Towards the application of case-based reasoning to a system for exploring cultural heritage corpus

Prunelle D. Treuil[1,*,†]

[1]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

Abstract

Harold is a conversational system developed in order to assist the historians working at Henri Poincaré's Archive to explore the mathematician's correspondence. This correspondence is currently stored in a knowledge base using a SPARQL endpoint, which is difficult to use for anyone unfamiliar with SPARQL. Using Harold, the historians are able to delimit any set of letters useful to answer the research question they are currently working on through a serie of interactions with the system. The results of every search are presented in a hierarchy of concepts using formal concept analysis. This allows the user to quickly select the letters or properties that interest them or, conversely, those that they want to remove from the results. Additionally, these concepts can also be the basis for constructing an ontology containing both a hierarchy of properties and a hierarchy of concepts. This ontology allows for the restructuring of the results to show more generalized concepts. A future work to be done on the system is to guide the user through all possible interactions with Harold using CBR. This could be done in several different ways: (1) case-based reasoning on traces could be applied to the serie of user interactions, (2) Harold being already a conversational system, conversational case-based reasoning principles could also be used.

Keywords

Ontology management, conversational system, formal concept analysis, cultural heritage collection, case-based reasoning on traces, conversational case-based reasoning

1. Introduction

The present paper presents work done to implement Harold, a system used to access and analyze a textual corpus represented by a knowledge graph and structured by an ontology, and to show how it could be enhanced thanks to the CBR methodology. Harold currently allows for the exploration of the correspondence of Henri Poincaré, a famous French mathematician (1854-1912). He excelled in most scientific fields of his time: mathematics, philosophy of science, physics, chemistry, celestial mechanics, etc. and was in contact with numerous scientific circles, making his correspondence particularly insightful to study. This correspondence is accessible both on a website and via a SPARQL endpoint. This endpoint gives direct access to the RDF triples structuring the corpus, the triples subject-property-object (or value) used to describe the knowledge graph. Using the SPARQL endpoint is a difficult task for the historians working on the letters, as they neither know how to write SPARQL queries nor the vocabularies used for both the knowledge graph and the ontologies. Harold allows non SPARQL specialists, such as Henri Poincaré Archives historians, to explore this corpus and to exploit the knowledge on each letter in the knowledge graph used to structure the corpus to improve their understanding of it. With this system, users can create their own domain-focused ontology and use it to fill the graph with new concepts and properties to better describe each letter.

Currently, Harold does not use case-based reasoning, but several ways to incorporate it are considered. (1) Since Harold is built to foster an iterative search process, the history of all user interactions could be exploited thanks to a case-based reasoning system to guide the user in their next step. The idea would be that Harold could give them suggestions such as "add the letters with these properties", "remove all the letters from this person", "regroup these two concepts into a more general concept", etc. (2) As Harold is already designed as a conversational system, an approach using conversational case-based

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France

*Corresponding author.

prunelle.daudre-treuil@univ-lorraine.fr (P. D. Treuil)

reasoning [1] principles could be imagined. Harold's users do not often have a clear idea of how to solve a given research question and need several steps to refine their initial query.

2. Research Plan

2.1. Research Objectives

Harold's goal is to accompany a historian during their exploration of a corpus, namely Henri Poincaré's correspondence, from their research question to the writing of a scientific publication. For example, a historian could have an interest in the Dreyfus affair, a French political scandal that straddles the 19^{th} and 20^{th} centuries. After using Harold, they could explain the involvement of Henri Poincaré in the trial that followed based on the letters that he exchanged with the main actors in this affair [2]. If the beginning of this thesis was focused on the implementation of Harold, the next step of this work should revolve around the integration of case-based reasoning into the system to improve its usefulness for historians of the Henri Poincaré Archives.

Harold provides multimodal access to the corpus, allowing the user to select a precise set of letters depending on the properties they should or should not have: metadata like the sender or recipient, the persons quoted, the terms used both in the text and in the annotations, or other data such as formulas, graphs, and drawings present in the original documents.

With the set of selected letters, Harold highlights new pieces of information on the corpus, shows patterns, and potential interesting concepts that the historian would not necessarily have seen by themselves. This analysis can be the basis for the construction by the user of an ontology, including both a hierarchy of properties and a hierarchy of concepts. This ontology then impacts the precision and usefulness of the results of every search done on the corpus in a positive feedback loop until the user improves their understanding of their data enough to answer their initial research question.

2.2. Approach / Methodology

The choice has been made for Harold to be a conversational system in the sense that it allows iterative interactions between the user and the system to solve a given problem. Indeed, working on a quite complex corpus and with often very short documents, most letters require the expertise of Harold's user in order to be correctly interpreted. For example, if someone is trying to find all the letters related to Henri Poincaré's work on physics, most of them will not contain the term "physics". The user will therefore need to search for additional letters by a new query or remove some letters or properties from the results, using their own knowledge. In the case of letters related to physics, a specialist could add letters exchanged with physicists and physics institutions that they already know to have a contact with the mathematician.

To do so, Harold's user interface is composed of several parts. First, a form containing different fields, for each property describing a document, allows the user to either start a new search from the ground up or to add new documents to the current search. The properties used are the metadata describing the letters: sender, recipient, writing place, language, and writing date, as well as some textual data: the persons and groups of persons quoted in the text and some candidate terms extracted from the letters and their annotations produced by the historian, called the critical apparatus. These candidate terms could be both automatically extracted or manually selected by historians. A second part of the interface shows in a synthetic way all the properties and letters selected by the user and all of them that should not be kept. From this history, SPARQL queries are built, and their results are displayed as a hierarchy of concepts obtained using formal concept analysis. In this context, a concept is understood as a pair (set of letters, set of properties) such that all letters possess all the properties and all the properties are possessed by all the letters. This allows the user to quickly discover the main properties shared by the letters they have selected.

A specific part of the user interface allows ontology management. **As seen in Figure 1,** it contains both a hierarchy of properties and a hierarchy of concepts. These properties and concepts can be

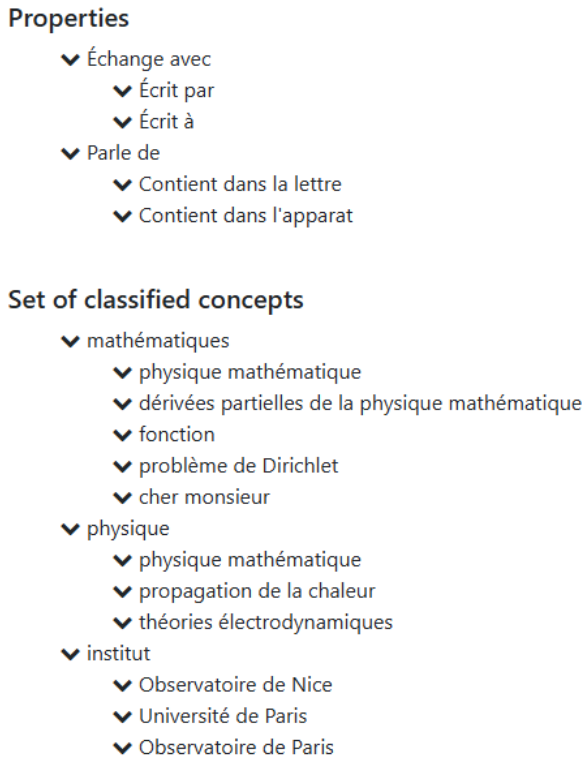
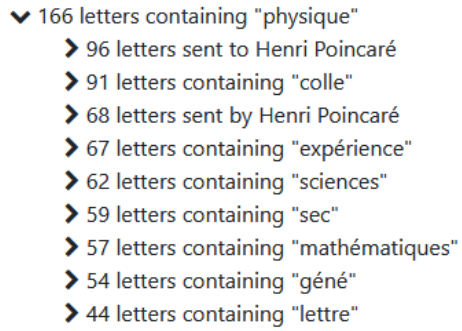
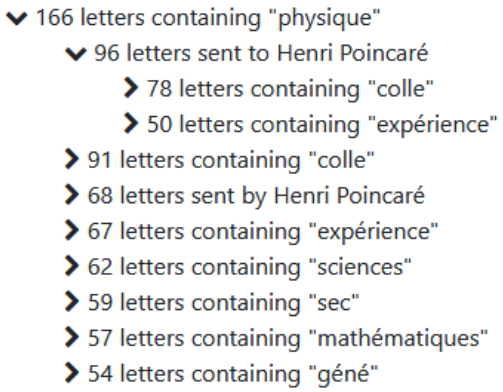


Figure 1: An example of hierarchies built by the user.



(a) Hierarchical presentation of the results after starting a new search on the letters containing the term "physique".



(b) The results after having taken into account the concept hierarchy of Figure 1.

Figure 2: An exemple of interaction with Harold.

retrieved from the results section or be created manually. Once the hierarchies are organized properly, they can be used to generalize the properties of every letter. For example, if a letter contains in its text the word "problème de Dirichlet" (in english "Dirichlet problem"), then it should be generalized that one of the letters' topics is related to mathematics. Using this generation before triggering the formal concept analysis allows Harold to find more general concepts. Even if the letters do not contain the words "mathematics" or "physics", the user can iteratively create both concepts to find all the letters related to both of these topics.Figure 2 shows the impact of the hierarchies on the letters containing the term "physique". Figure 2a presents the result of this search without taking the hierarchies into consideration, and in Figure 2b, the results that take them into account can be seen? Finally, a last part of the interface allows the user to visualize in more detail the letters found by their search with their data and a link to the Henri Poincaré Archives' website with the scan of the letter, its transcription and

possible additional images. This allows the user to study each individual letter in order to answer their initial research question.

The system should be able to guide the user in the different possible interactions. Indeed, Harold offers many different possible interactions and it is not always apparent which one would produce the best results. As such, the system should propose some possible next steps for the user in order to bring them closer to their research question. Different approaches are possible to do so, similarity search, using LLMs to propose new additions to the ontology, and so on. An interesting possibility would also be to use case-based reasoning to find next steps potentially useful to the user. The history of all the interactions performed by Harold's users in each of their sessions could give some insight as to which step the user could want to carry next. As this system has not yet been developed, no particular CBR technique has been selected to implement it. Some possibilities could be explored: (1) First, since Harold is already a conversational system, some principles from works on conversational CBR [3] could be reused to take advantage of the already existing interactivity of Harold. (2) Another possibility would be to apply case-based reasoning on traces approaches [4] on the list of successive interactions a user had with Harold to introduce them to potential new steps. This can also fall under the scope of process-oriented CBR [5]. (3) With the development of Natural Language Processing (NLP) techniques, new methods using CBR principles for question answering on knowledge base can be imagined [6, 7]. (4) Finally, textual CBR [8] could be used to improve text mining on the letters in order to better retrieve their underlying themes.

3. Progress Summary

So far, Harold allows users to explore iteratively and interactively Henri Poincaré's correspondence. The ontology management interface has been implemented, and the ontology can be used to create more general concepts shown in the results. Harold is accessible online for testing on Henri Poincaré corpus at https://harold.ahp-numerique.fr/.

Even if Harold could be applied to other corpus, it is primarily created for the Henri Poincaré correspondence, and as such the historians of the Henri Poincaré Archive are present throughout its development, guiding its design and functionalities and testing it thoroughly. Several evaluation sessions have been done and will be done with them. Their expertise ant time will be useful for any implementation of a CBR system in Harold.

4. Conclusion and Future Work

As Harold is now online, the historians working on the Henri Poincaré corpus have started to test the system. Harold will also be used on several other corpora, including (a) the corpus of the correspondence of another scientist and (b) a set of scientific papers on a given domain with the purpose of helping to build a state of the art if this domain.

To expand the multimodal access to the corpus provided by Harold, and since the corpus used with it is mainly scientific, an interesting next step would be to use the formulas presented in the documents. The formulas contained in the Poincaré corpus are currently encoded in LaTeX. The idea would be to automatically link a formula with some concept of the ontology. For example, a ∇ or ∂ would indicate a differential equation, $\int$, an integral, etc. This would allow us to extract more information from the letters.

Finally, the question of how to guide the user through all the different interaction possibilities in the interface is still open. The idea is that the historian has a research question to solve, and the system should suggest an efficient next step to complete this goal. As mentioned above, at least two CBR approaches can be considered to solve this: applying trace-based CBR on the user interaction history in order to propose relevant new interactions to the historian or developing the conversational approaches of Harold to both select the letters needed to answer the user's research question and the ontology used to structure them. Other CBR approaches could probably also be used to guide the user in their

interactions with Harold such as textual CBR, process-oriented CBR or CBR for question answering on knowledge base.

Acknowledgments

The author thanks the reviewers for their insightful suggestions and the papers they recommended, which have introduced us to new perspectives.

Declaration on Generative AI

During the preparation of this work, the author used Writefull for the purpose of: Grammar and spelling check, Paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. M. Richter, R. O. Weber, M. M. Richter, R. O. Weber, Conversational CBR, Case-Based Reasoning: A Textbook (2013) 465–485.

[2] L. Rollet, Autour de l'affaire Dreyfus. Henri Poincaré et l'action politique, Revue Historique 298 (1997) 49–101.

[3] D. W. Aha, L. A. Breslow, H. Muñoz-Avila, Conversational Case-Based Reasoning, Applied Intelligence 14 (2001) 9–32.

[4] A. Cordier, M. Lefevre, P.-A. Champin, O. Georgeon, A. Mille, Trace-Based Reasoning — Modeling Interaction Traces for Reasoning on Experiences, Twenty-Sixth International Florida Artificial Intelligence Research Society Conference (2013).

[5] M. Minor, S. Montani, J. A. Recio-García, Process-oriented case-based reasoning, Information Systems 40 (2014) 103–105.

[6] R. Das, A. Godbole, A. Naik, E. Tower, M. Zaheer, H. Hajishirzi, R. Jia, A. Mccallum, Knowledge Base Question Answering by Case-based Reasoning over Subgraphs, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 4777–4793.

[7] J. Li, X. Luo, G. Lu, GS-CBR-KBQA: Graph-structured case-based reasoning for knowledge base question answering, Expert Systems with Applications 257 (2024) 125090.

[8] R. O. Weber, K. D. Ashley, S. Brüninghaus, Textual case-based reasoning, The Knowledge Engineering Review 20 (2005) 255–260.

[9] A. Cordier, M. Lefevre, P.-A. Champin, A. Mille, Modéliser les traces d'interaction pour raisonner partir de l'expérience tracée ?, in: IC - 24èmes Journées francophones d'Ingénierie des Connaissances, 2013.