preferences for food-related items may also be dependent on user's context that can be social (*e.g.* diner with friends, beer with friends on Saturday, fish on Friday), geographical and seasonal (*e.g.* recommending menus with seasonal ingredients). Finally, when knowledge and/or data are available, other constraints may be interesting to be included such as ecological and ethical aspects that may concern the origin of the ingredients, their environmental impact (*e.g.* use of phytosanitary products, deforestation) and if their production is ethics compatible.

On top of all of these characteristics, a menu recommender system should be the less invasive possible and avoid the *cold-start problem*. This is the problem of not having enough information for making good recommendations to a new user and requires specific techniques to target new users [Kluver & Konstan 2014].

[9]The EXERSYS project is a research project I am managing in collaboration with Nicolas Darcel, Stephane Dervaux, Vincent Guigue, Fatiha Sais and Paolo Viappiani. With the support of DATAIA it financed the 6 months internship of Noémie Jacquet, the 2 months internship of Ayoub Hammal and the PhD thesis of Alexandre Combeau.

A menu recommender system has to match the users' dietary behaviours to find similarities and give recommendations. Menus recommendation is a novel challenging topic; recent works [Elsweiler & Harvey 2015, Cholissodin & Dewi 2017] have dealt with the problem of recommending a menu but they do not consider the complexity of the problem such as the context of the consumption or the past meals consumed. Only recently the research community has started considering sequence-aware recommender systems [Quadrana *et al.* 2018, Guàrdia-Sebaoun *et al.* 2015] but, to our knowledge, there is no work on sequences of complex items recommendation (such as menus are).

The main objective of the EXERSYS project is to develop a new recommender system that is able to suggest menus for users while considering three important aspects: official nutrition and healthy guidelines, user preferences and the user eating history. Moreover, such a system should be multi-scale and structured, computing profiles and providing suggestions for: ingredients, dishes, meals and menus for a day or weeks.

Based on the works done in Mélanie Münch's thesis [Münch *et al.* 2019a], the focus of the EXERSYS project is to tackle these challenges by considering an hybrid approach that merges machine learning and KGs. Combining machine learning and KGs has gained great interest in both communities. On one hand, deep learning techniques are more and more used in KGs especially for KG refinement [Nathani *et al.* 2019] thanks to the success of new KG embedding techniques (see [Zhang *et al.* 2019] for a survey) that exploit KG embeddings in a low dimensional space to measure the semantic similarity between entities. On the other hand, KGs and ontologies are used in machine learning for both injecting domain constraints and contextual knowledge in machine learning models [?] as well as for black-box models explainability [Ma *et al.* 2019]. KGs can bring to the recommender systems several benefits [Wang *et al.* 2018]: (i) the KG can introduce semantic relatedness among items to find their latent connections and improve the precision of the recommended items; (ii) various types of relations in the KG are helpful to extend a user's interests and increase the diversity of the recommended results and (iii) the KG can bring explainability to the recommendation via the connection between users' historical records and the recommended items.

A recent work, FoodKG [Haussmann *et al.* 2019b], combines machine learning and KGs for food recommendation. The authors use (1) the reasoning capabilities of KG for inferring alternative ingredients and (2) the latent semantics of those in the form of word embeddings using the Word2Vec algorithm to determine the best alternatives to specific recipe ingredients to meet user needs. Our aim is to develop a novel approach that combines machine learning and KGs in a way to provide diverse recommendations of sequences of meals and not only single isolated food items by exploiting heterogeneous sources where data descriptions can be complex, temporally ordered and may be limited in quantity.

The first recommender system we developed is the one resulted from the internships of Noémie Jacquet and Ayoub Hammal. Ayoub Hammal's 2 months internship main result has been a first attempt for the definition of a KG in the eating domain. During

her 6 months internship, Noémie Jacquet studied a recommender system that gives recommendations based on machine learning techniques. Those are "filtered" by the KG developped during Ayoub Hammal's internship to take into account diet and context constraints.

The combination of the two methods (learning users' food preferences based on machine learning techniques to deliver recommendations that are, then, filtered according to some rules defined on the KG) for generating a recommendation, enables to take into account both user's preferences and the nutritional balance of recommendations, whereas most existing food recommender systems take only one aspect into account [Trattner & Elsweiler 2017, Yera Toledo *et al.* 2019, Pecune *et al.* 2020, Silva *et al.* 2022]. Moreover, it allows to compensate for the weaknesses of CF approaches. Thanks to the KG, it is possible to define user profiles (based on their preferences or eating habits) and to attach a new user to a predefined profile after asking him a few concise questions. Finally it permits to avoid irrelevant recommendations for a user (*e.g.* offering meat to a vegetarian consumer), recommendations which could discredit the system and limit user support.

The combination of the two methods will also allow to consider the diversity of the recommended meals, by reasoning on the scale of a day or a week and to integrate the contextual elements of the meal, such as the place (home, restaurant), the time and the social context, which are strong determinants of taste and consumption, as pointed out in [Trattner & Elsweiler 2017].

During Noémie Jacquet's internship we applied existing machine learning algorithms to the food domain. At first, we applied the Word2Vec algorithm to find similarities between consumed food items for meals recommendation. A second result obtained during Noémie Jacquet's internship is the modelisation of the dynamics of the consumption to recommend meals. To do that, we proposed a method that learns Recurrent Neural Network (RNN) from consumption data.

3.3.1 The FilterCollab Model

The first result of Noémie Jacquet's internship is the FilterCollab model, a CF approach. It learns a food representation space and models users' preferences in this space. This model has been proposed, parameterised and tested on breakfast (a meal with less variety in terms of food items) and showed its limitations on more complex meals.

3.3.1.1 General Approach

The FilterCollab model is composed of 4 steps. (1) It uses the Word2Vec algorithm trained on individuals' meal sequences to find similar food items. We assumed that food items consumed in the same context are close in the representation space learnt by the Word2Vec algorithm and, for this reason, similar. (2) It creates categories of similar food

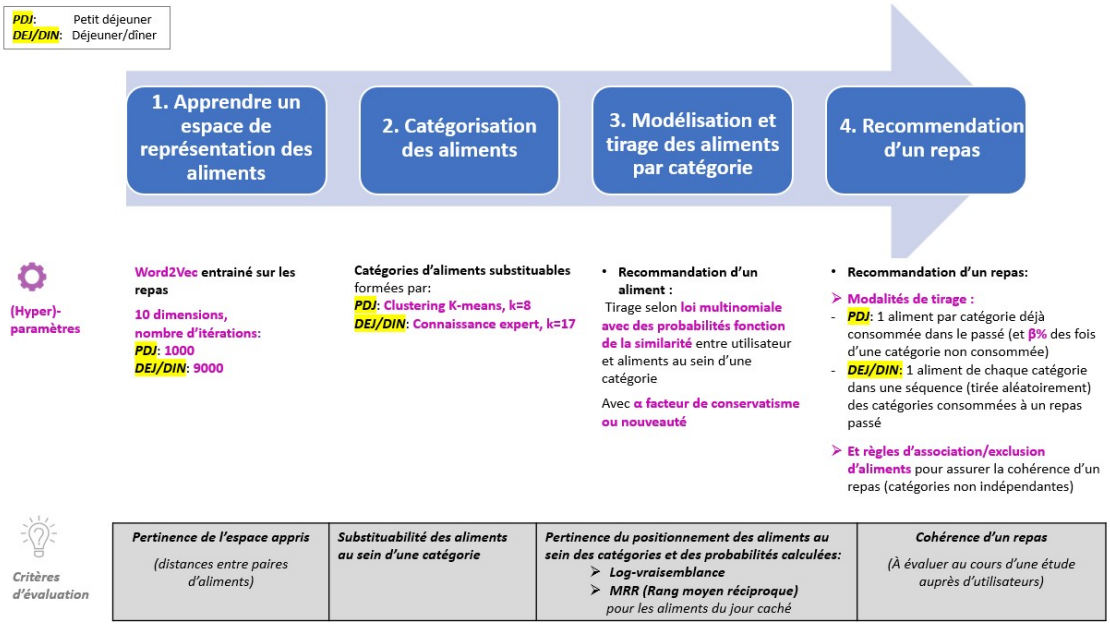


Figure 3.9: FiltreCollab parameters and evaluation criteria. Those are valid for breakfast and lunch/dinner if not specified.

items. (3) It draws food items within a category according to a probability distribution based on the distances (similarities) between users and food items within a category. We made the assumption that the distances learned when learning the food representation space made sense within each defined category. (4) It draws categories following some association/exclusion rules to form a coherent meal recommendation.

The overall approach is described in Figure 3.9. We applied FilterCollab first to breakfast and then to the more complex meals of lunch and dinner. We defined evaluation criteria at each step of the process to validate the choices of hyperparameters in the model and assessed its performance against recommender systems of reference.

Step 1: Learning the food representation space The algorithm used is the Word2vec algorithm [Mikolov *et al.* 2013b, Mikolov *et al.* 2013a]. This is a self-supervised machine learning algorithm that represents words as vectors in a space whose number of dimensions is defined by the user. This NLP technique is used for semantic analysis and text classification. The Word2vec algorithm learns vector representations of words from a text and predicts the context in which each word appears in the text. Vector representations of words have the interesting property that words appearing in the same context have "close" vectors in the learnt space and are characterised by high similarity. The measure of similarity between 2 words is computed as the cosine similarity between their 2 vectors.

Applying the Word2vec algorithm to sequences of food items consumed within a

meal by the individuals participating in the INCA2 study enabled us to identify food items within a meal that are similar because they are consumed in the same context ("surrounded" by the same foods). After excluding rare food items (consumed by less than 5% of individuals), we used the Word2Vec algorithm to learn a representation space for breakfasts (made up of 55 non-rare foods) or lunches and dinners (made up of 335 non-rare foods) from weekly meals consumed by an individual registered in the INCA2 study.

The main parameters to be set at this purpose are the number n of dimensions of the representation space (set to 10, the order of magnitude of the size of the smallest category), the number of iterations for learning (1000 for breakfast, 9000 for lunch/dinner) and the size of the window (set at the maximum length of a meal, excluding rare foods). After learning, each food item is represented by an n-dimensional vector in the representation space. The choice of the number of dimensions is a major criterion because a too high value for the parameter representing the dimensions can lead to over-fitting [Hung & Yamanishi 2021].

To evaluate the performance of this step we defined a specific criteria that evaluates the relavance of the representation space learnt. We compared the distances between pairs of food items in the representation space with regards to the INCA2 groups to which these items belong.

Step 2: Grouping food items We wanted to obtain categories of food items within a meal. To do that, at first, we used the K-means algorithm, for the data on which this approach failed, we created categories of food items using expert's knowledge.

The main parameter to be defined for the K-means algorithm is the number of groups K. We determined K using expert's knowledge (homogeneity of categories with respect to the INCA2 food groups) and to a lesser extent using the silhouette score[10].

We were able to obtain homogeneous categories by clustering with K-means on breakfast but not on lunches and diners data where clustering led to heterogeneous categories of food items. For this reason, we implemented an approach that uses an *a priori* based on expert's knowledge from the food groups defined in the INCA2 dataset to form categories. In this way, for lunches and dinners, a category of food items corresponds to one or more INCA2 food groups.

To evaluate the performance of this step we considered two kinds of scores. At first, we computed the degree of homogeneity of the categories in relation to the INCA2 food groups, then the similarity of the food items within a category was assessed by an expert.

Step 3: defining probabilities of food items We used the same approach for the breakfasts and the lunches and dinners. We chose to represent a category c (c

[10]The silhouette coefficient is calculated using the average intra-group distance, a, and the average distance to the nearest group, b, for each sample. The silhouette coefficient for a sample is $\frac{b-a}{max(b-a)}$.

belonging to C, the set of categories) by the average m_c of the food items belonging to this category.

We chose to model a user u:

- by its consumption frequencies f_c for each food category c; this is the number of food items in category c consumed by u, divided by the total number of food items consumed restrained to breakfast or lunch/dinner,
- within each category c, by the average u_c of food items of category c consumed during the meals consumed by u; u_c is the average of the food items consumed within the category c by user u,
- for each category c, for each food item k within category c, by the probability $P_{k|c,u}$ that k is recommended to u by randomly drawing in c according to a multinomial distribution.

We introduced a factor α to add the possibility to reduce or increase the difference of probabilities to recommend food items. By increasing α we aimed at favouring conservatism (recommending food items that the user already consumed); by decreasing α, on the contrary, we aimed at recommending food items that were new (never, or little, consumed by the user).

The probability that the item k within category c is recommended to user u ($P_{k|c,u}$) is defined by a softmax function:

- either as a function of the distance between the food item and the user (u_c is the average of the food items consumed within the category c by user u) if he has consumed an item in category c ($f_c \neq 0$),
- either as a function of the distance between the food item and the average of food items in the category c (m_c), if the user has not consumed any food in this category ($f_c = 0$).

Let I_c be the set of food items of category c and i a food belonging to I_c. S_{i,u_c} is the similarity between the user represented by u_c (the average of the food items consumed within the category c by u) and the food item i. S_{i,m_c} is the similarity between m_c (the average of the food items in c) and the food item i. The probability $P_{k|c,u}$ is defined as follows:

$$\begin{aligned} &\text{If } f_c \neq 0 : \\ &\qquad P_{k|c,u} = \frac{e^{\alpha * S_{k,u_c}}}{\sum_{i \in I_c} e^{\alpha * S_{i,u_c}}} \\ &\text{If} f_c = 0 : \\ &\qquad P_{k|c,u} = \frac{e^{\alpha * S_{k,m_c}}}{\sum_{i \in I_c} e^{\alpha * S_{i,m_c}}} \end{aligned} \tag{3.5}$$

S_{i,u_c} and S_{i,m_c} are cosine similarities. S_{i,u_c} is computed as follows:

$$S_{i,u_c} = \frac{u_c \cdot i}{\|u_c\|\|i\|} \tag{3.6}$$

with $u_c \cdot i$ the scalar product of vectors u_c and i, $\|u_c\|$ the norm of the vector u_c and $\|i\|$ the norm of the vector i. We are, thus, able to recommend a food item within each category by randomly drawing it according to a multinomial distribution with the probabilities defined above.

To evaluate the performance of the recommendation, we splitted the data-set into 2 sets: a training set consisting of the meals (breakfast or lunch/dinner) from the first 6 days of the collection and a test set consisting of the meals consumed during the last day. Since the individuals participating in the survey could have started the completion on any day of the week, the hidden day is a random day of the week. We measured the relevance of the recommendation (which is based on the training set) according to its ability to "predict" the hidden day. To do that, we used the log-likelihood score of the food items consumed during the hidden day because our model is probabilistic and the Mean Reciprocal Rank (MRR) because it is a rank metric commonly used in recommender systems[11]

For the log-likelihood score, we calculated the average log-likelihood score over all individuals according to our model. Being I_{uc} the set of food items in category c consumed by user u on the hidden day, C the set of breakfast (or lunch and dinner) food categories, $P_{c,u}$ the probability for user u to consume a food item of category c and $P_{i|c,u}$ the probability for user u to consume the food item i knowing it belongs to the category c, we computed the log-likelihood of consumption on the hidden day for a user u, $\mathcal{L}_u$ as

$$\mathcal{L}_u = \log\left(\prod_{c \in C}\prod_{i \in I_{uc}} P_{c,u} \cdot P_{i|c,u}\right) \tag{3.7}$$

With $P_{i|c,u}$ defined in equation 3.5 and $P_{c,u} = f_c * (1 - |C| * \beta) + \beta$, if $f_c \neq 0$ and $P_{c,u} = \beta$ if $f_c = 0$. β is an hyperparameter of the model that models the probability of an individual to consume an item in a category not yet consumed in the previous 6 days (arbitrarily set, initially, to a low value).

For the MRR, we computed the average over all individuals of the MRR according to our model: we considered MRR_u, the MRR of a user u, N the total number of food items consumed by u at breakfast (or lunch and dinner) on the hidden day, I_{uc} the set of food items in the category c consumed by u, $rang\ i_c$ the rank of the i^{th} food

[11] Different metrics have been applied, in the literature, to measure the performance of a recommender system. These generally reflect the error of the predicted scores (*e.g.* Mean Absolute Error (MAE) or Root Mean Square Error (RMSE)) or the quality of the list of the n first ranked items, *e.g.* MRR, recall, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG).

item consumed by the user on the hidden day within the recommendation made by the system for category c, we have

$$MRR_u = \frac{1}{N} * \left(\sum_{i \in I_{uc}} \frac{1}{rang\, i_c} \right) \tag{3.8}$$

The average scores for the two metrics were compared to the values found for reference models: random recommendation by category and recommendation of the most popular food items. In the case of the MRR metric, the score of our model was compared with a Bayesian Personalized Ranking (BPR) recommendation model, which is a matrix factorisation method [Rendle *et al.* 2009].

Step 4: recommending a meal There are several possible approaches to recommend a meal for a given user:

- we can draw a food item from the categories already consumed in the past and in $\beta\%$ of the cases from categories not yet consumed (favoured approach for breakfast);
- we can randomly draw a sequence of categories consumed during a meal in the previous days and draw a food within each of these categories (preferred approach for lunch and dinner).

The selection of a single food item within a category is justified by the frequency of consumption observed: zero or one food item from each category consumed by at least 98% of individuals for 7 of the 8 categories in the case of breakfast, for example.

For the draws, the categories are assumed to be independent, *i.e.* consumption of one category does not affect the probability of consumption of other categories (based on consumption history). To correct this approximation, we took into account the existing food associations within a meal by defining association rules (*e.g.* the consumption of bread if butter was consumed) or exclusion of certain categories according to their support, confidence and lift indices.

We defined an evaluation criteria for the performance of a meal recommendation. The usual criteria is the user's satisfaction following the recommendation, which we could not evaluate during the internship for lack of time. However, a major component of user's satisfaction is the overall consistency of the recommended meal. We, therefore, considered this as an evaluation criteria for step 4.

3.3.1.2 FilterCollab: Results and Discussion

The results are summarised in Figure 3.10 and detailed below.

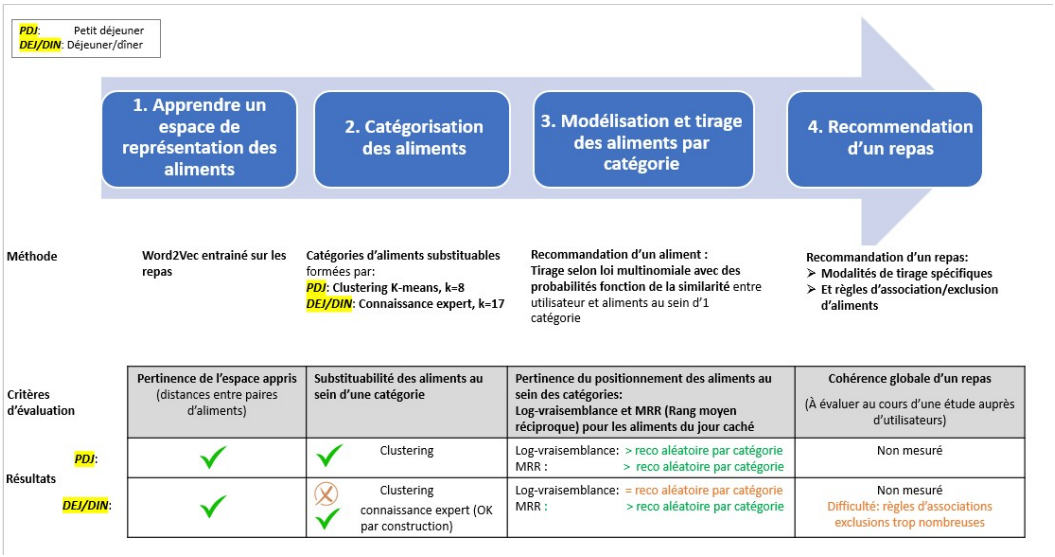


Figure 3.10: FilterCollab: evaluation criteria and results at each step. Comparison between breakfast and lunch/dinner

Steps 1 and 2, breakfast : The distances between pairs of food items showed the relevance of the representation space learned by the Word2Vec algorithm (distances interpreted with expert knowledge) and the K-means clustering combined with expert knowledge produced 8 food categories, the majority of which are similar food categories with a high degree of homogeneity compared to the INCA2 food groups. These categories are presented in Figure 3.11. The degree of homogeneity of the categories compared to the INCA2 groups is illustrated in the second column.

Steps 1 and 2, lunch and dinner: The distances between pairs of food items showed the relevance of the representation space learned with the Word2Vec algorithm (distances interpreted with expert knowledge) but clustering by K-means, whatever the number of categories K, lead to obtain certain categories that are heterogeneous compared to the INCA2 food groups or expert knowledge.

In the case of lunch and dinner, significant noise and a limited amount of data for learning meal sequences with the Word2Vec algorithm did not make it possible to correctly position all food items in the representation space and cluster similar food items. These results can be linked to (1) the great variety of food items consumed that results in a high number of possible combinations compared to the breakfast data-set, and (2) to the presence, in the lunch and dinner data-set, of underlined moments of consumption (starter, main courses, dessert) that structure the sequences of food items consumed but were not taken into account during learning.

We decided to integrate expert's knowledge to try to compensate for these limitations and created 16 categories of food items based on the INCA2 food groups and the study

N° de catégorie constituée par clustering (k-means) [1]	Groupes INCA[2] et pourcentages des aliments qui appartiennent à ces groupes dans la catégorie	Noms des aliments de la catégorie
0	**lait: 17%, eaux: 17%, boissons fraîches sans alcool: 33%, fruits: 33%**	['eau du robinet', 'eau de source', 'lait demi-écrémé uht', 'orange fraîche', "jus d'orange à base de concentré pasteurisé", 'pomme non pelée fraîche', 'lait écrémé uht', 'banane fraîche', "jus d'orange pressé maison", "pur jus d'orange pasteurisé", 'jus de fruits sans précision', 'kiwi
1	**pain et panification sèche: 100%**	['pain baguette', 'pain de campagne ou bis', 'pain courant français boule à la levure', 'pain complet ou intégral artisanal', 'biscotte classique type heudebert lu', 'pain grillé maison', 'pain de mie', 'autre biscotte', 'pain aux céréales artisanal']
2	**beurre: 50%, margarine: 50%**	['matière grasse allégée 60% m.g.', 'matière grasse allégée 55-60% m.g. riche en oméga 3 et 6', 'beurre doux', 'beurre demi-sel sel maxi 3%', 'matière grasse légère 38-41% m.g. à tartiner', 'beurre allégé sans précision']
3	**café: 67%, autres boissons chaudes: 33%**	['café au lait ou café crème ou cappuccino non sucré', 'café noir prêt à boire non sucré', 'thé infusé non sucré', 'café soluble reconstitué prêt à boire non sucré', 'poudre cacaotée et sucrée pour boisson au chocolat', 'café expresso non sucré']
4	**viennoiserie: 67%, chocolat: 11%, biscuits sucrés ou salés et barres: 11%, pâtisseries et gâteaux: 11%**	['brioche industrielle préemballée', 'croissant sans précision', 'pain au chocolat feuilleté artisanal', 'brioche sans précision', 'croissant au beurre artisanal', 'pâte à tartiner au chocolat et aux noisettes type nutella', 'madeleine', 'pain au lait artisanal', 'goûter sec fourré au chocolat type prince ou bn au chocolat']
5	**sucres et dérivés: 75%, aliments destinés à une alimentation particulière: 25%**	['sucre blanc', 'sucre blanc ajouté au service', "édulcorant à l'aspartame", 'sucre roux', 'aliment non codifié']
6	**sucres et dérivés: 71%, ultra-frais laitier: 14%, fruits: 14%**	['confiture ou marmelade tout type', 'clémentine ou mandarine fraîche', 'yaourt ou spécialité laitière nature', 'miel', 'confiture de fraise', "confiture d'abricot", 'confiture allégée']
7	**autres boissons chaudes: 50%, café: 50%**	['poudre soluble à base de chicorée et de café type ricoré', 'café soluble en poudre']

Figure 3.11: FilterCollab: the 8 breakfast food categories resulting after clustering with K-means, k mainly determined by expert knowledge.

of frequent patterns. We have, thus, grouped, within the same INCA2 food category, groups of food items that are similar from a nutritional point of view and which are not consumed together during a meal. For example, we grouped in the same category the 4 INCA2 food groups "meat", "chicken", "fish" and "eggs".

Step 3 We defined the probabilities of drawing a food item within a category from the distances in the representation space learned with the Word2Vec algorithm, whether the categories were learned by the K-means clustering (for breakfast) or by expert knowledge (for lunch and dinner). We observed that our approach has performances that are superior (in terms of total log-likelihood score and MRR score) to that of a model that randomly selects a category for the recommendation and of a method that recommends the most popular food item by category for both breakfast and lunch/dinner data. The MRR score of our approach is equal to the MRR of a BPR model.

Opposed to the performance of our method for the breakfast data-set, Word2Vec for lunch and dinner sequences only made it possible to learn partial information. It probably learned the relative positions of food items, *i.e.* their ordering within a category (because the MRR based on the ranking is superior to the random model per category), but not the "absolute" distances between food items (because the log-likelihood based on distances is equivalent to the random model by category). This seems to be linked to the greater variety of food items consumed at lunch and dinner and to the much greater number of possible food items combinations compared to breakfast.

Concerning the variety of food items consumed, the meal (lunch or dinner) of the hidden day is, on average, made up of 49% of food items already consumed in the past and 51% of new foods, while the hidden breakfast is, on average, constituted of 89% of food items already consumed in the past and only 11% of new food items. Due to this variety of consumption at lunch/dinner (that is intrinsically due to the limited data) we can, therefore, benefit much less of the information coming from the user's consumption history and the distances between food items were "less well learned" in the case of lunch/dinner, which explains the lower performance in terms of recommendation.

It can be noted that in both cases (breakfast and lunch/dinner), the performance in terms of MRR score of our model is comparable to the MRR score of the matrix factorization recommendation model (BPR). This reinforces our idea that the lower performance observed in the case of lunch/dinner is linked to the lack of data rather than to the method used, since two different CF methods (our model and BPR matrix factorization) lead to close average MRR scores.

It should be noted that in all the considered models we reasoned by categories. Those categories were defined in step 2, this questions the relevance of the categories formed and the possibility to optimize the performance by modifying those categories.

Step 4 The method for drawing food items from a category and then applying the rules of association/exclusion of food items defined by the study of frequent patterns of

step 4 are supposed to ensure a consistent meal recommendation. Even if the consistency of a recommended breakfast has not been quantified because we did not had time during Noémie Jacquet's internship to plan a user case type experiment, the few tests we accomplished tend to show that the association/exclusion rules identified are effective in ensuring the recommendation of a consistent breakfast.

On the other hand, due to their high number of food items, we were not able to identify the rules of association and exclusions between categories and/or food items that should be applied to ensure the coherence of a lunch/dinner recommendation. Ensuring the consistency of a meal recommendation resulting from food items draws within the different categories seems to be a major challenge.

3.3.1.3 FilterCollab: Final Remarks

We identified two ways to improve the performance of the recommendation for the lunch/dinner data-set. We could (1) Increase the size of the data-set with FilterCollab. We aim at increasing the quantity of training data and users' consumption history to reduce the proportion of new food items consumed on the 7th day. This supposes to obtain consumption data from new structured surveys collected in an identical way than the INCA2 survey (food items nomenclature, order of the registered consumed food items *etc.*). As seen, the INCA3 survey is referenced by a different food items nomenclature (CIQUAL for INCA2 and FoodEx2 for INCA3), one of the results of Ayoub Hammal's internship has been to link FoodEx2 and CIQUAL composition tables to have a common language between INCA2 and INCA3. A second way of improvement could be to (2) Exploit the order of food items consumption with a new model. The INCA2 survey provides the order of consumption of food items. With the GenSeqRNN model, a second model developed during Noémie Jacquet's internship and explained in Section 3.3.4, we hypothesized that we can enhance the sequential character of a meal and generate sequences of lunches/dinners by using RNNs.

3.3.2 A Knowledge Graph for the Eating Domain

The main goal of Ayoub Hammal's internship was to develop a knowledge base that we called BDNutri. This knowledge base is populated by data collected from heterogeneous sources and contains information on food items and their nutritional composition as well as, information relative to different consumer profiles such as consumption, allergies and different diets (vegetarian, vegan...). BDNutri is a set of KGs with the aim of modeling the knowledge we have on the food items, the users and the relations between them. During his internship, Ayoub Hammal, also, implemented a set of constraints extracted from expert's knowledge. These rules can be used to filter the recommendation generated according to the consumer's profile and his preferences.

The role of BDNutri is to formally describe and link the information that interact in

the recommender system. The advantage of a knowledge base over traditional databases is its inference capabilities: the inference is simplified thanks to the definition of a set of logical relations. This allows us to express constraints in a more declarative fashion rather than a procedural programming.

The KGs are expressed in the RDF framework, the SPARQL language and the RDF Query Language are used to query the KG and retrieve information. SHACL language is used to validate the conformity of the data against the defined constraints.

BDNutri is separated into units, with a large graph containing information about food and a separate graph for each consumer and their respective consumption.

Figure 3.12 summarizes the different OWL classes represented in BDNutri as well as their object and data properties. The main classes in this KG are those concerning *Food*, *Ingredients* and *Consumers*. Other information are attached to the ingredients; such as the allergies they cause and their origin. These information will come useful to define constraints based on allergies and/or vegetarian diets, for example. These information, not present initially in the INCA2 data-set, were fetched from the OpenFoodFact Ontology.

BDNutri separately models consumer profiles. Each profile has its own properties; for instance profiles associated to a certain allergy are linked to the Allergy class, whereas the vegetarian profile is linked to a set of prohibited origin of food items. Each consumer is linked to the consumption he made and the recommendations that were made to him.

The global KG is separated into elementary units, with one large graph containing food information (the top portion of the graph in Figure 3.12 from the Food class), and a separated graph for each consumer and their respective consumption. Only relevant parts of the graph are merged for inference purposes. This trick reduces not only the loading time of the graph (up to 10 times faster), but allows to diminish substantially (up to 100 times faster) the SHACL requests validation.

3.3.3 Making Informed Recommendations

As a first step, during Ayoub Hammal's internship, we used DBNutri to validate (or invalidate) the recommendations emitted by the FilterCollab approach thanks to the Python library *rdflib* that allows to query the RDF graph and, in this way, to link the recommendation resulted from FilterCollab and the KG. Given the difficulties we showed in the previous section to emit a recommendation for lunch and dinner, we tested our recommender system with constraints only on the breakfast data-set.

At inference time, the system fetches the food graph and the graph of the consumer concerned by the recommendation. It inserts the recommendation in the union of those two graphs and validates the resulting graph using a set of SHACL shapes.

A SHACL shape is a formal tool to ensure that a KG obeys a specific structure. It can be applied to both concepts and properties and generates detailed, customizable messages in case there is any violation. This helps improve the explainability of the

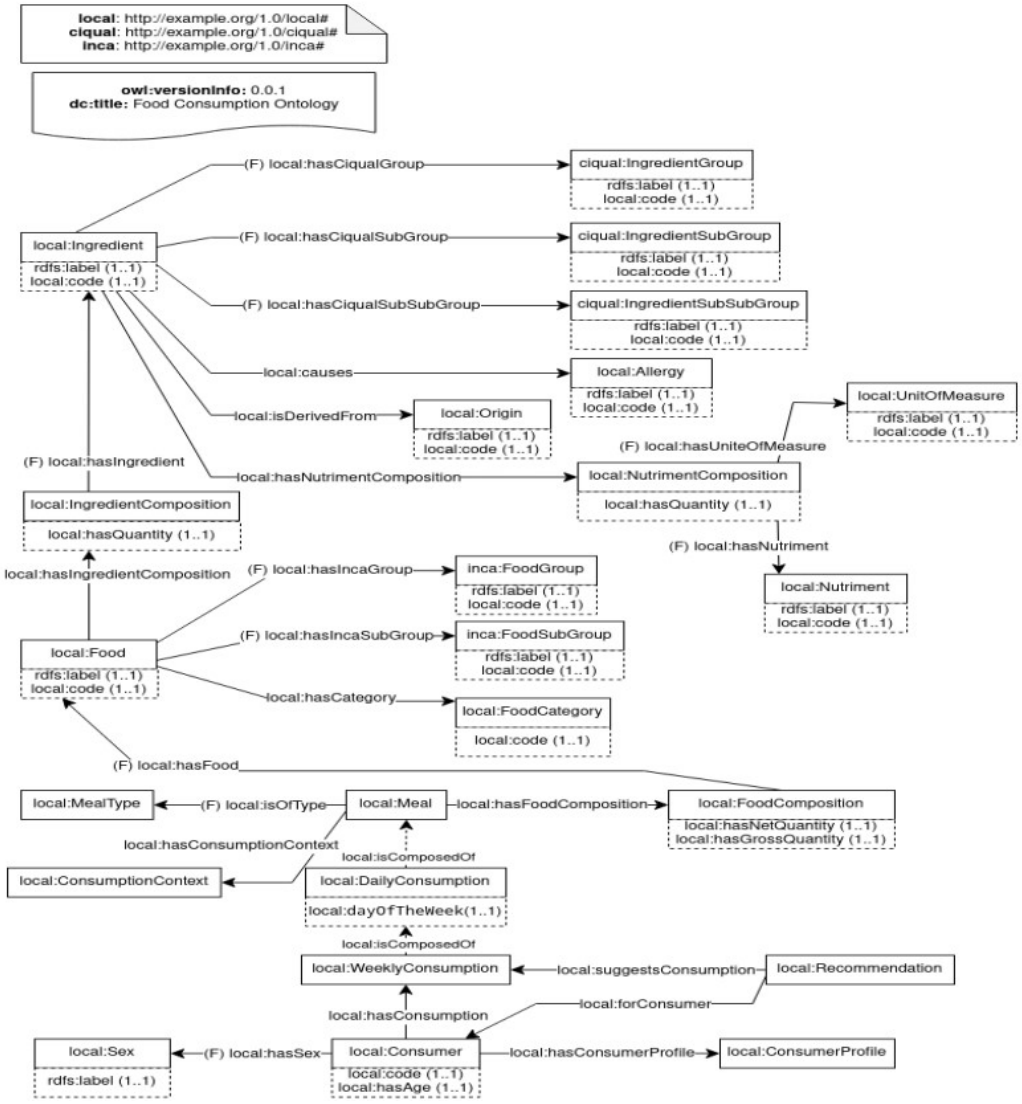


Figure 3.12: BDNutri schema.

recommender system and gives it more credibility.

Given a recommendation emitted by FilterCollab, we were able to verify three main types of constraints: expert's rules, nutriment scores and prohibited food. Expert's rules are a set of association, exclusion and cardinality rules; they are constructed from statistical observations drawn from the consumption data (for example bread and butter being always eaten together). Nutriment scores and prohibited food rules are profile specific rules, which means that they depend on the current consumer's profile. Nutriment scores are optimization constraints that rank a set of recommendations according to one or more objective functions (for example, an athlete has to maximize his protein intake). Prohibited food rules are a set of rules that verify if the consumer is allowed to eat the recommended food item, this filters out food items that cause allergies the consumer suffers from, for example.

In this way, our recommender system can issue a valid meal recommendation for a consumer with allergies or dietary restrictions applying the following principle: in case of non-compliance of the recommended food item linked to allergies, we draw a new food item within the same category as the prohibited item which had been recommended, in case of non-compliance of food items combinations or exclusions, we issu a new meal recommendation and we iterate till finding a recommendation that suits all the rules (fixing a maximum number of iterations).

Finally we were able to identify the percentages of problematic food items for each food category by type of allergy or dietary restriction as illustrated in Figure 3.13. This Figure shows the complexity of issuing a recommendation for profiles allergic to milk or avoiding dairy products.

These preliminary tests show the soundness of our approach that consists in filtering a recommendation emitted by a machine learning recommender system. These results are preliminary and the long-term ambition of the EXERSYS project is to use KGs also for other purposes, for example to integrate other contextual elements of the meal (time, place, *etc.*) or ensure dietary diversity on a weekly scale. In the EXERSYS project we would also like to integrate DBNutri into the machine learning based recommender system. This is something we plan to do during the PhD Thesis of Alexandre Combeau that started in October.

3.3.4 Recommendation by Sequence Generation

As an alternative to the FilterCollab model, we tested the learning of lunch/dinner sequences with an RNN model [Goodfellow *et al.* 2016], called GenSeqRNN. Unlike FilterCollab, this approach exploits the sequential nature of a meal. We did that with the aim of taking into account the sequential aspect present in the lunch and dinner data that may advantage the recommendation of sequences. RNNs are designed to take into account temporal dependencies in data by maintaining an internal memory. This makes them effective for tasks like machine translation and text generation. At the best of our

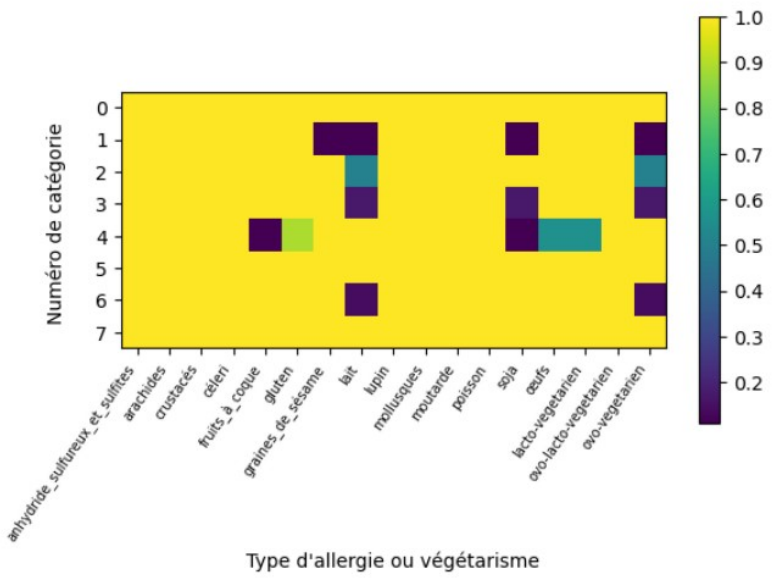


Figure 3.13: Percentage of food items, identified in each of the 8 breakfast categories, that are OK to be recommend in case of the given type of allergy or vegetarianism

knowledge, it was the first time they were used for meals recommendation.

3.3.4.1 Dynamics of Food Consumption Data

A menu is a complex item made up of different dishes that can be broken down into ingredients. In addition to this hierarchical representation, a sequence of menus is subject to a particular dynamic: the menus must be varied and their balance is judged over a certain period. Joint modeling of this hierarchy and dynamics is critical; in fact, constraints may apply to ingredients (allergies), cooking methods (frying), dishes (preference) or to the overall sequence, on a daily or weekly scale (respect for energy intake).

Thus, the application of sequential approaches in recommendation for nutrition poses a certain number of research questions, both on the modeling of data input to the system, on the learning of profiles and on the construction of a structured output:

RQ1 How can we characterize the sequentiality of the different meals of the day, from breakfast to dinner?

RQ2 Does sequential menu modeling make it possible to capture co-consumption (*e.g.* bread and butter for breakfast)?

RQ3 How to combine sequential modeling and user preferences?

RQ4 How to evaluate the quality of recommendations in a sequential context?

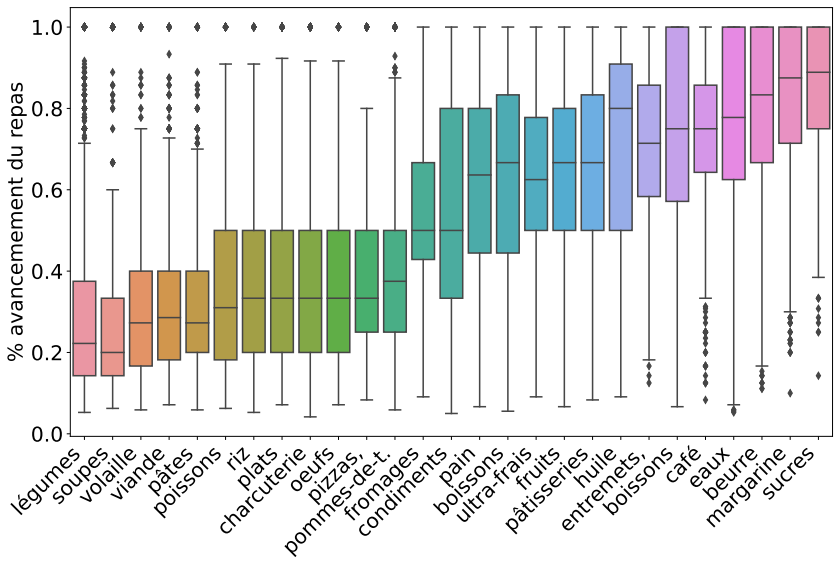


Figure 3.14: Temporal distribution of food during lunches and dinners. The abscissa designates the 44 INCA2 food groups, the ordinate indicates the period of consumption: soup, for example, is consumed at the start of lunch/dinner (first 20% of the meal). For sake of readability, INCA2 food groups have been reduced to their first word.

RQ5 How to mix user preferences and hard constraints (allergies, global energy envelopes, *etc.*) in a sequence generation context?

In [Jacquet *et al.* 2024] we focused on QR1, QR3 and QR4, marginally on QR2, we plan to deal with Q5 in perspective.

The participants of the INCA2 study entered their consumptions following the order in which they consumed them. To this order, we added the hypothesis that lunches and dinners follow a certain schema: *starter-dish-dessert*. In addition to this hypothesis, a rapid analysis of the INCA2 data shows that a certain number of cross-cutting elements (*e.g.* oils, sugars, drinks) are arbitrarily grouped at the end of the description.

In order to measure the sequentiality of the data, we studied the distribution of the INCA2 group food items over time. Each category, thus, becomes a distribution on the progress of the meal, which is expressed as a percentage to make sequences of different lengths comparable. Figure 3.14 illustrates this sequentiality on lunch/dinners: the categories are ordered according to their average and a trend emerges quite clearly. It appears in the same graph that this ordering is nevertheless noisy as shown by the numerous anomalies at the extremes of the distributions. Our preliminary experiments showed that a discrete Hidden Markov Model approach did not work well on this data, which is directly related to this level of noise.

This figure also shows that in the *starter-dish-dessert* hypothesis, the transition is

very marked between *dish* and *dessert* and much less between *starter* and *dish*, which, once again, explains the difficulty of discrete modeling the phenomenon. These analyses and preliminary experiments, therefore, lead us to consider continuous sequential modeling based on RNNs.

3.3.4.2 Modeling Food Consumption with RNNs

There exist numerous modeling tools for sequential recommendation, from Markov chains [Rendle *et al.* 2010] for discrete sequences to transformers which make it possible to introduce a global approach to better capture dependencies between distant events [Kang & McAuley 2018]. Our aim was to learn continuous representations for users and food items in order to ultimately study the topology of the population on one hand and the similarity of food items on the other. The data on which we work are qualitative but few in number, which pushed us to explore models that are parsimonious in parameters.

The flexibility of RNN architectures and the fact that they have largely proven itself on recommender systems pushed us to study them in detail. Furthermore, the sequences at hand are quite short, marked by clear transitions (starter-dish-dessert type for lunches and dinners) and, *a priori*, not subjected to long-term dependencies.

Sequential data The INCA2 data-set is a collection of consumed menus m, each menu being a sequence of food items a:

$$X = \{m^u_{d,r}\}, \qquad m^u_{d,r} = [a_1, \dots, a_t, \dots a_T] \tag{3.9}$$

Each menu is associated with a date d, a meal r (breakfast, dinner, *etc.*) and a user u.

Representations Learning The food item space $\mathcal{A}$ is discrete. Following the representation learning paradigm [Bengio *et al.* 2013], we projected the food items into a vector space of dimension z:

$$a \in \mathcal{A} \mapsto \mathbf{a} \in \mathbb{R}^z \tag{3.10}$$

Similarly, users were projected into the latent space: $u \in \mathcal{U} \mapsto \mathbf{u} \in \mathbb{R}^z$. We initialised and learnt these representations with a recurrent architecture considering different initialisation.

Recurrent Neural Networks and Meals Dynamics An RNN updates the hidden state at time t ($\mathbf{h}_t \in \mathbb{R}^h$) using the following function:

$$\mathbf{h}_t = g(W\,\mathbf{a}_t + U\,\mathbf{h}_{t-1}), \qquad W \in \mathbb{R}^{h\times z}, U\ in\mathbb{R}^{h\times h} \tag{3.11}$$

Where g is an hyperbolic tangent function, $\mathbf{a}_t \in \mathbb{R}^z$ is the input food item representation at time t and $\mathbf{h}_{t-1}$ is the hidden state at time $t-1$. The weight parameters are given by the matrices $W \in \mathbb{R}^{h \times z}$ and $U \in \mathbb{R}^{h \times h}$.

The RNN gives a prediction of the next food item in a sequence, always within an horizon time of 1. The prediction, $P(a_{t+1}|a_t)$, is estimated by a function `softmax`:

$$\hat{p} = f(\mathbf{h}_t) = \texttt{softmax}(V\,\mathbf{h}_t) \in \mathbb{R}^{|\mathcal{A}|}, \qquad \hat{a}_{t+1} = \arg\max \hat{p} \tag{3.12}$$

where the matrix $V \in \mathbb{R}^{|\mathcal{A}| \times h}$ collects the prediction parameters and the network learning criteria is the cross entropy: $H(\hat{p}) = -\log \hat{p}(a_{t+1})$. The samples are grouped into mini-batch to save calculation time, this parameter has very little influence on the performance.

Users Integration The architecture presented so far does not involve the user, it simply models the dynamics of the consumption. Two simple technical solutions made it possible to integrate the user's profile: (1) we played on the initial conditions; (2) we concatenated the user's profile to each of the entries.

Each meal starts with a special food item $\mathbf{a}_0 = DEB$, which allows the architecture to predict the first element in the sequence. The first idea we implemented is to replace $\mathbf{a}_0$ with $\mathbf{u}$ in order to make subsequent predictions user dependent and learn the user's profile in parallel. The main drawback of this solution is the fact that little gradient information will be transmitted until the end of the sequence. To solve this problem, we proposed a new architecture that provides, at each time step t, the concatenated input $[\mathbf{a}_t, \mathbf{u}] \in \mathbb{R}^{2z}$.

Initialisation The initialization of the representations is usually done randomly, the learning of the next item of the sequence allows, then, not only to optimize the weights U, V, W but also the food item representations $\{\mathbf{a}_t\}$ and user profiles $\mathbf{u}$.

However, the optimization process is non-convex and the little data available may encourage the search for finer initialization than random to facilitate and stabilize the learning. The most classic solution is to start from the profile learned by robust matrix factorization, for example by using BPR [Rendle *et al.* 2012].

Evaluation Evaluation is usually one of the most critical points in recommender systems [Said & Bellogín 2014]. To evaluate the performance of the GenSeqRNN model, we used the rate of correct classification on the prediction of the next food item, possibly relaxed in top-N (*i.e.* on the N most probable predictions made for the following food item).

3.3.4.3 Experiments

Our system should be able to suggest food items that are both relevant to the user and consistent with his consumptions. It is important to distinguish the performance of the

Metrique	Tx	Tx-Top3	Tx-CAT	Tx	Tx-Top3	Tx-CAT
Models	DEJ+DIN			PT-DEJ		
No user	0.11	0.25	0.20	0.38	0.60	0.47
Util. = $\mathbf{h}_0$	0.13	0.27	0.24	0.66	0.78	0.73
Concat. Util.+Food item.	0.16	0.30	0.24	0.71	0.83	0.77
Random	0.003	0.01	0.023	0.018	0.055	0.023

Table 3.4: Results for the good classification rate (Tx) on the prediction of the next food item in the meals of testing, Tx-Top3 is the recognition rate in Top3. Tx-Cat designates the recognition rate of the INCA2 food groups.

GenSeqRNN model on breakfast data from that of longer meals (lunch/dinner). In fact, the regularity of the users and the few food items involved in breakfast data, make the prediction much simpler for the morning meal.

Breakfast The sequentiality of breakfast is not obvious *a priori* (compared to the *starter-dish-dessert* structure of lunch/dinner) but we can observe it in the data: it is possible to predict the next element in a sequence at 60% in top-3 without even taking the user into account (see Table 3.4). This performance is directly linked to the presence of numerous recurring associations which are well predicted by the RNN architecture. The contribution of user modeling is very clear, as it is the superiority of repetitive architecture where the user profile is presented at each time step to avoid the problem of memorizing preferences at the end of the meal (+23% correct prediction in top-3).

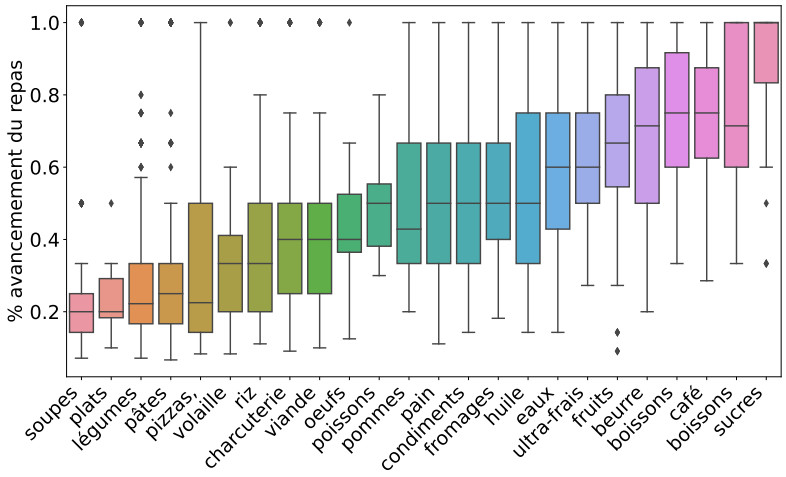


Figure 3.15: Food items temporal distribution in the predictions done by the RNN for lunch and dinner.

Lunch and Dinner The performances of the GenSeqRNN model are much more interesting on long meals. Although the performance is clearly below those on breakfast

data, 11% of the correct predictions is a satisfying result. This rate rises to 30% by integrating the user and relaxing to the top-3 metric. Figure 3.15 also illustrates the fact that the RNN architecture respects the general dynamics of the meal, even if the prediction horizon at 1, limits the conclusions on this point.

The impact of the user profile on the performance is significant but nevertheless disappointing. In our opinion, this point is directly linked to the variability of the data: the user test data (=7th day of the study) contains 51% of new food items (compared to the 11% for breakfast). This is a context, similar to a cold start, where the extraction of a profile is particularly delicate.

3.3.4.4 Final remarks

The work presented in [Jacquet *et al.* 2024] constitutes an important preliminary step towards building a recommender system in the field of nutrition. We demonstrated the weaknesses of traditional approaches, including session modeling, for profile learning and prediction when conditions are difficult (as they are in lunch and dinner data). In parallel, we compared three RNN-based approaches for taking into account dynamics and user preferences: this technical basis is interesting and now needs to be optimized and exploited.

The exploitation of RNNs for the generation of a meal is quite trivial, the main issue lies in the evaluation metric of such a system. More data will be needed to better model similarities between food items and to be able to evaluate propositions at the semantic level. Architecturally, the challenge consists in building a hybrid neural model integrating hard constraints from nutritional knowledge bases when generating a sequence of food items that also respects user preferences. This is one of the final goals of the ongoing EXERSYS project.

3.4 The Company as the Context of a Meal

If recommending a food item or a meal is important but raises a lot of difficulties, making a recommendation that suits a group of people hugely augments these problems, without being less important (think about cooking a meal for a family, or choosing a catering menus for a meeting). To address this issue we have to deal with the preference aggregation problem, that is an important task studied in social choice [Maudet *et al.* 2005] and in group recommender systems [A. Felfernig & Tkalcic 2018].

With Paolo Viappiani and Nicolas Darcel I supervised two internships on aggregating preferences and applied the approaches implemented to provide food related recommendation to a group of people. During the internship of Maéva Caillat we introduced a Bayesian elicitation paradigm for social choice. The system maintains a discrete probability distribution over the preferences (rankings) of the users. At each step the system asks a pair-wise question to one of the voters and updates the distribution conditioned

on this response. We considered strategies to pick the next question based on the value of information, the conditional entropy and a mix of these two notions. We developed these ideas focusing on scoring rules and compared different elicitation strategies in the case of the Borda rule.

The internship of Yuhan Wang, that I supervised with Paolo Viappiani, relaxed the hypothesis of having a discrete probability distribution over the preferences of the users. We investigated the cases where the rankings are not known precisely, but are known to belong to some probabilistic ranking model, we used the Plackett-Luce (PL) model and the expected Borda score. Moreover, we considered questions of the "next best" item kind.

Both internships have developed application independent approaches that we experimented on food related data. In both internships, we made some experiments on the sushi data-set[12]. During Maéva Caillat's internship we also collected some new data to prove the acceptability of the recommendations.

3.4.1 Bayesian Vote Elicitation for Group Recommendation

Preferences are not always readily available in social choice and group recommender systems. It is necessary to consider voting procedures with incomplete preferences and elicitation procedures for voting. It is known [Konczak & Lang 2005] that one can find an item that is likely to satisfy a group without knowing every user's preference.

During Maéva Caillat's internship we studied an incremental elicitation process that asks a reduced number of questions to the users until it finds an item suitable for the entire group [Caillat *et al.* 2020]. We considered a Bayesian approach, that provides a principled quantification of uncertainty and uses the notion of expected loss that allows an elegant termination condition.

3.4.1.1 Bayesian Vote Elicitation

Notation We assumed that $U = \{1, \ldots, n\}$ is the set of users and A the set of items (or candidates); with $|A| = m$. A profile $(r^1, \ldots, r^n)$ is a vector of linear orders on A, one for each voter $u \in U$. The set of all $m!$ linear orders on A is denoted as L; hence the set of all possible profiles is L^n. We used $r(x)$ to denote the position of item x in ranking r and r_i to denote the candidate in the i-th position in r. So if $r(x) = i$ then $r_i = x$ and vice versa.

A voting rule $f : L^n \to 2^A \setminus \emptyset$ maps a profile to a non-empty subset of items (the "winners"). Scoring rules are voting rules that rank candidates according to their score, computed by summing up the number of points they receive in each ranking; the score

[12] http://www.kamishima.net/sushi/

of x with respect to rankings $r^1, \ldots, r^n$ is:

$$s(x; r^1, \ldots, r^n) = \sum_{i=1}^{n} w(r^i(x)),$$

where $w()$ is a function that assigns each position $1, \ldots, m$ to a number of points. We used the Borda scoring rule, that assumes the weights to be $w(i) = m - i$: the first ranked item is assigned $m - 1$ points, the second obtains $m - 2$, *etc.*

The Bayesian Approach At each step of the elicitation, we faced a voting problem under incomplete preferences [Konczak & Lang 2005], where the preference profile consisted in partial orders. We adopted a Bayesian approach to preference elicitation and approximate winner determination.

Our system maintains a probability distribution $\mathbb{P}(r^1, \ldots, r^n)$ over the preferences (rankings) $r^1, \ldots, r^n$ of the voters. The distributions give zero probability to all rankings that are not completion of the known preferences of the users. We assumed that the voters preferences were independent, thus $\mathbb{P}(r^1, \ldots, r^n) = \mathbb{P}(r^1) \cdot \ldots \cdot \mathbb{P}(r^n)$. The incremental elicitation approach developed during Maéva Caillat's internship looped through the following steps:

- it computes the current best candidate, x^*, that achieves the highest score *in expectation*;
- if x^* meets the stopping criterion, the procedure stops and outputs x^*;
- otherwise, it selects an elicitation question to ask the user and asks it;
- it updates the probability distribution of the rankings conditioned on the obtained response.

In the following, we discuss these different steps.

Computing the Expected Scores We identified the "approximate winner" as the candidate that yields the highest expected score under the current probability distribution of rankings. Each alternative x is associated to its expected score $\bar{s}(x)$. For a scoring rule, $\bar{s}(x)$ can be computed as:

$$\bar{s}(x) = \sum_{i=1}^{n} \sum_{r^i \in L} \mathbb{P}(r^i) w(r^i(x))$$

For the Borda scoring rule, the expression simplifies to:

$$\bar{s}(x) = mn - \sum_{i=1}^{n} \sum_{r^i \in L} \mathbb{P}(r^i) r^i(x).$$

Let $s^* = \max_{x \in A} \bar{s}(x)$ be the maximum value of the expected score given the current uncertainty. The associated candidate, x^* (the "winner in expectation"), is the best item computed by the approach at this step: $s^* = \bar{s}(x^*)$.

Expected Loss and Stopping Criterion Once we have a best candidate, we estimate the regret (or loss) of stopping the elicitation and recommending x^*. The user's loss is the difference between his expected utility, under the true preferences, of the optimal alternative x^+, and his expected utility under the recommended alternative x^*.

The loss $\ell(r^1, \ldots, r^n)$ is the regret of choosing x^* that occurres when the true users' preferences are $(r^1, \ldots, r^n)$:

$$\ell(r^1, \ldots, r^n) = \max_{y \in A} s(y; r^1, \ldots, r^n) - s(x^*; r^1, \ldots, r^n).$$

Since we do not know the true preferences, but we know their distributions, we consider the expected loss $\mathbb{E}[\ell]$ (in a way analogous to [Chajewska *et al.*] that considered Bayesian elicitation in influence diagrams) that quantifies how far we are from the true optimum in expectation:

$$\mathbb{E}_{r^1 \sim \mathbb{P}(r^1), \ldots, r^n \sim \mathbb{P}(r^n)}[l(r^1, \ldots, r^n)] =$$
$$= \Big[\sum_{r^1 \in L} \cdots \sum_{r^n \in L} \mathbb{P}(r^1) \ldots \mathbb{P}(r^n) \max_{y \in A} s(y; r^1, \ldots, r^n) \Big] - s^*.$$

We approximated the above expression with a Monte Carlo method. We sampled the users' preference rankings from $\mathbb{P}(r^1), \ldots, \mathbb{P}(r^n)$, we computed the scores of alternatives and computed the loss for these preferences. We repeated the procedure N times and took the average. To set N (the number of samples) we used the Chebyshev inequality. N should be at least $\frac{b^2}{4\delta\varepsilon^2}$ where ε is the required precision, δ is the confidence and b is an upper bound of the variance; in our case we set $b = n(m-1) - s^*$ (the highest possible Borda score less the current best expected score).

The elicitation procedure continues until the expected loss is lower than a given threshold. If the goal is to find a necessary winner with certainty, the threshold can be set to zero.

Elicitation Strategies We considered three different strategies to decide which query to ask at any step of the elicitation process. The strategies aim at reducing uncertainty over the users' preferences to improve the quality of the approximated winner.

We focused on pairwise comparisons since it has been shown that it is easier for users to state opinions when the queries are pairwise [Balakrishnan & Chopra 2010]. In the following we denote $q^u_{a,b}$ the query asking user u to compare items a and b.

The first elicitation strategy we used is the Information Gain for Borda (IGB). The goal of the IGB strategy is to maximize the information in terms of entropy at each step till reducing the entropy to its minimum value [Naamani-Dery *et al.* 2015]. Given the query $q^u_{a,b}$, it, first, computes the information gain (IG) of every answer $q^u_{a \succ_u b}$. This is the difference between the prior entropy and the posterior entropy given this answer to the probability of winning for an item. The next selected query is the one maximizing

the weighted information gain. In case of equality, we choose the item with the smallest ID.

Let us call $P_{\text{win}}(a)$ the probability that a wins, this can be found by summing up the probability of all preference combinations that make a a winner. Let $H(W)$ be the entropy associated to the distribution P_{win}. The query $q^u_{a,b}$ is associated with its *information gain* (*i.e.* the conditional entropy):

$$\text{IG}(q^u_{a,b}) = p^u_{a \succ_u b} H(W|a \succ_u b) + p^u_{b \succ_u a} H(W|b \succ_u a).$$

where $p^u_{a \succ_u b}$ is the probability that user u prefers a to b and can be computed by marginalization. The chosen query following the IGB strategy is the one that maximizes IG.

The second strategy, we implemented, is the Expected Score Euristic for Borda (ESB) [Naamani-Dery *et al.* 2015]. It computes the *a posteriori* improvement of the maximum of P_{win}.

This strategy relies on the hypothesis that it is better to select a query $q^u_{a,b}$ where one item a or b is expected to win. If we pick an item that has significant chances to win, the possible minimum will increase faster and a necessary winner will stand out quickly. Thus, ESB selects the queries containing the item with the highest winning probability.

Given a query, $q^u_{a,b}$, the Expected Maximum (EM) of the answer $q^u_{a \succ_u b}$ represents the difference between the highest winning probability given user u prefers a over b and the highest winning probability without asking any query:

$$EM(q^u_{a \succ_u b}) = max(p^u_{a \succ_u b}) - max(P_{\text{win}}).$$

The Weighted Expected Maximum (WEM) is computed as:

$$WEM(q^u_{a,b}) = p^u_{a \succ_u b} EM(q^u_{a \succ_u b}) + p^u_{b \succ_u a} EM(q^u_{b \succ_u a}).$$

The query chosen by the ESB strategy is the one that maximizes the WEM.

The third strategy adopts the myopic Expected Value of Information (EVOI). EVOI has been shown, in [Viappiani & Boutilier 2020], to be very effective for single-user preference elicitation and is defined as:

$$\text{EVOI}(q^u_{a,b}) = p^u_{a \succ b} \max_{x \in C} \mathbb{E}[s(x)|a \succ b] + p^u_{b \succ a} \max_{x \in C} \mathbb{E}[s(x)|b \succ a] - s^*.$$

The selected query, following the EVOI strategy, is the one with the highest EVOI.

While EVOI can often identify very informative queries, in preliminary tests we realized that it can sometimes happen that *myopic* EVOI of all candidate queries is zero. Motivated by this observation, we designed the *EVOI+IGB* strategy that asks the query with the highest EVOI if its value is positive and follows the IGB strategy otherwise.

Updating the Distributions Whenever a query is answered, the distributions are updated using the Bayes theorem. Since there is no noise in user feedback, this means assigning zero probability to rankings that are inconsistent with the user's input and to renormalize.

3.4.1.2 Experiments

We carried out some experiments on the Sushi data-set [Kamishima *et al.* 2010] and on a data-set we collected, called CROUS. The experiments on the Sushi data-set allowed us to evaluate the performance of the Bayesian elicitation approach comparing the different elicitation strategies. The experiments on the CROUS data-set allowed to assess the accuracy of the food recommendations made by the system.

Experiments on the Sushi Data-set We examined a scenario of users who were required to decide between ten types of sushi. The Sushi data-set contains 5 000 preference rankings over ten kinds of sushi. We derived six different random matrices of the size of ten users $\times$ six sushi. These six sushi were chosen randomly among the ten items ranked in the data-set.

To create an initial permutation probability distribution, we aggregated the number of appearances of each permutation in the training set and divided it by the total number of users. Then, we normalized the initial distribution, so that there was no permutation receiving a null probability. Thus the initial permutation distribution was equal for all users. Then, we randomly selected users in the sushi data-set and extracted their rankings over the six previous sushi. As users answered more queries based on the selected preferences, the distributions were updated for each user. Over time, a unique permutation distribution pattern emerged for each user.

The experiments show that a necessary winner can be found with relatively few questions. Somewhat surprisingly, we found the ESB strategy to perform worse than IGB, contrary to what reported in [Naamani-Dery *et al.* 2015]. In our tests, EVOI+IGB was the most efficient query strategy.

Experiments on the CROUS Data-set We tested the approach in a realistic setting of food recommendations. To assess the accuracy of the food recommendations made, we examined a scenario of users required to make a common choice between dishes. We collected a food preferences data-set that we have called the CROUS data-set.

The CROUS data-set contains 130 preference rankings over five starters, five main courses and five desserts. These 15 dishes are top-ranked items from the menu of the CROUS of Versailles.

We created an online questionnaire. 130 French-speaking persons over the age of majority (75 women and 55 men) completed it. After ranking the five starters, five main courses and five desserts in order of preference, 18 of the participants gathered in "virtual tables" of four or five people on an instant messaging application. For each discussion group, the selection of a meal (starter, main course and dessert) was carried out both by the participants - by exchanging via instant messaging -, by the recommendation algorithm - by querying the database preferences of the guests -, as well as by an operator

having access to all the preferences of the participants. We evaluated the accuracy of our system by comparing the dishes selected by the virtual tables with those returned by the algorithm, as well as with the dishes suggested by the omniscient operator.

We compared the items selected by the participants of the experiment with the items having the highest Borda scores. 4 items out of 12 (33%) were the winning elements for both the Borda protocol and the users' experiment. 2 (17%) items selected by the virtual tables had the second best Borda score; 5 (42%) the third best Borda score and 1 the fourth best Borda score (8%). In summary, in this experiment, the Borda voting protocol matched reality half of the times. When the Borda model matched reality (6 cases out of 12), the variance of the Borda scores divided by the number of participants was high (this means that group preferences were clear). The cases where there were substantial mistakes had lower variance of the Borda scores divided by the number of participants (this is the case when group preferences were not clear).

We looked closely at the instant messaging chats, in particular at the cases for which there were huge difference between the Borda score and reality. We realized that, except for the cases where everyone agreed over an item, people did not state clearly their preferences. They would rather try to make everyone more or less happy than openly express their tastes. Sometimes, people might even agree over a particular dish regarding their hidden personal rankings, but the group would rather choose an item that the voters think others prefer. In that case, the Borda score was mistaken. This is one of the aspects we will investigate in the thesis of Thomas Dheilly, that started in November 2023 that I am co-supervising with Nicolas Darcel, Sabrina Teyssier, Paolo Viappiani and Patrick Taillandier.

3.4.2 Bayesian Preference Elicitation for Group Decisions with the Plackett-Luce Model

Since the representation of distributions used during Maéva Caillat's internship was not scalable, during Yuan Wang's internship, we adopted a probabilistic ranking model: the Plackett-Luce (PL) model. PL models a ranking process of a set of items with a vector of weights from which probabilities can be easily computed.

3.4.2.1 The Plackett-Luce Model

Probabilistic ranking models are used to compactly represent a probability distribution over rankings. A popular probabilistic ranking model is the PL model.

Definition 1 *Given a vector of weights $\gamma = (\gamma_a)_{a \in A}$, where γ_a is the weight of alternative $a \in A$, a ranking $r = \langle r_1, \ldots, r_m \rangle$ on m items is distributed according to $\mathrm{PL}(\gamma)$, denoted as $r \sim \mathrm{PL}(\gamma)$, if:*

$$\mathbb{P}(r) = \prod_{i=1}^{m-1} \frac{\gamma_{r_i}}{\sum_{j=i}^{m} \gamma_{r_j}}.$$

Obviously multiplying the vector γ by a constant does not change the distribution.

An important property of PL is the following:

Lemma 1 *Let r be a ranking sampled from $PL(\gamma)$. Let $a, b \in A$; The probability of the event "a is preferred to b" evaluates to:*

$$\mathbb{P}(a \succ_r b) = \frac{\gamma_a}{\gamma_a + \gamma_b}.$$

Sampling The definition of the PL model suggests a straightforward efficient way to sample from a PL distribution. The sampling proceeds in steps. In the first step, we sample from a cardinal distribution where, for all $a \in A$, the probability of picking item a is $\frac{\gamma_a}{\sum_{c \in A} \gamma_c}$. Then, in each of the following steps, we pick an item from those that have not been picked in previous steps, with the probability of picking item a proportional to γ_a.

Expected Rank When the ranking r is uncertain, the position $r(a)$ of an alternative a is a random variable. We are interested in evaluating the expected rank (position) of an alternative a, when the ranking r is not known, but we know its distribution ($\mathbb{E}[r(a)]$).

When adopting PL as a probabilistic ranking model for r (*i.e.* when $r \sim \mathrm{PL}(\gamma)$) the expected rank of an item, x, can be computed in a very convenient way:

$$\mathbb{E}_{r \sim \mathrm{PL}(\gamma)}[r(x)] = m - \sum_{y \neq x; y \in A} \frac{\gamma_x}{\gamma_x + \gamma_y}.$$

Expected Borda Score We were interested in the cases where the rankings are known to belong to some probabilistic ranking model. Since the Borda score cannot be computed, we evaluated the alternatives with respect to their expected Borda score. It is possible to efficiently compute the expected Borda score of an item when n rankings are sampled from a PL model.

For each user, let γ^i denotes the PL weights. Let $r^1 \sim \mathrm{PL}(\gamma^1), \ldots, r^n \sim \mathrm{PL}(\gamma^n)$. The expected Borda score of an alternative x can be computed as:

$$\mathbb{E}[s(x)] = \sum_{i=1}^{n} \sum_{y \neq x; y \in A} \frac{\gamma_x^i}{\gamma_x^i + \gamma_y^i}.$$

3.4.2.2 Bayesian Preference Elicitation with the PL Model

During Yuhan Wang's internship we focused on the estimation of the PL weights. The performances of three estimation algorithms that maximise the log-likelihood were compared. Those are: the Generalized Method-of-Moments (GMM) [Hansen 1982, Azari Soufiani *et al.* 2013], the Minorize-Maximization (MM) [Hunter *et al.* 2004] and the Luce-Spectral-Ranking (LSR) [Maystre & Grossglauser 2015]. Our experiments showed that GMM was able to best approximate the PL weights in different settings.

Another aspect on which Yuhan Wang's work focused has been the type of questions to pose. Instead of posing pair-wise questions such as "do you prefer a to b?" as done in [Caillat *et al.* 2020], the system poses questions like "which is the item you best like next?": we asked the user to choose the item he likes best among the items not ranked yet. To evaluate the quality of the question to choose, we implemented and compared the performance of the EVOI and the IGB as in [Caillat *et al.* 2020].

We computed an EVOI for each user, but, differently from Maéva Caillat's work, having a different kind of query to ask, we did not compute it for each triplet user-item1-item2. For each user, we divided the set of items A in two sets: the set of items not yet ranked by the user, A_n^u, and the set of items already ranked by the user, A_r^u.

$$\forall u \in U, \quad A_n^u \cup A_r^u = A, \quad A_n^u \cap A_r^u = \emptyset$$

The expected Borda score for an item x for a user u can be computed as:

$$\mathbb{E}_u[s(x)] = \sum_{y \neq x; y \in A_n^u} \frac{\gamma_x^i}{\gamma_x^i + \gamma_y^i} \text{ si } x \in \mathcal{A}_n^u.$$

The equation for the IGB for the query q^u to pose to user u becomes:

$$\mathrm{IG}(q^u) = \sum_{x \in A_n^u} p_x^u H(W|x \text{ has been selected})$$

As in the strategy that uses the EVOI, the approach asks to the user that has the most elevated IGB the item he prefers among the items still not ranked.

To obtain an estimation of the parameters γ for each user, that we need at the beginning of the process, we applied the GMM algorithm. The vector of parameters γ represents the distribution of the individual preferences *a priori*. It represents the initial uncertainty. No one has yet declaired any preferences, so all the permutations are still possible.

$$\forall u \in U, A_n^u = A, A_r^u = \emptyset$$

We first computed the expectation of the regret under the current uncertainty. If this expectation is less than or equal to the pre-defined threshold, it means that the quality of the current recommendation is satisfying, the procedure can be stopped and the recommendation sent back to the users. Otherwise, more information about preferences are needed. We will ask the "next best" type of question to the user with the highest EVOI, if there is a user with positive EVOI. If the EVOI are all negative, we calculate the IGB for each user and ask the question to the user who gives us more information. After the interaction, the uncertainty distribution is updated by the user's response and a new approximated winner is determined from this new distribution. We repeat these steps till obtaining an expectation bounded by the given threshold.

3.4.2.3 Experiments on the Sushi Data-set

We conducted some preliminary experiments on the Sushi data-set. We randomly took data from this data-set as historical data to build different PL models with different parameters to simulate users with different preferences. For the interaction, we took the n linear ordering as the true profile of the group. The user chosen by the elicitation strategy will answer the "next best" question following his pre-selected rankings. We applied our recommendation procedure to different scenarios where the number of candidates and the number of users and their behaviours vary.

With respect to the state of the art, the recommendation algorithm developed based on the PL model performed better for large groups (10-15 users) or many available items. The recommendation could be made in about ten interactions and in 995 experiments out of 1000, our system succeeded in returning the right candidate. The efficiency and quality of recommendation has been improved with respect to the state of the art.

One point on which the system needs to be improved is the Monte Carlo approximation. In Yuhan Wang's internship, we found that the performance of the Monte Carlo method for this settings is difficult to control: choosing a large number of samples slows down a lot the algorithm, while a small number of samples does not reach good performances.

3.5 Final Remarks

In June 2020 and June 2021 I co-supervised two internships on very similar topics. The work of Maéva Caillat has been a first attempt to the implementation of a group recommender system. She provided a working recommender system able to give recommendations to a group of users of which preferences are unknown at first and to whom the system asks pair-wise questions. The uncertainty on the individual preferences has been modeled by a discrete probabilistic model.

The system implemented during the internship of Yuhan Wang relaxed the hypothesis of a discrete probabilistic model and modeled the uncertainty on the preferences of the users with the PL model. The scalability of this model allowed us to investigate another kind of queries: instead of asking pair-wise questions, the system developed by Yuhan Wang asked "next best" questions. This provided a faster and more precise group recommender system.

These works are far to be completed. We would like to investigate other probabilistic models for the uncertainty of the users' preferences and other types of query. In both the works we aggregated the score obtained by the items with the Borda rule, it could be interesting to experiment other aggregation methods to better model user behaviours.

Another result of Maéva Caillat's internship is the CROUS data-set. A careful analysis of this data-set, showed that, while users have their own preferences, when in group, they tend to suit the others. Starting November 2023 I am co-supervising

the thesis of Thomas Dhelly on how interactions with other people influence our eating habits.

The GIFTED thesis project aims at understanding how social influences amplify or attenuate the effects of information policies about nutritional or environmental aspects of food. The first part of the thesis will investigate the use of tools from cognitive and experimental economics science to assess the effects of direct social influences on the decisions of consumption of products by consumers in the presence or absence of nutritional or environmental information. We will study how the preferences of the user, their perception of the rules and the direct influence of the social context influence and interact in the decision of the user. We will, then, focus on understanding the mechanisms leading to the achievement of a group consensus for sustainable food choices, with or without information and depending on the social context.

It will be necessary to determine if the fact of looking for a consensus by the user, can influence his future behaviour changing his preferences. This work will be carried out comparing the results obtained by models for single user preference elicitation and group preference elicitation (as in the works of Maéva Caillat and Yuhan Wang) with food choice observation data obtained in discussion groups or in real food choice situations. The answers that will be provided in the thesis will help in the implementation of new policy instruments that will take into account social influences to facilitate the transition towards more sustainable food consumption patterns.

The final goal of the EXERSYS project is to develop a meal sequential recommender system based on the combination of machine learning methods and methods of the knowledge representation domain. With the internships of Noémie Jacquet and Ayoub Hammal we did a first step towards this goal. The FilterCollab model merged with the BDNutri knowledge base is able to deliver recommendations that satisfy user's preferences and diet constraints for a breakfast meal.

Eating consumptions are (intrinsically) sequential. To model this aspect the GenSeqRNN model uses an RNN algorithm that well generates the next food item given an history of food consumptions. In Alexandre Combeau's PhD thesis we are currently investigating this aspect merging RNN algorithms and KGs.

Moreover, one of the points we address in the EXERSYS project is the context of consumption, modelled as constraints for the recommendation. In this setting, we could consider the observations done on the CROUS data-set or the ones we will do in Thomas Dhelly' thesis, to best model the context in a recommender system; this context being the interactions between individuals.

Chapter 4

Conclusions

Contents

In this manuscript I reviewed my most important research contributions, retracing my main activities since obtaining my PhD, mainly focusing on what I have done while assistant professor at AgroParisTech. These works are the results of collaborations (that I have often initiated) with other researchers, PhD students, master interns and postdocs.

My research concerns methods for integrating experts' knowledge in learning and inference in machine learning; I applied my research to various application domains and different classes of statistical learning approaches.

At AgroParisTech, I have focused my expertise in the field of life sciences. In the era of big data, one of the main issues raised by most life science applications is the lack of data. Indeed, in this field, experiments are often conducted with little repetitions, the materials used are often expensive and the knowledge available is seldom complete. This makes acquiring data in this field a very difficult task and, thus, the development of tools facilitating reasoning and learning with limited data is a research axis of critical importance. My research aims at contributing to this axis by proposing to integrate expert's knowledge into the system to ease the reasoning task: I formalised expert's knowledge within the ontology framework and we used knowledge graphs to formalise the expert's knowledge at our disposal making it exploitable by the automatic system.

A recurrent theme of my research is the presence of the *human-in-the loop*, as in most of my works I exploit the information provided by a human expert in order to improve an automated system using machine learning techniques. A second common theme of my research is that of providing systems that are *explicable* and *interpretable*, thanks to the integration of the expert's knowledge into machine learning techniques.

4.1 Summary of Contributions

In this manuscript I organised my works in two groups, corresponding to the two main research domains I dealt with. Those are detailed in the two main chapters of the manuscript.

4.1.1 Experts' Knowledge and PRMs

In Chapter 2, I presented how I proposed to map an ontology representing experts' knowledge to probabilistic relational models (PRMs). Motivated by the necessity of reasoning about transformation experiments and their results, I proposed methods that use data enriched with expert's knowledge formalised within the ontology framework to learn probabilistic models.

I showed how this approaches allow to deal with the problems of (1) modeling a transformation process, (2) reasoning with the uncertainty present on it, (3) causal discovery and (4) parameters control. I presented the POND framework that uses ontology axioms and properties to do inference on the domain and, when this is not enough, uses the PRM aligned from the ontology to do inference taking into account uncertainty in the domain.

4.1.2 Experts' Knowledge and Recommender Systems

The motivation for the second group of works, described in Chapter 3, is that unhealthy eating habits are widespread and contribute to the insurgence of many chronic diseases such as diabetes or cardiovascular diseases. In this context, I presented my work on food recommendations, aimed at providing healthy food recommendations that can be accepted by the user.

I presented the definition of dietary context and food intake context and I showed how considering these contexts in a recommender system could improve the recommendation task. In the EXERSYS project, based on the works we did on aligning ontologies and PRMs, I propose to formalise the definition of context within the ontology framework. I propose to integrate experts' knowledge expressed in a Knowledge Graph KG in a recommender system that uses a Recursive Neural Networks (RNN) based approach, to provide recommendations of sequences of menus that can respond to different nutritional constraints.

In this chapter I also presented the studies I supervised on methods to aggregate preferences for group recommendations, focusing on the food domain. These studies are connected with an ongoing PhD thesis on how interactions with other people influence our eating habits.

4.2 Future Works

In future years, I intend to continue working at the interface of machine learning and knowledge representation methods. In the following paragraphs, I present different directions for research, extending the discussion at the end of previous chapters. I also list some additional research avenues not previously covered.

These future directions could take place within my established research collaborations or with new collaborations at AgroParisTech or at the University of Paris-Saclay as well with old or new collaborations with researchers abroad. I am, also, motivated in developing new industrial collaborations.

4.2.1 Expert's Knowledge and PRMs

Giving Feedback to the Ontology A natural next step for my work on aligning ontologies and PRMs would be to make the system able to give feedback to the knowledge base to improve the data and, possibly, the ontology itself. We could use the ability of the POND workflow, to evaluate the quality of potential new data (probable, not probable, impossible) and help the expert finding outliers or to suggest relational constraints that are not present in the ontology so to improve the ontology itself according to new information gathered.

Transfer Learning and ontologies A possible future work of Mélanie Münch's PhD thesis could be an approach to transfer the knowledge we have over a knowledge base (mapped with a PRM) to another one. The idea could be to use data linking methods to ease the learning of a PRM from a new knowledge base by transfering the learning from an existing "close" one and its PRM. For this goal, data linking techniques (that use ontology alignment approaches whose objective is to detect the correspondences between the concepts and the relations of different ontologies) and transfer learning techniques (that reuse knowledge already acquired in one domain to improve or accelerate learning in new domains) could be merged to exchange information between ontologies aligned to some PRMs.

This approach could be used to:

- learn a new PRM from a new ontology, that is similar to another one, and has itself associated few data;
- merge different data-sets obtained from experiments with different settings;
- deal with domain evolution.

4.2.2 The EXERSYS Ongoing Project

Sequence Recommendation I intend to develop more the ongoing EXERSYS project. The research work on food recommender systems is currently being extended by considering the sequential aspects of the recommendation, since food consumption is intrinsically sequential.

This line of research is part of the ongoing PhD thesis of Alexandre Combeau, where we are currently investigating the use of the GenSeqRNN model with an RNN algorithm in order to generate the next food item given an history of food consumption. This work also aims at merging RNN algorithms and KGs, consistently with my focus on combining expert's knowledge with learning techniques.

Group Recommendations I would like to extend the work done during Yuhan Wang's internship by investigating other probabilistic models for the uncertainty of the users' preferences and other types of query. Until now, we aggregated the score obtained by the items with the Borda rule, but it could be interesting to experiment other aggregation methods to better model user behaviours. It would be interesting to consider the development of other elicitation strategies and, as well, other probabilistic models.

I also plan to study methods for groups recommendation taking into account user's preferences and the context of food consumption. This will be a natural follow up of the theses of Alexandre Combeau and Thomas Dheilly to obtain a recommender system able to provide suggestions that are understandable by the user and that take into account preferences, past consumptions and the social context of the consumption.

4.2.3 New Application Domains

I am interested in new application domains; some ideas are quickly mentioned below.

Information Visualisation Interactions with the experts cannot be disjointed from taking into account the human interface; for this purpose I intend to initiate collaborations with experts in information visualisation. My aim is to develop a system that is easy to use by the expert that provides knowledge to improve the system.

Tracking the Growth of Crops A possible relevant topic is the domain of following the growth of crops from drone's images (the expert knows if a particular speciem influences the growth of another, and this information can be taken into account to better follow the different speciems); with Jean-Marc Gilliot, we submitted a project on this topic at AgroParisTech.

Anomaly Detection A second possible direction concerning tracking is the detection of anomalies in time series taking into account expert's knowledge; a topic that is of interest in real applications

Bibliography

[A. Felfernig & Tkalcic 2018] M. Stettinger A. Felfernig L. Boratto and M. Tkalcic. Group recommender systems. an introduction. Springer, 2018. (Cited on page 104.)

[Achananuparp & Weber 2016] Palakorn Achananuparp and Ingmar Weber. *Extracting Food Substitutes From Food Diary via Distributional Similarity*. CoRR, vol. abs/1607.08807, 2016. (Cited on pages 66 and 69.)

[Adomavicius & Tuzhilin 2005] Gediminas Adomavicius and Alexander Tuzhilin. *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*. Knowledge and Data Engineering, IEEE Transactions on, vol. 17, pages 734–749, 07 2005. (Cited on page 60.)

[Adomavicius & Tuzhilin 2010] Gediminas Adomavicius and Alexander Tuzhilin. *Context-aware recommender systems*. In Recommender systems handbook, pages 217–253. Springer, 2010. (Cited on pages 60, 68 and 69.)

[Adomavicius *et al.* 2022] Gediminas Adomavicius, Konstantin Bauman, Alexander Tuzhilin and Moshe Unger. *Context-Aware Recommender Systems: From Foundations to Recent Developments*. In Francesco Ricci, Lior Rokach and Bracha Shapira, editors, Recommender Systems Handbook, pages 211–250. Springer US, 2022. (Cited on page 69.)

[Akkoyunlu *et al.* 2017] Sema Akkoyunlu, Cristina E. Manfredotti, Antoine Cornuéjols, Nicolas Darcel and Fabien Delaere. *Investigating Substitutability of Food Items in Consumption Data*. In David Elsweiler, Santiago Hors-Fraile, Bernd Ludwig, Alan Said, Hanna Schäfer, Christoph Trattner, Helma Torkamaan and André Calero Valdez, editors, Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th International Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017, volume 1953 of *CEUR Workshop Proceedings*, pages 27–31. CEUR-WS.org, 2017. (Cited on pages 5, 8, 63, 68, 69, 83 and 84.)

[Akkoyunlu *et al.* 2018] Sema Akkoyunlu, Cristina E. Manfredotti, Antoine Cornuéjols, Nicolas Darcel and Fabien Delaere. *Exploring Eating Behaviours Modelling for User Clustering*. In David Elsweiler, Bernd Ludwig, Alan Said, Hanna Schäfer, Helma Torkamaan and Christoph Trattner, editors, Proceedings of the 3rd International Workshop on Health Recommender Systems, HealthRecSys 2018, co-located with the 12th ACM Conference on Recommender Systems (ACM RecSys

2018), Vancouver, BC, Canada, October 6, 2018, volume 2216 of *CEUR Workshop Proceedings*, pages 46–51. CEUR-WS.org, 2018. (Cited on pages 64, 75, 76, 77 and 83.)

[Amirkhani *et al.* 2017] H. Amirkhani, M. Rahmati, P. J. F. Lucas and A. Hommersom. *Exploiting Experts' Knowledge for Structure Learning of Bayesian Networks*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pages 2154–2170, Nov 2017. (Cited on page 22.)

[Azari Soufiani *et al.* 2013] Hossein Azari Soufiani, William Ziwei Chen, David C Parkes and Lirong Xia. *Generalized method-of-moments for rank aggregation*. 2013. (Cited on page 111.)

[Balakrishnan & Chopra 2010] Suhrid Balakrishnan and Sumit Chopra. *Two of a Kind or the Ratings Game? Adaptive Pairwise Preferences and Latent Factor Models*. In 2010 IEEE International Conference on Data Mining, pages 725–730, 2010. (Cited on page 107.)

[Ben Messaoud *et al.* 2009] Montassar Ben Messaoud, Philippe Leray and Nahla Ben Amor. *Integrating Ontological Knowledge for Iterative Causal Discovery and Visualization*. In Claudio Sossai and Gaetano Chemello, editors, Symbolic and Quantitative Approaches to Reasoning with Uncertainty, pages 168–179, 2009. (Cited on page 22.)

[Bengio *et al.* 2013] Yoshua Bengio, Aaron Courville and Pascal Vincent. *Representation learning: A review and new perspectives*. IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pages 1798–1828, 2013. (Cited on page 101.)

[Besnard *et al.* 2014] Philippe Besnard, Marie-Odile Cordier and Yves Moinard. *Arguments Using Ontological and Causal Knowledge*. In Foundations of Information and Knowledge Systems - 8th International Symposium, FoIKS 2014, Bordeaux, France, March 3-7, 2014. Proceedings, pages 79–96, 2014. (Cited on page 22.)

[Bier *et al.* 2008] Dennis Bier, Doris Derelian, J. Bruce German, David L Katz, Russell R. Pate and Kimberly M. Thompson. *Improving Compliance With Dietary Recommendations: Time for New, Inventive Approaches?* Nutrition Today, vol. 43, pages 180–187, 2008. (Cited on page 61.)

[Bobillo *et al.* 2013] Fernando Bobillo, Miguel Delgado and Juan Gómez-Romero. *Reasoning in Fuzzy OWL 2 with DeLorean*. In Uncertainty Reasoning for the Semantic Web II, International Workshops URSW 2008-2010 Held at ISWC and UniDL 2010 Held at FLoC, Revised Selected Papers, pages 119–138, 2013. (Cited on page 16.)

[Boulos *et al.* 2015] Maged N. Kamel Boulos, Abdulsalam Yassine, Shervin Shirmohammadi, Chakkrit Snae Namahoot and Michael Brückner. *Towards an "Internet of Food": Food Ontologies for the Internet of Things*. Future Internet, vol. 7, no. 4, pages 372–392, 2015. (Cited on page 67.)

[Bron & Kerbosch 1973] Coenraad Bron and Joep Kerbosch. *Finding All Cliques of an Undirected Graph (Algorithm 457)*. Commun. ACM, vol. 16, no. 9, pages 575–576, 1973. (Cited on page 70.)

[Buche *et al.* 2005] Patrice Buche, Catherine Dervin, Ollivier Haemmerlé and Rallou Thomopoulos. *Fuzzy querying of incomplete, imprecise, and heterogeneously structured data in the relational model using ontologies and rules*. IEEE T. Fuzzy Systems, vol. 13, no. 3, pages 373–383, 2005. (Cited on page 16.)

[Caillat *et al.* 2020] Maeva Caillat, Nicolas N. Darcel, Cristina Manfredotti and Paolo Viappiani. *Bayesian Vote Elicitation for Group Recommendations*. In From Multiple Criteria Decision Aid to Preference Learning (DA2PL 2020), Trento, Italy, November 2020. Andrea Passerini (University of Trento) and Vincent Mousseau (Centrale Supelec). (Cited on pages 9, 105 and 112.)

[Caracciolo *et al.* 2023] Caterina Caracciolo, Sophie Aubin, Clément Jonquet, Emna Amdouni, Romain David, Leyla J. García, Brandon Whitehead, Catherine Roussey, Armando Stellato and Ferdinando Villa. *Correction: 39 Hints to Facilitate the Use of Semantics for Data on Agriculture and Nutrition*. Data Sci. J., vol. 22, 2023. (Cited on page 67.)

[Carvalho *et al.* 2013] Rommel N. Carvalho, Kathryn B. Laskey and Paulo Cesar G. da Costa. *PR-OWL 2.0 - Bridging the Gap to OWL Semantics*. In Fernando Bobillo, Paulo Cesar G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool, editors, Uncertainty Reasoning for the Semantic Web II, International Workshops URSW 2008-2010 Held at ISWC and UniDL 2010 Held at FLoC, Revised Selected Papers, volume 7123 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2013. (Cited on page 16.)

[Catherine *et al.* 2017] Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi and William Cohen. *Explainable entity-based recommendations with knowledge graphs*. arXiv preprint arXiv:1707.05254, 2017. (Cited on page 61.)

[Cattelani *et al.* 2014] Luca Cattelani, Cristina E. Manfredotti and Enza Messina. *A Particle Filtering Approach for Tracking an Unknown Number of Objects with Dynamic Relations*. J. Math. Model. Algorithms Oper. Res., vol. 13, no. 1, pages 3–21, 2014. (Cited on page 4.)

[Chajewska *et al.*] Urszula Chajewska, Daphne Koller and Ronald Parr. *Making Rational Decisions Using Adaptive Utility Elicitation*. In Proceedings of AAAI 2000. (Cited on page 107.)

[Chickering 2003] David Maxwell Chickering. *Optimal Structure Identification with Greedy Search*. J. Mach. Learn. Res., vol. 3, pages 507–554, March 2003. (Cited on page 21.)

[Cholissodin & Dewi 2017] Imam Cholissodin and Ratih Kartika Dewi. *Optimization of Healthy Diet Menu Variation using PSO-SA*. Journal of Information Technology and Computer Science, vol. 2, no. 1, page 28–40, Jun. 2017. (Cited on page 85.)

[Chung *et al.* 2015] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho and Yoshua Bengio. *Gated feedback recurrent neural networks*. In International conference on machine learning, pages 2067–2075. PMLR, 2015. (Cited on page 9.)

[Cohen *et al.* 2003] William W. Cohen, Pradeep Ravikumar and Stephen E. Fienberg. *A Comparison of String Distance Metrics for Name-Matching Tasks*. In Subbarao Kambhampati and Craig A. Knoblock, editors, Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico, pages 73–78, 2003. (Cited on page 58.)

[Cooper & Herskovits 1992] Gregory F. Cooper and Edward Herskovits. *A Bayesian method for the induction of probabilistic networks from data*. Machine Learning, vol. 9, no. 4, pages 309–347, Oct 1992. (Cited on page 40.)

[Ćutić & Gini 2014] Darija Ćutić and Giuseppina Gini. *Creating causal representations from ontologies and Bayesian networks*. 2014. (Cited on page 22.)

[da Costa *et al.* 2008] Paulo Cesar G. da Costa, Kathryn B. Laskey and Kenneth J. Laskey. *PR-OWL: A Bayesian Ontology Language for the Semantic Web*. In Paulo Cesar G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool, editors, Uncertainty Reasoning for the Semantic Web I, ISWC International Workshops, URSW 2005-2007, Revised Selected and Invited Papers, volume 5327 of *Lecture Notes in Computer Science*, pages 88–107. Springer, 2008. (Cited on page 16.)

[de Campos & Ji 2008] Cassio P. de Campos and Qiang Ji. *Improving Bayesian Network parameter learning using constraints*. In 19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA, pages 1–4. IEEE Computer Society, 2008. (Cited on page 23.)

[De Campos *et al.* 2009] C.P. De Campos, Z. Zhi and Q. Ji. *Structure Learning of Bayesian Networks Using Constraints*. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 113–120, New York, USA, 2009. ACM. (Cited on page 23.)

[Delporte *et al.* 2014] Julien Delporte, Stéphane Canu and Alexandros Karatzoglou. *Apprentissage et Factorisation pour la Recommandation*. Revue des Nouvelles Technologies de l'Information, vol. Apprentissage Artificiel et Fouille de Données, RNTI-A-6, pages 1–26, 2014. (Cited on pages 60 and 61.)

[Despres 2014] Sylvie Despres. *Construction d'une ontologie modulaire pour l'univers de la cuisine numérique*. In Catherine Faron-Zucker. IC - 25émes Journées francophones d'Ingénierie des Connaissances, May 2014, Clermont-Ferrand, France, number 1, pages pp.27–38, May 2014. (Cited on page 23.)

[Després 2016] Sylvie Després. *Construction d'une ontologie modulaire. Application au domaine de la cuisine numérique*. Rev. d'Intelligence Artif., vol. 30, no. 5, pages 509–532, 2016. (Cited on page 67.)

[Devitt *et al.* 2006] Ann Devitt, Boris Danev and Katarina Matusikova. *Constructing Bayesian networks automatically using ontologies*. Applied Ontology, vol. 0, 2006. (Cited on page 15.)

[Doan *et al.* 2012] AnHai Doan, Alon Y. Halevy and Zachary G. Ives. Principles of data integration. Morgan Kaufmann, 2012. (Cited on page 16.)

[Dong *et al.* 2017] Xin Dong, Lei Yu, Zhonghuo Wu, Yuxia Sun, Lingfeng Yuan and Fangxi Zhang. *A hybrid collaborative filtering model with deep structure for recommender systems*. In Proceedings of the AAAI Conference on artificial intelligence, volume 31, 2017. (Cited on page 61.)

[Dooley *et al.* 2018] Damion M. Dooley, Emma J. Griffiths, Gurinder Gosal, Pier Luigi Buttigieg, Robert Hoehndorf, Matthew Lange, Lynn M. Schriml, Fiona S. L. Brinkman and William W. L. Hsiao. *FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration*. In npj Science of Food, volume 2, page 23, 2018. (Cited on pages 9 and 67.)

[Dooley *et al.* 2021] Damion Dooley, Magalie Weber, Liliana Ibanescu, Matthew Lange, Lauren Chan, Larisa Soldatova, Hande K McGinty, Chen Yang and William Hsiao. *Food Process Ontology Requirements*. IFOW 2021 Integrated Food Ontology Workshop, September 15-18, Bolzano Italy, 2021. (Cited on page 9.)

[Elsweiler & Harvey 2015] David Elsweiler and Morgan Harvey. *Towards Automatic Meal Plan Recommendations for Balanced Nutrition*. In Hannes Werthner, Markus Zanker, Jennifer Golbeck and Giovanni Semeraro, editors, Proceedings of

the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015, pages 313–316. ACM, 2015. (Cited on pages 62 and 85.)

[Ettouzi *et al.* 2016] Nourhene Ettouzi, Philippe Leray and Montassar Ben Messaoud. *An Exact Approach to Learning Probabilistic Relational Model*. In Alessandro Antonucci, Giorgio Corani and Cassio Polpo Campos, editors, Proceedings of the Eighth International Conference on Probabilistic Graphical Models, pages 171–182, 2016. (Cited on page 22.)

[Fenz 2012] Stefan Fenz. *An ontology-based approach for constructing Bayesian networks*. Data Knowl. Eng., vol. 73, pages 73–88, 2012. (Cited on page 15.)

[Ferrara *et al.* 2013] Alfio Ferrara, Andriy Nikolov and François Scharffe. *Data Linking*. J. Web Semant., vol. 23, page 1, 2013. (Cited on page 57.)

[Freyne & Berkovsky 2010] Jill Freyne and Shlomo Berkovsky. *Intelligent food planning: personalized recipe recommendation*. In Charles Rich, Qiang Yang, Marc Cavazza and Michelle X. Zhou, editors, Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010, pages 321–324. ACM, 2010. (Cited on page 62.)

[Fridman Noy 2004] Natalya Fridman Noy. *Semantic Integration: A Survey Of Ontology-Based Approaches*. SIGMOD Record, vol. 33, no. 4, pages 65–70, 2004. (Cited on page 16.)

[Friedman *et al.* 1999] Nir Friedman, Lise Getoor, Daphne Koller and Avi Pfeffer. *Learning Probabilistic Relational Models*. In Thomas Dean, editor, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages, pages 1300–1309. Morgan Kaufmann, 1999. (Cited on pages 19 and 22.)

[Ge *et al.* 2015] Mouzhi Ge, Francesco Ricci and David Massimo. *Health-aware Food Recommender System*. In Hannes Werthner, Markus Zanker, Jennifer Golbeck and Giovanni Semeraro, editors, Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015, pages 333–334. ACM, 2015. (Cited on page 84.)

[Getoor & Taskar 2007] Lise Getoor and Ben Taskar. Introduction to statistical relational learning (adaptive computation and machine learning). The MIT Press, 2007. (Cited on pages 22 and 33.)

[Glorot *et al.* 2011] Xavier Glorot, Antoine Bordes and Yoshua Bengio. *Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach*. In Lise Getoor and Tobias Scheffer, editors, Proceedings of the 28th International

Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, pages 513–520. Omnipress, 2011. (Cited on page 58.)

[Gonzales *et al.* 2015] Christophe Gonzales, Séverine Dubuisson and Cristina E. Manfredotti. *A New Algorithm for Learning Non-Stationary Dynamic Bayesian Networks With Application to Event Detection*. In Ingrid Russell and William Eberle, editors, Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida, USA, May 18-20, 2015, pages 564–569. AAAI Press, 2015. (Cited on pages 7 and 58.)

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep learning. MIT Press, 2016. http://www.deeplearningbook.org. (Cited on page 98.)

[Grüninger & Fox 1995] Michael Grüninger and Mark S. Fox. The role of competency questions in enterprise engineering. 1995. (Cited on page 54.)

[Guàrdia-Sebaoun *et al.* 2015] Elie Guàrdia-Sebaoun, Vincent Guigue and Patrick Gallinari. *Latent trajectory modeling: A light and efficient way to introduce time in recommender systems*. In Proceedings of the 9th ACM Conference on Recommender Systems, pages 281–284, 2015. (Cited on page 85.)

[Guarino *et al.* 2009] Nicola Guarino, Daniel Oberle and Steffen Staab. *What Is an Ontology?* In Steffen Staab and Rudi Studer, editors, Handbook on Ontologies, International Handbooks on Information Systems, pages 1–17. Springer Berlin Heidelberg, 2009. (Cited on page 17.)

[Guo *et al.* 2020] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong and Qing He. *A survey on knowledge graph-based recommender systems*. IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pages 3549–3568, 2020. (Cited on page 61.)

[Hageman PA 2014] Hertzog M Boeckner LS. Hageman PA Pullen CH. *Effectiveness of tailored lifestyle interventions, using web-based and print-mail, for reducing blood pressure among rural women with prehypertension: main results of the Wellness for Women: DASHing towards Health clinical trial.* Int J Behav Nutr Phys Act., vol. Dec 6, pages 11–148., 2014. (Cited on page 60.)

[Hansen 1982] Lars Peter Hansen. *Large sample properties of generalized method of moments estimators*. Econometrica: Journal of the Econometric Society, pages 1029–1054, 1982. (Cited on page 111.)

[Harvey *et al.* 2013] Morgan Harvey, Bernd Ludwig and David Elsweiler. *You Are What You Eat: Learning User Tastes for Rating Prediction*. In Oren Kurland, Moshe Lewenstein and Ely Porat, editors, String Processing and Information Retrieval

- 20th International Symposium, SPIRE 2013, Jerusalem, Israel, October 7-9, 2013, Proceedings, volume 8214 of *Lecture Notes in Computer Science*, pages 153–164. Springer, 2013. (Cited on page 62.)

[Hauser & Bühlmann 2014] Alain Hauser and Peter Bühlmann. *Two optimal strategies for active learning of causal models from interventional data*. Int. J. Approx. Reasoning, vol. 55, pages 926–939, 2014. (Cited on page 21.)

[Haussmann *et al.* 2019a] Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness and Mohammed J. Zaki. *FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation*. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois and Fabien Gandon, editors, The Semantic Web – ISWC 2019, pages 146–162, Cham, 2019. Springer International Publishing. (Cited on page 67.)

[Haussmann *et al.* 2019b] Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James V. Codella, Ching-Hua Chen, Deborah L. McGuinness and Mohammed J. Zaki. *FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation*. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois and Fabien Gandon, editors, The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II, volume 11779 of *Lecture Notes in Computer Science*, pages 146–162. Springer, 2019. (Cited on page 85.)

[Helsper & van der Gaag 2002] Eveline M. Helsper and Linda C. van der Gaag. *Building Bayesian Networks through Ontologies*. In Frank van Harmelen, editor, Proceedings of the 15th Eureopean Conference on Artificial Intelligence, ECAI'2002, Lyon, France, July 2002, pages 680–684. IOS Press, 2002. (Cited on page 16.)

[Hobbs & Pan 2004] Jerry R. Hobbs and Feng Pan. *An ontology of time for the semantic web*. ACM Trans. Asian Lang. Inf. Process., vol. 3, no. 1, pages 66–85, 2004. (Cited on page 26.)

[Hoffmann 2003] I. Hoffmann. *Transcending reductionism in nutrition research*. Am J Clin Nutr.h, vol. 78(3 Suppl), pages 514S–516S, 2003. (Cited on page 63.)

[Hung & Yamanishi 2021] PT Hung and K Yamanishi. *Word2vec Skip-Gram Dimensionality Selection via Sequential Normalized Maximum Likelihood*. Entropy, vol. 23, page 997, Jul 2021. (Cited on page 88.)

[Hunter *et al.* 2004] David R Hunter*et al. MM algorithms for generalized Bradley-Terry models*. The annals of statistics, vol. 32, no. 1, pages 384–406, 2004. (Cited on page 111.)

[Ibanescu *et al.* 2016] Liliana Ibanescu, Juliette Dibie, Stéphane Dervaux, Elisabeth Guichard and Joe Raad. *PO^2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels*. In Emmanouel Garoufallou, Imma Subirats Coll, Armando Stellato and Jane Greenberg, editors, Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings, volume 672 of *Communications in Computer and Information Science*, pages 155–165, 2016. (Cited on pages 6, 17 and 35.)

[Ishak *et al.* 2011] Mouna Ben Ishak, Philippe Leray and Nahla Ben Amor. *A Two-way Approach for Probabilistic Graphical Models Structure Learning and Ontology Enrichment*. In Joaquim Filipe and Jan L. G. Dietz, editors, KEOD 2011 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Paris, France, 26-29 October, 2011, pages 189–194. SciTePress, 2011. (Cited on page 15.)

[Ivens 2016] Smith Edge M. Ivens B. J. *Translating the Dietary Guidelines to Promote Behavior Change: Perspectives from the Food and Nutrition Science Solutions Joint Task Force.* J Acad Nutr Diet, vol. 116, no. 10, pages 1697–1702, 2016. (Cited on pages 8 and 60.)

[Jaccard 1912] Paul Jaccard. *The distribution of the flora in the alpine zone. 1*. New phytologist, vol. 11, no. 2, pages 37–50, 1912. (Cited on page 72.)

[Jacquet *et al.* 2024] Noémie Jacquet, Vincent Guigue, Cristina Manfredotti, Fatiha Saïs, Stephane Dervaux and Paolo Viappiani. *Modélisation du caractère séquentiel des repas pour améliorer la performance d'un système de recommandation alimentaire*. In Jérôme Gensel and Christophe Cruz, editors, 24èmes conférence Francophones sur l'Extraction et la Gestion des Connaissances, EGC 2024, Dijon, France, 22-26 Janvier, 2024, Revue des Nouvelles Technologies de l'Information. Hermann-Éditions, 2024. (Cited on pages 100 and 104.)

[Jaeger 1997] Manfred Jaeger. *Relational Bayesian Networks*. In Dan Geiger and Prakash P. Shenoy, editors, UAI '97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Brown University, Providence, Rhode Island, USA, August 1-3, 1997, pages 266–273. Morgan Kaufmann, 1997. (Cited on page 2.)

[Kamishima *et al.* 2010] Toshihiro Kamishima, Hideto Kazawa and Shotaro Akaho. *A Survey and Empirical Comparison of Object Ranking Methods*. In Johannes Fürnkranz and Eyke Hüllermeier, editors, Preference Learning, pages 181–201. Springer, 2010. (Cited on page 109.)

[Kang & McAuley 2018] Wang-Cheng Kang and Julian McAuley. *Self-attentive sequential recommendation*. In 2018 IEEE international conference on data mining (ICDM), pages 197–206. IEEE, 2018. (Cited on page 101.)

[Kluver & Konstan 2014] Daniel Kluver and Joseph A. Konstan. *Evaluating recommender behavior for new users*. In Alfred Kobsa, Michelle X. Zhou, Martin Ester and Yehuda Koren, editors, Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014, pages 121–128. ACM, 2014. (Cited on page 84.)

[Koller & Friedman 2009] Daphne Koller and Nir Friedman. Probabilistic graphical models: Principles and techniques - adaptive computation and machine learning. The MIT Press, 2009. (Cited on pages 2 and 19.)

[Konczak & Lang 2005] Kathrin Konczak and Jerome Lang. *Voting procedures with incomplete preferences*. Proceedings of the Multidisciplinary IJCAI-05 Workshop on Advances in Preference Handling, 01 2005. (Cited on pages 105 and 106.)

[Lau & Baldwin 2016] Jey Han Lau and Timothy Baldwin. *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation*. In Phil Blunsom, Kyunghyun Cho, Shay B. Cohen, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston and Scott Wen-tau Yih, editors, Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016, pages 78–86. Association for Computational Linguistics, 2016. (Cited on page 76.)

[Lee & Seung 1999] D. D. Lee and H. S. Seung. *Learning the parts of objects by non-negative matrix factorization*. Nature, vol. 401(6755), page 788–791, 1999. (Cited on page 75.)

[Li & Zhou 2007] Xiao-Lin Li and Zhi-Hua Zhou. *Structure Learning of Probabilistic Relational Models from Incomplete Relational Data*. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic and Andrzej Skowron, editors, Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings, volume 4701 of *Lecture Notes in Computer Science*, pages 214–225. Springer, 2007. (Cited on page 22.)

[Lipton 2018] Zachary C. Lipton. *The mythos of model interpretability*. Commun. ACM, vol. 61, no. 10, pages 36–43, 2018. (Cited on page 3.)

[Lops *et al.* 2019] Pasquale Lops, Dietmar Jannach and Cataldo Musto. *Trends in content-based recommendation*. User Model User-Adap Inter, vol. 29, pages 239 – 249, 04 2019. (Cited on page 61.)

[Lukasiewicz & Straccia 2008] Thomas Lukasiewicz and Umberto Straccia. *Managing uncertainty and vagueness in description logics for the Semantic Web*. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, no. 4, pages 291 – 308, 2008. Semantic Web Challenge 2006/2007. (Cited on page 16.)

[Luo *et al.* 2014] Xin Luo, Mengchu Zhou, Yunni Xia and Qingsheng Zhu. *An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems*. IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pages 1273–1284, 2014. (Cited on page 76.)

[Ma *et al.* 2019] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma and Xiang Ren. *Jointly Learning Explainable Rules for Recommendation with Knowledge Graph*. In The World Wide Web Conference, WWW '19, page 1210–1221, New York, NY, USA, 2019. Association for Computing Machinery. (Cited on page 85.)

[Madigan *et al.* 1996] David Madigan, Steen A Andersson, Michael D Perlman and Chris T Volinsky. *Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs*. Communications in Statistics–Theory and Methods, vol. 25, no. 11, pages 2493–2519, 1996. (Cited on page 19.)

[Manfredotti *et al.* 2011] Cristina E. Manfredotti, David J. Fleet, Howard J. Hamilton and Sandra Zilles. *Simultaneous Tracking and Activity Recognition*. In IEEE 23rd International Conference on Tools with Artificial Intelligence, ICTAI 2011, Boca Raton, FL, USA, November 7-9, 2011, pages 189–196. IEEE Computer Society, 2011. (Cited on page 4.)

[Manfredotti *et al.* 2013] Cristina E. Manfredotti, Kim Steenstrup Pedersen, Howard J. Hamilton and Sandra Zilles. *Learning Models of Activities Involving Interacting Objects*. In Allan Tucker, Frank Höppner, Arno Siebes and Stephen Swift, editors, Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. Proceedings, volume 8207 of *Lecture Notes in Computer Science*, pages 285–297. Springer, 2013. (Cited on page 4.)

[Manfredotti *et al.* 2015] Cristina E. Manfredotti, Cédric Baudrit, Juliette Dibie-Barthélemy and Pierre-Henri Wuillemin. *Mapping Ontology with Probabilistic Relational Models*. In Ana L. N. Fred, Jan L. G. Dietz, David Aveiro, Kecheng Liu and Joaquim Filipe, editors, KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 2, Lisbon, Portugal, November 12-14, 2015, pages 171–178. SciTePress, 2015. (Cited on pages 4, 6, 15, 22, 23, 28, 33 and 35.)

[Manfredotti 2009] Cristina E. Manfredotti. *Modeling and Inference with Relational Dynamic Bayesian Networks*. In Yong Gao and Nathalie Japkowicz, editors, Advances in Artificial Intelligence, 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009, Kelowna, Canada, May 25-27, 2009, Proceedings, volume 5549 of *Lecture Notes in Computer Science*, pages 287–290. Springer, 2009. (Cited on page 2.)

[Massimo *et al.* 2017] David Massimo, Mehdi Elahi, Mouzhi Ge and Francesco Ricci. *Item Contents Good, User Tags Better: Empirical Evaluation of a Food Recommender System*. In Mária Bieliková, Eelco Herder, Federica Cena and Michel C. Desmarais, editors, Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP 2017, Bratislava, Slovakia, July 09 - 12, 2017, pages 373–374. ACM, 2017. (Cited on page 62.)

[Maudet *et al.* 2005] Nicolas Maudet, Ulle Endriss and Yann Chevaleyre. *A Short Introduction to Computational Social Choice*. In Proceedings of the 33rd conference on Current Trends in Theory and Practice of Computer Science, 01 2005. (Cited on page 104.)

[Maystre & Grossglauser 2015] Lucas Maystre and Matthias Grossglauser. *Fast and accurate inference of Plackett-Luce models*. Technical report, 2015. (Cited on page 111.)

[McAuley *et al.* 2015] Julian J. McAuley, Rahul Pandey and Jure Leskovec. *Inferring Networks of Substitutable and Complementary Products*. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu and Graham Williams, editors, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 785–794. ACM, 2015. (Cited on page 68.)

[Mijangos *et al.* 2017] Victor Mijangos, Gerardo Sierra and Azucena Montes. *Sentence level matrix representation for document spectral clustering*. Pattern Recognit. Lett., vol. 85, pages 29–34, 2017. (Cited on page 78.)

[Mikolov *et al.* 2013a] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*, 2013. (Cited on page 87.)

[Mikolov *et al.* 2013b] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado and Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality*. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani and Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems

2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 3111–3119, 2013. (Cited on pages 9, 75, 76 and 87.)

[Min *et al.* 2022] Weiqing Min, Chunlin Liu, Leyi Xu and Shuqiang Jiang. *Applications of knowledge graphs for food science and industry*. Patterns, vol. 3, no. 5, page 100484, 2022. (Cited on page 67.)

[Mosqueira-Rey *et al.* 2023] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán and Ángel Fernández-Leal. *Human-in-the-loop machine learning: a state of the art*. Artificial Intelligence Review, vol. 56, no. 4, pages 3005–3054, Apr 2023. (Cited on page 3.)

[Münch *et al.* 2021] Mélanie Münch, Patrice Buche, Cristina E. Manfredotti, Pierre-Henri Wuillemin and Hélène Angellier-Coussy. *A Process Reverse Engineering Approach Using Process and Observation Ontology and Probabilistic Relational Models: Application to Processing of Bio-composites for Food Packaging*. In Emmanouel Garoufallou, María Antonia Ovalle-Perandones and Andreas Vlachidis, editors, Metadata and Semantic Research - 15th International Conference, MTSR 2021, Virtual Event, November 29 - December 3, 2021, Revised Selected Papers, volume 1537 of *Communications in Computer and Information Science*, pages 3–15. Springer, 2021. (Cited on page 7.)

[Münch *et al.* 2022] Mélanie Münch, Patrice Buche, Stéphane Dervaux, Juliette Dibie, Liliana Ibanescu, Cristina E. Manfredotti, Pierre-Henri Wuillemin and Hélène Angellier-Coussy. *Combining ontology and probabilistic models for the design of bio-based product transformation processes*. Expert Syst. Appl., vol. 203, page 117406, 2022. (Cited on pages 7 and 53.)

[Murena & Cornuéjols 2016] Pierre-Alexandre Murena and Antoine Cornuéjols. *Minimum Description Length Principle applied to structure adaptation for classification under concept drift*. In 2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016, pages 2842–2849. IEEE, 2016. (Cited on page 58.)

[Murphy 2002] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002. (Cited on pages 2 and 30.)

[Münch *et al.* 2017] Melanie Münch, Pierre-Henri Wuillemin, Cristina E. Manfredotti, Juliette Dibie and Stéphane Dervaux. *Learning Probabilistic Relational Models Using an Ontology of Transformation Processes*. In Hervé Panetto, Christophe Debruyne, Walid Gaaloul, Mike P. Papazoglou, Adrian Paschke, Claudio Agostino Ardagna and Robert Meersman, editors, On the Move to Meaningful Internet Systems. OTM 2017 Conferences - Confederated International Conferences: CoopIS,

C&TC, and ODBASE 2017, Rhodes, Greece, October 23-27, 2017, Proceedings, Part II, volume 10574 of *Lecture Notes in Computer Science*, pages 198–215. Springer, 2017. (Cited on pages 6, 22, 32, 33, 35, 38, 40, 49 and 50.)

[Münch *et al.* 2018a] Melanie Münch, Pierre-Henri Wuillemin, Juliette Dibie, Cristina E. Manfredotti, Thomas Allard, Solange Buchin and Elisabeth Guichard. *Identifying Control Parameters in Cheese Fabrication Process Using Precedence Constraints*. In Larisa N. Soldatova, Joaquin Vanschoren, George A. Papadopoulos and Michelangelo Ceci, editors, Discovery Science - 21st International Conference, DS 2018, Limassol, Cyprus, October 29-31, 2018, Proceedings, volume 11198 of *Lecture Notes in Computer Science*, pages 421–434. Springer, 2018. (Cited on pages 6, 22, 46, 47 and 48.)

[Münch *et al.* 2018b] Melanie Münch, Pierre-Henri Wuillemin, Cristina E. Manfredotti and Juliette Dibie. *Towards interactive causal relation discovery driven by an ontology*. Technical report, https://hal.archives-ouvertes.fr/hal-01823862v1, 2018. (Cited on page 47.)

[Münch *et al.* 2019a] Melanie Münch, Juliette Dibie, Pierre-Henri Wuillemin and Cristina E. Manfredotti. *Towards Interactive Causal Relation Discovery Driven by an Ontology*. In Roman Barták and Keith W. Brawner, editors, Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019, pages 504–508. AAAI Press, 2019. (Cited on pages 4, 6, 22, 40, 43, 47, 48, 49, 51, 55 and 85.)

[Münch *et al.* 2019b] Melanie Münch, Juliette Dibie-Barthélemy, Pierre-Henri Wuillemin and Cristina E. Manfredotti. *Interactive Causal Discovery in Knowledge Graphs*. In Elena Demidova, Stefan Dietze, John G. Breslin, Simon Gottschalk, Philipp Cimiano, Basil Ell, Agnieszka Lawrynowicz, Laura Moss and Axel-Cyrille Ngonga Ngomo, editors, Joint Proceedings of the 6th International Workshop on Dataset PROFILing and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 27, 2019, volume 2465 of *CEUR Workshop Proceedings*, pages 78–93. CEUR-WS.org, 2019. (Cited on pages 44 and 55.)

[Naamani-Dery *et al.* 2015] Lihi Naamani-Dery, Inon Golan, Meir Kalech and Lior Rokach. *Preference Elicitation for Group Decisions Using the Borda Voting Rule*. Group Decision and Negotiation, vol. 24, no. 6, pages 1015–1033, 2015. (Cited on pages 107, 108 and 109.)

[Nanba *et al.* 2014] Hidetsugu Nanba, Toshiyuki Takezawa, Yoko Doi, Kazutoshi Sumiya and Miho Tsujita. *Construction of a cooking ontology from cooking*

recipes and patents. In A. J. Brush, Adrian Friday, Julie A. Kientz, James Scott and Junehwa Song, editors, The 2014 ACM Conference on Ubiquitous Computing, UbiComp '14 Adjunct, Seattle, WA, USA - September 13 - 17, 2014, pages 507–516. ACM, 2014. (Cited on page 67.)

[Nathani *et al.* 2019] Deepak Nathani, Jatin Chauhan, Charu Sharma and Manohar Kaul. *Learning attention-based embeddings for relation prediction in knowledge graphs*. arXiv preprint arXiv:1906.01195, 2019. (Cited on page 85.)

[Neapolitan 2003] Richard E. Neapolitan. Learning bayesian networks. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. (Cited on page 39.)

[Newby & Tucker 2004] P. K. Newby and Katherine L. Tucker. *Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review*. Nutrition Reviews, vol. 62, no. 5, pages 177–203, 05 2004. (Cited on page 63.)

[Niles & Pease 2001] Ian Niles and Adam Pease. *Towards a standard upper ontology*. In Formal Ontology in Information Systems, 2001. (Cited on page 23.)

[O'Callaghan *et al.* 2017] Tom F. O'Callaghan, David T. Mannion, Deirdre Hennessy, Stephen McAuliffe, Maurice G. O'Sullivan, Natasha Leeuwendaal, Tom P. Beresford, Pat Dillon, Kieran N. Kilcawley, Jeremiah J. Sheehan, R. Paul Ross and Catherine Stanton. *Effect of pasture versus indoor feeding systems on quality characteristics, nutritional composition, and sensory and volatile properties of full-fat Cheddar cheese*. Journal of Dairy Science, vol. 100, no. 8, pages 6053 – 6073, 2017. (Cited on page 48.)

[Pan *et al.* 2005] Rong Pan, Zhongli Ding, Yang Yu and Yun Peng. *A Bayesian Network Approach to Ontology Mapping*. In Yolanda Gil, Enrico Motta, V. Richard Benjamins and Mark A. Musen, editors, The Semantic Web - ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings, volume 3729 of *Lecture Notes in Computer Science*, pages 563–577. Springer, 2005. (Cited on page 16.)

[Parviainen & Koivisto 2013] Pekka Parviainen and Mikko Koivisto. *Finding Optimal Bayesian Networks Using Precedence Constraints*. Journal of Machine Learning Research, vol. 14, pages 1387–1415, 2013. (Cited on page 40.)

[Pearl 2009] Judea Pearl. Causality: Models, reasoning and inference. Cambridge University Press, New York, USA, 2nd édition, 2009. (Cited on page 21.)

[Pecune *et al.* 2020] Florian Pecune, Lucile Callebert and Stacy Marsella. *A Recommender System for Healthy and Personalized Recipe Recommendations*. 09 2020. (Cited on page 86.)

[Qi *et al.* 2010] Guilin Qi, Qiu Ji, Jeff Z. Pan and Jianfeng Du. *PossDL - A Possibilistic DL Reasoner for Uncertainty Reasoning and Inconsistency Handling*. In The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Proceedings, Part II, pages 416–420, 2010. (Cited on page 16.)

[Quadrana *et al.* 2018] Massimo Quadrana, Paolo Cremonesi and Dietmar Jannach. *Sequence-Aware Recommender Systems*. ACM Comput. Surv., vol. 51, no. 4, pages 66:1–66:36, 2018. (Cited on page 85.)

[Rand 1971] William M. Rand. *Objective Criteria for the Evaluation of Clustering Methods*. Journal of the American Statistical Association, vol. 66, no. 336, pages 846–850, 1971. (Cited on page 81.)

[Reedy *et al.* 2009] Jill Reedy, Elisabet Wirfält, Andrew Flood, Panagiota N. Mitrou, Susan M. Krebs-Smith, Victor Kipnis, Douglas Midthune, Michael Leitzmann, Albert Hollenbeck, Arthur Schatzkin and Amy F. Subar. *Comparing 3 Dietary Pattern Methods—Cluster Analysis, Factor Analysis, and Index Analysis—With Colorectal Cancer Risk: The NIH–AARP Diet and Health Study*. American Journal of Epidemiology, vol. 171, no. 4, pages 479–487, 12 2009. (Cited on page 63.)

[Rendle *et al.* 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner and Lars Schmidt-Thieme. *BPR: Bayesian Personalized Ranking from Implicit Feedback*. UAI '09, page 452–461, Arlington, Virginia, USA, 2009. AUAI Press. (Cited on page 91.)

[Rendle *et al.* 2010] Steffen Rendle, Christoph Freudenthaler and Lars Schmidt-Thieme. *Factorizing personalized markov chains for next-basket recommendation*. In Proceedings of the 19th international conference on World wide web, pages 811–820, 2010. (Cited on page 101.)

[Rendle *et al.* 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner and Lars Schmidt-Thieme. *BPR: Bayesian personalized ranking from implicit feedback*. arXiv preprint arXiv:1205.2618, 2012. (Cited on page 102.)

[Rep 2003] *World Health Organization. Diet, nutrition and the prevention of chronic diseases: report of a joint who/fao expert consultation.*, 2003. (Cited on pages 8 and 59.)

[Ricci *et al.* 2015] Francesco Ricci, Lior Rokach and Bracha Shapira. *Recommender systems: introduction and challenges*. Recommender systems handbook, pages 1–34, 2015. (Cited on pages 60 and 61.)

[Rousseeuw 1987] Peter J. Rousseeuw. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, vol. 20, pages 53–65, 1987. (Cited on page 76.)

[Said & Bellogín 2014] Alan Said and Alejandro Bellogín. *Comparative recommender system evaluation: benchmarking recommendation frameworks*. In Proceedings of the 8th ACM Conference on Recommender systems, pages 129–136, 2014. (Cited on page 102.)

[Saïs & Thomopoulos 2014] Fatiha Saïs and Rallou Thomopoulos. *Ontology-aware prediction from rules: A reconciliation-based approach*. Knowl.-Based Syst., vol. 67, pages 117–130, 2014. (Cited on page 16.)

[Santiago-López *et al.* 2018] Lourdes Santiago-López, Jose E. Aguilar-Toalá, Adrián Hernández-Mendoza, Belinda Vallejo-Cordoba, Andrea M. Liceaga and Aarón F. González-Córdova. *Invited review: Bioactive compounds produced during cheese ripening and health effects associated with aged cheese consumption*. Journal of Dairy Science, vol. 101, no. 5, pages 3742 – 3757, 2018. (Cited on page 48.)

[Sarwar *et al.* 2001] Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl. *Item-based collaborative filtering recommendation algorithms*. In Proceedings of the 10th international conference on World Wide Web, pages 285–295, 2001. (Cited on page 60.)

[Sheridan *et al.* 2019] Paul Sheridan, Mikael Onsjö, Claudia Jeanneth Becerra, Sergio Jiménez and George Dueñas. *An Ontology-Based Recommender System with an Application to the Star Trek Television Franchise*. Future Internet, vol. 11, no. 9, page 182, 2019. (Cited on page 61.)

[Shim JS 2014] Kim HC. Shim JS Oh K. *Dietary assessment methods in epidemiologic studies.* Epidemiol Health., vol. Jul 22, page 36, 2014. (Cited on page 65.)

[Shin 2021] Donghee Shin. *The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI*. International Journal of Human-Computer Studies, vol. 146, page 102551, 2021. (Cited on page 61.)

[Silva *et al.* 2022] Vanderlei Carneiro Silva, Bartira Gorgulho, Dirce Maria Marchioni, Sheila Maria Alvim, Luana Giatti, Tânia Aparecida de Araujo, Angelica Castilho Alonso, Itamar de Souza Santos, Paulo Andrade Lotufo and Isabela Martins Benseñor. *Recommender System Based on Collaborative Filtering for Personalized Dietary Advice: A Cross-Sectional Analysis of the ELSA-Brasil Study*. International Journal of Environmental Research and Public Health, vol. 19, no. 22, 2022. (Cited on page 86.)

[Singh & Deepak 2022] Siddhant Singh and Gerard Deepak. *OntoRecipe: An Ontology Focussed Semantic Strategy for Recipe Recommendation*. In Saad Motahhir and Badre Bossoufi, editors, Digital Technologies and Applications, pages 21–33, Cham, 2022. Springer International Publishing. (Cited on page 67.)

[Snae & Bruckner 2008] Chakkrit Snae and Michael Bruckner. *FOODS: A Food-Oriented Ontology-Driven System*. In 2008 2nd IEEE International Conference on Digital Ecosystems and Technologies, pages 168–176, 2008. (Cited on page 67.)

[Spirtes *et al.* 2000] P. Spirtes, C. Glymour and R. Scheines. Causation, prediction, and search. MIT press, 2nd édition, 2000. (Cited on pages 21 and 49.)

[Staab & Studer 2009] Steffen Staab and Rudi Studer, editors. Handbook on ontologies, International Handbooks on Information Systems. Springer, 2009. (Cited on page 17.)

[Sutskever *et al.* 2014] Ilya Sutskever, Oriol Vinyals and Quoc V Le. *Sequence to sequence learning with neural networks*. Advances in neural information processing systems, vol. 27, 2014. (Cited on page 61.)

[Teng *et al.* 2012] ChunYuen Teng, Yu-Ru Lin and Lada A. Adamic. *Recipe recommendation using ingredient networks*. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy and Wolfgang Nejdl, editors, Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012, pages 298–307. ACM, 2012. (Cited on page 62.)

[Thorpe MG 2016] Crawford D McNaughton SA. Thorpe MG Milte CM. *A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians.* Int J Behav Nutr Phys Act., vol. Feb 29, pages 13–30, 2016. (Cited on page 63.)

[Thrun *et al.* 2005] Sebastian Thrun, Wolfram Burgard and Dieter Fox. Probabilistic robotics (intelligent robotics and autonomous agents). The MIT Press, 2005. (Cited on page 35.)

[Torti *et al.* 2010] Lionel Torti, Pierre-Henri Wuillemin and Christophe Gonzales. *Reinforcing the Object-Oriented Aspect of Probabilistic Relational Models*. In Proceedings of the 5th Probabilistic Graphical Models, pages 273–280, 2010. (Cited on pages 6 and 19.)

[Trattner & Elsweiler 2017] Christoph Trattner and David Elsweiler. *Food Recommender Systems: Important Contributions, Challenges and Future Research Directions*. CoRR, vol. abs/1711.02760, 2017. (Cited on pages 62 and 86.)

[Truong *et al.* 2005] Binh An Truong, Young-Koo Lee and Sungyoung Lee. *A Unified Context Model: Bringing Probabilistic Models to Context Ontology*. In

Tomoya Enokido, Lu Yan, Bin Xiao, Daeyoung Kim, Yuan-Shun Dai and Laurence Tianruo Yang, editors, Embedded and Ubiquitous Computing - EUC 2005 Workshops, EUC 2005 Workshops: UISW, NCUS, SecUbiq, USN, and TAUES, Nagasaki, Japan, December 6-9, 2005, Proceedings, volume 3823 of *Lecture Notes in Computer Science*, pages 566–575. Springer, 2005. (Cited on page 15.)

[Tumnark *et al.* 2019] Piyaporn Tumnark, Paulo Cardoso, Jorge Cabral and Filipe Conceição. *An Ontology to Integrate Multiple Knowledge Domains of Training-Dietary-Competition in Weightlifting: A Nutritional Approach: Nutritional Approach*. ECTI Transactions on Computer and Information Technology (ECTI-CIT), vol. 12, no. 2, page 140–152, Mar. 2019. (Cited on page 67.)

[Vandeputte *et al.* 2022] Jules Vandeputte, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere and Christine Martin. *Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits*. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud and Matthew E. Taylor, editors, 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022, pages 1292–1300. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022. (Cited on page 84.)

[Vandeputte *et al.* 2023] Jules Vandeputte, Pierrick Herold, Mykyt Kuslii, Paolo Viappiani, Laurent Muller, Christine Martin, Olga Davidenko, Fabien Delaere, Cristina Manfredotti, Antoine Cornuéjols and Nicolas N. Darcel. *Principles and Validations of an Artificial Intelligence-Based Recommender System Suggesting Acceptable Food Changes*. Journal of Nutrition, 2023. (Cited on pages 8 and 83.)

[Verny *et al.* 2017] Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh and Hervé Isambert. *Learning causal networks with latent variables from multivariate information in genomic data*. PLOS Computational Biology, vol. 13, no. 10, page e1005662, 2017. (Cited on page 21.)

[Viappiani & Boutilier 2020] Paolo Viappiani and Craig Boutilier. *On the equivalence of optimal recommendation sets and myopically optimal query sets*. Artificial Intelligence, vol. 286, page 103328, 2020. (Cited on page 108.)

[von Luxburg 2007] Ulrike von Luxburg. *A Tutorial on Spectral Clustering*, 2007. (Cited on page 77.)

[Wang *et al.* 2018] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie and Minyi Guo. *RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems*. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster and

Haixun Wang, editors, Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pages 417–426. ACM, 2018. (Cited on page 85.)

[Wang *et al.* 2019] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie and Minyi Guo. *Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation*. In The World Wide Web Conference, WWW '19, page 2000–2010, New York, NY, USA, 2019. Association for Computing Machinery. (Cited on page 61.)

[Webb 2015] Byrd-Bredbenner C. Webb D. *Overcoming consumer inertia to dietary guidance.* Advances in Nutrition, vol. 6, no. 4, pages 391–396, 2015. (Cited on pages 8 and 60.)

[Wendel *et al.* 2013] Sonja Wendel, Benedict Dellaert, Amber Ronteltap and Hans CM Trijp. *Consumers' Intention to Use Health Recommendation Systems to Receive Personalized Nutrition Advice*. Bmc Health Services Research, vol. 13, no. 126, pages 1–21, 2013. (Cited on page 63.)

[Woolhead *et al.* 2015] Clara Woolhead, Michael J Gibney, Marianne C Walsh, Lorraine Brennan and Eileen R Gibney. *A generic coding approach for the examination of meal patterns*. The American journal of clinical nutrition, vol. 102, no. 2, page 316—323, August 2015. (Cited on page 64.)

[Wuillemin & Torti 2012] Pierre-Henri Wuillemin and Lionel Torti. *Structured probabilistic inference*. Int. J. Approx. Reasoning, vol. 53, no. 7, pages 946–968, 2012. (Cited on page 19.)

[Yang & Calmet 2005] Yi Yang and Jacques Calmet. *OntoBayes: An Ontology-Driven Uncertainty Model*. In 2005 International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA 2005), International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2005), 28-30 November 2005, Vienna, Austria, pages 457–463. IEEE Computer Society, 2005. (Cited on page 16.)

[Yera Toledo *et al.* 2019] Raciel Yera Toledo, Ahmad A. Alzahrani and Luis Martínez. *A Food Recommender System Considering Nutritional Information and User Preferences*. IEEE Access, vol. 7, pages 96695–96711, 2019. (Cited on page 86.)

[Zanzotto 2019] Fabio Massimo Zanzotto. *Viewpoint: Human-in-the-loop Artificial Intelligence*. J. Artif. Intell. Res., vol. 64, pages 243–252, 2019. (Cited on page 3.)

[Zetlaoui *et al.* 2011] Mélanie Zetlaoui, Max Feinberg, Philippe Verger and Stephan Clémençon. *Extraction of Food Consumption Systems by Nonnegative Matrix*

Factorization (NMF) for the Assessment of Food Choices. Biometrics, vol. 67, no. 4, pages 1647–1658, 2011. (Cited on pages 63 and 76.)

[Zhang *et al.* 2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie and Wei-Ying Ma. *Collaborative Knowledge Base Embedding for Recommender Systems*. New York, NY, USA, 2016. Association for Computing Machinery. (Cited on page 61.)

[Zhang *et al.* 2019] Si Zhang, Hanghang Tong, Jiejun Xu and Ross Maciejewski. *Graph convolutional networks: a comprehensive review*. Computational Social Networks, vol. 6, no. 1, pages 1–23, 2019. (Cited on page 85.)

[Zheng *et al.* 2009] Jiaqian Zheng, Xiaoyuan Wu, Junyu Niu and Alvaro Bolivar. *Substitutes or complements: another step forward in recommendations*. In John Chuang, Lance Fortnow and Pearl Pu, editors, Proceedings 10th ACM Conference on Electronic Commerce (EC-2009), Stanford, California, USA, July 6–10, 2009, pages 139–146. ACM, 2009. (Cited on page 68.)

Curriculum Vitae

Présentation

Situation actuelle

Nom et prénoms	**MANFREDOTTI Cristina Elena**
Date et lieu de naissance	07 août 1979, Milan, Italie
Statut marital et nombre d'enfants	mariée, 2 enfants
Section CNECA	3
Grade et date d'accès au grade	maîtresse de conférences, classe normale, depuis le 01 septembre 2017
Echelon et date d'accès à l'échelon	échelon 7, depuis 01 juillet 2023
Établissement de rattachement pour les activités d'enseignement	UFR d'Informatique, Département MMIP (Modélisation Mathématiques Informatique Physique), AgroParisTech
Unité ou UMR de rattachement pour les activités de recherche	MIA Paris-Saclay, (Mathématique et Informatique Appliquées, MIA 518), département MATHNUM (MATHematique et NUMerique), INRAE
Coordonnées professionnelles	22, place de l'Agronomie 91120 Palaiseau Tel. (+33)1 89 10 0959 Bureau, E.3.712 E-mail : cristina.manfredotti@agroparistech.fr https://mia-ps.inrae.fr/cristina-manfredotti
Existence de temps partiel(s) et la ou les périodes concernées	20 septembre - 15 novembre, 2020 : arrêt maladie 25 février - 31 août, 2019 : congé maternité

Formations depuis l'entrée en fonction

Février 2024	**Agir contre les violences sexistes et sexuelles.** 1 heure et demi, module e-learning, Université Paris-Saclay.
Octobre 2017 - Juin 2018	**Cycle de formation des enseignants-chercheurs des établissements publics de l'enseignement supérieur agronomique, vétérinaire et de paysage**. 4 semaines, AgroParisTech.

Diplômes

2017 **Qualification**, Section 27 - Informatique, qualification n.17227241682

2010 **Doctorat en informatique**, mention *Eccellente* (excellent, en italien)
Università di Milano-Bicocca, Milan, Italie

Sujet	**Modelling and Inference with Relational Dynamic Bayesian Networks**
Dirigé par	*Enza Messina*, professeur, Università di Milano-Bicocca, Milan, Italie
	David J. Fleet, professeur, University of Toronto, Toronto, Canada
Jury	Francesco Archetti (President), professeur, Università di Milano-Bicocca, Milan, Italie
	Giovanni Gallo, professeur, Università di Catania, Catane, Italie
	Paolo Boldi, professeur, Università di Milano, Milan, Italie
Rapporteurs	*Cory Butz*, professeur, University of Regina, Regina, Canada
	Hans W. Guesgen, professeur, Massey University, Massey, Nouvelle-Zélande

2004 **Diplôme post-master en Mathématiques pour l'Industrie**
Istituto Nazionale di Alta Matematica, Milan, Italie

Sujet	**ImageJ plug-in for polyharmonic wavelet transform**
Encadrant	*Cédric Vonesch*, Biomedical Imaging Group (BIG), Ecole Polytechnique Fédérale du Lausanne (EPFL), Lausanne, Suisse

2003 **Laurea (diplôme de master) en Mathématiques**, mention 95/110
Università di Milano-Bicocca, Milan, Italie

Sujet	**Problemi matematici inerenti la Risonanza Magnetica**
Encadrant	*Mira Bozzini*, professeur, Università di Milano-Bicocca, Milan, Italie

Cursus Professionnel

Septembre 2017 - maintenant	**maîtresse de conférences de l'enseignement supérieur agricole**, Institut des sciences et industries du vivant et de l'environnement (AgroParistech), département MMIP, UFR Informatique, UMR MIA Paris-Saclay.
Septembre 2014 - Août 2017	**maîtresse de conférences contractuel de l'enseignement supérieur agricole**, Institut des sciences et industries du vivant et de l'environnement (AgroParistech), département MMIP, UFR Informatique, UMR MIA Paris.
Mars 2013 - Juillet 2014	**Postdoctorante** Département d'Informatique de Paris 6 (LIP6), Université Pierre et Marie Curie, Paris, France

Mars 2011- Octobre 2012	**Postdoctorante** Datalogisk Institut Kobenhavns Universitet (DIKU), Copenhague, Danemark
Janvier - Décembre 2010	**Postdoctorante** Department of Computer Science, University of Regina, Regina, Canada
Janvier 2007 - Décembre 2009	**Vacataire** pendant mes études de doctorat Dipartimento di Sistemistica e Comunicazione (DiSCo), Università di Milano-Bicocca (UniMiB), Milan, Italie
Janvier 2005 - Décembre 2006	**Assistante Chercheuse** Projet FIT COMERSON, Dipartimento di Sistemistica e Comunicazione (DiSCo), Università di Milano-Bicocca (UniMiB), Milan, Italie

Mobilité

Mai 2008 - Octobre 2009	**Doctorante visiteur**, Computer Science Department, University of Toronto, Toronto, Canada
Juin - Septembre 2004	**Stagiaire post-master**, Biomedical Imaging Group (BIG), Ecole Polytechnique Fédérale du Lausanne (EPFL), Lausanne, Suisse

Connaissance des langues

- **français** : niveau intermédiaire/avancé
- **anglais** : niveau supérieur/courant
- **italien** : langue maternelle

Activités d'enseignement

Cadre structurel de conduite des activités d'enseignement

Depuis septembre 2017, je suis maîtresse de conférences à **AgroParisTech**. Je suis membre du département MMIP (Modélisation, Mathématiques, Informatique et Physique) qui est constitué de 3 Unité de Formation et de Recherche (UFR) : l'UFR de Mathématiques Appliquées, l'UFR de Sciences Physiques pour l'Ingénieur et l'UFR d'Informatique. Le président du département MMIP est Christophe Doursat.

Je fais partie de l'UFR d'Informatique. Jusqu'à décembre 2019, cette UFR a été composée de deux professeurs, Antoine Cornuéjols et Juliette Dibie et de quatre maîtres de conférences, Michel Cartereau, Liliana Ibanescu, Christine Martin et moi même. En 2020, suite au départ de Juliette Dibie, une maîtresse de conférences contractuel, Chloé Vigliotti, nous a rejoint. Chloé a été, ensuite, confirmé, suite au départ à la retrait de Michel Cartereau. En Septembre 2022, Vincent Guigue a pris la place de Juliette Dibie et en Septembre 2023 Chloé a decidé de quitter AgroParisTech. L'UFR d'Informatique d'AgroParisTech est actuellement composé de deux professeurs, Antoine Cornuéjols et Vincent Guigue et de trois maîtres de conférences, Liliana Ibanescu, Christine Martin et moi même.

Depuis septembre 2019 la responsable de l'UFR d'Informatique d'AgroParisTech est Liliana Ibanescu. Entre septembre 2014 et août 2017, j'ai été maîtresse de conférences contractuel dans la même UFR.

Avant de venir à AgroParisTech, j'ai enseigné comme vacataire dans l'UFR d'Informatique de l'**Université Pierre et Marie Curie**[1] en tant que postdoctorante au Laboratoire d'Informatique de Paris 6 (LIP6), à l'Université de Copenhague au **Danemark** et pendant plusieurs années à l'Université de Milano-Bicocca en **Italie**, comme vacataire pendant mes études de doctorat.

Démarches pédagogiques, responsabilités assumées

Depuis septembre 2014, mon activité d'enseignement porte essentiellement sur les trois années du cursus des ingénieurs d'AgroParisTech. Elle est centrée autour de l'**apprentissage automatique** et des **langages de programmation** (Python et EXCEL VBA) :

- en première année, cursur ingénieurs : *Système d'Information et Programmation* (PHP, Python, SQLite) et *Système d'Information Geographiques* (QGis);
- en première année, cursus ingénieurs apprentis : *Système d'Information et Programmation* (Python, SQLite), *Bureautique Excel* et *Programmation en VBA pour EXCEL*;
- en deuxièmme année : *Programmation en VBA pour EXCEL*;
- en troisième année : *Algorithmiques*, *Apprentissage Automatique* et *Raisonnement bayésien*.

[1] Actuellement Sorbonne Université.

Je suis **responsable de deux unités d'enseignement qui impliquent la participation d'autres collègues**[2] :

- depuis septembre 2016, de l'unité d'enseignement de première année (cursus ingénieurs) *Système d'Information et Programmation*[3] (360 étudiants, 18 TDs, 4 enseignants en parallèle) et
- depuis septembre 2022, de l'unité d'enseignement de première année, cursus apprentis, programmation VBA pour Excel (60 étudiants 4 TDs, 2 enseignants en parallèle).

Services d'enseignement et réalisé pédagogique

Le tableau 6 donne pour chaque année universitaire depuis 2017-18 (l'année de mon recrutement comme maîtresse de conférences à AgroParisTech) le nombre d'heures d'enseignement effectué (colonne **Total**) et l'obligation de service (colonne **OS**). Le nombre d'heures est exprimé en heures équivalent TD (h éq TD) et un regroupement est fait selon le type d'enseignement : C (Cours), TD (Travaux Dirigés), Autre (encadrement, évaluation de stages et organisation). La grille d'équivalence est 128 h de cours = 192 h de TD (ou 1h de cours = 1,5 h éq TD).

Mon deuxième enfant est né en mars 2019 et mon obligation de service pour l'année 2018-2019 a été réduite à 128 h éq TD. En septembre 2020 il a été hospitalisé pour plusieurs mois, j'ai pris un congé de 38 jours pour rester à ses cotés, ce qui a donné une réduction de mon OS de 28 h éq TD. Étant donné l'emploi du temps de nos enseignements (la plupart au début de l'année universitaire) ma charge pour l'année universitaire 2020-2021 (mon OS de 164h éq TD) n'a pas pu être effectuée dans sa totalité.

	Nb heures en h éq TD					
Année	C	TD	Autre	**Total**	OS	Différence
2022–23	12.75	188.5	26.37	**227.18**	192	+35.18
2021–22	11.25	162	27.19	**200.44**	192	+8.44
2020–21	11.25	117	28.93	**157.18**	164	-6.82
2019–20	15.75	131	44.45	**192.20**	192	+0.20
2018–19	15.75	99	18.27	**133.02**	128	+5.02
2017–18	42.75	150	32.06	**224.81**	192	+32.81
TOTAL	**109.5**	**847.5.5**	**177.27**	**1134.83**	**1060**	+74.83
moyenne par année	15.6	121.07	25.32	162.11	151.42	10.69

Table 6: Nombre d'heures d'enseignement effectués depuis septembre 2017.

[2] J'ai choisi, ici, de distinguer les responsabilités pour les quelles je suis amenée à l'organisation et l'encadrement d'autres collègues et les responsabilités pour les quelles je suis toute seule.

[3] Depuis l'année universitaire 2022–2023 ce cours a changé de nom; il s'appelle *Informatique : Programmation e Bases de Données*.

Publications d'enseignement

- **Cristina Manfredotti**. *Programmation en VBA pour EXCEL*. Cours et fiches d'exercices, 70 pages. Années 2015–16, 2016–17. Document créé à partir des cours de Juliette Dibie (2008) et Christine Martin (2013)
- **Liliana Ibănescu et Cristina Manfredotti**. *Informatique: Programmation et bases de données. Bases de données et SQLite*. Cours et fiches d'exercices, 61 pages. Première année, tronc commun. Année 2021–22.

Encadrement de stage court de 2A En 2018 **j'ai co-encadré** avec Juliette Dibie le **stage facultatif d'élèves ingénieurs d'AgroParisTech en deuxième année** de Camille Bardon qui a développé le cours en-ligne sur la programmation VBA pour Excel.

Tutorat de stage de fin d'études Depuis septembre 2014 j'ai été tutrice école de **17** stages de fin d'études de 6 mois d'élèves ingénieurs en 3A:

2023 **Jules MARCAIS**, *Apprentissage machine et inférence de réseaux pour l'analyse de données multi-omiques et multi-échelles. Application à la prédiction de sévérité de brûlures radiologiques*. Entreprise: IRNS-Institut de radioprotection et de sûreté nucléaire.

2022 **Mathilde GUYOT**, *B*ig data : Optimisation d'un système de production pénéicole malgache. Entreprise: OSO Farming – Les Gambas de l'Ankarana.

2021 **Cecile CAUMETTE**, *E*tude de la dynamique de population de Bactrocera dorsalis, en lien avec la matrice paysagère d'un bassin de production de mangues. Entreprise: CIRAD

Alexis VERGNE, *D*eepBeesAlert : Vers un système de gestion et de protection durable des ressources de pollinisation. Entreprise: UMR botAnique et Modélisation de l'Architecture des Plantes et des végétations (AMAP)

Theophile ADOUARD, *D*étection automatique de qualité d'image subjective de coroscaners, quantification automatique de la réserve coronaire à partir de MPR (multi-planar reconstruction) curvilignes de coroscaners et développement d'une plate-forme d'apprentissage pour les internes en radiologie.. Entreprise: Spimed-AI

Aurelien BEAUDE, *D*éveloppement et amélioration des algorithmes de traitement d'images pour la classification des coroscanners. Entreprise: Spimed-AI

2020 **Clemence CHATUE**, *S*ummarization on Short Dialogues. Entreprise: Praelexis Stellenbosch, South Africa

Naomi BERDA, *E*stimation par satellite et drone des rendements céréaliers à l'échelle des paysages dans un système agro-forestier sénégalais. Entreprise: CIRAD (Sénégal) Centre de Suivi écologique de Dakar

Pauline MATHIEU, *P*rédiction de tendances dans differents secteurs de l'economie mondiale. Entreprise: Intellimind

Anaelle BADIER, *I*mplémentation d'Algorithmes d'Adaptive Learning Proposer un parcours pédagogique personnalisé au sein de l'application mobile Nomad Education. Entreprise: Nomad Education

2018 **Antoine MONIOT**, *I*nclusion of omics data in prediction of fluxes closest to biochemical constraints in a metabolic network in E.coli. Entreprise: Max Planck Institute of Molecular Plant Physiology

Benjamin DENEU, *P*rédiction géo-localisée de communautés végétales par apprentissage profond. Entreprise: INRIA

2017 **Quentin FALCAND**, Spécialisation en Informatique MISI (Management et Ingénierie du Système d'Information). *A*pplication de méthodes de machine learning et deep learning dans le secteur de l'agriculture. Etude à partir de ces outils de la propagation de maladies dans les exploitations viticoles en fonction de différents facteurs. Entreprise: Quantcube.

Riheng ZHU, Spécialisation en Informatique MISI (Management et Ingénierie du Système d'Information). *M*ise en oeuvre et évaluation de modèles statistiques et d'apprentissage automatique basés sur les données pour la gestion d'actifs des réseaux d'eau. Entreprise: Veolia Environment SA.

2015 **Louis Victor PASQUIER**, Spécialisation en Informatique MISI (Management et Ingénierie du Système d'Information). *A*nalyse et fouille de données chez COFELY. Entreprise: Cofely, GDF Suez.

Etienne DAVID, Spécialisation en Informatique MISI (Management et Ingénierie du Système d'Information). *D*etenction and segmentation with machine learning: a supervised method and basis for an unsupervised one. Entreprise: IPAL, Singapeur.

Adrien GIRAUD, Spécialisation en Informatique MISI (Management et Ingénierie du Système d'Information). *A*nalyse de comportements de consommation d'eau potable, du quartier au compteur. Entreprise: Veolia.

Tutorat de stage court de 2A Depuis septembre 2014 j'ai été tutrice école de **1** stages court de 3 mois d'un élève ingénieur en 2A:

2016 **Quentin FALCAND**, *Amélioration de l'outil Biodi(V)strict: Developpement d'un program informatique*. Entreprise: Laboratoire Ecologie, Systématique et Evolution.

Enseignant-référent Depuis septembre 2023 je suis enseignant-référent de **10** élèves du premier année : CARO Robin, BOURGET Merlin, BRIOT Agathe, ALEXANDRE Solene, BISCUIT Louison, BOUCHAMA Tifenn, DANDOY Timothe, DENEUVILLE Rosalie, DROUIN Emilie et COMBET Emma.

Participation et missions d'enseignement hors de l'établissement

Entre septembre 2014 et juin 2018 **j'ai participé à des enseignements en anglais** dans le master M2 en *Decision Support and Business Intelligence* à l'École Centrale de Paris, dont la responsable était Nacera Bennacer, professeur en Informatique à Centrale Supélec. Ce master

est cohabilité par AgroParisTech. Dans ce master j'ai assuré entre 60 et 75 h éq TD sur l'**apprentissage automatique** (supervisé et non) et le **raisonnement bayésien** dans l'UC *Data Mining and Machine Learning*.

Le master *Erasmus Mundus* met ensemble des étudiants ayant suivi un parcours d'étude différent. Le défi de ce type d'enseignement est d'intéresser les étudiants, tout en expliquant des concepts que certains d'entre eux ont déjà vu et, donc, d'être à la fois clair, simple et précis.

Activités d'intérêt général

Instances, commissions et groupes de travail

En 2022 j'ai été **élue au Conseil de l'enseignement et de la Vie Étudiante** (CEVE) comme suppléante de Jade Giguelay, maîtresse de conférences dans l'UFR de Mathématiques Appliquées.

Depuis septembre 2022, je fais partie de la **commission de titularisation pour les maîtres de conférences**.

Depuis septembre 2022 je suis représentante pour l'équipe Ekinocs au Conseil du Laboratoire MIA Paris-Saclay.

Depuis septembre 2016, je fais partie du **réseau Prisme** qui s'occupe de la stratégie internationale d'AgroParisTech et depuis janvier 2024 je fais partie du **groupe de travail pour la mobilité d'échange entrante à AgroParisTech**.

Participation à des jurys

En 2022 et 2021 j'ai participé aux jurys de concours pour le recrutement d'un maître de conférences contractuelle en Informatique.
J'ai participé à **3** jurys de thèse : **1** comme examinatrice (T. Hoa en 2018) et **2** comme encadrant (M. Bouyrie en 2016 et M. Munch en 2020).
J'ai participé à **15** jurys du concours d'admission pour les apprentis (en moyenne 3 jurys par an).
J'ai participé à **86** jurys de stage de troisième année pour la dominante IODAA (en moyenne 12 étudiants-jury/an).
En janvier 2024, j'ai été recruté pour la **jury du concours agronomiques et vétérinaires**. J'ai examiné 47 dossiers.

Activités de recherche et de développement

Cadre structurel

Après avoir obtenu mon doctorat, en 2010, et avoir suivi trois postdoctorats, depuis septembre 2014, je suis membre de l'**équipe EkINocs** (Expert Knowledge, INteractive modellING and

learnINg for understandINg and decisiOn makINg in dINamic Complexe Systems)[4]. L'équipe Ekinocs est l'une des deux équipes composant l'UMR MIA Paris-Saclay (Mathematique et Informatique Appliqué).

L'UMR Paris-Saclay est associée aux tutelles AgroParisTech, INRAE et Université Paris Saclay. Elle développe des méthodes mathématiques et informatiques avec une visée applicative, particulièrement dans les sciences du vivant, de l'environnement, l'agronomie et les sciences de l'alimentation. L'unité est rattachée au département MATHNUM d'INRAE et au département MMIP d'AgroParisTech. Le directeur de l'unité MIA Paris-Saclay est Julien Chiquet.

Au moment de mon recrutement, en 2014, l'équipe Ekinocs (à l'époque équipe LInK) comptait 5 personnels AgroParistech et 1 personnels INRAE : 2 professeurs, Antoine Cornuéjols (responsable de l'équipe) et Juliette Dibie, 3 maîtres de conférences, Liliana Ibanescu, Christine Martin et moi même et un ingénieur d'étude INRAE, Stéphane Dervaux. Dans les derniers années, les effectif de l'équipe ont changé beaucoup. En 2017, 2 chercheurs INRAE, Joon Kwon et George Katsirelos, ont rejoint l'équipe. En 2019, 4 chercheurs INRAE ont été rattaché à l'équipe Ekinocs : 1 chercheuse Nadia Boukhelifa et 3 directeurs de recherche Evelyne Lutton, Nathalie Mejean et Alberto Tonda. En 2020, suite au départ de Juliette Dibie (qui est maintenant chercheuse associée), une nouvelle maîtresse de conférences contractuel, Chloé Vigliotti a été recruté et elle a, ensuite, démissionné en 2023. En 2022 un nouveau professeur, Vincent Guigue, a été recruté et 2 chercheurs IPEF Sophie Martin et Isabelle Alvarez, ont été rattaché à l'équipe. Aujourd'hui, l'équipe Ekinocs est composée de 5 personnels AgroParisTech, 7 personnels INRAE, 2 IPEF et 1 chercheuse associé.

Alors qu'Antoine Cornéjols et Christine Martin sont experts en apprentissage automatique (en particulier apprentissage non supervisé, apprentissage par transfert et apprentissage en ligne, et systèmes de recommandation), l'expertise de Juliette Dibie, Stephane Dervaux et Liliana Ibanescu est dans le domaine de la représentation des connaissances et, en particulier, dans le domaine de l'intégration des données. Depuis mon recrutement, en 2014, **j'ai mis mon expertise au service de ce deux principales thématiques de recherche de l'équipe facilitant leur collaboration dans des projets communs**. **Étant données les nouvelles dimensions de d'équipe, le défi est, maintenant, de trouver des nouveaux intérêts pour des collaborations plus élargis**, chose qui est facilité par le déménagement à Palaiseau où nous sommes tous dans les même locaux.

Thème(s), projets (genèse, état actuel, perspective)

Mon domaine de recherche est l'**apprentissage artificiel**. En particulier, je m'intéresse à des systèmes automatiques qui doivent pouvoir raisonner efficacement sur les interactions entre plusieurs objets en tenant compte du contexte et en présence d'incertitude, apprendre du passé et s'adapter à la situation actuelle.

[4]L'equipe a changé de nom en 2019, anciennement elle s'appellait LInK (Learning and INtegration of Knowledge)

Avant de venir à AgroParisTech, j'ai étudié des modèles et des algorithmes pour résoudre des problèmes dans des domaines comportant de nombreuses relations entre différentes entités. En particulier j'ai travaillé sur les problèmes de suivi simultané de plusieurs objets et de reconnaissance d'activités (principalement dans des systèmes de vidéo-surveillance) en utilisant des méthodes d'inférence probabiliste.

Au cours des dernières années à AgroParisTech, j'ai concentré mon expertise sur les modèles probabilistes dans le domaine des sciences du vivant et de l'alimentation. En particulier, j'ai travaillé sur des processus de transformation et sur des systèmes de recommandation alimentaire.

Théme(s)

Depuis septembre 2014, à AgroParistech, je me suis intéressée principalement à deux thématiques de recherche : (1) **comment modéliser l'incertitude dans des processus de transformation** avec des modèles probabilistes en couplant une ontologie qui représente un processus de transformation et les modèles probabilistes relationnel, et (2) **comment améliorer un système de recommandations nutritionnel** en modélisant les interactions entre l'utilisateur et le contexte de sa consommation et la caractéristique séquentielle de la prise de décision dans ce contexte.

Projets

Au sein de l'équipe Ekinocs, j'ai participé au montage de trois projets AgroparisTech, différents projets ANR (un retenu) et un projet européen. J'ai également participé à la demande de différentes bourses de thèse. Dans la suite je liste ces projets en ordre chronologique. Certains sont en ligne avec mes thèmes de recherche dans des autres j'ai un rôle marginale.

Project ANR JCJC - Process-WAvE (2015, non retenu). *Studying Microorganisms Stabilization Processes during time and at different granularity With ProbAbilistic Relational ModEls*, Project ANR Jeunes Chercheuses Jeunes Chercheurs (JCJC).

Collaborateurs : **Cristina Manfredotti (porteuse)**, Juliette Dibie, Caroline PÉNICAUD (UMR782 GMPA Génie et Microbiologie des Procédés Alimentaires), Pierre-Henri Wuillemin (Sorbonne Université), Liliana Ibanescu, Fernanda Fonseca (UMR 782 Génie et Microbiologie des Procédés Alimentaires), Cedric Baudrit (Institut de Mécanique et d'Ingénierie).

Rôle personnel : **Écriture, Soumission**

Projet ANR DeliciouS (2015, non retenu). *A Data drivEn ontoLogy data warehouse: from healthy food to perception by specific population through teChnological, physIOlogical and nutritional transformation processeS*.

Projet AgroParisTech Transform (2015-2017). *Une nouvelle approche pour modéliser/ représenter des processus de Transformation combinant Ontologie et Modèles Relationnels probabilistes. Application à la stabilisation de micro-organismes : levures et bactéries*. Projet

de recherche collaboratif entre deux équipes d'AgroParisTech : l'équipe Ekinocs et l'équipe "Bio-produits, Aliments, Micro-organismes et Procédés" (BioMiP).

Participants : **équipe Ekinocs** et équipe "Bio-produits, Aliments, Micro-organismes et Procédés" (BioMiP).

Rôle personnel : **écriture et demande de subvention**, participante chercheuse, **encadrant du stage** de fin d'études de Melanie Munch. Publications : [P9, P13]

Projet financé par Danone Nutricia Research (2017-2021). *Conception et validation d'un système de recommandations alimentaires à partir de données de consommation*. Projet de recherche collaboratif entre deux unités mixtes de recherche INRA/AgroParisTech, l'équipe "Mathématiques et informatique Appliquées" (MIA-Paris) et l'équipe "Physiologie de la Nutrition et du Comportement Alimentaire" (PNCA) et Danone Nutricia Research.

Participants : **équipe Ekinocs**, équipe PNCA et Danone Nutricia Research.

Rôle personnel : chercheuse participant et **co-encadrant de la thèse** de Sema Akkoyunlu (33%).

Publications : [W6, W7]

A partir des données de consommation alimentaires journalières[5], nous avons étudié les co-occurrences de différents aliments pour trouver les contextes alimentaires dans lesquels un aliment est le plus souvent consommé. Une fois les contextes alimentaires d'un aliment découverts et étant donné un souhait d'un utilisateur à manger quelque chose, on peut donner des recommandations de substitutions de cet aliment, acceptables car ils respectent les contextes de l'aliment souhaité. Malheureusement, Sema a décidé d'interrompre sa collaboration avec nous avant sa soutenance.

Demande de thèse : alignement PRM-ontologies (2017-2020). *Améliorer le raisonnement dans l'incertain en combinant les modèles relationnels probabilistes et la connaissance experte*. École ABIES.

Participants : Juliette Dibie, Pierre-Henri Wuillemin (maître de conférences à l'Université Paris Sorbonne), **Cristina Manfredotti**.

Rôle personnel : **écriture et demande de subvention**, participant chercheuse, **co-encadrant de thèse** (33%).

Publications : [P12, P14, W8]

L'objectif de la thèse de Mélanie Munch qui a été soutenue en 2020, a été de guider l'apprentissage des relations probabilistes avec les connaissances d'experts dans des domaines décrits par des ontologies. Pour ce faire, des bases de connaissances ont été couplées avec les PRMs dans l'objectif de compléter l'apprentissage statistique par des connaissances expertes afin d'apprendre un modèle aussi proche que possible de la réalité et de l'analyser quantitativement (avec des relations probabilistes) et qualitativement (avec la découverte causale).

[5]Nous avons utilisé le données issues de l'étude INCA (étude individuelle nationale des consommations alimentaires)https://www.anses.fr/fr/glossaire/1205

Projet GRAPH-MATCHING (2018, 2019). Financement pour deux stages de fin d'étude.

Participants : Juliette Dibie, Fatiha Sais (Univesrité Paris-Saclay), **Cristina Manfredotti**.

Rôle personnel : **écriture et demande de subvention**, participant chercheuse, **co-encadrant des deux stages** (33%).

Dans les deux stages nous avons étudié des méthodes de *Graphs Matching* pour faire du transfert entre des domaines différents représentées par une même ontologie.

Projet ANR SHIFT (2018-2022). *Étude des dynamiques de changement de comportement alimentaire vers des régimes de meilleure qualité - Approches interdisciplinaires de la notion d'acceptabilité d'une proposition de substitution entre aliments*.

Participants : ISIR Institut des Systèmes Intelligents et Robotiques, INRA-ALISS Alimentation et Sciences Sociales, Danone-GND DANONE RESEARCH, University of Birmingham / School of Psychology, **MIA-Paris Mathématiques et Informatique Appliquées**, PNCA Physiologie de la Nutrition et du Comportement Alimentaire.

Rôle personnel : collaboration à l'écriture du projet, participant chercheuse.

Publications : [P5]

Projet DATAIA WarmRules (2019-2022). *Gradual Causal Rules Detection in Knowledge Graphs - Applications to Plant Development*.

Participants : **MIA-Paris**, LRI/LaHDAK, GQE - Le Moulon, INRA

Rôle personnel : participant chercheuse.

Projet MANGER ENSEMBLE (2020 et 2021). Collaboration de recherche entre Nicolas Darcel, Paolo Viappiani et moi même qui a permis de financer deux stages de fin d'étude.

Participants : Nicolas Darcel, Paolo Viappiani, **Cristina Manfredotti**.

Rôle personnel : **co-encadrante des stages** (33%).

Publications : [W9]

On mange rarement seul et satisfaire les préférences de plusieurs personnes ensemble est un des défis des systèmes de recommandation alimentaires. Dans le stage de Maeva Caillat, nous avons étudié des méthodes bayésiennes pour la recommandation à des groupes et l'elicitation interactive de préférences. Nous avons comparé différentes stratégies d'elicitation et, dans des simulations, on a amélioré la performance des algorithmes de recommandation aux groupes par rapport à l'état de l'art. Dans le stage de Youhan Wang, nous avons poursuivi cette étude et nous avons proposé une approche capable de passer à l'échelle basée sur les modelés de Placket Luce.

Projet DECHETS (2020-2022). *Modélisation des processus de Transformation combinant Ontologie et Modèles Relationnels probabilistes. Application à la production d'emballages alimentaires*. Projet de recherche collaboratif qui a permise le financement d'un stage de fin d'études et un postdoc.

Participants : Patrice Buche (ingénieur de recherche IATE-INRAE Montpellier) Stéphane Dervaux, Liliana Ibanescu, Melanie Munch (en postdoc) et **Cristina Manfredotti**

Rôle personnel : participant chercheuse, **co-encadrant du stage** de fin d'études (33%) de Alan Kabbouh.

Publications : [J3, J4, P16, W10, N1]

Avec cette collaboration nous avons étendue le travail de thèse de Melanie Munch afin de modéliser un processus pour la transformation de déchets urbains pour la production de matériel d'emballage. Dans ce contexte, les déchets urbains (feuilles sèches, petits morceaux de bois, ...) sont déchiqueté et, ensuite, mixé avec un polymère que rend le produit finale imperméable et résistante. Le défi est, alors, de trouver le juste compromis entre la qualité du pre-traitement des déchets et la quantité de polymère utilisé (qui est coûteux). Nous avons utilisé un des algorithmes présenté dans la thèse de Mélanie Munch couplé avec l'ontologie PO^2 pour apprendre un PRM qui a été, ensuite, utilisé avec des algorithmes d'inférence à l'état de l'art pour répondre à ce défi.

Projet AgroParisTech ToUHR-DRONAé (2020-2023). *Télédétection à Ultra-Haute-Résolution par Drone pour l'adaptation à des pratiques Agroécologiques*.

Rôle personnel : participant chercheuse.

Projet EXERSYS (2022-2025). *An EXplainable RecommandER SYStem for the Nutrition Domain, combining Knowledge Graphs and Machine Learning*, projet de recherche qui a financé un stage de fin d'études (financement DATAIA) et une thèse (financement DATAIA et école doctorale STIC).

Participants : Nicolas Darcel, Stephane Dervaux, Vincent Guigue, Fatiha Sais (LISN, Université Paris Saclay), Paolo Viappiani (CNRS, Université Paris Dauphine) et **Cristina Manfredotti**.

Rôle personnel : **Écriture, Soumission et Direction du projet**, **co-encadrement du stage** (25%) **et de la thèse** (50%). Co-encadrement de un stage de master M1 (50%) et suivi d'un étudiante en *Parcours Recherche* (33%).

Publications : [P17]

Projet GIFTED (2023-2026).*Prise de décision collective et recherche de consensus pour des choix alimentaires durables - GIFTED (Group Influences on Food Transition Eating Decisions)*, Projet de financement d'une thèse. École doctorale ABIES

Participants : Nicolas Darcel, Sabrina Teysser, Patrick Taillandier, Paolo Viappiani et **Cristina Manfredotti**.

Rôle personnel : participant chercheuse, **co-encadrant de thèse** (10%).

Le projet vise à poursuivre nos travaux sur des stratégies de recommandation pour groupes d' utilisateurs.

Projet AgroParisTech PrediMix (2023-2026). *méthodologie de prédiction et sélection pour des mélanges céréale-légumineuse permettant une gestion azotée agroécologique grâce à la modélisation et au phénotypage à haut débit par drone*.

Rôle personnel : participant chercheuse.

Projet ANR FRIEND (2023, non financé) *Food Recommendation Intelligent Engines for Nutritionally-improved Dietary Swaps*.

Rôle personnel : participant chercheuse.

Projet Européen SWAPS (soumis novembre 2023). *Swaps With Alternative Protein Sources*, MARIE SKŁODOWSKA-CURIE ACTIONS Doctoral Networks Call : HORIZON-MSCA-2023-DN-01-01.

Partners : **AgroParisTech**, Politecnico Di Torino, Universitair Medisch Centrum, Sorbonne Université, University of Birmingham, Université Grenoble Alpes, Universitaet Regensburg, Universitetet Tromsoe-Norges Arktiske Universitet.

Rôle personnel : participant chercheuse, **co-encadrant d'une thèse** (25%).

Encadrement de la recherche

Théses de doctorat Depuis septembre 2014, j'ai co-encadré **3** thèses de doctorat et j'en co-encadre **2** actuellement.

Étudiant	**Thomas DHEILLY**
Sujet	**Prise de décision collective et recherche de consensus pour des choix alimentaires durables - GIFTED (Group Influences on Food Transition Eating Decisions)**
Encadrants	Nicolas Darcel, Sabrina Teyssier, Cristina Manfredotti (10%), Paolo Viappiani, Patrick Taillandier
Status	en cours, date du debut 2 Novembre 2023
Etablissement	AgroParisTech, Ecole doctorale ABIES

Étudiant	**Alexandre COMBEAU**
Sujet	**Un système de recommandation explicable pour le domaine de la nutrition, combinant les graphes de connaissances et l'apprentissage automatique**
Encadrants	Fatiha Sais, Cristina Manfredotti (50%)
Status	en cours, date du debut 1 Septembre 2023
Etablissement	LISN (Laboratoire Interdisciplinaire des Siences du Numérique), Ecole doctorale STIC

Étudiant	**Melanie MUNCH**
Sujet	**Améliorer le raisonnement dans l'incertain en combinant les modèles relationnels probabilistes et la connaissance experte.**
Encadrants	Juliette Dibie, Pierre Henri Wuillemin, Cristina Manfredotti (33%)
Soutenue	le 17 novembre 2020
Etablissement	AgroParisTech, Ecole doctorale ABIES
Prod. scientifique	[P12, P13, P14, W8]
Position actuelle	Research Engineer, UMR STLO, PFS team, INRAE Rennes, France.

Étudiant	**Sema AKKOYUNLU**
Sujet	**Compréhension des dynamiques de consommation alimentaires et système de recommandation alimentaire.**
Encadrants	Antoine Cornuéjols, Nicolas Darcel, Cristina Manfredotti (33%)
Période	de février 2017 à janvier 2020
Etablissement	AgroParisTech, Ecole doctorale ABIES
Prod. scientifique	[W6, W7]
Position actuelle	Partie avant terminer la thèse.

Étudiante	**Mathieu BOUYRIE**
Sujet	**Restauration d'images de noyaux cellulaires en microscopie 3D par l'introduction de connaissance *a priori***
Encadrants	Antoine Cornuéjols, Nadine Peyriéras, et Cristina Manfredotti (33%)
Soutenue	le 29 novembre 2016
Etablissement	AgroParisTech, Ecole doctorale ABIES
Prod. scientifique	[P11]
Position actuelle	Chercheur chez Niji.

Stages de Master Recherche M2 Depuis septembre 2014 j'ai co-encadré **7** stages de Master Recherche M2.

2023 **Noémie JACQUES**, (25%), étudiante en double diplôme IODAA et Master M2 AMI2B (biologie computationnelle : analyse, modélisation et ingénierie de l'information biologique et médicale) à Paris-Saclay. *E*XERSYS: un système de recommandation pour le domaine de la nutrition, combinant graphes de connaissance, ontologies et apprentissage automatique.. Co-encadrament avec Vincent Guigue, Fatiha Sais et Paolo Viappiani. Production scientifique : [P17, W11]

2021 **Yuhan WANG**, (50%), Master M2 en Informatique: ANDROIDE Sorbonne Université. *B*ayesian Preference Elicitation for Group Decisions with the Plackett Luce Model. Co-encadrement avec Paolo Viappiani.

2020 **Maeva Caillat**, (33%), Master M2 en Informatique Ecole Centrale Nantes. *B*ayesian Elicitation for Group Decisions with Monte Carlo Filtering Methods. Co-encadrement avec Paolo Viappiani et Nicolas Darcel. Production scientifique : [W9]

Allan Kabbouh, (33%), Master M2 en science des données et systèmes complexes à l' Université de Strasbourg. *M*odélisation de processus de transformation combinant Ontologie et PRMs - applications à la production d'emballages alimentaires. Co-encadrement avec Mélanie Munch et Patrice Buche.

2019 **Serife AKKOYUNLU**, (33%), Master M2 en Informatique: Systemes Intelligents de l'Université Paris-Dauphine. *P*rofiling des data-sets dans un objectif d'exploitation des dans outils d'AD. Co-encadrement avec Juliette Dibie et Fathia Sais.

2018 **Jiang YOU**, (33%), Master M2 en Informatique: ANDROIDE, Sorbonne Université. *G*raph Matching and Transfer Learning : How to learn a new model from an existing one. Co-encadrement avec Juliette Dibie et Fathia Sais.

2017 **Melanie MUNCH**, (25%), Master M2 en Informatique: Systemes Intelligents de l'Université Paris-Dauphine. *M*odélisation des processus de Transformation Combinant Ontologie et Modèles Relationnels probabilistes Application à la stabilisation de micro-organismes. Co-encadrement avec Pierre-Henri Wuillemin, Juliette Dibie et Stephane Dervaux.

Stages–courts, élève ingénieur 2A Depuis septembre 2014, j'ai co–encadré **3** stages–courts, élève ingénieur 2A.

2023 **Ayoub HAMMAL**, (50%), étudiant de master M1 en Intelligence artificielle à l'université de Paris-Saclay. *E*nrichir les données alimentaires pour faire des recommandations informées

2018 **Camille BARDON**, (50%), élève ingénieur 2A AgroParisTech. *E*ntre web design et pédagogie : développement d'un cours en ligne. Co-encadrement avec Juliette Dibie.

2016 **May-Line GADONNA**, (50%), élève ingénieur 2A AgroParisTech. *E*tude et amélioration de l'algorithme de filtrage à particules. Co-encadrement avec Pierre-Henri Wuillemin.

Contribution au rayonnement international de la discipline

Participation aux jurys de thèse de Doctorat et de HDR J'ai participé comme **co-encadrant** aux jurys de thèse de Mélanie Munch et de Mathieu Bouyrie et j'ai participé en tant qu'**examinatrice** au jury de thèse de Tran Thi Nhu Hoa.

Organisation colloques, conférences, journées d'étude, programmes de coopération scientifique en réseau J'ai été **co-chair** du *Canadian Artificial Intelligence Graduate Students Symposium* en 2014 et en 2011. En 2014, j'ai obtenu un montant de 3000 euro par le comité de **sponsoring** du Journal de l'Intelligence artificielle (AIJ), pour financer le voyage des orateurs et des invités à la conférence.

Expertise J'ai été membre du comité de programme des conférences suivantes:

- FLAIRS de 2011 à 2016 (6 ans),
- Canadian AI en 2011 et 2013,
- IJCAI en 2011.

J'ai participé à la relecture de plusieurs articles:

- pour la conférence nationale EGC en 2016 et la conférence canadienne Canadian AI en 2009
- pour les conférences européennes ECSQARU en 2015 et ECAI en 2010
- pour trois revues internationales : Data in Brief Journal en 2021, Journal of Approximate Reasoning en 2011 et Journal of Knowledge and Information Systems en 2010,
- pour les conférences internationales AAAI en 2013, AAMAS en 2012, CAIP en 2011 et AKDM en 2010.

Dans les dernières années, en raison aussi la charge familiale et la crise sanitaire, j'ai préféré me focaliser sur l'encadrement des étudiants et ma recherche.

Publications scientifiques et valorisations

Depuis l'année 2014, à AgroParisTech, j'ai publié 4 articles scientifiques dans des revues d'audience internationale, 9 papiers dans des conférences avec comité de sélection et actes et 6 dans des conférences sans actes. La table suivante synthétise toutes mes publications groupées par audience et venue. Ensuite je rapporte la liste complète de mes publications.

	Audience		
	Internationale	**Nationale**	**Total**
Articles scientifiques dans des revues avec comité de lecture	5	0	5
Chapitre d'ouvrage	0	0	0
Communications avec comité de sélection et avec publication dans des actes	17	0	17
Communications avec comité de sélection	11	1	12

Articles dans des revues à comité de lecture

J5 Jules Vandeputte, Pierrick Herold, Mykyt Kuslii, Paolo Viappiani, Laurent Muller, Christine Martin, Olga Davidenko, Fabien Delaere, Cristina Manfredotti, Antoine Cornuéjols, Nicolas Darcel. Principles and Validations of an Artificial Intelligence-Based Recommender System Suggesting Acceptable Food Changes. *Journal of Nutrition*, 153, 2 (2023).

J4 Mélanie Münch, Patrice Buche, Cristina E. Manfredotti, Pierre-Henri Wuillemin, Hélène Angellier-Coussy. Formalizing Contextual Expert Knowledge for Causal Discovery in linked Knowledge Graphs about Transformation Processes: Application to processing of bio-composites for food packaging *International Journal of Metadata, Semantics and Ontologies* IJMSO-349696

J3 Mélanie Münch, Patrice Buche, Stéphane Dervaux, Juliette Dibie, Liliana Ibanescu, Cristina E. Manfredotti, Pierre-Henri Wuillemin, Hélène Angellier-Coussy. Combining ontology and probabilistic models for the design of bio-based product transformation processes. *Expert Systems for Applications* 203: 117406 (2022)

J2 Luca Cattelani, Cristina E. Manfredotti, Enza Messina. A Particle Filtering Approach for Tracking an Unknown Number of Objects with Dynamic Relations. *Journal of Mathematical Modeling and Algorithms in OR* 13(1): 3-21 (2014)

J1 Elisabetta Fersini, Enza Messina, Francesco Archetti, Cristina E. Manfredotti. Combining Gene Expression Profiles and Drug Activity Patterns Analysis: A Relational Clustering Approach. *Journal of Mathematical Modeling and Algorithms*, 9(3): 275-289 (2010)

Communications à des congrès avec actes et comité de lecture

P17 Noémie Jacquet, Vincent Guigue, Cristina Manfredotti, Fatiha Saïs, Stéphane Dervaux, Paolo Viappiani. Modélisation du caractère séquentiel des repas pour améliorer la performance d'un système de recommandation alimentaire. 24eme Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC 2024), Jan 2024, Dijon, France.

P16 Mélanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin, Helene Angellier-Coussy. A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models: application to processing of bio-composites for food packaging. 15th International Conference on Metadata and Semantics Research 2021.

P15 Cristina E. Manfredotti, Paolo Viappiani. A Bayesian Interpretation of the Monty Hall Problem with Epistemic Uncertainty. MDAI 2021: 93-105

P14 Melanie Munch, Juliette Dibie, Pierre-Henri Wuillemin, Cristina E. Manfredotti. Towards Interactive Causal Relation Discovery Driven by an Ontology. FLAIRS Conference 2019: 504-508

P13 Melanie Munch, Pierre-Henri Wuillemin, Juliette Dibie, Cristina E. Manfredotti, Thomas Allard, Solange Buchin, Elisabeth Guichard. Identifying Control Parameters in Cheese Fabrication Process Using Precedence Constraints. DS 2018: 421-434

P12 Melanie Munch, Pierre-Henri Wuillemin, Cristina E. Manfredotti, Juliette Dibie, Stéphane Dervaux. Learning Probabilistic Relational Models Using an Ontology of Transformation Processes. OTM Conferences (2) 2017: 198-215

P11 Mathieu Bouyrie, Cristina E. Manfredotti, Nadine Peyrièras, Antoine Cornuéjols. Denoising 3D Microscopy Images of Cell Nuclei using Shape Priors on an Anisotropic Grid. *International Conference on Pattern Recognition Applications and Methods (ICPRAM)* 2016: 291-298

P10 Christophe Gonzales, Sèverine Dubuisson, Cristina E. Manfredotti. A New Algorithm for Learning Non-Stationary Dynamic Bayesian Networks With Application to Event Detection. *Florida Artificial Intelligence Research Society Conference (FLAIRS)* Conference 2015: 564-569

P9 Cristina E. Manfredotti, Cèdric Baudrit, Juliette Dibie, Pierre-Henri Wuillemin. Mapping Ontology with Probabilistic Relational Models. *International Conference on Knowledge Engineering and Ontology Development (KEOD)* 2015, pp. 171-178.

P8 Cristina E. Manfredotti, Kim Steenstrup Pedersen, Howard J. Hamilton, Sandra Zilles. Learning Models of Activities Involving Interacting Objects. Proceedings of *Advances in Intelligent Data Analysis (IDA)* 2013, pp. 285-297.

P7 Luca Cattelani, Cristina E. Manfredotti, Enza Messina. Multiple Object Tracking with Relations. *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2012: 459-466.

P6 Cristina E. Manfredotti, David J. Fleet, Howard J. Hamilton, Sandra Zilles. Simultaneous Tracking and Activity Recognition. Proceedings of the *23rd IEEE International Conference on Tools with Artificial Intelligence ICTAI* 2011: 189-196.

P5 Cristina E. Manfredotti, David J. Fleet, Enza Messina. Relations to improve Multi-Target Tracking in an Activity Recognition System. *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP-09)*, December, 2009.

P4 Cristina E. Manfredotti, Enza Messina. Relational Dynamic Bayesian Networks to Improve Multi-target Tracking. *Advanced Concepts for Intelligent Vision Systems (ACIVS)* 2009: 528-539

P3 Cristina E. Manfredotti. Modeling and Inference with Relational Dynamic Bayesian Networks. *Canadian Conference on AI*, 2009, pp. 287-290.

P2 Francesco Archetti, Cristina E. Manfredotti, Vincenzina Messina, Domenico G. Sorrenti. Foreground-to-Ghost Discrimination in Single-Difference Pre-processing. *Advanced Concepts for Intelligent Vision Systems (ACIVS)* 2006: 263-274.

P1 Francesco Archetti, Cristina Manfredotti, Matteo Matteucci, Enza Messina, Domenico G. Sorrenti. Parallel first-order Markov Chain for on-line Anomaly Detection in traffic video surveillance. *IET Conference on Crime and Security: the Technical Fight*, June, 2006.

Communications à des congrès sans actes

W11 Noémie Jacquet, Cristina Manfredotti, Vincent Guigue, Fatiha Saïs, Paolo Viappiani An EXplainable RecommandER SYStem for the Nutrition Domain, combining Knowledge Graphs and Machine Learning CoCoA-BeANS workshop: Cognitive and COmputational Approaches of Behaviour and Nutrition Studies. Paris, May 11-12, 2023.

W10 Melanie Munch, Cristina Manfredotti, Liliana Ibanescu, Patrice Buche. An ontology-based pipeline to support the design of technical itineraries: application to composite food packaging. Integrated Food Ontology Workshop 2021, Sept 2021, Bolzano, Italy.

W9 Maeva Caillat, Nicolas Darcel, Cristina Manfredotti, Paolo Viappiani. Bayesian Vote Elicitation for Group Recommendations. DA2PL 2020, Nov 2020, Trento, Italy.

W8 Melanie Munch, Juliette Dibie-Barthélemy, Pierre-Henri Wuillemin, Cristina E. Manfredotti. Interactive Causal Discovery in Knowledge Graphs. PROFILES/SEMEX@ISWC 2019: 78-93

W7 Sema Akkoyunlu, Cristina E. Manfredotti, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere. Exploring Eating Behaviours Modelling for User Clustering. HealthRecSys@RecSys 2018: 46-51

W6 Sema Akkoyunlu, Cristina E. Manfredotti, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere. Investigating Substitutability of Food Items in Consumption Data. HealthRecSys@RecSys 2017: 27-31

W5 Luca Cattelani, Cristina Manfredotti, Enza Messina. Multiple objects tracking with probabilistic relationships. 1st interdisciplinary workshop on Mathematics of Filtering and its Applications (MFA2011), July, 2011.

W4 Cristina Manfredotti, David J. Fleet, Howard J. Hamilton, Sandra Zilles. Relational Particle Filtering. NIPS workshop on Monte Carlo Methods for Bayesian Inference in Modern Day Applications, December 2010.

W3 Cristina Manfredotti, Sandra Zilles, Howard J. Hamilton. Learning RDBNs for Activity Recognition. NIPS Workshop on Learning and Planning in Batch Time Series Data, December 2010.

W2 Cristina Manfredotti, Enza Messina, David Fleet. Relations as Context to improve Multi Target Tracking and Activity Recognition. First International Workshop on Logic-Based Interpretation of Context: Modeling and Applications, September, 2009.

W1 Elisabetta Fersini, Cristina Manfredotti, Enza Messina, Francesco Archetti. Relational Clustering for Gene Expression Profiles and Drug Activity Pattern Analysis. SysBioHealth Symposium, October, 2007.

Communications d'audience nationale avec comité de sélection

N1 M. Münch, P. Buche, C. Manfredotti, P-H. Wuillemin, H. Angellier-Coussy. Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes: application aux emballages bio-composites IC-Ingénierie des connaissance, affilié à PFIA 2022.

Publications dans des congrès internationals sans comité ni actes

C5 Cristina Manfredotti, Kim S. Pedersen, Howard J. Hamilton, Sandra Zilles. Learning Models of Activities Involving Interacting Objects. *European Conference on Operation Research*, July, 2013.

C4 Enza Messina, Giorgio Consigli, Cristina Manfredotti. A sequential learning method for tracking stochastic volatility. *European Conference on Operation Research*, July, 2010.

C3 Cristina Manfredotti, Enza Messina, Francesco Archetti. Improving Multiple Traget Tracking with RDBNs. *AIRO winter 2009, Conference of the Italian Operations Research Society*, January, 2009.

C2 Elisabetta Fersini, Cristina Manfredotti, Enza Messina. Relational K-means for gene expression profiles and drug activity patterns analysis. *AIRO 2007, XXXVIII Annual Conference of the Italian Operations Research Society*, September, 2007.

C1 Cristina Manfredotti, Francesco Archetti. Anomaly detection in traffic video surveillance: a multiple hypotesisMarkov chain approach. *AIRO 2006, XXXVII Annual Conference of the Italian Operations Research Society*, September, 2007.

Thèses

T3 Cristina Manfredotti. Modelling and Inference with Relational Dynamic Bayesian Networks. PhD Thesis. Università di Milano-Bicocca, 2009.

T2 Cristina Manfredotti. ImageJ plug-in for polyharmonic wavelet transform. Diplôme post-master Thesis. Università di Milano-Bicocca, 2004.

T1 Cristina Manfredotti. Problemi Matematici inerenti la Risonanza Magnetica. Master Thesis. Università di Milano-Bicocca, 2003.

Rapports techniques

TR2 Cristina E. Manfredotti, David J. Fleet, Howard J. Hamilton, Sandra Zilles. Simultaneous Tracking and Activity Recognition with Relational Dynamic Bayesian Networks. Technical Report TR 2011-01, Department of Computer Science, University of Regina, 2011. ISBN 978-0-7731-0694-9.

TR1 Francesco Archetti, Elisabetta Fersini, Ilaria Giordani, Cristina Manfredotti, Enza Messina, Daniele Toscani, A Unifying view of Probabilistic Relational Models and Applications. In Emerging Paradigms in Informatics, Systems and Communication, C. Batini, R. Schettini (eds.). DiSCo Quaderni di Dipartimento. Report n. 2009-01, June 2009