

# Automatic clusterization of knowledge bases

Catherine Shchurevich, Elena Kruchkova

*Chair of Applied Mathematics, Altai State Technical University, Lenina str. 46, Barnaul, 656038, Russia*

[scey.barn@gmail.com](mailto:scey.barn@gmail.com), [kruchkova\\_elena@mail.ru](mailto:kruchkova_elena@mail.ru)

**Abstract.** *In the paper are considered intellectual systems having the knowledge bases, generated on the basis of texts in a natural language. The model of systems of such type is described and the algorithm of automatic clusterization of knowledge bases is offered. Bases of clusterization are results of work of an ant algorithm on a semantic network of the knowledge base of the system. The primary analysis of clusterization results it is offered to receive by means of neural networks. Results of testing of the described algorithm are shown, directions of the further work are designated and prospects of a method are considered. The way of visualization of knowledge by means of cutting off the insignificant information and displaying of structure of knowledge on a plane is offered.*

## Keywords

Knowledge base, clusterization, ant algorithm, neural network, knowledge visualization

## 1 Introduction

Now there appear more and more intellectual systems working with texts in a natural language (NL). The area of their application is wide, beginning from search in the Internet, finishing complicated self-trained complexes for the analysis of the information. In such systems a greater role plays their ability to reveal semantic part of the processable text. Complexity of recognition of a semantic component of data is caused, in particular, by impossibility to formalize completely human language, that's why intellectual systems usually strongly depend on the expert training them, especially at the initial stage of development. In the presented work the technique of automatic knowledge clusterization on the basis of texts on NL is considered.

## 2 Theoretical Part

The technique of knowledge clusterization was developed in the assumption, that all phrases of NL can be present in the form of superposition of functions and concepts. Superpositions play a role of communications between the concepts, creating a semantic network for some text or a set of texts. For decomposition of the NL text on superposition with the purpose of extraction of concepts and communications between them we used development of AOT project [1]. A degree of nearness of two concepts of a network we shall define as **distance** — size, inversely proportional to quantity of communications between them, revealed as a result of analysis of the text on NL. In the given work we consider a technique of knowledge clusterization of such network — allocation of groups of the concepts outlining some fields of knowledge.

The offered clusterization process consists of two parts. At the first stage to a system's semantic network of concepts is applied **the ant algorithm** which purpose is to define the areas of a network most closely connected with each other. From these areas we shall try to generate clusters in the further. Ant algorithm is the method of optimization which is based on modeling of behaviour of an ant's colony. Basis of "social" behaviour of ants makes self-organizing — set of the dynamic mechanisms providing achievement by system of the global purpose as a result of low-level interaction of its elements. Basic feature of such interaction is use by elements of system only the local information. Thus any centralized management and the reference to a global image, representing system in an external world is excluded. Self-organizing grows out of interaction of following four components:

- Randomness;
- Recurrence;
- Positive feedback;
- Negative feedback.

As an output of ant algorithm we have the weighted graph of ants' movements over communications of a semantic network and a final arrangement of ants in the graph. Each edge of the graph of movements is more short (length is characterized by a quantity of pheromone left on edge), as more often ants used it for transitions from one concept to another.

At the second stage of clusterization the model of the knowledge base simplifies — communications of a semantic network on which the quantity of pheromone is less than some limited value  $M_0$  are deleted. Next process of cluster formation begins. In the given work we offer two ways of clusterization:

- on the basis of the graph of movements of ant algorithm;
- on the basis of the distance between concepts.

In both cases formation of cluster begins from communication on which there was left maximal, not less than some value  $F_{min}$ , quantity of pheromone in the graph of movements. Concepts which are connected with this communication are marked belonging to one cluster. Then cluster's distribution occurs on the communications which are starting from concepts marked on the previous step of algorithm. For the first way this distribution is limited by the sum of quantity of pheromone on a way from one of initial cluster's concepts up to considered concept, and for the second — by the sum of distances between concepts. When a limit of cluster's distribution will be reached, attempt to form next cluster is undertaken. For this purpose we search communication with quantity of pheromone on it not less  $F_{min}$ , connecting two concepts which still have been not marked by an accessory to earlier generated cluster.

On the basis of data of ant algorithm by means of neural networks it is possible to carry out also the primary analysis of generated clusters to estimate a degree of their completeness, to define “weak” places, to learn possible ways of there development, to refer cluster to some type.

Also in given work is considered the algorithm of visualization of multivariate space of the system's knowledge base by its displaying on a plane. For a filtration of displayed concepts and communications between them are used results of ant algorithm's work, and preservation of proportional distances between concepts during transformation is reached owing to antigradient method.

### 3 Conclusion

In the given research we put before ourselves a task to develop algorithms of clusterization which could work almost without participation of the expert on the empty system's knowledge base, and also on texts of small size.

The presented algorithms of clusterization have been checked up on texts of different stylistics and different size. On short texts (3-5 sentences) the system has shown the results similar to an estimation of the expert in 60% of cases, we consider that as a good factor. As sizes of texts increase so quality indicators grow with different speed depending on stylistics of used articles.

In the further it is planned to automate calculation of values  $M_0$  and  $F_{min}$  depending on fullness of the system's knowledge base and size of the processable text.

Visualization of the system's knowledge base due to intellectual algorithm of reduction of volume of the deduced information in aggregate with obviousness and intuitive clearness facilitates to the expert the analysis of a current condition of the system. At the same time, graphic representation of the text allows to make quick insight about its content without perusal.

### References

- [1] “Automatic text processing” project — <http://www.aot.ru>