Twenty-Eighth International Conference on
Case-Based Reasoning
(ICCBR 2020)

Doctoral Consortium Proceedings

Michael W. Floyd and Stewart Massie (Editors)

Proceedings of the ICCBR 2020 Doctoral Consortium

Michael W. Floyd and Stewart Massie (Editors)

Students and *Mentors*

Christopher L. Bartlett	State University of New York at Oswego, USA
Kerstin Bach	*Norwegian University of Science and Technology, Norway*
Ciara Feely	University College Dublin, Ireland
Cindy Marling	*Ohio University, USA*
Nicolas Lasolle	Université de Lorraine, France
David B. Leake	*Indiana University, USA*
María Eugenia Pérez-Pons	University of Salamanca, Spain
Isabelle Bichindaritz	*State University of New York at Oswego, USA*
Marta Plaza-Hernández	University of Salamanca, Spain
Stelios Kapetanakis	*University of Brighton, UK*
Luis Raúl Rodríguez Oconitrillo	Universidad de Costa Rica, Costa Rica
David W. Aha	*Naval Research Laboratory, USA*
Niloufar Shoeibi	University of Salamanca, Spain
Barry Smyth	*University College Dublin, Ireland*
Ashish Upadhyay	Robert Gordon University, UK
Sarah Jane Delany	*Technological University Dublin, Ireland*
Xiaomeng Ye	Indiana University, USA
Mark Keane	*University College Dublin, Ireland*

Program Chairs

Stewart Massie	Robert Gordon University, UK
Michael W. Floyd	Knexus Research Corporation, USA

Program Committee

Agnar Aamodt	Norwegian University of Science and Technology, Norway
David Aha	Naval Research Laboratory, USA
Klaus-Dieter Althoff	DFKI / University of Hildesheim, Germany
Kerstin Bach	Norwegian University of Science and Technology, Norway
Ralph Bergmann	University of Trier, Germany
Isabelle Bichindaritz	State University of New York at Oswego, USA
Sutanu Chakraborti	Indian Institute of Technology Madras, India
Sarah Jane Delany	Technological University Dublin, Ireland
Stelios Kapetanakis	University of Brighton, UK
Mark Keane	University College Dublin, Ireland
David Leake	Indiana University, USA
Cindy Marling	Ohio University, USA
Mirjam Minor	Goethe University, Germany
Stefania Montani	University of Piemonte Orientale, Italy
Barry Smyth	University College Dublin, Ireland

Table of Contents

ICCBR 2020 Doctoral Consortium

Stewart Massie[1] and Michael W. Floyd[2]

[1] Robert Gordon University, Aberdeen, Scotland, UK
[2] Knexus Research Corporation, Springfield, USA

Preface

This year marks the twelfth anniversary of the ICCBR Doctoral Consortium (DC). The DC was designed to nurture PhD candidates by providing them with opportunities to obtain feedback on their research, future work plans, and career objectives from senior case-based reasoning (CBR) researchers and practitioners. We are proud to carry on the tradition with a cohort of nine doctoral students from six different countries.

PhD candidates who applied to the program submitted summaries of their doctoral research. In their research summaries, they detailed the problems they are addressing, outlined their proposed research plans, and described progress to date. Accepted applicants were paired with mentors who helped them to refine their research summaries in light of reviewer feedback. The updated research summaries, which appear in this volume, were then orally presented at the ICCBR DC on June 9, 2020 in our first virtual event conducted fully online.

This year's participants presented a broad array of ongoing CBR research. Christopher Bartlett presented his work on using a CBR approach that couples clinical covariate data with epigenetic data in an analysis of breast cancer. Ciara Feely studied the use of recommender systems to support marathon runners during training and in the race itself. Marta Plaza-Hernández discussed how CBR can be used in the management of submerged archaeological complexes by serving as a support tool for decision making. Eugenia Pérez-Pons explained her research on finding the most efficient capital investment for Spanish companies not operating in the stock market. Luis R. Rodríguez Oconitrillo discussed how a CBR system can be designed to capture and represent what a judge understands of the information contained in a legal case file. Xiaomeng Ye presented an approach to robust adaptation that employs a chain of adapted cases or rules to form an adaptation path. Nicolas Lasolle discussed the development of a dedicated tool that employs a range of methods to populate and search an RDF database representation of a digital Humanities corpus. Niloufar Shoeibi described an approach in which a recommender system uses information extracted from social media to inform fundamental stock market analysis. Finally, Ashish Upadhyay presented his work on employing hybrid textual CBR and deep learning approaches for natural language generation from structured data.

We particularly want to thank all of the students, mentors, and program committee members who worked so hard to make the DC a success.

Clinical Covariate Based Survival Prediction in Breast Cancer

Christopher L. Bartlett*

Intelligent Bio Systems Laboratory, Biomedical and Health Informatics
State University of New York at Oswego, 7060 NY-104, Oswego, NY 13126
cbartle3@oswego.edu

1 Abstract

In the new era of personalized medicine where clinicians and researchers alike are seeking to custom-tailor treatment plans to individuals, the integration of clinical data with DNA microarray data is surprisingly absent. While clinicians call upon clinical data to apply similar treatments to similar patients, it's usually not the case within epigenetic or genetic research where only biochemical or expression differences are used for differentiation. With this in mind, this project proposes using a case-based reasoning system that couples clinical covariate data with epigenetic data in a comprehensive analysis of breast cancer. Previous work utilized a novel confidence-based method to extract cases by age group, racial group and therapy method before further narrowing down the case base using similar epigenetic profiles. Results within this work showed performance measures that were similar or greater than traditional classification. Now we are seeking to extend this work into ventures of survival analysis.

1 Introduction

In this new era of personalized medicine, clinicians have sought after methods which specifically target the patient through carefully tailored treatment plans. Throughout this movement, clinical and molecular profiles are constructed and managed in unison for advanced treatment. While this is becoming more prevalent on the frontlines of healthcare, the integration is surprisingly absent in 'omics research. The term 'omics collectively refers to genomics, proteomics, epigenomics and similar fields. Here, analysts are typically focused on a specific subtype of 'omics data while paying little attention to the clinical information that define the research sample. Even in studies that span across 'omics, these primary variables are neglected. As these clinical variables are more descriptive, they increase focus and lend to a more explainable outcome. Therefore, it has been the intent of our research to couple a stable biomarker, DNA methylation, with clinical data through a case-based reasoning structure. After several studies

* The author acknowledges the mentoring of his advisor Dr. Isabelle Bichindaritz at the State University of New York at Oswego

found age [2] [4], therapy [7], and race [6] to have a significant effect on DNA methylation levels, we used these confounding variables to draw an initial set of cases from a case base. We then narrowed this case base further using distance measures derived from DNA methylation values. In a study of discerning breast cancer tissue from normal breast tissue, this methodology retrieved balanced accuracy results up to 96.79%. We then validated these results using a dataset of ER positive and triple negative tissue samples where we saw balanced accuracies up to 77.55%. Further details on this project are provided in Section 5. We are now seeking to extend this project to a study of survival analysis.

2 Methylation

Epigenetics is the study of external modifications that alter gene expression without changing the DNA sequence. One such epigenetic modification is methylation. Methylation is a covalent attachment of a methyl group to cytosine. Cytosine (C) is one of the four bases that construct DNA and one of only two bases that can be methylated. While adenine can be methylated as well, cytosine is typically the only base that's methylated in mammals. Once this methyl group is added, it forms 5-methylcytosine where the 5 references the position on the 6-atom ring where the methyl group is added. Under the majority of circumstances, a methyl group is added to a cytosine followed by a guanine (G) which is known as CpG. While the methyl group is added onto the DNA, it doesn't alter the underlying sequence but it still has profound effects on the expression of genes and the functionality of cellular and bodily functions. Methylation at these CpG sites has been known to be a fairly stable epigenetic biomarker that usually results in silencing the gene. Further, the amount of methylation can be increased (known as hypermethylation) or decreased (known as hypomethylation) and improper maintenance of epigenetic information can lead to a variety of human diseases.

The most widely studied aberrations of methylation occur within the domain of cancer. While there's different molecular mechanisms that can impact the progression of cancer such as the loss of genetic materials or gene mutations, disruption of the epigenome can alter the onset of cancer as well. These disruptions can occur at either a global level, or a precise locus. To study these disruptions, researchers use a microarray chip that detects the amount of the methyl chemical at specific locations (usually CpG sites) within the DNA. Through samples extracted from control samples, and samples extracted from a target group (such as cancer), locations in the DNA with differentially methylated signatures can potentially lead to determining impacted genes or functional pathways.

3 Research Plan

Li, Liu, Chen and Rudin [5] built a deep learning neural network for case-based reasoning that employs prototypes and explains its predictions. Built for classification, observations are classified by their proximity to one of the prototypes. An

autoencoder reduces dimensionality and learns the features that will be useful for prediction before using the encoded input to produce a probability distribution over the classes through the prototype network. An additional advantage of their constructed network is that the reconstructed input is passed to a separate layer which allows the user to view how the network is constructing its predictions and prototypes; effectively reducing the black-boxed nature of most machine learning algorithms.

While Li, Liu, Chen and Rudin's [5] architecture learns the features of inputted images, carrying out tasks such as classifying a numeral in one font based on its appearance to the same numeral in other fonts, a similar methodology is desired. The network learns the relevant features and becomes guided by knowledge as it moves towards its prediction. The desired pipeline is to incorporate knowledge that aides in the survival prediction. Ideally, the network will utilize Enhancer Linking by Methylation/Expression Relationships (ELMER), an R package that reconstructs gene regulatory networks by combining gene expression and DNA methylation data. Methylation changes at cis-regulatory modules are the central hub of these networks, and correlation analysis is used to associate them with both upstream master regulator transcription factors, and downstream target genes. The package locates probes with significantly different DNA methylation levels between two groups, and identifies the putative target genes for these differentially methylated probes. Then it seeks enriched motifs for the probes which are significantly differentially methylated and linked to the putative target genes. Last, it identifies the regulatory transcription factors whose expression associate with DNA methylation at enriched motifs.

At these aforementioned steps, the samples can be clustered by similar clinical covariates. A prototypical representation of each cluster can then be formed, through which novel cases can be compared to. Then, a prediction of survival can be made for these novel cases through their proximity to the prototype. Doing so for each of these steps will allow for comparative analyses and a clear idea of how the varying levels of knowledge assist in the survival prediction. It is hypothesized that transitioning from differentially methylated positions to associated genes and the master regulators will not only increase the explainability, but also increase the predictive power. Associating the genes to their functional pathways will also add an additional level.

4 Domain Problems

Case-based reasoning (CBR) within the domain of microarray analysis is mostly unexplored, especially for epigenetic data. The primary foundation for CBR is its ability to consistently update with new cases, and adapt prior solutions to fit a new problem. Within microarray analysis, however, problems exist that make updating and adaptation particularly difficult. The first problem is the high dimensionality with few samples. There are thousands of features for a small subset of samples (specifically 485,000 for the standard chipset used in DNA methylation), and these samples are often imbalanced between cases and

controls. In other domains, class balancing is possible through algorithms such as the *S*ynthetic *M*inority *O*versampling *TE*chnique (SMOTE), though artificially generating samples is ill-advised when dealing with diseased tissue. There are also additional problems when using clinical covariates, as this data cannot be artificially generated.

A second problem is that technical variations, called "batch effects", often exist. Batch effects are alterations of the data that occur when different laboratories, technicians, or different equipment collects the data. Even when the same technicians operate on the same equipment within the same laboratory, subtle varying factors such as the amount of humidity can alter expression levels. These effects can be controlled to some degree through pre-processing, but need to be performed again when new cases are introduced. This makes the creation of an ongoing system problematic, and requires the researcher to operate on one dataset at a time.

A third problem is centered on case adaptation. For some domains, the data can be modified to bring cases in the case base closer to the query case (or the query case closer to the cases), but modifications in the microarray domain mean introducing artificial values to biologically derived data. This could easily mean attributing the presence of a disease to an incorrect locus, or believing the disease is present when it is not and other such outcomes.

5 Research Progress

Table 1: Testing results for classifying cancer tissue versus normal tissue using either a two-stage process or microarray data alone. (8,722 features).

Data Used	*Feature Selection*	*Categorical distance measure*	*Balanced Accuracy*	*F Statistic*	*Kappa Statistic*
Microarray	No	Lin	97.45%	0.98	0.88
Microarray	No	Goodall	97.15%	0.98	0.89
Microarray	BIRF	Lin	96.94%	0.98	0.89
Microarray	BIRF	Goodall	96.94%	0.98	0.89
Microarray	rKNN	Lin	96.75 %	0.99	0.91
Microarray	rKNN	Goodall	96.75 %	0.99	0.91
Two-Stage	No	Lin	94.25%	0.98	0.84
Two-Stage	No	Goodall	95.04%	0.98	0.84
Two-Stage	BIRF	Lin	95.96%	0.98	0.88
Two-Stage	BIRF	Goodall	96.68%	0.98	0.88
Two-Stage	rKNN	Lin	95.05%	0.98	0.85
Two-Stage	rKNN	Goodall	96.79%	0.98	0.85

Previous work conducted used three clinical covariate variables: age group, therapy method and racial group, to differentiate cancerous breast tissue and normal breast tissue. As age group, therapy and racial group are categorical variables, we utilized two categorical distance measures, Goodall3 and Lin (discussed in [3]). Euclidean distance was used for finding the distance among samples based on their DNA methylation values. We tested using two feature selection algorithms, a balanced iterative random forest (BIRF) [1] and a random kNN (rKnn) algorithm.

During this project, we constructed a method that would compute a confidence metric composed of each case base sample's average distance among

members of a different class minus the average distance among members of the same class. In this manner, we could locate a single value that specified how influential each sample in the case base was. Then, when drawing samples from the case base, we could continue to draw the nearest samples until a threshold was met. The single values were normalized between 0 and 1, with the threshold typically being 1.0 to indicate 100% confidence. Further, we drew based on the similarity among the clinical covariates first before narrowing down our retrieved cases using the similarity among DNA methylation values. We compared this two-stage process of using clinical covariate date prior to microarray data with using microarray data alone. The results are shown in Table 1.

6 Conclusion

To summarize, previous work was performed integrating clinical covariate data with microarray data to classify breast cancer samples. The current project seeks to extend this work to predict survival outcomes. Future work will extend further into highly specific feature selection to locate impacted functional pathways.

References

1. Anaissi, A.: Case-Base Retrieval of Childhood Leukaemia Patients Using Gene Expression Data By (January) (2013)
2. Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., et al.: Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genetics **8**(4) (2012). https://doi.org/10.1371/journal.pgen.1002629
3. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. Proceedings of the 2008 SIAM International Conference on Data Mining (2008). https://doi.org/10.1137/1.9781611972788.22
4. Horvath, S., Zhang, Y., Langfelder, P., Kahn, R.S., Boks, M.P., Eijk, K.V., Berg, L.H.V.D., Ophoff, R.A.: Aging effects on dna methylation modules in human brain and blood tissue. Genome Biology **13**(10) (2012). https://doi.org/10.1186/gb-2012-13-10-r97
5. Li, O., Liu, H., Chen, C., Rudin, C.: Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions (2017), http://arxiv.org/abs/1710.04806
6. Song, M.A., Brasky, T.M., Marian, C., Weng, D.Y., Taslim, C., Dumitrescu, R.G., Llanos, A.A., Freudenheim, J.L., Shields, P.G.: Racial differences in genome-wide methylation profiling and gene expression in breast tissues from healthy women. Epigenetics **10**(12), 1177–1187 (Feb 2015). https://doi.org/10.1080/15592294.2015.1121362
7. Yang, G.S., Mi, X., Jackson-Cook, C.K., Starkweather, A.R., Kelly, D.L., Archer, K.J., Zou, F., Lyon, D.E.: Differential dna methylation following chemotherapy for breast cancer is associated with lack of memory improvement at one year. Epigenetics p. 1–12 (2019). https://doi.org/10.1080/15592294.2019.1699695

Using Case-Based Reasoning to Support Marathon Runners

Ciara Feely

ML Labs, University College Dublin, Dublin, Ireland
Ciara.Feely@ucdconnect.ie

Abstract. This document describes the proposed research plan for the project "Using Recommender Systems Techniques to Support Endurance Athletes, in particular Marathon Runners". This project is in its first year but builds upon work that has been previously presented in ICCBR. The research problems being addressed relate to useful supports that could be available to runners including training session classification, personalised training plan recommendation, injury prediction and prevention, and race-time prediction and pacing plan recommendation. The progress to date includes an initial publication involving race-time prediction and training plan recommendation, as well as the early-stage development of a Strava companion app that will be used to collect user-labelled training session data. Future work will involve using that labelled data alongside case-based reasoning to provide solutions for the research problems, and evaluating these solutions with a live user-study.

Keywords: CBR for health and exercise; marathon running

1 Research Problems

The aim of this research is to utilize personal activity data that is being collected by wearable sensors and fitness apps to provide beneficial supports to endurance athletes, in particular marathon runners. Since the first ever marathon in ancient Greece, millions of people have tackled the iconic race as a feat of human endurance. The number of recreational runners participating in endurance events is increasing every year. Training for a long-distance event is difficult; hence there is an increasing population of recreational endurance athletes seeking advice on how to train, recover, and plan for their race. Simultaneously, there has been an influx in the availability and adoption of wearable sensors and mobile fitness applications that can track every training session.

These developments provide great opportunity to utilise techniques such as case-based reasoning to solve a variety of problems including:

1. Training session classification
2. Personalised training plan recommendation
3. Injury prediction and prevention
4. Race-time prediction and pacing plan recommendation

1.1 Training Session Classification

Marathon training plans are comprised of a variety of session types including long runs, tempo runs, interval runs, recovery runs. The ability to classify these sessions automatically as a given run-type would allow for improved training plan recommendation since then a recommendation could be a "tempo-run with average pace 4.24 minutes per kilometre". A case-based classification approach could be utilised since the different run-types might manifest differently in different runners. Thus being able to find similar runners and then classifying the run-types based on the similar runners will lead to a more accurate classifier.

1.2 Personalised Training Plan Recommendation

Training plans for marathoners are usually 12-16 weeks in duration and are broken up into 3-4 week training blocks. While professional runners may have a team of personal trainers assisting them in their training, recreational runners usually rely on online sites and one-size-fits-all training plans. Runners would benefit from having a tailored plan that takes into consideration their fitness, goals, and other personal characteristics.

1.3 Injury Prediction and Prevention

The incidence of running related injuries (RRIs) is high for all runners, but the proportion of injuries is especially high for novice runners and those at the lower and higher end of the distance curve [1]. The only risk factor that is consistently linked to the development of an RRI is having a history of RRI [2]. Yet, no definitive injury prevention procedures exist. Employing CBR to predict whether a runner is likely to become injured based on the previous injuries that developed in similar runners with a similar training load would assist runners in preventing injuries. For this, injury-labelled data is required – training data for which runners identify training breaks as being caused by injuries.

1.4 Race-time prediction and pacing-plan recommendation

In the lead up to a race, a runner will begin to plan for their race which may involve a nutrition plan, pacing plan, and any strategies they will use to stay motivated. In order to do this, they must have an estimate for their race duration. Many race-time prediction equations exist, however many require laboratory equipment or do not work equally for professional and recreational runners [3]. Utilising case-based reasoning, the authors of [4, 5] used previous marathon finish-times to predict future personal best finish times and provide an accompanying pacing plan. While they found that using multiple races improved prediction accuracy [6, 7], it naturally excluded more novice runners who would benefit the most from these predictions.

2 Research Plan

2.1 Progress To Date

This research project commenced in October 2019, and the intended finish date is December 2022. A first publication [8] is being presented in this year's ICCBR conference precedings. The work aimed to address research problem 4, by using CBR to build a prediction model that provided runners with estimated marathon finish-time, based on their training to date, at different points in the lead up to the marathon. Additionally, a proof-of-concept was demonstrated for research problem 2, indicating that when runners adjust their goal-time, faster times lead to training plans that have a faster average pace and greater total distance. Since then, different training representations, prediction models and training plan recommendation procedures have been explored. Other progress includes the development of a companion app for the popular fitness app Strava that will allow for labelled data collection. The app is still under construction but it is aimed to be completed in the summer of 2020. The current dissertation status is that the literature review is being written up, and subsequently, the first main chapter on the aforementioned publication will be written.

2.2 Labelled data collection

For some of the tasks mentioned (research problems 1 and 3), acquiring labelled data is key to finding a solution. The fitness application *Strava* allows developers to build companion apps that can be linked to Strava. A companion app allowing for training session data collection while asking the users to label their sessions as a given run-type, to give a history of previous injuries and to track if they subsequently become injured, is currently being developed.

2.3 Training Session Classification

Once labelled training-session data is collected, a run-type classification model can be built such that future sessions can be automatically labelled. The proposed method is to employ motif detection to find recurring motifs that distinguish the sessions as a specific run-type. This has been done before for classifying physical activities [9].

2.4 Personalised Training Recommendations

Thus far, the concept of using CBR to recommend training by selecting similar runners who have achieved that time, and recommending their training plan has been demonstrated [8]. To extend this, training a system to provide a set of training plans that are diverse across different features such as average pace, weekly distance, longest run — would allow users to have more choice and flexibility in their training. Achieving training session classification would facilitate more detailed recommendations. To evaluate the training session recommendation, a live user-study in which the finish-times of runners who adopted the recommended training interventions were compared to those who did not is necessary.

2.5 Injury Prediction and Prevention

Obtaining injury labelled data – a runner's injury history and whether training breaks were due to injury – would allow for the development of an injury prediction model. Additionally, what type of training leads to secondary injuries could be determined for runners with a history of injury. Then, training that does not lead to injury could be recommended when it seems that a runner is behaving in a way that could lead to injury. Again, to evaluate whether the provision of injury risk information alters a runner's training and whether this leads to injury prevention, a live user-study is required.

2.6 Race-Time Prediction and Pacing Plan Recommendation

The work achieved so far [8] has allowed for predictions to be made earlier in the training than in [4–7]. However, the race-time predictions were less accurate. Future work will involve incorporating some of the features used in [4–7] alongside the method in [8] in an attempt to improve the accuracy, while allowing for predictions to be made earlier in training. Further exploration of different features that may act as predictors for marathon-time is also required.

2.7 Timeline

The aim is to release the companion application to collect data for the 16 weeks leading up to the Dublin marathon which takes place at the end of October 2020. Following that, this data would be utilised in the previously described proposed solutions. The subsequent year would involve testing out the personalised training plan recommendation and whether it led to the achievement of a goal-time in a live user-study. After that, the models could be adjusted and another live-user study could be carried out in preparation for the 2022 Dublin marathon. At this point the final chapters of the thesis could be written up with goal of finishing in December 2022.

2.8 Expected Contributions

Since the number of people participating in marathons is ever-increasing, fulfilling the research goals identified in this document would be of enormous benefit to runners in improving how they train, recover, and plan for races. Additionally, this work could be extended to provide similar perks to recreational athletes in other endurance sports including cycling, swimming, and triathlons.

2.9 Conclusions

There are many open questions remaining in this work. Some relate to the target domain of marathon running, others relate to the suitability of the approach to be taken and the use of case-based reasoning techniques with time-series data. My hope is that participating at the ICCBR DC will provide me with a unique

opportunity to present my recent work and current plans in a forum that is designed specifically to support PhD students. I expect that case-based reasoning will remain an important feature of my research and as such, the opportunity to engage with senior members of the community and fellow CBR students is most welcome and likely to be very useful at this point in my PhD programme.

2.10 Acknowledgements

This work is supported by Science Foundation Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

References

1. B. Kluitenberg, M. van Middelkoop, R. Diercks, and H. van der Worp, "What are the Differences in Injury Proportions Between Different Populations of Runners? A Systematic Review and Meta-Analysis," *Sports Medicine (Auckland, N.Z.)*, vol. 45, pp. 1143–1161, Aug. 2015.
2. M. P. van der Worp, D. S. M. ten Haaf, R. van Cingel, A. de Wijer, M. W. G. Nijhuis-van der Sanden, and J. B. Staal, "Injuries in runners; a systematic review on risk factors and sex differences," *PloS One*, vol. 10, no. 2, p. e0114937, 2015.
3. A. Keogh, B. Smyth, B. Caulfield, A. Lawlor, J. Berndsen, and C. Doherty, "Prediction equations for marathon performance: A systematic review," *International Journal of Sports Physiology and Performance*, vol. 14, no. 9, pp. 1159–1169, 2019.
4. B. Smyth and P. Cunningham, "Running with cases: A CBR approach to running your best marathon," in *Case-Based Reasoning Research and Development - 25th International Conference, ICCBR 2017, Trondheim, Norway, June 26-28, 2017, Proceedings*, pp. 360–374, 2017.
5. B. Smyth and P. Cunningham, "A novel recommender system for helping marathoners to achieve a new personal-best," in *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*, pp. 116–120, 2017.
6. B. Smyth and P. Cunningham, "An analysis of case representations for marathon race prediction and planning," in *Case-Based Reasoning Research and Development - 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings*, pp. 369–384, 2018.
7. B. Smyth and P. Cunningham, "Marathon race planning: A case-based reasoning approach," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pp. 5364–5368, 2018.
8. C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, "Using case-based reasoning to predict marathon performance and recommend tailored training plans," in *Case-Based Reasoning Research and Development - 28th International Conference, ICCBR 2020, Online June 8-12, 2020, Proceedings*, 2020.
9. E. Berlin and K. V. Laerhoven, "Detecting leisure activities with dense motif discovery," in *The 2012 ACM Conference on Ubiquitous Computing, Ubicomp '12, Pittsburgh, PA, USA, September 5-8, 2012*, pp. 250–259, 2012.

Indexing and Exploring a Digital Humanities Corpus

Nicolas Lasolle

Université de Lorraine, CNRS, Université de Strasbourg, AHP-PReST, F-54000 Nancy, France
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
nicolas.lasolle@univ-lorraine.fr

Abstract. Semantic Web technologies have been chosen to structure and publish data of the Henri Poincaré correspondence corpus and to provide a set of tools for exploiting it. Two major issues have arisen during this work. The first one is related to the manual editing of triples in order to populate an RDF database. It is a tedious task for users with an important risk of error. That justifies the development of a dedicated tool to assist them during this process. It uses and combines different methods which are presented in this article. The second issue is related to the exploration of the corpus database. The need of an approximate and explained search has emerged. An engine based on the application of SPARQL query transformation rules has been designed. Different research tracks are considered to pursue this work which intends to apply in other contexts than the Henri Poincaré correspondence corpus.

Keywords: case-based reasoning, Semantic Web, content annotation, explained and approximate search, RDF(S), digital humanities

1 Introduction

The Henri Poincaré correspondence corpus is composed of around 2100 letters and gathers scientific, administrative and private exchanges. These letters are available on the website `http://henripoincare.fr` which was created and is managed with the CMS[1] Omeka S [1]. In addition to the Omeka S environment, Semantic Web technologies have been applied to this corpus.

Resource Description Framework (RDF [4]) allows the representation of data by using a labeled directed graph. It is based on the use of triples of the form $\langle subject\ predicate\ object \rangle$. For instance, $\langle$`letter11 sentBy henriPoincaré`$\rangle$ states that letter 11 has been sent by Henri Poincaré. RDFS Schema is extending the RDF model by introducing a new set of classes and properties. It allows to create a hierarchy between classes (using `rdfs:subclassof`) and properties (using `rdfs:subpropertyof`). `rdfs:domain` (resp. `rdfs:range`) applies for a property and adds a constraint about the type of the resource which is in

[1] Content Management System

subject (resp. *object*) position for a triple. Different inference rules which use these properties are considered [2]. SPARQL is a query language for RDF data which is a W3C[2] recommendation [5]. For the sake of readability, its specific syntax is not presented here. In the remainder of this article, queries are presented in an informal way. Manual Semantic Web data editing (i.e. creating the triples) is often a tedious task which may cause errors of different kinds. A suggestion tool combining the use of Semantic Web technologies with a case-based reasoning (CBR [6]) methodology is explained in Section 2. The use of SPARQL querying is sometimes not satisfactory to exploit this corpus. A tool associated with a language, named SQTRL (SPARQL Query Transformation Rule Language) have been designed to enable flexible querying [3]. This mechanism as well as different future works are presented in Section 3. Section 4 concludes.

2 Assisting human annotation for Semantic Web data editing thanks to a CBR methodology

2.1 Proposal of an annotation tool

A tool including an autocomplete mechanism has been created to assist users during the annotation process. It enables the visualization and the update of RDF databases. An *annotation question* corresponds to a triple for which 1,2 or the 3 fields are unknown, and for which a field is currently being edited. For example, consider the annotation question ⟨`letter11 sentTo` `?o`⟩. The objective here is to provide appropriate suggestions by listing and ranking the potential values for the object field. 4 versions of the editor have been implemented and use different approaches to order the suggestion list.

The *basic editor* uses the alphabetical order to rank the potential values.

The *deductive editor* benefits from the knowledge about the `rdfs:domain` and `rdfs:range` of the properties defined within the ontology. For the running example, as `Person` is `range` of the property `sentTo`,[3] this knowledge is used to favor the instances of this class.

The *case-based editor* follows the case-based reasoning methodology(CBR [6]). Information from situations similar to the current annotation question is reused. An annotation problem $\mathtt{x}^{\mathtt{tgt}}$ is composed of an annotation question and a *context* which corresponds to the set of edited triples related to the resource currently being edited. The annotation problem is defined as follows:

$$\mathtt{x}^{\mathtt{tgt}} = \begin{array}{|ll|}\hline \textbf{question:} & \langle \mathtt{letter11\ sentTo\ ?o} \rangle \\ \textbf{context:} & \langle \mathtt{letter11\ sentBy\ henriPoincaré} \rangle \\ & \langle \mathtt{letter11\ hasTopic\ écolePolytechnique} \rangle \\ & \langle \mathtt{letter11\ quotes\ paulAppell} \rangle \\ \hline \end{array}$$

A solution $\mathtt{y}^{\mathtt{tgt}}$ corresponds to a value for `?o`. The RDF database $\mathcal{D}$ is defined as the case base. Each resource of $\mathcal{D}$ is defined as a source case $\mathtt{x}^{s}$ and is used

[2] World Wide Web Consortium

[3] According to the Henri Poincaré corpus ontology.

to propose a candidate solution $\mathtt{y}^s$. The method consists in retrieving the most similar resources in the case base. SPARQL querying is used in that context. An initial query `Q` is created to retrieve the resources whose context is the same as the one of `letter11`. However, it is uncommon to find two resources with the exact same context. The SQTRL tool [3], whose functioning is detailed in the following section, is reused in this application framework. Using this tool allows the generation of a new set of queries which can be executed to find resources that partially match the context of `letter11`.

The *combination editor* combines the use of RDFS deduction with CBR.

2.2 Evaluation and results

Two different evaluations, which include a comparison of the different versions of the suggestion system, have been carried out. The RDF graph of the Henri Poincaré correspondence corpus $\mathcal{G}_{\mathrm{HP}}$ has been used as a test set.

The first evaluation is human-based: a user works with the dedicated tool to edit a set of unpublished letters. The user feedback has been collected through a survey in which he has been asked to attribute a score to measure the efficiency of the different versions. The second evaluation is automatic through a program which computes measures to compare the different versions for similar annotation questions. Several edited triples have been randomly extracted from this database graph and have been used to simulate annotation questions for which the answers are already known. The results show that the combination editor is the most efficient, and this for all the properties of the evaluation. Combining the use of RDFS knowledge with a CBR methodology is an appropriate solution to rank the potential values in a efficient way.

3 Approximate and explained search

3.1 SQTRL

SQTRL is a tool which has been designed to manage approximate search and has proven useful in different contexts including the search in the Henri Poincaré correspondence corpus and the case-based cooking system Taaable [3]. Users can configure SPARQL query transformation rules which can be general (e.g. property and class generalization, triple removal, etc.) or context-dependent (e.g. exchange of the sender and the recipient of a letter, substitution of a person by one of his/her colleagues, etc.). An application of a rule r to a SPARQL query $\mathtt{Q}$ for a given RDF database $\mathcal{D}$ could generate a new query $\mathtt{Q}_N$. The execution of $\mathtt{Q}_N$ on $\mathcal{D}$ may return a different set of results than the execution of $\mathtt{Q}$ on that same database. A search tree can be explored, starting from the initial query $\mathtt{Q}$, by applying one or several successive transformation rules. To each rule is associated a cost (defined as a constant integer) which corresponds to a query transformation cost. By defining a maximum cost, it is possible to limit the

depth of the search tree exploration. Let $\mathcal{Q}$ be an example of a query formulated by a historian to query the database of the correspondence corpus $\mathcal{D}_{\text{HP}}$:

$$\mathcal{Q} = \left| \begin{array}{l} \text{“Give me the letters sent by } \texttt{henriPoincaré} \\ \text{ to a physicist with } \texttt{optics} \\ \text{ as a topic between 1880 and 1890.”} \end{array} \right.$$

The queries $\{\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3\}$ correspond to queries generated at depth 1:

$$\mathcal{Q}_1 = \left| \begin{array}{l} \text{“Give me the letters sent by} \\ \underline{\text{ a physicist to } \texttt{henriPoincaré}} \\ \text{with } \texttt{optics} \text{ as a topic} \\ \text{between 1880 and 1890.”} \end{array} \right. \quad \mathcal{Q}_2 = \left| \begin{array}{l} \text{“Give me the letters sent} \\ \text{ by } \texttt{henriPoincaré} \text{ to a} \\ \underline{\text{scientist}} \text{ with } \texttt{optics} \text{ as a} \\ \text{topic between 1880 and 1890.”} \end{array} \right.$$

$$\mathcal{Q}_3 = \left| \begin{array}{l} \text{“Give me the letters sent by } \texttt{henriPoincaré} \\ \text{ to a physicist with } \texttt{optics} \\ \text{ as a topic } \underline{\text{between 1875 and 1895}}\text{.”} \end{array} \right.$$

$\mathcal{Q}_1$ is generated by applying a rule which exchanges the sender and the recipient of the letter. An object class generalization rule replaces `Physicist` in `Scientist` and generates the query $\mathcal{Q}_2$. $\mathcal{Q}_3$ is generated by applying a rule which extends the temporal bounds of the query. Other rules could be applied to Q. Moreover, depending on the maximum cost which has been set, the tree exploration can be continued to generate other queries.

3.2 Future works and research plan

Works related to this approximate and explained search mechanism are frequently being discussed with historians of science. The objective is to provide tools which would assist them efficiently for the study of the Henri Poincaré corpus and which may be applied in other contexts. At this stage, different research tracks are considered and these discussions should be pursued in order to identify the main concerns and specify the research plan.

A first issue is related to the representation of transformation rules which does not insist on the explainability of this mechanism. Moreover, it could be relevant to include user preferences to favor the application of some rules in given situations. The application of a CBR methodology could be useful to store and reuse successful transformation paths. Another issue is related to the management of the costs associated to transformation rules. These have been set subjectively (defined as constant integers) for a first set of rules. It could be relevant to consider it as dependent on the resources occurring in the SPARQL query on which the rule is applied. The CBR community could be helpful here: the issue is to find how to assess similarity between resources to propose a method adapted for the context of the Henri Poincaré correspondence corpus graph. Other issues are related to the application of these transformation rules. As an example, a limitation is related to the occurrence of some unnecessary compositions of rules.

Consider the example presented in Section 3 in which the query $\mathcal{Q}$ is generated into a query $\mathcal{Q}_1$ by applying a rule which exchanges the sender and recipient of the letter. Currently, nothing is done to prevent the application of the same rule on $\mathcal{Q}_1$. It is thus possible to generate, at depth 2, a query $\mathcal{Q}_{11}$ such as $\mathcal{Q}_{11} = \mathcal{Q}$. A way to prevent such a situation should be explored.

4 Conclusion

Semantic Web technologies have been used to exploit the Henri Poincaré correspondence corpus. In that context, several issues have arisen. A tool has been proposed to assist users during the human annotation process. It benefits from the use of RDFS deduction and from the application of a CBR methodology. A human and an automatic evaluation have been carried out for this system and have shown positive results. Another issue is related to the need of an approximate search mechanism. The SQTRL tool proposes a method based on the use of SPARQL query transformation rules. However, the work related to this mechanism should be pursued in order to propose tools and methods suitable for the Henri Poincaré correspondence corpus and which can be reused in other contexts. The use of a CBR methodology is considered to achieve this goal and respond to some issues which have arisen.

Acknowledgement. This work is supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

References

1. Boulaire, C., Carabelli, R.: Du digital naive au bricoleur numérique les images et le logiciel Omeka. In: Expérimenter les humanités numériques. Des outils individuels aux projets collectifs. Les Presses de l'Université de Montréal (2017)
2. Brickley, D., Guha, R.V.: RDF Schema 1.1, `https://www.w3.org/TR/rdf-schema/`, W3C recommendation, last consultation: February 2020 (2014)
3. Bruneau, O., Gaillard, E., Lasolle, N., Lieber, J., Nauer, E., Reynaud, J.: A SPARQL Query Transformation Rule Language — Application to Retrieval and Adaptation in Case-Based Reasoning. In Aha, D., Lieber, J., eds.: Case-Based Reasoning Research and Development. ICCBR 2017. Lecture Notes in Computer Science, Springer (2017) 76–91
4. Lassila, O., Swick, R.R., et al.: Resource description framework (RDF) model and syntax specification. (1998)
5. Prud'hommeaux, E.: SPARQL query language for RDF, W3C recommendation (2008)
6. Riesbeck, C.K., Schank, R.C.: Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey (1989)

Methodology for efficient capital investments in private sector companies

María Eugenia Pérez-Pons[1][0000−0002−2194−572X]

[1] BISITE Research Group, Salamanca University
Edificio Multiusos I+D+I, Calle Espejo 2, 37007 Salamanca (Spain)
eugenia.perez@usal.es

Abstract. The research aims to create a platform that will enable finding the most efficient capital investment in Spanish companies that do not operate in the stock market. The objective is to maximize the possibilities of good investments through different Artificial Intelligence (AI) techniques as well as to allow diversification according to investments made by similar investors. Currently, a wide variety of methodologies are used for company valuation at investment level, especially those that take into account financial statements oriented to fulfill the investor's preferences. However, there is no method that would be capable of predicting, with full certainty, the future success of an investment.

Keywords: Company valuation · Capital Investment · Machine Learning.

1 Research Summary

The project aims to create a system for the identification and recommendation of efficient capital investments. The platform will be oriented either to potential investors and companies; the platform architecture is described in figure 1. In the defined architecture, recommendations are made according to an investor's profile and the companies interested in receiving investment capital. To this end, it is important to identify the preferences of an investor taking into account that they differ greatly from one investor to another; and also depending on the stage of the company in which they usually invest [6]. Currently, a wide variety of methodologies are used for company valuation [2], especially those that take into account financial statements. However, there is no method that would be capable of predicting with full certainty, the future success of an investment or the most appropriate investment for a given investor. The purpose is also to recommend investments based on previous investments made by investors with a similar profile and economic interest.

A problem that tries to overcome that research is to propose a platform with proven information to avoid the lack of transparency [4] and to uniform information of private companies in the Spanish enterprise investment environment. Companies that do not operate in the stock market are not exposed to a public audit and valuation of their financial statements, which makes it more

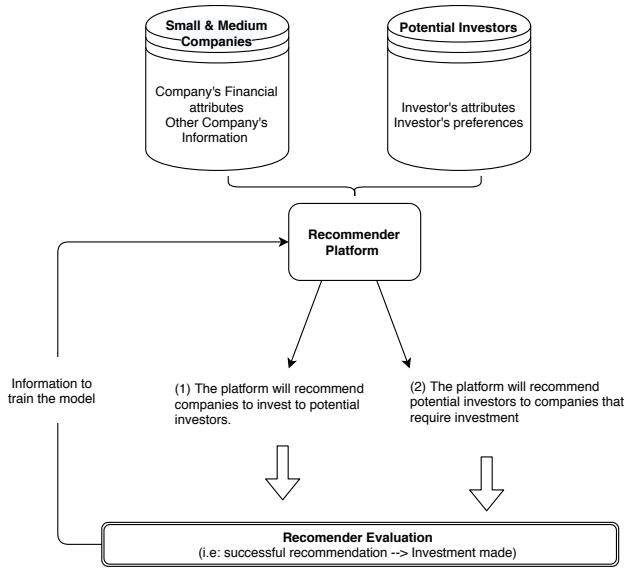

Fig. 1. Platform phases

complex to be able to analyze them in a transparent way. Therefore the platform will allow potential investors and companies that require investment to obtain information from many sources which will permit to contrast and compare information in order to find the most verified information. Today's investment environment is very dynamic, and different types of investors have appeared on the market. Business angels (BA), for example, are a type of investors that view investment in companies as the alternative to investment in other fields [3]. BA can be taken as a representative sample for different industries because they are a rising type of investor in Spain and also the one that can have a lack of understanding of investment processes [9] and also due to the variety of ways they consider the investment process. Many variables influence whether a company is valued for investment or not; those financial methods can be classified in six groups as proposed by Fernández [1] such as: (1) Balance Sheet, (2) Income statement, (3) Goodwill, (4) Cash flow discounting, (5) Value creating or (6) Options. For company valuation, some studies focus on the company's product or on its net value derived from cash flows [8], which also could be considered as an alternative methods. At the product target level, there are currently two product criteria under which the profitability of an acquisition or holding can be analyzed [7], namely, the uniqueness of the product on the market and product-market fit, that is, the degree to which a product satisfies demand. However, in a business acquisition, merely calculating the viability of the product would limit the number of potential investments for investors.

2 Research Plan

The initial defined stages to fulfill the project in terms of data are displayed in the diagram 2. The first step is the the identification of different sources and the data storage, the next steps are to pre-process the data, the implementation of machine learning algorithms and finally to create the visualization.

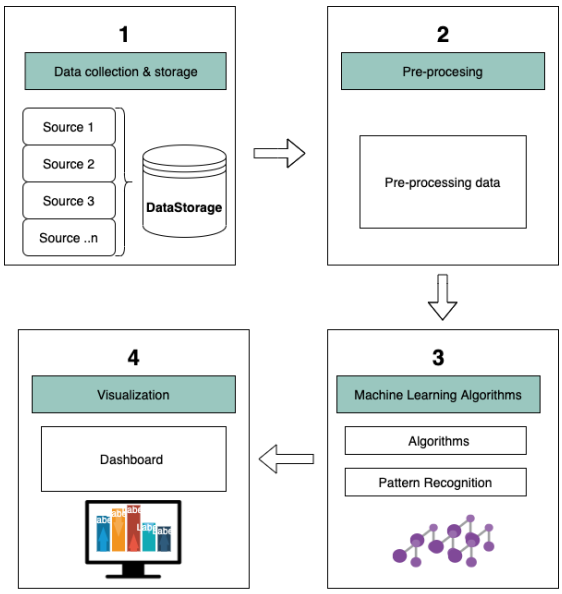

Fig. 2. Data Research Plan

2.1 Description of progress and challenges

To this point, regarding the first step of the Research Stages, the main sources that will be used have been identified. Those sources contain three types of information:

1. General information about the company
2. Financial information
   (according to the classification proposed by Fernández [1])
3. List of investors and their attributes

One of the main problems is that the data has been captured from different sources, and information differs from one to another. For example; a given company number of employees or description is described in one source with some numbers, and then in another site, it has another information. This dilemma

can lead to confusion in the application of algorithms as it would be incorrect to average the data from different sources. To this end, the use of Natural Language Processing (NLP) methodologies to merge fields as a description of companies is proposed for the text fields. Nevertheless, when considering financial companies' values, it is not clear which would be the most accurate way to proceed, and that is where the possibility of incorporating a Case-Base Reasoning (CBR) methodology for the union of the different fields comes in. This could be treated as multi-case base reasoning in a similar method as the one proposed by [5] or to incorporate a hybrid approach that includes Rule-Case Reasoning (RCR) with CBR as [11] did for diagnosis with expert knowledge. Another interesting approach which has had very interesting results is to incorporate fuzzy neural networks into the CBR methodology [10].

3 Acknowledgements

The directors of the PhD program are: Juan Manuel Corchado[0000−0002−2829−1829] and Javier Parra[0000−0002−1088−9152].

This research has been supported by the project "INTELFIN: Artificial Intelligence for investment and value creation in SMEs through competitive analysis and business environment", Reference: RTC-2017-6536-7, funded by the Ministry of Science, Innovation and Universities (Challenges-Collaboration 2017), the State Agency for Research (AEI) and the European Regional Development Fund (ERDF).

References

[1] Fernández López Fernández et al. *Valuation methods and shareholder value creation*. Academic Press, 2002.

[2] Pablo Fernández et al. "Company valuation methods. The most common errors in valuations". In: *IESE Business School* 449 (2007).

[3] Jon Hoyos Iruarrizaga, Marıa Saiz Santos, et al. "The informal investment context: specific issues concerned with business angels". In: (2013).

[4] Gerald H Lander and Kathleen A Auger. "The economic impact of the lack of transparency in financial reporting". In: *Atlantic Economic Journal* 36.1 (2008), pp. 105–116.

[5] David B Leake and Raja Sooriamurthi. "When two case bases are better than one: Exploiting multiple case bases". In: *International Conference on Case-Based Reasoning*. Springer. 2001, pp. 321–335.

[6] Alison Mackey, Tyson B Mackey, and Jay B Barney. "Corporate social responsibility and firm performance: Investor preferences and corporate strategies". In: *Academy of management review* 32.3 (2007), pp. 817–835.

[7] Ian C MacMillan, Lauriann Zemann, and PN Subbanarasimha. "Criteria distinguishing successful from unsuccessful ventures in the venture screening process". In: *Journal of business venturing* 2.2 (1987), pp. 123–137.

[8] Matthias Meitner. *The market approach to comparable company valuation.* Vol. 35. Springer Science & Business Media, 2006.

[9] Amparo San José, Juan Roure, and Rudy Aernoudt. "Business angel academies: unleashing the potential for business angel investment". In: *Venture capital* 7.2 (2005), pp. 149–165.

[10] Kim Hua Tan et al. "An intelligent decision support system for manufacturing technology investments". In: *International journal of production economics* 104.1 (2006), pp. 179–190.

[11] Yuan-Hsin Tung et al. "A rule-based CBR approach for expert finding and problem diagnosis". In: *Expert Systems with Applications* 37.3 (2010), pp. 2427–2438.

An intelligent platform for the management of underwater cultural heritage

Marta Plaza-Hernández[1]

[1] BISITE Research Group, University of Salamanca,
Edificio Multiusos I+D+i, Calle Espejo 2, 37007 Salamanca (Spain).
martaplaza@usal.es

Abstract. Conservation of Underwater Cultural Heritage is crucial to preserving society's history. This research proposal aims to study and develop Artificial Intelligence techniques for the management of submerged archaeological complexes, serving as a supporting tool for decision making.

Keywords: Underwater Cultural Heritage, Artificial Intelligence, Underwater Internet of Things, Edge Computing.

1 Introduction

The documentation and conservation of Underwater Cultural Heritage (UCH) are crucial to preserving society's identity and memory, ensuring its accessibility to present and future generations. Conservation methods and processes, from the evaluation and analysis of the state of the heritage to restoration activities, still present multiple challenges, including the complexity of operating underwater, the lack of regulation, policies and resources to cope with the effects of climate change, ineffective protection of cultural heritage (unsustainable tourism) or the elevated costs. To overcome these obstacles, the UNESCO created a treaty, *The Convention on the Protection of Cultural Heritage Underwater 2001*, which establishes basic principles for protection, rules for heritage treatment and a system of international cooperation. So far, only 63 countries have ratified or accepted this document [1].

The conservation of submerged archaeological complexes requires the adoption of innovative and sustainable solutions that aim not only at preserving them in-situ but also at using the available information for decision-making. The use of sensors could be one of the most cost-effective practices for assessing the state of tangible heritage, facilitating the monitoring of environmental changes. The Internet of Things (IoT) refers to the connection of multiple and heterogeneous objects (buildings, machinery, vehicles, etc.) with electronic devices (sensors and actuators) through different communications to collect and provide data. This new technology has grown rapidly, finding applications in multiple sectors such as energy efficiency, health care, industry 4.0, security and public protection logistics and transport, etc.

The concept of IoT adapted to marine environments is known as the Internet of Underwater Things (IoUT) [2-5] and consists of a network of interconnected and intelligent underwater objects, such as sensors, probes or autonomous vehicles. To support the IoUT, the Underwater Wireless Sensor Networks (UWSN) [6-7] are considered a promising network system. Since the technologies of communication and waterproofing of equipment are in a mature phase, it is an appropriate time to investigate in this field.

2 Proposed research plan

This PhD project aims to research on Artificial Intelligence (AI) techniques that support coordination and management of UCH. These algorithms and models, capable of generating knowledge, will be incorporated into a platform based on IoUT technologies and the Edge Computing paradigm. The platform will integrate information stored in databases with data acquired in real time. The different stages of the data are displayed in the figure below (Fig. 1).

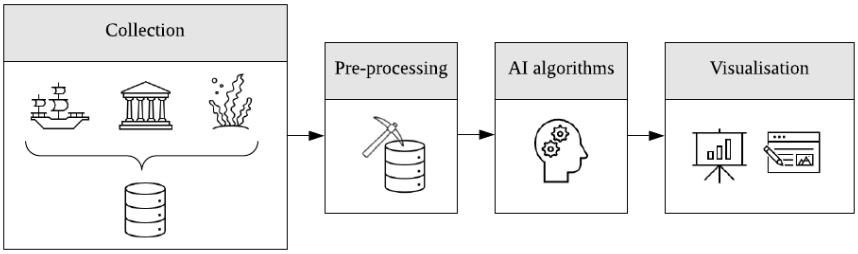

Fig. 1. Proposed data stages for Underwater Cultural Heritage management.

The general objective is divided into the following specific objectives:

- To design a database containing environmental factors and parameters involved in the deterioration of UCH (temperature, pH, salinity, conductivity, marine currents, biological growth, etc.), including existing databases and information measured on-site.
- To study the application of IoUT in cultural heritage conservation.
- To research on AI algorithms, and to develop predictive models capable of quantifying UCH degradation phenomena in a changing environment.
- To develop a platform integrating a Service-Oriented Architecture (SOA) on underwater networks of IoUT sensors.

This research is linked to the project *Technological Consortium TO develop sustainability of underwater cultural heritage (TECTONIC)*, which started in February 2020 [8]. Funded by the European Union's Horizon 2020 programme, its main objective is the implementation, improvement and evaluation of innovative materials, techniques, instruments and methodologies for conservation, restoration and management of UCH.

For the obtention of the PhD degree, the modality initially chosen is "Thesis by Compendium of Articles/Publications", in which the student must submit at least three articles or chapters published or accepted in journals in the field of research chosen. The estimated duration of this research is three years. An Action Plan is being defined to ensure compliance with the objectives on time, and a Quality Plan will be prepared for the monitoring and evaluation of the research.

During the first year, a literature review will be conducted on the following topics: UCH degradation phenomena, AI techniques, IoUT, UWSN, SOA and Multi-Agent Systems (MAS), the Edge Computing paradigm, norms and standards for UCH conservation. Additionally, a database including critical environmental factors in UCH degradation phenomena will be created. Participation in 2-month research exchanges (as part of the secondment plan of the TECTONIC project) are planned for the whole PhD duration. The first year these will be dedicated to the collection of data at the TECTONIC's pilot sites: Italy, Greece and Argentina. During the second year, AI algorithms (classification, clustering and regression, initially) will be studied for predicting the status of the underwater environment. Moreover, the UCH management platform proposed will be designed. The third year will be dedicated to implement and evaluate the platform at the pilot sites.

Several resources will be available for the development of the proposed platform. On the one hand, the existing resources in the BISITE Research Group, which count with an extensive research career in IoT and its applications in different fields [9-14]. On the other hand, the connection of this PhD research with the TECTONIC project will allow access to its resources, the possibility of implementing the proposed platform in the pilot sites, and the access to educational and training activities.

3 The role of CBR in UCH management

The seabed is a giant museum. Underwater Cultural Heritage includes three million shipwrecks, cities and ruins, and thousands of prehistoric sites [15]. Conservation methods and restoration activities are still complex and expensive, making almost impossible to individually analyse the vast number of submerged historical items/sites. In addition, the increasing impacts of climate change continuously threat submerged heritage.

To cope with these challenges, the Case-Based Reasoning (CBR) methodology is proposed as a sustainable approach. When a new UCH case is found, it will be compared with a UCH *case base* of global items/sites previously analysed and restored.

The most similar cases will be *retrieved* from the UCH *case base*, and the conservation approach and restoration treatment will be *reuse* for the new UCH case discovered (after *revision*).

4 Progress to date

This research is at a very early stage. The past few months have been dedicated to reviewing the literature, although several more months will be needed to complete this phase. The first research exchange was programmed for the summer of 2020. However, due to the health crisis caused by the COVID-19, the secondment plan of the TECTONIC project will change, affecting the PhD timing schedule.

So far, dissemination activities include the submission of a paper for the 17th International Conference on Distributed Computing and Artificial Intelligence (DCAI); and the preparation of two Doctoral Consortiums (DCAI and ICCBR).

Acknowledgments. This research has been supported by the project "Technological Consortium TO develop sustainability of underwater Cultural heritage (TECTONIC)", financed by the European Union (Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 873132).

References

1. UNESCO: LEGAL INSTRUMENTS, Convention on the Protection of the Underwater Cultural Heritage, last accessed 2020/06/03.
2. Kao, C-C.; Lin, Y-S.; Wu, G-D. and Huang, C-J.: A Comprehensive Study on the Internet of Underwater Things: Applications, Challenges, and Channel Models, Sensors, 17(7), 1477 (2017).
3. Nordrum, A.: A language for the internet of underwater things [News]. IEEE Spectrum, vol. 54, no. 9, pp. 9-10 (2017).
4. Liou, E.; Kao, C.; Chang, C.; Lin, Y. and Huang, C.: Internet of underwater things: Challenges and routing protocols, 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, pp. 1171-1174 (2018).
5. Xu, G.; Shi, Y.; Sun, X. and Shen, W.: Internet of Things in Marine Environment Monitoring: A Review. Sensors, vol 19, issue 7, 1711 (2019).
6. Jouhari, M.; Ibrahimi, K.; Tembine, H. and Ben-Othman, J.: Underwater Wireless Sensor Networks: A Survey on Enabling Technologies, Localization Protocols, and Internet of Underwater Things. IEEE Access, vol. 7, pp. 96879-96899 (2019).
7. Urunov, K.; Shin, S.; Namgung, J. and Park, S.: High-Level Architectural Design of Management System for the Internet of Underwater Things. 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), Prague, pp. 326-331 (2018).
8. TECTONIC, https://www.tectonicproject.eu/, last accessed 27/04/2020
9. Sittón-Candanedo, I.; Alonso, R.S.; Corchado, J.M.; Rodríguez, S.; Casado-Vara, R.: A review of edge computing reference architectures and a new global edge proposal. Future Generation Computer Systems, vol 99, pp. 278-294 (2019)

10. Sittón-Candanedo, I.; Hernández-Nieves, E.; Rodríguez-González, S.; Santos-Martín, M.T.; González-Briones, A.: Machine learning predictive model for industry 4.0. International Conference on Knowledge Management in Organizations, Springer, pp. 501–510 (2018).
11. Tapia, D. I., Fraile, J. A., Rodríguez, S., Alonso, R. S., & Corchado, J. M.: Integrating hardware agents into an enhanced multi-agent architecture for Ambient Intelligence systems. Information Sciences, 222, 47-65 (2013).
12. De la Prieta, F.; Bajo, J.; Rodríguez, S. and Corchado, J.M.: MAS-based self-adaptive architecture for controlling and monitoring Cloud platforms. Journal of Ambient Intelligence and Humanized Computing, vol 8, no 2, pp. 213-221 (2017).
13. Alonso, R.S.; Tapia, D.I.; Bajo, J.; García, O.; de Paz, J.F.; Corchado, J.M. (2013) Implementing a hardware-embedded reactive agents platform based on a service-oriented architecture over heterogeneous wireless sensor networks. Ad Hoc Networks, 11, 151–166 (2013).
14. González-Briones, A.; De La Prieta, F.; Mohamad, M.; Omatu, S.; Corchado, J.M. Multi-agent systems applications in energy optimization problems: A state-of-the-art review. Energies, 11, 1928 (2018).
15. UNESCO, The World's Underwater Cultural Heritage, http://www.unesco.org/new/en/culture/themes/underwater-cultural-heritage/underwater-cultural-heritage/, last accessed 2020/06/03.

An Explainable Case-Based Recommender System for Legal Reasoning*

Luis Raúl Rodríguez Oconitrillo[1]

Universidad de Costa Rica, San José, CR
luis.rodriguezoconitrillo@ucr.ac.cr.com

Abstract. In this work design science research process is used to create a novel computational platform called RYEL to captures and processes the interpretation and assessment made by a judge about facts and proof to analyze a case. We investigate about explanatory computational and legal explanations models related to the inferences obtained by the platform. Case-Based Reasoning (CBR) is used for handling cases.

Keywords: Case-Based Reasoning (CBR) · Explainable Artificial Intelligence (XAI) · Interpretable Artificial Intelligence · Semantic Networks (SN) · Knowledge Graph (KG)

1 Introduction

Design science research process proposed in [1], was used to generate RYEL wich is platform based on analysis and research develop in [2], [3], [4], [5], within a legal context. RYEL captures and processes the interpretation and evaluation made by a judge about facts and proof, stores it in a Knowledge Representation (KR) [6], and information about laws and norms is generated in order to support decision making [7].

RYEL implements Semantic Networks (SN) [8], [9] through Knowledge Graphs (KG) [10], [11], [11], [12], [13] graphically in order to capture and represent the judge's knowledge structures [14]. The graphics techniques are interactive to represent and process information related to the interpretation and assessment made by a judge. The techniques allowed us to investigate about interface accessibility problems related to explanatory computational models [15]. In this way we have been able to investigate about the legal explanations related to the inferences [15] obtained by the platform. RYEL incorporates granular data into nodes and edges and graph algorithms; This allows us to investigate about interpretation [16] made by a human in a given context, a legal one in our case. Case-Based Reasoning (CBR) [17], [18], [19], [20] is used in order to have a methodological process to manage the case library.

* Acknowledge supervisors from Costa Rica: PhD. Juan José Vargas (Computing Science - AI) - director, PhD. Arturo Camacho (Computing Science - AI) - advisor, PhD. Alvaro Burgos (Law and Criminal Sciences) - advisor. Acknowledge supervisors from Spain: PhD. Juan Manuel Corchado (Computing Science - AI). Acknowledge institutions: BISITE-USAL, Spain and UCR, Costa Rica.

A new method is develop called Interpretation-Assessment / Assessment-Interpretation (IA-AI) consisting in explaining why a machine inferred information in the way it did and why a user interpreted and assessment something in the way he did. The platform has an explanatory and interpretive nature and could be used in other domains of discourse, some examples are: (1) the interpretation a doctor has about a disease and the assessment of using certain medicine, (2) the interpretation a psychologist has from a patient and the assessment for a psychological application treatment, (3) or how a mathematician interprets a real world problem and makes an assessment about which mathematical formula to use. However, now we focus on the legal domain. An exploratory search of bibliographic information between 1986 and 2017 was carried out, and to date no work similar to ours was found, but related works.

2 Problem outline

The proposed system contributes to answering the problem of representing the teleological structure in case-based legal reasoning from Berman and Hafner in [21] where Ronald Loui handle it like a computational "challenge", which we understand as an open problem and he explains it as "... an ontological challenge for the next decades of AI and Law" and consists of processing legal cases using "patterns of interpretation" which are data with a structure that is repeated within a set of information and refers to the way in which a judge understands the laws and factors [16]. Factors can be facts and proof, and the "ontological" concept refers to the formal definition of properties, types and relationships between entities, that is, between objects, existing in the domain of legal discourse. Baude and Sachs restates the problem in [22] as "the law of interpretation" which consists of trying to explain the norms, precepts or principles that govern human interpretation and determine if they are valid and how much influence they have in a case.

Research question: How to capture and represent the processes of interpretation and assessment that a judge makes of the annotations of a file and apply Case-Based Reazoning (CBR) on these processes to generate recommendations prior to the resolution of a case related to jurisprudence, doctrine and norms in different legal contexts? The analysis a judge performs is absolutely subjective, and may differ, depending on the person who performs it, the time it is carried out, the stage of the judicial process in which it is carried out, the type of judicial office in which it is carried out, type of country and type of case (just to mention a few examples).

Traditional forms of Case-Based Reasoning (CBR) have been successful in areas of Common Law or in similar legal fields for legal cases [18] [19] [17], creating algorithms for indexing legal cases [23], working with legal arguments [24], making comparison of legal concepts [25], using case-based argumentation [26], or in applications where a solution for a case is somehow previously outlined [27] [28] [29] [30] [31], or working with case events [32], therefore the consequent is to follow its trend to solve future problems. However, when a legal

framework such as Civil Law takes place, and cognitive process of judge's interpretation is demanding when working in-depth with facts and evidence, the solution depends on a subjective professional criterion making traditional application of case-based reasoning methods difficult to apply in the judge dynamic environment; it is when interpretation and assessment come into play, which are fundamental for legal case reasoning and explanation. These are the bases that promote the Explainable Artificial Intelligence (XAI) in our investigation about interface accessibility [15] and explanation [33] related to the facts and evidence on Civil Law framework.

Traditional machine learning algorithm, for example, Bayesian Networks for legal cases explained by Kevin A. in [20] is not sufficient due to the bias that occurs when working with uncertainty. Lets consider a disease and symptoms, with a Bayesian Network we could represent the probabilistic relationship between them, so, given some clinical or pathological picture we could compute the probability of the existence of various diseases. Now, Bayesian Network can be seen as an extension of propositional logic allowing reasoning with hypotheses, that is, propositions whose truth or falsity are uncertain. Thus, How can be obtained the doctor's interpretation about the patient condition and disease, and how can be explained the assessment does the doctor make to prescribe a medicine?, the answer is, it can't be.

3 Methodology and evaluation

For building the case-base library using knowledge extraction is employed "Grover" methodology [34] to build a semantic structure [35], [9] based on knowledge graphs [36], [6]. Design science research process [1] is used to complete retrive (script patterns, graph algorithms), reuse (laws and norms), revise (knowledge graph) and retain (graph features) process.

4 Conclusions and further work

The platform is approaching to answer the research question posed by Ronald Louis [16]. Currently it is necessary to increase the testing with more judges and rule out any false positive results and include more countries for testing.

A disruptive data analysis strategy uses ordinary technology in a non-ordinary way for the legal domain. These helps to promotes and generates scientific research contributions to the computing field.

In future the use of Case-Based Reasoning and Community Detection in the graph will help to evaluate facts or proofs clustering to improve retrieval process. Case-Based Reasoning and Partitioned Distributed node could improve the reuse laws related to a case, for example, Triangle Counting/Clustering Coefficient algorithms.

References

1. Offermann, P., Levina, O., Schönherr, M., and Bub, U., "Outline of a design science research process," *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, pp. 7–1, 7–11, 2009.
2. Rodríguez, L. R., "Jurisdictional normalization of the administration of justice for magistrates, judges using artificial intelligence methods for legal guidance systems," *II Central American and Caribbean Congress on Family Law, Panamá*, pp. 1–10, 2016.
3. Rodríguez, L. R., "Jurisdictional normalization based on artificial intelligence models," *XX Iberoamerican Congress of Law and Informatics (FIADI) - Salamanca, Spain, october 2016*, pp. 1–16, 2016.
4. Rodríguez, L. and Osegueda, A., "Business intelligence model to support a judge's decision making about legal situations," *IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI), Costa Rica*, pp. 1–5, 2016.
5. Rodríguez, L. R., "Artificial intelligence applied in procedural law and quality of sentences," *XXI Iberoamerican Congress of Law and Informatics (FIADI) - San Luis Potosí, México, october 2017*, pp. 1–19, 2017.
6. Bonatti, P., Cochez, M., Decker, S., Polleres, A., and Valentina, P., "Knowledge graphs: New directions for knowledge representation on the semantic web," *Report from Dagstuhl Seminar 18371*, pp. 2–92, 2018.
7. Khazaii, J., "Fuzzy logic," *Advanced Decision Making for HVAC Engineers, Springer*, pp. 157–166, 2016.
8. Florian, J., *Encyclopedia of Cognitive Science: Semantic Networks.* Hoboken, NJ: Wiley and Sons, 2006. ISBN: 9780470016190.
9. Noirie, L., Dotaro, E., Carofiglio, G., Dupas, A., Pecci, P., Popa, D., and Post, G., "Semantic networking: Flow-based, traffic-aware, and self-managed networking," *Bell Labs Technical Journal, Volumen 14*, 2009.
10. De Riet, R., "Linguistic instruments in knowledge engineering (like)," *Data and Knowledge Engineering, Volume 8*, pp. 187–189, 1992.
11. Zhang, L., *Knowledge Graph Theory and Structural Parsing.* Enschede: Twente University Press, 2002.
12. Singhal, A., "Introducing the knowledge graph: things, not strings." https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.htm, 2012. [accedido: 2018-12-03].
13. Paulheim, H., "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web 0 (2016) 1–0*, pp. 1–23, 2016.
14. Gasevic, D., Djuric, D., and Devedzic, V., "Model driven architecture and ontology development," *Springer-Verlag, Berlin Heidelberg*, pp. 1–310, 2006.
15. Wolf, C. and Ringland, K., "Designing accessible, explainable ai (xai) experiences," *ACM SIGACCESS Accessibility and ComputingMarch 2020 (6)*, pp. 1–5, 2020.
16. Loui, R., "From berman and hafner's teleological context to baude and sachs' interpretive defaults: an ontological challenge for the next decades of ai and law," *Artificial Intelligence and Law*, pp. 371–385, 2016.
17. Ashley, K., "Reasoning with cases and hypotheticals in hypo," *International Journal on Man-Machine Studies 34 (6)*, p. 753–796, 1991.
18. Ashley, K. and Rissland, E., "But, see, accord: generating blue book citations in hypo," *ICAIL '87 Proceedings of the 1st international conference on Artificial intelligence and law*, pp. 67–74, 1987.

19. Ashley, K. and Rissland, E., "A case-based system for trade secrets law," *ICAIL '87 Proceedings of the 1st international conference on Artificial intelligence and law*, pp. 60–66, 1987.
20. Ashley, K., "Case-based models of legal reasoning in a civil law context," *International Congress of Comparative Cultures and Legal Systems of the Instituto de Investigaciones Jurídicas, Universidad Nacional Autónoma de México*, 2004.
21. Berman, D. and Hafner, C., "Representing teleological structure in case-based legal reasoning: the missing link," *ICAIL '93 Proceedings of the 4th international conference on Artificial intelligence and law*, pp. 50–59, 1993.
22. Baude, W. and Sachs, S., "The law of interpretation," *HARVARD LAW REVIEW*, pp. 1079–1147, 2017.
23. Briininghaus, S. and Ashley, K., "Toward adding knowledge to learning algorithms for indexing legal cases," *ICAIL '99 Proceedings of the 7th international conference on Artificial intelligence and law*, pp. 9–17, 1999.
24. Lynch, C., Ashley, K., Pinkwart, N., and Aleven, V., "Toward assessing law students' argument diagrams," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, (Barcelona, Spain), pp. 222–223, ACM, 2009.
25. Grabmair, M. and Ashley, K., "Facilitating case comparison using value judgments and intermediate legal concepts," in *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, (New York, NY, USA), pp. 161–170, ACM, 2011.
26. Aleven, V., *Teaching case-based argumentation through a model and examples, Doctoral Dissertation.* Pittsburgh, PA, USA: University of Pittsburgh, 1997.
27. Branting, K., "Building explanations from rules and structured cases," *International Journal of Man-Machine Studies, 34 (6*, pp. 797–837, 1991.
28. Rissland, E. and Skalak, D., "Cabaret: rule interpretation in a hybrid architecture," *International Journal of Man-Machine Studies, 34 (6*, pp. 839–887, 1991.
29. Pal, K. and Campbell, J., "An application of rule-based and case-based reasoning within a single legal knowledge-based system," *SIGMIS Database journal*, pp. 48–63, 1997.
30. Chorley, A. and Bench-Capon, T., "Agatha: Using heuristic search to automate the construction of case law theories," *Artificial Intelligence and Law, Volume 13, Springer*, pp. 9–51, 2005.
31. Conrad, J. and Al-Kofahi, K., "Scenario analytics: analyzing jury verdicts to evaluate legal case outcomes," *ICAIL '17: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pp. 29–37, 2017.
32. Rodrigues, C., Goncalves, F., and Ribeiro, R., "An ontology for property crime based on events from ufo-b foundational ontology," *5th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 331–336, 2016.
33. Ha, T., Lee, S., and Kim, S., "Designing explainability of an artificial intelligence system," *TechMindSociety '18: Proceedings of the Technology, Mind, and Society*, pp. 1–1, 2018.
34. Plant, R. and Gamble, R., "Methodologies for the development of knowledge-based systems, 1982–2002," *The Knowledge Engineering Review*, pp. 47–81, 2003.
35. Lehmann, F., "Semantic networks," *Computers and Mathematics with Applications*, pp. 1–50, 1992.
36. Yan, J., Wang, C., Cheng, W., Gao, M., and Aoying, Z., "A retrospective of knowledge graphs," *Front. Comput. Sci.*, pp. 55–74, 2018.

A Novel Recommender System Using extracted information from Social Media for doing Fundamental Stock Market Analysis

Niloufar Shoeibi[1]

[1] BISITE Research Group, Universidad de Salamanca, Salamanca, Spain
Niloufar.shoeibi@usal.es

Abstract. In the stock market, there is a chain of factors impacting each other, and this complex network affecting the trends and values in the stock market. When a new event happens, the experts will analyze the effect of this new event on the environment and predict the action with the maximum of winning or the minimum of losing. Hence, the *time* and *the amount of available information* plays the most significant role in having a better analysis. In this paper, I am proposing to consider this tremendous complex environment of the stock market into a case-based reasoning model as a knowledge-base, designed by the expert. It holds a set of rules defined by the previously taken actions in different situations and circumstances of the stock market. The aim is to use this CBR model for automatizing the process of decision making. Social media platforms are an enormous source of information by allowing users to share news and express their thoughts and feelings. By understanding this data and evoke the concepts, the knowledge will be derived. I propose a tool for obtaining information acquired from social media for fundamental analysis, which is relying on the information related to the actual events happening in the world and microeconomics and macroeconomics indicators. The purpose is to maintain more knowledge extracted from social media as an additional source of information to make a more precise analysis, and therefore, better recommendations.

Keywords: Case Based Reasoning, CBR, Social Media, Stock Market, Fundamental Analysis, Social Network Analysis, Influencers, Prediction, Machine Learning, Natural Language processing, Text Mining, Recommender Systems, Knowledge-Based Systems.

1 Introduction and Related Work

For understanding the idea which we will present in this paper, some primary knowledge needs to be discussed for better understanding and in each section, some related work has been reviewed. First, in subsection 1.1, Cased-Based Reasoning has been discussed. In the next subsection, 1.2, we are going to give an introduction about the stock market analysis and the different techniques for doing that. Next, in subsection 1.2, we discuss the social media information and the importance of using this data. And in the last subsection, we will introduce NLP (Natural Language Processing) and talk about its abilities.

1.1 Case-Based Reasoning

Case-Based Reasoning (CBR) is an automated reasoning and decision-making process which solve new problems through the previous experiences that has been done before in solving the past problems. CBR model is a set of simple queries and the possible actions. The new cases like the query will be retrieved from the database and then their solutions will be applying to the new problem [1].

Hiral R. Patel et al. proposed a recommendation model gaining profit the facility of e-Social network analytical study. In this work they focus on the network of economic consultants to build the case-based reasoning model to help the non-professional investors continue investing considering the knowledge of the experts [2].

1.2 Stock Market Analysis

Stock market prediction is all about the methods to use for predicting the value of a company stock. Predicting the trend of the stock market is one of the hot topics of all time. this predictive analysis can be done in 3 ways: Data Mining Analysis, Technical Analysis, and Fundamental Analysis techniques [3]. We will discuss them all below.

1.2.1 The Data Mining analysis

Case-Based Reasoning (CBR) is automated reasoning and decision-making process which solves new problems through the previous experiences that have been done before in solving the past problems. CBR model is a set of simple queries and possible actions. The new cases like the query will be retrieved from the database and then their solutions will be applying to the new problem [1].

Hiral R. Patel et al. proposed a recommendation model gaining profit from the facility of e-Social network analytical study. In this work, they focus on the network of economic consultants to build the case-based reasoning model to help the non-professional investors continue investing considering the knowledge of the experts [2].

1.2.2 The Fundamental Analysis

The Fundamental Analysis is relying on the information received from news, profitability, and macroeconomic indicators like EBITDA, P/E, income, return on equity, and dividend yield. So, the fundamental analysts will buy/sell the stock when the intrinsic is greater/lower than the market price; even though, the defenders of EMH argue that the intrinsic value of a stock is always equal to its current price [7].

1.2.3 The Technical Analysis

The Technical Analysis refers to analysis methodology for forecasting the trend and direction of stock market prices, using the historical data of the stock market, primarily price and volume. In other words, technical analysis makes predictions based on mathematical indicators built from the stock market [8]. It does not use the intrinsic value of

the stock market, instead, it manipulates tables, graphs, and analytical tools for predicting the movements in the price [9].

1.3 Social Media

By the advancement of technologies and artificial intelligence, social media frameworks got more and more popular. It's an easy way for the users to share the moments of their lives day by day [10]. Therefore, it is understandable that nowadays it's a part of the routine life of almost everybody. In other words, the content on social media can be considered as an infinite source of information. Understanding the content and extracting knowledge is a huge advantage because all these interactions happen in real time. Therefore, it is possible to consider all the events without missing any information.

1.4 Text Mining and Natural Language Processing (NLP)

Text mining is deriving meaningful information from a natural language text. On the other hand, Natural Language Processing (NLP) is a branch of artificial science. It analyzes the written texts automatically without human intervention [11]. There are two major components of NLP; Natural Language Understanding and Natural Language Generation.

Natural Language Understanding refers to mapping input into natural language into useful representations and analyzing those aspects of language. However, Natural Language Generation is the process of generating meaningful phrases and sentences in the form of natural language from some internal representations. In general, the purpose of NLP is to make the machine read, understand, and derive meaning from the human language. For doing this, the texts need to go through different processing, such as Tokenization, Stemming, Lemmatization, POS Tags, Named Entity Recognition, and chunking [12]. After applying all these steps, the proper algorithms should be considered to obtain the logic and the true meaning of the text. For instance, Semantic analysis describes the process of understanding natural language [13] or Sentiment Analysis which is the process of extracting the polarity of the text; how positive, negative, or neutral is the text [14].

This paper has been organized as follows: In Section 2, I describe the proposed method, I present a novel architecture for adding more information to the Technical and Fundamental Analyses. And in section 3, I talk about the conclusion and the results of using these architectures. And finally, I will have the References.

2 Proposed Method

As has been discussed in the previous section, the content on social media can be considered as an infinite source of information. This design is focusing on Twitter which is the most news-friendly social media platform. Twitter has a unique characteristic among all social media platforms. In Twitter, users can share their opinions and

viewpoints by tweeting or reacting to each other's tweets pushing like button or leaving a comment. The tweets may contain text, pictures, videos, or links. Also, users can share other users' tweet statuses by retweeting. The data related to the tweet and some information about the profile is available in an entity called Tweet Object, on the JSON format. In this paper, I am going to present an architecture for extracting data and analyze the data into meaningful information to understand the semantic of the tweets for analyzing the stock market, using this interpreted information [15].

The proposed architecture is presented in Fig. 1. This design aims to detect an event from social media in real-time moreover make convenient recommendations due to the detected event. This architecture consists of four different stages. The first step is to keep track of information by monitoring the public content news profiles share. on the next step, a new event will be detected, then the event will be processed for understanding the sense of the event. This step contains a few steps, text preprocessing, Text preparing and Exploratory Data Analysis, Text Representation and Feature Engineering, Modeling or Pattern Mining, and Evaluation and Deployment. and on the last step, the recommendation system will suggest actions based on the combination of the discovered information and the features related to the fundamental analysis like microeconomics and microeconomics indexes and so on.

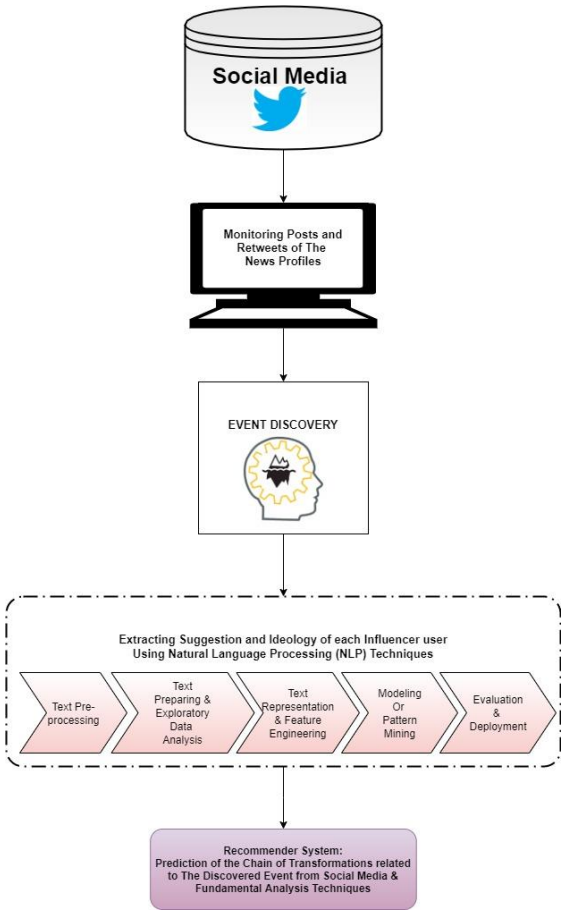

Figure 1 *Proposed Architecture of the Recommender system doing Fundamental Analysis*

3 Conclusion and Future Work

In this paper, I explained the idea of using social media as a great source of information. Also, I proposed a novel architecture to obtain extra knowledge in Fundamental Analysis. Our method is linking NLP techniques, Graph Network Analysis, Behavior Mining, Feature Extraction Strategies, and Machine Learning and Reasoning. After analyzing additive data, new features will be associated with the current methods of Fundamental Analysis, and finally the recommendation system will give pieces of advice based on the derived information and the knowledge-base made by an expert of fundamental analysis.

I believe that by applying this additional source of information and using extra information from social media, understanding the flow of information, discovering new events, and automatizing the procedure in a time so close to the real-time, I will have more precise predictions, and therefore more beneficial recommendations as well as fewer costs of human resources by making a knowledge-based model trained with the knowledge of the experts.

References

1. Lamy, J. B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial intelligence in medicine, 94, 42-53.
2. Patel, H. R. (2019). Analytical Study for Hybrid Method based Stock Recommendation. Journal of the Gujarat Research Society, 21(6), 227-234.
3. Sethi, K. K., Ramesh, D., Rathore, A., & Sarin, S. (2019, July). HUIM-SMP: High utility itemset mining based stock market analogy. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
4. X.-Y. Qian and S. Gao, "Financial series prediction: Comparison between precision of time series models and machine learning methods," Mathematics, Computer Science, Economics, 2017.
5. R. Singh and S. Srivastava, "Stock prediction using deep learning," Tools Application, vol. 76, pp. 18569–18584, 2017.
6. Montenegro, C., & Molina, M. (2020). Improving the Criteria of the Investment on Stock Market Using Data Mining Techniques: The Case of S&P500 Index. International Journal of Machine Learning and Computing, 10(2).
7. Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications, 135, 60-70.
8. Sang, C., & Di Pierro, M. (2019). Improving trading technical analysis with tensorflow long short-term memory (LSTM) neural network. The Journal of Finance and Data Science, 5(1), 1-11.
9. Sagala, T. W., Saputri, M. S., Mahendra, R., & Budi, I. (2020, January). Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis. In Proceedings of the 2020 2nd Asia Pacific Information Technology Conference (pp. 123-127).
10. Dolan, R., Conduit, J., Frethey-Bentham, C., Fahy, J., & Goodman, S. (2019). Social media engagement behavior. *European Journal of Marketing*.
11. Juhn, Y., & Liu, H. (2019). Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. The Journal of Allergy and Clinical Immunology, 145(2), 463-469.
12. Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International journal of medical informatics.
13. Salloum, S. A., Khan, R., & Shaalan, K. (2020, April). A Survey of Semantic Analysis Approaches. In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 61-70). Springer, Cham.
14. Rajput, A. (2020). Natural Language Processing, Sentiment Analysis, and Clinical Analytics. In Innovation in Health Informatics (pp. 79-97). Academic Press.
15. Haustein, S. (2019). Scholarly twitter metrics. In Springer Handbook of Science and Technology Indicators (pp. 729-760). Springer, Cham.

Natural Language Generation for Business Processes

Ashish Upadhyay

Robert Gordon University,
Aberdeen, UK
a.upadhyay@rgu.ac.uk

Abstract. Natural Language Generation from structured information is a common use-case in various business processes. Automatically generating human-like text is a challenging task because grammatical rules are complex and evaluation of the generated text is also not easy. A text generation system is evaluated on the grounds of accuracy, readability and diversity at the same time. Apart from all this, businesses processes suffer from the problem of cold-start as well. In this research, I aim to develop hybrid systems for text generation from structured information using Textual Case-Based Reasoning and Deep Learning. The generated texts will be evaluated using novel metrics capable of measuring accuracy, readability and diversity. Also, I aim to develop novel Information Extraction techniques for case-base generation from unlabelled data.

Keywords: Natural Language Generation • Deep Learning • Textual Case-Based Reasoning • Text Evaluation • Case-Base Development

1 Introduction

The evolution of the web has led to an enormous growth in the amounts of data and information available for organisations to employ in their everyday operations. Different business processes such as medical analysis, compliance requirement management, anomaly reporting require a lot of textual documents to be processed in order to produce effective results. Applying novel Natural Language Processing and Text Mining techniques to these document analysis tasks can reduce human effort and error, thus reducing the overall cost.

Automatic generation of natural language text can be one such task. Natural Language Generation (NLG) can be broadly categorised into two different tasks: text-to-text generation, where natural language text is generated from a textual input only; and data-to-text generation, where text is generated from a non-linguistic and structured input [3]. In this PhD project I aim to work towards the data-to-text generation tasks. Through academic research I wish to solve the problems involved in real-world applications, such as smart home automation - where the requirement can be to summarise a huge amount of tabular data generated from sensors in textual format.

One approach for such tasks can be using a standard abstract template with some pre-defined attributes available as slots to be filled by corresponding values. But a generic template results in repetitive text and is unsuitable for complex scenarios. Textual Case Based Reasoning (TCBR) provides an opportunity to develop dynamic templates with diverse text using previous solutions on similar problems [1,4]. After the advancements in Deep Learning (DL), neural methods capable of learning the semantic relation between data and text through iterative training on large datasets have also become popular [5,7].

The effectiveness of a data-to-text generation system is measured in terms of accuracy, readability and the diversity of texts generated from the system [3]. The generated text should be accurate in terms of including the information provided while maintaining diversity in different generations for similar kind of input. The system should also respect the grammar rules of natural language and texts should be readable as if it was written by a human.

Apart from this, the real-world applications suffer from the problem of cold-start as well. When the employed systems receives the data from a category that wasn't seen during training phase - it may perform inadequately. For example, a system trained for summarising the regulatory information of industrial equipment may not perform very well for a new equipment which wasn't available during training time.

Most of the methods proposed in academic research don't consider these special cases and thus often suffer to accomplish the desirable result when applied to real-world applications. Through this research, I wish to push the boundaries of academic research and develop effective NLG systems that can work in a cold-start scenario.

2 Research Question and Objectives

This research will explore the problems involved in Natural Language Generation (NLG) from structured information (or **data-to-text generation**) for business processes in a cold-start scenario. The primary research question for this project will be - **"Can Deep Learning improve NLG performance for business processes in a cold-start scenario?"**. I hope to meet the following objectives in order to answer the research question:

O1 Develop novel **data-to-text generation** techniques for cold-start scenario.
- TCBR methods offer accurate text generations with limited data but may fail in complex scenarios.
- DL methods offer diverse and fluent texts but require huge amount of labelled data for training.

O2 Propose new **evaluation techniques** keeping accuracy, readability and diversity into account.
- Human evaluation is always best but very expensive to get - need of automated evaluation metrics.
- Case Alignment techniques used in CBR methods are specific to CBR and cannot be applied to other generation methods.

- Popular metrics based on n-gram overlap such as BLEU and ROGUE scores prefer readability over accuracy.

O3 Develop effective methods to **generate case base**.
- Different techniques will have different data requirements which needs to be catered with minimal cost involved.
- Manual labelling is very expensive and also time consuming, whereas automated labelling techniques are fast but lack precision.

3 Proposed Research Plan

3.1 Text Generation

TCBR methods often generate accurate texts with small amount of labelled data given for training. But these templating methods often tend to make grammatical mistakes such as: using 'he' or 'she' as pronoun when the subject of the sentence is 'female' or 'male' respectively; or, using one '.' instead of two '..' at the end of 'His name is Neymar Jr.' [6]. Although most of these errors can be handled by adding different rules for adaptation process, but it's hard to define rules for every possible scenario.

On the other hand, **DL methods** trained on lots of labelled samples offer much more diverse and grammatically correct texts. But these DL systems may hallucinate by generating misleading and non-cohesive texts [7]. DL methods can also be used for solution adaptation after generating text using TCBR. Neural models trained for grammatical correction can be transferred into our use-cases to eliminate the requirement of defining rules for solution adaptation. A survey on the developments in DL methods for data-to-text generation tasks can be explored to identify their drawbacks in a cold-start scenario.

In a cold-start scenario, where labelled data is collected as a part of the process - lazy learned TCBR systems can be a better option instead of end-to-end trained Deep Learning systems. Since, TCBR systems learn from the previous experience when a target problem is received, it will be easier to incorporate new labelled samples (generated during the process) into the decision making process, rather than retraining the whole model from scratch as required for DL systems. Thus, a two-step approach can also be explored where at the initial stage - TCBR methods are used for text generation with few labelled data, and as the labelled data is generated by time - DL methods can be employed for diverse generations.

3.2 Evaluation of Generated Text

Human evaluation is undoubtedly the best evaluation metric for NLG, but is expensive and inconvenient. Automated evaluation metrics for different NLG systems still remain an ongoing challenge. For this, novel **Case Alignment** techniques for CBR methods will be explored addressing the current drawbacks. One such measure using discounted cumulative gain is proposed in recently accepted paper at ICCBR 2020.

The case alignment techniques have a demerit - they can only be used for the CBR systems. Accordingly we plan to develop evaluation metrics that are agnostic of the text generation method. i.e. metrics can be used with both neural and CBR methods. A new family of **Extractive Evaluation** methods cab be developed. The motivation is to use Information Extraction (IE) techniques to extract the useful entities from generated text and matching them with the structured information provided as input. They have recently gained some popularity for the evaluation of different neural text generation systems [7,5]. A similar technique, "**Average Attribute Error**", was used in the accepted paper at ICCBR 2020, where we compared the number of features included in the generated text with the provided input.

3.3 Case-Base Development

For the initial development of case-base, a domain expert may require to manually label few samples from the pool of unlabelled samples. Selecting the most informative samples from the pool of unlabelled samples for manual labelling will result in the gain of better performance with lesser data, instead of random sampling [2]. For this objective, **Active Learning** can be explored to gain better performance with lesser data labelled data. After we have few labelled samples, different IE techniques such as: text classification; or named entity recognition can be used for **Automated Sequence Labelling** to generate more labelled samples from unlabelled documents.

3.4 Data

There are several public datasets available for data-to-text generation problems. These public datasets can be used to simulate a cold-start scenario where each sample in the training set will be marked with a timestamp. A small list of publicly available datasets is given here:

1. **Boxscore**: The dataset consists of NBA basketball game box and line scores with a summary of the game [1].
2. **E2E**: The dataset is a crowd-sourced description of restaurants aligned with the summary of that description [2].
3. **WebNLG**: RDF triple aligned with textual paragraphs summarising the information in those triples [3].

I'll also try to contact industrial experts for gaining access to their real-world data. It might not be possible to publicly distribute the data but the different algorithms developed using that data can be surely published. As of now, I have an **Obituary Dataset** available which is a collection of obituaries

[1] `https://github.com/harvardnlp/boxscore-data/`
[2] `http://www.macs.hw.ac.uk/InteractionLab/E2E`
[3] `https://webnlg-challenge.loria.fr/challenge_2017/#data`

from Scotland aligned with the structured representation of details about the corresponding deceased person. The data is used in my current accepted paper at ICCBR 2020, more details about the paper in next section.

4 Current Progress

I am currently at the initial stage of my research: reviewing the current literature for all three objectives parallely and developing my research methodology. A paper on automated **Natural Language Generation** for obituaries has been accepted at **ICCBR 2020**. The paper presents a case-based method for generating obituaries which is then evaluated using a novel case alignment metric. The CBR method is developed using an initial case-base containing 100 manually labelled samples. Using this obituary generation data I plan extend the case-base with the help of active learning as well as different information extraction techniques.

Another paper on **Text Classification** has been accepted at **IEEE CEC 2020**. The paper uses an ensemble of different machine/deep learning classifiers and representation weighted using PSO for predictions. Experimental results exhibit the advantage of our method over several state-of-the-art text classification algorithms on smaller datasets. This idea will be particularly helpful in the automated labelling of textual documents for the expansion of case-base and aligns with the third objective.

References

1. Adeyanju, I.: Generating weather forecast texts with case based reasoning. International Journal of Computer Applications 975, 8887
2. Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C., Xu, H.: A study of active learning methods for named entity recognition in clinical text. Journal of biomedical informatics 58, 11–18 (2015)
3. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research 61, 65–170 (2018)
4. Massie, S., Wiratunga, N., Craw, S., Donati, A., Vicari, E.: From anomaly reports to cases. In: International Conference on Case-Based Reasoning. pp. 359–373 (2007)
5. Puduppully, R., Dong, L., Lapata, M.: Data-to-text generation with content selection and planning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6908–6915 (2019)
6. Reiter, E.: Nlg vs templates: Levels of sophistication in generating text (Dec 2016), `https://ehudreiter.com/2016/12/18/nlg-vs-templates/`
7. Wiseman, S., Shieber, S., Rush, A.: Challenges in data-to-document generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2253–2263. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017), `https://www.aclweb.org/anthology/D17-1239`

Robust Adaptation in CBR

Xiaomeng Ye

Luddy School of Informatics, Computing, and Engineering
Indiana University, Bloomington IN 47408, USA
xiaye@iu.edu

Abstract. Case adaptation largely determines the flexibility of a Case-based Reasoning (CBR) system by potentially providing unseen solutions adapted from stored knowledge. Traditional case adaptation is done a single-step process, attempting to modify the retrieved case to be closer to the query at hand by certain adaptation rules. A CBR system can modifying a retrieved case multiple times using multiple rules, resulting in a chain of adapted cases and adaptation rules, or an adaptation path. This research proposes techniques to enable to robust adaptation (ROAD) for the building of adaptation paths. Moreover, the adaptation paths built offers introspective learning opportunities to improve aspects of the CBR system other than the case adaptation process.

Key words: adaptation path

1 Introduction

The coverage of case base and the potent of case retrieval underline the performance of a CBR system by correctly finding a similar case for a given query. On the other hand, case adaptation is a critical component determining the flexibility of the system, by adapting the retrieved case when its solution does not solve the query [6]. Case adaptation is also one of the more difficult and less studied processes in CBR. One major breakthrough in case adaptation is the introduction of case difference heuristic [4], which uses pairs of cases to generate adaptation rules. A rule is composite of the problem difference and the solution difference between a pair of cases. If the problem difference between a retrieved case and a query matches that of a rule, then the CBR system can apply the solution difference to the retrieved case and generate a potential solution for the query.

However, the traditional adaptation process by applying single rules might not suffice in many situations. This can happen when not enough adaptation rules are stored or when the query is novel such that even the most similar case in the case base cannot be adapted to solve the query [5]. A multi-step adaptation is one way to extensively adapt cases toward the query where a case is adapted by multiple rules, one at a time, resulting in a chain of adapted cases and the corresponding adaptation rules used. Such a chain is the adaptation path. Adaptation paths are previously discussed by D'Aquin, Lieber, and Napoli [2], Badra, Cordier, and Lieber [1], and Fuchs et al. [3].

2 Proposed Approach

Following the existing works, this research proposes robust adaptation (ROAD) to aid the building of adaptation paths. After the case retrieval process, the case retrieved is supposedly the nearest neighbor to the query. Based on the assumption that each adaptation rule may introduce error and computational cost, shorter adaptation paths are preferred over longer ones. Therefore the problem of adapting a retrieved case to a query can be transformed into a shortest path finding problem from the retrieved case to the query.

The core of ROAD (in its current design) has three aspects: 1) Heuristics for initializing and pursuing multiple adaptation paths concurrently; 2) Heuristics for extending a single adaptation path and terminating it when appropriate; 3) Heuristics for upholding the reliability of adaptation paths by resetting a path to a stored case [5].

The resulted adaptation paths from ROAD also provides introspective learning opportunity for the CBR system: 1) When the heuristics correct an adaptation path by resetting to another stored case, it shows that the initially retrieved case is harder to adapt to the query than the reset case; 2) The accuracy of the final result of an adaptation path can indicate the reliability of the adaptation rules involved in the path. Moreover, the more paths an adaptation rule is involved in, it is more compatible with other rules; 3) A high-traffic region of adaptation paths indicate region of interest in the case space. Retaining cases in such a region can benefit the paths passing the region thus increasing the performance of the CBR system.

3 Current Progress

3.1 Robust Adaptation Algorithm

The core aspects of ROAD is implemented and tested in Leake and Ye [5]. Given a query, ROAD first performs k-nn to retrieve k cases but starts paths from a subset of those k cases selected for diversity. Filtering for a subset of diverse cases can reduce the number of adaptation paths. Moreover, a diverse set of initial cases indicate that the paths are going to work toward the query from different directions using different set of adaptation rules, potentially providing more benefit as an ensemble of solutions.

To expand each adaptation path, ROAD applies an adaptation rule to the current case at the head of the path. A greedy search is used to find the rule that results in a case most similar to the query case. If the result case has already been considered along an adaptation path, or if the result is impossible in the task domain (predetermined by domain knowledge), then the next best rule is chosen. In contrast to a greedy search, a complete search is possible but too costly, considering that a path can involve any number of steps within its length limit and each step can be one of many adaptation rules.

During the expansion of an adaptation path, it may go astray or the path might grow too long. In a domain where the adaptation rule is not perfectly

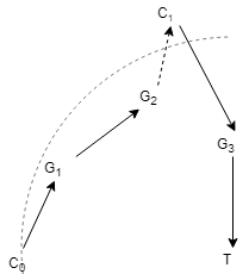

Fig. 1. Illustration of Path Resetting, from Leake and Ye [5]

reliable and might introduce error, a long adaptation path compounds errors of the rules involved, therefore degrading the quality of the final solution. To decrease path lengths, ROAD uses greedy search of adaptation rules and limits the number of rules in an adaptation paths. Moreover, ROAD also resets an adaptation path to a stored case, effectively resetting the reliability of the path by restarting from a stored case of ground truth.

The resetting mechanism triggers if an adaptation path gets close to a stored case other than the initially retrieved case. As illustrated in Figure 1: "C_0 and C_1 are two cases in the case base. Given a query Q, for which the solution case would be T, the source case C_0 is retrieved. Here C_0 is closer to T than C_1 according to the similarity metric (indicated by the dashed arc indicating a radius of equal similarity values). As adaptation rules are applied successively, ghost cases G_1 and G_2 are generated. The ghost case G_2 is found to be more similar to C_1 than C_0. In this situation, ROAD resets the path to C_1. This is expected to increase reliability, because C_1's reliability is guaranteed (because it is a real case), while G_2 is a case produced after adapting C_0 twice. At this point, the path continues from C_1 and yields the ghost case G_3, which is then adapted to T [5]".

The resetting mechanism also triggers when two adaptation paths meet each other (the problem descriptions of their head cases are similar) but disagree on the solutions. In this situation, at least one of the solutions provided by the two adaptation paths are inaccurate, and a reset is applied to the path of longer length.

In both scenarios, the resetting mechanism searches for the nearest neighbor of the current head case of the adaptation path. If the nearest neighbor is closer to the head case than the initially retrieved case of the path, then the adaptation path is reset by appending the nearest neighbor as the new head of the path.

3.2 Hindsight of Robust Adaptation

A current study (to be presented on ICCBR2020) is focused on using the hindsight of ROAD to improve the system efficiency.

Improve Similarity Metric to Reduce Resets A path is reset when its head gets close to a stored case. This happens when the similarity metric does not align

with the adaptability of the domain. In other words, the retrieved most similar case is not the easiest to adapt to the query.

After using ROAD, the adaptation paths built can provide insight on the room of improvement for the similarity metric. Given a query, an adaptation path with reset indicates a reset case should have been retrieved in place of the actually retrieved case, and an adaptation path without reset indicates a case is correctly retrieved. If the similarity metric allows modification based on such feedback, then it can be improved. With the improved similarity metric, the CBR system is less likely trigger the reset mechanism, therefore leading to shorter paths and higher efficiency.

Filter for More Reliable Rules The accuracy of the final result of an adaptation path can reflect the compatibility and reliability of the adaptation rules involved in the path. If the final result is accurate, then the adaptation rules involved are compatible with each other and they provide reliable results after being applied in a sequence.

In this experiment, the system attempts to build adaptation paths by combining any two rules and applying them to all applicable cases. If the combination of the two rules can be applied to many cases, then the two rules are compatible in relatively larger scope of the task domain. If the final results after the adaptation is accurate (evaluated by comparing the final solution with its nearest neighbor in the case base), then the two rules are reliable when used together. By studying the compatibility and reliability of pairs of rules, researcher can find the most compatible or the most reliable rules and filter out the less useful ones. A preliminary experiment shows that filtering for more compatible/reliable rules produces shorter paths and reduces the computation overhead of ROAD. The result also shows that the accuracy of ROAD may improve due to usage of more reliable rules, however, the accuracy may degrade when too many rules are removed and the adaptation rule set fails to provide good coverage of the task domain.

4 Future Directions

4.1 Using Hindsight of ROAD to Facilitate Case Retaining

A high-traffic region of adaptation paths is a region in the case space where multiple adaptation paths get close to or cross each other. Many adapted cases are generated in such a region as paths pass through. If there is a stored case in this region, then the paths can potentially reset to the stored case to uphold the quality of the final solutions. If the region is less populated with cases, then it is beneficial to retain cases for reasons described above, as also suggested by Mathew and Chakraborti [7].

4.2 Applying ROAD to Existing CBR Systems

Case adaptation determines the flexibility of the CBR system. ROAD, as a general scheme, is applicable to almost all CBR systems and can enhance their

flexibility. One future direction is to apply ROAD scheme to existing CBR systems and examine its potential in improving the adaptability as well as other powers of the systems.

Currently the ROAD experiments are conducted on task domains where adaptation rules are generated using case difference heuristics. In certain CBR systems (e.g. the CookingCAKE system [8]), adaptations can also be done by generalization, specialization, and composition. It is worth exploring how ROAD may interact with various kinds of adaptations.

Acknowledgment: I want to thank Dr. David Leake for his guidance on the incubation and throughout the development of this research idea.

References

1. Badra, F., Cordier, A., Lieber, J.: Opportunistic adaptation knowledge discovery. In: Case-Based Reasoning Research and Development, ICCBR 2009. pp. 60–74. Springer, Berlin (2009)
2. D'Aquin, M., Lieber, J., Napoli, A.: Adaptation knowledge acquisition: a case study for case-based decision support in oncology. Computational Intelligence 22(3/4), 161–176 (2006)
3. Fuchs, B., Lieber, J., Mille, A., Napoli, A.: Differential adaptation: An operational approach to adaptation for solving numerical problems with CBR. Knowledge-Based Systems 68, 103–114 (2014)
4. Hanney, K., Keane, M.: Learning adaptation rules from a case-base. In: Proceedings of the Third European Workshop on Case-Based Reasoning. pp. 179–192. Springer, Berlin (1996)
5. Leake, D., Ye, X.: On combining case adaptation rules. In: Case-Based Reasoning Research and Development, ICCBR 2019. pp. 204–218. Springer (2019)
6. López de Mántaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M., Cox, M., Forbus, K., Keane, M., Aamodt, A., Watson, I.: Retrieval, reuse, revision, and retention in CBR. Knowledge Engineering Review 20(3) (2005)
7. Mathew, D., Chakraborti, S.: Competence guided model for casebase maintenance. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. pp. 4904–4908. International Joint Conferences on Artificial Intelligence (2017)
8. Mller, G., Bergmann, R.: CookingCAKE: A framework for the adaptation of cooking recipes represented as workflows (01 2015)