Knowledge density in raw text: a criterion for assessing the usefulness of texts for expert systems

Olegs Verhodubs
oleg.verhodub@inbox.lv

Abstract. Raw (unprocessed) text can serve as a source of rules for a rule-based expert system [1]. Several types of sentences from which rules can be generated have been described [2] [3], but this list is far from exhaustive. A prototype Python package for generating rules from raw text has been developed with great potential for further development [4]. Different texts yield different numbers of generated rules. The more rules that can be generated from a text, the more valuable the text is for an expert system and the greater the likelihood that the user will receive a high-quality answer from the expert system. To evaluate a text for the number of rules it contains, and therefore for its usefulness for an expert system, a criterion called knowledge density is introduced. This paper is devoted to familiarization with the knowledge density criterion. This paper describes: the knowledge density of the whole text, point or local knowledge density, text integrity or the emergence criterion of raw text, as well as various properties of the knowledge density of raw text.

Keywords: knowledge density, knowledge, rules, expert systems, knowledge from text, natural language processing

I. Introduction

Heterogeneity is a fundamental quality of the surrounding world. Any material, substance, or matter possesses heterogeneity, which is the source of diversity. More generally, any sufficiently large set of similar elements exhibits heterogeneity. The profound fundamentality of the principle of heterogeneity lies in its applicability not only to tangible but also to intangible objects. For example, a text of the same length may contain different amounts of information, while a text of different lengths may contain the same amount of information. Knowledge and information are not exactly the same, but they are essentially related, and it is plausible to assert that knowledge is a derivative of information. Knowledge is a product of information processing, but the process of processing does not eliminate heterogeneity: both before and after information processing, heterogeneity remains. Heterogeneity of knowledge consists of its uneven distribution, meaning that one source provides more knowledge, while another provides less. Moreover, knowledge generated from one source may be distributed unevenly within it, namely, such that one part of the source may provide more knowledge than another.

Knowledge generation is crucial because, firstly, it enables the transformation of a large volume of information into a smaller volume of knowledge, and secondly, it enables the use of acquired knowledge in expert systems to generate expert conclusions. Knowledge can be generated from various sources, namely ontologies, databases, and texts. Knowledge generation from texts is particularly promising given the large number of useful texts available on the web. Different texts can generate varying amounts of knowledge, so it is necessary to have a criterion for assessing the potential of a given text for knowledge generation, compared to other similar texts. In other words, a criterion is needed to enable an informed decision about choosing one text over another for generating knowledge from the selected text. This article proposes a knowledge density criterion that allows for the preference of one text over another. The higher the knowledge density, the better, and the more useful a given source is for an expert system.

Moreover, the more knowledge a given text provides, the greater the likelihood of continuous reasoning and the unnecessary need for additional sources for knowledge generation. Here, an expert system is meant to be a Keyword Search Engine Enriched by Expert System Features [5] that is capable of reasoning based on ontologies and texts from the Web, but it is assumed that the subsystem for generating knowledge and the ability to reason based on this knowledge will be useful in various applications.

This paper is organized into several sections. The next section, consisting of two subsections, describes knowledge density itself. Then comes the conclusion.

II. Knowledge density criterion

Knowledge density is discussed here, but the concept of density has long been known and is not the author's invention. Before exploring the concept of knowledge density, let us examine the concept of density in general.

A: The essence of the concept of density

Matter density is a fundamental physical property that is defined as the ratio of mass to volume of a substance [6]. The Greek letter ρ is often used to denote density, but the Latin letter d can also be used for this purpose. Thus, the formula for calculating density looks like this [7]:

$$d = M / V, \quad \text{where } d \text{ is density, } M \text{ is mass, and } V \text{ is volume.} \tag{1}$$

This formula for calculating density (1) is basic and cannot be valid always and under all circumstances. The formula given above is widely used in physics, however, in other fields of science, modified formulas for calculating density can be used, and this applies primarily to chemistry. For example, a completely different formula should be used to calculate the density of solutions. There are also other special states of matter where the basic formula either does not work at all or is inaccurate. Nevertheless, for stable conditions and homogeneous matter, the formula above is entirely valid, which makes it similar to the knowledge density criterion being developed.

Breaking this formula down, we can say that formula (1) represents the ratio of the quantity of something in a limited volume. Matter consists of atoms, and since their number in material objects is difficult to calculate and large numbers are inconvenient to manage, the mass parameter is used, which is directly related to the quantity parameter.

B: Knowledge density

Knowledge can be generated from different types of sources. The possibilities of generating knowledge from ontologies [8] [9], databases [10], and raw texts [2] [3] have already been partially explored, although this work is not yet fully complete. In addition, knowledge can be generated from other sources, which will be the subject of subsequent research. Obviously, knowledge density is related to the source of that knowledge, but this paper will only consider knowledge density in raw text.

The bulk of raw text consists of sentences, and it is the sentence that is the smallest unit from which knowledge is generated. Knowledge can be represented in different ways, but here knowledge is discrete and consists of “IF...THEN” rules. Therefore, the knowledge density of

raw text is defined as the amount of knowledge (rules) per entire text (number of sentences), and as a formula it looks like this:

$$d_t = R / S\,, \tag{2}$$

where d_t is knowledge density in raw text, R is the number of generated rules, and S is the total number of sentences from which knowledge is generated.

A sentence is the smallest unit for generating knowledge in the form of If...Then rules. However, it is possible to generate not just one, but several rules from the single sentence. For example, it is possible to generate several rules namely:

IF *is a sports car* THEN *Lamborghini* (3)

IF *is a fast car* THEN *a sports car* (4)

from the sentence:

Lamborghini is a sports car because a sports car is a fast car. (5)

Since there can theoretically be more rules than sentences from which they are generated, the knowledge density of the raw text (d_t) can be greater than 1.

So, when we want to calculate the knowledge density of a raw text, we get it by dividing the number of generated rules by the total number of sentences in the text. In other words, we take the entire text. However, we can take not the entire text, but a portion of it, thereby obtaining the knowledge density at a specific point in the text. Let us call the knowledge density at a specific point in the text d_{tp}, then the formula for calculating the knowledge density at a specific point in the text will look like this:

$$d_{tp}^{n..m} = R_{tp}^{n..m} / S_{tp}^{n..m}\,, \tag{6}$$

where $d_{tp}^{n..m}$ is knowledge density at a specific point in the text (in other words, local knowledge density), $R_{tp}^{n..m}$ is the number of generated rules in a given range [n..m] of sentences, and $S_{tp}^{n..m}$ the total number of sentences in the given range [n..m] from which the rules were generated. Here, the given range [n...m] denotes which specific sentences are taken from the text. For example, $d_{tp}^{7..10}$ denotes the point density of knowledge in the text, starting from sentences 7 to 10 inclusive.

It must be said that the knowledge density of the entire text can be greater than the sum of all the point knowledge densities of non-intersecting fragments of the text, that is:

$$d_t \geq d_{tp}^{1..i} + d_{tp}^{i+1..j} + \ldots + d_{tp}^{y+1..z}, \tag{7}$$

where z is the number of all sentences in the text, and [1 .. i], [(i+1) .. j] ... [(y+1) .. z] is the set of continuous fragments of sentences.

Inequality (7) primarily refers to the fact that each raw text can yield not only simple rules obtained through simple comparison with patterns defined by parts of speech, but also rules that

express a certain general meaning, or, so to speak, spirit, contained within several sentences. Let us give an example of this. An example of a simple rule and the proposition from which it is derived is shown in Table I. Here we call a simple rule a rule that is generated from a single sentence.

TABLE I. Simple form of the rule

Sentence	Simple rule
An apple is a green fruit.	**IF** is green fruit **THEN** apple

The simple form of the rule in the right column of Table I is generated based on the parts of speech of the sentence in the left column [1].

An example of a complex rule and the proposition from which it is derived is shown in Table II. Here, the so-called complex rule is a rule that is formed on the basis of several sentences in raw text.

TABLE II. Complex form of the rule

Sentence	Complex rule
He did not win a single battle during the entire campaign. He was forced to retreat constantly. During this retreat, he inflicted maximum damage on the enemy. This earned him respect and recognition in professional circles.	**If** you lose the battle **THEN** you can win the war

In general, a classification is possible for complex rules. One group of complex rules are those that are generated from several sentences through logical reasoning. For example, from the sentence group in Table II, one could derive the rule: “if you lose a battle but inflict maximum damage on the enemy, then you earn respect”. The second group of rules are rules that are generated from several sentences, but it is either impossible to derive them explicitly through reasoning, or it is very difficult.

Since not just one sentence but several sentences can serve as the basis for generating rules, and a complex rule can be generated not only through logical reasoning but also as the general meaning, the essence, of these several sentences, it makes sense to introduce a new criterion that will demonstrate the integrity, or cohesion, of the text. Let us call this criterion the integrity of the text and denote it as Int:

$$\text{Int} = d_t - (d_{tp}^{1..i} + d_{tp}^{i+1..j} + \ldots + d_{tp}^{y+1..z}) \quad (8)$$

Formula (8) defines the text integrity criterion as the difference between the knowledge density of the entire text and the sum of the knowledge densities of the text fragments. Clearly, when a text fragment is a sentence, the knowledge density of the entire text is equal to the sum of the

knowledge densities of the text fragments. This means that the text integrity criterion equals zero that is:

$$\text{Int} = 0, \tag{9}$$

when text fragment in the formula (8) is one sentence.

It is well known that any system is greater than the sum of its parts. In other words, the system possesses emergent properties. When applied to raw text, the knowledge density of the entire text is always higher than the sum of the knowledge densities of its fragments. Therefore, the text integrity criterion (Int) can also be conditionally called the emergence criterion of raw text.

It is important to clarify that using the raw text knowledge density criterion has its caveats. Firstly, the raw text knowledge density criterion directly depends on the number of rule templates involved, that is:

$$d_t\,(n) \geq d_t\,(m),\ if\ n > m, \tag{10}$$

where d_t is knowledge density of raw text, n and m are numbers of involved rule templates.

When we talk about rule templates, we mean a specific order of parts of speech in a sentence from which rules can be generated.

Let us clarify this with an example. In paper [4] five types of rules (5 rule templates) were described. So, the knowledge density of five rule templates is greater than or equal to the knowledge density of only three rule templates:

$$d_t\,(5) \geq d_t\,(3), \tag{11}$$

where d_t is knowledge density of raw text.

Moreover, formula (10) degenerates into formula (12) as the number of sentences in the raw text increases.

$$d_{tp}^{1..k}(n) > d_{tp}^{1..k}(m),\ if\ n > m\ and\ k \to \infty, \tag{12}$$

where $d_{tp}^{1..k}$ is point or local knowledge density (knowledge densities of text fragments), n and m are numbers of involved rule templates, k is the number of sentences in the text.

Secondly, the knowledge densities of the same text, but in the case where different rule templates are involved (even if their number is the same), are not equal (13):

$$\begin{cases} d_{tp}^{1..k}(X) \neq d_{tp}^{1..k}(Y), \\ X, Y \subset R, \\ X \neq Y, \\ k \to \infty \end{cases}, \tag{13}$$

where $d_{tp}^{1..k}$ is knowledge density of text fragments, which consists of k sentences, R is a set of rules, X and Y are subsets of R set.

For example, suppose we have a text consisting of 100 sentences. Then the knowledge density based on the first rule type [4] will not be equal to the knowledge density based on the third rule type [4]. To be more precise, theoretically this is not excluded, since everything depends on the text, but in principle the density of knowledge based on different rules is not equal.

III. Conclusion

Raw text can serve as a source of knowledge. More specifically, “IF...THEN” rules can be generated from raw text. These rules can then be used in expert systems to enable even unskilled users to generate expert decisions in various domains.

Different texts have varying potential for generating knowledge in the form of rules. The knowledge density metric was introduced to enable selection of texts with a larger number of rules and a smaller volume among similar texts. Knowledge density in raw text, point or local density in raw text, and related properties and dependencies were discussed in this paper. However, further research in this area remains untapped. For example, one could study the knowledge density of a single sentence and determine the maximum number of rules it can store. Another broad topic is the study of the raw text integrity (in the terms of this paper), that is, how many rules a text can produce, where the basis for each such rule is 2 or more sentences. This is a big topic for research that will be carried out in due course.

Acknowledgments

Gratitude is expressed to friends and family for their long-term moral and material support in the face of numerous illegal and discriminatory actions by the Latvian state, namely the Latvian police, the Latvian prosecutor's office, the Latvian State Security Service, and the Riga Technical University, against the author in particular and against national minorities living in Latvia in general.

The illegal and discriminatory actions by the aforementioned Latvian institutions against the author can be proven by independent, qualified, legal, non-Latvian expert analysis.

References

[1] O. Verhodubs, Prerequisites for Fuzzy Inference on Raw Text Using Semantic Reasoner, 2025.

[2] O. Verhodubs, Rule Mining from Raw Text, 2021.

[3] O. Verhodubs, Experiments of Rule Extraction from Raw Text, 2021.

[4] O. Verhodubs, ruft: A Python Package for Rule Generation from Raw Text, 2025.

[5] O. Verhodubs, Keyword Search Engine Enriched by Expert System Features, 2020.

[6] https://www.sciencedirect.com/topics/physics-and-astronomy/matter-density [Accessed: 07.12.2025]

[7] https://www.britannica.com/science/gram-measurement [Accessed: 07.12.2025]

[8] O. Verhodubs, J. Grundspenkis, Evolution of ontology potential for generation of rules, 2012.

[9] O. Verhodubs, Ontology as a source for rule generation, 2014.

[10] O. Verhodubs, Generation of Rules from the Relational Database Structure, 2023.