N. A. Stokolova
Philadelphia, Pa.

Paradigmatic Relations (Pt. III of 'Elements of a Semantic Theory of Information Retrieval')

Stokolova, N. A.: **Paradigmatic relations.** Pt. III of 'Elements of a semantic theory of information retrieval'.
In: Intern. Classificat. 4 (1977) No. 1, p. 11–19.
Investigation into the paradigmatic tools of information retrieval systems (IRS) and their role in the algorithmic reproduction of a relevance bigraph. This bigraph is considered as a model of an ideally functioning IRS. An ideal and practically feasible procedure for establishing and quantitatively estimating the usefulness of paradigmatic relations is given. A method for the construction of information languages based on the model described here and in the previous parts I and II is outlined.
(Author)

In parts I and II of this study (1, 2) a model of an ideally functioning information retrieval system (IRS) was considered in the form of an oriented bigraph representing "strict" and "probable" relevance relations on a set T of natural language texts of documents and requests dealt with in an IRS. The function of a real IRS dealing with a given set T is seen to be the algorithmic reproduction of the bigraph of relevance relations on T by processing the "indexes" of documents and requests (i.e. the translation of these texts into an information language (IL)).

Our purpose is to investigate the role of the different semantic components of an IRS in the algorithmic reproduction of the relevance bigraph. In Part II emphasis was laid on the role of the syntax of the IL. In this Part (III) the role of the paradigmatic tools will be considered.

As the result of constructing a satisfactory vocabulary of IL and choosing a suitable syntax for it, the fundamental requirements of the IL will be met; all texts of T will have nonempty translations into the IL_T insofar as all the necessary keywords will be included in the thesaurus of IL_T; all texts of T with identical meanings will have identical translations into IL_T because the suitable quasi-synonymic keywords will be put together in the thesaurus in one set in order to translate them into the same descriptor; all synonymous semantic relations between keywords in texts of T which are expressed as by certain words, as well as by grammatical tools of the natural language, will be translated identically using the syntactic tools of the IL_T which, in particular may form a special part of the vocabulary tools of the IL_T (such vocabulary tools of some ILs are special types of descriptors termed "aspect" descriptors corresponding to predicates or predicates' places) or descriptors corresponding to some multi-word keywords; texts with different meanings will be translated into different representations (particularly by distinguishing of homographs, and by using the appropriate vocabulary and syntactic tools).

Hence all nodes of the graph depicting relevance relations between the texts of T will be correctly reproduced by a graph, where nodes correspond to expressions of T. But in order to reproduce the whole relevance graph correctly, it is necessary to reproduce algorithmically all the arrows of the relevance graph (i.e. to reflect not only the mere existence of all the relevance relations between texts of T but also the corresponding coefficients of probable relevance).

1. A usefulness measure for paradigmatic relations

As was argued in Part I, the relevance relations (and their corresponding comparison rules) between expressions of an IL have to be explicitly defined for each IL. In any IL there are special tools for modelling relevance relations between texts of T, these are the so-called *"paradigmatic relations"*, which are established between descriptors in the thesaurus of the IL. Using paradigmatic relations the semantic relationships between different texts are determined through context independent relationships between words contained in the different texts.

It is necessary to note here that in such a way it is impossible to determine precisely the texts' relations in all cases. Such mistakes are inevitable, for example, in the case of the usage of such a wide-spread paradigmatic relation type as the "whole-part" relationship.

For instance, in the case of extralinguistic situations which can be expressed by the predicate: *"Process* x *takes place with the object* y" – $P_1(x,y)$ – the inference "$P_1(a,b) \rightarrow P_1(a,c)$", where b is a part of c, is true. (Here a,b,c are the descriptors corresponding to the names of the objects; this inference – according to the definition given in Part I – corresponds to the strict relevance relationship of $P_1(a,b)$ to $P_1(a,c)$.)

So the following concrete inferences are true: *"Destruction of the foundation"* (1) → *"Destruction of the building"* (2), *"Diseases of blood vessels"* (3) → *"Disease of the circulatory system"* (4) where *foundation* is a part of *building* and *blood vessels* are part of the *circulatory system.*

As a matter of fact in the text *"Destruction of the foundation* (1)" the following meaning is implied: *"There is a place (or places) in the foundation where destruction is occurring"*. But any place in the foundation is, at the same time, a site of the building (as the foundation is a part of the building) and, therefore, there is a site of the building where destruction is occurring (it is just the meaning of the text (2)). So if the text (1) is true, then the text (2) is true too, i.e. the text (2) can be inferred from the text (1).

In the case of situations which can be expressed by the predicate "Object x is made of material y" – $P_2(x,y)$ – the inference "$P_2(b,d) \rightarrow P_2(c,d)$", where b is a part of c is untrue. So, the following concrete inference may be erroneous: *"The foundation is made of reinforced concrete"* (5) → *"The building is made of reinforced concrete"* (6) (because, for instance, notwithstanding the concrete foundation, the building itself may be made

mainly of bricks). I.e., in the case when between descriptors *"foundation"* and *"building"* in the thesaurus the relation *"a foundation is part of a building"* is established, this relation will cause according to the comparison rules frequently used in IRS the output of text (5) in reply to request (6), which is incorrect and so the precision ratio will decrease. Similarly erroneous results will be obtained with the predicates "Object x has size y" insofar as the size of an object and the size of a part of it are not coinciding, "Object x has shape y" etc.

Nevertheless if this relationship is not established in the thesaurus, then text (1) and (3) will not be included in the output of the IRS in reply to the corresponding requests (2) and (4) accordingly that is also incorrect and will cause a decrease in the recall ratio.

Let us define the notion of the *"usefulness"* of a paradigmatic relation. It is such a quantitative characteristic of any such relationship (for instance of the relation of the descriptor d_p to d_q : $d_p \rightarrow d_q$) that the greater the degree of relevance of the text t_i^p to t_j^q in such pairs (t_x^p, t_y^q) of T whose representations differ only by the corresponding descriptors d_p and d_q, the greater is the usefulness of this paradigmatic relation. Then, if this measure of usefulness of the relationship $d_p \rightarrow d_q$ is recorded in the thesaurus, it will allow the algorithmic determination of the degree of relevance between all such text pairs of T as (t_i^p, t_j^q) (i.e. to determine that t_i^p is relevant to t_j^q and to give the degree of this relationship measured by the coefficient of relevance of t_i^p to t_j^q) by means of the usefulness measure of the paradigmatic relationship $d_p \rightarrow d_q$.

It seems reasonable that the contribution of the value of the usefulness measure of the relationship between a given descriptor pair to the degree of relevance of texts, has to be the same in the case of texts with representation differing only by this descriptor pair as (t_i^p, t_j^q), as well as those with representations differing by several descriptor pairs. This assumption determines the mode of application of usefulness measures within the IRS comparison rules, which are used for the algorithmical determination of relevance degrees (coefficients) between texts from T.

In order to determine the usefulness measure of a given paradigmatic relation (d_p to d_q) – as is obvious from the foregoing discussion – it is necessary to select from T such text pairs (t_i^p, t_j^q) whose representations differ only by this given descriptor pair. Such text pairs will be called *"demonstrative" for the descriptor pair* (d_p, d_q) (because the relevance degree of t_i^p to t_j^q in this case depends just upon the paradigmatic relation $d_p \rightarrow d_q$). The usefulness measure of this paradigmatic relation will be determined through the relevance degrees of all such demonstrative pairs $(t_i^p, t_j^q), (t_k^p, t_l^q), \ldots, (t_m^p, t_n^q)$ of texts of T. These relevance degrees, in their turn, are previously determined by means of the method, based on the explication of the notion of probable relevance described in Part I.

If there is a single demonstrative text pair for d_p, d_q in T – t_i^p, t_j^q – the usefulness measure of the paradigmatic relationship of d_p to d_q – $U^{p,q}$ – is taken as equal to the coefficient of relevance of t_i^p to t_j^q, i.e. to $k_{p,q}^{i,j}$. If there are N such demonstrative pairs in T – it is natural to calculate $U^{p,q}$ as the arithmetical mean (average) of the coefficients of relevance for all these pairs, i.e.

$$U^{p,q} = \frac{\sum_{l=1}^{N} k_{p,q}^{l}}{N} \qquad (\alpha)$$

So returning to our example mentioned above, in order to determine the usefulness of the relation between descriptors *"foundation"* and *"building"* it is necessary to select from T text pairs demonstrative for these descriptors (such are above-mentioned text pairs (1), (2) and (5), (6)). For all such pairs it is necessary to determine their relevance degrees (coefficients) and the usefulness of the relation *"foundation → building"* has to be calculated according to formula (α). As is ovbious from the foregoing discussion of the concrete texts, demonstrative for these descriptors, the usefulness measure of this paradigmatic relationship is not maximal i.e. ($\leqslant 1$), although some of such text pairs (as (1), (2)) are connected by strict relevance relationships.

2. Types of paradigmatic relations

2.1 Paradigmatic relations for modelling strict relevance relationships

The strict relevance relationships between nonsynonymous texts often correspond to such paradigmatic relationships between descriptors as the "species → genus" and "consequence → cause" relations. The first relation is recorded in thesauri as "broader term (BT) – narrower term (NT)" relation.

The inference: "$P_k(a,c) \rightarrow P_k(b,c)$" where a is a species of b (the corresponding paradigmatic relation is "a → b") is true for almost any predicate P_k because the statement $P_k(a,c)$ is often meant in the following way: "there are some a which are in relation P_k to c" i.e. – using logical symbolism – in the sense "$\exists x[P_k(x,c) \,\&\, (x \in a)]$"[1].

In fact one can see, that, if the foregoing statement is true and any element of the class a is an element of the class b (what is true by the definition of the "genus–species" relation) then necessarily there are also some b which are in relation P_k to c which is the intended meaning of the statement $P_k(b,c)$.

So, this kind of paradigmatic relationship is likely to have maximum usefulness measure ($\cong 1$): all (or the overwhelming majority) of the texts which are demonstrative for the corresponding descriptor pairs are connected by the strict relevance relationship.

As previously mentioned, if text t_i is strictly relevant to t_j and t_j is not strictly relevant to t_i then t_j is probably relevant to t_i. One can see that if a is species of b unlike $P_k(a,c)$ which is practically always strictly relevant to $P_k(b,c)$, the text $P_k(b,c)$ will not be strictly relevant to $P_k(a,c)$, so by means of the paradigmatic relation "b is the genus of a" the probable relevance of text $P_k(b,c)$ to text $P_k(a,c)$ is modelled.

The functions of paradigmatic relations "consequence → cause" and – partially – of the relation *"thing (material, process) → its property or characteristics"* are analogous to the function of the relation "species → genus": by means of these relations also strict relevance relationships between texts of T (and the reciprocal relations of probable relevance) can be modelled. Then this relation means *"all* a *have property* b" the inference

"$P_k(a,c) \rightarrow P_k(b,c)$" is true and $P_k(a,c)$ is strictly relevant to $P_k(b,c)$.

2.2 *Paradigmatic relations for modelling probable relevance relationships*

The principal IL tools for modelling probable relevance relations are the so-called *"associative relations"*. This term is used particularly in the UNISIST "Guidelines" (3). Sometimes the associative relations on the one hand and the relations *"cause–consequence"*, *"whole–part"* and *"thing–property"* on the other hand are not distinguished and often are denoted in thesauri by the common reference *"related terms"* (RT). As will be seen, the functions and the meanings of the associative relationships differ from the functions of above-mentioned other relations and it is reasonable to distinguish them.

It is implied from the foregoing explication of the probable relevance relation (the case when text t_i is probably relevant to t_j and reciprocally t_j is probably relevant to t_i), that this relation takes place between such texts (t_i, t_j), which are different incomplete descriptions of some situation (or situations) for which more complete (more precise) descriptions exist in T and the texts of these more complete descriptions are strictly relevant to both t_i and to t_j.

For example, the text *"Chemical reaction of type* a *yielding substances* b *and* c *is carried out in apparatus* e *using catalyst* f" (1) is strictly relevant to the texts:
"Producing substance b *in apparatus* c" (2)
"A chemical reaction of type a *yielding substance* b" (3)
"Producing substance b *using catalyst* f" (4)
The texts (2), (3), (4) are mutually probably relevant, because for each pair containing two of these three texts there is a text (1) which is strictly relevant to both texts of such a pair.

It is seen that text pairs (2), (3); (3), (4); (2), (4) are demonstrative for descriptors a,e; a,f; e,f respectively. So, if paradigmatic relations "a $\rightleftharpoons$ e", "a $\rightleftharpoons$ f", "e $\rightleftharpoons$ f" will be established in the thesaurus using them it will be possible to determine algorithmically the relevance relations (2) $\rightleftharpoons$ (3); (3) $\rightleftharpoons$ (4); (2) $\rightleftharpoons$ (4).

It is seen that these associative paradigmatic relations are relations between the participants of one situation. And the more texts there are in T similar to such texts as (1), which are strictly relevant to a given demonstrative text pair (2) (3) the greater is $k^{2,3}$ and hence the greater is the usefulness measure $U^{a,e}$. I.e. the greater the frequency of co-occurrence of objects a and e in some situations, described in texts from T the more $U^{a,e}$ has to be.

It is necessary to take into consideration that such co-occurrence is not always expressed through the co-occurrence of the corresponding keywords in texts from T (as it was in the case of text (1)).

For example the text *"Sulphuration of aromatic compounds"* (5) is strictly relevant to the texts:
"Chemical production of sulphurated aromatic compounds" (6)
"The application of sulphurators in chemical production" (7)
although text (5) contains neither the descriptor *"sulphurated aromatic compounds"* nor *"sulphurators"*.

The text (5) is strictly relevant to (6) because – as is well known by a chemist – the product of the sulphuration reaction of aromatic compounds are always sulphurated aromatic compounds; and text (5) is strictly relevant to (7) because the reaction mentioned in it when used in production is always carried out in special apparatus, called sulphurators.

From this example one can see that some texts (e.g. (5)) not containing a given keyword pair, can nevertheless be strictly relevant to other texts (e.g. to (6) and (7)), demonstrative for the descriptor pair corresponding to this keyword pair.

On the contrary a text not strictly relevant to two other texts, which are demonstrative for some descriptor pair, can contain both corresponding keywords.

One can see that the inferences, modelled by associative relationships have the following appearance (A):

$$P_k(a^i, b^j) \rightarrow P_k(a^i, c^m) \qquad \text{where } j \neq m$$

i,j,m – the places of variables of the predicate P_k filled in by the descriptors a,b,c and the paradigmatic relation $b \rightarrow c$ is an "associative" one. I.e. the corresponding descriptors in the case of "associative" relations – unlike other paradigmatic relations – are values of different variables in predicates, by means of which the significant situations of a given subject field are described.

In the case of the non-associative paradigmatic relationships the inference, modelled by them, as was shown for "genus–species" "whole–part" and "thing–property" relations, have the following appearance (B):

$$P_k(a^i, b^j) \rightarrow P_k(a^i, c^m) \qquad \text{where } j = m$$

and the paradigmatic relationship $b \rightarrow c$ is a non-"associative" one. As a matter of fact, in this case descriptors b and c are different names of the same object, therefore the objects denoted by descriptors, which are connected by any of the non-associative relations – unlike the case of the associative relations – do not co-occur in any situations.

Not all possible kinds of paradigmatic relationships are exhausted by those which were already discussed ("genus–species", "consequence–cause", "process–apparatus" etc.). There are many other kinds of such relationships. What we are suggesting is to study each of these other kinds of relationships in order to establish what kind of relevance and what kind of inference ((A) or (B)) is modelled by them. This will allow the use for these other relations of the same approximate techniques, (corresponding to inference types either (A) or (B)) which will be described later for the estimation of the usefulness measures.

The algorithmical procedure, described in 1. for recognising paradigmatic relations on the basis of the analysis of the complete graph of relevance for T (and evaluating these relations' usefulness measures) is an ideal procedure, which is not meant to be followed literally.

On the one hand, the ideal algorithm described permits one to use some approximate simplified ways for such calculations. On the other hand, the described semantic typology of inferences, which are the basis of the relevance relationships recorded in the complete relevance graph, permits us to develop the described typology of the paradigmatic relationships and to clarify their different functions in modelling the corresponding inferences; as a result sometimes, particularly for the

majority of "non-associative" relations, one can determine the usefulness measures of paradigmatic relations without tedious calculations.

All the corresponding practical recommendations will be described below in 5.

3. Comparison rules

The aim of establishing paradigmatic relationships between descriptors and recording in the thesaurus measures of their usefulness, is to enable the algorithmic determination of the relevance (and of the relevance degree) between any given arbitrary text pair of T, and hence, the full reconstruction of the complete relevance graph, whose arrows and the coefficients marked on these arrows are depicting all the relevance relationships between texts of T.

The algorithmic determination of the relevance coefficient of a text t_i to a text t_j is accomplished by comparing their representations by $IL_T - n_i$ and $n_j -$ according to formal comparison rules.

These rules depend upon the syntactic tools of IL_T. In the case of using grammar tools, they indicate for each descriptor occurring in the representation n_j of request's text t_j which is the corresponding descriptor of the document's text representation n_i.

For instance, if grammar tools of multiplace predicate type are used, the corresponding descriptors are those occurring in the same places of the same predicates; if "roles" are used, corresponding descriptors are descriptors with the same "role" indicators, etc. But in the case of use of approximate grammar tools, causing "cohesion", it is possible that a descriptor occurring in some different texts (which are "cohered") in different predicate places or even in different predicates, will have as a result of "cohering" a same place, a same "role" or will be in a same predicate in the approximate representations obtained using such an IL. In such cases the approximate grammar will indicate as corresponding one to another such descriptors which really occur in different places of predicates, have different "roles" etc.; (the algorithmic comparison of such erroneously correlated descriptors will cause the precision ratio to decrease). It is natural that in the case of IL_T without grammar such erroneous correlations would increase.

After detecting corresponding descriptor pairs the next step in the algorithmic comparison of text representations is the comparison of such descriptor pairs.

If for each request descriptor an identical corresponding descriptor is found in the document's representation, the document's text is recognised as strictly relevant to the request's text. If the corresponding descriptors are not identical, the paradigmatic relations between descriptors recorded in the thesaurus (and the measures of their usefulness) are used for making the comparison. If a document's descriptor d_k is not connected by any paradigmatic relation to the corresponding descriptor d_l of the request (in other words, if between these descriptors only paradigmatic relations with zero usefulness measures are found) then the document's text is not relevant to this request (relevance degree is zero).

In the intermediary cases when for each request descriptor the corresponding document descriptor is connected to it by a paradigmatic relation with a positive usefulness measure, the document's text is recognised as relevant to the request's text by a degree which has to be calculated as a function of the usefulness measures of all these paradigmatic relations. These usefulness measures, which were calculated through the relevance measures of the text pairs demonstrative for the corresponding descriptor pairs, are naturally considered as measures of the contributions of different corresponding descriptor pairs' relevances to the resulting relevance measure of the text pair.

Insofar as the coefficients of probable relevance and hence, the measures of usefulness of the paradigmatic relations are probability measures of the existence of the true relevance relationships between corresponding texts, it is natural to calculate the above-mentioned function as the aritlmetical product of the usefulness measure of the paradigmatic relations.

An analogous suggestion, concerning the quantitative representation of paradigmatic relations (by using so-called "similarity factors") and calculating the relevance degree of a document as the aritlmetical product of similarity factors of the corresponding descriptor pairs was made by Rolling (4). Documents in an IRS's output are ranked according to their relevance degrees calculated in this way.

Let us denote the descriptors of the request's representation as $d_{j_1}, d_{j_2}, \ldots, d_{j_m}$ and their corresponding document representational descriptors as $d_{i_1}, d_{i_2}, \ldots, d_{i_m}$, then the relevance degree (coefficient) $k^{i,j}$ of document text t_i to request text t_j will be:

$$k^{i,j} = U^{i_1,j_1} \times U^{i_2,j_2} \times \ldots \times U^{i_m,j_m}$$

where $U^{ie,je}$ is the measure of usefulness of the paradigmatic relation $d_{ie} \rightarrow d_{je}$.

The coefficients of relevance are used for ranking the output of the IRS in order of decreasing relevance coefficients, which is supposed to correspond to the order of decreasing relevance of the documents answering a given request.

4. Remarks about the significance of the ideal IRS model

Up to this moment in parts I, II, and the previous sections of this part of our study we have described a model of a perfectly performing IRS using an ideal IL (perfectly meeting the fundamental requirements of IL's) and have drawn from this model an ideal algorithmic procedure for recognising quasi-synonymic and paradigmatic relationships between descriptors.

As any model and any ideal procedure described from it this model and procedure are not intended to be neither interpreted nor applied to real IRS literally. In the next section 5, we are going to outline a practically feasible method of IL constructing, based on the insight into the work of IRS gained from the described model and algoritlm. Before doing so in this section some necessary further evaluating remarks about the essential features of the model and algorithm will be presented.

In order to reveal and evaluate the paradigmatic relations by following literally the procedures described above, one has to begin from an already recorded relevance graph, depicting all the strict relevance relations existing between texts of T. By the procedure described

in section 2.1 of Part I the quasi-synonymic relations between keywords can be found from this graph. Then – according to the procedure described in section I of Part I – we have to reveal the probable relevances and their corresponding coefficients and to transform, using them, the strict relevance graph into the complete relevance graph, which has to be used for revealing the paradigmatic relations between descriptors and for the calculation of their usefulness measures. These paradigmatic relationships and their usefulness measures enable us – using text representations in IL_T – to recreate the original relevance graph, which, – if all the construction were meant literally – doesn't contain anything new, because this graph was already known.

Nevertheless if the file T, being a relatively small one, is statistically representative of a large (and increasing) document file, the results of the procedures mentioned above when carried out on the file T (specifically revealing the paradigmatic relationships and their usefulness measures) will be likely valid for the large file too.

But the literal realisation of these ideal procedures will be very tedious even for small files.

The really important implications of the ideal model we have described, and all ideal procedures drawn from it, consist in the insights gained from them into the nature of relevance, paradigmatic and syntactic relationships. From this point of view, the identification of the relevance relations with inference relations between texts, and the described correspondence between different types of a paradigmatic relation to different types of inference schemes, is essential.

But the single most important practical implications of the ideal model is the fact that the described ideal procedures are valuable reference points for the evaluation of different simpler existing procedures, which are easily recognisable as approximations to the ideal ones. Moreover, some new, more effective practically feasible approximations can be recommended.

5. Outline of a method for constructing IL

It is seen from the foregoing consideration that the most important problems, the decisions concerning which influence IRS performance level (i.e. the adequacy of the recreation of the complete relevance graph) are the following ones:

1. The compiling of the set of keywords (including single and multi-word keywords).
2. The revealing of synonymy and quasisynonymic relations between keywords.
3. The distinguishing of homographs.
4. The establishing of paradigmatic relationships.
5. The constructing of the IL's syntax.

We will consider in this section some methods for solving these problems and will make some new suggestions, drawn from the previously described theoretical model of information retrieval.

5.1 Compiling of the set of keywords

Our[2], and many other authors' experience of IL construction confirms, that an IL and particularly its thesaurus must expediently be constructed on the basis of a representative file of documents and requests. This technique, which is called by Lancaster (5) the *"empirical"* approach has to be supplemented by using the corresponding text books, encyclopaedias etc., i.e. by some elements of the "gestalt" method.

The most difficult problem at the stage of selecting the terms and on the following stage of classifying the terms selected from the representative text file is the estimation of term significance (importance). The privious systematization by estimating the most significant types of information occurring in the corresponding subject field, may be of great benefit for this purpose.

Different techniques of such systematization are recommended by many investigators (6–11, 17).

The approach connected with earlier facet analyses was advocated by Vickery who identified 18 facets and subfacets in his classification for soil science (11).

Campbell (7) has proposed a set of facets (categories) applicable for science and technology.

Aitchison and Gilchrist wrote in (12)

"Prior to term selection, break down the subject field into main groups or facets. This may well have been done already during the definition of the subject field, but finer divisions may be necessary."

The present author's experience in IL construction presented in particular in (13) has also confirmed the usefulness of the previous systematization of information in the subject field, covered by the representative text file.

Nevertheless, instead of the systematization of the terms which are contained in these texts, it was found to be easier and more useful to begin with the systematization of the typical situation (i.e. with the systematization of the facet types) which are described in the text file.

It is easier to estimate the significance of the facts (for the users) than to estimate the terms (or terms' category) significance. Besides, as will be shown below, this systematization of the situations occurring in a particular subject field, proves to be necessary also in the following stages of the IL construction.

After the systematization of the situations, term selection becomes a simpler and less ambiguous procedure.

Our experience in systematization of typical situations in several subject fields (organic chemistry, biology, geology, chemical machine building, and some others) has shown, that it is possible to distinguish the following five main types of significant information in these fields:

(1st) Information aiming at the identification of objects which are dealt with in a given study (chemical substances, mixtures, biological organisms or organs, chemical apparatus and parts of them, etc). Such identification is accomplished by means of denomination of an object and the enumeration of some characteristics of the object, for instance, the chemical structure and physical state of a compound; the organisms' species, its development phase, sex, etc.; apparatus type, size, etc.).

(2nd) Information about newly discovered or newly described properties of the objects identified by the type 1st information (new chemical, physical and other properties of substances and mixtures; different newly described properties of organisms and their organs; details about the apparatus' construction etc.).

(3rd) Information about the process the objects are subject to, or used for, particularly about the modifications of properties described by the type 2nd information (for instance chemical reactions, feeding, digestion,

substance transfer, particularly transportation in organisms or apparatus etc.) and information about characteristics of these processes (for instance process velocity and duration, conditions of realization, place and time etc.).

(4th) Information about the studies or investigations carried out upon objects and processes, described by the 1st, 2nd or 3rd information types, particularly about the method of investigation, instruments used for, etc.

(5th) Information about the comparison of facts, described by the 1st, 2nd or 3rd information types and recognizing different connections between these facts (for instance cause – consequence connection, results of comparison of objects properties or processes characteristics etc.)

The analysis of information contained in the texts of a representative file allows one to recognize the fact types, specific for a given subject field (i.e. to recognize the typical situations in a field corresponding to the foregoing information types).

Examples of information type 3rd specific to chemistry are: – "chemical reaction", "modifications of mixtures' composition", "alteration of substances physical states", "mechanical processes", "physio-chemical processes".

After recognizing these typical situations specific for a given subject field it is necessary to make clear what kinds of significant information are contained in the descriptions of the corresponding specific situations in texts of the file.

For example; it is found that descriptions of chemical reactions include information about reagents, products and by-products; reaction conditions; apparatus used, etc. Descriptions of procedures of chemical analysis (another typical situation in chemistry referred to the 4th information type) include information about the object (chemical compound) to be identified by the analysis, the nature (composition) of the analysed mixture, the method and apparatus used etc.

A simple and standard way for describing such typical situations is the use of some multiplace predicates, the variable places of which are reserved for the different kinds of information included in the situations' descriptions[3]. Some examples of such multiplace predicates were given in section I of Part II of this study. Another possible way of displaying multiplace relations is a graph representation such as the "Structural Formula-Like Representation" used in TOSAR (14, 19).

The multiplace predicates, corresponding to the important situations typical of a given subject field, were called "standard phrases".

The set of terms (keywords) which may be put in a specific place of a standard phrase are made of a particular *"category"* of terms. So to each specific place of a given standard phrase a certain category of terms corresponds, but a single category can correspond to several different places in a particular standard phrase, and/or to different places in several standard phrases.

The total set of keywords which may be put in all the different places of all kinds of standard phrases used for a subject field is the set of keywords which have to be included in the thesaurus of the IL for this field.

Examples of term categories for chemistry are the following ones: chemical compounds; such different kinds of the properties of compounds as: physical properties, chemical properties, biological properties; such physical and physico-chemical states and conditions under which substances and their mixtures can occur as: conditions of temperature, pressure, concentration conditions in solutions; types of physico-chemical processes; types of analytical methods and others.

The way described above of keyword systematization has the following advantages:

A. After the full list of standard phrases and term-categories involved by them for a given subject field is established one can use the following simple and unambiguous criterion for deciding whether single word or multiword combinations should be included as keywords in the IL thesaurus: a single or multi-word term has to be included in the thesaurus if it is contained in one or more texts of the representative file and belongs to some category established as mentioned above. I.e. this term might be put in some place of some standard phrase (or standard phrases) by means of which information, significant for the subject field can be described. Due to this criterion multi-word combinations have to be included as keywords when in the IL there are no grammar tools for the regular description of semantic relations between words of a given word combination.

B. One can review all possible context types for each keyword: these are all the standard phrases in which the corresponding category occurs. These context types are necessary for decisions about synonymy and quasi-synonymic paradigmatic relations between terms, and also in distinguishing homographs. These contexts also allow one to establish precisely the meanings of terms. The use of such contexts will be discussed later.

C. The list of standard phrases is a valuable basis for the construction of the future IL's syntax independently of the concrete grammar type which will be chosen for it.

As was discussed in Part II the semantic analysis which has to be carried out for the objective choice of the syntax type and particularly for the choice of grammar tools, appropriate for a given IL is facilitated by envisaging such a semantically powerful IL as the IL using standard phrases. In particular even when constructing an IL with the simplest syntax (i.e. without grammar tools) the prior establishment of standard phrases will allow one to carry out reasonably the selection of multiword keywords, of predicate type descriptors and other substitutes for grammar tools.

5.2 Establishing synonymy and quasi-synonymy relationship

It was noted among the foregoing considerations that the aim of establishing synonymy and quasi-synonymic relationship between keywords is to avoid the description of identical meanings by different expressions of IL, which would result in a decrease of the recall ratio.

As was shown above the synonymy and quasi-synonymic relationship have to be recognized only if there are such texts in T, in which the given keywords play the same semantic role (i.e. they are the values of the same variables in corresponding standard phrases) and these texts are mutually strictly relevant; as is seen for the establishment of equivalence relations the context analysis is necessary.

For example only context analysis will permit one to distinguish the terms *"vaporization"* and *"evaporation"* – the first term in some given subject field is used only in situations of intentional production of vapour. In other fields these two terms might be used as synonyms.

Only context analysis permits one to establish or not the equivalence relation between, for example, the terms *"heredity"* and *"genetics"*, *"change"*, and *"alteration"*.

This procedure will permit one to establish this kind of relation between some antonyms e.g. such keywords as *"conductivity"*, *"hardness"*, *"dryness"*, *"accuracy"* are often contained in the same contexts as the corresponding keywords *"resistance"*, *"softness"*, *"wetness"*, *"errors"* and the corresponding text pairs prove to be mutually strictly relevant.

It is worthwhile to note that contexts may exist in which such antonyms prove to be non-equivalent. For example the text *"Usage of materials with high hardness"* is not strictly relevant to the text *"Usage of materials with high softness"*. The text *"The ways of reducing the hardness of steel"* is not strictly relevant to the text *"The ways of reducing steel softness"*. In order to decide this issue for given antonyms it is necessary to estimate the proportion of the contexts in which they are equivalent and non-equivalent. Another possible decision in this case is the establishment of some paradigmatic relationships between given antonyms, with a usefulness measure less than 1.

A detailed discussion of the synonymy problem is provided by Sparck Jones (15).

5.3 Distinguishing homographs

The aim of distinguishing homographs in IL consists in avoiding the identical description by an IL of different meanings and the concomitant decrease of the precision ratio.

A characteristic sufficient for the recognition of homographs is that the corresponding keywords belong at the same time to several different categories.

In such a case this keyword can express different meanings, corresponding to the several variant places, to which these different categories correspond.

For example the keyword "MERCURY" can appear in the categories "metals" and "planets". "MERCURY" as a metal can figure in the situations "Substance property" (information type 2nd), "Chemical reaction" and "Modifications of the physical states of substances" (information type 3rd) etc. "MERCURY" as a planet can figure in the situations "Newly found properties of objects" (2nd information type), "Movement processes of objects" (3rd information type) etc.

In order to produce different representations for the texts *"Analysis of Mercury in the Atmosphere"* and *"Analysis of the Atmosphere of Mercury"* it is necessary to distinguish between these homographs. One can see that in the case of an IL with grammar tools, the representations of these two texts would not be identical: although the described situations are identical ("Chemical analysis"), the "roles" of "MERCURY" in them are different.

5.4 Establishing paradigmatic relations

As was discussed above there are two different inference types (see type (A) and (B) in 2.2) which are modelled by means of paradigmatic relationships and hence two different roles which these relationships play in modelling the relevance relation.

Therefore the establishment of paradigmatic relations has to consist of two processes:

(1) the recognizing of "genus–species (hierarchical)", "thing–property", "part–whole", "cause–consequence" relationships;
(2) the establishment of "associative" relationships.

5.4.1 Establishing non-associative relations

As was shown above the typical inference scheme by means of which strict relevance is modelled using a descriptor-type of IL is the following one:

$$P_k(a^i, b^j) \rightarrow P_k(a^i, c^m)$$

where j = m and the paradigmatic relationship b → c is a relation "species b → genus c", "part b → whole c", "thing b → property c (all things b have property c) "consequence b → cause c" etc. Insofar as both the descriptors b,c are values of the same variable in predicate, they belong to the same descriptor category.

In cases when the relation "thing b → property c" is established the descriptor b is usually one that has meaning "thing (material) with a given property", such for instance are the descriptors "conductors", "ferromagnetic materials", "catalyst" etc.

The usefulness measure of any paradigmatic relationship can be calculated by means of formula (α).

Concerning this calculation it is useful to add that it can occur that in T the demonstrative text pairs for some descriptor pairs b,c cannot be found. In such cases it is possible to use instead of demonstrative pairs (in the strict sense of this term defined in section 1) such text pairs, whose representations differ not only by this descriptor pair b,c but also by such other descriptor pairs, for which the usefulness measures were already determined (particularly are equal to 1 as in case of the majority of the relationships "species → genus", "consequence → cause" and others).

Nevertheless on the basis of the knowledge of a given subject field it is possible to simplify this evaluative process recognizing such relationships – having usually the maximum usefulness measure (=1) – as being of one of the following types; "genus–species", "cause–consequence", "thing–property" (in the sense that all such things have this property) and also "whole–part".

But, as it was shown for the relationships "genus–species", "whole–part" (and could be shown for some other of these relationships) their usefulness measures are not always = 1, and therefore some supplementary control is necessary. That is because there can be such a demonstrative text pair t_i^b, t_j^c of T that t_i^b and t_j^c are not relevant one to another although b and c are connected by one of these relationships.

Therefore it is necessary to examine different contexts (i.e. descriptions of different situations) in which the corresponding descriptors can occur.

For descriptors which have to be connected by the "species → genus" relationship it is necessary to check whether their meanings in these contexts correspond to the use of the existential or universal quantifier (see section 2.1).

For the "part → whole" relationship it is necessary to

check how often such predicates occur in texts from T, with which this relationship does not model strict relevance.

In evaluating the usefulness measures of these paradigmatic relationships the frequency of such context pairs, corresponding to the absence of the strict relevance relationship, has to be taken into account. On can see that approximate estimations on the basis of the analysis of contexts from T is simplified if some semantically powerful IL, using multiplace predicates, is previously constructed for T. In this case, it is possible to predict all context types in which a given descriptor can occur: these context types correspond to predicates, in which this term's category is a domain of one or more variables.

After the establishment of the "species b → genus c" "part b → whole c", "thing b → property c", relationships it is possible to estimate approximately the usefulness measures of reciprocal relationships "genus c → species b", "whole c → part b", "property c → thing b": the lower is the total number of such mutually exclusive descriptors b,e, ..., m (corresponding to subdivisions by a same *"facet"*) that each of them is species of c, part of c or thing (material) with property c, the greater is the usefulness measure of the relation c → b.

5.4.2 Establishing associative relationships

In section 1. a technique was described for the exact calculation of the usefulness measures of the paradigmatic relationships.

We will consider here only techniques for approximate evaluation of the usefulness measures of associative relationships based on this technique and also on remarks contained in section 2.2 discussing the role of the associative relationships in modelling probable relevance relationships. According to this discussion the usefulness measure of the associative relationship b ⇢ c has to correspond to the probability that in each situation in which object b is a participant, the object c is a participant too.

For approximate determination of this probability, knowledge of a given subject field is necessary, but only the knowledge of that fragment of this field which is reflected in T need to be used.

A previously compiled list of standard phrases for this fragment of the subject field allows one to select the pairs of co-participants of the most significant situations in this field (corresponding to pairs of positions in these predicates).

For example, for chemistry such are the following place position pairs: "type of chemical reaction – apparatus used for", "apparatus used for – reaction conditions", "mixture separating method – apparatus used for", "reagent – process type", etc. These place positions correspond to certain descriptor categories, so the approximate determination of the usefulness measures of these relationships consists in the comparison of descriptor pairs b,c of these category pairs in order to estimate whether the participation of an object b in a situation corresponding to a certain standard phrase, always implies the participation of the object c.

If the answer is positive the usefulness of the relation b ⇢ c has to be evaluated as = 1 for this situation. If not, it is necessary to estimate approximately the frequency of cases in which when b is a participant of a given situation c is also a participant.

For example, if the descriptor *"benzine"* is a product, *"oil"* is always a raw material in situations of *"Benzine production"*. In the case of hydrolysis reactions, water is always one of the reactants.

As a result of establishing these relationships in the thesaurus, the text *"Benzine production"* will be algorithmically recognized as relevant to the request *"Oil processing"* and the text *"Using water as reactants"* will be included in IRS output to the request *"Hydrolysis reactions"*.

But in other types of situation these relations are not useful. For example the text *"Property of oil"* will be incorrectly included in IRS' output to the request *"Property of benzine"* and the text *"Purification of water"* to the request *"Hydrolysis reactions"*. Therefore in order to estimate correctly the usefulness measure of these relationships it is necessary to take into account the frequency of all situations in which the corresponding two objects do not co-occur.

In the foregoing examples two objects always co-occur in a certain situation (case α) but more often two given objects co-occur in some situations only sometimes (case β).

So, furniture is only sometimes made of wood, sulphuric acid is only sometimes produced from pyrites etc. In these cases it is necessary to estimate how often a given reaction is carried out in a certain apparatus, or a given manufactured article is made of particular material, or a given product is produced by a certain process, etc.

It is possible in this latter case to estimate co-occurrences of the corresponding objects in the same situations by calculating the co-occurrences of corresponding descriptors in the texts of T. In the former case (case α) such a calculation is useless because the fact of such a co-participation of the corresponding objects in a situation is a trivial one for this subject field and therefore very often not all such co-participants of a situation are indicated in texts (for instance in a text describing a hydrolysis reaction, water is not likely to be mentioned, in a text describing benzine production, the oil might not be mentioned as raw material, etc.).

Some of the relationships of the "thing ⇢ property", "material → property" and "process → property" type are modelling the probable relevance relationship like the associative relationship.

This is the case when only some things, materials or processes b have the property c. But unlike associative relations, the corresponding objects in this case are not co-participants of same situations.

In the case (described in 2.1) when all things, materials or processes b have the property c the substitution of descriptor b in some text t_i by the description *"Things with property c"* yields such a text t_j that the text t_i is strictly relevant to t_j.

Unlike this case if only some things, materials or processes b have the property c, the text t_i proves to be probable relevant to t_j obtained by such a substitution. For approximate determination of the usefulness measue of such a paradigmatic relationship b ⇢ c, it is necessary to estimate how often a given material (thing, process) has a certain property. For the determination of the usefulness measure of reciprocal relationship c ⇢ b it is

necessary to estimate how often a material with this given property c is a certain type of material b (and not some other material type, mentioned in texts of T).

In concluding this short treatise of the practical procedure for constructing the semantic tools of IRS we can see that the theoretical insight provided by the presented model enables us to understand much better the essence and the actual goals when applying existing procedures and to suggest on this basis various ways for their improvement. In some cases the theoretical insight enables us to explain why some purely empirically developed existing techniques give highly ambiguous results. For example while the above considerations in 5.4.2, concerning the co-participation of objects in typical situations, offer a theoretical foundation for methods based on term co-occurrence frequency counts ("clustering"); at the same time they do indicate the limitations inherent to such methods. In particular, as it was indicated, neither do the names of objects which co-participate in typical situations co-occur in texts, nor does the co-occurrence of terms within (more or less limited) portions of texts correspond to the real co-participation of the designated objects within typical situations. (The co-participation of the latter type does correspond to really useful associative relationships). Therefore the revealing of really useful associative relationships by mere clustering seems to be unachievable.

Of course by no means is the development of the suggested model and the practical conclusions which can be drawn from it exhausted by our presentation. However it is felt that beyond the exciting possibilities offered by the continuous development of the new information processing and access technologies, in particular by the more and more widely used on-line techniques, the future progress of information retrieval needs a semantic-theoretical framework appropriate to give eventually practical guidance to constructing and evaluating semantic retrieval tools in particular classifications and ILs. The main intention of this study was to bring some support to the view according to which the above goal is achievable.

References

(1) Stokolova, N. A.: The concepts of relevance and information language. Pt. I of "Elements of a semantic theory of information retrieval". In: Inform. Process. & Management 13 (1977) No. 3, p. 153–160.

(2) Stokolova, N. A.: Syntactic tools and semantic power of information languages. Pt. II of "Elements of a semantic theory of information retrieval". In: Intern. Classificat. 3 (1976) No. 2, p. 75–81.

(3) UNISIST. Guidelines for the Establishment and Development of Monolingual Thesauri. Paris: Unesco 1973. 35 p.

(4) Rolling, L.: Compilation of thesauri for use in computer systems. In: Inform. Storage & Retrieval 6 (1970) p. 341–350.

(5) Lancaster, F. W.: Vocabulary Control for Information Retrieval. Washington, D.C.: Information Resources Press 1972. 233 p.

(6) Brenner, E. H., Mulvihill, I. G.: American Petroleum Institute Information Retrieval Project Subject Authority List. In: Bull. de l'Assoc. Intern. des Documentalistes et Techn. de l'Inform. 5 (1966) p. 81–84.

(7) Campbell, D. I.: Making your own indexing system in science and technology. In: Aslib Proc. 15 (1963) p.282–302.

(8) London, G.: A classed thesaurus as an intermediary between textual, indexing, and searching languages. In: Rev. Intern. Doc. 32 (1965) No. 4, p. 145–149.

(9) Mulvihill, J. G., Brenner, E. H.: Faceted organization of a thesaurus vocabulary. In: Proc. Amer. Doc. Inst. 3 (1966) p. 175–183.

(10) Vickery, B. C.: Classification and Indexing in Science. London: Butterworth 1958. 3. 1975. 228 p.

(11) Vickery, B. C.: Faceted classification. London: Aslib 1960.

(12) Aitchison, J., Gilchrist, A.: Thesaurus Construction. A practical manual. London: Aslib 1972. 95 p.

(13) Vleduts, G. E., Stokolova, N. A.: A methodology for the construction of information retrieval thesauri. Moscow: VINITI 1973. 151 p.

(14) Fugmann, R., Nickelsen, H., Nickelsen, I., Winter, J. H.: TOSAR – a system for the structural formula-like representation of concept connections in chemical publications. In: J. Chem. Inform. & Comp. Sci. 15 (1975) p. 52–55.

(15) Sparck Jones, K.: Synonymy and semantic classification. Cambridge, England: Cambridge Language Research Unit 1964. M. L. 170.

(16) Stokolova, N. A., Vleduts, G. E.: On the basic features of an information language for searching according to titles of chemical publications. In: Naučno-techn. Inform. (1966) No. 10, p. 19–25, No. 11, p. 25–30.

(17) Soergel, D.: Indexing languages and thesauri: construction and maintenance. Los Angeles: Melville Publishing Company 1974. 632 p.

(18) Petöfi, J. S.: Some aspects of a multi-purpose thesaurus. In: Intern. Classificat. 1 (1974) No. 2, p. 69–76.

(19) Fugmann, R.: The glamour and the misery of the thesaurus approach. In: Intern. Classificat. 1 (1974) No. 2, p. 76–86.

Notes

1 If the $P_k(a,c)$ statement means *"All* a *are in relation* P_k *to* c", the corresponding inference is untrue; but assertions of this type (with a universal quantifier instead of an existential quantifier) are rare in documents and in information requests.

2 See references 1–6 in (2) and 16.

3 See reference 18.