Preface: The ICCBR 2014 Doctoral Consortium

The objective of a doctoral consortium (DC) is to nurture the interests of newcomers who recently started studying a specific research field. A DC provides participants with an opportunity to describe and obtain feedback on their research, future work plans, and career objectives from senior researchers and peers. For the case-based reasoning (CBR) community, the DC at ICCBR is *very* important because it provides a unique forum for the community to welcome, guide, and encourage junior researchers who may become active (and even leading) community members. The ICCBR held its first DC at ICCBR 2009. The DC events are now an established and expected part of ICCBR. The submission numbers are now consistent with the submission level of the conference.

The DC call was widely advertised to identify prospective participants, asking applicants to submit: (1) a 1-page Application Cover Page; (2) a 3-page Research Summary; (3) a 1-page Resumé; and (4) an acknowledgement from the student's advisor. The summary requires students to describe their objective, progress, and plans using the conference's publishing format, the Resumé describes the applicant's experience, and the acknowledgement ensures that advisors are aware of this event.

The DC program committee (PC) reviewed and accepted applications inviting them to participate. PC members provided feedback to research summaries submitted by all applicants, with each applicant receiving at least feedback from two PC members. We assigned one on-site mentor per student, matching mentors who could provide valuable feedback from multiple perspectives. Mentors also guided students to prepare presentations for the DC.

The ICCBR 2014 DC program started on September 28th with a meet and greet session, followed by an invited talk by David Wilson (University of North Carolina), and dinner. On September 29th, the program started with an invited talk by David Leake (Indiana University), entitled Career Cases: Tips for CBR Doctoral Students. The rest of the program consisted of 15-minute talks presented by the 14 doctoral. Presentations were followed by a question-answer session on presentation skills and content led by their mentor. Each mentor was asked to attend at least 2 students' presentations, thus allowing them to also attend co-timed events.

We want to thank David Wilson and David Leake for their invited talks. Thanks to the 14 DC students and their advisors. We are indebted to all PC members who provided important and useful guidance to DC students, either as reviewers or as mentors. Thank you CBR researchers for being mentors, providing this invaluable service to the community. Thank you to the Artificial Intelligence Journal for a donation that allowed a substantial reduction on student's fee. Thank you to the conference chair Derek Bridge for his suggestions, guidance, and efforts in support of the 2014 DC.

We hope that the ICCBR 2014 DC enhanced each student's interest in studying CBR. We expect that students felt welcomed, supported, and encouraged to pursue their research in CBR further. We strongly encourage them to participate in future ICCBR conferences and related venues. We wish the students well!

September 2014

Rosina Weber
Nirmalie Wiratunga

Organization

ICCBR 2014 DC Chairs

Rosina Weber	College of Information Science and Technology, Drexel University; Philadelphia, PA, USA
Nirmalie Wiratunga	IDEAS Research Institute, Robert Gordon University, Aberdeen, Scotland

ICCBR 2014 DC Program Committee

David Aha, Naval Research Laboratory, USA
Klaus-Dieter Althoff, DFKI / University of Hildesheim, Germany
Derek Bridge, University College Cork, Ireland
Amelie Cordier, LIRIS, France
Sarah Jane Delany, Dublin Institute of Technology, Ireland
Luc Lamontagne, Laval University, Canada
David Leake, Indiana University, USA
Jean Lieber, LORIA - INRIA Lorraine, France
Cindy Marling, Ohio University, USA
Stewart Massie, The Robert Gordon University, UK
David McSherry, University of Ulster, UK
Stefania Montani, Università del Piemonte Orientale A. Avogadro, Italy
Mirjam Minor, Goethe University Frankfurt, Germany
Santiago Ontanon, Drexel University, USA
Pinar Ozturk, Norwegian University of Science and Technology, Norway
Miltos Petridis, Brighton University, UK
Enric Plaza, IIIA-CSIC, Spain
Thomas Roth-Berghofer, University of West London, UK
David Wilson, University of North Carolina, USA

Sponsoring Institutions

We would like to thank the donation of 2,000 Euros from the Artificial Intelligence Journal (AIJ) division of the non-profit organization IJCAI.

DC Participants, Institutions

Xavier Ferrer Aran, IIIA-CSIC, Spain
Luca Canensi, Università del Piemonte Orientale A. Avogadro, Italy
Yoke Yie Chen, The Robert Gordon University, UK
Dileep Kvs , Indian Institute of Technology Madras
Pádraig Ó Duinn, University College Cork, Ireland
Emmanuelle Gaillard, Universit e de Lorraine, LORIA
Hugo Hromic, National University of Ireland Galway
Racha Khelif, FEMTO - ST Institute, France
Gulmira Khussainova, Nottingham University Business School, UK
Khalil Muhammad, University College Dublin, Ireland
Gilbert Müller, University of Trier, Germany
Pol Schumacher, Goethe University Frankfurt, Germany
Egon Sewald Jr, Federal University of Santa Catarina, Florianópolis, SC, Brazil
Stefan Wender, University of Auckland, New Zealand

DC Participants, Advisors

Xavier Ferrer Aran, Enric Plaza
Luca Canensi, Stefania Montani
Yoke Yie Chen, Nirmalie Wiratunga
Dileep Kvs , Sutanu Chakraborti
Pádraig Ó Duinn, Derek Bridge
Emmanuelle Gaillard, Emmanuel Nauer
Hugo Hromic, Conor Hayes
Racha Khelif, Brigitte Morello
Gulmira Khussainova, Sanja Petrovic
Khalil Muhammad, Barry Smyth
Gilbert Müller, Ralph Bergmann
Pol Schumacher, Mirjam Minor
Egon Sewald Jr, Aires Jose Rover
Stefan Wender, Ian Watson

Table of Contents

The Web of Experience: Reusing Other Peoples Experiences in Analytical Tasks

Xavier Ferrer Aran

Artificial Intelligence Research Institute (IIIA),
Spanish Scientific Research Council (CSIC),
Campus UAB, Bellaterra, Spain
{xferrer}@iiia.csic.es

1 Introduction

People share experiences and opinions of the real world on the web. For instance, product reviews on e-commerce sites or forums where people request and offer service on everyday tasks. People read and rate comments created by others users about products they want to acquire. A user may have a task –for example, to buy a camera, to book a hotel room or to cook a healthy meal– in which she would benefit from the experience of others. The user needs to discover relevant experiences and reuse them to obtain actionable knowledge to accomplish his task. And internet is full of people's experiences. However, existing tools such as Internet search engines or document classification algorithms, treat experiences on the web no differently from the way they treat other web content. For them, content is treated as documents.
In this project we will treat textual records of experiences as a special kind of content. We want to present an approach to analyze and discover the practical knowledge tacitly contained in analytical tasks of records of people's experiences present for a particular domain with a finite and enumerable set of items. Our research will produce concepts, methods and tools that will support organizing people's experiences in a way that they could be reused as actionable knowledge. Similar work in this field can be seen in [1, 2]. Although the results look promising, there is much potential to improve and extend this work by, for example, exploring different techniques for topic mining, sentiment aggregation and different feature weighting models.

2 Research Aim

The aim of this research is to reuse people's textual experiences to help other people. The first step in this task is to discover what are the arguments that define an experience. To this end, we need to extract and analyze the information presented in the text to build a concept vocabulary formed by important aspects contained in experiences. The extraction of the vocabulary is one of the most important parts of this research: a rich vocabulary will lead to a better

definition of arguments, improving our precision when retrieving similar experiences. Using this vocabulary, we create a prototype of each entity based on the aspects used by people in their experiences about that entity. After this step, we search for positive and negative arguments expressed by users in their experiences considering all different entities. Sentiment analysis can be used for that purpose. Finally, those prototype's arguments are used to recommend to new users depending on the importance a user gives to the different aspects present in the vocabulary.

3 Proposed Plan of Research

1. **Creation of vocabulary.**
   Aspect extraction. Explore knowledge-light and knowledge-rich approaches to extract meaningful aspects and relations from opinion text. This step is of vital importance in the project, since a precise vocabulary is necessary to define experiences correctly. This step aims to explore different statistical methods such as word co-occurence or word frequency and compare them with other knowledge-rich approaches sich as *Wikipedia*[4], and *WordNet*[3, 5].
   Aspect clustering. Explore clustering strategies to group similar entities that refer to the same concept. This part is important to get rid of redundant aspects that can add noise to our system. Here we want to study different clustering methods such as top-down approaches or hierarchical bottom-up clustering.
2. **Detect the arguments that define an entity's prototype and the polarity associated**.
   Polarity of the sentiment. Once the vocabulary is defined, we need to identify people's positive/negative sentiment about certain arguments of their experiences. In this step, we will check current state of the art techniques to identify whether the arguments in user's experiences are positive or negative.
   Aspect weighting. Not all the aspects are equally important for users. In this part we want to explore techniques that combine user information, such as collaborative filtering or content based filtering, to determine which aspects are most interesting.
3. **Argument structure**. Based on the analysis of experiences, we want to explore the possibility of creating an argument tree using the vocabulary found in the first and second step and the sentiments associated. Based on their experiences, users can have different opinions about the same aspect and those different opinions can be aggregated and organized using an argument tree.
4. **Reuse of previous experiences**. This step aims to make use of all the practical information obtained in previous steps and presented in the argument structure to help other individuals to make informed decisions. We plan to do this by creating a personalized recommender system able to personalize recommendations given a set of user experiences and preferences.

4 Current Progress

Progress made to date is presented below:

- **Study of the state of the art.** During my first year I have been mainly focused in studying the current state of the art to approach the topic. I have been studying different areas related to *Natural Language Processing*, *Case Based Reasoning* and *Information Extraction*, paying special attention to recommender systems based in users reviews.
- **Aspect extraction.** As one of the most important parts, we proposed a novel algorithm that combines dependency extraction rules with word frequency filtering. Potential aspects are extracted by means of the dependency rules and are filtered depending on a frequency-cut off. The proposed approach was evaluated against state of the art techniques and obtained positive results, notably improving the *precision* of the aspects extracted compared to other techniques. Note that precision is important in aspect extraction and specially for our system, since further steps (such as determining the arguments of an experience) clearly depend on achieving a certain level of proficiency on topic identification.
- **Weighted sentiment scoring**. We proposed an algorithm that incorporates aspect importance weight and sentiment distribution. We investigated two different resources that infer the importance of product aspects: product preference and temporal dynamics. We studied how user preferences over a product change over time and how our system recommendations are affected by those changes.
- **Evaluation metric for recommender system**. In the absence of ground truth data, we derived approximations from the Amazon data we had crawled and proposed an evaluation metric derived from Amazon's reviews, questions and timeline data.

References

1. Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, and Alberto Díaz. A joint model of feature mining and sentiment analysis for product review rating. In *Advances in information retrieval*, pages 55–66. Springer, 2011.
2. R. Dong, M. Schaal, M. OMahony, K. McCarthy, and B. Smyth. Opinionated product recommendation. In *Inter. Conf. on Case-Based Reasoning*. 2013.
3. Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.
4. Peter Schönhofen. Identifying document topics using the wikipedia category network. *Web Intelligence and Agent Systems*, 7(2):195–207, 2009.
5. Sabrina Tiun, Rosni Abdullah, and Tang Enya Kong. Automatic topic identification using ontology hierarchy. In *Computational Linguistics and Intelligent Text Processing*, pages 444–453. Springer, 2001.

A tool for mining and checking processes

Doctoral Consortium ICCBR 2014

L. Canensi

Department of Computer Science, Università di Torino, Italy
canensi@di.unito.it

1 Research Summary

Today, organizations are beginning to emphasise the corporate governance, risk management and compliance. Various regulations are focused specifically on the compliance of business management rules and shares norms. The techniques of Process Mining offers tools to tighten up conformity of services and to ascertain the validity and reliability of the information relating to the organization's key process. For example, hospitals try to streamline their process. In order to do so, it is essential to have an accurate view of the "careflows" under consideration. Starting from data related to patient care process for a certain disease, might be useful verify whether different hospitals implement different processes even when following the same guideline, or whether different patient categories are differently treated. My PhD Thesis aims to improve the mining of the model process [1] and the comparison between models. Therefore, my Phd Thesis is composed of two steps:

- the first step is focused on build the model process
- the second step concerns the comparison of the model processes learned

1.1 Process Mining

The main part of the work is Process Mining (PM). This discipline focuses on the study of models of real processes starting from log data. In particular, it comes to handling a data collection (event log), where each element is related to a process instance in the real world. Each event consists of an ordered sequence of activities. The mined model process can be used to understand, adapt and modify the real process to increase performance and become an high quality process.

Process Mining include (automated) process discovery, conformance checking (monitoring deviations by comparing model and log), social network/organization mining, automated construction of simulation models, model extension, model repair, case prediction and history-based recommendations [2]. My research, however, we will only deal with process discovery and conformance checking techniques. Discovery Process is the most relevant and widely used process mining activity. A conformance checking that we will used in my work takes in input

two model process and compare them, to measure the similarity between them. This process mining activity wil be discussed in section 1.2.

Most of the phases of a process life cycle [3] can benefit of process mining techniques, they can be adopted to analyse an existing model, to diagnose problem, and possibly to adapt/redesign/tune the process model itself.

All these considerations lead to define process mining as a very important instrument for modern organizations that need to manage non-trivial operational processes. It is obviously vital that process mining results are correct and reliable as much as possible, in order to facilitate the work of process engineers and company decision makers. Indeed, a miner may produce a model process that includes a path never observed in the event log. This can be very harmful in some applications (like, e.g., patient management / disease treatment), and, generally, in all those cases in which the quality of the process has to be assessed.

My research is focused on the creation of a new process mining tool based on algorithm studied to guarantee an output model only including paths that actually correspond to a trace in the event log. In order to realize this objective the miner takes into account of the different execution contexts and makes an intensive use of all the frequency information that appear in the log. The process model will be represented as a graph, where the nodes are events and the edges represent a direct sequence. In an abstract way the miner could have a granularity problem (i.e. a model process is described in more detail than another one), but the problem will be solved using semantical analysis. However, this issue are not addressed in my Phd Thesis, because in the specific application in which it will be tested, the terms and activities are very homogeneous.

1.2 Comparison of mined process

After that the model process was built, start the conformance checking step where the model process is compared with another one. Generally, this technique takes an existing process model and compares it with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. One of them could be the real process. Hence, conformance checking may be used to detect, locate and explain deviations, and to measure the severity of these deviations. Sometimes, however, it can be useful to compare two models of the same process. This is especially useful, if the processes are executed in different organizations or by different resources. To this end, Case Based Reasoning (CBR) can been proposed as a natural methodological solution. In particular, as mentioned earlier, the mined process models are represented in the form of graph (where nodes represent activities and edge provide information about the control flow) and in order to compare two models will be used appropriate metrics for the comparison of graphs. My aim is to retrieve similar processes and to quantify the differences between them through an appropriate metric. In order to retrieve process models and order them on the basis of their distance with respect to a given query model, I will analyse the distance definition that are used by the CBR community for a graph data structure. Normally, to compare this kind

of data are used metrics based on the graph edit distance [4, 5]. Such a notion calculates the minimal cost of transforming one graph into another by applying insertions/deletions and substitutions of nodes, and insertions/deletions of edges. The distance should be adapted to consider particular process information, such as time, resource, duration, etc... As clearly stated in [6, 7], three classes of similarity metrics can be considered to deal with process model comparisons: (i) node (or edge) matching similarity, which basically compares the labels attached to process model; (ii) structural similarity, which compares node labels, as well as graph topology; (iii) behavioural similarity, which compares node labels, as well as the behavioural/causal relations captured in the process models. My research is related to class (ii). The state of the art on structural similarity on process models is represented by the work by Dijkman et al. [4]. This metrics is "blind" and completely context-independent since that is independent of the application domain. On the other hand, in particular domains (including medical applications) there is available more knowledge and its exploitation can surely improve the quality of any automated support to the expert's work [8]. However, my research is at a preliminary step and this question will be deepened after the part of process discovery.

References

1. C. Rolland, "A comprehensive view of process engineering," in *Advanced Information Systems Engineering, 10th International Conference CAiSE 98, Pisa, Italy, June 8-12, 1998, Proceedings* (B. Pernici and C. Thanos, eds.), vol. 1413 of *Lecture Notes in Computer Science*, pp. 1–24, Springer, 1998.
2. "$http://www.win.tue.nl/ieeetfpm$." IEEE Taskforce on Process Mining: Process Mining Manifesto (last accessed on 4/11/2013).
3. W. Scacchi and P. Mi, "Process life cycle engineering: A knowledge-based approach and environment.," *Int. Syst. in Accounting, Finance and Management*, vol. 6, no. 2, pp. 83–107, 1997.
4. R. Dijkman, M. Dumas, and R. Garcıa-Banuelos, "Graph matching algorithms for business process model similarity search," in *Proc. International Conference on Business Process Management* (U. Dayal, J. Eder, J. Koehler, and H. Reijers, eds.), vol. 5701 of *Lecture Notes in Computer Science*, pp. 48–63, Springer, Berlin, 2009.
5. H. Bunke, "On a relation between graph edit distance and maximum common subgraph," *Pattern Recognition Letters*, vol. 18(8), p. 689–694, 1997.
6. R. Dijkman, M. Dumas, B. VanDongen, R. Kaarik, and J. Mendling, "Similarity of business process models: metrics and evaluation," *Information Systems*, vol. 36, pp. 498–516, 2011.
7. M. Becker and R. Laue, "A comparative survey of business process similarity measures," *Computers in Industry*, vol. 63, pp. 148–167, 2012.
8. R. Basu, N. Archer, and B. Mukherjee, "Intelligent decision support in healthcare," *Analytics*, vol. jan-feb 2012, pp. 33–38, 2012.

Aspect-based Sentiment Analysis for Social Recommender Systems

Yoke Yie Chen

IDEAS Research Institute,
Robert Gordon University,
Aberdeen, Scotland
{y.y.chen}@rgu.ac.uk

1 Introduction

Recommender systems aim to provide users with a list of recommended items by exploiting knowledge from user preferences [7], their information needs [3] or by exploiting similar behavior of other users [1]. Representation, similarity and ranking algorithms from the Case-Based Reasoning (CBR) community has naturally made a significant contribution to recommender systems research [5]. The social web has create opportunities for new recommendation algorithms to utilise knowledge from such resources and so the emergence of social recommender systems.

Social recommender system harness knowledge from user generated content to generate better recommendation. This is because a significant component of knowledge about user preferences is implicit in online product reviews. One method to harness knowledge from product reviews for recommendation task is by incorporating sentiment expressed in opinions to bias the recommendation list [2]. The product aspects (e.g. picture, price and weight) and sentiments described in the product reviews are served as the basis to represent a product case for recommendation that emphasis on similarity and sentiment. Dong et al.[2] uses shallow NLP techniques and sentiment informed frequency pruning to extract aspects from reviews. A sentiment lexicon was then used to identify potential aspects. However, the use of sentiment lexicon to identify sentiment words has its limitations. Sentiment word in sentiment lexicon may not bears sentiment. For example, "I am looking for a *good* point and shoot camera". The word *good* here bears neither positive nor negative sentiment to the camera. Therefore, the set of aspects generated contain noisy aspects (non-genuine aspects).

A user's purchase decision hints at the aspects that are likely to have influenced their decision and as such be deemed more important. To understand the importance of an aspect to users, it is necessary to further reveal the importance weight that users placed on an aspect. Additionally, user preferences change over time. Term frequency (TF) is the naive approach for this task where the weight of an aspect to be equal to the number of occurrence of an aspect [2]. However, this approach is not able to capture users' preference that change over time and so the importance of the aspects.

The proportion of positive and negative sentiments for an aspect is an important piece of information. Lexicons are often used to ascertain the polarity (positive or negative) and strength of sentiment expressed at word-level. However, averaging sentiment scores for an aspect does not reflect users' consensus of a sentiment when sentiment scores contain extreme values (e.g. strong positive or negative). Sophisticated methods are needed to aggregate these scores at the sentence, paragraph and document level to account for the distribution of sentiments.

2 Research Aim and Proposed Plan of Research

The aim of this research is to investigate and define new techniques for high quality recommendation to be made through social content. Our particular focus will be on using product reviews to develop novel algorithms for product recommendation. For this purpose, we intend to:

1. **Analyze the advantage and disadvantages of using product reviews for product recommendation.** The potential of using product reviews for product recommendation is yet to be fully explored. Therefore, we intend to study on the metadata of product reviews (e.g. review helpfulness, temporal information). Specifically, we want to know what are the similarity knowledge resources that can be exploited for product case representation?
2. **Develop product case representation that take advantage of the social knowledge available in product reviews.** Aspects of a product case generated using shallow NLP techniques have some ability to extract potential aspects. However, this approach produce high amount of potential aspects which leads to high computational costs and noisy aspects (non-genuine aspects). Furthermore, the importance of an aspect is more effectively derive from product reviews where users explicitly expressed their preferences. Thus, we aim to answer the following questions:
   - How to infer importance of product aspect using temporal information?
   - How to aggregate sentiment scores from individual aspect to account for sentiment distribution and aspect importance?
3. **Design, develop and evaluate a social recommender system.** The product case representation developed in objective 2 will be used in the ranking strategy of a social recommender system. Products are recommended based on similarity and user preference. Recommendation quality will be evaluated based on accuracy of the recommender system in predicting users preferred products.

3 Current Progress

Designed and developed novel algorithms in the following areas:

- **Aspect extraction**. Our proposed aspect extraction algorithm (FqDPrules) integrate semantic relationship and frequency cut off. We compare our proposed approach with the following alternative extraction algorithms:
  - FqItemsets uses Apriori algorithm to identify candidate aspects that are then pruned using a frequency cut-off threshold [4].
  - FqPos uses Part-of-Speech(POS) extraction patterns that are then pruned using sentiment informed frequency cut-off threshold [2].
  - DPrules uses the dependency extraction rules [6].

  Our results show that FqDPrules outperforms other alternative algorithms by gaining a higher precision value. We advocate higher precision because this would mean we are able to identify true aspects of a product and this will lead to better learning of users preference.
- **Aspect-based sentiment scoring**. The proposed algorithm incorporate aspect importance weight and sentiment distribution. We investigated two different resources to infer aspect importance: preferences from purchase summary statistics and temporal information. Initial results suggest that both knowledge sources are able to improve recommendations.
- **Ranking Strategy**. Using top n recommendation approach, we compare the performance of two different knowledge sources for recommendation: aspect similarity and sentiment scores. Our initial results show that product aspect similarity plays an important role in improving recommendation quality.

References

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17, 2005.
2. R. Dong, M. Schaal, M. OMahony, K. McCarthy, and B. Smyth. Opinionated product recommendation. In *Inter. Conf. on Case-Based Reasoning*. 2013.
3. S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. In *The adaptive web*, pages 54–89. 2007.
4. M. Hu and B. Liu. Mining and summarising customer reviews. In *Proc. of ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, 2004.
5. L. McGinty and B. Smyth. Collaborative case-based reasoning: Applications in personalised route planning. In *Case-Based Reasoning Research and Development*, pages 362–376. 2001.
6. S. Moghaddam and M. Ester. On the design of lda models for aspect-based opinion mining. In *Proc. Inter. Conf. on Information and Knowledge Management*, CIKM '12, 2012.
7. S. Vasudevan and S. Chakraborti. Mining user trails in critiquing based recommenders. In *Proc. Inter. Conf. on World Wide Web Companion*, pages 777–780, 2014.

Intelligent Integration of Knowledge Sources for Textual Case Based Reasoning(TCBR)

KVS Dileep

Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600036
kvsdilip@cse.iitm.ac.in

Abstract. Wikipedia, WordNet and word co-occurrences are knowledge sources that have been used in text processing to solve various problems. In text classification, these knowledge sources are used to enrich the document representation and increase classification accuracy. Easy articles can be classified correctly by looking at just the title or author while others might require more processing to classify correctly. Such an approach has interesting correspondences to how humans progressively resort to more involved processing once shallower approaches fail. As part of doctoral research, we attempt to propose a method, where we try to suggest a principled combination of knowledge sources for each document that uses minimal knowledge to solve the problem. This is done by estimating the difficulty in processing the documents and suggesting the right knowledge source that would reduce this difficulty.

1 Introduction

A pivotal problem in TCBR is case representation. The bag-of-words approach, where a case is represented as a vector of keywords, is generally used for this purpose. Similarity measures devised on such representations might be insufficient in retrieving cases of high utility. By high utility, we mean that the solution of the retrieved case will most probably be reused to solve the query problem. To address this problem, the system needs access to additional 'knowledge', which leads to a richer case representation thus improving case base competence. The knowledge sources generally used to enrich the textual representation are Wikipedia, WordNet [1] and word co-occurrences.

Let us now consider a simple task like reading a newspaper article. We often non-consciously integrate various sources of knowledge to understand the article [2]. While classifying the article for example, it is usual for us to use minimal knowledge to arrive at our decision, say a title or first few lines of the article. Only when the task is quite difficult, like trying to distinguish between articles on closely related themes such as IBM hardware and Apple hardware, we read deeper into the article to reach our decision. This dynamic integration of knowledge on an on-demand basis is an important tool for humans. Is it possible to replicate the same in machines? Or more formally, given the different knowledge sources, when and how can these knowledge sources be integrated? Can we derive important insights from the characterization of the data for efficient integration?

To the best of our knowledge, existing techniques process all documents in a uniform manner and perform an ad hoc integration of knowledge. We try to propose an alternate method as part of our work, where we perform a principled integration of knowledge sources as demanded by the estimated difficulty of the document being processed. With this framework, we wish to draw a parallel to how humans resort to additional knowledge and deeper processing when shallow approaches fail.

2 Research Plan and Methodology

The research goal we wish to achieve is to support the following hypothesis - *"documents easy w.r.t a given task can be processed with lesser knowledge compared to those that are difficult"*. CBR has a rich literature on complexity that we wish to use for our work. We aim to quantify 'ease' or 'difficulty' by estimating complexity of documents. Currently we focus on text classification due to the ease in evaluation, but in future we might take up other tasks as well. The initial part of work consists of studying the problem of estimating complexity. For our method to be successful, the complexity measure must be able to predict the document difficulty accurately. The next step would be to use the complexity measure as a guiding tool to decide whether to integrate knowledge or not.

We now define the notions of 'breadth' and 'depth' in knowledge sources. 'Breadth' refers to features of the document, while 'depth' is used to denote knowledge sources used for enriching the features. A document's sections like title, author, introduction form the breadth while knowledge like Wikipedia and word co-occurrences constitute the depth. For example, an easily classifiable newspaper article may be classified correctly with just the words in title, while a hard article might be correctly classified only after reading deeper into the article along with background knowledge. Our method aims to find the right feature-knowledge combination for classification. In the framework of intelligent integration, we would like to demonstrate through experimentation that documents easy to classify require shallower feature-knowledge combination for correct classification compared to documents that are difficult to classify.

We wish to extend our framework to other problems related to textual CBR. Repositories like stack overflow, yahoo answers and quora contain a rich collection of questions and answers spread over different topics. A recent work tries to suggest query expansion for ill-formed queries on textual CBR systems using complexity measure in the backend [3]. Intelligent integration of knowledge sources will have many interesting applications for retrieval on textual CBR systems. Further we can extend the method to process datasets that are non-textual in nature.

3 Research Progress

In the initial part of our work, we looked at complexity measures defined in literature. We proposed a new complexity measure based on estimating intrinsic dimension of data as opposed to estimating difficulty based on extrinsic dimension of data. Extrinsic dimension is number of dimensions used to represent the data while intrinsic dimension is the minimum number of dimensions data can reside without losing relative

distance information. The new measure based on fractal dimension has been applied to both textual and non-textual data with satisfactory results. The new measure estimates collection complexity and we assume that datasets with more intrinsic dimensions are complex compared to those with less intrinsic dimension. This work has been accepted in ICCBR 2014.

For the initial work on complexity guided integration of knowledge sources, we chose the complexity measure proposed by Massie et al. [4]. This measure estimates how well similar problems in the case base correspond to similar solutions or in other words how 'well aligned' are they. High alignment implies low complexity and vice versa. For classification, when similar documents have similar labels the dataset is well aligned and it does not depend on the dimensionality. But if approaches like Latent Semantic Analysis bring documents with similar labels closer, then alignment will increase and in turn the classification accuracy too will increase.

We proposed a method of selective integration of knowledge, SelInt. We take a fixed combination of representations like title, title with background knowledge, whole document in bag of words representation, whole document enriched with Wikipedia knowledge and so on. SelInt takes a test document as input and estimates its difficulty in a shallow representation like words from titles only. If we find that the expected alignment of test document is high, we proceed with classification in the current representation and add no further knowledge. Otherwise, we integrate knowledge and re-estimate its alignment and proceed as before. We also choose the right knowledge to integrate by using ideas in reminding. We look at the representations of the neighbours of query which yielded a correct label during training in a leave-one-out scenario. We use a majority voting scheme with alignments as weights to choose the right knowledge source.

Our work estimates query complexity to decide on the appropriate choice of knowledge representation, with the basic premise that shallow representations should be preferred over deep ones, unless the latter promises substantial gains based on precomputed estimates of neighbourhood complexity. We have experimented with several textual datasets of varying complexity. It was encouraging to observe that not only were redundant query revisions avoided, there were also effectiveness gains in certain cases.

References

1. Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
2. Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.
3. Deepak P, Sutanu Chakraborti, and Deepak Khemani. Query Suggestions for Textual Problem Solution Repositories. In *ECIR*, pages 569–581, 2013.
4. Stewart Massie, Nirmalie Wiratunga, Susan Craw, Alessandro Donati, and Emmanuel Vicari. From Anomaly Reports to Cases. In Rosina O. Weber and Michael M. Richter, editors, *Case-Based Reasoning Research and Development*, volume 4626 of *Lecture Notes in Computer Science*, pages 359–373. Springer Berlin Heidelberg, 2007.

Classification of Posts to Internet Forums

Pádraig Ó Duinn*

Insight Centre for Data Analytics,
School of Computer Science and Information Technology,
University College Cork, Ireland
padraig.oduinn@insight-centre.org

1 Background

Internet forums are a popular place for people to discuss a wide variety of topics. Forums comprise threads, each of which is a dialogue on a particular topic. Users post messages in order to contribute to the dialogue. Thread topics vary from the quite general, such as current affairs, to the more specific, such as technical support for a particular device.

Forum users assume different roles on different occasions. Sometimes a user may be seeking information, either by posting a question or reading existing threads. Other times, users contribute information by posting messages that respond to questions posted by other users. Forums often appoint human moderators whose task it is to intervene to enhance thread quality and enforce forum policies.

Forums store posts to threads in chronological order and present them linearly in the same order (or in reverse order). But the argument structure of the thread may be other than linear. Subtopics and digressions are really branches from the main topic, meaning that, semantically, the thread structure is actually tree-like. As the thread grows, users may find it increasingly difficult to extract information or make relevant, new contributions since the linear presentation obscures the argument structure.

We are looking at developing tools to support the different users of forums.

2 Progress

We have focused so far on the task of post classification in technical question-and-answer (Q&A) forums. The goal of post classification is to predict a label for each post, according to its role within its thread. These role labels are based on the main dialog-act present in the post. In the context of Q&A forums, post labels might include Question, Answer, Clarification, and so on. We believe that identifying the role of each post is a first step to revealing argument structure and thereby improving the usability of Internet forums.

* The authors gratefully acknowledge the financial support of the Irish Research Council. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

For the task of post classification, we have looked at using *collective classification*. Traditional classifiers predict the labels of objects independently of predicting the labels of related objects. Collective classifiers, on the other hand, predict the labels of related objects simultaneously: prediction of one object's label makes use of information from the related objects, including their currently-predicted labels. Collective classifiers repeatedly classify an object, each time incorporating features or information derived from the previous classifications. In many domains, collective classification is more accurate than traditional classification. While there has been some work that develops collective classifiers for tasks such as web page classification [3], these collective classifiers have not been applied to forum post classification.

We based our initial experiments on the work of *Kim et al.* [2]. We used their CNET dataset, along with the FIRE dataset used by *Bhatia et al.* [1]. Most importantly, as far as possible we used the same experimental methodology as they did: 10-fold cross-validation, based on partitioning the dataset into training and test sets, where the partitioning was performed at the thread level, i.e. the training set contained complete threads, as did the test set. We compared a collective classifier (namely, the Iterative Classification Algorithm, ICA) with the Linear-Chain Conditional Random Field (CRF) classifier that *Kim et al.* found to have highest accuracy on their dataset. We found that ICA's accuracy was comparable with that of the CRF.

A methodology that partitions forums at the thread level lacks realism. In an active forum, threads grow over time. A classifier will be called upon to classify posts incrementally. One of our main contributions has been to design a number of experiment protocols that, while they are still dataset-driven simulations of forum activity, we believe are more realistic. In particular, they treat posts in chronological order; they use the first 100 days' posts as a training set; they test predictions for the remaining days' posts, producing a daily accuracy figure. They also allow for the possibility of daily re-training of the classifiers.

Our different protocols simulate different levels of intervention on the part of human experts. For example, in one protocol, daily re-training uses the true labels of each day's posts — this simulates, the somewhat unrealistic scenario in which a human moderator provides the correct label for every newly-arrived post. In another protocol, daily re-training uses the *predicted* labels of each day's posts — this simulates, the possibly also unrealistic scenario in which the human moderator play no part in labelling posts. These two protocols are extremes, and we have defined others that lie between these extremes.

Encouragingly, ICA has significantly higher accuracy than the CRF across all our protocols and for both datasets. We also experimented with cautious collective classification techniques. While these have been shown to improve the accuracy of collective classifiers in some cases, we did not see any significant improvements.

More recently, we have been experimenting with different feature sets for representing the posts. In particular, we have tried incorporating the lexical content of the posts using tf-idf term vectors. We had not done this in the earlier

experiments for comparability with *Kim et al.*'s work, where lexical content was not used. Instead we represented a post as a vector of numerical features derived from the post itself (such as the post's position in thread, and the number of hyperlinks in the text) and from the post's relation to other posts in the data set (including information about the creators post history on the forum).

3 Plan

For the task of post classification, there is still a number of areas to explore. Collective classifiers use a notion of neighbourhood: an object is classified using the predicted labels of other objects in its neighbourhood. An object's neighbourhood consists of other objects which are in some way related. In the context of post classification, there are several ways we could potentially consider two posts to be related (for example two posts from the same users). So far, our definition of neighbourhood has been the previous posts in the thread. It would be interesting to extend this. For example, we could revise a post's predicted label in the light of objects that come subsequent to it in the thread. Perhaps most interesting is to extend collective classification to objects other than posts, such as threads and users.

The overall goal of this research is to develop tools that assist the users of Internet forums. So far, we have been using data taken from Q&A forums. In these forums the threads usually consist of an intial question, followed by several answers and posts providing additional information. In more general forums this structure may not be present, and posts may fill a different set of roles. It remains to be seen if our approach will as be effective on non-Q&A forums.

As discussed earlier, the argument structure is tree-like, not linear. Forum users may find it easier to assimilate what is going on in a thread whose structure is presented in this fashion. It would also be interesting to investigate if there are other representations which would increase accessibility for users.

To facilitate the presentation of thread argument structure, a next step in our research may be prediction of link structure. This task involves identifying a 'link' between two posts. If we were to present threads in tree structure, for example, the goal would be classify a post as a child of a particular earlier post.

References

1. Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. Classifying user messages for managing web forum data. In Zachary G. Ives and Yannis Velegrakis, editors, *Procs. of the 15th International Workshop on the Web and Databases*, pages 13–18, 2012.
2. Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Procs. of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202. ACL, 2010.
3. Luke McDowell, Kalyan Moy Gupta, and David W. Aha. Case-based collective classification. In David Wilson and Geoff Sutcliffe, editors, *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, pages 399–404. AAAI Press, 2007.

Building a Case-Based Reasoning System Exploiting Knowledge Coming from the Web

Emmanuelle Gaillard

Université de Lorraine, LORIA — 54506 Vandœuvre-lès-Nancy, France
CNRS — 54506 Vandœuvre-lès-Nancy, France
Inria — 54602 Villers-lès-Nancy, France
Emmanuelle.Gaillard@loria.fr

1 Introduction

The aim of my thesis is to exploit knowledge coming from the Web in a case-based reasoning (CBR) system. CBR systems solve new problems by analogy with past experiences, or cases. This process uses knowledge units (KUs) usually separated in several containers: the case base, the domain knowledge, the adaptation knowledge and the similarity knowledge. Acquiring enough knowledge to perform quality reasoning is cumbersome and tedious. This is the reason why the Web, and more specifically, e-communities, is more and more explored to build knowledge bases. For example, [10] presents tools to extract knowledge from the Web, tools which are integrated to the *my*CBR 3 system. However, due to several factors (e.g. the expertise level of users, their points of view), the quality of this knowledge is questionable.

In order to ensure the quality of CBR results, the reliability of the KUs has to be managed in order to use the most reliable KUs. The first part of my work, presented in [3] and evaluated in [4], was to design a meta-knowledge model (MKM) which can be seen as a new (meta-knowledge) container to manage the knowledge reliability. The second part of my work is to study how this new container may be plugged into a CBR system.

2 Research Plan and Progress

State of the Art about Meta-Knowledge. A bibliography work has been realized on notions such as quality, belief, trust and reputation which influence reliability. I studied how these notions are used and linked in computer systems, in particular in reputation-based trust systems [1, 6, 11] and in recommender systems [5], and CBR systems. We have highlighted that no one CBR approach represents meta-knowledge for all KUs of the different containers. In some CBR approaches, meta-knowledge like belief, trust or reputation are only used to describe cases for recommendations [8] or to give adaptation explanations (e.g. [7, 2]).

Use Case and Community Constitution. For my research work, the use case is a French version of the CBR system TAAABLE (`http://taaable.loria.fr`) in the cooking domain. The knowledge used by TAAABLE is shared in a collaborative web site (in French) called ATAAABLE (`http://ataaable.loria.fr`). According to user contraints, TAAABLE searches, in the recipe base of ATAAABLE, whether some recipes satisfy these constraints. Recipes, if they exist, are returned to the user; otherwise the system search similar recipes guided by the domain knowledge edited by ATAAABLE users, in order to relax user contraints. Similar recipes are adapted, creating new ones.

Meta-Knowledge Model. At the end of the first year of my PhD, a meta-knowledge model (MKM) has been proposed in order to manage meta-knowledge such as quality, belief, trust and reputation [3]. Users of the ATAAABLE community may rate KUs representing the belief score and to other users representing the *a priori* trust score. Theses scores are the foundations of the model and allow to compute the intermediate meta-knowledge:

- Quality score of a KU ku represents the global quality of ku for the community, inferred from belief scores about ku.
- Trust score from a user u towards a user v, represents how u trusts v, inferred from belief scores that u has assigned to KU's created by v and the *a priori* trust score of u towards v.
- Reputation score of a user u, represents reputation of u in the system, inferred from all the trust scores about u.

The final score is the reliability score between a user u and a KU ku computed from the intermediate meta-knowledge. Taking into account the trust score between u and v to compute a reliability score between u and ku enables personalized results to be returned to the user u. The reliability score in MKM is used upstream and downstream of the reasoning process, to select relevant knowledge to conduct the reasoning process, and to rank results provided by the CBR engine. All the KUs with a reliability score higher than a given threshold are selected to be used by the CBR engine. Returned CBR results ranked based on the associated reliability of the results.

Reliability of a CBR Result. In the CBR system, results are ranked thanks to the *inferred reliability* score. The *inferred reliability* score of a result is computed from all KUs involved in the computation of this result. Currently, the reliability score is seen as the probability that the result will be satisfactory (e.g., for a cooking application, the probability that this is the recipe of a tasty dish). This probability depends on the probabilities that the retrieved case is satisfactory, and that each KU used in the adaptation is satisfactory (with the assumption that these probabilities are independent one from another).

Methodology Evaluation. A methodology evaluation was proposed and used to evaluate a CBR system using knowledge from the Web with and without the MKM model. This methodology is based on methods used in the evaluation

of recommender systems evaluation (e.g., [9]). A first evaluation was performed on the Taaable system and has concluded that the MKM model integration in Taaable improves the quality of solutions provided by the system [4].

3 Future Plan

- **September 2014**: extending a CBR system with personalization of results, by taking into account the trust propagation.
- **December 2014**: intregration of the MKM model in the CBR engine. The retrieve and adaptation steps will be guided by the reliability of KUs.
- **February 2015**: the theories of fuzzy logic and possibilistic logic will be explored as a means to refine the meta-knowledge representation.
- **June 2015**: final evaluation of the MKM model and of its impact on the CBR results.
- **September 2015**: PhD defense.

References

1. Artz, D., Gil, Y.: A Survey of Trust in Computer Science and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (2007)
2. Briggs, P., Smyth, B.: Provenance, Trust, and Sharing in Peer-to-Peer Case-Based Web Search. In: Proceedings of the 9th European conference on Advances in Case-Based Reasoning. pp. 89–103 (2008)
3. Gaillard, E., Lieber, J., Naudet, Y., Nauer, E.: Case-Based Reasoning on e-community Knowledge. In: Case-Based Reasoning Research and Development, pp. 104–118. Springer Berlin Heidelberg (2013)
4. Gaillard, E., Lieber, J., Nauer, E., Cordier, A.: How Case-Based Reasoning on e-community Knowledge can be Improved thanks to Knowledge Reliability. In: International Conference on Case-Based Reasoning (2014)
5. Golbeck, J.: Computing and Applying Tust in Web-based Social Networks. Ph.D. thesis, University of Maryland (2005)
6. Jø sang, A., Ismail, R.: The Beta Reputation System. In: Proceedings of the 15th Bled Electronic Commerce Conference. pp. 324–337 (2002)
7. Leake, D., Whitehead, M.: Case Provenance: The Value of Remembering Case Sources. In: Proceedings of the 7th international conference on Case-Based Reasoning: Case-Based Reasoning Research and Development. pp. 194–208 (2007)
8. Quijano-Sánchez, L., Bridge, D., Díaz-Agudo, B., Recio-García, J.: A Case-Based Solution to the Cold-Start Problem in Group Recommenders. In: International Conference for Case-Based Reasoning. pp. 342–356 (2012)
9. Quijano-Sánchez, L., Recio Garcia, J., Díaz-Agudo, B.: Using Personality to Create Alliances in Group Recommender Systems. In: Ram, A., Wiratunga, N. (eds.) Case-Based Reasoning Research and Development, Lecture Notes in Computer Science, vol. 6880, pp. 226–240. Springer Berlin Heidelberg (2011)
10. Sauer, C., Roth-Berghofer, T.: Extracting knowledge from web communities and linked data for case-based reasoning systems. Expert Systems pp. n/a–n/a (2013)
11. Young, A.K., Muhammad, A.A.: Trust, Distrust and Lack of Confidence of Users in Online Social Media-sharing Communities. Knowledge-Based Systems 37(0), 438 – 450 (2013)

News Filtering using Structural Approaches

Hugo Hromic

Insight Centre for Data Analytics
National University of Ireland Galway (NUIG)
hugo.hromic@insight-centre.org
http://www.insight-centre.org

Abstract. This document presents a summary of research done by Hugo Hromic, PhD student based in the Insight Centre at the National University of Ireland Galway (NUIG). His research topic is News Filtering and Event Detection in Twitter using structural-based approaches. Current progress and future plans are described.

1 Summary of Research

Social media services are widely integrated into our modern digital lives. They developed from simple blogs into complex multi-functional platforms like Facebook, Twitter, or Flickr. These systems us to almost instantly connect with people and share content, thoughts and personal moments. They also fuel the information overload that is ubiquitous in our global world. Most of these platforms have millions of active users around the globe constantly generating content in real-time. Moreover, common people are accompanied by news agencies, advertising companies and self-marketing celebrities as well.

It is widely understood that the resulting massive stream contains mostly noisy information, but also contains timely and crucially relevant information about news and events of interest happening in real life around us. One particularly hard challenge is to *discover* and *filter* this relevant content.

Many corresponding approaches have been proposed in the literature. One popular class of solutions is based on mining the streams for bursts in activity, usually in the context of monitoring certain predefined keywords [1, 2]. While these approaches are very fast and often effective, they are limited on only the a priori chosen keywords and they can perform very poor when significant signal noise is present. Another class of approaches is based on analysing the content of user posts, e.g., by on-the-fly maintenance and analysis of probabilistic topic models [3, 4]. Such approaches are inevitably sensitive to language, writing style and are often expensive to process. These characteristics can cause significant problems in fast-paced streams shaped by colloquial language, such as Twitter.

We propose to follow a purely structural approach to address the problem of information filtering in social media streams. Our hypothesis is that bursts in user interest – i.e., in certain topics, news, events, etc. – correlate with observable bursts in the interaction dynamics between users. Thus, we focus on mining the graphs built from these interactions. For example, Twitter usually "reacts" to

breaking news by showing a high number of retweet, mention and reply activities that can all be modelled as edges in a graph. In these graphs, we can mine for groups of tightly connected users and assume that all users of such groups are interested in the same topic, which in turn is related to the particular news in question. Likewise, Facebook users post opinions and thoughts, share photos and links, etc.

This structural approach is independent of analysing the actual content, but rather focuses on the implicit behavioural and communication dynamics to extract this content. Based in this assumption, in 2012 we successfully developed a prototype system called **Whassappi** [5], a mobile application for the visitors of the final leg of the Volvo Ocean Race 2012 in Ireland. To the best of our knowledge, this is the very first system for filtering information about news and events in Twitter following a purely structural approach.

We believe that the Case-based Reasoning (CBR) workflow can be applied as well to our research. The problem to be solved is then *how to identify emerging topics in social streams*, and the knowledge base instances are *the different structural dynamics observed that describe events* (see Item 2a in Section 3). Events could be described through different graph and community characteristics that can be compared on-the-fly with incoming real-time dynamics. A key CBR challenge is then how to validate new events observations for storing.

2 Current Progress

Whassappi is currently composed of three main modules: the *Twitter listener*, the *Community Analytics* and the *Mobile Web Application*.

For the Twitter Listener, we have developed a system that connects to the Twitter Streaming API with a fixed set of *seed terms* and then adapts itself according to observed co-occurrence of hashtags and users [5]. The listener also implements a basic spam filtering technique based on user accounts age and a blacklist of words. Moreover, before extending any new hashtag or user the listener checks for potential introduction of noise by performing burst detection of the new candidate terms.

For the Community Analytics, the heart of Whassappi, we designed *GraFEN* (*Graph-based Filtering of Events and News*). With GraFEN we propose a generalised approach for any social media service that supports interactions between users based on exchanging or posting content, however our current implementation is focused on Twitter. We have done exhaustive experimentation based on an innovative approach for generating synthetic data by "injecting" relevant pieces of information using a set of captured Tweets [6]. So far we have used synthetic as well as non-annotated real-world data to assess our approach in comparison to two state-of-the-art methods. The first, TwitInfo [2], is based on identifying activity bursts in sub-streams extracted via filtering of predefined keywords. The second utilises a streaming version of LDA (Latent Dirichlet Allocation) topic models [6]. We are planning to use a different evaluation based on annotated real-world Twitter data [7].

Finally, the *Mobile Web Application* module is a basic user interface that displays the outcomes of the Community Analytics module in real-time. No further research is planned for this component.

3 Future Research

The following is a list of planned upcoming research tasks.

1. Research on the Twitter Listener module.
   (a) Evaluate the self-adapting approach for terms extension.
   (b) Investigate crawling of followers networks to assist GraFEN.
2. Research on the Community Analytics module (GraFEN).
   (a) Investigate methods for event detection based on evolution of the real-time graphs and communities. Usage of signals captured from structural and statistical features – i.e. size, modularity, density, degrees, user and hashtags frequencies, etc. – is planned. A tentative method is to do Curve fitting of these signals to ADSR (attack-decay-sustain-release) envelopes.
   (b) Investigate methods for improving community selection and ranking. Our current proposal uses naive methods for these tasks but gives reasonable results. We believe this can be further improved by investigating hybrid content-based and structural-based methods.
   (c) Use standard and formal methods for evaluating our community finding and event detection capabilities against state-of-the-art approaches.

References

1. Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.
2. Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.
3. Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 536–544. Association for Computational Linguistics, 2012.
4. Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.
5. Hugo Hromic, Marcel Karnstedt, Mengjiao Wang, Alice Hogan, Václav Belák, and Conor Hayes. Event panning in a stream of big data. In *LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML)*, 2012.
6. Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, volume 12, pages 1519–1534, 2012.
7. Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 409–418. ACM, 2013.

Similarity based RUL Prediction Approach

Racha KHELIF

FEMTO - ST Institute, 24 rue Alain Savary, 25000 Besançon, France,
racha.khelif@femto-st.fr

1 Research Scope and Challenges

Prognostics as defined by the 2004 international organization for standardization, is an estimation of time to failure and risks of one or more existing or future failure modes. In our research, we are implementing a prognostic approach for predicting the Remaining Useful Life (RUL) of critical components, where a critical component is a component that might result in big damage for the system when it breaks down. This RUL estimation is a useful information that helps scheduling predictive maintenance actions. Knowing the time needed for an engine to break down will give the maintenance agent time to act and helps avoiding catastrophic situations. In our approach we work with time series data called instances representing health indicators of components and obtained by the fusion of the sensory data while taking into account the order of the input data.

A perspective of our research work is to build a knowledge oriented prognosis system using Case Based Reasoning. The first step consisted in constructing an Instance Based Learning (IBL) that can evolve to CBR. In fact, IBL and CBR share the same basic idea that is learning from seen cases. However, the difference between the two relies in how instances are indexed, retrieved and reused. In IBL a case is a data record and the retrieval step is based on a global similarity metric with no knowledge being generated. Learning is a storage process.

In IBL, we don't acquire any specific knowledge and we believe that CBR can enhance our results by using the learnt knowledge. This transition rises a lot of questions among which: How to acquire knowledge from the sensory data in the absence of any information about the physical system being considered? How to include time, an important notion in prognostics, in the knowledge and adaptation learning?

At this current stage, we tried to overcome two challenges with the objective of building a reliable predictive system.

The first challenge dealt with the system design i.e. instance formalization. The second challenge is to provide the system with the ability to retrieve similar instances by implementing similarity metric takes into account the whole history of data while giving more importance to late observations.

The remaining of this paper describes what has been done so far and the ongoing work.

2 Implemented Approach

IBL approaches for RUL prediction go through three steps: instance formalization, retrieval step and RUL prediction.
The purpose of instance formalization is to construct a library of instances that represents the progression of the components health status and model the degradation evolution. At the retrieval step, problem instances are identified to elements of the library by conducting a similarity measure with the aim of retrieving the most similar instances. The information available in these instances is used for RUL prediction.

2.1 Instance Formalization

Two different approaches have been considered for instance formalization.
Supervised Instance Formalization: Health indicator trajectories are obtained using linear regression [1]. These HI are bounded between '1' and '0', representing respectively healthy and faulty states. The linear regression model was trained using only boundary data (early and late moments of the component's life).
In order to apply such an approach, two more hypotheses need to be satisfied:
- One degradation model for all the train components.
- Train components are in a good health at the beginning and end up developing a fault prior to failure.

Unsupervised Instance Formalization: This time, health indicator trajectories are obtained in an unsupervised way using Unsupervised Kernel Regression (UKR). The application of this approach does not require any additional hypotheses. For each component, a degradation model is learned. [2]

2.2 Retrieval Step

The purpose of the system is to predict remaining useful life of critical components by retrieving the most similar instances and projecting the current time of the test instance on the time axis of similar instance, the projection is assumed to be the end of similarity. RUL is then computed as: End of Life-End of Similarity. This is done by conducting a similarity test between train and test instances with the aim of attributing a similarity score to each train instance.

2.3 Application and Results

The described approach was tested on the diagnostics and prognostics dataset generated by the NASA prognostics center [3]. The dataset simulates the damage propagation of aircraft gas turbine engines. It comes in two files, "train file" and "test file". For the train file, engines are run until failure is reached. Engines of the test file,

on the other hand, are stopped sometime prior to failure. The task is thus to predict the RUL for test engines. The sensory measurements contain information such as the fan temperature, the physical fan speed and static pressure. The two first lines of the table below summarize the results obtained using our approach while the last two ones show results found in the literature from authors who worked on the same dataset. In the future, we are planning to compare results obtained CBR approach with the one we are having with the IBL approach.

Table 1. Results comparison table.

Approach	Rate of right predictions
Supervised instance formalization [1]	54%
Unsupervised instance formalization [2]	**57%**
Ramasso et al.[4]	53%
Javed el al. [5]	48%

3 Ongoing Work

The unsupervised instance formalization shows promising results. However, the way the approach is constructed is heavy in terms of calculation and time consuming. For each test component, all training models are used to obtain the appropriate health indicators. The idea is to acquire knowledge that allows selecting appropriate models instead of using all models available in the library of instances. This is done by constructing cases as clusters of similar trajectories. The first step consists in clustering the trajectories into groups and then learning rules that allows assigning test trajectories into one of the clusters and adaptation rules for changing clusters

References

1. R. Khelif, S. Malinowski, B. Chebel-Morello, & N. Zerhouni (2014). “RUL prediction based on a new similarity-instance based approach”. In Proceeding of the International Symposium on Industrial Electronics. Istanbul, June 2014.
2. R. Khelif, S. Malinowski, B. Chebel-Morello, & N. Zerhouni,(2014). “Unsupervised Kernel Regression Modeling Approach for RUL Prediction”. In Proceeding of the annual conference of PHM; Nantes, July 2014.
3. Saxena, A., Goebel, K., Simon, D.,& Eklund, N. (2008, October). “Damage propagation modeling for aircraft engine run-to-failure simulation”.In International Conference on Prognostics and Health Management, 2008. PHM 2008. (pp. 1-9). IEEE.
4. Ramasso, E., Rombaut, M., & Zerhouni, N. (2013). “Joint prediction of continuous and discrete states in time-series based on belief functions”. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 43(1), 37-50.
5. Javed, K. (2014). “A robest and reliable data-driven prognostics approach based on extreme learning machine and fuzzy clustering”. Ph.D. dessertation, University of Franche-Comt.

Knowledge-Light Adaptation Approaches in Case-Based Reasoning for Radiotherapy Treatment Planning

Gulmira Khussainova

Operational Management and Information Systems, Business School,
University of Nottingham, UK
psxgk@nottingham.ac.uk

1 Background

The aim of my research is to develop and adaptation part to currently existing CBR retrieval only system for radiotherapy treatment planning. An adaptation is often needed to change the retrieved solution to meet the needs of the new case [3], [5], [1], [2]. It often requires domain specific knowledge in a large amount, therefore the most common type of adaptation approaches found in the literature are based on acquiring adaptation knowledge from domain experts. However, expert knowledge can be expensive and is not always readily available [4]. Radiotherapy Treatment Planning: Radiotherapy treatment planning aims at delivering a sufficient radiation dose to a tumour while sparing the healthy organs in the tumour surrounding area. Radiotherapy treatment planning is often is a time consuming trial and error process which can take from 2-3 hours to a few days. Therefore, facilitating the treatment plan generation will result in a faster planning process and thus, can save the medical experts' valuable time.

The research is conducted in collaboration with the Nottingham University Hospitals Trust, NHS, Nottingham City Hospital Campus.

2 Research Question and CBR Adaptation Research Plan

CBR system: A CBR system for brain cancer radiotherapy treatment planning was developed by [5]. The current CBR system retrieves a case suggesting two main parameters: Beam Number and Beam Angle. My work builds up on the previously built CBR retrieval system [5]. Experts' knowledge was not readily available, therefore, the objective was to develop an adaptation approach by using the knowledge available within the case base. It was also hoped to gain an insight into the effect of the case attributes on the output. Machine-learning based adaptation approach was implemented using Neural Networks (NN), which considers input values en masse as a matrix, and Naive Bayesian, which considers every input individually.

Smyth and Keane [7] claimed that it is unwarranted to assume that the most similar case is also the most appropriate for the adaptation. They suggested adaptation-guided

retrieval, where retrieval mechanism retrieves a case that is easy to adapt. Based on that notion a knowledge-light adaptation method (KLAM), which is based on a modified version of the knowledge-light method described in [6] was developed. The results of the proposed methods are compared.

2.1 Knowledge-Light Adaptation Approaches

Machine Learning Based Adaptation. In the developed NN input neurons contain the attribute differences between the new case and the retrieved case while the output neuron contains the class which denote the required change to the beam number in the retrieved case. The purpose of using Naive Bayesian in our experiments was to see if there is a presence of a particular individual input attribute's affecting the output more significantly than others.

Methodology: In every new run weights are assigned to input attributes according to the input training set classifiers. Experiments were run 50 times with random cases selected for both training and test sets in every new run.

Results: The retrieval success rate was 61% and NN model has improved it to 73% which shows that NN can be useful for beam number adaptation purposes. Applying Naive Bayesian gave results lower than the success rate of the retrieval.

Knowledge-Light Adaptation Method. This method does not rely on the use of domain-specific knowledge as opposed to 'knowledge-intensive' approach. Adaptation is performed by using the knowledge contained in the case base itself. In KLAM the retrieval phase searches for both a case similar to the new case and cases appropriate for adaptation, and they are all used in solving the problem in the new case.

Methodology: The aim of the KLAM is to predict a change in the output value of the beam number of the retrieved case by using a triplet of cases (C1, C2, Cr) where (Cr) is the case retrieved for the new case (Cn) using weighted nearest neighbour similarity measure [5], and (C1-C2) - is a pair of cases which have similar attribute differences as the new and retrieved case, and is in the same region of the domain space as the (Cn) and (Cr). This means that changes to the attribute values should have similar impact on the output value. CBR system looks for a triplet where the attributes difference between pairs (C1-C2) and (Cn-Cr) is minimised and similarity between Cn and Cr cases is maximised. Cr is retrieved in conjunction with (C1-C2) in a triplet, therefore it is not necessarily the most similar case to Cn, rather it is the most similar case that that can also be adapted. The difference in beam number between C1 and C2 is used to change the beam number of Cr for the new case.

Results: KLAM has improved the retrieval results from 61% to 90%.

Conclusion: Experiments conducted have shown that by using only the knowledge available in the case base without including domain experts in the adaptation process was successful. In machine learning based adaptation NN performed well by improving the retrieval success rate up to 73%. Naive Bayesian's low success rate confirms that attributes interplay between each other in forming the output rather than individually affecting it. KLAM performed best in improving the results of the retrieval phase with a success rate of a 90%.

2.2 'Knowledge-Intensive' Adaptation-Guided Retrieval

Isocentre match clustering (Current work). Medical physicists used CBR retrieval only system to retrieve cases for the new patients. Retrieved cases were often not a good start for adaptation. Medical physicists stated that isocentre, the point where all beams intersect, ideally should coincide with tumour centre, should be also considered in the case selection as it could be seen from the suggested retrieved cases that it was not considered. Therefore, non-adaptable cases with different tumour position were retrieved for the new case.

Progress and Preliminary results: Filtering of cases according to the tumour position was incorporated into the retrieval phase. The preliminary results are encouraging: with the adapted retrieval mechanism, case retrieved has now matching isocentre and the same tumour position as the new case. The retrieved case therefore, should be a good starting point for further adaptation.

Future Work: In order to verify the validity of the adapted beam angles, suggested beams will be projected on the patient image file to determine if it crosses organs-at-risk to facilitate beam angle adaptation and adapted case validation.

3 References

1. Bergmann, R., Wilke, W.: Towards a new formal model of transformational adaptation in case-based reasoning (1998).
2. Craw, S., Wiratunga, N., Rowe, R.C.: Learning adaptation knowledge to improve case-based reasoning. Artificial Intelligence 170(1617), 1175 – 1192 (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(0), 103 – 114 (2014).
4. Hanney, K., Keane, M.: The adaptation knowledge bottleneck: How to ease it by learning from cases. In: Leake, D., Plaza, E. (eds.) Case-Based Reasoning Research and Development, Lecture Notes in Computer Science, vol. 1266, pp. 359–370. Springer Berlin Heidelberg (1997.
5. Jagannathan, R., Petrovic, S., McKenna, A., Newton, L.: A novel two phase retrieval mechanism for a clinical case based reasoning system for radiotherapy treatment planning. International Journal on Artificial Intelligence Tools 21(04), 1240017 (2012).
6. McDonnell, N., Cunningham, P.: A knowledge-light approach to regression using case-based reasoning. In: Roth-Berghofer, T., Gker, M., Gvenir, H. (eds.) Advances in Case-Based Reasoning, Lecture Notes in Computer Science, vol. 4106, pp. 91–105. Springer Berlin Heidelberg (2006).
7. Smyth, B., Keane, M.T.: Adaptation-guided retrieval: Questioning the similarity assumption in reasoning. Artificial Intelligence 102, 249–293 (1998).

Explanation-driven Product Recommendation from User-Generated Reviews

Khalil Muhammad

Insight Centre for Data Analytics
University College Dublin,
Belfield, Dublin 4, Ireland.
{firstname.lastname}@insight-centre.org

1 Introduction

The rise of user-generated content in the form of product reviews presents a unique opportunity for recommender systems. These reviews encode real user opinions and product experiences, which can be mined as the basis for novel product descriptions. The challenge is to ensure that only relevant knowledge is extracted from such reviews. This is important for case-based recommenders which typically rely on more structured product representations while borrowing notions of similarity from traditional case-based reasoning research. Research by Moghaddam et al. [1] shows how to extract useful information from user-generated content such as reviews, and Dong et al. [2] adapted the aforementioned opinion mining technique to extract product descriptions that can drive a case-based recommender.

Studies reported in [3], [4] and [5] claim that the effectiveness of a recommender system is influenced by how the system presents and explains recommendations to the user. In this context, an explanation is any additional information that help users better understand a recommendation [6]. It provides transparency into the reasoning process of recommenders that normally operate as *black boxes*. Explanations based on opinions may be more natural and convincing. In a hotel domain, for example, there may be an important feature, such as *cleanliness* that strongly dictates our decision not to choose a hotel even if the candidate hotel is highly rated for other features like *location* and *sleep quality*. So there might be some value in explanations of why an hotel is not recommended, e.g. *"This hotel is not recommended because it was negatively rated for cleanliness in many reviews"*.

Extensive research has been carried out on explanations in recommender systems: Tintarev et al. [7] define the possible goals of explanation facilities in recommender systems and show that the presentation of recommendations influences the effectiveness of explanations. Friedrich et al. [6] propose a taxonomy for categorising explanations based on major design principles: reasoning model, recommendation paradigm and the exploited information categories.

The starting point for this work is the idea that we can mine novel opinionated product descriptions from user-generated reviews as proposed in [8], [1] and

[9]. Furthermore, we will investigate the potential of these product descriptions in a variety of recommendation tasks including case-based product recommendation, summarisation of reviews and explanation of recommendation. We aim to use a case-based approach to reuse the knowledge from explanations to refine recommendations.

2 Research Plan

The core focus of this work is to continue to explore the potential of user-generated reviews as a source of knowledge for recommendation. Accordingly, we have identified the following areas of interest:

Harnessing opinionated product descriptions in recommendation. Continuing from the work of [2], we will explore the issues associated with mining opinionated product descriptions. Our aim is to develop methods for evaluating their quality, and how they can be used in different recommendation tasks.

Explanation-driven recommendation. We are concerned about how opinion information influences the generation, presentation and explanation of recommendation. Also, we plan to investigate the role of explanations in improving the internal process of recommendation.

Our expected contribution is to develop a new technique for evaluating the quality of automatically mined features from reviews, and a new approach for using explanations to support opinionated product recommendation.

2.1 Methodology

Firstly, we identify and evaluate the quality of features extracted from user-generated reviews. We define a quality feature to be one that is relevant to products as well as their users. This involves improving the opinion mining approach described in [2] to reduce the amount of noisy features it extracts.

Secondly, we investigate techniques for summarising the opinionated features of interest. These summaries will be used in generating recommendations and explanations.

Thirdly, we will develop a approach for generating explanations using the raw and summarised opinionated features. Additionally, we will employ a case-based approach to reusing knowledge from explanations to improve subsequent recommendations.

Finally, we evaluate the influence of opinionated product descriptions in recommendation, and also the effectiveness of the proposed explanation-driven recommender system.

3 Progress

In our experiments with feature quality metrics, we adapted the approach in [2] to mine opinionated features from a dataset of TripAdvisor[1] hotel reviews. We

[1] http://www.tripadvisor.ie/Hotels

use various lexical and frequency-based filters to remove noisy, less opinionated and unpopular features. The remaining features are assigned relevance scores based on their degree of relatedness to hotels and users.

We also considered opinion summarisation similar to [10]. We used simple statistical techniques to compute the positivity, interestingness and controversiality scores of features; these scores will be used in generating explanations.

Although we are currently running recommendation experiments using the aforementioned features, the next stage of this research is to focus on generating explanations with the opinionated features, and then to use the explanation structures to improve the recommendation process.

References

1. Moghaddam, S., Ester, M.: Opinion digger: An unsupervised opinion miner from unstructured product reviews. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10, New York, NY, USA, ACM (2010) 1825–1828
2. Dong, R., Schaal, M., O'Mahony, M.P., McCarthy, K., Smyth, B.: Opinionated product recommendation. In: Case-Based Reasoning Research and Development. Volume 7969 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 44–58
3. Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the users perspective: survey of the state of the art. User Modelling and User-Adapted Interaction **22** (2012) 317–355
4. Knijnenburg, B.P., Schmidt-Thieme, L., Bollen, D.G.: Workshop on user-centric evaluation of recommender systems and their interfaces. In: Proceedings of the fourth ACM conference on Recommender systems, ACM (2010) 383–384
5. Konstan, J.A., Riedl, J.: Recommender systems: from algorithms to user experience. User Modeling and User-Adapted Interaction **22** (2012) 101–123
6. Friedrich, G., Zanker, M.: A taxonomy for generating explanations in recommender systems. AI Magazine **32** (2011) 90–98
7. Tintarev, N., Masthoff, J.: Designing and evaluating explanations for recommender systems. In: Recommender Systems Handbook. Springer US (2011) 479–510
8. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artifical Intelligence. AAAI'04, AAAI Press (2004) 755–760
9. Dong, R., Schaal, M., O'Mahony, M.P., McCarthy, K., Smyth, B.: Harnessing the experience web to support user-generated product reviews. In: Case-Based Reasoning Research and Development. Volume 7466 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 62–76
10. Huang, J., Etzioni, O., Zettlemoyer, L., Clark, K., Lee, C.: Revminer: An extractive interface for navigating reviews on a smartphone. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology. UIST '12, New York, NY, USA, ACM (2012) 3–12

Research Summary
Process-oriented Case-based Reasoning

Gilbert Müller

Business Information Systems II
University of Trier
54286 Trier, Germany
muellerg@uni-trier.de,
http://www.wi2.uni-trier.de

Process-oriented Case-based Reasoning (POCBR) supports the creation and adaptation of processes [6] that are, e.g., represented as workflows. Workflows are "the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules" [4]. Thus, workflows consists of a structured set of tasks and data objects shared between those tasks. Although, POCBR is of high relevance only little research exist so far.

The presented research is part of the EVER (Extraction and Processing of Procedural Experience Knowledge in Workflows) project funded by the German Research Foundation (DFG). It deals with the extraction of procedural experience knowledge available in the Internet and further to process this knowledge by aid of CBR methods.

1 Research Questions

This section presents the research questions addressed by my doctoral thesis in note form, which focus on the development of new POCBR methods.

1. How can workflows be efficiently retrieved?
2. How can workflows be adapted regarding defined preferences or restrictions?
3. How can abstracted workflows be generated?
4. How can workflows be automatically optimized considering available resources w.r.t. parallel executions of tasks?
5. How can a new workflow be generated from scratch?
6. How can generalized workflows be generated from a case base of workflows?
7. How can similarity measures be learned that reflect the adaptability of workflows?

The approaches to address some of the presented research questions are described in the next section and section 3 describes how the remaining open research questions are going to be investigated.

2 Current state of research

The CAKE (Collaborative Agent-based Knowledge Engine) framework[1] developed at the University of Trier, offers a Process-oriented Case-based Reasoning framework, which is able to retrieve the most similar semantic workflows to a given query according to the semantic similarity measure defined by Bergman and Gil [1]. CAKE as well as semantic workflows are used to implement and evaluate the approaches addressing the research questions given above. The approaches will be illustrated and investigated in the cooking domain, which means that the workflows represent cooking recipies.

Due to the increasing number of workflows and workflow repositories, it was investigated whether and how it is possible to cluster semantic workflows guided by the semantic similarity measure [2]. This can be useful to ease the navigation or to support the analysis of the workflow repository. Recently, approaches adressing the first two research questions have currently been investigated:

1. Based on research about clustering of workflows, the problem of improving retrieval performance by developing a cluster-based retrieval method for semantic workflows [7] was addressed. To achieve this, a new clustering algorithm, which constructs a binary tree of clusters was developed. The binary tree is used as index structure during a heuristic search to identify the most similar clusters containing the most similar workflows in a top-down fashion. As it is not ensured that the most similar workflows are found, retrieval errors might occur. However, the investigation revealed that the presented approach is able to decrease the retrieval time without a considerable number of retrieval errors compared to an A* retrieval approach. Furthermore, the approach was able to compete with a domain specific MAC/FAC approach [3]. In contrast to this MAC/FAC approach the developed cluster-based retrieval approach can potentially be used with any kind of workflow representation or similarity measure.
2. Recently, a compositional adaptation approach for semantic workflows was investigated [8]. In contrast to most adaptation approaches, no adaptation knowledge is required. Instead, the available case base of workflows is analyzed and each case is decomposed into meaningful subcomponents, called *workflow streams*. During adaptation, deficiencies in the retrieved case are incrementally compensated by replacing fragments of the retrieved case by appropriate workflow streams. The investigation demonstrated the feasibility of the approach and showed that the quality of adapted workflows is very close to the quality of the original workflows contained in the case base.

3 Future Work

In future work, the workflow streams developed for the compositional adaptation approach [8] will be applied to further POCBR approaches:

[1] cakeflow.wi2.uni-trier.de

- Identification of subworkflows as workflow streams represent meaningful subcomponents
- Construction of abstract workflows by replacing each workflow stream of a workflow with an abstract task
- Optimization of workflows w.r.t. parallel execution, i.e., identify which workflow streams can be executed in parallel
- Usage of abstract workflows as a skeleton to create new workflows from scratch, e.g., by replacing abstract tasks with concrete tasks contained in workflow streams.

Moreover, it will be investigated how to generalize workflows, i.e. to generalize tasks or data objects given in a workflow, by comparing the workflow to other workflows contained in the case base using an ontology of tasks and data objects.

Additionally, a transformational adaptation approach will be developed for semantic workflows similar to the adaptation technique presented by Minor et. al. [5], which is based on adaptation cases describing how to transform a particular workflow to a target workflow.

Furthermore, methods to learn similarity measures that reflect the adaptability of semantic workflows w.r.t. the developed adaptation approaches will be investigated based on methods presented by [9].

References

1. Bergmann, R., Gil, Y.: Similarity assessment and efficient retrieval of semantic workflows. Inf. Syst. 40, 115–127 (Mar 2014)
2. Bergmann, R., Müller, G., Wittkowsky, D.: Workflow clustering using semantic similarity measures. In: Timm, Thimm (eds.) KI 2013: Advances in Artificial Intelligence. LNCS, vol. 8077, pp. 13–24. Springer (2013)
3. Bergmann, R., Stromer, A.: Mac/fac retrieval of semantic workflows. In: Boonthum-Denecke, C., Youngblood, G.M. (eds.) Proceedings of FLAIRS 2013, St. Pete Beach, Florida. AAAI Press (2013)
4. Hollingsworth, D.: Workflow management coalition glossary & terminology. `http://www.wfmc.org/docs/TC-1011_term_glossary_v3.pdf` (1999), last access on 04-04-2014
5. Minor, M., Bergmann, R., Görg, S., Walter, K.: Towards case-based adaptation of workflows. In: Case-Based Reasoning. Research and Development, pp. 421–435. Springer (2010)
6. Minor, M., Montani, S., Recio-Garca, J.A.: Process-oriented case-based reasoning. Information Systems 40(0), 103 – 105 (2014)
7. Müller, G., Bergmann, R.: A cluster-based approach to improve similarity-based retrieval for process-oriented case-based reasoning. In: Proceedings of ECAI 2014. Prague, Czech Republic (2014)
8. Müller, G., Bergmann, R.: Workflow streams: A means for compositional adaptation in process-oriented cbr. In: Proceedings of ICCBR 2014. Cork, Ireland (2014)
9. Stahl, A.: Learning of knowledge-intensive similarity measures in case-based reasoning. Ph.D. thesis, University of Kaiserslautern (2004), http://d-nb.info/972459111

Workflow extraction from textual process descriptions

Pol Schumacher

Goethe University Frankfurt - Institute for Computer Science
D-60325 Frankfurt am Main, Germany
schumacher@cs.uni-frankfurt.de

Abstract. This thesis is on workflow extraction from textual process descriptions. A lot of procedural knowledge is formulated in natural language in textual process description (e.g. cooking recipes or howtos). Workflow extraction can be used to transform those textual process descriptions into formal workflow models and thereby enable automatic reasoning on it.

1 Introduction

For my thesis I investigate the area of workflow extraction from textual process descriptions. Workflow extraction is the transformation of a process description formulated in natural language into a formal workflow model.

Recently Process oriented Case-Based Reasoning (POCBR) emerged and approaches to handle procedural knowledge were developed [1]. My work is part of a joint project of the University of Trier and the Goethe University Frankfurt.

A lot of how-to communities raised in the internet [2]. A how-to is an instruction to perform a certain task. They describe a process and contain therefore procedural knowledge. Unfortunately this knowledge is stored in natural language. Current approaches to use this knowledge e.g. for retrieval or automatic adaptation process these texts as general texts and do not take into account their procedural character.

A workflow is a set of activities which are ordered by a control-flow. A control-flow can be parallel, disjunctive, or repetitive. An activity can have a set of input- and output-products. Input products are consumed and output products are produced by a task. Workflow extraction can be divided into three phases. In the first phase we employ standard natural language processing (NLP) software to perform a linguistic analysis. In the second phase we try to identify the different elements of a workflow. At the end we try to build the control- and data-flow. The data-flow defines the flow of products or information through the workflow.

Workflow extraction faces several challenges. First, textual descriptions of processes are frequently incomplete. People often omit certain details because they can be inferred with implicit knowledge for example, anaphoras. A second problem relates to the type of content which we are processing, the user generated content. This content contains more grammatical and orthographic

errors than authored content (e.g. newspaper articles). One of the main challenges in workflow extraction is the evaluation. Due to different granularities and the paraphrasing problem, it is necessary to employ a human expert for the evaluation which makes it expensive. The paraphrasing problem is the problem that the same process can be described by different workflows. The granularity problem is the problem of handling the different levels of abstraction which can be used to formalize a process using a workflow. The automatically extracted workflows can be used in different scenarios. A workflow execution engine can be used to execute the workflows, they can be used as knowledge for reasoning and they can support the evaluation of new reasoning approaches. The first domain which is investigated is the domain of cooking as it is frequently used in artificial intelligence research, especially the CCC. A second domain which is investigated is the domain of fault isolation manuals.

For my thesis I formulated the following research questions. Can workflow extraction be used to automatically transform textual process descriptions into formal workflows? Can the resulting workflows be used for reasoning, execution or authoring support? How can workflow extraction be evaluated? What is the domain dependency?

2 Research plan

Current forms of procedural knowledge: It is necessary to get an overview of the existing forms of textual process descriptions. For example for howtos it is necessary to determine which communities exists and what the domains are. Another point is to determine under which form these descriptions are published and if they can be retrieved automatically.

Build a repository with textual description: For the development and the evaluation we need a repository of process descriptions for at least two domains.

Experiment with different NLP tools: Different types of NLP software exist. Before developing the prototype the appropriate software must be determined.

Build prototype: The prototype is used to generate workflows for evaluation. In addition it can be used to create test repositories for our partners in the project.

Develop evaluation approaches: To assess the result of the different aspects of workflow extraction, different evaluation approaches need to be developed.

Evaluate prototype: The prototype is evaluated in the cooking domain. In a second step the software is adapted to another domain and evaluated again. Are there any differences in the result and how big is the effort to adapt the software to the new domain.

3 Progress

Two prototypes with reduced functionality were evaluated. One was built using GATE as NLP tool and the other one using SUNDANCE as NLP tool. The software which used SUNDANCE performed better [3]. Several how-to communities were investigated. For two communities the how-tos were crawled and transformed to an easy to handle xml format. The first community was the cooking community allrecipes.com with about 37 000 cooking recipes. The second one was the general purpose community wikihow.com with about 140 000 how-tos. A framework for workflow extraction was developed on top of the SUNDANCE NLP tool. The framework is flexible and allows to adapt the software for a new domain. A prototype for the cooking domain was build using the framework. A research project with an industry partner which allowed us access to maintenance manuals was conducted. Maintenance manuals contain a lot of textual process description. Workflow extraction was applied and evaluated on fault isolation procedures from those manuals Three different approach for workflow extraction evaluation had been presented. The first approach evaluates the advantage of workflow extraction with manual correction compared to a complete manual workflow creation [4]. The second one can assess the quality of the automatically created data-flow [5]. The precision and recall measures were adapted to measure the quality of the extracted data-flow. In an extension of the later [6] a support for semantic distances of the data objects was added to the measures. The last one focuses on the extracted control-flow [7]. A trace-index was used to assess the quality of the control-flow.

References

1. Minor, M., Montani, S., Recio-Garca, J.A.: Editorial: Process-oriented case-based reasoning. Information Systems (2014)
2. Plaza, E.: On reusing other peoples experiences. Künstliche Intelligenz **9**(1) (2009) 18–23
3. Schumacher, P., Minor, M., Walter, K., Bergmann, R.: Extraction of procedural knowledge from the web. In: Workshop Proceedings: WWW'12, Lyon, France (2012)
4. Schumacher, P., Minor, M.: Hybrid extraction of personal workflow. In: Konferenzbeiträge der 7. Konferenz Professionelles Wissenmanagement, Passau, Germany (2013)
5. Schumacher, P., Minor, M., Schulte-Zurhausen, E.: Extracting and enriching workflows from text. In: Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration. (2013) 285–292
6. Schumacher, P., Minor, M., Schulte-Zurhausen, E.: On the use of anaphora resolution for workflow extraction. In: Integration of Reusable Systems [extended versions of the best papers which were presented at IEEE International Conference on Information Reuse and Integration and IEEE International Workshop on Formal Methods Integration, San Francisco, CA, USA, August 2013]. (2013) 151–170
7. Schumacher, P., Minor, M.: Extracting control-flow from text. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, San Francisco, CA, USA (accepted for publication)

Textual Case Based Reasoning and Semantic Similarity Comparison of Documents to Decision-making of the court judgment

Egon Sewald Junior

Department of Knowledge Engineering, Federal University of Santa Catarina
Florianópolis, Santa Catarina, Brazil

egon@egc.ufsc.br

Introduction

The duty of maintaining the rule of law applied by the judiciary board, watching the constitution and its laws, judging conflicts of interest and maintaining the social order, must be provided to meet the citizens and businesses needs properly. One of the main points to be considered is the fight against the slowness of the judiciary; in other words, achieve reasonable quickness in the processes.

The electronic process reduced the judicial proceeding time, allowing to diminish “dead time”, for example, assembly the process files, page numbering, and mainly with the physical transit of procedural parts. Notice that, to a greater reduction in length of these pending actions, it is necessary a correct use of resources of the court and increase the performance in decision-making judgment, intensive knowledge activities.

This work starts from researching the organizational context, identifying the intensive knowledge activities and define the Knowledge System pattern to support the decision-making process within the state courts. It has brought up engineering applications expertise applied to the judiciary and from this literature review, to generate a pattern to the knowledge system model.

For the organizational system pattern "Court of the State of Amazonas" (TJ/AM) the CommomKADS methodology [1] was applied to raise the organizational context as well as the concepts to be designed to solve problems and exploit opportunities, as well as the design definitions of the artifact, in other words, the Knowledge System.

From the application of this methodology, it was established organization models, where it can define problems and opportunities, which may or may not be related to knowledge intensive activities, so identifying applied knowledge assets. The task model describes the processes of the TJ/AM, identifying the use of knowledge and a

review about its correct application. The instrument model provides a study about the human agents or software. Based on these models, the context is presented and it is defined the concept of the knowledge system, by the use of knowledge systems models, which describes its application during the tasks, as well as the communication models that define the interaction between the agents. Thenceforth, it is defined the software design, outlining an artifact that I should develop to support the decision of judges.

Work Definition

Research Activity: sentencing (court judgment)

Domain Context: State Courts of Justice, first instance, natural language processing, representation cases, case law, information retrieval, template mining, evaluation of similarity, text retrieval, decision-making process of the court

Scientific contribution: A model to support the sentencing of the Judge

Support: This study should present one quantitative evaluation of effectiveness of decisions and, a qualitative evaluation considering the perception of judges that will answer the survey.

Research Plan and Progress

Needs assessment of the judiciary: The activity sentencing was observed. We realize that judges are convinced based Walk-in law enforcement in similar cases, adapting similar decisions [2].

Thus, we observe that knowledge of court proceedings can be formalized in the form of cases and apply CBR [3]. The comparison uses contextual or semantic approach (that are not structured data are compared). We observe that for each type of classes and procedural matters, the judge makes a series of comparisons between cases. This phase of the work is in process of completing.

Case extraction: The extraction of the texts of cases should be performed using an engine of knowledge discovery based on ontologies[4][5][6][7], which must respect the context and analyze the use of the phrases second context, in order to reduce the ambiguity problem of under Brazilian law and enable automatic extraction of cases based on documents generated in the courts and decisions already handed down.

CBR System: The comparison of cases should point out the similarity between them, allowing comparison on the basis of trial, stating and supporting the judge's decision

[8]. The comparison must take into account cases of the same subject and procedural class as well as adjust the items and weights for comparison.

End-user evaluation: quantitative evaluation of effectiveness of decisions and qualitative evaluation considering the perception of judges that will answer the survey.

References

1. Sewald Junior, E., Rotta, M., Vieira P., Rover, A., Sell, D.: Modelagem de Sistema baseado em Conhecimento em um Tribunal de Justiça utilizando CommonKADS. In: Revista Democracia Digital. 7Ed. Florianópolis, (2012) 160-189

2. Schreiber, G.; Akkermans, H.; Anjewierden, A.; Hoog, R.; Shadbolt, N.; De Velde, W. V.; And Wielinga, B.: Knowledge Engnineering and Management: the CommonKADS Methodology. MIT Press. Cambridge. Massachussets. 2002.

3. Ruschel, A. J.: Modelo de conhecimento para apoio ao juiz na fase processual trabalhista. Tese (Doutorado) Curso de Pós-Graduação em Engenharia e Gestão de Conhecimento. Universidade Federal de Santa Catarina - UFSC, Florianópolis (2012)

4. Weber, R., Ashley, K. D, Brüninghaus S.: Textual case-based reasoning. In: Knowledge Engineering Review. Cambridge, New York, Cambridge University Press (2005) 255-260

5. Smyth, B., Keane, M.T.: Using adaptation knowledge to retrieve and adapt design cases. Knowledge-Based Systems 9(2) (1996) 127–135

6. Bergmann, R., Gil, Y.: Retrieval of semantic workfows with knowledge intensive similarity measures. In: Case-Based Reasoning. Research and Development, 19th International Conference on Case-Based Reasoning, ICCBR 2011, Springer (2011) 17–31

7. Roth-Berghofer, T. R., Cassens, J.: Mapping Goals and Kinds of Explanations to the Knowledge Containers of Case-Based Reasoning Systems. Héctor Muñoz-Avila and Francesco Ricci, editors, Case Based Reasoning Research and Development – ICCBR 2005, volume 3630 of LNAI, Chicago, Springer. (2005) 451– 464

8. Weber, R: Pesquisa jurisprudencial inteligente. Tese (Doutorado) Curso de Pós-Graduação em Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina - UFSC, Florianópolis (1998)

9. Cassens, J., Kofod-Petersen, A.: Explanations and Case-Based Reasoning in Ambient Intelligent Systems. David C. Wilson and Deepak Khemani, editors, ICCBR-07 Workshop Proceedings, Belfast, Northern Ireland, (2007) 167–176

A multi-layer hybrid CBR/RL approach to micromanagement in RTS games

Stefan Wender

Department of Computer Science, The University of Auckland, New Zealand
s.wender@cs.auckland.ac.nz

1 Problem Outline

The research problem addressed in my dissertation is the creation of a hybrid case-based reasoning (CBR)/reinforcement learning (RL) technique that uses a hierarchical approach to address multiple levels of the problem of combat simulation in a commercial real-time strategy (RTS) game. RTS games offer a polished environment that includes numerous properties that are interesting for AI research, such as imperfect information, spatial and temporal reasoning as well as learning and opponent modeling [1]. However, these characteristics also make RTS game environments very complex. Even sub-problems, such as the control of units in combat situations, can not be completely solved by brute force algorithms. For these reasons, I chose StarCraft as a test bed for a machine learning (ML) approach that tries to learn how to manage combat units on a tactical level ("micromanagement"). Micromanagement requires a large number of actions over a short amount of time. It requires very exact and prompt reactions to changes in the game environment. The problem involves concepts like damage avoidance, target selection and, on a higher, more tactical level, squad-level actions and unit formations [2].

The inherent complexity in RTS games means, that a single, holistic approach cannot solve the entire problem. However, RTS games can be split into logical tasks that fall into distinct categories such as strategy, economy, tactics or reactive maneuvers (see Figure 1). These tasks can in turn be grouped into layers [2]. For the creation of an AI that plays a game, these tasks and layers are used to subdivide the overall problem into areas of responsibility for certain parts of a bot architecture. This layering in RTS games leads to most RTS agents being inherently hierarchical.

2 Progress to Date

As a first step to create a hybrid RL/CBR agent an evaluation of the suitability of simple reinforcement learning algorithms to perform the task of micro-managing combat units in StarCraft was done [3]. The techniques applied are variations of the common Q-learning and Sarsa algorithms [4]. The aim was the design of an agent that addresses the lowest level of combat in the game, involving

only one unit that is controlled by the RL agent in a very simplified state- and action-space.

Subsequently, I extended the simple RL agent to integrate CBR for memory management, furthermore using a relational database management system (RDBMS) to store the large amounts of data that the problem produces [5]. The agent devised for [5] uses a CBR-based memory which utilizes RL to learn the fitness of its case solutions. The model of the game world is based on two different case-bases for different levels of abstraction of the current game state. RL is used to update the value of unit actions. Those unit actions represent the case solutions. Using this hybrid approach, the initial agent was extended from controlling a single unit to controlling multiple units in different scenarios. This approach produced promising results and insights into the development of the case-bases using the model I designed.

Based on this CBR/RL integration, I then created a hierarchical architecture that uses a number of case-bases (see Figure 2) to address the entire micromanagement components as described in Section 1. This hierarchical CBR (HCBR) approach is inspired by the layered learning method [6] and also by [7], which describes the creation of an HCBR system for software design. However, I use a more complex approach with case-bases that are not strictly partonomic, since lower-level case-bases use other inputs besides solutions from higher level states as their case descriptions.

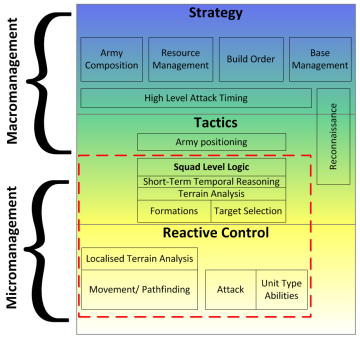

Fig. 1: RTS Game Layers and Tasks

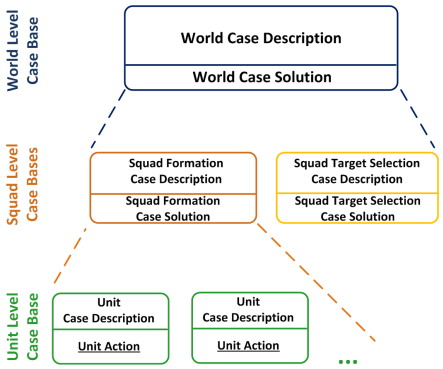

Fig. 2: Hierarchical Case-Base Structure

This architecture covers the micromanagement component in Figure 1, more specifically, all tasks enclosed by the dashed red square. Its aim is, while not addressing the entire problem through a single approach, to closely integrate several levels of reasoning into one approach that addresses the entire micromanagement problem. The architecture has been implemented in the StarCraft RTS game. The pathfinding component on the lowest level has been implemented using a combination of CBR and RL and shows good performance and extendability for bigger tasks [8].

3 Planned Research

Basic components for the top levels of the multi-layer architecture and a more complex component for the bottom level navigation have been created. I am currently in the last stages of further optimizing the navigation by using a variation of locality-sensitive hashing (LSH) [9] to enable more complex navigation scenarios through optimized case retrieval. I am also planning to use LSH for case-bases at higher levels. Should there be enough time, I would like to compare LSH for case retrieval against other common approaches such as kd-trees [10]. While higher-level components and case-bases have been created, their implementation and performance has not yet been sufficiently evaluated. The transfer of knowledge between layers, an integral part of the overall architecture, is also yet to be finished.

The eventual aim is to create an AI agent that is able to handle any possible combat situation in StarCraft successfully, independent of the involved numbers and types of units. To properly evaluate this and prove its viability, in a final step my layered hybrid RL/CBR approach will be evaluated against existing StarCraft agents that have been used in competitions co-located with research conferences [11].

References

1. Buro, M., Furtak, T.: Rts games and real-time ai research. In: Proceedings of the Behavior Representation in Modeling and Simulation Conference (BRIMS). (2004)
2. Weber, B., Mateas, M., Jhala, A.: Building human-level ai for real-time strategy games. In: 2011 AAAI Fall Symposium Series. (2011)
3. Wender, S., Watson, I.: Applying reinforcement learning to small scale combat in the real-time strategy game starcraft:broodwar. In: IEEE Conference on Computational Intelligence and Games 2012. (2012)
4. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press (1998)
5. Wender, S., Watson, I.: Integrating case-based reasoning with reinforcement learning for real-time strategy game micromanagement. In: Accepted for Presentation at the 13th Pacific Rim International Conference on Artificial Intelligence. (2014)
6. Stone, P.: Layered Learning in Multiagent Systems: A Winning Approach to Robotic Soccer. MIT Press (1998)
7. Smyth, B., Cunningham, P.: Déjà vu: A hierarchical case-based reasoning system for software design. In: ECAI. Volume 92. (1992) 587–589
8. Wender, S., Watson, I.: Combining case-based reasoning and reinforcement learning for unit navigation in real-time strategy game ai. In: Accepted for Presentation at the 22nd International Conference on Case-Based Reasoning (ICCBR). (2014)
9. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM (1998) 604–613
10. Wess, S., Althoff, K., Derwand, G.: Using k-d trees to improve the retrieval step in case-based reasoning. Topics in Case-Based Reasoning (1994) 167–181
11. Ontanón, S., Synnaeve, G., Uriarte, A., Richoux, F., Churchill, D., Preuss, M.: A survey of real-time strategy game ai research and competition in starcraft. (2013)