UDC 004.827

O. Siedushev, E. Burov
Lviv Polytechnic National University,

Information Systems and Networks Department

FORMS OF FUZZINESS IN DATA AND KNOWLEDGE BASES

This paper describes and analyzes the whole variety of forms of fuzziness in data and knowledge bases. All kinds of data that is imprecise, vague, uncertain, incomplete etc. have been researched and compared. It is shown that fuzzy knowledge is represented through fuzzy production rules, which contain fuzzy data in the antecedent and parts.

Keywords – forms of fuzziness, fuzzy data, fuzzy knowledge, fuzzy production rules.

Introduction and problem statement

Classical models of data often suffer from their own inability to display or manipulate the data and knowledge that is imprecise, uncertain, vague etc. However, such data and knowledge are increasingly found in today's information systems, databases, data warehouses, knowledge bases as many subject areas that are sources of data constitute a manifestation of different forms of fuzziness.

There is a strong need to identify the forms of such data and knowledge, to distinguish between them, to use them as input in a particular form, to transform and even to remove if necessary.

In this article, the entire set of heterogeneous forms of data that are inaccurate, vague, uncertain, inconsistent, incomplete, etc. will be called non-classical data. The reason for this is that the nature and origin of all such forms of data varies, therefore there is a need for varying techniques to work with every form of non-classical data. Classical data can be considered as accurate, defined, coherent, clear, that has no lost or missing values etc. Classical data prevailed in many sciences for a long time before L. Zadeh began to form mathematical foundations of fuzzy logic (in the 1960s) [1].

As of today, the range of problems solved by the use of the apparatus of fuzzy sets and fuzzy logic greatly expanded and includes areas such as data analysis and data mining, pattern recognition, operations research, modeling of complex systems, decision support and more.

In most cases, important information for the system comes from two sources:

1) from people and experts that describe their knowledge of the subject area, etc., using natural language, which means generating subjectivity and ambiguity;

2) from the instruments, sensors, meters, mathematical models, which means generating uncertainty and inconsistency.

Therefore, conservation of expert assessments and opinions, and inaccurate data requires knowledge of non-classical data, and the ability to work with it. This includes mining and interpreting data from fuzzy databases and knowledge bases, which are often contain non-classical data.

In general, most modern systems, database and knowledge base management systems, controllers, devices and software applications require a mechanism to maintain and manage non-classical data and fuzzy knowledge, plus the ability to mine and analyze them.

Analysis of recent research and papers

In [2] states that storing non-classical data in databases allows DBMS better respond to user requests, as user queries are usually vague and imprecise. The authors also give a first attempt to classify the manifestations of fuzzy data. Similarly, in [3] authors describe an ontology to represent fuzzy knowledge and non-classical data using classes, slots, instances. This ontology is an intuitive tool that allows users who are not experts to query specific information without the aid of expert who knows the directory structure.

Today the urgent task is a creation of diagnostic expert systems that can work out fuzzy diagnostic information. Knowledge bases of such systems contain qualitative and quantitative information (represented mainly by linguistic variables), which describes the state of the object of diagnosis. In [4] the author describes a mechanism of fuzzy inference for expert computer based diagnostic system, which takes into account the fuzzy qualitative knowledge in the diagnosis, which improves the quality of a system under incompleteness of the current situation.

In [5] are shown manifestations of non-classical forms of data that can be stored in the current fuzzy databases. There are given various examples in relation to different types of fuzzy databases. The emphasis is towards fuzzy object-oriented databases, which can cope with complex objects as well as with fuzzy queries and different forms of ambiguity and uncertainty in the tuples and entities.

The authors of [6] consider the problem of decision making under uncertainty on the basis of fuzzy production rules. They describe the structure and functions of fuzzy decision-making system.

Goal of the article

The goal of this paper is to build a taxonomy framework of forms of non-classical data and fuzzy knowledge, the research and comparison of such forms. Construction of such a taxonomy framework, determining the characteristics of the origin and forms of fuzziness are necessary for analysis and design of methods for mining, interpretation, preservation and processing of fuzzy data and knowledge.

Main material

Incomplete, inconsistent, uncertain, ambiguous, fuzzy or vague, imprecise, null data represent possible non-classical forms of data. Below are listed the characteristics, examples and sub-forms of each of the above forms. It should be mentioned that every non-classical form of data is in itself a manifestation of something (e.g., incompleteness, inconsistency, imprecision, ambiguity, vagueness etc.).

1. Incomplete data

Incomplete usually means the absence or lack of value, or inaccurate information, where the set of possible values covers the entire domain of possible values. Incomplete data processing occurs when the process of data acquisition did not take place in time. Incomplete data is generated by the lost update, bad reading, missing values, passing data file for insufficient number of cycles [7].

2. Inconsistent data

The concept of inconsistency is rather applicable for storing the data in different models than the data themselves. Inconsistency is a semantic conflict, which means that the same aspect or the same value of the data is displayed differently, often by mistake. For example, employee's X salary is stored both in the form of $1000 and 2000$ in the same database or in different databases. Information's inconsistency is usually caused by the process of integration (combination) of information from different input sources. In general inconsistent data can be treated as mutually contradictory and unreliable.

- Data inconsistency caused by unreliable sources

This kind of inconsistency may exist if happens the merge of information from the data sources that are not reliable or corrupted (Table 1).

Table 1.

Sample from the database

Owner	Car	License plate	Release year	Kilometrage
Gnativ Ivan	BMW X5	AA 6549 LO	2001	100000
Ivan Gnativ	BMW X5	AA 6549 LO	2003	200000

According to Table 1 the inconsistency lies in the values of attributes *Release Year* and *Mileage*. In this case, it is difficult to know which of the tuples (first or second) is correct. If the client of the database wants to buy a car, then the data from which tuple he has to focus on? This inconsistency could arise when the owner of the car has made the information with a variety of errors in the different databases.

3. Uncertain data

Uncertainty is generated when the expert forms the subjective opinion or gives the subjective assessment of the truth of a fact in which he is uncertain to all 100%. Distortion of this type of information makes it impossible to precisely determine its truth or falsity. The one and only thing that could be done in this case is to estimate the probability of such information to be true or false on some infinite interval of values (usually intervals [0, 1] and [0, 100] are used, where the first and last values identify 100% true and 100% false information, respectively) [8].

Thus, the uncertainty of the data is associated with the degree of truth of its values. For example, the probability (degree of accuracy) that the car BMW X5 in Table 1 was released in 2001 equals to 50%.

- *Data uncertainty caused by statistical indicators or analysis*

Many data (measurements, survey data etc.) are obtained statistically and in fact are always questionable. In addition, the gauges can measure something with some error (it is caused by imperfection of devices), and thus produce the measured data, although with minimal, but uncertainty.

- *Data uncertainty associated with the classification and protection of information*

Many data are uncertain or doubtful, because their disclosure is undesirable.

4. Ambiguous data

These data represent a certain ambiguity, because they can be interpreted (treated) differently. In general, ambiguity means that some data due to certain circumstances devoid of some semantic independence and uniqueness, which leads to additional interpretations.

- *Data ambiguity caused by using abbreviations*

It often happens that the use of abbreviations leads to confusion in the interpretation of data. In this case, the stored value needs to be clearly deciphered, but not abbreviated.

- *Data ambiguity caused by incomplete context*

Consider a database that stores certain weather data [2]. There can be stored daily average temperature values for a particular city (Table 2).

Table 2.

Sample from the database

City	Date	Temperature
Lviv	09.09.2013	20
Lviv	10.09.2013	22

As can be seen from Table 2, the temperature values are integer values. However, there are no specifications whether the temperature is stored in degrees of Celsius or in degrees of Fahrenheit.

In the same Table 2 the dates are shown that could be stored in any permitted format. Thus, how to know where are the days and where are the months? From those data we can summarize two conclusions: either the dates are equal to 9 September 2013 and 10 September 2013, or they are equal to 9 September 2013 and 9 October 2013. In Ukraine the first number usually indicates a day, and the second - a month. In the USA – vice versa.

- *Data ambiguity caused by different words order*

Such data occur in cases when one single data is represented semantically correct, but in different ways. If we look at the values of the *Owner* attribute in Table 1, we can see that they are the same but written differently. In this particular case it does not play a big role (unless some inconsistency and lack of higher normal forms for relations are manifested), because the name *Gnativ* does not exist, and we clearly

understand that this is a surname. However, there are exist such full names as *Bogdan Boris*. Here it is clearly not possible to say what is a name and what is a surname.

5. Fuzzy or vague data

Fuzzy or vague implies a certain degree of expression when you cannot say about the meaning, value and significance of the data clearly and precisely. In order to translate (defuzzify) fuzziness (vagueness) in the plane of sharpness and clarity one can apply certain mathematical mechanisms such as fuzzy logic.

Fuzzy data contains a fuzzy predicate (e.g., "old", "young", "low", "high"). Typically, these predicates are modeled by fuzzy linguistic variables.

The linguistic variable takes the value of the fuzzy set of words or phrases of some natural or artificial language. The set of values of linguistic variable is called a term set. A term is any element of term set. In the theory of fuzzy sets a term is formalized by means of fuzzy sets membership function M_F.

If to simulate any linguistic variable as a fuzzy subset of values in the interval $[0, \infty)$ with the membership function M_F on the interval [0, 1], then the projection $[0, \infty) \rightarrow [0, 1]$ is a mathematical description of the value of linguistic variable.

For example, let the linguistic variable L = {low, high, very high}. Then the graphical visualization of the linguistic variable L using fuzzy logic is shown in Fig. 1.

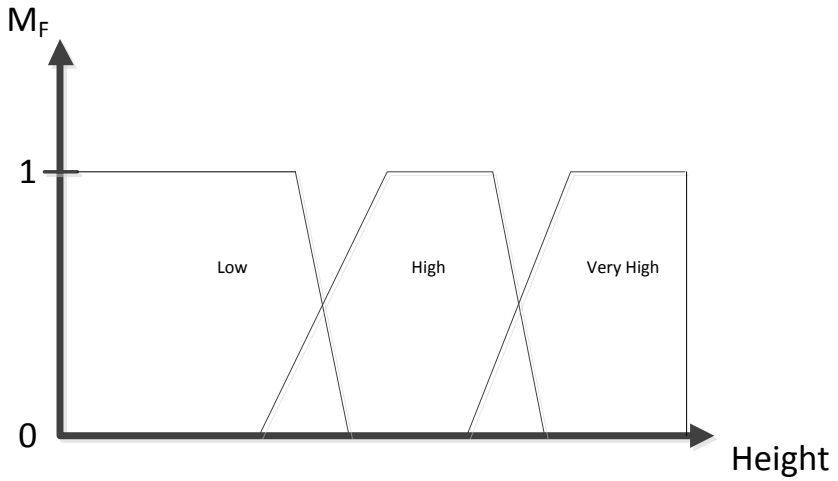

Fig. 1. Graphical visualization of values of the linguistic variable L using fuzzy logic

Intervals and spaces in Fig. 1 indicate only that everyone understands and interprets the low, high, very high values differently. It should be noted that when defining the values of linguistic variable, the latter should comply with certain restrictions, including the ordering, completeness, consistency, normality.

6. Imprecise data

Imprecise data is not false or erroneous data and does not violate the integrity of the information system. Imprecision arises as a product of the existence of a value that cannot be measured with adequate precision.

- Disjunctive imprecise data

- *Always true imprecise data*

This kind of imprecise data can take alternative values from a set of values, but the accuracy and certainty of the data values will always be equal to one, because their probabilities and chances are equal and always true.

The best example of such a data is suitable route data that can be seen in Table 3.

Table 3.

Route data

Destination	Route bus, №
Shevchenko's grove	15 or 24 or 30 or 40

To get to the Shevchenko's grove, one must use one of the above mentioned route buses. There are no differences which route bus to choose, because all of them are equal in the meaning of how to get to the Shevchenko's grove. But this case creates inaccuracy (imprecision), because it is necessary to choose an option from a set of options to get the exact option (value). This situation can be modeled by fuzzy logic in Fig. 2.

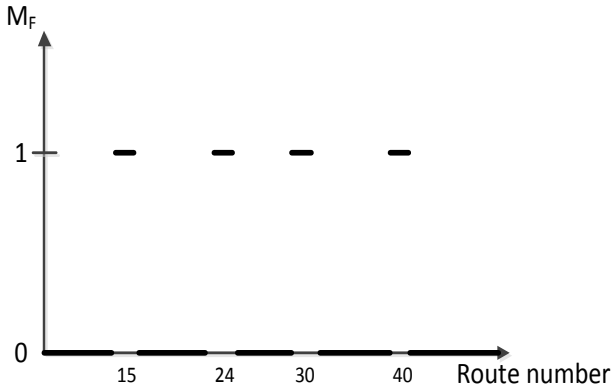

Fig. 2. Representation of imprecise disjunctive data using fuzzy logic

- *Probabilistic imprecise data*

These data are similar to always true imprecise data, since the value should be chosen from a provided discrete set (integer interval). The difference is that the accuracy and certainty of the data values are not equal to one (and may not ever be equal) and thus generates different mathematical probability. It can be said that the value of such data vary within certain limits, furthermore each tag within the following limits has its own chance or probability. For example, let the actual number of years of the planet Earth be in a range from 3 to 6 billion years (Table 4). Then each alternative has its own mathematical probability. Based on the probabilities one must choose one of the options.

Table 4.

Test data

Number of Earth's years, billions of years	Probability
3	70%
4	90%
5	92%
6	73%

- Imprecise interval data

Imprecise interval data means that their values are true on a certain interval rather than a specific point. Visual representation of such a data is shown below in Fig. 3.

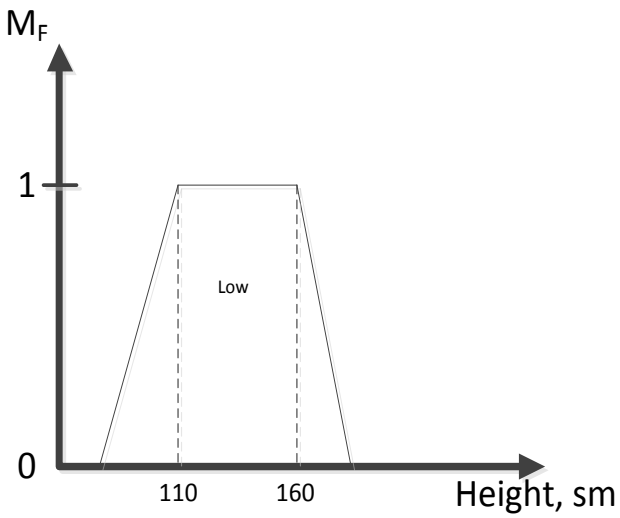

Fig. 3. Visual representation of imprecise interval data using fuzzy logic

- *Imprecise data caused by error margins*

Like imprecise interval data, imprecise data caused by error margins may acquire its values only within a certain interval. But the main difference is that the interval should be a fuzzy singleton. Fuzzy singleton means that only one value from a finite range of values is true and precise. The membership function returns one only for that value, while the membership of the other values is always less than one. Mathematically, this kind of data can be represented as $D \pm \delta$, where D is a value, and δ - allowable and possible error margin.

Suppose there is a power supply with a voltage of 400 ± 10 V. Then this data can be represented visually in Fig. 4.

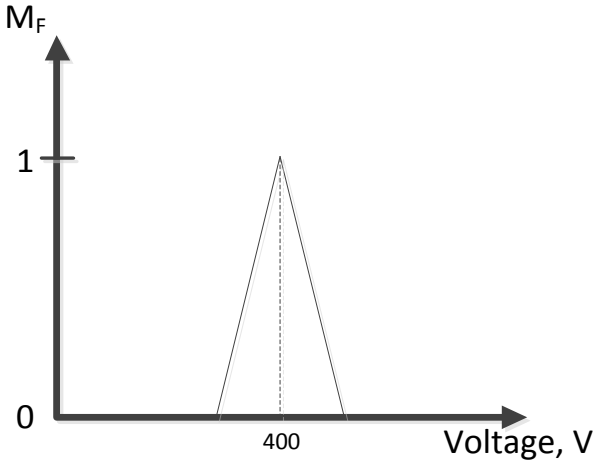

Fig. 4. Representation of imprecise data caused by error margins using fuzzy logic

- *NULL data*

This data is a critical case of imprecise data. NULL values usually refer to the lack of information. The main problem of such data is that they can be interpreted in different ways. The most common interpretations are:

• value is unknown (it exists, but it is unknown);
• value does not exist.

Summarize the results in Table 5 after examining and analyzing non-classical forms of data.

Table 5.

The forms of non-classical data and their characteristics

№	Form of non-classical data	Manifestation in data	The most characteristic features	Source causes
1	2	3	4	5
1	Incomplete data	Incompleteness	The absence or lack of value	• Data acquisition did not take place in time • Lost updates • Lack of values in the data source
2	Inconsistent data	Inconsistency	The same data has a different value in the same or different data sources	• Combining (integrating) information from different data sources • Fallacy of values
3	Uncertain data	Uncertainty	The possibility of truth (falsity) of data is unknown	• Subjective opinions, propositions, etc. • Statistical errors or measurement errors
4	Ambiguous data	Ambiguity	A number of possible interpretations	• Usage of abbreviations • Lack of contextual information • Different words orderings
5	Fuzzy (vague) data	Fuzziness (vagueness)	The meaning, the value of data is unclear, blurred	• Certain subjective representations and evaluations of a fact, phenomena, etc.
6	Imprecise data	Imprecision	The value can not be measured with some pre-defined accuracy (precision)	• Option to choose from a set of values, that have different mathematical probabilities • Value is represented by the interval of values • Possible error margins

The majority of non-classical forms of data can be incorporated in fuzzy knowledge. Fuzzy knowledge is primarily the information obtained from experts. Inaccurate and vague expert opinions and assessments, not completely and precisely defined concepts and terms, possible informality of natural language lead to knowledge fuzziness.

Fuzzy knowledge could be represented as experts assessments, for example: "expiration date of nearly 4-5 years", "old device", "air temperature from 10 to 15" and so on. These estimates can be easily treated as containers for non-classical forms of data. Even more obvious manifestations of non-classical data are seen in fuzzy production rules through which knowledge can be represented in fuzzy knowledge bases. Using fuzzy production rules is a very convenient way to represent fuzzy knowledge by experts and it greatly simplifies their analysis by computational machines.

Fuzzy production rules have the following abstract form:

$$IF < \text{fuzzy proposition} >, THEN < \text{fuzzy proposition} > \quad (1)$$

In (1) fuzzy proposition in conditional part (IF) is called the antecedent, while fuzzy proposition in resulting part (THEN) – the consequent.

Fuzzy propositions can be of two types, namely:

1) atomic fuzzy propositions that contain only one condition. For example, the temperature is low;

2) compound fuzzy propositions that contain *n* conditions connected by conjunctions "AND", "OR", "NOT". For example, temperature is NOT low AND humidity is average.

The most common forms of fuzzy production rules are rules that contain atomic or compound fuzzy propositions in the antecedent part, and only atomic propositions in the consequent part. If the consequent of a fuzzy rule contains a compound fuzzy proposition, then such rule can be decomposed into multiple rules with consequents containing only one atomic fuzzy proposition.

The most used forms are:

1) Canonical fuzzy production rules

They have the following form:

$$\text{Rule } R_i : \textit{IF } x_1 \text{ is } A_1 \textit{ AND } \dots \textit{ AND } x_n \text{ is } A_n, \textit{ THEN } y \text{ is } B \quad (2)$$

where R_i – a tag (a label) of rule *i* in fuzzy knowledge base (here and next),

$x = (x_1, x_2, \dots, x_n)$ – an *n*-dimensional input vector of linguistic variables (here and next),

A, B – the fuzzy sets respectively (here and next),

y – an output linguistic variable.

This form of rule was introduced by M. Mamdani. The main advantage of (2) is clear and transparent linguistic interpretation.

Example: *IF* the road is slippery *AND* the road is steep, *THEN* driving is dangerous.

There are other forms of canonical representation of fuzzy production rules. In the mid 80-ies of XX century Takagi and Sugeno proposed the use of fuzzy production rules that include linear functions in their consequent part [9]. That is, the output result was not defined as a linguistic variable, but as a certain linear function.

$$\text{Rule } R_i : \textit{IF } x_1 \text{ is } A_1 \textit{ AND } \dots \textit{ AND } x_n \text{ is } A_n, \textit{ THEN } y = f_i(x)$$

where $f_i(x) = b_{i0} + b_{i1}x_1 + \dots + b_{in}x_n$,

b_{iq} – a real number (here and next).

Fuzzy rules represented in this format have the property of high proximity to the desired result, however linguistic interpretation of these rules is somewhat worse than (2).

Example: *IF* x_1 is small *AND* x_2 is small, *THEN* $y = 0.7 - 0.6x_1 + 0.8x_2$

Also widely used in classification and pattern recognition tasks the following canonical form of fuzzy rules [10]:

$$\text{Rule } R_i : \textit{IF } x_1 \text{ is } A_1 \textit{ AND } \dots \textit{ AND } x_n \text{ is } A_n, \textit{ THEN } y = \text{Class}_q$$

where Class_q – a class label which the object to be classified or recognized belongs to.

Example: IF x_1 is small *AND* x_2 is big, *THEN* y = Class 1

2) Fuzzy production rules with "OR" conjunction

The form of these rules is described below.

$$\text{Rule } R_i : \textit{IF } x_1 \text{ is } A_1 \textit{ AND } \dots \textit{ AND } x_m \text{ is } A_m \textit{ OR } x_{m+1} \text{ is } A_{m+1} \textit{ AND } \dots \textit{ AND } x_n \text{ is } A_n, \textit{ THEN } y \in B \quad (3)$$

(3) can be divided into the following two fuzzy rules:

IF x_1 is A_1 *AND* … *AND* x_m is A_m, *THEN* y is B

IF x_{m+1} is A_{m+1} *AND* … *AND* x_n is A_n, *THEN* y is B

In general, this format is used in order to reduce the number of rules in the fuzzy knowledge base, since in both cases the consequent part is shared.

Summary

Illustrating a set of non-classical forms of data and fuzzy knowledge, let us summarize the following:

- every non-classical form of data has its own approach for the interpretation and processing that should be considered in such areas as data mining;

- given set of non-classical forms of data is not final and can be expanded in the future;

- fuzzy knowledge tightly intertwined with non-classical data (i.e. the former includes the latter), what affects the interpretation of the former and the possible scope of use of fuzzy production rules.

The taxonomy of forms of fuzzy data and knowledge that has been shown is the basis for the analysis and generalization of data mining methods and tasks that could be applied to fuzzy knowledge bases and will be investigated in the future.

1. Zadeh L.A. Fuzzy Sets // Information and Control, Vol.8, 1965. – P. 338-353. 2. Popat D. Classification of Fuzzy Data in Database Management System / Popat, D., Sherda, H., Taniar, D. // Proceedings of 8th International KES Conference, Wellington, New Zealand, 2004. - P. 691-697. 3. Blanco I.J. About the Use of Ontologies for Fuzzy Knowledge Representation / Blanco, I.J., Marin, N., Martinez-Cruz, C., Vila, M.A. // Proceedings of the Joint 4th Conference of the European Society for Fuzzy Logic and Technology, Barcelona, Spain, 2005. – P. 106-111. 4. Гнатчук Є.Г. Моделювання нечіткого логічного висновку процесу діагностування комп'ютерних засобів // Вісник Вінницького політехнічного інституту. – Вінниця: ВНТУ. – 2005. - №6(63). – С. 220-224. 5. Ma Z.M. A Literature Overview of Fuzzy Database Models / Ma, Z.M., Yan, L. // J. Inf. Sci. Eng. №24, 2008. – P. 189-202. 6. Кравець П. Системи прийняття рішень з нечіткою логікою / П. Кравець, Р. Киркало // Вісник Національного університету "Львівська політехніка". – Львів. – 2009. - №650. – С. 116-123. 7. Motro A. Uncertainty Management in Information Systems: From Needs to Solutions / Motro, A., Smets, P. – Springer, 1997. – 464 p. 8. Parsons S. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases // Knowledge and Data Engineering IEEE, Vol.8, №3, 1996. - P. 483-488. 9. Sugeno M. Fuzzy Identification of Systems and It's Applications to Modeling and Control / Sugeno, M., Takagi, T. // IEEE Trans. On Systems, Man, and Cybernetics №15, 1985. – P. 116-132. 10. Ishibuchi H. Pattern Classification with Linguistic Rules / Ishibuchi, H., Nojima, Y. // Fuzzy Sets and Their Extensions: Representation, Aggregation and Models Studies in Fuzziness and Soft Computing, Vol. 220, 2008. – P. 377-395.