Development of the formal model for the presentation of poorly structured and unstructured information

A V Rabin and A A Petrushevskaya
Saint-Petersburg State University of Aerospace Instrumentation (SUAI), ul. Bolshaya Morskaya, 67, lit. A, St. Petersburg, 190000, Russia
E-mail: alexey.rabin@guap.ru

Abstract. The article describes the tasks of data mining, data modeling, presents a classification of data models, and also develops a formal model for representing various types of weakly structured and unstructured information resources. The developed formal model for representing various types of weakly structured and unstructured information resources within the framework of the software environment represents the theoretical basis for creating tools for solving the applied problem of automatic analysis of documentation on paper and digital media for the subsequent cataloging of poorly structured information.

1. Introduction
Currently, within the framework of the «Presentation of Knowledge» direction, which is one of the areas of artificial intelligence (AI), problems are being solved related to the formalization and presentation of knowledge in the memory of an intelligent system (IS). For this, special models of knowledge representation and languages for describing knowledge are developed, various types of knowledge are distinguished. We study the sources from which IP can draw knowledge, and create procedures and techniques by which it is possible to acquire knowledge for IS. The problem of knowledge representation for IS is extremely relevant, since IS is a system whose functioning is based on knowledge about the problem area that is stored in its memory.

In an electronic computer, knowledge, like data, is displayed in sign form - in the form of formulas, text, files, information arrays, etc. Therefore, in the narrowest sense, we can say that knowledge is a specially organized data.

In intelligent systems, knowledge is the main object of formation, processing and research. The knowledge base, along with the database, is a necessary component of the software complex of intelligent systems.

2. Tasks of data mining
Data mining is one of the most relevant and sought-after areas at present. Modern business and manufacturing processes generate huge amounts of data. How to extract the maximum of useful knowledge from multidimensional, heterogeneous, incomplete, inaccurate, contradictory, indirect data, and how to use the available information to automate decision-making in various professional fields - at least these and many other problems are considered in this direction.

Classification refers to a supervised learning strategy, which is also called supervised learning.
The classification task is often called the prediction of a categorical dependent variable (i.e., a dependent variable that is a category) based on a selection of continuous and/or categorical variables. The task of clustering is to divide many objects into groups (clusters) of similar parameters. Moreover, in contrast to the classification, the number of clusters and their characteristics can be unknown in advance and determined during the construction of clusters based on the degree of proximity of the objects to be combined according to the set of parameters.

A cluster can be characterized as a group of objects having common properties. The characteristics of the cluster can be called two signs:

- internal uniformity;
- external isolation [1-3].

Table 1 shows a comparison of some parameters of the classification and clustering problems.

| Feature                  | Classification | Clustering          |
|--------------------------|----------------|---------------------|
| Learning controllability | Supervised training | Unsupervised learning |
| Strategy                 | Teacher training | Teacherless learning |
| Class label              | The training set is accompanied by a label indicating the class to which the observation belongs | Learning set class labels unknown |
| The basis for classification | New data classified based on training set | Given a lot of data to establish the existence of classes or clusters of data |

It should be noted that as a result of applying various methods of cluster analysis, clusters of various shapes can be obtained. For example, “chain” type clusters are possible when clusters are represented by long “chains”, elongated clusters, etc., and some methods can create clusters of arbitrary shape.

Different methods may seek to create clusters of certain sizes (for example, small or large), or assume the presence of clusters of different sizes in the data set.

3. Tasks of data modeling. Data Model Classification

Creating an information system in order to support activities in any field allows a person working in this field to interact not with the real object itself, but with the information system. This approach allows you to get an idea of the state of objects and ongoing processes of the subject area not by direct measurement in the real world, but using the information model. The advantage of the model approach is that it provides opportunities for analysis and assessment of the consequences of various management decisions without conducting a full-scale experiment. Information models provide a formalized representation (algebraic, graphic, etc.) of the studied objects and their relationship. Data models will never provide a comprehensive knowledge of the subject area, but it is important that the amount of knowledge and semantics of the data is consistent with the intended use of the data.

The essential feature by which operations are distinguished is determined by the method of obtaining the result. On this basis, navigation and specification operations are distinguished. An operation is called navigation if its result is obtained by going through the relationships implemented in the data structure. The result of its execution is a single object, for example, attribute value, entity implementation [4].

If operations can specify a new structure defined on the database structure, which generally corresponds to many objects, then such an operation is called specification. Navigation operations always involve sampling through the current, while for specification operations this is not mandatory.

The operations defined above are performed according to the “Choice-Action” scheme, however, there are other operations performed according to other schemes that implement more complex
functions, for example, the integrity support function. Such generalized operations are called database procedures. This is a highly general mechanism for changing the state of a database. The procedure is considered as a single macro operation. Using procedures significantly expands the dynamic properties of the data model [4-5].

Operations can lead to a change in the database or leave it in the current state (for example, operations to update field values and select according to a specific user request). Performing operations may violate explicit integrity constraints; internal constraints, as already mentioned, are automatically maintained by the database management system.

Thus, the data model can be defined as «a set of rules for structuring data in databases, permissible operations on them, and integrity constraints that they must satisfy» (figure 1).

![Data model structure](image)

**Figure 1. Data model structure.**

When describing the subject area, we use the term data model at different levels. These levels are not always associated with the classic three-tier database architecture and can be represented as the following diagram.

At the very lowest level, it is said that the current state of a particular database is a model of reality, because this is a record of selected facts that are true at the moment (a requirement of the relevance of the databases). As an example, we can cite the situation when there is «Product «B» in the amount of 5 units in the warehouse. In case of a change in the quantity of goods, the database must change in order to remain a true model of reality.

At a higher level, the schema describing the structure of the database is a model of a collection of models (i.e., it is a model of a collection of database states). The scheme models a wide variety of
database states, determining the general characteristics of all these states. So, «Client Name» and his «Address» are highlighted as characteristics applicable to the quality of various people, and may change from time to time.

At the highest level, the database design methodology describes the designs and rules that apply when creating the schema. Thus, this level is also a model of many models (possible data schemes). This design methodology, such as a conceptual or relational data model, is a top-level model and describes in general terms many schemes.

Thus, speaking of a data model, it implies a methodology for creating database schemas for practical situations. These database schemas are, in turn, models that define the logical structure that captures facts about some part of the real world. When these facts are recorded and recorded in the computer database system, the database itself becomes a model of the current state of reality. Each of the two upper levels becomes a model of the level preceding it (figure 2).

The subject area of economic information processing systems is considered defined if the objects existing in it, their properties and relationships are known. Modeling of the economic system begins with a preliminary structuring of the subject area: real-world objects are classified, a set of object types to be displayed in the database is fixed.

![Figure 2](image-url)

*Figure 2. Three levels of models.*

For each type of object, a set of properties is revealed by which specific objects of this type will be described in the database and the types of relations (relationships) between these objects. Then questions regarding the information presented in the database are resolved.

Analysis of the main types of models of economic systems allows us to formulate the basic requirements that the modeling apparatus must meet:

- ensure that the characteristics of economic information are taken into account;
- provide means for determining internal integrity constraints that are relevant to the main practical situations;
- based on the use of objects that are complex structures;
- provide an opportunity to describe a wider class of concepts used to adequately represent the subject area;
- provide a variety of types of representations of objects of the subject area (graphic, textual, algorithmic);
- allow to build a multidimensional display of the subject area, on the basis of which the requirements for a data processing system can be formalized;
• ensure the sustainability of design decisions to changes in the information sphere of the subject area.

In the modeling process, as the main information unit, it is necessary to have one that most adequately reflects the essence of the object and the laws of development of the subject area. At the initial stage of modeling the economic structure, where the procedures of substantive analysis are dominant, it is preferable to use models with a high level of semantic mapping. At the subsequent stages of modeling, where the syntactic structure of the data becomes decisive, it is necessary to use models that provide means of formalizing the structural relationships of the data. Moreover, they are already abstracting from their semantics and other substantive aspects [6].

4. The formal model for representing various types of weakly structured and unstructured information resources

In the framework of the study, under-structured and unstructured information resources are understood as text data of various formats, represented by text in a natural language. In fact, a textual information resource is a container of poorly structured and / or unstructured information.

To solve the problem of semantic analysis of text information resources, it is necessary to use several different representations of text information resources:

1. Statistical index – this presentation of text information resources is formed in the process of statistical analysis of the content of text information resources.

Formally, the statistical index model can be represented using the following expression (1):

$$I_{Stat}^D = \{(w_i^D, freq_i^D), (w_j^D, freq_j^D), K, (w^D, freq^D)\},$$

where \(w_i^D \in W^D\) – the i-th term of text information resources D; \(f_i^D\) – frequency of occurrence of the i-th term \(w_i^D\) of text information resources D.

2. The linguistic index is formed in the process of morphological analysis of the content of text information resources and can be represented in the form of the following expression (2):

$$H_{Lin}^D = \left\{ \left( \text{par}_1^D, \text{sent}_1^D, w_1^D, M_{Lin}^{NL_P} \right), \left( \text{par}_1^D, \text{sent}_1^D, w_2^D, M_{Lin}^{NL_P} \right), K \right\},$$

where \(w_k^D \in W^D\) – k-term of text information resources D; \(M_{Lin}^{NL_P} \subseteq M_{Lin}\) many morphological features of the k-th term \(w_k^D\) in the j-th sentence \(sent_j^D\) paragraph i – \(par_i^D\) text information resources D.

3. Structural index – a structural representation of text information resources containing information about the boundaries of paragraphs, sentences and the position of words in them.

Formally, the structural index can be written as follows (3):

$$H_{Struct}^D = \left\{ \left( \text{par}_1^D, \text{sent}_1^D, w_1^D, \text{pstrn}_1^D \right), \left( \text{par}_1^D, \text{sent}_1^D, w_2^D, \text{pstrn}_2^D \right), \ldots \right\},$$

where \(w_k^D \in W^D\) – k-term \(w_k^D\) i-th paragraph \(par_i^D\) j -th sentence \(sent_j^D\) text information resources D; \(\text{pstrn}_k^D\) – position of the k-th term \(w_k^D\) in the j-th sentence \(sent_j^D\) i of the paragraph \(par_i^D\) of textual information resources D.

Thus, the use of several types of representations of poorly structured and unstructured information resources allows you to take into account various features of text information resources in the process of further analysis.

5. Conclusions

When solving complex problems in the face of uncertainty, the user of the software environment should receive comprehensive support from the system, having the opportunity to choose a solution not only
through the use of expert knowledge stored in the knowledge base, but also taking into account the experience of solving this problem in the past. Such experience can be formalized in the form of a set of precedents for the ontology of precedents, which should be processed and taken into account as part of the solution of the problem.

The development of a unified environment for semantic analysis of flows of poorly structured information that implements modern intelligent algorithms for processing text information will greatly facilitate the decision-making process by a specialist in the time constraint mode, due to the possibility of using a single unified bank of expert knowledge in the work of the question-answer system, and will also allow automated semantic verification of information flows in order to provide information ion safety organization.

Acknowledgments
The paper was prepared with the financial support of the Ministry of science and higher education of the Russian Federation in the course of the applied research «The comprehensive project to create high-tech production of software tools for automatic analysis of documentation on paper and digital media using semantic-cognitive technologies for cataloging poorly structured information» (unique identifier of the project 075-11-2019-055, Decree of the Government of the Russian Federation N 218, 09.04.2010).

References
[1] Rajpathak D, Chougule R and Bandyopadhyay P 2012 A domain-specific decision support system for knowledge discovery using association and text mining Knowledge and Information Systems 31 405–32
[2] Renu R S, Mocko G and Koneru A 2013 Use of big data and knowledge discovery to create data backbones for decision support systems Procedia Computer Science 20 446–53
[3] Chen D and Manning C D 2014 A fast and accurate dependency parser using neural networks Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing pp 740–50
[4] Ltifi H, Kolski C, Ayed M B and Alimi A M 2013 A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases Journal of Decision Systems 22 69–96
[5] Zarubin A A, Koval A R, Moshkin V S and Filippov A A 2017 Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources Conference: Information Technology and Nanotechnology 2017 Retrieved from http://ceur-ws.org/Vol-1903/paper26.pdf
[6] Rabin A V, Petrushhevskaya A A and Sinitsin O V 2020 Methods and formal models of intelligent analysis of weakly structured data IOP Conference Series: Materials Science and Engineering 450 012159