Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

An Empirical Study of Knowledge Tradeoffs in Case-Based Reasoning

Devi Ganesan and **Sutanu Chakraborti**
Indian Institute of Technology Madras, Chennai, India
{gdevi, sutanuc}@cse.iitm.ac.in

Abstract

Case-Based Reasoning provides a framework for integrating domain knowledge with data in the form of four knowledge containers namely Case base, Vocabulary, Similarity, and Adaptation. It is a known fact in Case-Based Reasoning community that knowledge can be interchanged between the containers. However, the explicit interplay between them, and how this interchange is affected by the knowledge richness of the underlying domain is not yet fully understood. We attempt to bridge this gap by proposing footprint size reduction as a measure for quantifying knowledge tradeoffs between containers. The proposed measure is empirically evaluated on synthetic as well as real-world datasets. From a practical standpoint, footprint size reduction provides a unified way of estimating the impact of a given piece of knowledge in any knowledge container, and can also suggest ways of characterizing the nature of domains ranging from ill-defined to well-defined ones. Our study also makes evident the need for maintenance approaches that go beyond case base and competence to include other containers and performance objectives.

1 Introduction

Case-Based Reasoning (CBR) is a paradigm of reasoning inspired by the human way of using past experiences to solve new problems [Kolodner, 1992; Aamodt and Plaza, 1994]. Unlike most machine learning algorithms that are purely data-driven, CBR provides a framework for combining domain knowledge with data. Problem-solving in CBR involves the use of four knowledge containers namely *Vocabulary, Case Base, Similarity* and *Adaptation*. The effectiveness of the reasoner can be improved by carefully handling the knowledge containers [Richter, 1995], i.e. interchanging the knowledge between containers to improve its performance. While past work on CBR has realized the interplay between the containers, the impact of this interchangeability across a diversity of domains has not been studied in isolation.

The interplay between containers also has an important role to play in the evolution of a CBR system. An initial CBR system could be just a large collection of cases. As the system evolves, knowledge can be shifted from case-base to similarity or adaptation containers and can even lead to a revised vocabulary. Learning to improve vocabulary is very hard to be automated and is still largely dependent on the domain experts. However, there has been past work on inducing similarity measures [Stahl, 2001; Cheng and Hüllermeier, 2008] and adaptation rules [Hanney and Keane, 1997; Craw *et al.*, 2006] from case base either with minimal or no feedback from domain experts. These learning techniques have largely focussed on individual knowledge containers, thus not laying adequate emphasis on the interdependence between containers. Realizing the utility of cases present in a case base is impossible without a good similarity measure. Similarly, adaptation knowledge is effective only when the relevant case is retrieved from the case base in the first place, and this, in turn, is dependent on the similarity measure.

The observation above motivates the need to seek a unified ground for studying the interchangeability between CBR knowledge containers. More specifically, in this work we propose a novel measure based on footprint set [Smyth and McKenna, 1999] to quantify knowledge tradeoffs between containers. The proposed idea is evaluated using synthetic and real-world datasets, and the tradeoffs are visually illustrated using parallel coordinate plots [Inselberg, 2009]. Our work also includes a study of the following two factors that influence tradeoffs between containers: nature of underlying domain and user demands on solution quality. The paper is concluded with a discussion on practical implications and future extensions of the proposed idea. In particular, we envisage that our interpretation of footprint cases and knowledge transfers will streamline and motivate new avenues of cross-container maintenance activities in CBR.

2 Background

Knowledge Containers The four knowledge containers and problem-solving methodology in CBR are briefly explained in this section. *Case base* is the repository of experiences known to a case-based reasoner. Each experience is stored as a problem-solution pair known as a case. *Vocabulary* specifies the choice of language that is used to gather, organize and represent the cases. *Similarity* knowledge guides the retrieval of past experiences that are potentially useful in solving the target problem. *Adaptation* knowledge is used to modify the retrieved solution to address the specific needs of

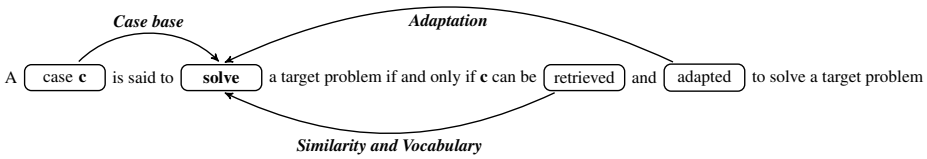


Figure 1: The *Solves* function in footprint algorithm connects the footprint set to all four knowledge containers.

the target problem. A typical problem-solving cycle involves four steps – Retrieve, Reuse, Revise and Retain [Aamodt and Plaza, 1994]. In Retrieve step, the reasoner searches its repository for the cases most similar to the target problem. In Reuse step, the retrieved cases are adapted to propose a solution for the target problem. The proposed solution is optionally revised by a domain expert and retained in case base if deemed useful.

Competence and Footprint Set Competence of a CBR system is the range of target problems it can solve. It depends critically on the competence of its case base, which in turn, is estimated from the competence of its individual cases. Smyth and McKenna [1998] propose a model of case competence which assumes that the cases in the case base are a representative sample of the target problems. Under this model, the local competence of a case is computed from its *coverage* and *reachability* sets, which are defined based on the *solves* function. A case c is said to solve a target problem t if and only if c can be retrieved and adapted to *solve* t. *Coverage set* includes those target problems t that c can solve. *Reachability set* includes those cases that can solve c. *Related set* is the union of the coverage and reachability sets. Some cases may exhibit *shared coverage* due to overlap of their related sets. While calculating the global competence of a reasoner, the cases are grouped into *competence groups* to avoid the problem of duplicate coverage.

Footprint set proposed by Smyth and McKenna [1999] is a minimal set of cases that has the same competence (problem-solving ability) as the entire case base. To compute the footprint set, the authors define a measure called *relative coverage*. For a case c, its relative coverage is computed by weighing the contribution of each of its covered cases by the degree to which these cases are themselves covered. By definition, each competence group makes a unique contribution to the competence of the case base. Hence, footprint set of the case base is the union of the footprint cases of all its competence groups. The non-redundant cases within each competence group constitute its *footprint cases*.

3 Approach

Footprint Size as a Unit of Knowledge In our work, we hypothesize that knowledge contained in a case base is only as good as the knowledge contained in its footprint cases. *We use footprint size to quantify the knowledge contained in case base*. One simplifying assumption is the equivalence of footprint cases in their contribution to case base competence.

Iglezakis and Roth-Berghofer [2000] discuss the centrality of case base in maintenance activities and one of their hypotheses is that *cases are natural crystallization points for the knowledge in case-based reasoning systems*. Further, according to Smyth and McKenna [1998], the *competence group is a fundamental unit of competence in a case base*. Their views reinforce the use of footprint size to quantify the knowledge contained in the case base.

Footprint Size Reduction to Quantify Knowledge Tradeoffs The function *solves* in footprint algorithm connects the footprint set to the four knowledge containers (Figure 1). Adding or removing knowledge from containers impacts the footprint set through the *solves* function. This motivates our second hypothesis that *adding useful knowledge to the Vocabulary, Similarity or Adaptation container leads to a reduction in footprint size*, i.e. knowledge is traded off between case base and other containers. Under this hypothesis, we propose *footprint size reduction* as a unit for measuring knowledge tradeoffs between containers.

Let V, CB, S, R represent the Vocabulary, Case Base, Similarity and Adaptation (Reuse) containers and $|FP_{(V,CB,S,R)}|$ be the size of footprint set. The knowledge added by changing V to V', S to S' and R to R' can be quantified as below.

$$\Delta Knowledge \approx |FP_{(V,CB,S,R)}| - |FP_{(V',CB,S',R')}| \quad (1)$$

Pairwise Tradeoffs To keep the discussion simple while retaining the essence of the idea proposed, we focus on the following four types of knowledge tradeoffs in a case-based reasoner.

- Case base and Vocabulary tradeoff
- Case base and Similarity tradeoff
- Case base and Adaptation tradeoff
- Similarity and Adaptation tradeoff

In practice, it may be necessary to update more than one container simultaneously. For example, changing the vocabulary may trigger a corresponding change in similarity measure. This paper focusses only on the above pairs of isolated containers as it serves as a proof of concept for the proposed measure.

4 Experiments and Results

In this section, we use *footprint size reduction* to *quantify* tradeoffs between containers on synthetic and real world datasets.

4.1 Synthetic Dataset

We generated a synthetic case base (Table 1) using the function $D = 6A + 3B + C$. (A, B, C) represents the problem component of a case while D is the solution component.

A,B,C,D	A,B,C,D	A,B,C,D
1,1,1,10	1,1,2,11	1,1,3,12
1,2,1,13	1,2,2,14	1,2,3,15
1,3,1,16	1,3,2,17	1,3,3,18
2,1,1,16	2,1,2,17	2,1,3,18
2,2,1,19	2,2,2,20	2,2,3,21
2,3,1,22	2,3,2,23	2,3,3,24
3,1,1,22	3,1,2,23	3,1,3,24
3,2,1,25	3,2,2,26	3,2,3,27
3,3,1,28	3,3,2,29	3,3,3,30

Table 1: Synthetic case base (27 cases) for regression.

Vocabulary	Footprint Size	Tradeoff (CB, V')
V: A,B,C as attributes	18.0	-
V': X,C as attributes	9.2	8.8

Table 2: Tradeoffs between case base and vocabulary, calculated using Equation 2.

A case-based reasoner to predict D given the target problem (A, B, C) is said to solve it when the predicted solution is within 10% of D. Effects of change in the acceptable prediction error are studied in Section 5. In all the experiments on synthetic case base, the results are averaged from 5 fold train-test splits, and the relation between footprint size reduction and knowledge transfers is tested for statistical significance.

Case base and Vocabulary Tradeoff

Tradeoff between case base and vocabulary container is measured by fixing the knowledge in similarity and adaptation containers and revising V to V' as given in the equation below.

$$\text{Tradeoff}(CB, V') = |FP_{(V,CB,S,R)}| - |FP_{(V',CB,S,R)}| \quad (2)$$

In the synthetic case base, we fix S and R as uniform global similarity and null adaptation respectively. Vocabulary revision changes the problem representation from (A, B, C) to (X, C). X is a virtual attribute calculated from A, B as $X = 6A + 3B$. From Table 2, vocabulary revision from V to V' is equivalent to a knowledge tradeoff of 8.8 footprint cases between case base and vocabulary container. We found that the reduction in footprint size with increase in vocabulary knowledge is statistically significant ($p < 0.001$). A designer can choose to have a case base with 18 footprint cases and low vocabulary knowledge or a case base with 9 footprint cases and high vocabulary knowledge since both configurations have the same competence.

Case base and Similarity Tradeoff

Tradeoff between case base and similarity container is measured by fixing the knowledge in vocabulary and adaptation containers and revising S to S' as given in below equation.

$$\text{Tradeoff}(CB, S') := |FP_{(V,CB,S,R)}| - |FP_{(V,CB,S',R)}| \quad (3)$$

In the synthetic dataset, similarity knowledge is represented using the global weight vector. For example, weight vectors $(3, 2, 1)$ and $(6, 3, 1)$ are in line with the relative importance of attributes in domain theory. Table 3 shows five different similarity knowledge settings S through S_4; the column named domain knowledge is a qualitative description of similarity knowledge based on how reflective the global weight vector is, of the underlying domain theory.

Similarity	Domain Knowledge	Footprint Size	Tradeoff (CB, S')
S: 1,1,1	Low	18.0	-
S_1: 2,1,1	Mid	15.2	2.8
S_2: 3,1,1	Mid	15.0	3.0
S_3: 3,2,1	High	9.4	8.6
S_4: 6,3,1	High	9.4	8.6

Table 3: Tradeoffs between case base and similarity, calculated using Equation 3 where S' varies from S_1 to S_4.

Adaptation	Domain Knowledge	Footprint Size	Tradeoff (CB, R')
R	Null	18.0	-
R_1	Low	17.4	0.6
R_2	Mid	14.6	3.4
R_3	High	11.0	7.0

Table 4: Tradeoffs between case base and adaptation knowledge, calculated using Equation 4 where R' varies from R_1 to R_3.

When S is revised to S_4, knowledge equivalent to 8.6 footprint cases is traded off between case base and similarity container. The reduction in footprint size with increase in domain knowledge in the weight vectors is statistically significant ($p < 0.001$). All 5 design choices corresponding to each row in Table 3 have the same competence but varying levels of knowledge in similarity container and case base.

Case base and Adaptation Tradeoff

Tradeoff between case base and adaptation container is measured by fixing the knowledge in vocabulary and similarity containers and revising R to R' as given in the equation below.

$$\text{Tradeoff}(CB, R') := |FP_{(V,CB,S,R)}| - |FP_{(V,CB,S,R')}| \quad (4)$$

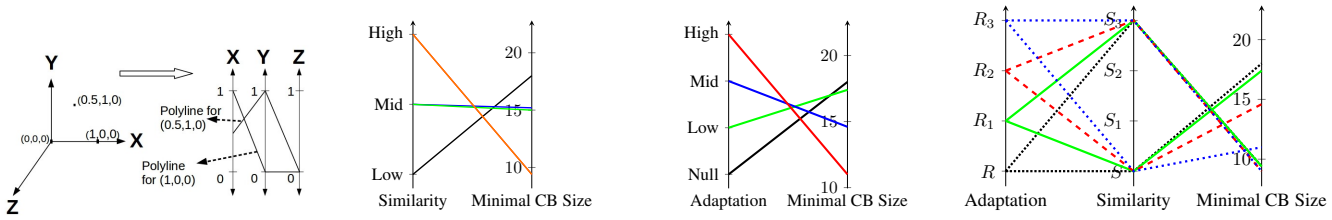
Let Q be the query problem and N be the nearest case in terms of absolute distance. Adaptation rules used are null adaptation R; R_1: add $3 \times (N_A - Q_A)$ to N's solution; R_2: add $3 \times (N_A - Q_A) + 2 \times (N_B - Q_B) + (N_C - Q_C)$ to N's solution; R_3: add $6 \times (N_A - Q_A) + 3 \times (N_B - Q_B) + (N_C - Q_C)$ to N's solution. As in previous cases, reduction in footprint size with richer adaptation knowledge is statistically significant ($p < 0.001$).

Similarity and Adaptation Tradeoff

The tradeoff between similarity and adaptation containers is measured by fixing the knowledge in vocabulary and case base (Equation 5). In Table 5, same adaptation knowledge

Knowledge	R: Null	R_1	R_2	R_3
S: 1,1,1	18.0 (-)	17.4 (0.6)	14.6 (3.4)	11.0 (7.0)
S_1: 2,1,1	15.2 (2.8)	15.2 (2.8)	10.0 (8.0)	10.0 (8.0)
S_2: 3,1,1	15.0 (3.0)	15.0 (3.0)	9.8 (8.2)	9.8 (8.2)
S_3: 3,2,1	9.4 (8.6)	9.4 (8.6)	9 (9.0)	9 (9.0)
S_4: 6,3,1	9.4 (8.6)	9.4 (8.6)	8.6 (9.4)	8.6 (9.4)

Table 5: Tradeoffs between similarity and adaptation containers, measured using Equation 5 where S' varies from S_1 to S_4 and R' from R_1 to R_3. A value of 15.0 (3.0) stands for footprint size of 15 and Tradeoff(S', R') of 3 footprint cases.



(a) Parallel Coordinates (b) Similarity versus CB (c) Adaptation versus CB (d) Similarity versus Adaptation

Figure 2: Visualization of knowledge tradeoffs in synthetic dataset. Minimal CB size is the effective case base required and is the same as the footprint size. Each polyline represents a choice of knowledge distribution in containers and *has the same competence as all other polylines*

gives different tradeoffs with different similarity measures. For example, R_1 yields a tradeoff of 0.6 with S; this increases to 8.6 for S_3 and S_4. This shows that the effectiveness of adaptation knowledge depends on the similarity measure as the retrieved case must be appropriate for adaptation.

$$\text{Tradeoff}(S', R') := |FP_{(V,CB,S,R)}| - |FP_{(V,CB,S',R')}| \quad (5)$$

Parallel Coordinates for Visualization

Parallel Coordinates (|| coords) is a visualization technique introduced by Alfred Inselberg [Inselberg, 2009], popularly used for discovering visual patterns in multivariate datasets. A dataset of n dimensions is plotted on n equally spaced and parallel axes. As shown in Figure 2a, a point in n dimensional space becomes a polyline in || coords. Figure 2 shows the || coords plots for pairwise tradeoffs. We experimented with alternate visualizations like 3D scatter plots and bubble charts, and found || coords to be most expressive in demonstrating the trends as well as extents of tradeoffs. Figures 2b and 2c show that as *knowledge in one container decreases, knowledge in the other needs to increase to maintain the same competence*. Figure 2d shows the interaction between similarity and adaptation containers as described in previous section.

4.2 Real-World Datasets

Next, we discuss empirical results on three real world datasets taken from UCI machine learning repository [Dheeru and Karra Taniskidou, 2017] namely *Iris*, *Auto-MPG* and *Boston Housing* and two textual datasets based on 20 Newsgroups [Lang, 1999]. Our choice of datasets is guided by characteristics such as size of case base, number of attributes, choice of representation and our knowledge about their domains. Primarily, our emphasis is on demonstrating knowledge tradeoffs in representative datasets from diverse CBR settings like regression, classification, conversational CBR and textual CBR.

Iris Dataset

Iris is a widely used benchmark dataset for classification and has 150 cases with 4 attributes each. Initial exploratory analysis of data revealed that the attributes *petal length* and *petal width* are more important than others for Iris species classification. We incorporated this knowledge in two ways - into similarity measure and into the vocabulary. We found that this knowledge is best represented in vocabulary container as it gives the maximum tradeoff with case base.

Vocabulary	Similarity	Footprint Size	Tradeoff
V_1: Flat Attribute	S_1:(3,2,0,0)	87	63
	S_2:(1,1,1,1)	64	86
	S_3:(0,0,3,2)	57	**93**
V_2: Decision Tree + Flat Attribute	S_1:(3,2,0,0)	43	107
	S_2:(1,1,1,1)	30	120
	S_3:(0,0,3,2)	27	**123**

Table 6: Iris dataset: Tradeoffs among Case base, Vocabulary and Similarity containers, with null adaptation. Tradeoff is (case base size − footprint size).

Domain Knowledge in Similarity Container As in the case of the synthetic dataset, the global weight vector is modified to impart domain knowledge. The vector $(0, 0, 3, 2)$ emphasizes the importance of *petal length* and *petal width* while $(3, 2, 0, 0)$ downplays them. S_3 gives the maximum tradeoff with case base (Table 6). One can also compare two similarity measures in terms of their footprint size reduction. Knowledge in S_3 is more than in S_1 by 30 footprint cases.

Domain Knowledge in Vocabulary Container From data analysis using decision trees, we found that the species *Iris Setosa* always has petal width lesser than 1 while *Iris Versicolor* always has petal length greater than 5.1. We added two binary valued attributes one each for $petallength < 1$ and $petalwidth > 5.1$ and used a Conversational Case Base Reasoning (CCBR) [Aha and Muñoz-Avila, 2001] style approach to return the corresponding label based on two successive questions. Only when the user answers negatively to both these questions, does the retrieval step proceed to search the case base. Search space is reduced because the first two questions have eliminated some portion of the case base. Hence, a richer vocabulary further reduces the footprint size as empirically confirmed by the results reported in Table 6.

Auto-MPG Dataset

The Auto-MPG dataset contains 392 cases with 8 attributes and the task is to predict the miles per gallon (mpg) of a car. The attributes *displacement* and *weight* are important factors affecting the fuel consumption. S_0 is a similarity measure that ascribes equal weights to all attributes. S_1 emphasizes the weight attribute and S_2 emphasizes both weight and displacement attributes. Adaptation knowledge is R_0 which is null adaptation, or R_1 which is adapting the mpg value of the 1-Nearest Neighbour proportionate to the differences in *weight* and *displacement* attributes. Acceptable Prediction

Adaptation →	R_0-Low		R_1-High	
Similarity	Footprint Size	Tradeoff	Footprint Size	Tradeoff
S_0-Low	304	88	302	90
S_1-Mid	309	83	291	**101**
S_2-High	300	92	298	94

Table 7: Auto-MPG dataset: Tradeoffs among Case base, Similarity and Adaptation containers. Vocabulary is flat attribute representation. Tradeoff is (case base size − footprint size).

Adaptation →	R_0-Low		R_1-High	
Similarity	Footprint Size	Tradeoff	Footprint Size	Tradeoff
S_0-Low	413	93	382	**124**
S_1-Mid	407	99	394	112
S_2-High	405	101	389	117

Table 8: Boston housing dataset: Tradeoffs among Case base, Similarity and Adaptation containers. Vocabulary is flat attribute representation. Tradeoff is (case base size − footprint size). Acceptable Prediction Error is 5%.

Error is 5%. From Table 7, the combination of S_1 and R_1 gives the maximum tradeoff with case base. It is interesting to note that the pair S_2, R_1 is outperformed by S_1, R_1; this confirms that a chosen similarity knowledge shows preferential attachment to a certain adaptation knowledge; best performance is not necessarily obtained when both similarity and adaptation knowledge are rich, since they interact with each other non-linearly.

Boston Housing Dataset

The housing dataset contains 506 cases and the task is to predict the price of a house given 14 other attribute values. Exploratory data analysis revealed the attributes *INDUS, RM, DIS, PTRATIO, LSTAT* to be important for predicting the house price. *INDUS* is the proportion of non-retail business acres per town; *RM* is the average rooms per house; *DIS* is the weighted distances to employment centers; *PTRATIO* is the pupil to teacher ratio in the neighborhood; *LSTAT* is the percentage of lower status of the population. As before, the global weight vector was used to incorporate domain knowledge into the similarity container. S_0 is a similarity measure that uniformly weighs the attributes, S_1 emphasizes *RM, PTRATIO, LSTAT* uniformly and S_2 emphasizes *INDUS, DIS* additionally over S_1. Adaptation was either null adaptation (R_0) or included three rules (R_1) based on the following domain knowledge: *If the number of rooms increases, the house price increases*; *If the distance to employment centers increase, the house price decreases*; *If the proportion of lower status population in the housing location is high, the house price decreases*. From Table 8, the combination of S_0 and R_1 gives the maximum tradeoff with case base. In addition to the results reported above, we have also examined the impact of knowledge transfer and footprint reduction on generalization over test data. In the Iris dataset, test accuracies improved conspicuously by 12.66% and 23.67% as similarity knowledge was enriched from S_1 through S_3 under settings of V_1 and V_2 respectively (Table 6). Similar trends were observed for Auto-MPG and Boston, though the accuracy improvements due to footprint size reduction were less conspicuous in the relatively less knowledge-rich domains like Boston.

	Relpol		Hardware	
Vocabulary	Footprint Size	%CB Compression	Footprint Size	%CB Compression
Count Vectors	1276	57.9	564	51.7
Tfidf Vectors	1259	58.4	563	51.8
LSA vectors	670	**77.9**	497	**57.4**

Table 9: Textual CBR: Tradeoffs between Vocabulary and Case base on Relpol and Hardware datasets. Cosine similarity measure and null adaptation were used.

Textual CBR

Relpol and Hardware are two text classification datasets based on 20 Newsgroups [Lang, 1999] with 3031 and 1168 cases respectively. Relpol has two classes religion and politics and is relatively easy to classify. In contrast, the Hardware domain is harder to classify since there is considerable vocabulary overlap between documents of the two classes IBM and Mac. On these two datasets, we experimented with different representations: count-based, Term Frequency Inverse Document Frequency (TFIDF)[1] and Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990] vectors. In Table 9, we observe that footprint size decreases as the representation becomes richer. LSA gives the maximum reduction in footprint size on both datasets. We also measured the extent of compression of case base by footprint set. The case base compression ratio in the Hardware dataset is lower than that of the Relpol dataset; this is in line with their complexity estimates, 2.0358 and 1.0028, reported in [Chakraborti *et al.*, 2008].

5 Characterization of Domains

Though CBR is intended to operate over ill-defined domains [Kolodner, 1992], the general CBR paradigm does not place any restrictions on modeling of a wide spectrum of domains. At one extreme are the well-defined domains where we employ knowledge-rich approaches. At the other extreme are the ill-defined domains, where we employ knowledge-light or data intensive approaches. We see a CBR system that relies only on cases as the equivalent of a data-driven reasoner as opposed to a knowledge-driven reasoner that uses rich domain knowledge in its containers. We expect that in well-defined domains, the benefit of using domain knowledge is very high whereas the benefit fades as we move towards ill-defined domains. In an attempt to demonstrate this trend, we represent the benefit of domain knowledge (referred henceforth as Benefit) by the maximum reduction in footprint size.

To simulate a spectrum of domains, we employ a noisy channel model. The synthetic case base in Table 1 illustrates a well-defined domain where perfect domain theory is available. We assume that the more ill-defined the domain is, the less faithful is the domain model we have to explain our observations. To simulate this scenario in a controlled setting, we make the data points deviate from domain theory (D = 6A+3B+C). The extent of this deviation is called 'noise', which we vary to simulate domains of different knowledge levels. This notion of 'noise' may be contrasted with the usual notion of noise in a non-synthetic setting.

[1] https://en.wikipedia.org/wiki/Tf-idf

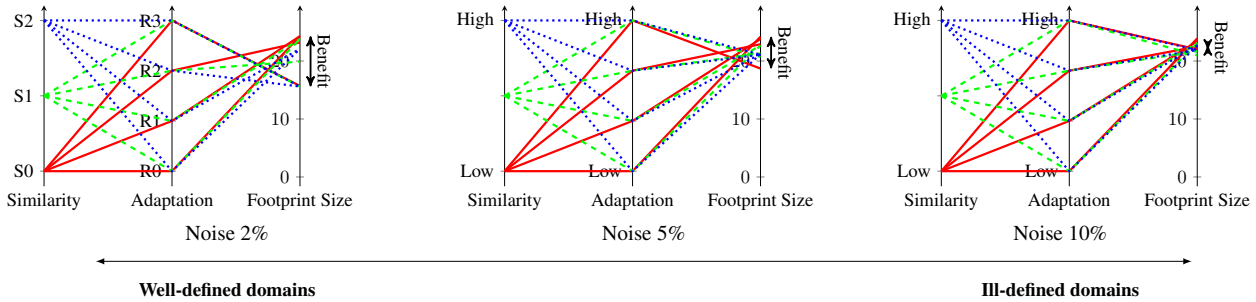


Figure 3: Data and knowledge tradeoffs decrease as we move from well-defined to ill-defined domains. In each || coord plot, the polylines correspond to different combinations of similarity and adaptation knowledge and their footprint sizes. Each color corresponds to a distinct similarity configuration.

We measure the footprint size across the simulated domains and the maximum reduction is marked by *double-headed arrows* on || coords in Figure 3. As the scale of noise is increased from 1% to 10%, the maximum footprint reduction decreases and a corresponding reduction in Benefit (see Figure 3). In ill-defined domains (noise levels close to 10%), while global knowledge pertaining to underlying domain theory becomes less useful, any local knowledge (i.e. knowledge from the cases and/or adaptation knowledge pertaining to the local neighbourhood of cases) may still be useful. This is indicative of the fact that we need more and more data to get good predictions in ill-defined domains.

Impact of User Demands on Knowledge Tradeoffs The design of a case-based reasoner is influenced not only by the nature of underlying domain but also by the user requirements on solution quality. Footprint algorithm naturally accommodates user demands on solution quality by way of its *Solves* function. In regression settings, the allowable prediction error represents the quality demands. Each row in Figure 4 shows the trend that a given piece of knowledge (similarity or adaptation) is likely to result in a higher reduction in footprint size if the user is less stringent (APE: 10%), than in a scenario where she is stringent (APE: 2%).

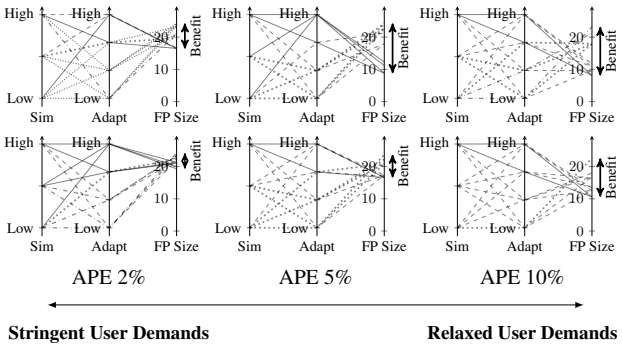


Figure 4: Impact of user demands on knowledge tradeoffs. Acceptable Prediction Error (APE) in *Solves* function is the user demand in regression. Benefit increases with increase in APE. The top row shows this trend in a well defined setting (noise 2%) and the bottom row shows the same trend in a relatively ill defined setting (noise 5%).

6 Discussion

The idea of footprint set has been used extensively in the context of case base maintenance. In our work, we have brought out its connection to all knowledge containers through the *Solves* definition (Figure 1). Competence of a CBR system arises jointly from the vocabulary, case base, similarity and adaptation knowledge. Hence, maintenance of case-based reasoners needs to take into account the interaction between knowledge containers. Current research in CBR maintenance is skewed towards case base maintenance [Iglezakis and Roth-Berghofer, 2000]. Though case base is a central source of knowledge in a CBR system, maintenance of other knowledge containers is also equally important from the point of view of other performance objectives. [Portinale *et al.*, 1999] discusses the tradeoffs between performance goals in multi-modal diagnostic systems that combine case-based reasoning and model-based reasoning. [Leake and Wilson, 2000] highlights the importance of considering adaptation cost in case base maintenance and propose quantification metrics for the same. Mathew and Chakraborti [2017] propose a modification of Smyth's footprint algorithm [Smyth and McKenna, 1999] called footprint$_{CA}$ which accounts for single-case as well as compositional adaptation. We intend to use footprint$_{CA}$ for analysing knowledge tradeoffs in our future work.

Footprint size reduction is a unified measure that allows a maintenance engineer to compare design choices that are roughly equivalent in terms of competence. The idea can further be extended to compare design choices in terms of response time. It is well known that while knowledge rich systems like model-based or rule-based reasoners are relatively less suited for ill-defined domains, they can have faster response times compared to a case-based reasoner that operates over a large number of cases and a complex similarity measure. Since incorporating richer domain knowledge containers like adaptation and vocabulary knowledge can lead to a reduction in the effective case base size, the proposed measure can aid a maintenance engineer to explore options for trading off between generalization and (time) efficiency. We can also position model-based systems as an extremely special case of CBR systems where domain knowledge is so rich that we can entirely do away with cases. This idea of using the proposed measure to guide design choices demands more

elaboration. But, this paper restricts its scope to systematizing the analysis of cross-container knowledge shifts.

With respect to general Artificial Intelligence (AI) audience, our work reinforces the potential of using domain knowledge to effectively prune hypotheses spaces induced by inductive learners [Aamodt and Plaza, 2017]. Levesque [2014] argues that despite remarkable successes in data driven tasks, it is an illusion to believe that a silver bullet can solve all problems in AI. On similar lines, we envisage that in the near future, we would increasingly witness the need for top-down (knowledge based) approaches to complement bottom up (data-driven) approaches in solving real world problems. In this context, it is interesting in principle to quantify the impact of knowledge and this paper is a modest effort towards that initiative in a restricted (CBR) setting.

7 Conclusion

The central contribution of this paper is in proposing an approach that analyses the knowledge tradeoffs between containers in CBR. It uses the pivotal idea of reduction in footprint set size that is effected by each knowledge container. It also throws light on a way of empirically positioning domains in a spectrum ranging from ill-defined domains to well-defined ones. To the best of our knowledge, such a quantitative characterization has not been attempted before in the CBR community. We also studied the influence of user demands on such tradeoffs. We expect that our interpretation of footprint cases and knowledge transfers will streamline and motivate new avenues of cross-container maintenance activities in CBR. We also envisage that this line of thinking can be extended to closely examine the nature of data-knowledge tradeoffs in domains outside CBR as well.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59, 1994.

[Aamodt and Plaza, 2017] Agnar Aamodt and Enric Plaza. Case-based reasoning and the upswing of ai. 2017.

[Aha and Muñoz-Avila, 2001] David W Aha and Héctor Muñoz-Avila. Introduction: Interactive case-based reasoning. *Applied Intelligence*, 14(1):7–8, 2001.

[Chakraborti *et al.*, 2008] Sutanu Chakraborti, Ulises Cerviño Beresi, Nirmalie Wiratunga, Stewart Massie, Robert Lothian, and Deepak Khemani. Visualizing and evaluating complexity of textual case bases. In *European Conference on Case-Based Reasoning*, pages 104–119. Springer, 2008.

[Cheng and Hüllermeier, 2008] Weiwei Cheng and Eyke Hüllermeier. Learning similarity functions from qualitative feedback. *Advances in Case-Based Reasoning*, pages 120–134, 2008.

[Craw *et al.*, 2006] Susan Craw, Nirmalie Wiratunga, and Ray C Rowe. Learning adaptation knowledge to improve case-based reasoning. *Artificial Intelligence*, 170(16-17):1175–1192, 2006.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[Dheeru and Karra Taniskidou, 2017] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[Hanney and Keane, 1997] Kathleen Hanney and Mark T Keane. The adaptation knowledge bottleneck: How to ease it by learning from cases. In *International Conference on Case-Based Reasoning*, pages 359–370. Springer, 1997.

[Iglezakis and Roth-Berghofer, 2000] Ioannis Iglezakis and Thomas Roth-Berghofer. A survey regarding the central role of the case base for maintenance in case-based reasoning. In *ECAI Workshop Notes*, pages 22–28, 2000.

[Inselberg, 2009] Alfred Inselberg. Parallel coordinates: Visual multidimensional geometry and its applicationss, 2009.

[Kolodner, 1992] Janet L Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1):3–34, 1992.

[Lang, 1999] Ken Lang. 20 newsgroups data set, 1999.

[Leake and Wilson, 2000] David B Leake and David C Wilson. Guiding case-base maintenance: Competence and performance. In *Proceedings of the 14th European Conference on Artificial Intelligence Workshop on Flexible Strategies for Maintaining Knowledge Containers*, 2000.

[Levesque, 2014] Hector J Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[Mathew and Chakraborti, 2017] Ditty Mathew and Sutanu Chakraborti. Competence guided model for casebase maintenance. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4904–4908. AAAI Press, 2017.

[Portinale *et al.*, 1999] Luigi Portinale, Pietro Torasso, and Paolo Tavano. Speed-up, quality, and competence in multimodal case-based reasoning. *Lecture notes in computer science*, 1650:303–317, 1999.

[Richter, 1995] Michael M. Richter. The knowledge containers in similarity measures. slides of invited talk at the first international conference of case-based reasoning (iccbr-95), 1995.

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

[Smyth and McKenna, 1999] Barry Smyth and Elizabeth McKenna. Footprint-based retrieval. In *ICCBR*, volume 1650, pages 343–357. Springer, 1999.

[Stahl, 2001] Armin Stahl. Learning feature weights from case order feedback. In *International Conference on Case-Based Reasoning*, pages 502–516. Springer, 2001.