Circularity in Case-Based Reasoning

Adwait P. Parsodkar[1,2,*,†]

[1]Indian Institute of Technology, Madras, Chennai, 600036, India

Abstract

We attempt to bring to the forefront the often-overlooked *circularity* in a wide variety of tasks linked to Case-Based Reasoning (CBR). This motivates the need for a methodical examination of circularity to arrive at new paradigms in CBR. In our previous work on *RelCBR*, we have illustrated that the current CBR paradigm can benefit from the consideration of circularity. We propose several other directions of research wherein the state-of-the-art approaches are still agnostic to circularity and argue why the incorporation of circularity can yield improvement in performance. Approaches proposed for resolving the diverse variety of circularities can serve as templates for circularities of similar nature that emerges in different contexts.

Keywords

Case-Based Reasoning, Circularity, Truth Discovery

1. Introduction

Case-Based Reasoning (CBR) is an inherently cyclical problem-solving paradigm wherein learning and reasoning go hand in hand. The effectiveness of the reasoning mechanism heavily relies on the knowledge contained, however, the knowledge itself is obtained through the process of reasoning. The interdependence between problem-solving and learning in CBR is just one example of *circularities* that occur in various research sub-themes of CBR. We use the term *circularity* to indicate the existence of a (direct or indirect) dependence of a property P_1 on another property P_2 when P_2 in itself requires the knowledge of P_1 to be determined. In our work, we aim to identify instances of circularities that manifest in diverse contexts in the CBR literature, followed by proposing circularity resolution approaches. These approaches can potentially be reused to resolve the circularities in apparently unrelated problems.

Several works in the Machine Learning literature have identified circularities in a variety of tasks and have proposed approaches to resolve them. For instance, an iterative solution has been employed to arrive at the *PageRank* scores [1] of web pages in a network of hyperlinked web pages. The circularity emerges from the fact that the determination of the score of a page depends on that of pages that hyperlink to it, and is given by the definition - *'A page is important if it is pointed to by important pages.'* Related algorithms such as Hyperlink Induced Topic Search (HITS) [2] and SimRank [3] are also based on circular definitions.

ICCBR DC'23: Doctoral Consortium at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.

cs20d404@cse.iitm.ac.in (A. P. Parsodkar)

https://sites.google.com/smail.iitm.ac.in/pmrf2502402/home (A. P. Parsodkar)

0000-0002-9408-0843 (A. P. Parsodkar)

The well-known k-Means algorithm also deals with an inherent circular problem: *'In order to identify the cluster centroids, the instances have to be assigned to clusters. But such an assignment is not possible unless the cluster centroids are known.'* The involved circularity is resolved using the Expectation Maximization (EM) algorithm [4].

In the context of CBR, there have been past works that have benefited from an appreciation of circularity. The document segmentation approach for the construction of cases by segmenting raw text documents, as presented in [5], uses the EM algorithm to resolve the proposed circularity. An approach inspired by the PageRank algorithm has been employed to assign *Retention Score* [6] in the context of case base maintenance. Product preference information based on interactions with users has been leveraged to enrich case descriptions [7] in case base recommender systems by means of the PageRank algorithm.

Our recent work on quantifying the reliability of cases in the case base [8] is based on a circular theme. It takes inspiration from the Truth Discovery [9] literature that is fundamentally concerned with a circular problem while aggregating knowledge from an array of knowledge sources. The following statement of circularity constitutes the foundation of the Truth Discovery literature.

> *A knowledge source is considered reliable if it provides trustworthy solutions to problems, and a solution is deemed to be trustworthy if it is supported by reliable knowledge sources.*

Our experiments have shown improvements due to the consideration of circularity. We further posit that a systematic study of circularity can impact research in the following ways:

1. Identifying tasks that involve inherent circularity but have been traditionally addressed via approaches that are agnostic to such circularity.
2. Proposing novel approaches to resolve the involved circularities and identifying situations in which these can be extended to circularities that manifest in other contexts.
3. A paradigm level shift might be necessary that inherently respects circularity, as a result of which the various manifestations of circularity might disappear. The Holographic paradigm [10] holds promise in this regard.

2. Research Plan

In this section, we describe our research objectives and a rough outline of our roadmap towards realizing our goals.

2.1. Research Objectives

Several approaches in the CBR literature have not paid explicit attention to the underlying circularity in the tasks they attempt to address. Our objective is to highlight the importance of circularity by identifying its underlying nature in CBR tasks such as retrieval, adaptation, and maintenance in the traditional CBR. We further hypothesize the need for a novel paradigm in CBR that may inherently handle such instances of circularity at a deeper level.

Finally, the techniques proposed for the resolution of circularity hold the potential to serve as a template for circularity resolution in situations that arise in contexts that appear to be distinct on the surface but share similar inherent structures. For instance, the circularity resolution technique employed in the PageRank algorithm can be used for arriving at preferability scores of products in recommender systems domain, as illustrated in [7]. Another instance is the task involving document segmentation [5] that resolves the proposed circularity using the EM algorithm, much like that used to address the circularity in the $k-$Means clustering.

2.2. Approach / Methodology

In our recent submission, we have highlighted the importance of circularity in a variety of tasks relating to CBR in an unsupervised setting, some of which are listed below.

2.2.1. Retrieval and Adaptation

When a reasoner is equipped with a multitude of similarity functions (possibly learned bottom-up; see [11] for one such approach), it might be beneficial to assign different degrees of reliability to these similarity functions (can be seen as *rankers* that rank cases in order of relevance to the queries). We propose the following statement of circularity towards realizing this:

> *A ranker is reliable if it produces trustworthy rankings for several queries. A ranking is trustworthy if it is supported by reliable rankers.*

A statement of circularity, with nature very similar to the one above, can be proposed when a reasoner learns adaptation rules in a data-driven fashion (say using the case difference heuristic [12], for instance).

2.2.2. Maintenance

One aspect of the maintenance of a case-based reasoner is the removal of *noisy* and *redundant* cases from the case base. Literature for the detection of noisy cases is often based on a non-circular definition, the limitations of which are discussed in Section 3. Our approach for the identification of noisy cases (via estimating their *reliability*) is grounded upon the following circular definition, which has been shown to outperform its non-circular counterparts.

> *A case is reliable if it can be solved by its reliable neighbors.*

The Footprint Algorithm [13] is typically employed for the detection and removal of redundant cases by means of estimating the competence of cases in the case base. The competence of a case is often quantified using its *Relative Coverage* [13] that assigns a high score to a case if the case solves cases that are not solved by many *solver* cases. Such a measure, however, is agnostic to the competence values of the solver cases. In other words, regardless of whether the solver cases are retained or discarded by the footprint algorithm, their contributions to the Relative Coverage of the case remain unaffected.

A possible circular extension for arriving at the competence of cases is presented. Notice that it makes use of the heuristic that a case with a higher competence score is likely to be retained.

> *A case is likely to be retained if the cases it solves are not already solved by cases that are likely to be retained.*

We would like to emphasize the fact that the identified circular dependencies disappear in the presence of knowledge from a domain expert. In particular, if top-down knowledge is available to the reasoner concerning the reliability of similarity functions (or adaptation strategies) in an ensemble, or the competence of cases in the case base, the proposed circularities cease to exist.

2.2.3. Experimental Methodology

We aim to demonstrate the effectiveness of methods that address circularity over those that are agnostic of it. For instance, the problem-solving ability of a reduced case base constructed using a non-circular definition to eliminate cases can be compared against that when a method that takes the underlying circularity into account is employed. In settings involving multiple similarity functions (or adaptation strategies), the closeness of the predicted solution to the ground truth can be quantified in order to compare approaches that make use of circularity with those that do not.

We also intend to investigate whether the proposed solution techniques can be reused to resolve circularities of similar nature that manifest in a different context.

3. Progress Summary

In our work submitted to ICCBR 2023, we demonstrate that diverse CBR research contexts share an underlying circularity. We examine the root cause of such circularities and present fundamental impossibility results in this context. We show how a systematic study of circularity can motivate the quest for novel CBR paradigms and lead to novel approaches that address circularities in traditional CBR retrieval, adaptation, and maintenance tasks. Furthermore, such an analysis can help in extending the solution of one problem to solve an apparently unrelated problem, once we discover the commonality they share deep down in terms of the circularities they address.

Our prior work, *RelCBR* [8], serves as a motivating scenario that illustrates the importance of realizing circularity over approaches that disregard it. The task is aimed towards the identification of cases that provide *trustworthy* solutions when a query in close proximity is presented to the reasoner. This tendency of cases providing trustworthy solutions is characterized by their *reliability score*. Although a domain expert can potentially inspect the entire case base in order to assign a reliability score to each case, the impracticality particularly associated with large-scale systems is apparent. To circumvent the potential absence of such top-down assistance, several bottom-up strategies have been proposed in the literature for the identification of unreliable cases. These approaches build upon the non-circular definition - *'A case is reliable if it can be solved by its neighbors.'* Notice that this definition is agnostic to the reliability of the neighbors leading to two classes of problematic situations:

1. A truly reliable case situated in a neighborhood of predominantly unreliable cases is deemed unreliable due to its disagreement with its neighbors.

2. A truly unreliable case in a predominantly unreliable neighborhood can be considered reliable by virtue of the agreement with its neighbors.

We have proposed RelCBR, which takes the reliability of the neighbors of cases into account. The underlying circular definition states that

A case is reliable if it can be solved by its reliable neighbors.

Obtaining reliability based on such a circular definition serves two key benefits. First, cases with reliability under a predefined threshold can be presented to a domain expert for inspection. Further, these reliability scores can be used to undermine the contribution of unreliable cases when addressing a new query. Our experiments have shown statistically significant improvements over the baseline methods [14, 15] in regression and classification tasks.

4. Conclusion and Future Work

Our work is centered around the claim that diverse research problems in CBR such as retrieval, adaptation, and maintenance can benefit from an appreciation of the underlying circularity. RelCBR is a case in point. In the long run, a systematic study of circularity can motivate the exploration of novel CBR paradigms. Finally, we foresee wider implications of circularity in the context of Artificial Intelligence.

References

[1] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford Digital Library Technologies Project, 1998.

[2] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (1999) 604–632. doi:10.1145/324133.324140.

[3] G. Jeh, J. Widom, Simrank: a measure of structural-context similarity, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 538–543. doi:10.1145/775047.775126.

[4] X.-L. Meng, D. Van Dyk, The em algorithm—an old folk-song sung to a fast new tune, Journal of the Royal Statistical Society Series B: Statistical Methodology 59 (1997) 511–567.

[5] D. P., K. Visweswariah, N. Wiratunga, S. Sani, Two-part segmentation of text documents (2012). doi:10.1145/2396761.2396862.

[6] D. Mathew, S. Chakraborti, An optimal footprint method for case-base maintenance, in: The Thirty-First International Flairs Conference, 2018.

[7] S. R. Vasudevan, S. Chakraborti, Enriching case descriptions using trails in conversational recommenders, in: Case-Based Reasoning Research and Development: 22nd International Conference, ICCBR 2014, Cork, Ireland, September 29, 2014-October 1, 2014. Proceedings 22, Springer, Springer International Publishing, 2014, pp. 480–494.

[8] A. P. Parsodkar, D. P., S. Chakraborti, Never judge a case by its (unreliable) neighbors: Estimating case reliability for cbr, in: M. T. Keane, N. Wiratunga (Eds.), Case-Based Reasoning Research and Development, Springer International Publishing, Cham, 2022, pp. 256–270.

[9] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, ACM Sigkdd Explorations Newsletter 17 (2016) 1–16.

[10] D. Ganesan, S. Chakraborti, Holographic case-based reasoning, in: I. Watson, R. Weber (Eds.), Case-Based Reasoning Research and Development, Springer International Publishing, Cham, 2020, pp. 144–159.

[11] A. Jaiswal, K. Bach, A data-driven approach for determining weights in global similarity functions, in: Case-Based Reasoning Research and Development, Springer International Publishing, 2019, pp. 125–139. doi:10.1007/978-3-030-29249-2_9.

[12] K. Hanney, M. T. Keane, Learning adaptation rules from a case-base, in: Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1996, pp. 179–192. doi:10.1007/bfb0020610.

[13] B. Smyt, E. McKenna, Footprint-based retrieval, in: Case-Based Reasoning Research and Development, Springer Berlin Heidelberg, 1999, pp. 343–357. doi:10.1007/3-540-48508-2_25.

[14] L. Lamontagne, Textual cbr authoring using case cohesion, in: Proceedings of 3rd Textual Case-Based Reasoning Workshop at the 8th European Conf. on CBR, 2006, pp. 33–43.

[15] S. Massie, N. Wiratunga, S. Craw, A. Donati, E. Vicari, From anomaly reports to cases, in: International Conference on Case-Based Reasoning, Springer, 2007, pp. 359–373.