ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ

ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

43

ТАРТУ
1980

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

БАНК ДАННЫХ ДЛЯ ЕС ЭВМ

ТРУДЫ
ВЫЧИСЛИТЕЛЬНОГО
ЦЕНТРА

ВЫПУСК 43

ТАРТУ 1980

ВЫДЕЛЕНИЕ НАИБОЛЕЕ СУЩЕСТВЕННЫХ КЛАССОВ ДАННЫХ

К. Ээремаа

1. База данных, как правило, предназначена для хранения большого количества сведений о некоторых объектах и для представления связей между этими объектами. В некоторых случаях исследование всего набора объектов можно заменить исследованием его некоторого, наиболее существенного подмножества. В настоящей статье и рассматривается вопрос о выделении таких наиболее существенных подмножеств объектов. Для этого конструируется система, элементами которой являются рассматриваемые объекты, внутренние связи между элементами выбраны в соответствии с заданными отношениями между объектами, а значимость (вес) каждого элемента оценивается некоторым числом. Если в такой системе выполняется принцип монотонности в определенном ниже смысле, то по теории выделения экстремальных подсистем монотонной системы [2,3] можно найти наибольшее ядро системы, задающее искомое подмножество.

Описываемый ниже метод может быть рассмотрен частным случаем классифицирования объектов, в котором наряду с разбиением на классы искомым является и максимально возможный уровень. В конце статьи показывается, что ядро системы, при соответствующем задании весов элементов можно рассматривать

© *Kuldev Ääremaa*

DETERMINING THE MOST IMPORTANT DATA CLASSES [*]

Kuldev Ääremaa

Abstract. The method described below can be considered a special case of object classification, in which, along with the division into classes, the required level is also the maximum possible level. At the end of the article, it is shown that the kernel of the system, with appropriate assignment of weights of elements, can be considered as a K-cluster (in the sense of Ling, 1971), where k is the maximum number at which clusters are formed.

1. Introduction. The database, as a rule, is designed to store a large amount of information about some objects and to represent the relationships between these objects. In some cases, the study of the entire set of objects can be replaced by the study of some of the most essential subset. This article also considers the issue of identifying such most essential subsets of objects. For this, a system is constructed, the elements of which are the objects under consideration, the internal connections between the elements are selected in accordance with the given relations between the objects, the significance (weight) of each element is estimated by a certain number. If in such a system the principle of monotonicity is fulfilled in the sense defined below, then the theory of singling out extreme subsystems of a Mullat's monotonic system, a) 1971, one can find the largest kernel of the system that defines the desired subset.

The following discussion is mainly based on examples. The analyzed example is selected from the field of information search—methods of setting the links between documents and indices and the selection on their basis of the most essential classes are considered. The presentation of the method on specific examples does not limit the generality—the method is applicable for the analysis of various data structures.

2. Monotone System. First of all, we briefly give the definitions of a monotone system and a kernel of monotone system, and also describe the algorithm for calculating the largest kernel. A more detailed presentation of these questions can be found in Mullat, 1976-1977. Let some finite set of

[*] Translatedd from К. Ээремаа, (1980) ВЫДЕЛЕНИЕ НАИБОЛЕЕ СУЩЕСТВЕННЫХ КЛАССОВ ДАННЫХ, ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, ТРУДЫ ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА, ВЫПУСК 43, ТАРТУ, стр. 74-90.

elements $W = \{x_1, x_2, ..., x_n\}$ be given, on which the weight function $g \equiv g_w$ is defined. The system S is called the pair $S = (W, g)$ and the value $g_w(x)$ is called the weight of the element $x \in W$ in the system S. Let's say that'd for any subset $W' \subseteq W$ the restriction of $g_{w'}$ the weight function g is defined. Then the system $S' = (W', g_{W'})$ is considered a subsystem of the system S. Since the weight function is fixed for the considered system S. then any subsystem S' is determined by the set of its elements W'.

Definition 1. *A system* $S = (W, g)$ *is called monotone if for any two of its subsystems* $S^1 = (W^1, g^1)$ *and* $S^2 = (W^2, g^2)$ *for* $x \in W^2$ *and* $W^2 \subset W^1$ $g_{W^2}(x) \geq g_{W^1}(x) \geq g_w(x)$ *take place or* $g_{W^2}(x) \leq g_{W^1}(x) \leq g_w(x)$. *In the first case, we denote the system by type* $S^{\oplus}$, *and in the second, by type* $S^{\ominus}$.

It is easy to prove that the subsystem of a monotone system is monotone in the same direction. Among all possible subsystems of a monotonic system, of particular interest are those in which the weight function g takes extreme values in the sense defined below. Such subsystems are called system kernels. The kernel of the system can be considered its most essential part.

Definition 2. *The kernel of the system* $S^{\oplus} = (W, g)$ *is called its subsystem* $H^{\oplus}$, *for which the function* $F^{+}(H) = \max\limits_{x \in w'} g_{W'}(x)$ *found among all subsystems* $H = (W', g_{W'})$ *reaches its global minimum, i.e.,* $H^{\oplus} = \arg\min F^{+}(H)$. *Analogically, the kernel of the system* $S^{\ominus} = (W, g)$ *is its subsystem* $H^{\ominus}$ *at which* $F^{-}(H) = \min\limits_{x \in w'} g_{W'}(x)$ *reaches its global maximum,* $H^{\ominus} = \arg\max F^{-}(H)$.

This article further discusses only the system $S^{\ominus} = S = (W, g)$ and, accordingly, the kernel $H^{\ominus} \equiv H = (W', g)$. In cases where the minimum weight of the element of a kernel $s = \min\limits_{x \in w'} g_{W'}(x)$ is emphasized the kernel is denoted by H^S.

The kernel of the system is not necessarily defined unambiguously: the function F^{-} can reach the maximum on several subsystems. Mullat, a) 1976, proved that if $H^1 = (W^1, g^1)$ and $H^2 = (W^2, g^2)$ are the kernels of the given system, then the subsystem $H = (W', g_{W'})$, where $W' = W^1 \cup W^2$ is also a kernel. The union of all the kernels of the system is called the largest kernel.

The algorithm for calculating the largest kernel consists in finding such a numerical value $u \in [L, M]$, where $L = \min_{x \in w} g_W(x)$ and $M = \max_{x \in w} g_W(x)$, which the special $LAYER$ procedure highlights the greatest kernel. The "LAYER" procedure is combined into sequence of "levels" sub-procedures and is described by $LAYER(u, W')$ with $W' \subset W$ as sequential application of the auxiliary procedures $layer(u, W^i)$ as follows:

$$LAYER(u, W') = layer(u, W^n), \text{ where}$$
$$W^i = layer(u, W^{i-1}) = \{x : x \in W^{i-1}, g_{W^{i-1}}(x) > u\},$$

$i = \overline{1,n}$, $W^0 = W'$, and the value of n is determined by the condition $W^n = W^{n+1}$. The algorithm for calculating the largest kernel is now described by the following steps:

1. $L := \min_{x \in W} g_W(x)$, $M := \max_{x \in W} g_W(x)$;
2. $u := \varphi(L, M)$, *where* u *fixed function calculating the value* u;
3. $W' := layer(u, W')$ *if* $W' = \varnothing$ *place* $M := u$ *and return to step 2;*
4. $u := \min_{x \in W'} g_{W'}(x)$;
5. $W'' := layer(u, W')$; *if* $W' \neq \varnothing$ *place* $L := u$ *nd return to step 2 and return to step 2, otherwise the largest kernel* $H^u = (W', g)$ *= is found.*

An example of calculating the largest kernel is given below.

In this article, the allocation of the "sub-kernels" of the largest kernel is considered only in a particular case. The following theorem holds.

Theorem 4. *If in the largest kernel* $H^S = (W', g)$ *of the system* $S = (W, g)$ *there exists a subsystem* $H^1 = (W^1, g)$, *where* $W^1 \subset W'$ *such that for any element* $x \in W^1$ *holds* $g_{W^1}(x) = g_{W'}(x)$, *then* H^1 *is the kernel of the system* S.

To prove the theorem, it must be shown that H^1 is one of those subsystems on which $s^1 = \min_{x \in W^1} g_{W^1}(x)$ reaches the maximum value of s. It is clear that s^1 cannot be less than s. Indeed, if $s^1 < s$ and the value of s^1 is attained for an element $\bar{x}$. In this case by the premises of the theorem $s^1 = g_{W^1}(\bar{x}) = g_{W'}(\bar{x}) < s$. and the value of s is no longer the minimum weight. That means that $s = s^1$ and the subsystem H^1 is the kernel.

3. Documents. Let a set of documents (objects) $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ and a set of indices (attributes) $\mathcal{X} = \{x_1, x_2, ..., x_m\}$ be given. Each document $d \in \mathcal{D}$ is described by a fuzzy set D in the index space (Ääremaa,1979):

$$D = \left\{ x_1 \middle| f(d, x_1), x_2 \middle| f(d, x_2), x_3 \middle|, ..., x_m \middle| f(d, x_m) \right\}, \text{ where}$$

the given membership function f takes values on the segment $[0,1]$

Similarly, each index $x \in \mathcal{X}$ can be associated with a fuzzy set:

$$X = \left\{ d_1 \middle| f(d_1, x), d_2 \middle| f(d_2, x), d_3 \middle|, ..., d_n \middle| f(d_n, x) \right\}$$

and thus the set of documents and indices is described by the matrix $\{D_i\} \times \{X_j\}$, where $i = \overline{1,n}$; $j = \overline{1,m}$. In general, this approach corresponds to the description of a set of objects by some features, in which the role of objects and features is interchangeable.

One of the most important tasks in the field of information retrieval is the division of a set of documents into some thematic classes (Solton, 1979). The matter is easier when there are initial considerations by which you can determine the number of classes and those of each of them. In the general case, it is required to find the classification of documents in advance of unknown topics: the topic of the class is determined during the classification. The latter case is analyzed in this article as well.

Let us first set up a narrower task: to find a set of the most clearly expressed thematic classes. Let us assume that the topic of a document is determined by a set of indices for this document, i.e. a multi-index document description is a thematic description. Then the generality of the topics of two documents $d_i, d_j \in \mathcal{D}$ is characterized by the relationship between these documents by indices and can be calculated by the formula

$$R(d_i, d_k) = \left|D_i \bigcap D_k\right|, \qquad (1)$$

where $D_i \bigcap D_k$ is the intersection of fuzzy sets, i.e.

$$D_i \bigcap D_k = \left\{x_1 \middle| \min\left(f(d_i, x_1), f(d_k, x_1)\right), \ldots, x_m \middle| \min\left(f(d_i, x_m), f(d_k, x_m)\right)\right\}$$

while a $\left|D\right| = \sum_{j=1}^{m} f(d, x_j)$ is the power of the fuzzy set D. In the particular case when documents are described as not fuzzy sets, formula (1) gives the number of common indices in the description of documents.

For each document $d_i \in \mathcal{D}$ we assign the number

$$g_{\mathcal{D}}(d_i) = \sum_{j \neq i} R(d_i, d_j) \qquad (2)$$

depending on d_i and $\mathcal{D}$, and characterizing the thematic connectivity of the document d_i with all other documents d_j.

By this, indeed has been constructed the system $S = \left(\mathcal{D}, g_{\mathcal{D}}\right)$. It is easy to verify that this system is monotonic. Having in mind, the content of the weight function g, the largest kernel of the system, is the desired set of the most related documents. Let's look at an example.

Important Data Classes

Let the set of documents $\mathcal{D} = \{d_1, d_2, ..., d_8\}$ be described by the indices $\mathcal{X} = \{x_1, x_2, ..., x_8\}$ as given in Table 1 (the table indicates the values of the membership function f). Then, using formula (1), it is possible to establish

$\mathcal{D}$ \ $\mathcal{X}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
d_1	0,8	0,9						
d_2	0,4		1,0					
d_3			0,2	1,0				
d_4			0,8	0,2				
d_5		0,3			0,5	1,0		
d_6					0,3	0,2	1,0	
d_7					0,5			1,0
d_8							0,2	0,2

Table 1. "Document-index" matrix

connections between the separate documents (see Fig. 1), and by (2) find the weight of each document. Calculating the largest kernel for the resulting system (see Table 2), we obtain the set $H = \{d_2, d_4, d_5, d_6, d_7\}$ as the kernel. The links between the documents of the largest kernel are shown in Fig. 2. Judging by the figure, the largest kernel splits into two unconnected parts: $W^1 = \{d_2, d_4\}$ and $W^2 = \{d_5, d_6, d_7\}$.

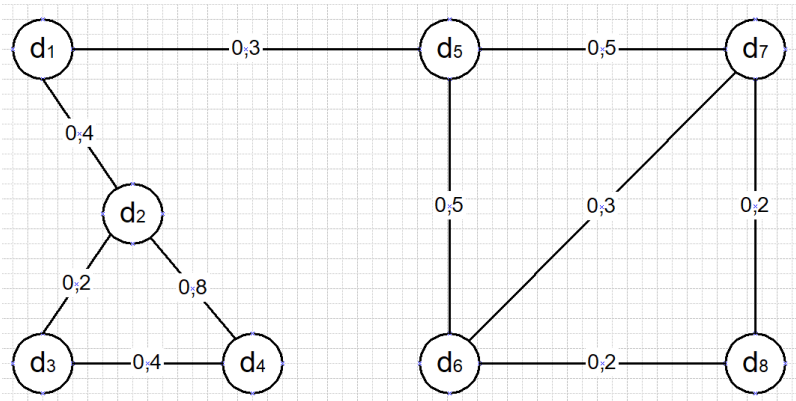


Figure 1. Graph of links between documents

In accordance with the premise of the Theorem 1: $H^1 = (W^1, g)$ and $H^2 = (W^2, g)$ are kernels. The minimum weight of a kernel element is 0.8.

u \ g	$g(d_1)$	$g(d_2)$	$g(d_3)$	$g(d_4)$	$g(d_5)$	$g(d_6)$	$g(d_7)$	$g(d_8)$	Comments
	0,7	1,4	0,6	1,2	1,3	1,0	1,0	0,4	L=0,4;M=1,4;u=(L+M)/2
0,9	-	0,8	-	0,8	1,0	0,8	0,8	-	LAYER(0,9; $\mathfrak{D}$)
		-		-	0	-	-		$\mathfrak{D} = \varnothing$;M:=0,9
0,65	0,7	1,2	-	0,8	1,3	0,8	0,8	-	$\mathfrak{D} \neq \varnothing$; u = min $g_{\mathfrak{D}'}(d_i)$
0,7	-	0,8		0,8	1,0	0,8	0,8		LAYER(0,7; $\mathfrak{D}$) $\neq \varnothing$;L=0,65
0,77	-	0,8	-	0,8	1,0	0,8	0,8	-	LAYER(0,77; $\mathfrak{D}$) $\neq \varnothing$;u=min
0,8		-		-	0	-	-		$\mathfrak{D}$ - kernel

Table 2. Progress of calculating the largest kernel.

But given the prerequisites, each kernel can be considered a thematic class of documents, characterized by the indices used there. In this case, the kernel $\{d_2, d_4\}$ is described in decks x_1, x_3 and x_4, and the kernel $\{d_5, d_6, d_7\}$ by indices x_2, x_5, x_6, x_7 and x_8.

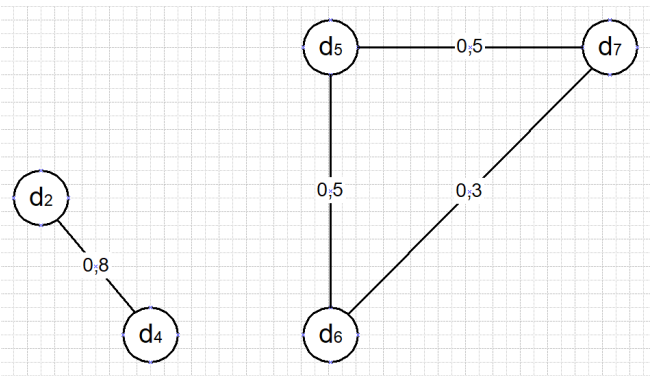


Figure 2. The largest kernel of documents

It is clear that among the indices describing a certain thematic class, there may be "more important" and "less important" indices. To obtain an assessment of the importance of indices, we will consider their joint occurrence in documents. In this case, we obtain formulas similar to formulas (1) and (2)

Important Data Classes

$$R(x_k, x_j) \tag{3}$$

and

$$g_x(x_j) = \sum_{k \neq j} R(x_k, x_j). \tag{4}$$

Formulas (3) and (4) can be used to assess the importance of the indices of the kernel of documents. However, simply "dropping" less important indices in order to find the main topics of the kernel will affect the structure of the kernel and the considered set of documents may no longer be the kernel in the sense of Definition 2.

To avoid such a situation, we construct a system $S = (W, g))$, the elements of which are both documents and indices $(W = \mathcal{D} \cup \mathcal{X})$, and the weight of an element is determined either by the formula (2) or (4):

$$G_w(y) = \begin{cases} g_{\mathcal{D}}(y) \text{ if } y \in \mathcal{D}, \\ g_{\mathcal{X}}(y) \text{ if } y \in \mathcal{X}. \end{cases}$$

It is clear that the system $S = (W, G)$ falls apart subsystems that are superimposed on each other. In the case of the considered example to the structure of documents shown in Fig. 1, the structure of the indices is also added (see Fig. 3). Using the data in the table 1 we see that the largest kernel of the constructed system $S = (W, G)$ is the set $[d_5, d_6, d_7, x_5, x_6]$.

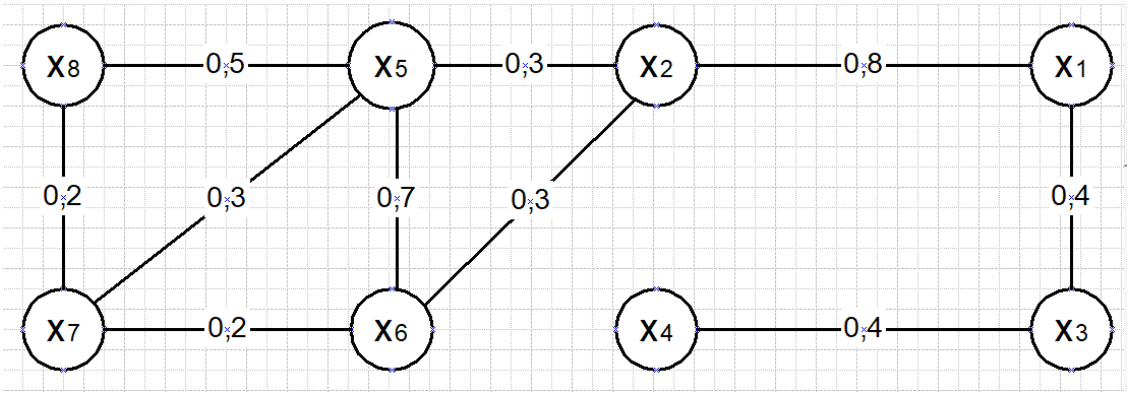


Figure 3. Graph of connections between indexes of table 1.

The resulting kernel (with a minimum weight of 0.7) can be considered as consisting of two parts (see Fig. 4): the set of documents $\{d_5, d_6, d_7\}$ and the set of indices $\{x_5, x_6\}$ characteristic of these documents. We emphasize that in this case neither the set of documents $\{d_5, d_6, d_7\}$ nor the set of indices $\{x_5, x_6\}$, taken separately, are not kernels. The kernel of the system $S = (W, G)$ should be interpreted as the most clearly expressed thematic class $\{x_5, x_6\}$ with the main theme

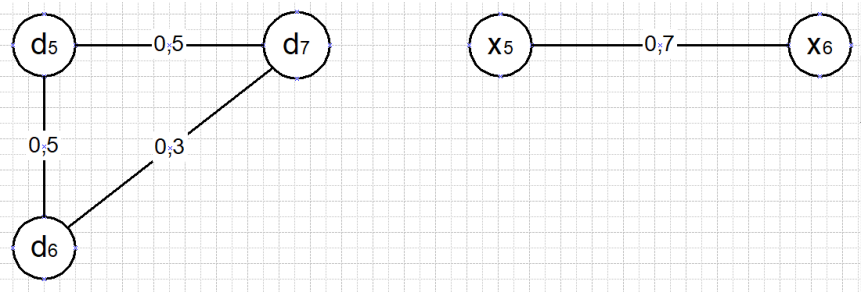


Figure 4. The largest kernel of the system $S = (W, G)$.

The classes of the main topics obtained in the analysis of the system $S = (W, G)$ can be considered in information retrieval systems as associative classes of indices. Obviously, similar index classes can also be obtained by analyzing the system $S = (\mathcal{X}. g_x)$.

4. Sequential kernels. Let the largest kernel $H^{S_1} = (W^1, g)$ be found for the system $S = (W, G)$. Then the process of computing the kernel can be repeated by subsystem $S' = (W \setminus W^1, g)$ since by definition for the function g is defined on any subset and the subsystem of the monotonic system is monotonic. Minimum weight s_2 elements of the largest kernel H^{S_2} of subsystem S'' will, of course, be less than s_1.

As a result of the sequential application of such a process for isolating the largest kernels, a sequence is obtained of kernels $H^{s_1}, H^{s_2}, \ldots, H^{s_m}$ to which there correspond a sequence of sets $W^1, W^2, \ldots, W^m$ and weights $s_1, s_2, \ldots, s_m$. If a certain value of s is fixed, which is considered the level of

classification, then it is natural to terminate the process at such a step k for which $s_k \geq s > s_{k+1}$. In this case, the set is the set $W^1 \bigcup W^2 ... \bigcup W^k$ of minimally "k-essential" elements, on which a partition in terms of "degrees of essentiality" is given.

Turning again to the example considered above, we find for the system $S = (\mathcal{D} \bigcup \mathcal{X}, G)$ the second importance of the kernel. This shows the set of elements $\{d_1, d_2, d_3, d_4, x_1, x_2, x_3, x_4\}$ with a minimum weight of $0{,}4$. Satisfied with the value of the level $0{,}4$, from the set $\mathcal{D} \bigcup \mathcal{X}$ the set $\{d_1, d_2, ..., d_7, x_1, x_2, ..., x_6\}$ with an inner bundle at the level $0{,}7$ was separated. The remaining set $\{d_8, x_7, x_8\}$ can be regarded as the set of the least essential elements.

5. General considerations. In the above presentation, we proceeded from a specific foot example in order to demonstrate the possibilities of applying the theory of monotonic systems to classify objects. This technique, however, can be generalized to any investigated objects for which it is possible to construct a monotonic system. Below, the possibility of such a construction based on the difference matrix is considered in general terms and the identification of the kernels with clusters in the sense of R.F. Ling, 1972.

Let a finite set of objects $W = \{x_1, x_2, ..., x_n\}$ and a difference matrix be given $R = (r(x_i, x_j))$, where $i, j = \overline{1, n}$. Let us associate each element $x_i \in W$ with its weight $g_W(x_i) = F(x_i, R)$ as a function of the matrix R. If, in addition, F is defined in such a way that for any $W' \subset W'' \subseteq W$ with difference matrices R' and R'' respectively, for all $x \in W'$ holds

$$g_{W'}(x) = F(x, R') \leq F(x, R'') = g_{W''}(x)$$

then the pair $S = (W, F)$ defines a monotone system. The meaningful value of the largest kernel of the system S is determined by the semantics of the function F. For example, if we put

$$g^1_W(x_i) = \sum_{i=1}^{n} r_{i,j},$$

then the elements of the kernel of the system are those objects at which the minimum total difference from other objects reaches its maximum value. Thus, the kernel is the set of the most distant objects. On the other hand, if for a fixed r we put

$$g^2_W(x_i) = \sum_{j=1}^{n} r^*_{i,j} \tag{5}$$

where

$$r^*_{i,j} = \begin{cases} 1, \text{if at } i \neq j \ r_{i,j} \leq r, \\ 0. \text{if } r_{i,j} > r \text{ , or } i = j, \end{cases} \tag{6}$$

then the weight of an object x in the system $S = \left(W, g^2\right)$ evaluates (the number of "likelihood") i.e., the differences by which the object do not exceed the specified limit. In this case, some kernel $H^k_i = \left(W, g^2\right)$ represents a subset of objects W_i that have the least k similar ones in W_i. Moreover, by the definition of the weight function, k is an integer and, by the definition of the kernel, it reaches on W_i the maximum value.

By its internal structure, the kernel of the system $S = \left(W, g^2\right)$ resembles a k-cluster built at a fixed level of the difference r with the maximum possible number of connections k. In order to show that the kernel is indeed a k-cluster, let us first recall the definition of the notion of a cluster.

Let a set of objects $W = \{x_1, x_2, ..., x_n\}$ and a difference matrix $R = \left(r(x_i, x_j)\right)$ be given, where $i, j = \overline{1, n}$. Then the subset $W' \subset W$ is called a k-cluster for a given value of r, if:

1° *for any* $x, y \in W$ *there is a chain* $x = x_1, x_2, ..., x_m = y$ *such that* $r(x_i, x_{i+1}) \leq r$, *where* $i = \overline{1, m-1}$;

2° *for any* $x \in W'$ *is at least a subset* $W^X \subset W' \ \left(x \neq W^X\right)$ *of k-elements such that* $r(x.y) \leq r$ *for* $y \in W^X$;

3° *The subset* W^* *is maximal in the sense that there is no set* $W'' \supset W'$ *"* such that conditions 1° and 2° are satisfied on W''.

The above definition can be easily reformulated for the matrix $R^*(r_{ij}^*)$ obtained from the difference matrix at using formula (6). Indeed, in the first condition one should only write an equivalent to the inequality $r(x_i, x_{i+1}) \le r$ the equality $r(x_i, x_{i+1}) = 1$, and in the second $r^*(x, y)$ instead of $r(x, y) \le r$. Thus, as a basis the allocation of clusters and the largest kernel are the same initial data.

It is clear that, in general, the largest kernel is not a k-cluster, since the fulfillment of the condition 1° is not guaranteed at all. Let us first assume that in the particular case the largest kernel $H^k = \left(W', g^2\right)$ is a connected set in the sense of condition 1° in the definition of a cluster. Then, by the definition of the kernel, for any element $x \in W'$ we have

$$g^2_{W'}(x) = \sum_{y \in W'} r^*(x, y) \ge k,$$

herby there might be found at least a k-element set W^X such that $r^*(x, y) = 1$ for $y \in W^X$. Thus, condition 2° in the definition of a cluster is satisfied. The fulfillment of condition 3° is ensured by the fact that W' is the largest set where inequality (7) holds.

Hence, if the set W' is a connected set, then it is a k-cluster. In addition, by the definition of the kernel, there is no subsystem $H^{k'} = \left(w'', g^2\right)$ such that $k' > k$, while by that this connected largest kernel is a cluster with the maximum possible connectivity of the elements.

Now suppose that condition 1° is not satisfied for the largest kernel $H^k = \left(w', g^2\right)$. We split the set of elements W' into subsets $W' = W_1' \bigcup ... \bigcup W_m'$ in such a way that $x, y \in W_i'$ if there is a chain between them defined by condition 1°. It is clear that with such a partition

$W_i' \cap W_j' = \varnothing$ if $i \neq j$. It turns out that in this case the system $s_i = \left(W', g^2\right)$ defined by the set W_i' is the kernel of the system $s = \left(W, g^2\right)$. Indeed, since for any $x \in W_i'$

$$g^2_{W'}(x) = \sum_{i=1}^{m} \left(\sum_{y \in W_i'} r^*(x, y) \right)$$

then the sets W_i' $(i = \overline{1, m})$ satisfy the assumptions of the theorem 1, that is, $H_i^k = \left(W_i', g^2\right)$ is a kernel, and taking into account the construction of the set, it is a connected kernel. For a connected kernel, as was shown above, condition 2° is satisfied, while condition 3° is satisfied by the fact that for any element $x \in W_i'$ there is no $y \in W' \setminus W_i'$ such that $r^*(x, y) = 1$.

Thus, the partition of the largest kernel $H^k = \left(W', g^2\right)$ of the system $S = \left(W, g^2\right)$, where the function g^2 is defined by formula (5), corresponds to the selection of k-clusters of the set W for given r and R. In this case, k is the maximum number of such elements at which clusters are formed.

It should be noted that the system $S' = \left(W \setminus W', g^2\right)$ cannot be used to select clusters at a lower level $k' < k$ as it was done above, since the system S' does not take into account the similarity of elements from $W \setminus W'$ with elements from W'. Therefore, the method of isolating nuclei cannot be consider as a generalization of the method clustering. Both of these methods have their own specifics and their own field of application, although there are some docking points.

Literature

Ääremaa K. A., (1979) The IPS model with a thesaurus based on the theory ofwashed sets. Uch. app. TSU. Proceedings on Artificial Intelligence II. Semantics and knowledge representation. Tartu, 145-155.7.

Ling, R.F., (1972) On the theory and construction of k-clusters. The Computer Journal, 1972, vol. 15, No. 4, 326-332.

Mullat J. E., a) (1971) On a Maximum Principle for certain Functions of Sets, in: Notes on Data Processing and Functional Analysis, Proceedings of the Tallinn Polytechnic Institute (in Russian), Series A, No. 313, Tallinn Polytechnic Institute, pp. 37–44;.
b) (1976) Extremal subsystems of monotone systems. I. Autom. and Telem., 1976, 5, 130-139; II., 1976, 169-178; III. 1977, 109-119;
c) (1977) Andmestruktuuri Avamismeetodid, metoodiline juhend, Tallinna Polütehniline Instituut, Informatsioonitöötlemise kateeder, TRÜ Arhhivikogu;
d) Mullat J. E and L. K. Võhandu, (1979) Monotonic Systems in Scene Analysis, - Symposium, Mathematical Processing of Cartographic Data, Tallinn, pp. 63-66;
e) (1995) A Fast Algorithm for Finding Matching Responses in a Survey Data Table, Mathematical Social Sciences, 30, 195 – 205.

Salton D. Zh., (1979) Dynamic library and information systems. M., 1979.

Vyhandu L.K., (1979) Monotone methods of data analysis. Proceedings of TPI,1979, 468, 15-26. Proceedings No. 468 are devoted to "*Theoretical and Experimental Methods for the Analysis of Optimal Systems of Structural Mechanics*" and have nothing to do with data analysis.