Weakly Supervised Learning Algorithm to Eliminate Irrelevant Association Rules in Large Knowledge Bases

Bruno B. Cifarelli, Rafael G. L. Miani

Instituto Federal de São Paulo, Votuporanga, Brazil
brunocifarelli@gmail.com, rafael.miani@ifsp.edu.br

Abstract. The construction and population of large knowledge bases have been widely explored in the past few years. Many techniques were developed in order to accomplish this purpose. Association rule mining algorithms can also be used to help populate these knowledge bases. Nevertheless, analyzing the amount of association rules generated can be a challenge and time-consuming task. The technique described in this article aims to eliminate irrelevant association rules in order to facilitate the rules evaluation process. To achieve that, this article presents a weakly supervised learning technique to prune irrelevant association rules. The proposed method uses irrelevant rules already discovered in past iterations and prunes off those with the same pattern. Experiments showed that the new technique can reduce and eliminate the amount of rules by about 60%, decreasing the effort required to evaluate them.

Categories and Subject Descriptors: H.4 [**Information Systems**]: Data Mining—*Association Rules*; H.4 [**Information Systems**]: Databases—*Knowledge Base construction*; I.2 [**Computer Methodologies**]: Machine Learning—*Learning paradigms*

Keywords: Association Rules, Irrelevant Rules, Large Knowledge Base, Weak Supervision

1. INTRODUCTION

In the past few years many systems, such as Cyc [Matuszek et al. 2006], DBpedia [Bizer et al. 2009], YAGO [Suchanek et al. 2007] and NELL [Carlson et al. 2010] have been developed in order to construct and populate large knowledge bases (KBs). The main purpose of such systems is to create techniques to assist the KB extension, which can be performed by (i) populating the KB with instances or (ii) by increasing the relations between the KB categories.

This article used NELL's KB to perform its experiments. NELL (Never-Ending Language Learning) works extracting texts from the Web uninterruptedly, with the goal of populating and extending its own KB. NELL has an ontology composed of categories (domains), relations and their instances. CPL (Coupled Pattern Learner) [Carlson et al. 2009], CSEAL [Carlson et al. 2010], Prophet [Appel and Hruschka Jr 2011] and CL (Conversing Learning) [Pedro and Hruschka Jr 2012] are examples of NELL's components that assist the KB construction.

In previous work [Miani and Hruschka Junior 2015], an association rule (AR) [Agrawal et al. 1993] mining algorithm was used to help NELL's large-growing KB population. All relevant association rules were used to populate the KB. For example, imagine the association rule AR1. It states that if the athlete plays in team *Bears*, then he plays *football*. Automatically, the algorithm fills the knowledge base with the *football* (*sports*) category whenever it finds the *Bears* (*Team*) category in the data set, contributing to an increase in the KB and a decrease in the amount of missing values.

$$AR1\text{: } \textit{AthletePlaysForTeam (X, Bears)} \rightarrow \textit{AthletePlaysSport (X, football).}$$

Table I. Association Rules generated in **Experiment 1** with support 0.04

Number	Domains	Association Rule	Relevant
R1	*Athlete, sport → Athlete, league*	*athletePlaysSport(X, basktball) → athletePlaysLeague(X, nba)*	NO
R2	*Athlete, league → Athlete, sport*	*athletePlaysLeague(X, nba) → athletePlaysSport(X, basktball)*	YES
R3	*Athlete, sport → Athlete, league*	*athletePlaysSport(X, football) → athletePlaysLeague(X, nfl)*	NO
R4	*Athlete, league → Athlete, sport*	*athletePlaysLeague(X, nfl) → athletePlaysSport(X, football)*	YES
R5	*Athlete, sport → Athlete, league*	*athletePlaysSport(X, hockey) → athletePlaysLeague(X, nhl)*	NO
R6	*Athlete, league → Athlete, sport*	*athletePlaysLeague(X, nhl) → athletePlaysSport(X, hockey)*	YES

However, association rule algorithms usually bring a large quantity of rules, and the effort spent on analyzing each one can be an exhausting task. Imagine that an association rule mining algorithm generated 1000 rules. If no automatic technique is used to evaluate or to remove redundant or irrelevant rules, a large amount of time is necessary to evaluate them. Therefore, the goal of this article is to decrease the number of association rules in order to facilitate the evaluation step. Thus, this article presents a weakly supervised learning algorithm to automatically eliminate irrelevant association rules. It is called weakly supervised learning as it requires minimal external supervision. The proposed approach eliminates rules that have the same categories of irrelevant rules already discovered before in other algorithms' iterations. This procedure contributes to a decrease in the number of rules and is the main contribution of this article.

Consider Table I as an example, which brings some association rules (third column) and the categories of such rule (second column). Imagine that rules *R1* and *R2* were generated in the first iteration of the algorithm. They were analyzed and classified as irrelevant (*R1*) and as relevant (*R2*). Then, *R2* is used to populate NELL's KB and *R1* is stored as irrelevant and will be used in other iterations to eliminate irrelevant rules with the same categories. *R1* has *athlete* and *sports* as categories in the antecedent side of the rule, and *athlete* and *league* in the consequent side. This is the only time a rule containing those categories is analyzed. In Table I, *R3* and *R5* have the same categories as *R1* and the proposed method will automatically prune those rules from the final set of generated rules. If rules from 3 to 6 were discovered in a second iteration, for example, only rules 4 and 6 would be analyzed, reducing by 50% the effort required on evaluate them. Experiments performed showed that the number of rules decreased by about 60% with the new approach.

Therefore, the main contribution of this article is a new technique to automatically remove irrelevant association rules with minimal supervision. The remainder of this article is organized as follows: Section 2 brings some related work. The complete description of the new technique is in Section 3 in the *Irrelevant Rules Elimination with Weak Supervision* subsection. Section 4 depicts the experiments.

2. RELATED WORK

The problem of large knowledge bases construction has been widely explored in the past few years. Many systems, such as Cyc [Matuszek et al. 2006], DBpedia [Bizer et al. 2009], YAGO [Suchanek et al. 2007] and NELL [Carlson et al. 2010] were developed in order to construct these knowledge bases.

Association rules [Agrawal et al. 1993] can also be a useful technique to assist KB construction. AMIE [GalÁrraga et al. 2013] is a generalized association rule mining algorithm [Srikant and Agrawal 1995] which works under incomplete evidences in ontological knowledge bases. AMIE extracts association rules like *motherOf(m,c)* $\wedge$ *marriedTo(m,f)* $\rightarrow$ *fatherOf(f,c)*. In [Miani and Hruschka Junior 2015], association rules were used to help populate NELL's KB, generating rules like the *AR1* in the previous section, which is different from AMIE that uses generalized association rules. Wahyudi et al.

[2019] made use of Graph-Pattern Association Rules (GPARs) [Fan et al. 2015] to extract rules from YAGO knowledge base. Their research results in 1114 association rules, which can be used to predict new facts that are not in the dataset.

However, a tricky problem with association rule mining algorithms is how to evaluate the large amount of generated rules. In this way, great efforts have been made to develop techniques (at the pre-processing or post processing step of the algorithm) to decrease the number of association rules, eliminating those considered irrelevant or redundant. The weakly supervised learning algorithm runs at the post processing step.

Some approaches constructed methods to eliminate redundant item sets, such as *Closed Itemsets* [Pasquier et al. 1999] and *Maximal Itemsets* [Burdick et al. 2001], for example. Zaki [2000] used the concept of frequent closed itemsets, reducing the number of rules without any loss of information. CHARM is an efficient algorithm for mining closed itemsets, which enumerates closed sets using a dual itemset-tidset search tree. It also uses a fast hash-based approach to remove any "non-closed" sets found during computation [Zaki and Hsiao 2002]. *A-Close* is an algorithm proposed by [Pasquier et al. 1999] that also decreases the number of rules without any loss of information, reducing the algorithm computation cost. *A-Close* is one of the best known algorithms to mine frequent closed itemsets. Xiao and Hu [2019] developed MRClose, a frequent closed itemset algorithm that uses the MapReduce technique under big data. The authors justified the use of a parallel method such as MapReduce as most frequent closed itemset algorithms can no longer meet the requirements of big data mining.

As with the closed itemset technique, Maximal Frequent Itemset algorithms are implemented in the candidate generation step of the algorithm. A Maximal Frequent Itemset (MFI) is a frequent itemset X with no frequent super set of X [Burdick et al. 2001]. The authors developed MAFIA, an algorithm to mine MFIs, which integrates a variety of algorithm ideas into a practical algorithm. FPMax [Grahne and Zhu 2003] uses a FP-tree structure to store the frequency information of the whole dataset. To test if a frequent itemset is maximal, another tree structure, called a Maximal Frequent Itemset tree (MFI-tree), is utilized to keep track of all MFIs. In [Sinthuja et al. 2019], a maximal frequent itemset mining algorithm based on linear prefix-tree was proposed. This approach is grounded on array and there is no need for each node to carry a pointer. It thus consumes a lower amount of memory and has a better run time if compared to other algorithms. *GenMax* [Gouda and Zaki 2005] and *MaxMiner* [Bayardo Jr 1998] are also well-known algorithms in this field.

At the post-processing step many techniques were developed to prune the number of redundant and irrelevant rules. Marinica and Guillet [2010] proposed an interactive approach where human domain experts filter the rules extracted in order to decrease the amount of association rules. CoGAR [Baralis et al. 2012] is a generalized association rule algorithm that introduced two new measures: (i) a schema constraint is created by an analyst and drives the itemset mining phase and (ii) opportunistic confidence constraint that identifies significant and redundant rules at the post-processing phase.

PNAR_IMLMS [Swesi et al. 2012] and MIPNAR_GA [Rai et al. 2014] are algorithms that discover both positive itemsets (frequent itemsets) and negative itemsets (infrequent itemsets). They created some measures to prune rules and generate positive and negative association rules. Dong et al. [2019] explored redundancy when mining positive and negative association rules (PNARs). They analyzed what kinds of PNARs are redundant and then proposed a method called LOGIC by using logical reasoning to prune redundant PNARs.

Djenouri et al. [2014] explored meta-rule extraction in order to prune irrelevant rules. First, they cluster association rules for large data sets. Then, different dependencies between rules of the same cluster are extracted using meta-rule algorithms, and the prune algorithm uses these dependencies to delete the deductive rules and keep just the representative rules for each cluster. The PVARM algorithm was proposed by [Rameshkumar et al. 2013]. It used the n-cross validation technique to

Table II. Comparison Among Algorithms

Algorithm / Technique	Large KBs	Association Rules	Prune Itemsets	Prune Association Rules
AMIE	X	X		
GPARs	X	X		
Closed and Maximal itemsets algorithms		X	X	
CoGar, PNAR_IMLMS, PVARM		X		X
NEW TECHNIQUE	X	X		X

reduce the amount of irrelevant association rules.

Miani and Hruschka Jr [2018] introduced the concept of super (sub) consequent rules and super (sub) antecedent rules. Basically, the method removes irrelevant super consequent rules discovered based on a sub consequent rule considered irrelevant in a previous iteration. The algorithm also eliminates super antecedent association rules considered redundant. With both methods combined, it was possible to reduce the amount of extracted rules by more than 30% without any loss of information. The main difference between this and the new approach is that in [Miani and Hruschka Jr 2018] association rules with items are eliminated, without considering the categories of the rule. This is better explained in the Rules Elimination subsection.

The algorithm introduced in this article made use of the following topics depicted in this section:

—Large knowledge bases;

—Association rules;

—Pruning of irrelevant association rules.

In a nutshell, the new algorithm eliminates association rules obtained from a large knowledge base at post-processing step, with minimal supervision. Table II provides a comparison of some of the techniques described in this section and the new technique. In the first column of the table is the algorithm (the approach introduced in this article is referred as NEW TECHNIQUE) name. In the first line of Table II is the technique explored by the algorithms. An *X* mark in the table indicates that an algorithm has accomplished a specific technique.

3. METHODOLOGY

This section describes the association rule mining algorithm used, with emphasis on the *Irrelevant Rules Elimination with Weak Supervision* method proposed in this article to eliminate irrelevant association rules discovered using NELL's large-growing knowledge base.

Figure 1 illustrates the complete system architecture. The new approach developed in this article is highlighted in the Rules Generation step. To sum up, a subset of NELL's KB is extracted and used as an input to the association rule mining algorithm to extract new patterns. The association rule mining algorithm used is based on NARFO [Miani et al. 2009]. NARFO was already modified in previous work to deal with missing values [Miani and Hruschka Junior 2015] and to eliminate redundant and irrelevant association rules [Miani and Hruschka Jr 2018]. This algorithm was first selected once it was an available algorithm and it has some important characteristics that are helpful to this article. NARFO is based on *Apriori*, but it can also navigate through an ontology structure, which is very important in this article to identify the correspondent domain of an item in the data set.

After data preparation, the algorithm begins generating candidate itemsets. In this step, a missing value treatment is performed due to an incomplete KB. Then, the frequent itemsets are used to generate association rules. The set of rules with confidence value greater than the *minimum confidence*

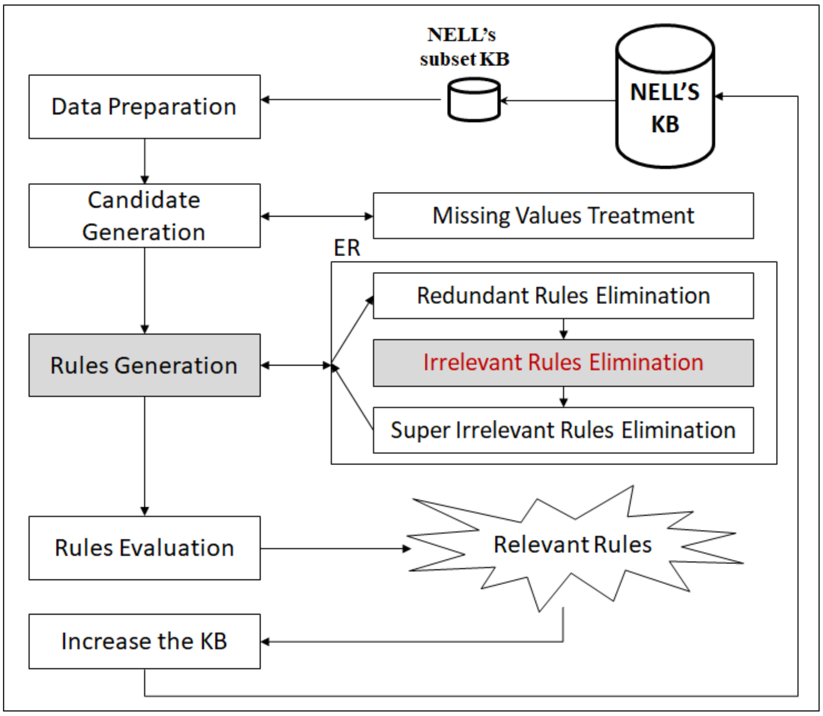

Fig. 1. System Architecture

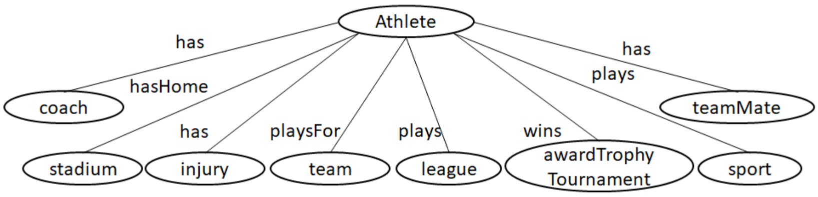

Fig. 2. NELL's KB subset

defined are used by the ER component. ER has three modules in order to remove association rules: (i) *Redundant Rules Elimination*, (ii) *Irrelevant Rules with Weak Supervision* (in which is the main contribution of this article) and (iii) *Super Consequent Irrelevant Rules Elimination*. The final set of rules (the ones that were not eliminated) are evaluated and the relevant rules are used to increase NELL's KB.

3.1 Data Preparation

Figure 1 shows the system architecture. As can be seen, a subset of NELL's KB is selected. Figure 2 presents the chosen subset of NELL's ontology, which is used to describe and explain the algorithm steps and it is also the subset used in the experiments. This subset is composed of a sports domain. For each athlete in the data set, each item corresponding to the respective domain (sports or league, for example) to that athlete is filled. If the value is not discovered yet by NELL's components, it is filled with a missing value, which is represented by *mv* in this article.

3.2 Candidate Generation

This step is similar to *Apriori.* After a proper data preparation, the algorithm begins looking for frequent itemsets, *i.e.*, those having their support value greater than the *minimum support* defined. All frequent itemsets will be considered to generate association rules.

However, the algorithm is performed in a large-growing KB, resulting in a missing value data set. To perform an association rule mining algorithm in such an environment, this problem needs to be tackled with care. In previous work [Miani and Hruschka Junior 2015], a new measure, called *MSC*, was developed to deal with missing values. To sum up, *MSC* discards an itemset, during support calculation, if all items of the categories in the itemset are missing. This results in itemsets with higher support values, which could generate more helpful association rules to fill NELL's KB.

3.3 Rules Generation: ER Component

This section describes the main contribution of this article in the *Irrelevant Rules Elimination with Weak Supervision* subsection. Like the *Apriori* algorithm, rules are generated based on all frequent itemsets, combining all possibilities of antecedent/consequent. Rules that have their confidence value higher than the *minimal confidence* defined are generated.

Unfortunately, association rule algorithms usually produce a large amount of rules. In addition, due to the incomplete data of large-growing knowledge bases like NELL, many rules generated may be irrelevant and/or redundant. In [Miani and Hruschka Jr 2018], an ER (Rules Elimination) component was developed to deal with such problems. ER had two modules to decrease the number of rules: (i) Redundant Rules Elimination and (ii) Super Consequent Irrelevant Rules Elimination, which cut out redundant and irrelevant rules, respectively.

In this article, a new module that eliminates irrelevant rules with minimal supervision was added to the ER component. Basically, the new procedure gets the set of rules after redundant rules elimination and prunes those irrelevant ones with the same pattern as another irrelevant rules discovered before. This is better explained in the *Irrelevant Rules Elimination with Weak Supervision* subsection. After that, the *Super Consequent Irrelevant Rules Elimination* method is used to remove other kinds of not useful rules. Experiments showed that executing the new approach before the other irrelevant rules technique generates fewer rules than the contrary, without any loss of information.

To sum up, the ER component has three modules:

(1) Redundant Rules Elimination;
(2) Irrelevant Rules Elimination with Weak Supervision;
(3) Super Consequent Irrelevant Rules Elimination.

Sub components 1 and 3 were already developed in [Miani and Hruschka Jr 2018], and the module that eliminates irrelevant rules with weak supervision is the main contribution of this article.

3.3.1 *Redundant Rules Elimination.* After generating association rules, ER checks for redundant and irrelevant rules. The *Redundant Rules Elimination* method gets the set of rules generated and tries to remove redundant *super antecedent rules.*

A rule X is a *super antecedent rule* of a rule Y if both rules have the same consequent and the antecedent of X is a super set of Y.

Consider Table III, which contains some examples of association rules. Rule number 3 is considered a *super antecedent rule* of rules 1 and 2. Notice that all three rules have the same consequent (*Football*) and all itemsets of the antecedent of rule 3 (*super_ bowl, nfl*) are in rules 1 and 2. The antecedent

Table III. Generated Association Rules

Number	Association Rule
1	*athleteWinsAwardTrophyTournament(X, super_ bowl) → athletePlaysSport(X, Football)*
2	*athletePlaysLeague(X, nfl) → athletePlaysSport(X, Football)*
3	*athleteWinsAwardTrophyTournament(X, super_ bowl), athletePlaysLeague(X, nfl) → athletePlaysSport(X,Football)*
4	*athletePlaysSport(X, Football) → athletePlaysLeague(X, nfl)*
5	*athletePlaysSport(X, Basketball) → athletePlaysLeague(X, nba)*
6	*athletePlaysSport(X, Football) → athleteWinsAwardTrophyTournament(X, super_ bowl)*
7	*athletePlaysSport(X, Tennis) → athleteWinsAwardTrophyTournament(X, us_ open)*
8	*athletePlaysSport(X, Football) → athletePlaysLeague(X ,nfl), athletePlaysForTeam (X, giants)*
9	*athletePlaysSport(X, Football), athletePlaysForTeam (X, giants) → athletePlaysLeague(X, nfl)*

of rule 3 is a super set of the antecedent of rules 1 and 2. The algorithm removes a *super antecedent rule* if all its *sub antecedent rules* are generated, avoiding loss of information in the process of filling the KB.

Sub antecedent rules are a more efficient way of populating a KB because fewer items in the antecedent size are required to fill the data set with the consequent items. For rules 1 and 2, only necessary the items *super_ bowl* or *nfl* are requeiredin the data set to fill the sports category with *football*. The strategy of filling the KB using association rules is better described in the *Increase the KB* subsection.

3.3.2 *Irrelevant Rules Elimination with Weak Supervision.* This subsection describes the main contribution of this article. The goal is to automatically remove irrelevant rules, based on the set of irrelevant already discovered in past iterations of the algorithm. The method uses weak learning supervision, because it is only necessary to analyze new rules that do not have same patterns as the irrelevant rules already discovered.

Let's take Table III as an example. Consider that rule 4 was discovered in a previous iteration, for example in the first iteration of the algorithm. This rule has *(athlete, sports)* in the antecedent and *(athlete, league)* in the consequent side. The first time this kind of rule was generated, it had to be analyzed, because there were no discovered rules with those domains before. The association rule generated in this article can be evaluated manually or by Conversing Learning (a NELL component). In this article, Conversing Learning was used to evaluate unknown rules.

CL classified rule 4 as irrelevant, as not all athletes who play *Football* also play the *nfl* league. In this way, in future algorithm's iterations, this rule will be used to eliminate irrelevant association rules that have the same domains (in this case*(athlete, sports)* and *(athlete, league)* in the antecedent and consequent side, respectively). It can be noticed in Table III that rule 5 has the same antecedent/consequent as rule 4. Thus, the algorithm automatically classified it as irrelevant, which contributes to a decrease in the number of rules to be analyzed.

The same behavior happens with rules 6 and 7. If rule 6 was discovered in an iteration before rule 7, it would be used to remove rule 7 from the set of rules. Rule 6 has *(athlete, sports)* in the antecedent and *athlete, TrophyTournament* in the consequent. As rule 7 has the same categories in both antecedent/consequent sides of the association rule, the algorithm, automatically, removes it from the final set of rules. It is important to notice that an irrelevant rule had to be generated in a previous iteration so the algorithm can use it to eliminate others with the same pattern. Only rules with different patterns from those already discovered need to be evaluated, requiring minimal external evaluation. This is why it is called weakly supervised learning algorithm.

Algorithm 1 shows the pseudo code of this procedure. It uses the set of rules generated after removing the redundant ones as input. For each rule, it checks if there is a correspondent irrelevant rule with the same domains as the current rule. If so, this rule is not added to the set of rules without irrelevant rules. Otherwise, the algorithm adds the current association rule to the set of generated rules that will be used in the last component that eliminates irrelevant *super consequent rules.*

Algorithm 1 Eliminating Irrelevant Rules with Weak Supervision

```
$allDomains = false$;
for $i = 0$ to $numberOfNonRedundantRules - 1$ do
  $currentRule = getAssociationRule(i)$;
  $allDomains = findCorrespondentIrrRule(currentRule)$;
  if NOT($allDomains$) then
    $finalRules.add(currentRule)$;
  end if
  $allCombinations = false$;
end for
```

3.3.3 *Super Consequent Irrelevant Rules Elimination.* The set of association rules obtained after removing the irrelevant ones using weak supervision is used as an input to this step. The procedure consists of eliminating irrelevant *super consequent rules* based on an irrelevant *sub consequent rules* already discovered before.

A rule X is a *super consequent rule* of Y, if both have the same antecedent items and the consequent of Y is a subset of the consequent of X. Also, Y is a *sub consequent rule* of X.

In Table III, rule 8 is a *super consequent rule* of rule 4. Both have the same antecedent (*Football*) and the consequent of the association rule 4 is a subset of rule 8's consequent.

Considering Table III and the 3 techniques of the ER component combined, rules 3, 5, 7 and 8 would be automatically removed. In this example, this contributes to a decrease of almost 50% in the amount of rules, which in turn reduces the effort required to evaluate them.

3.4 Rules Evaluation

The ER component has modules that remove redundant and irrelevant rules. After its execution, the final set of association rules is ready to be analyzed. It can still contain new redundant and irrelevant rules as well as relevant rules that need evaluation.

As mentioned before, there are two common ways to analyze them:

(1) Conversing Learning;

(2) Manually.

Conversing Learning is a NELL component that uses Twitter and Yahoo Answers to validate some patterns and relations discovered by other NELL components. In this article, only Twitter was used to evaluate produced rules. Basically, the rules generated are displayed in NELL's Twitter to be analyzed by its users.

If CL is not available, each rules has to be checked manually.

3.5 Increase the KB

This section describes the strategy to fill NELL's KB using the relevant association rules generated. After evaluating the rules produced, for each useful rule, the algorithm checks the KB subset looking at the values in the antecedent side of the rule, filling the instance with the values in the consequent side if they are still missing. For example, consider rule 2 in Table III, the algorithm will fill the KB with *Football* (consequent) in sports category every time it finds *nfl* (antecedent) in the league category.

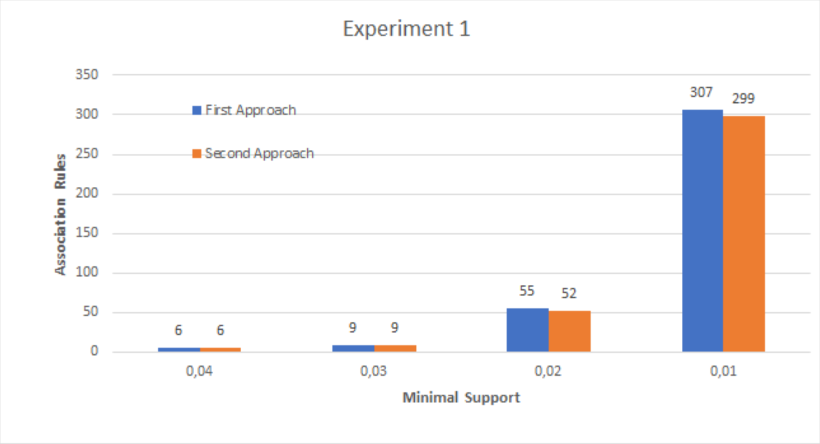

Fig. 3. Experiment to check the order of methods

4. EXPERIMENTS

This section describes the experiments performed in order to verify the results obtained due to the addition of the *Irrelevant Rules Elimination with Weak Supervision* method in the ER component, which was introduced and is the main contribution of this article. A real data set, extracted from NELL's KB, was used in Figure 2. It contains data related to sports, having domains like *athletes*, *sports*, *teams*, *leagues* and other sports categories.

The *minimum confidence* was set to 0.3 and the *minimum support* was set from 0.04 to 0.01, decreasing 0.01 in each algorithm iteration. The *minimum support* has low values because (i) the KB characteristic has lots of missing values, and (ii) when it was set to 0.05 no association rule was discovered.

The experiments were executed with three iteration cycles. Each cycle was performed with *minimum support* set from 0.04 to 0.01 as mentioned before, and used a data set of NELL's KB with additional new facts in each data set, to simulate a growing knowledge base. Consider *data set 1*, *data set 2* and *data set 3* the ones used in cycle 1, 2 and 3, respectively. Thus, three experiments were performed:

—**Experiment 1**: data set 1 and cycle 1;
—**Experiment 2**: data set 2 and cycle 2;
—**Experiment 3**: data set 3 and cycle 3.

In each experiment, the number of rules extracted was compared in the following aspects:

(1) Generated rules without any redundancy/irrelevancy treatment;
(2) Generated rules with redundant and super consequent irrelevant rules elimination;
(3) Generated rules with irrelevant rules elimination with weak supervision (complete ER).

The first aspect compared displays the number of association rules extracted without any redundancy or irrelevant treatment, *i.e.*, without any of ER components. The second one shows the reduction of rules after ER implementation to eliminate redundant and irrelevant rules, but without the irrelevant rules elimination with weak supervision method. The last one aims to show the reduction of rules after introducing the approach described in this article.

Before performing the experiments, two tests were executed to establish the best order in which to apply the methods to obtain maximum efficiency in reducing the amount of generated rules: (i) running

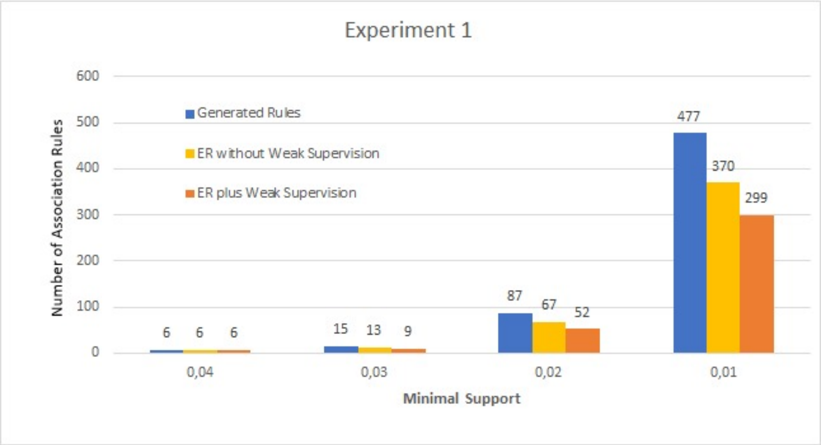

Fig. 4. Generated Rules by **Experiment 1**

the algorithm with the *Irrelevant Rules Elimination with Weak Supervision* as the last technique to eliminate rules and (ii) performing the algorithm using the new method between the redundant and the super consequent irrelevant rules elimination component. In Figure 3 it can be noticed that the second approach generated fewer rules than the first one, and it was therefore chosen to realize the experiments. The same behavior described in Figure 3 was repeated with data sets used in **Experiments 2** and **3**.

Figure 4 shows the results of **Experiment 1**. The first column on the left describes the amount of rules discovered without any redundancy or irrelevant treatment. The second one shows ER without the new approach, and the last column gives the number of rules using the complete ER. It can be observed that, as expected, more rules were extracted as the minimum support value decreased. However, using ER plus the method with weak supervision described in this article, the reduction of rules was more than 35% in comparison to the number of rules with no redundancy or irrelevancy treatment, and about 20% if compared to ER without weak supervision.

In Figure 4, the amount of rules generated is the same with minimum support 0.04 in all scenarios, even applying the complete ER. There are two possible reasons for this: (i) the irrelevant methods of ER only take effect if there were previous known irrelevant rules and (ii) the redundant technique eliminates redundant super antecedent rules. All association rules extracted in the first iteration of **Experiment 1** had only one itemset in the antecedent side and there was no known irrelevant association rule. In this case, all association rules needed to be evaluated by Conversing Learning. Those considered irrelevant must be used by the irrelevant modules of ER in futures iterations of the algorithm. Nevertheless, the relevant ones are used to help populate NELL's KB.

Table IV has all discovered association rules in **Experiment 1** with minimum support 0.04. The second column represents the rule's domains. The association rule column brings all generated rules, and the last one indicates if the rule is relevant or not after CL evaluation. In Table IV, rules 1, 2 and 5 were evaluated as irrelevant. These were used by all other iterations of the algorithm and those with the same domains were automatically pruned off the final set of rules. For example, consider rules 1 and 2 of Table IV, they have athlete, sport in the antecedent and athlete, league in the consequent side. All rules with these domains will be automatically removed in other iterations by the *Irrelevant Rules Elimination with Weak Supervision* method. The minimal supervision is due to the fact that only new association rules need evaluation. It is important to emphasize that the irrelevant association rules with those items were also considered to eliminate *super consequent irrelevant* rules.

In addition, all relevant rules discovered had only one itemset in the antecedent side. Thus, the ER component was not able to eliminate any redundant rules.

Table IV. Association Rules generated in **Experiment 1** with supprt 0.04

Number	Domains	Association Rule	Relevant
1	*Athlete, sport → Athlete, league*	*athletePlaysSport(X, basktball) → athletePlaysLeague(X, nba)*	NO
2	*Athlete, sport → Athlete, league*	*athletePlaysSport(X, football) → athletePlaysLeague(X, nfl)*	NO
3	*Athlete, league → Athlete, sport*	*athletePlaysLeague(X, nba) → athletePlaysSport(X, basktball)*	YES
4	*Athlete, league → Athlete, sport*	*athletePlaysLeague(X, nfl) → athletePlaysSport(X, football)*	YES
5	*Athlete, sport → Athlete, awardtrophytournament*	*athletePlaysSport(X, tennis) → athletePlaysLeague(X, us_open)*	NO
6	*Athlete, awardtrophytournament → Athlete, sport*	*athletePlaysSport(X, us_open) → athletePlaysLeague(X, tennis)*	YES

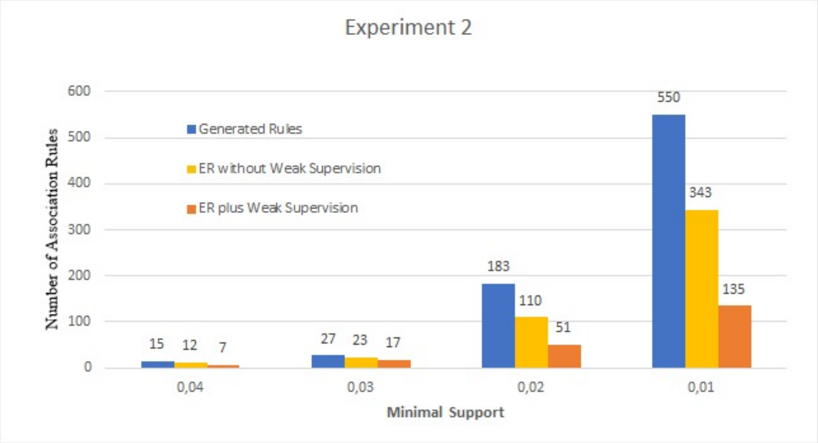

Fig. 5. Generated Rules by **Experiment 2**

Figure 5 shows the results of **Experiment 2**. As mentioned above, in **Experiment 2** the data set was increased by, approximately, 10%, which might have resulted in more generated rules. All irrelevant rules discovered before by ER methods were used in this experiment in order to remove new irrelevant ones. Those irrelevant rules discovered and evaluated by CL in the previous experiment were also used in **Experiment 2**.

It can be noticed in Figure 5 that with all minimum support iterations there was a reduction in the number of association rules due to ER's irrelevant and redundant methods. Most of the irrelevant rules discovered with minimum support of 0.01 in **Experiment 1** were very helpful, especially in the last iteration of **Experiment 2**. With all methods combined, the reduction of rules was, approximately, 75%. If compared only to ER without the method introduced in this article, the final set of rules decreased by about 60%, which shows how important the new approach is.

Comparing **Experiment 1** to **Experiment 2** with *minimum support* set at 0.01, it can be observed that **Experiment 2** has fewer rules than **Experiment 1**. This is the result of the *Irrelevant Rules Elimination with Weak Supervision* method. All irrelevant association rules discovered in **Experiment 1** with minimum support set at 0.01 were used in **Experiment 2**. Those rules as well as the irrelevant ones that would be extracted because they had the same patterns were not generated in **Experiment 2**. Without the new technique development the number of rules to be analyzed would be about 2.5 times higher.

The results of **Experiment 3** are in Figure 6. The same behavior as in those two previous experiments was noticed. As the minimum support value decreases the number of rules increases. The data set size was also incremented by about 10%, which contributed to the generation of more association

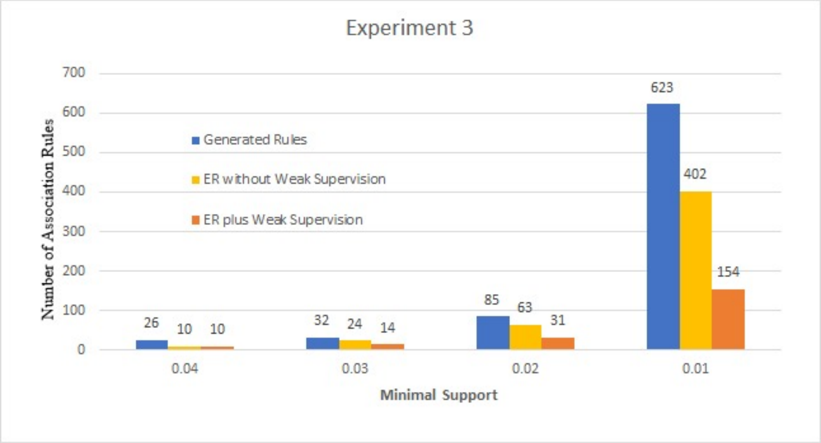

Fig. 6. Generated Rules by **Experiment 3**

rules.

As **Experiment 2**, redundant and/or irrelevant rules were removed since the first iteration (0.04 minimum support). Considering the minimum support value of 0.01, the reduction was 75%, approximately, if compared to the total number of generated rules without any redundancy/irrelevancy treatment, and 55% compared to ER without the irrelevant rules with weak supervision method. The amount of rules applying the complete ER was a little bigger in **Experiment 3** if compared to **Experiment 2**. This is the result of the new association rules discovered, probably due to the characteristic of the data added.

It is important to point out that the reduction in the amount of generated rules to be evaluated does not impact on the processing of NELL's KB population, because the algorithm only eliminates irrelevant and redundant rules. All redundant association rules were used to populate NELL's KB without any loss of information.

5. CONCLUSION AND FUTURE WORKS

This article presented a weakly supervised learning algorithm to eliminate irrelevant association rules in NELL's large-growing knowledge base. The *Irrelevant Rules Elimination with Weak Supervision* method was included in the ER component [Miani and Hruschka Jr 2018] to help in the reduction of rules.

The method consists of removing from the final set of rules, irrelevant association rules with the same pattern as another irrelevant association rule discovered before. Minimal supervision is required to evaluate only new association rules. Experiments showed that the new technique decreased by about 60% the amount of rules extracted in comparison to the ER component without the new technique, which contributes to a reduction in the time spent evaluating them.

However, it is still necessary to analyze redundant rules and new irrelevant ones. In this way, in future research, it is intended to develop techniques to:

(1) Automatically identify relevant association rules, based on those already discovered;
(2) Improve the evaluating process developing methods to automatically classify any rules (irrelevant or not), without any supervision.

To accomplish the first technique, it is expected that an algorithm similar to the *Irrelevant Rules Elimination* method will be used. The approach might consider relevant association rules already discovered and will automatically classify rules with the same patterns as relevant too. The second idea aims to evaluate the new rules discovered without any supervision.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In *IN: PROCEEDINGS OF THE 1993 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, WASHINGTON DC (USA*. pp. 207–216, 1993.

Appel, A. P. and Hruschka Jr, E. Prophet–a link-predictor to learn new rules on nell. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, pp. 917–924, 2011.

Baralis, E., Cagliero, L., Cerquitelli, T., and Garza, P. Generalized association rule mining with constraints. *Information Sciences* vol. 194, pp. 68–84, 2012.

Bayardo Jr, R. J. Efficiently mining long patterns from databases. *ACM Sigmod Record* 27 (2): 85–93, 1998.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. Dbpedia - a crystallization point for the web of data. *Web Semant.* 7 (3): 154–165, Sept., 2009.

Burdick, D., Calimlim, M., and Gehrke, J. Mafia: A maximal frequent itemset algorithm for transactional databases. In *Data Engineering, 2001. Proceedings. 17th International Conference on*. IEEE, pp. 443–452, 2001.

Carlson, A., Betteridge, J., Hruschka Jr, E. R., and Mitchell, T. M. Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*. Association for Computational Linguistics, pp. 1–9, 2009.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. Toward an architecture for never-ending language learning. In *In AAAI*, 2010.

Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr, E. R., and Mitchell, T. M. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 101–110, 2010.

Djenouri, Y., Drias, H., and Bendjoudi, A. Pruning irrelevant association rules using knowledge mining. *International Journal of Business Intelligence and Data Mining* 9 (2): 112–144, 2014.

Dong, X., Hao, F., Zhao, L., and Xu, T. An efficient method for pruning redundant negative and positive association rules. *Neurocomputing*, 2019.

Fan, W., Wang, X., Wu, Y., and Xu, J. Association rules with graph patterns. *Proceedings of the VLDB Endowment* 8 (12): 1502–1513, 2015.

GalÁrraga, L. A., Teflioudi, C., Hose, K., and Suchanek, F. Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Int. World Wide Web Conf. Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 413–422, 2013.

Gouda, K. and Zaki, M. J. Genmax: An efficient algorithm for mining maximal frequent itemsets. *Data Mining and Knowledge Discovery* 11 (3): 223–242, Nov., 2005.

Grahne, G. and Zhu, J. High performance mining of maximal frequent itemsets. In *6th International Workshop on High Performance Data Mining*, 2003.

Marinica, C. and Guillet, F. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering* 22 (6): 784–797, 2010.

Matuszek, C., Cabral, J., Witbrock, M., and Deoliveira, J. An introduction to the syntax and content of cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. pp. 44–49, 2006.

Miani, R. G., Yaguinuma, C. A., Santos, M. T., and Biajiz, M. Narfo algorithm: Mining non-redundant and generalized association rules based on fuzzy ontologies. In *Enterprise Inf. Systems*. Springer, pp. 415–426, 2009.

Miani, R. G. L. and Hruschka Jr, E. R. Eliminating redundant and irrelevant association rules in large knowledge bases. In *ICEIS (1)*. pp. 17–28, 2018.

Miani, R. G. L. and Hruschka Junior, E. R. Exploring association rules in a large growing knowledge base. *Int. J. of Comp. Info. Syst. and Ind. Mangt Apps*, 2015.

Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory*. ICDT '99. Springer-Verlag, London, UK, UK, pp. 398–416, 1999.

Pedro, S. D. and Hruschka Jr, E. R. Conversing learning: Active learning and active social interaction for human supervision in never-ending learning systems. In *Advances in Artificial Intelligence–IBERAMIA 2012*. Springer, pp. 231–240, 2012.

Rai, N. S., Jain, S., and Jain, A. Mining interesting positive and negative association rule based on improved genetic algorithm (mipnar_ga). *International Journal of Advanced Computer Science and Applications* 5 (1), 2014.

Rameshkumar, K., Sambath, M., and Ravi, S. Relevant association rule mining from medical dataset using new irrelevant rule elimination technique. In *Information Communication and Embedded Systems (ICICES), 2013 Int. Conf. on.* IEEE, pp. 300–304, 2013.

Sinthuja, M., Puviarasan, N., and Aruna, P. An efficient maximal frequent itemset mining algorithm based on linear prefix tree. In *Communication and Computing Systems: Proceedings of the 2nd International Conference on Communication and Computing Systems (ICCCS 2018), December 1-2, 2018, Gurgaon, India.* CRC Press, pp. 92, 2019.

Srikant, R. and Agrawal, R. Mining generalized association rules. In *Proceedings of the 21th International Conference on Very Large Data Bases.* VLDB '95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 407–419, 1995.

Suchanek, F. M., Kasneci, G., and Weikum, G. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web.* WWW '07. ACM, New York, NY, USA, pp. 697–706, 2007.

Swesi, I. M. A. O., Bakar, A. A., and Kadir, A. S. A. Mining positive and negative association rules from interesting frequent and infrequent itemsets. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on.* IEEE, pp. 650–655, 2012.

Wahyudi, W., Khodra, M. L., Prihatmanto, A. S., and Machbub, C. Using graph pattern association rules on yago knowledge base. *Journal of ICT Research and Applications* 13 (2): 162–175, 2019.

Xiao, W. and Hu, J. Mrclose: A parallel algorithm for closed frequent itemset mining based on mapreduce. In *Proceedings of the 2019 International Conference on Robotics Systems and Vehicle Technology.* pp. 7–13, 2019.

Zaki, M. J. Generating non-redundant association rules. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '00. ACM, New York, NY, USA, pp. 34–43, 2000.

Zaki, M. J. and Hsiao, C.-J. Charm: An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining.* SIAM, pp. 457–473, 2002.