European Journal of Innovation in Nonformal Education (EJINE)
Volume 2 | Issue 3 | ISSN: 2795-8612

Application of Rule-Based Reasoning in the Automatic Extraction of Metadata from Academic Articles

Rashid Turgunbaev
Tashkent University of Information Technologies

A B S T R A C T

This article provides detailed information on rule-based reasoning, rule-based systems and their main features, general view of rule-based systems and methods of their organization, forward chaining and backward chaining systems and algorithms, rule selection, application of rule-based reasoning in automatic metadata extraction given.

A R T I C L E I N F O

Article history:
Received 14 January 2022
Received in revised form
11 February 2022
Accepted 07 March 2022

Keywords: rule-based reasoning, rule-based systems, forward chaining systems, backward chaining systems

Hosting by Innovatus Publishing Co. All rights reserved. © 2022

INTRODUCTION

Certain rules are widely used in the expression of knowledge and reasoning. Known rules naturally make it easier to understand the knowledge presented. The basic view of the rule looks like this

if <conditions> then <conclusion>

where, <conditions> represents the conditions or basis of the rule, <conclusion> represents its result. The terms of the rule form logical functions by connecting to each other with logical connectors. When sufficient conditions of the rule are met, a conclusion is drawn and the rule is said to be fulfilled. The rules represent general knowledge of the field [1].

Rule-based systems are one of the most common and easy-to-implement tools for human experts to encode problem-solving skills. Experts tend to express most problem-solving methods through a set of situational rules, suggesting that rule-based systems should be the choice in the organization of expert systems that require knowledge.

Rule-based systems have the following main features [2]:

1. Introduces knowledge into if-then conditional rules.
2. System skills increase in proportion to the expansion of the knowledge base.
3. Problems of varying complexity can be solved by selecting appropriate rules and combining them accordingly.
4. The system can be flexibly designed to define a sequence of best rules for execution.
5. The system can review the stages of reasoning and justify the conclusions drawn. Every rule that is followed in reconsidering the stages of thinking can be illustrated using natural language.

GENERAL VIEW OF RULE-BASED SYSTEMS

The rule-based system includes a knowledge base and a reasoning mechanism. A knowledge base is a set

of facts and rules. A fact is a simple statement with a list of constant or variable objects, while a rule represents a conditional sentence with one or more bases and a subsequent component. The rules must be interpreted, i.e. if the basis can be satisfied, the result will also be satisfied. The result can be a conclusion or an application. When the result is valid, the effect of the satisfaction of the basis will be to perform that action. If the result determines the conclusion, the effect of the satisfaction of the basis will be to draw a conclusion [3].

There are several basic ways to set up rule-based systems. Rules can be used to express deductive knowledge, such as logical relationships, and thereby support the functions of inference, verification, or evaluation. Conversely, rules can be used to express goal-oriented knowledge that rule-based systems can apply to problem solving, and can be cited in justifying aspirational behaviors for their own purposes. Rules can also be used to express cause-and-effect relationships, and these relationships can be used by rule-based systems to answer "what-if" questions or to identify possible causes of defined events.

In a rule-based system, each if part can correspond to one or more statements in a set of statements. A set of statements is called a working memory. The total statements can be based on one or more rules in the knowledge base. All rules that match the statements in the working memory are placed in the problem set. From there, one of the rules is selected based on some problem-solving criteria, and then the selected rule is executed. The conclusions drawn are placed in the working memory to form a new statement, and this process continues until the desired result is achieved. Such a system is called a deduction system because it draws new conclusions from rules and statements.

Facts are a type of data that represents statements about features, relationships, suggestions in a knowledge base. Unlike rules that rule-based systems interpret as imperative, facts are usually immutable and indirectly inactive. In addition, the fact is pragmatic value and quiet in relation to the dynamic use of its knowledge. Although facts and rules logically replace each other in many contexts, in rule-based systems they differ from each other.

In addition to static memory for facts and rules, rule-based systems use working memory to store temporary statements. These statements record conclusions based on the previous rule. The contents of the working memory can be described as problem-solving case information. The information in the working memory applies to the syntactic convention of facts. Interim statements are consistent with dynamic facts.

The main task of rule-based systems is to develop results. The main result may be the result of solving a problem, answering a question, or analyzing some information. A rule-based system uses several basic processes that determine its overall performance.

In deduction and reaction systems, the correct connection is the process of going from part to then part if if the part determines the appropriate situations to deduce a new statement and the action to be performed in the case of reaction systems.

FORWARD CHAINING

In forward chaining systems, the process begins with the initial facts and new conclusions are drawn using the rules, taking these facts into account. Forward chaining systems are largely data-driven. At any time, the previous condition is satisfied when the if part corresponds to the statement. If an initiated rule establishes a new statement or performs an action, that rule is said to have been complied with. This process is repeated until other rules apply.

The selection process is carried out in two stages:

1. Preliminary selection: identify a set of appropriate rules called a set of problems.

2. Selection: Select a rule from a set of problems using problem-solving strategy tools.

Forward chaining algorithm

A simple forward chaining algorithm starts with certain facts in the knowledge base, executes all the rules that are the basis of certain facts, and then adds the results of all of them to the knowledge base. This process is repeated until a response is received or a conclusion is added to the knowledge base.

Dispute resolution

If more than one rule can be applied to data-based reasoning, it will not be necessary to follow all the rules. The removal of some of the applicable rules requires a dispute resolution mechanism to implement the most appropriate rule. In deduction systems, all rules are usually fulfilled, but in reaction systems, only one of the possible actions is considered acceptable when more than one rule is triggered at the same time [4].

Dispute resolution strategies prevent redundant rules from being enforced. The most obvious strategy is to randomly select one rule from these possible rules. The following approaches are available in determining the properties of a rule to be followed.

Selection by order

The selection of the order can be by order of the rule in the knowledge base or by the order of final application of the rule. The first rule that can be applied to the selection by order is selected in the order provided. When selecting the last application order, the rule is selected based on the final application.

Selection of a rule by its syntactic structure

The syntax of a rule can be the conditions and their specificity or the size of the rule. In the case of specificity, the rule whose terms can be most clearly applied is selected. In terms of size, the rule with the largest syntactic, that is, the most propositions, applies first.

Selection through additional knowledge

Additional knowledge about rules can be in the form of rule priority or in the form of meta knowledge. In the form of priority, each rule is assigned a level of priority in numerical form, and the rule with the highest priority is executed first.

In the simplest way to implement a proper link interpreter, the preconditions of all rules are checked, a list of applicable rules is created based on the criteria given in the dispute resolution strategy, then the leading rule in the list is executed, the current list is deleted, and the process restarts.

Efficiency in choosing rules

There are several criteria for choosing rules to implement the thinking process.

Similarities between rules

This approach is based on the similarity between the preconditions of the different rules. The rules that have a general proposal can be removed altogether if these proposals are not followed. To achieve this, the rules are arranged in a tangled hierarchy by creating a network of interconnected trees. In it, the internal decision nodes are formed by the propositions of the rules and the leaves of the tree. The path from the root of the tree to the rule contains all the propositions of the precondition of this rule.

Investigate the applicability of the new rules

This approach is based on the fact that a full re-calculation of a list requires more effort than a change-based editing. From one cycle to another, the knowledge base is changed only by the action of a selected rule, in which knowledge is added or deleted. It will be necessary to verify that the deleted knowledge eliminates the applicability of any rule in the old list and that the added knowledge makes the new rules applicable. The next check is done through a network of rules with new knowledge as a starting point.

Indexing

In this approach, all rules are indexed, and each parameter of the definition forms a reference to its rule. When a parameter accepts a new value or its value changes, this link allows you to find and effectively check the relevant rules.

The complexity of the preconditions

Different formalisms have different expressive powers in the solution of the description of preconditions and rules. The difference is the difference between the logic of the proposition and the logic of the

description. Ideally, each relevant state of the domain should be covered by only one rule. Describing a situation with different rules has a negative effect on the modularity of the knowledge base, as these rules can only be meaningful as a group. On the other hand, the high expressive power of rule formalism is complex and in many cases requires an inefficient rule interpreter. Therefore, the degree of complexity required for a domain should be examined [4].

Searching the knowledge base is the simplest way to assess preconditions. However, in digital or time relationships, the search alone is not considered sufficient and additional calculations are required.

One of the disadvantages of forward chaining is that it draws all possible conclusions based on existing facts and rules, regardless of the objectives being taken. This in turn suggests that it is an ineffective approach to thinking about the right connection. To avoid this shortcoming, it is handled with a selected subclass of rules. Backward chaining is another approach that only leads to the basics that lead to the goal. This is why such a system is called goal-oriented.

BACKWARD CHAINING

The backward chaining process in the system begins with some hypothesis that needs to be proven and continues with the search for rules that allow conclusions to be drawn from that hypothesis. Small goals can also be set to prove it during the process. To do this, the target statement is first placed in working memory, after which the system performs the following two operations:

1. The result of the rule is compared with the goal.

2. Selects the appropriate rule and places its basis in the working memory.

The second operation above refers to the process of dividing the problem into smaller goals. The process continues in the next iteration, in which these foundations become new goals that are consistent with the outcome of other rules. Hence, backward chaining systems can also be understood as test hypotheses.

Backward chaining uses stacks. First, the target is placed on the stack, then all the rules that lead to that goal are searched and placed on the stack. These become small goals and the process continues until all the facts have been proven or are available as key points. Unification is carried out each time a decision is made on the purpose and grounds.

Backward chaining algorithm

The backward chaining algorithm returns a set of substitutions, which makes the target a reality. The replacement package will initially be empty. The input data for the algorithm serves as a knowledge base, targets, and current substitutions. The algorithm returns a summary set of targets.

Goal setting

If the target is not known in the knowledge base, the rule interpreter decides whether it can be obtained first or whether it should be requested by the user. In order to achieve the goal, all the rules contained in the result section of this goal are processed. If rules are indexed based on result components, they can be effectively identified.

Thus, the backward chaining has an indirect dialog control, where the order of the questions determines the order of the parameter retrieval rules. Dependence on such an arrangement reduces the modularity of the rule-based system. The effectiveness of the backward chaining rule interpreter is determined by goal setting. The clearer the goal, the smaller the rules that need to be checked and the tree of questions that need to be asked.

APPLICATION OF RULE-BASED REASONING IN AUTOMATIC EXTRACTION OF METABOLES

Various researches have been conducted on the application of rule-based thinking to automatic extraction of data, including automatic extraction of metadata from academic articles [10] [11] as in machine learning [12] and case-based reasoning [13].

Waltl, B. et al. [4] suggest a method of automatic extraction of semantic data from official documents. The interaction of data processing and the extraction of semantic data based on a large body of documents

are demonstrated. It shows the capabilities of rule-based text analysis in the automatic extraction of semantic data. In addition, the legislation provides for the extraction of legal definitions and their use in a context that defines terms. In the analysis of the documents, a combination of structural and linguistic indicators was selected. Inside the case is written a special commentator to classify the relevant sentences. It also provides a complete set of logically related rules. 94% accuracy was achieved in the extraction of the identified data.

Azimjanov, J. et al. [6] suggest a rule-based system for extracting metadata from academic articles. It offers a free and open source, high-performance system based on the Java programming language. The extraction rate of the system is 9-10 times higher than existing metadata extraction systems. In this approach, the titles of academic articles are extracted according to location features, text size, and font. Other metadata, such as annotation, body of the article, keywords, conclusion, and references used, are extracted according to a clearly defined set of rules.

Makhija, V., and others [7] suggest a rule-based method of extracting texts from a bibliographic database. It suggests the process of extracting rule-based data from selected data extracted from a bibliographic database of published academic articles. The established rule-based system not only allows to create a field of knowledge, but is also useful in improving the data quality of the database. The data extraction algorithm is modified separately for each document type. Some algorithms were relatively simple and allowed to obtain accurate results. Some algorithms were complicated and gave erroneous and unexpected results. General rules have been developed for non-field analysis, covering all types of documents. In addition, the system allows you to synchronize and update the rules with samples in the field of knowledge. The system allows you to create a set of knowledge, delete duplicate data, create semantics using Bul logic, and translate knowledge into rules using iterative verification.

Shigarov, A. et al. [8] propose a software platform for extracting and modifying data from rule-based spreadsheets. This software platform allows you to extract data from spreadsheets and convert them to a structured view. It is based on a flexible table object model and field-specific rule language for table analysis. They serve to express knowledge about the structure and content properties of a table, as well as to interpret them according to the purpose of the change. This allows you to process tables from different fields. Empirical results show that a single set of rules can be used to process arbitrary tables with the same structure, style, or content characteristics.

ALAMOUDI, A., and others [9] suggest a rule-based approach to extracting data from PDF books. This approach proposes an intelligent, rule-based approach to error-free extraction of logical metadata from PDF books. In addition, a set of rules and templates have been established to increase the level of accuracy. The results of the experiment show that this approach is capable of successfully extracting metadata from PDF books with 94.62% and 90.27% accuracy, respectively, for training and test sets.

References

1. Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. Springer Nature.
2. Hayes-Roth, F. (1985). Rule-based systems. Communications of the ACM, 28(9), 921-932.
3. Patterson, D. (1990). Introduction to artificial intelligence and expert systems. Prentice-Hall, Inc..
4. Puppe, F. (2012). Systematic introduction to expert systems: Knowledge representations and problem-solving methods. Springer Science & Business Media.
5. Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, T., Stocker, C., & Schneider, C. (2017). Automated extraction of semantic information from german legal documents. In IRIS: Internationales Rechtsinformatik Symposium.
6. Azimjonov, J., & Alikhanov, J. (2018). Rule based metadata extraction framework from academic articles. arXiv preprint arXiv:1807.09009.
7. Makhija, V., & Ahuja, S. (2018). Rule based Text Extraction from a Bibliographic Database. DESIDOC Journal of Library & Information Technology, 38(1).
8. Shigarov, A., Khristyuk, V., & Mikhailov, A. (2019). TabbyXL: Software platform for rule-based

spreadsheet data extraction and transformation. SoftwareX, 10, 100270.

9. ALAMOUDI, A., ALOMARI, A., & ALWARTHAN, S. (2021). A Rule-Based Information Extraction Approach for Extracting Metadata From PDF Books. ICIC express letters. Part B, Applications: an international journal of research and surveys, 12(2), 121-132.
10. Turgunbaev, R. (2021, September). Metadata in Data Search. In " ONLINE-CONFERENCES" PLATFORM (pp. 93-96).
11. Тургунбаев, Р. (2021). Метамаълумот: хусусиятлари, турлари ва стандартлари. Science and Education, 2(5), 167-175.
12. Тургунбаев, Р. (2021). Машина ўрганиши ва уни академик мақолалардан метамаълумотларни автоматик экстракция қилишда аҳамияти. Science and Education, 2(11), 302-316.
13. Turgunbaev, R. (2021). Keysga asoslangan fikrlash va uni akademik metama'lumotlarni avtomatik ekstraksiya qilishda tadbiq qilinishi. Science and Education, 2(9), 129-144.