Competence Guided Model for Casebase Maintenance

Ditty Mathew and **Sutanu Chakraborti**
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600036
{ditty,sutanuc}@cse.iitm.ac.in

Abstract

A competence guided casebase maintenance algorithm retains a case in the casebase if it is useful to solve many problems and ensures that the casebase is highly competent. In this paper, we address the compositional adaptation process (of which single case adaptation is a special case) during casebase maintenance by proposing a case competence model for which we propose a measure called retention score to estimate the retention quality of a case. We also propose a revised algorithm based on the retention score to estimate the competent subset of a casebase. We used synthetic datasets to test the effectiveness of the competent subset obtained from the proposed model. We also applied this model in a tutoring application and analyzed the competent subset of concepts in tutoring resources. Empirical results show that the proposed model is effective and overcomes the limitation of footprint-based competence model in compositional adaptation applications.

1 Introduction

Case Based Reasoning(CBR) [Riesbeck and Schank, 1989] systems solve new problems by retrieving similar past problems from a casebase and adapting their solutions. Casebase Maintenance [Reinartz *et al.*, 2001] is a branch of CBR, which aims at looking into the quality of cases that should be retained in the casebase; the goal is often to maintain a compressed casebase that can solve new problems effectively. We need to ensure that the cases in the compressed casebase can be retrieved and adapted to solve a wide range of problems in the casebase. Thus, the competence of a casebase can be determined by the ability of cases in the casebase to solve a large number of problems. A competence guided casebase maintenance algorithm retains a case in the casebase if it is useful to solve many problems and ensures that the casebase is highly competent [Smyth and Keane, 1995].

Footprint-based retrieval [Smyth and McKenna, 1999] is an efficient retrieval approach in CBR, which guides the search procedure using a case competence measure called relative coverage [Smyth and McKenna, 1998]. This approach identifies a compact competent subset of the casebase called footprint set. However, the relative coverage measure used in this approach covers only the situation where a single retrieved case is adapted to solve a problem. It turns out that many CBR applications require compositional adaptation where the solution from multiple cases are combined to obtain a new solution [Wilke and Bergmann, 1998]. To the best of our knowledge, no previous work has attempted to address the maintenance of casebase which requires compositional adaptation. So, we are motivated by the research question, *"How can we model a competence guided casebase maintenance model where the adaptation process involves both single case and compositional adaptation?"*

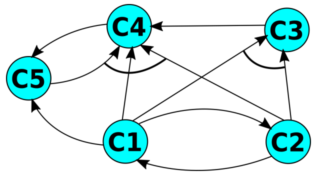

Figure 1: Example of casebase where compositional adaptation is involved

We illustrate the drawback of the competence model in footprint-based approach [Smyth and McKenna, 1999] when the adaptation process involves compositional adaptation. Fig 1 shows a network of cases where each node represents case and an edge from one case (say c_1) to another case (say c_2) indicates that the case c_1 can be retrieved and its solution can be adapted to solve c_2. As per [Smyth and McKenna, 1999], edge $c_1 \rightarrow c_2$ implies c_1 *solves* c_2. The arc (*AND* arc) between edges represents compositional adaptation. For example, the arc between edges $c_1 \rightarrow c_3$ and $c_2 \rightarrow c_3$ in the network indicates that a composition of c_1 and c_2 can solve the problem c_3. It is to be noted that neither case c_1 nor c_2 can solve c_3 in isolation. The footprint-based approach [Smyth and McKenna, 1999] does not consider the *AND* arcs between incoming edges, and outputs a footprint set $\{c_1\}$ corresponding to this network. So, the footprint set identified for the casebase in Fig 1 solves all cases in the network only when compositional adaptation is not taken into consideration. For example, case c_3 cannot be solved by this footprint set as it requires case c_2 which is not present in the footprint set, along with c_1 to solve it. The current competence model has to be enhanced to include compositional adaptation.

2 Approach

Compositional adaptation proposes a new solution by combining the solutions of multiple cases; cases which are used for adapting the new solution form an *AND* relation. It is possible to have multiple adapted solutions (either single case or compositional) for a target problem. These multiple solutions for a target problem shape an *OR* relation. The *AND* relation implies all cases that are part of this relation are required to adapt a new solution and the *OR* relation indicates any cases that involve in this relation can solve the target problem. The casebase is comprised of *AND-OR* relations between cases. We assume that the compositional adaptation operator is a disjunction over conjunctions.

We propose a measure called *retention score* to order cases in the casebase based on the extent to which a case is to be retained in the casebase. This measure can be applied in both single case and compositional adaptation applications. Using this measure, we propose a modification of Smyth's footprint [Smyth and McKenna, 1999] identification algorithm called footprint$_{CA}$ algorithm which accounts for compositional adaptation.

2.1 Retention Score

The retention score is a measure which quantifies the importance of a case in terms of whether it is required to be retained in the casebase or not. To illustrate the idea of retention score, consider the synthetic casebase graph in Fig 1. Here, case c_1 requires c_2 to solve c_3, and both c_2 and c_5 to solve c_4. The factors that determine the retention quality of a case are the range of problems that it solves and the number of cases that are required in conjunction with this case to solve those problems. In a casebase, we would like to retain fewer good retention quality cases that solve more useful cases. We define two terms to estimate retention score - covered cases and support cases.

The covered cases of a case c ($CoveredCases(c)$) include all cases that c can be used to solve either on its on, or in conjunction with other cases. For example, $CoveredCases(c_1)$ in the network shown in Fig 1 is $\{c_2, c_3, c_4, c_5\}$.

The support cases of a case c_i to solve the problem c_j ($SupportCases(c_i, c_j)$) is a set of cases that the case c_i requires to solve c_j. For example, in Fig 1 $SupportCases(c_1, c_3)$ is $\{c_2\}$ and $SupportCases(c_1, c_4)$ is $\{c_2, c_5\}$.

The proposed measure for retention score is based on these two sets and it is based on the idea that *a case has high retention score if it can solve several cases that have high retention score that is supported by less number of cases that have low retention score*. Using this idea we formulated a recursive formulation like PageRank [Page *et al.*, 1999] as given in Equation 1.

$$RetentionScore_{k+1}(c) = \sum_{c_i \in CoveredCases(c)} \frac{RetentionScore_k(c_i)}{1 + \sum_{c_j \in SupportCases(c, c_i)} RetentionScore_k(c_j)} \quad (1)$$

where $RetentionScore_{k+1}(c)$ is the retention score of a case c at $(k+1)^{th}$ iteration. The addition of 1 in the denominator is to handle the situation when a case does not require any support case to solve the corresponding covered case. The retention score of a case c for the first iteration is given as,

$$RetentionScore_0(c) = \sum_{c_i \in CoveredCases(c)} \frac{\frac{1}{1+|\{\mathbb{C}' \,:\, \mathbb{C}' \text{ solves } c_i \text{ and } c \notin \mathbb{C}'\}|}}{1 + |SupportCases(c, c_i)|} \quad (2)$$

For each covered case c_i in Equation 2, the numerator captures the individual contribution of c in solving c_i. The contribution of c in solving c_i is high if c is involved in all solutions of c_i, and the individual contribution of c to solve c_i is less when c_i can be solved without using c also. The denominator of Equation 2 ensures that the retention score increases with decrease in the number of support cases that c requires to solve c_i and vice versa. The addition of 1 in the denominator handles the situation when there are no supporting cases.

The retention score recursively measures the global competence of each case in the casebase. But, relative coverage measure used in the footprint-based approach [Smyth and McKenna, 1999] expresses only the individual contribution of each case irrespective of the requirements of other cases in solving a target problem.

2.2 Footprint$_{CA}$ Algorithm

The footprint algorithm proposed by [Smyth and McKenna, 1999] does not consider compositional adaptation while estimating the footprint set. We modified Smyth's footprint algorithm to obtain the footprint$_{CA}$ set and the algorithm is given below.

Algorithm 1: Footprint$_{CA}$ algorithm

```
Input: SortedCases : cases sorted based on retention score
Output: Footprint$_{CA}$ (FP)
FP $\leftarrow$ {}, Changes $\leftarrow$ true
while Changes do
    Changes $\leftarrow$ false
    for each $c \in$ SortedCases do
        if $c$ cannot be solved by any subset of cases in FP then
            Changes $\leftarrow$ true
            Add $c$ to FP
```

This algorithm processes each case in the decreasing order of retention score and each case is added to the footprint set only if it cannot be solved by any subset of cases in the footprint set. Thus the cases with high retention quality are added before the cases with less retention quality, and thus help to keep the good quality cases in the footprint set. We preserve the retention score ordering of cases in the final footprint set. In this way, the footprint$_{CA}$ set for the example in Fig 1 is obtained as $\{c_1, c_3\}$. The Smyth's footprint set [Smyth and McKenna, 1999] for the same graph includes only c_1. It may be noted that the footprint$_{CA}$ set can cover all concepts in the given network whereas the Smyth's footprint set $\{c_1\}$ cannot cover all cases in the network.

3 Evaluation

We empirically tested the proposed competence model by using synthetic datasets. The dataset generation process is illustrated below.

1. *Synthetic data 1:* $y = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$ +noise

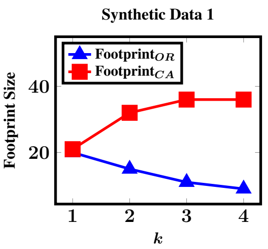

Figure 2: Footprint size analysis

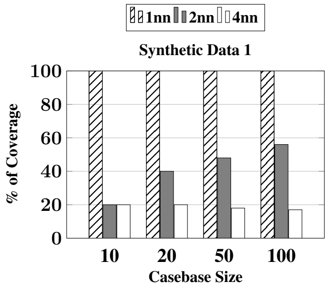

Figure 3: Casebase Coverage by Footprint$_{OR}$

2. *Synthetic data 2:* $y = x_1^4 + x_2^3 + x_3^2 + x_4 + \cos^2(x_5)$ +noise
3. *Synthetic data 3:* $y = \sin(x_1 x_2) + \sqrt{x_3 x_4} + \cos^2(x_5) + x_6 x_7 + x_8 + x_9 + x_{10}$ +noise

For each dataset, the x_i values of a data instance are sampled uniformly with values between 0 and 10; we added a random Gaussian noise with mean 0 and standard deviation 10. The y value corresponding to the x_i values of the data instance is computed from the formula. Each data instance is considered as a case in the casebase and each case is assumed to be solved by the compositional adaptation solution of its k-nearest neighbor cases. Thus the casebase graph contains cases as nodes, and edges from k-nearest neighbors of each case are connected to it by an AND arc. Then the footprint$_{CA}$ set is estimated using this graph and is compared with the footprint$_{OR}$ set (Smyth's footprint set [Smyth and McKenna, 1999]) which is obtained from the same graph by removing the composition (AND) condition. The experiments are done with k=1, 2 and 4 and by varying the number of instances (casebase size) from 10 to 100. At k=1, the adaptation process uses a single case; multiple cases are used when $k > 1$.

The analysis of the footprint size is one of the common criteria for evaluation. However, the sizes of both footprint sets are not strictly comparable as the footprint$_{CA}$ is expected to have more cases than the footprint$_{OR}$ set due to composition condition in the former set. The Fig 2 illustrates that the size of footprint$_{OR}$ decreases with increase in the value of k whereas the size of footprint$_{CA}$ increases with increase in the value of k. For a high value of k, more cases are involved in compositional adaptation during which the footprint$_{OR}$ size compresses more and thereby loses composition knowledge of adaptation.

We analyze the casebase coverage of footprint$_{CA}$ and footprint$_{OR}$ to estimate the effectiveness of footprint$_{CA}$ in compositional adaptation applications. The casebase coverage of a footprint set FP is the ratio of the number of cases that are solved by FP to the casebase size. The footprint$_{CA}$ obtained from all datasets have full casebase coverage. However, the footprint$_{OR}$ set covers the entire casebase only when k=1. The analysis of coverage on footprint set in *Synthetic Data 1* is illustrated in Fig 3. We can observe that the percentage of coverage increases with increase in the number of data instances when k=2. Also, the coverage percentage decreases with increase in the value of k. The reason behind this is that the increase in the number of neighbors decreases the size of footprint set and thereby reduces its effectiveness.

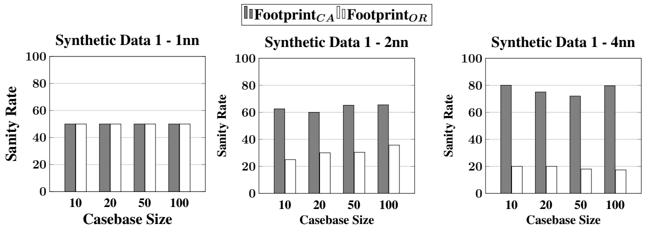

Figure 4: Sanity Rate of footprint cases in Synthetic Datasets

This indicates the ineffectiveness of the footprint$_{OR}$ set in compositional adaptation applications.

To measure the sanity of the footprint set, we found a method to identify a set of cases that can cover the entire casebase using a graph-theoretic approach. This set is identified from the casebase network where footprint$_{CA}$ is estimated. In this network, if we repeatedly remove the cases that do not solve any other cases until there are no such cases, the final network turns out to be a compressed set of cases that can solve all cases in the casebase transitively. This final network is called the *kernel* of the case network. Though there is no ordering of cases provided within the kernel, the cases in the kernel are the potential cases that can be presented in a footprint set. So, we compare the cases in the footprint set and kernel. The sanity measure is defined as,

$$\text{Sanity rate} = \frac{|\text{footprint cases} \cap \text{kernel cases}|}{|\text{kernel cases}|} \times 100 \qquad (3)$$

This idea is adapted from [Massé *et al.*, 2008] where Masse et. al estimates the grounding kernel of a dictionary graph where the graph is constructed from word definitions. Here the grounding kernel turns out to be the set of words using which the entire dictionary words have been defined.

The sanity rate of footprint$_{CA}$ and footprint$_{OR}$ are compared for 1nn, 2nn, 4nn and various casebase sizes. The results obtained for *Synthetic data 1* is depicted in Fig 4 and we have observed similarly results in other datasets. We can observe that footprint$_{CA}$ has significantly higher sanity rate when k >1. At k=1 (single case adaptation), both methods are performing similar which indicates that footprint$_{CA}$ is as good as footprint$_{OR}$ in the single case adaptation process.

4 Footprint$_{CA}$ in Tutoring Application

Encyclopedic resources like Wikipedia and general dictionaries do not have rich pedagogical content, tailored to suit the users learning goals [Mathew *et al.*, 2015]. The concepts in Wikipedia (articles) as well as in dictionary (words) are not arranged in a learning order whereas an ideal textbook explains a concept before referring it which results in a sequential order for learning [Agrawal *et al.*, 2012]. So, sequencing concepts in Wikipedia like resources may help online learners to fulfill their learning goal. Here we illustrate the usefulness of retention score in ordering concepts in Wikipedia.

We draw an analogy of the Wikipedia network to a casebase, in order to sequentially order Wikipedia articles (concepts) such that the ordering satisfies user's learning goal. At a high level, a Wikipedia page corresponds to a case, where

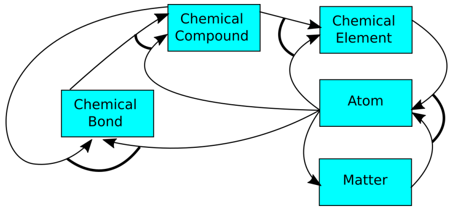

Figure 5: An example of Wikipedia network

the article name is the problem definition, and the solution corresponds to the representation of its meaning. The meaning of a Wikipedia article is given by its definition where it is approximated as the first sentence in the article [Ye *et al.*, 2009]. In order to construct a Wikipedia network analogous to casebase graph, we see a case c_A solves c_B as analogous to Wikipedia article A helps in understanding another article B. The articles pointed to by hyperlinks in the first sentence are assumed as concepts that help in understanding the corresponding article. For example, in Fig 5 the concept *atom* is explained in terms of *chemical element* and *matter*. Hence, the arc between the edges from *chemical element* and *matter* to *atom* which forms an *AND* relation indicates that the concepts *chemical element* and *matter* are composed together to explain *atom*. The resulting network resembles a casebase graph where compositional adaptation is used. We use the idea of retention score as used in casebase maintenance to arrive at an appropriate ordering of concepts.

A concept with high retention score is likely to be a basic concept as its coverage will be high due to its repetitive usage in defining other concepts. Thus, the ordering based on retention score provides an order in which one can learn entire concepts under a specific topic. The ordering of elements in the footprint$_{CA}$ set indicates the learning order where the position in the order implies the level of completion of learning. To satisfy the learning goal, one can learn concepts in footprint$_{CA}$ in the retention score ordering. While learning each concept in the footprint$_{CA}$, concepts that are solved by elements in footprint$_{CA}$ can be learned. Note that these concepts may not be present in the footprint$_{CA}$ set. Thus, footprint$_{CA}$ and the retention score ordering helps a learner to satisfy his/her goal.

4.1 Empirical Results

The effectiveness of retention score and footprint$_{CA}$ set is analyzed on the network extracted from the Wikipedia and dictionary. For Wikipedia, the network is constructed using the articles in Wikipedia Artificial Intelligence(AI) category and its sub-categories(wikiAI). In a dictionary, concepts are words that are defined in it. The content words in the definition of each word are marked as the concepts that help in understanding that word. These content words form an *AND* relation with the word being defined. We make a simplifying assumption that the words in the dictionary are sense disambiguated. So, the content words present in the first definition of the first sense is considered as the composed solution for each word. Thus, we have taken definitions from the WordNet(wn+ld) and Longman dictionary of contemporary English and the corresponding network results in an *AND-OR* graph due to the presence of multiple compositional solutions. Similarly, other networks are constructed using only WordNet(wn) and only Longman dictionary(ld).

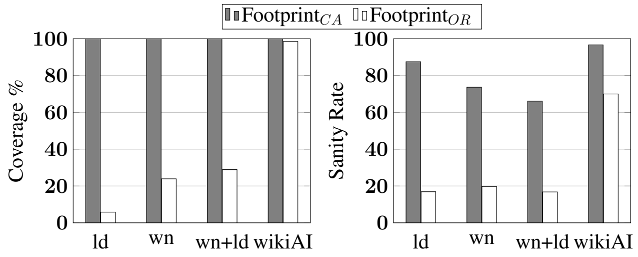

Figure 6: Coverage and Sanity Rate Analysis of Wikipedia and Dictionary Networks

We analyzed the coverage of concepts in the network by footprint$_{CA}$ set and footprint$_{OR}$ set in all networks and this is shown in Fig 6. This measure is similar to casebase coverage which is used for evaluating synthetic datasets. From Fig 6, it can be observed that the footprint$_{CA}$ covers all concepts in all networks whereas the footprint$_{OR}$ set covers only less than 30% of concepts in all networks except wikiAI. The higher coverage of footprint$_{OR}$ in wikiAI can be due to the less number of hyperlinks in the first sentence of each article.

The sanity of footprint$_{CA}$ and footprint$_{OR}$ are analyzed using the sanity rate formulated in Equation 3. In Fig 6, we can observe that the sanity rates of footprint$_{CA}$ concepts in all networks are more than 65% and that of footprint$_{OR}$ concepts are less than 20% except the wikiAI dataset which might be due to the lack of compositional information in the dataset. This indicates that the footprint$_{CA}$ set is useful for compositional adaptation applications.

5 Conclusion

We start with the observation that Smyth's footprint-based approach [Smyth and McKenna, 1999] is not designed for compositional adaptation applications. We proposed a measure called retention score to estimate the retention quality of a case that involves compositional adaptation. Using this score, we proposed a revised approach to identify the footprint$_{CA}$ set where compositional adaptation is required. We tested the effectiveness of the footprint$_{CA}$ using synthetic datasets and compared it with the Smyth's footprint set. The empirical results demonstrated the improved performance of our model when compositional adaptation is required; the proposed model performs equally well as Smyth's model during single case adaptation process. We also illustrated and tested the effectiveness of our method in tutoring application.

Acknowledgments

This paper is a an abridged report of the paper entitled ”Competence Guided Casebase Maintenance for Compositional Adaptation Applications” that appears in the *Proceedings of International Conference on Case-Based Reasoning*, [Mathew and Chakraborti, 2016].

References

[Agrawal *et al.*, 2012] Rakesh Agrawal, S Chakraborty, S Gollapudi, A Kannan, and K Kenthapadi. Quality of Textbooks: An Empirical Study. *ACM Symposium on Computing for Development*, 2012.

[Lekkas *et al.*, 1994] Georgios P Lekkas, Nicholas M Avouris, and Loizos G Viras. Case-Based Reasoning in Environmental Monitoring applications. *Applied Artificial Intelligence An International Journal*, 8(3):359–376, 1994.

[Massé *et al.*, 2008] A. Blondin Massé, G. Chicoisne, Y. Gargouri, S. Harnad, O. Picard, and O. Marcotte. How is Meaning Grounded in Dictionary Definitions? In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for NLP*, pages 17–24, 2008.

[Mathew and Chakraborti, 2016] Ditty Mathew and Sutanu Chakraborti. Competence Guided Casebase Maintenance for Compositional Adaptation Applications. In *International Conference on Case-Based Reasoning*, pages 265–280. Springer, 2016.

[Mathew *et al.*, 2015] Ditty Mathew, Dhivya Eswaran, and Sutanu Chakraborti. Towards Creating Pedagogic Views from Encyclopedic Resources. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 190–195, 2015.

[Page *et al.*, 1999] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[Reinartz *et al.*, 2001] Thomas Reinartz, Ioannis Iglezakis, and Thomas Roth-Berghofer. Review and Restore for Case-Base Maintenance. *Computational Intelligence*, 17(2):214–234, 2001.

[Riesbeck and Schank, 1989] Christopher K Riesbeck and Roger C Schank. Inside Case-Based Reasoning. Lawrence Erlbaum, Hillsdale, NJ, 1989.

[Smyth and Keane, 1995] Barry Smyth and Mark T Keane. Remembering to Forget. In *Proceedings of the 14th International Joint Conference on Artificial intelligence*, pages 377–382, 1995.

[Smyth and McKenna, 1998] Barry Smyth and Elizabeth McKenna. Modelling the Competence of Casebases. In *European Workshop on Advances in Case-Based Reasoning*, pages 208–220, 1998.

[Smyth and McKenna, 1999] Barry Smyth and Elizabeth McKenna. Footprint-based Retrieval. In *International Conference on Case-Based Reasoning*, pages 343–357, 1999.

[Wilke and Bergmann, 1998] Wolfgang Wilke and Ralph Bergmann. Techniques and Knowledge used for Adaptation during Case-Based Problem Solving. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 497–506, 1998.

[Ye *et al.*, 2009] Shiren Ye, Tat-Seng Chua, and Jie Lu. Summarizing Definition from Wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on NLP of the AFNLP*, pages 199–207, 2009.