Evaluation of Axiom Selection Techniques

Qinghua Liu[a], Zishi Wu[b], Zihao Wang[b] and Geoff Sutcliffe[b]

[a]Southwest Jiaotong University, China
[b]University of Miami, USA

Abstract

"Large theory" problems in Automated Theorem Proving (ATP) have been defined as having many functions and predicates, and many axioms of which only a few are required for the proof of a theorem. One key to solving large theory problems is selecting a subset of the axioms that is adequate for finding a proof. The main contribution of this paper is metrics for evaluating axiom selection techniques without having to run an ATP system on the problem formed from the selected axioms and the problem's conjecture. This paper additionally presents three new axiom selection techniques. The new techniques, and the axiom selection in the Vampire ATP system, are evaluated using the metrics.

Keywords

Automated theorem proving, axiom selection

1. Introduction

"Large theory" problems in Automated Theorem Proving (ATP) have been defined [1] as having many functions and predicates, and many axioms of which only a few are required for the proof of a theorem. Large theory problems are often found in corpora that contain very many problems, e.g., the MPTP2078 corpus [2], the Mizar 40 corpus [3], and the GRUNGE corpus [4]. Large theory problems present challenges to ATP systems, mainly because of the large search space generated by the large number of axioms. Thus, one key to solving large theory problems is selecting a subset of the axioms that is adequate for finding a proof. There has been significant and successful research on this topic, e.g., [5, 6, 7, 8, 9, 10, 2, 11, 12]. Many techniques are based on the occurrences of symbols in the formulae, e.g., the SInE method [9] and its derivatives. The fact that large theory problems often occur in large corpora makes the application of machine learning techniques [13] viable, e.g., as in the MaLARea system [14].

Evaluation of axiom selection techniques is typically done by:

1. Choosing a corpus of large theory problems that are known to be theorems.
2. For each problem in the corpus, selecting a subset of the problem's axioms.
3. Running an ATP system on the reduced problem formed from the selected axioms and the problem's conjecture.

In the third step of this process, a proof indicates an adequate selection (the quality of which can then be evaluated), and a countermodel indicates an inadequate selection. As the development of axiom selection techniques progresses, this approach requires repeating steps 2 and 3 to evaluate

The 7th Workshop on Practical Aspects of Automated Reasoning, June 29-30 2020, Online, Earth
qhliu678@gmail.com (Q. Liu); ry04ert39@miami.edu (Z. Wu); zxw526@miami.edu (Z. Wang); geoff@cs.miami.edu (G. Sutcliffe)
https://www.cs.miami.edu/home/geoff/ (G. Sutcliffe)
0000-0003-2271-2669 (Q. Liu); 0000-0002-1039-4264 (Z. Wu); 0000-0002-5559-7452 (Z. Wang); 0000-0001-9120-3927 (G. Sutcliffe)

progress. Evaluation needs to be done on large corpora, making this a time consuming process. As a consequence, to move things along faster, it is often necessary to impose a small time limit on each ATP system run. Time limits (and other practical forms of incompleteness) cause timeouts, and a timeout in the third step provides no information - the selection might be inadequate, or the selection might be adequate but the reduced problem is too hard because too many (unnecessary) axioms were selected, or the time limit is too small. The results are also influenced by the choice of ATP system. Overall, this process for evaluating axiom selection techniques is problematic.

The main contribution of this paper is metrics for evaluating axiom selection techniques without having to run an ATP system on the reduced problems. While the "proof is in the pudding" and it is eventually necessary to evaluate by running an ATP system, the metrics described in this paper provide a first-pass evaluation that allows axiom selection techniques to be rapidly tested and refined. The approach has the advantage of being independent of a chosen ATP system. This paper additionally presents three new axiom selection techniques. The new techniques and the axiom selection [9] in the Vampire [15] ATP system are evaluated using the proposed metrics.

This paper is structured as follows: Section 2 describes the evaluation metrics. Section 3 describes a relatedness measure between formulae, and the three new axiom selection techniques that are based on that measure. Section 4 provides evaluation results, including an initial evaluation of the metrics themselves. Section 5 concludes.

2. Axiom Selection and the Evaluation Metrics

The axiom selection evaluation metrics measure how precisely a set of selected axioms matches a known minimally adequate set of axioms. Two types of axiom selection techniques are considered:

- *Ranking* techniques, which take the problem's axioms as input, rank them according to how likely they are judged to be necessary for a proof, and select the axioms that are ranked above some cut-off criterion. This technique is used in, e.g., Isabelle's Sledgehammer [16].
- *Projection* techniques, which take the problem's axioms as input and directly return a selection of axioms. (Ranking is thus a special case of projection.) This technique is used in, e.g., Vampire.

2.1. The MPTP2078 Problem Corpus

In this work the MPTP2078 corpus[1], based on the Mizar Mathematical Library [17], was used for development. The MPTP2078 has two versions of each of its 2078 problems: the *bushy* (small) versions that contain only the Mizar axioms that a knowledgeable user selected for finding a proof of the conjecture, and the *chainy* versions that contain all the axioms that precede the conjecture in the Mizar library order. The bushy problems have between 10 and 67 axioms, while the chainy versions have between 10 and 4563 axioms.

In order to extract minimally adequate sets of axioms for each problem, Vampire and E [18] were run on the problems, using the StarExec [19] Miami[2] cluster with a 300s CPU time limit. The computers have an octa-core Intel(R) Xeon(R) E5-2667 3.20GHz CPU, 128GB memory, and run the CentOS Linux release 7.4.1708 operating system. This produced proofs for 1486 of the bushy problems (1474 by Vampire and 1263 by E) and 1345 of the chainy problems (1333 by Vampire and 815 by E). For each proof found, the axioms used in the proof were extracted as an minimally adequate set of axioms, and

[1] https://github.com/JUrban/MPTP2078
[2] https://starexec.ccs.miami.edu

a new problem was formed from that minimally adequate set with the problem's conjecture. These new problems were dubbed the *pruney* problems. Additionally, in testing the new axiom selection techniques described in Section 3, some further different minimally adequate sets were found and further pruney problems were created. This resulted in a further 65 pruney problems for the bushy problems, and a further 24 pruney problems for the chainy problems. For some problems multiple minimally adequate sets of axioms were found, resulting in a total of 1829 pruney problems for the 1551 bushy problems, and a total of 3093 pruney problems for the 1369 chainy problems.

The pruney problems provide minimally adequate sets of axioms against which selected sets of axioms can be compared. The smallest fraction of axioms in a minimally adequate set ranges from 0.01 to 1.0 for the bushy problems, and from 0.0002 to 0.36 for the chainy problems. The average fractions are 0.29 and 0.01, respectively. These fractions show that precise axiom selection could significantly reduce the number of axioms that need to be used, which in turn normally significantly reduces the search space of an ATP system. As might be expected, the numbers are more extreme for the larger chainy problems.

2.2. The Metrics

Define the following basic measures:

- The *N*umber of *Ax*ioms in a *P*roblem: *NAxP*.
- The *N*umber of axioms *Sel*ected: *NSel*.
- The *N*umber of axioms *U*sed *i*n a *P*roof, i.e., the cardinality of a minimally adequate set: *NUiP*. *NUiP* is 0 if no axioms need to be used, i.e., the conjecture is a theorem.
- In a ranked list of axioms, the number of axioms down to the lowest ranked axiom in a minimally adequate set, i.e., the *N*umber of axioms *N*eeded from the *Ra*nking: *NNRa*.

The axiom selection evaluation metrics are listed below. In all cases the range is $[0.0, 1.0]$.

Precision. If the axiom selection technique selects an adequate set of axioms, i.e., a superset of one or more of the known minimally adequate sets of axioms, then:

- If the minimum *NUiP* = 0, and *NSel*= 0, then 1.00
- Else the maximum *NUiP*/*NSel*

If the selection technique selects an inadequate set then the precision is 0.00. Larger values are better. The intuition here is of a probability that using the selection will result in a proof (which is 0.00 if an inadequate set is selected).

Selectivity. *NSel*/*NAxP*. This measures the fraction of axioms selected. 1.00 results from selecting all the axioms - the base case. Smaller values are better, provided an adequate set of axioms is selected.

Ranking precision. (Applicable to only ranking techniques.) If the axiom selection technique selects an adequate set of axioms, then:

- If the minimum *NUiP* = 0, and *NSel*= 0, then 1.00
- Else the maximum *NNRa*/*NSel*

If the technique selects an inadequate set, then 0.00. This measures how precisely the technique chooses the "best ones" from the ranked list of axioms. Larger values are better.

Ranking density. (Applicable to only ranking techniques.) If *NUiP* = *NNRa* = 0, then 1.00, else *NUiP*/*NNRa*. This measures the quality of the ranking - axioms in a minimally adequate set should be ranked highly, which in turn allows a smaller number to be selected if the ranking precision is good. Larger values are better.

Average precision/selectivity/ranking precision/ranking density. For a set of problems, the average over the problems.

Adequacy. For a set of problems, the fraction of problems for which the axiom selection technique selects an adequate set of axioms. Larger values are better.

Adequate precision/selectivity/ranking precision/ranking density. For a set of problems, the average over the problems for which the axiom selection technique selects an adequate set of axioms.

3. Our Axiom Selection Techniques

The motivation for developing these metrics was based on a need to evaluate new axiom selection techniques being developed by the first three authors. These techniques are described in this section, and their performance according to the metrics is evaluated in Section 4. It turns out that a very simple technique performs surprisingly well on the MPTP2078 corpus.

All of the new techniques are based on how strongly two formulae are related (as is the case in many axiom selection techniques). In this work a novel measure of relatedness is used, which is described in Section 3.1. The individual new techniques are then described in Sections 3.2 to 3.4.

3.1. Formula Dissimilarity and Similarity

The relatedness between two formulae is computed first as a *dissimilarity* between the two formulae, which is also later converted to a *similarity*.

The dissimilarity between two terms or atoms is an extended version of the Hutchinson distance [20]. For two terms or atoms Δ_1 and Δ_2, their *least general generalization* $\Delta = lgg(\Delta_1, \Delta_2)$, if it exists, is a term or atom Δ such that there are substitutions θ_1 and θ_2, $\Delta\theta_1 = \Delta_1$ and $\Delta\theta_2 = \Delta_2$, and there is no term or atom Δ' and substitutions $\sigma, \sigma_1, \sigma_2$ such that σ is not just a renaming substitution, $\Delta\sigma = \Delta'$, $\Delta'\sigma_1 = \Delta_1$, and $\Delta'\sigma_2 = \Delta_2$. If $lgg(\Delta_1, \Delta_2)$ does not exist, e.g., neither Δ_1 nor Δ_2 are variables and their principal symbols are different, then the dissimilarity $dsim(\Delta_1, \Delta_2) = \infty$. Otherwise:

Divide θ_i into two parts, θ_i^v and θ_i^f:

$$\theta_i^v = \{X_{i,1} \mapsto Z_{i,1}, \ldots, X_{i,m_i} \mapsto Z_{i,m_i}\} \qquad \theta_i^f = \{Y_{i,1} \mapsto f_{i,1}, \ldots, Y_{i,n_i} \mapsto f_{i,n_i}\}$$

where $X_{i,j}$ and $Y_{i,j}$ are the substituted variables, $Z_{i,j}$ are substituting variables, and $f_{i,j}$ are substituting functional terms. Let

- w_v be a weight for variables (currently set to 1[3]),
- w_f be a weight function for non-variable symbols (currently set to 2 for all symbols),
- $V(X, \Delta)$ be the set of occurrences of the variable X in Δ,

[3]The values of 1 and 2 for w_v and w_f were adopted from their use in E, place greater emphasis on non-variable substitutions, and are small enough to avoid large dissimilarity values. The use of $ln(x + 1)$ for $g(x)$ below was motivated by it's common adoption as a continuous increasing function, to usefully dampen the impact of variable substututions on dissimilarity values.

- $N(X, \Delta)$ be the depth of a particular occurrence of the variable X in Δ, e.g., $N(X, p(X, f(X)) = 1$ for the first occurrence and 2 for the second occurrence,
- $W_v(X, \Delta)$ be $w_v \times (|V(X, \Delta)| + \sum_{U \in V(X,\Delta)} N(U, \Delta))$, i.e., the weighted sum of the number of occurrences of X in Δ and the depths of the occurrences of X in Δ, e.g., $W_v(X, p(X, f(X)) = w_v \times (2 + (1 + 2)) = 5$,
- $W_f(\Delta)$ be the sum of the weights (using w_f) of the non-variable symbols occurring in Δ,
- $g(x)$ be a continuous increasing function $\mathbb{R} \mapsto \mathbb{R}$ such that $g(0) = 0$ and $g(x_1 + x_2) \leq g(x_1) + g(x_2)$ for $x_1, x_2 \geq 0$ (currently set to $ln(x + 1)$).

Then, S_f and S_v are functions that map θ_i^v and θ_i^f to real numbers:

$$S_v(\theta_i^v) = \sum_{j=1}^{m_i} g(W_v(Z_{i,j}, \Delta_i) - W_v(X_{i,j}, \Delta)) \tag{1}$$

$$S_f(\theta_i^f) = \sum_{j=1}^{n_i} (|V(Y_{i,j}, \Delta)| \times W_f(f_{i,j})) \tag{2}$$

$S_v(\theta)$ measures the difference between the usage of substituting variables in Δ_i and the usage of substituted variables in Δ, over the substitutions made in Δ by θ_i^v. $S_f(\theta)$ measures the total function symbol weight of the substituting terms, over the substitutions made in Δ by θ_i^f. The dissimilarity between two terms or atoms Δ_1 and Δ_2 is then:

$$dsim(\Delta_1, \Delta_2) = \sqrt{[S_v(\theta_1^v) + S_v(\theta_2^v)]^2 + [S_f(\theta_1^f) + S_f(\theta_2^f)]^2} \tag{3}$$

The dissimilarity measures the combined substitution "effort" required to convert the least general generalization to those terms or atoms.

Let Φ_1 and Φ_2 be two formulae, containing the atoms $\{\Delta_{1,1}, \ldots, \Delta_{1,n}\}$ and $\{\Delta_{2,1}, \ldots, \Delta_{2,m}\}$ respectively. Let $S_{\neq\infty}$ be the set of pairs $(\Delta_{1,i}, \Delta_{2,j})$ for which $dsim(\Delta_{1,i}, \Delta_{2,j}) \neq \infty$. If $S_{\neq\infty} = \emptyset$, i.e., all pairs of atoms in Φ_1 and Φ_2 are infinitely dissimilar, then the dissimilarity $dsim(\Phi_1, \Phi_2) = \infty$. Otherwise:

$$dsim(\Phi_1, \Phi_2) = \frac{\sum_{(\Delta_{1,i}, \Delta_{2,j}) \in S_{\neq\infty}} dsim(\Delta_{1,i}, \Delta_{2,j})}{|S_{\neq\infty}|} \times \frac{n \times m}{|S_{\neq\infty}|} \tag{4}$$

The first term is the average dissimilarity between pairs of atoms that are not infinitely dissimilar. The second term penalizes the first by the excess of atom pairs whose dissimilarity is infinite. Intuitively, the dissimilarity between two formulae measures the extent to which inference between them might not be possible by virtue of the dissimilarity of their atoms.

For two formulae Φ_1 and Φ_2, their similarity $sim(\Phi_1, \Phi_2)$ is defined in the context of a set of formulae $\mathcal{F}$, $\Phi_1, \Phi_2 \in \mathcal{F}$:

$$maxdsim(\mathcal{F}) = max_{\phi_i, \phi_j \in \mathcal{F}}(dsim(\phi_i, \phi_j) \neq \infty) \tag{5}$$

$$sim(\Phi_1, \Phi_2, \mathcal{F}) = \max(0, maxdsim(\mathcal{F}) - dsim(\Phi_1, \Phi_2)) \tag{6}$$

If the dissimilarity is 0, then the similarity is the largest dissimilarity that is not ∞. If the dissimilarity is greater than 0 and not ∞, then the similarity is the difference between the largest dissimilarity that is not ∞ and the dissimilarity. If the dissimilarity is ∞ or is equal to the largest dissimilarity that is not ∞, then the similarity is 0.

3.2. Q∞

This axiom selection technique takes a problem consisting of a conjecture $\mathcal{C}$ and axioms $\mathcal{A}$, and selects all axioms $\Phi \in \mathcal{A}$ such that $dsim(\mathcal{C}, \Phi) \neq \infty$. This simply means that each axiom contains at least one atom whose predicate symbol matches that of an atom in the conjecture.

3.3. Spectral Clustering

This axiom selection technique views the formulae of a problem as nodes of a complete undirected graph, with the similarities between the formulae as the edge weights. Spectral clustering [21] is then used to cluster together similar nodes, and the axioms in the cluster containing the conjecture are selected. This process requires three steps: determining the number of clusters, choosing the initial centroids for the clusters, and applying k-means clustering.

The initial centroids for the clustering are extracted from a feature matrix consisting of the eigenvectors of the normalized graph Laplacian matrix [21], computed from the graph's adjacency matrix with edges weighted by similarity. The rows of the feature matrix corresponding to the conjecture node and to the $k - 1$ (see below for how k is computed) axiom nodes with highest degree centrality are used as the initial centroids for the clustering.

To determine the number of clusters for a problem in the MPTP2078 corpus, the three step process was applied to each problem with the number of clusters ranging from two to half the number of formulae in the problem. The precision was computed each time, and the best number of clusters recorded. A median regression line was fitted to this data (separately for the bushy and chainy problems), as shown in Figure 1. The upper two plots are for a problem set of 325 smaller problems (see Section 4 for details of the testing problem sets), and the lower plot is for the 1551 bushy problems. While the fit of the regression lines is awful, they were used to set the number of clusters k for problems in the corpus, based on the number of formulae in each problem. For the 1551 bushy problems it turns out that two clusters is always optimal, regardless of the number of axioms. It was not possible to compute a regression line for the 1369 chainy problems with the time and computing resources available.

3.4. Greedy Tree plus Nearest Neighbours

This axiom selection technique views the formulae of a problem as nodes of a complete undirected graph, with the similarities between the formulae as the edge weights. Figure 2 provides a simple illustrative example, in which C is the conjecture and As are the axioms (the figure omits edges that do not affect the example). Starting at the conjecture, a greedy tree is built by iteratively visiting all the axioms most similar to the current leaves of the tree, until axioms that have no similarity (or equivalently, infinite dissimilarity) to the conjecture are reached. These infinitely dissimilar nodes are ignored. In Figure 2 the thicker edges are those that are followed, and the axioms with thicker circles are those at which the tree growth stops because they have no similarity to the conjecture. At that stage the light grey axioms are in the tree. As a final step, all the unvisited axioms most similar to the axioms in the greedy tree are added to the tree. In Figure 2, that adds the dark grey axioms to the tree. The axioms in the tree are selected - those are the non-white axiom in the Figure 2..

4. Evaluation Results

The new axiom selection techniques and the axiom selection in the Vampire ATP system were evaluated using the metrics. No particular effort was moade to optimize the implementation of the new

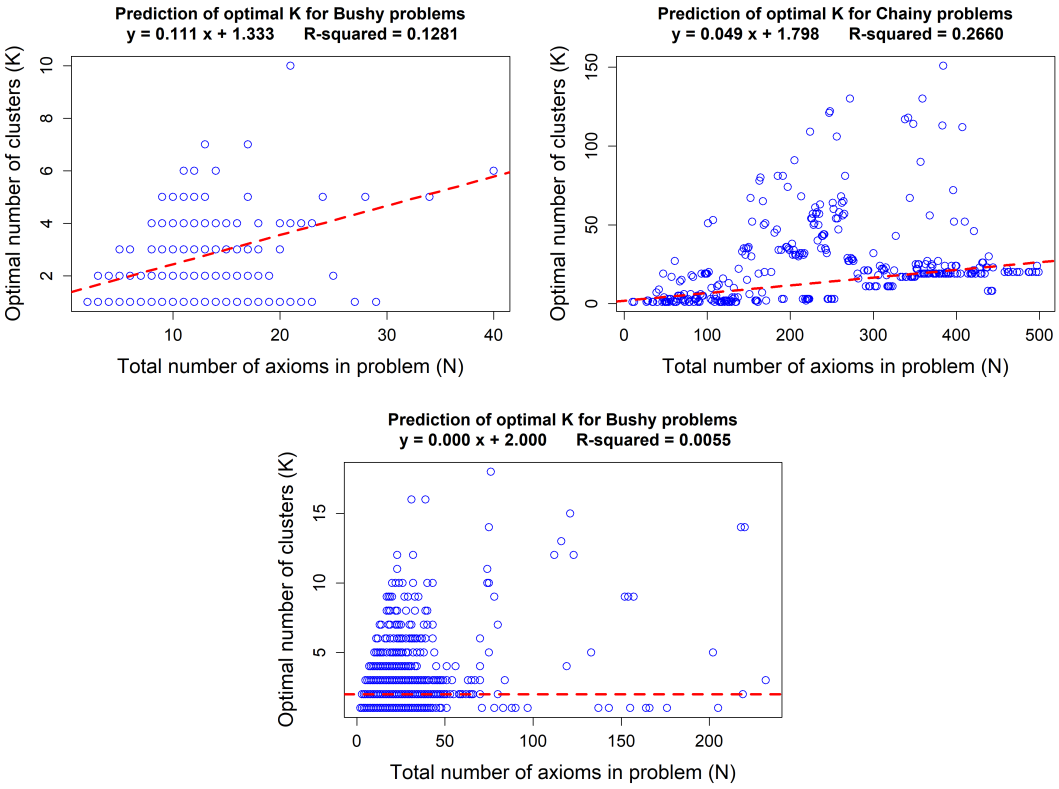

Figure 1: Best number of clusters vs. Number of formulae

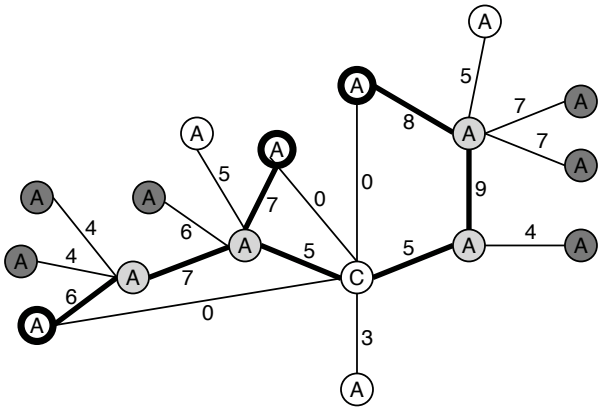

Figure 2: A greedy tree

selection techniques; rather, the evaluation aimed to demonstrate the usefulness of the metrics, and as such no CPU times are presented below. Two problem sets from each of the bushy and chainy parts of the MPTP2078 corpus were used. The first set was selected by taking the 325 chainy problems with less than 500 axioms and for which minimally adequate axiom sets of axioms (pruney problems) are known, and taking the corresponding 325 bushy problems. The second set was the 1551 bushy problems and 1369 chainy problems for which minimally adequate axiom sets (pruney problems) are known. The smaller set was useful for initial quick testing, and also necessary because some of the

325 Bushy problems	Average				Adequate				
Technique	Prcn	Sely	RPrn	RDen	Adeq	Prcn	Sely	RPrn	RDen
Base	0.35	1.00	-	-	1.00	0.35	1.00	-	-
Vampire 4.4	0.32	0.80	-	-	0.80	0.39	0.84	-	-
Q∞	0.43	0.54	0.62	0.52	0.74	0.58	0.61	0.84	0.71
Spectral Cl.	0.24	0.57	-	-	0.66	0.36	0.79	-	-
Greedy Tree+NN	0.36	0.57	-	-	0.66	0.36	0.79	-	-
325 Chainy problems	Average				Adequate				
Technique	Prcn	Sely	RPrn	RDen	Adeq	Prcn	Sely	RPrn	RDen
Base	0.06	1.00	-	-	1.00	0.06	1.00	-	-
Vampire 4.4	0.08	0.55	-	-	0.94	0.09	0.56	-	-
Q∞	0.08	0.53	0.57	0.21	0.85	0.09	0.56	0.67	0.25
Spectral Cl.	0.05	0.48	-	-	0.65	0.08	0.63	-	-
Greedy Tree+NN	0.05	0.79	-	-	0.86	0.06	0.85	-	-

Table 1
Results for the 325 smaller problems

1551 Bushy problems	Average				Adequate				
Technique	Prcn	Sely	RPrn	RDen	Adeq	Prcn	Sely	RPrn	RDen
Base	0.30	1.00	-	-	1.00	0.30	1.00	-	-
Vampire 4.4	0.21	0.69	-	-	0.58	0.35	0.76	-	-
Q∞	0.33	0.54	0.55	0.41	0.68	0.48	0.59	0.80	0.60
Spectral Cl.	0.25	0.69	-	-	0.76	0.33	0.83	-	-
1369 Chainy problems	Average				Adequate				
Technique	Prcn	Sely	RPrn	RDen	Adeq	Prcn	Sely	RPrn	RDen
Base	0.02	1.00	-	-	1.00	0.02	1.00	-	-
Vampire 4.4	0.03	0.40	-	-	0.87	0.04	0.41	-	-
Q∞	0.03	0.53	0.46	0.07	0.81	0.04	0.56	0.56	0.09

Table 2
Results for all problems with known minimally adequate axiom sets

new techniques could not analyse the chainy problems set with the time and computing resources available.

Tables 1 and 2 show the results, including a row for the base case in which all axioms are selected. The columns are the Precision (Prcn), Selectivity (Sely), Ranking precision (RPrn), Ranking density (RDen), Adequacy (Adeq), and the Adequate precision/selectivity/ranking precision/ranking density.

For the smaller set of 325 bushy problems Q∞ performs the best, with the highest precision and lowest selectivity. It also has a 74% adequacy. Moving up to the 325 chainy problems Q∞ and Vampire have the highest precision, with Q∞ having slightly higher selectivity, and Vampire better adequacy. For the 1551 bushy problems Q∞ is again the top performer, now in terms of precision, selectivity, and adequacy. Spectral clustering also performs reasonably well, with a better precision than Vampire, and with the cluster containing the conjecture (recall from Section 3.3 that two clusters is optimal, and hence used here) containing 69% of the axioms on average. Finally, for the 1369 chainy problems Q∞ and Vampire again have the highest precision, with Vampire having the best selectivity and adequacy. Overall, Q∞ performs the best on the bushy problems, with Vampire doing slightly better on the chainy problems. It is surprising that the very simple Q∞ technique performs as well as it does, when compared to the mature (and more complex) Vampire selection technique.

The adequacy numbers have to be viewed carefully - it is easy to get an optimal adequacy simply

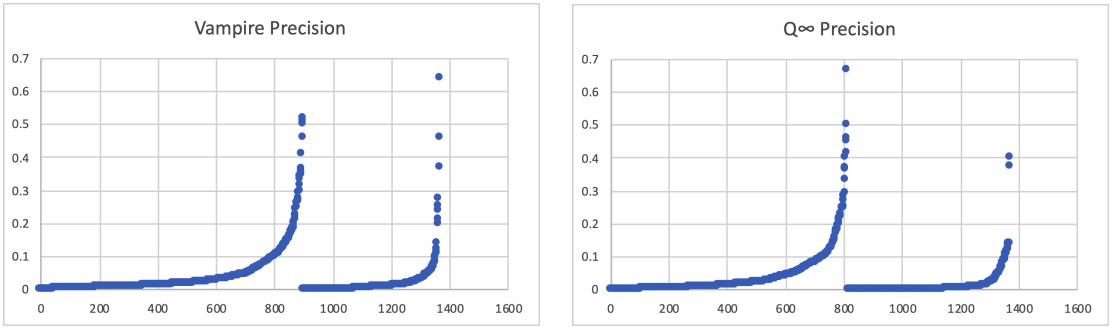

Figure 3: Distribution of precision values for chainy problems

by selecting all axioms and thus having pessimal selectivity. It is interesting to contrast the average and adequate precision values. The average precision for the base case is better than that of some of the selection techniques because of the zero precision assigned when an inadequate set of axioms is selected. An adequate precision close to the base case, as for, e.g., Spectral Clustering and Greedy Tree+NN on both sets of 325 problems, indicates that doing the axiom selection is probably wasted effort.

There is a correlation, but not a particularly strong correlation, between the number of axioms in a problem and the precision of the Vampire and Q∞ axiom selections, e.g., around 40% for the chainy problems. This suggests that the axiom selection techniques might scale well to larger problems.

Overall, the precision values are disappointingly low, especially for the chainy problems. For Q∞ the ranking density is similarly low, so that even with perfect ranking precision the raw precision cannot be very good. Figure 3 shows the distributions of the precision values for the chainy problems. In each plot the left set of points are for those problems where an adequate set of axioms was selected, and the right set of point are for those problems where an inadequate set of axioms was selected. For the majority of problems the precision in less that 0.1, i.e., less than 10% of selected axioms are used in a proof.

4.1. Evaluating the Metrics

As was noted in the introduction to this paper, while metrics such as those presented in this paper are useful for quick evaluation of axiom selection techniques, the “proof is in the pudding”. The quality of the metrics needs to be established by comparing them with eventual ATP system performance. An initial evaluation of the precision metric has been performed by using the Vampire and Q∞ axiom selections, and looking at the performance of E on the resultant reduced problems. This was done for the 1551 bushy problems and 1369 chainy problems for which there are known minimally adequate axiom sets, so that the precision could be computed. The testing was done with a 300s CPU time limit on computers with a dual-core Intel(R) Xeon(R) E5-2609 2.50GHz CPU, 8GB memory, and running the CentOS Linux release 7.6.1810 operating system. The results are shown in Table 3.

In all four cases the precision of the unsolved problems is much lower than for the solved problems - the precision metric aligns correctly with the solvability of the reduced problems. The numbers of problems with precision 0.00 and 1.00 also align with the numbers of solutions, with higher numbers of solved problems with precision 1.00, and higher numbers of unsolved problems with precision 0.00. Overall, these results indicate that the precision metric does correctly evaluate the quality of the axiom selection.

1551 Bushy problems	Vampire selection				Q∞ selection			
	Solved		Unsolved		Solved		Unsolved	
# Problems	848	(55%)	703	(45%)	969	(62%)	582	(38%)
Avg Prcn	0.35		0.03		0.48		0.07	
# Prcn 0.00	5	(1%)	646	(92%)	3	(0%)	487	(84%)
# Prcn 1.00	7	(1%)	0	(0%)	63	(7%)	0	(0%)
1369 Chainy problems	Vampire selection				Q∞ selection			
	Solved		Unsolved		Solved		Unolved	
# Problems	901	(66%)	468	(34%)	812	(59%)	557	(41%)
Avg Prcn	0.04		0.02		0.04		0.01	
# Prcn 0.00	19	(2%)	159	(34%)	11	(1%)	243	(44%)
# Prcn 1.00	0	(0%)	0	(0%)	0	(0%)	0	(0%)

Table 3
Evaluation of Vampire and Q∞ axiom selection wrt E

The attentive reader might have noticed that there are some solved problems with precision 0.00, which is weird because the precision should be 0.00 only if the selected axioms are not an adequate set (Section 2.2). This indicates that these experiments revealed some new minimally adequate sets of axioms for some problems in the MPTP2078 corpus, and thus that some further pruney problems need to be created.

5. Conclusion

The main contribution of this paper is metrics for evaluating axiom selection techniques without having to run an ATP system on the reduced problems. Three new axiom selection techniques have also been presented. Results from evaluating the new techniques and the axiom selection in the Vampire ATP system have been presented.

Future work includes a more comprehensive evaluation of the metrics, to confirm that the metrics do provide a meaningful evaluation of the axiom selection techniques. It will also be interesting to apply the metrics to the axiom selection techniques found in other ATP systems such as E, GKC [22], iProver [23], Leo-III [24], and MaLARea.

The results show that there is lots of room for further research and improvement in axiom selection. We are currently investigating the use of network propagation [25], adopted from the analysis of protein-protein interaction networks [26], for analysing the formula similarity graphs. One interesting possibility is pipelining axiom selection tools so that the first in the pipeline receives the original problem, it's output becomes the input to the second in the pipeline, and so on. Finally, the three new axiom selection techniques all build on the dissimilarity and similarity measures described in Section 3.1. Further development and improvement of those measures, or their complete replacement, would have consequential impacts downstream.

References

[1] G. Sutcliffe, The CADE-27 Automated Theorem Proving System Competition - CASC-27, AI Communications 32 (2020) 373–389.

[2] J. Alama, T. Heskes, D. Külwein, E. Tsivtsivadze, J. Urban, Premise Selection for Mathematics by Corpus Analysis and Kernel Methods, Journal of Automated Reasoning 52 (2014) 191–213.

[3] C. Kaliszyk, J. Urban, MizAR 40 for Mizar 40, Journal of Automated Reasoning 55 (2015) 245–256.
[4] C. Brown, T. Gauthier, C. Kaliszyk, G. Sutcliffe, J. Urban, GRUNGE: A Grand Unified ATP Challenge, in: P. Fontaine (Ed.), Proceedings of the 27th International Conference on Automated Deduction, number 11716 in Lecture Notes in Computer Science, Springer-Verlag, 2019, pp. 123–141.
[5] Y. Puzis, G. Sutcliffe, Y. Zhang, Y. Gao, Using Relevance Measures for Axiom and Lemma Selection in Automated Theorem Proving, in: V. Barr, Z. Markov (Eds.), Proceedings of the 17th International FLAIRS Conference, AAAI Press, 2004, p. Submitted.
[6] G. Sutcliffe, Y. Puzis, SRASS - a Semantic Relevance Axiom Selection System, in: F. Pfenning (Ed.), Proceedings of the 21st International Conference on Automated Deduction, number 4603 in Lecture Notes in Artificial Intelligence, Springer-Verlag, 2007, pp. 295–310.
[7] J. Meng, L. Paulson, Lightweight Relevance Filtering for Machine-generated Resolution Problems, Journal of Applied Logic 7 (2009) 41–57.
[8] D. Külwein, M. Cramer, P. Koepke, B. Schröder, Premise Selection in the Naproche System, in: J. Giesl, R. Haehnle (Eds.), Proceedings of the 5th International Joint Conference on Automated Reasoning, number 6173 in Lecture Notes in Artificial Intelligence, 2010, pp. 434–440.
[9] K. Hoder, A. Voronkov, Sine Qua Non for Large Theory Reasoning, in: V. Sofronie-Stokkermans, N. Bjœrner (Eds.), Proceedings of the 23rd International Conference on Automated Deduction, number 6803 in Lecture Notes in Artificial Intelligence, Springer-Verlag, 2011, pp. 299–314.
[10] D. Külwein, T. van Laarhoven, E. Tsivtsivadze, J. Urban, T. Heskes, Overview and Evaluation of Premise Selection Techniques for Large Theory Mathematics, in: B. Gramlich, D. Miller, U. Sattler (Eds.), Proceedings of the 6th International Joint Conference on Automated Reasoning, number 7364 in Lecture Notes in Artificial Intelligence, 2012, pp. 378–392.
[11] T. Gauthier, C. Kaliszyk, Premise Selection and External Provers for HOL4, in: X. Leroy, A. Tiu (Eds.), Proceedings of the 4th ACM SIGPLAN Conference on Certified Programs and Proofs, ACM Press, 2015, pp. 49–57.
[12] B. Piotrowski, J. Urban, ATPboost: Learning Premise Selection in Binary Setting with ATP Feedback, in: D. Galmiche, S. Schulz, R. Sebastiani (Eds.), Proceedings of the 9th International Joint Conference on Automated Reasoning, number 10900 in Lecture Notes in Computer Science, 2018, pp. 566–574.
[13] D. Külwein, J. Blanchette, A Survey of Axiom Selection as a Machine Learning Problem, in: S. Geschke (Ed.), Computability and Metamathematics : Festschrift Celebrating the 60th birthdays of Peter Koepke and Philip Welch, College Publications, 2014, pp. 1–15.
[14] J. Urban, G. Sutcliffe, P. Pudlak, J. Vyskocil, MaLARea SG1: Machine Learner for Automated Reasoning with Semantic Guidance, in: P. Baumgartner, A. Armando, G. Dowek (Eds.), Proceedings of the 4th International Joint Conference on Automated Reasoning, number 5195 in Lecture Notes in Artificial Intelligence, Springer-Verlag, 2008, pp. 441–456.
[15] L. Kovacs, A. Voronkov, First-Order Theorem Proving and Vampire, in: N. Sharygina, H. Veith (Eds.), Proceedings of the 25th International Conference on Computer Aided Verification, number 8044 in Lecture Notes in Artificial Intelligence, Springer-Verlag, 2013, pp. 1–35.
[16] L. Paulson, J. Blanchette, Three Years of Experience with Sledgehammer, a Practical Link between Automatic and Interactive Theorem Provers, in: G. Sutcliffe, E. Ternovska, S. Schulz (Eds.), Proceedings of the 8th International Workshop on the Implementation of Logics, number 2 in EPiC Series in Computing, EasyChair Publications, 2010, pp. 1–11.
[17] P. Rudnicki, An Overview of the Mizar Project, in: Proceedings of the 1992 Workshop on Types for Proofs and Programs, 1992, pp. 311–332.
[18] S. Schulz, S. Cruanes, P. Vukmirovic, Faster, Higher, Stronger: E 2.3, in: P. Fontaine (Ed.),

Proceedings of the 27th International Conference on Automated Deduction, number 11716 in Lecture Notes in Computer Science, Springer-Verlag, 2019, pp. 495–507.

[19] A. Stump, G. Sutcliffe, C. Tinelli, StarExec: a Cross-Community Infrastructure for Logic Solving, in: S. Demri, D. Kapur, C. Weidenbach (Eds.), Proceedings of the 7th International Joint Conference on Automated Reasoning, number 8562 in Lecture Notes in Artificial Intelligence, 2014, pp. 367–373.

[20] A. Hutchinson, Metrics on Terms and Clauses, in: M. van Someren, G. Widmer (Eds.), Proceedings of the 9th European Conference on Machine Learning, number 1224 in Lecture Notes in Artificial Intelligence, Springer-Verlag, 1997, pp. 138–145.

[21] U. von Luxburg, A Tutorial on Spectral Clustering, Statistics and Computing 17 (2007) 395–416.

[22] T. Tammet, GKC: a Reasoning System for Large Knowledge Bases, in: P. Fontaine (Ed.), Proceedings of the 27th International Conference on Automated Deduction, number 11716 in Lecture Notes in Computer Science, Springer-Verlag, 2019, pp. 538–549.

[23] K. Korovin, iProver - An Instantiation-Based Theorem Prover for First-order Logic (System Description), in: P. Baumgartner, A. Armando, G. Dowek (Eds.), Proceedings of the 4th International Joint Conference on Automated Reasoning, number 5195 in Lecture Notes in Artificial Intelligence, 2008, pp. 292–298.

[24] A. Steen, C. Benzmüller, The Higher-Order Prover Leo-III, in: D. Galmiche, S. Schulz, R. Sebastiani (Eds.), Proceedings of the 8th International Joint Conference on Automated Reasoning, number 10900 in Lecture Notes in Artificial Intelligence, 2018, pp. 108–116.

[25] A. Stojmirovic, Y. Yu, Information Flow in Interaction Networks, Journal of Computationl Biology 14 (2007) 1115–1143.

[26] P. Devkota, M. Danzi, V. Lemmon, J. Bixby, S. Wuchty, Computational Identification of Kinases that Control Axon Growth in Mouse, SLAS Discovery (2020) To appear.