Personalization of Social Networks: Adaptive Semantic Layer Approach

Alexander Ryjov

Abstract This work describes the idea of an adaptive semantic layer for large-scale databases, allowing to effectively handle a large amount of information. This effect is reached by providing an opportunity to search information on the basis of generalized concepts, or in other words, linguistic descriptions. These concepts are formulated by the user in natural language, and modelled by fuzzy sets, defined on the universe of the significances of the characteristics of the data base objects. After adjustment of user's concepts based on search results, we have "personalized semantics" for all terms which particular person uses for communications with data base or social networks (for example, "young person" will be different for teenager and for old person; "good restaurant" will be different for people with different income, age, etc.).

Keywords Personalization · Adaptive semantic layer · Fuzzy linguistic scales · Measure of fuzziness · Loss of information and information noise for fuzzy data

1 Motivation

Social Networks (SN) is one of the striking phenomena of the last decade. This is one of the dynamic and fast-growing segments of Information and Communications Technologies, which will largely determine the landscape of the industry in the coming decades [1]. SN is very diverse, and we do not have their universally recognized classification for now. They differ in size (ranging from mini networks—for example, corporate or project network to wide area networks such as Facebook, Linked In), the breadth of expertise (from highly specialized—for

A. Ryjov (✉)
Department of Mechanics and Mathematics, Chair MaTIS, Moscow State University, Moscow 119899, Russia
e-mail: alexander.ryjov@gmail.com

W. Pedrycz and S.-M. Chen (eds.), *Social Networks: A Framework of Computational Intelligence*, Studies in Computational Intelligence 526,
DOI: 10.1007/978-3-319-02993-1_2,

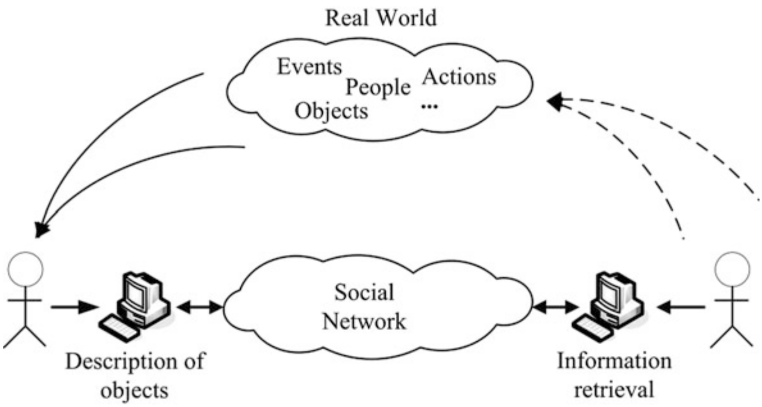


Fig. 1 SN as an information model of real world

example, the Berkeley Initiative in Soft Computing, to general ones—for example, Facebook), life cycle (writing of this book was supported narrow specialized mini networks who "will die" on the day the book will be published; Facebook can accompany a person throughout his conscious life), and many other parameters. Tasks that are important for networks of the particular type may be out of interest to other types of networks.

For sufficiently large networks (most of the participants do not know each other) important is the task of finding the "right" items—partners, restaurants, theaters, things—depending on the "specialization" of the network. Some of these objects are described by the "natural" values (for example, gender, age, price, average bill); other ones—through rubricates, classifiers, category and other tools.

SN is an information model of the real world (Fig. 1). All important from user's point of view objects from a real world are presented in SN as an "information image". We describe the real objects, search their images in SN, and use the results for operation in a real world.

From technical point of view, social networks could be interpreted as a collection of interrelated data bases. The present data processing technology only allows us to search information using concepts (words, symbols, figures) which are present in the data base descriptions of objects. This leads to difficulties in those situations where the information we need is not expressed unequivocally in the language of significances of attributes of object descriptions. The translation of such queries towards the latter search language tends to deform their meaning, and, hence, to reduce the efficiency and the quality of using these data bases.

One of the important properties of the information we need which distinguishes it from the information in the data base, is the fuzziness of the concepts of the user. The user, like any human being, thinks in qualitative categories [13-15], whereas the data base information is basically clear (sharp, non-fuzzy). This is of one of the main problems in the "translation" of user's need of information towards the data base query.

We can illustrate this with the following example.

Example 1 Choosing a car.

We consider the situation of a costumer choosing an car from a data base (electronic catalogue), which contains the following information:

- Price
- Model
- Year of issue
- Fuel consumption.

For the formulation of a request in such a data base, the user is forced to present his query in the language of concrete, definite numbers and models. If our user has in mind a very specific car, for which all the above-mentioned attributes have definite values, and if he has decided that other cars (similar, or for instance a little more economical) are not suitable, the standard technology of data bases solves all his problems. But, he wants to buy a car which is “not very expensive”, “prestigious”, “economical”, and “not old”, the formulation of such a query to the data base is not a very simple task.

A similar example can be given for the task of choosing a flat or other real estate property using a data base, containing the description of particular flats in a city. The processing of queries of the type “comfortable, not very expensive flat, close to a park, and not very far from an underground station” is thus possible. Applications of this approach for political problems (for example, monitoring and evaluation of State’s nuclear activities [5], department of safeguards, International Atomic Energy Agency) have been described in [10].

In general, it can be stated that the problem described above is very important when using automated (electronic) catalogues of goods and services, i.e., data bases, containing particular information about particular objects of the same general type. Especially this way could be effective for several tasks in big data analysis—the next frontier for innovation, competition, and productivity which was introduced by McKinsey Global Institute in May 2011 [1] and was supported by leading companies [2–4].

2 The Concept of an Adaptive Semantic Layer

The structure of an adaptive semantic layer is shown in Fig. 2. Here we use data base as a simplified model of SN. The idea of the adaptive semantic layer is to provide by user an interface which allows:

- define user’s concepts;
- search an information by this concepts;
- adjustment of user’s concepts based on search results.

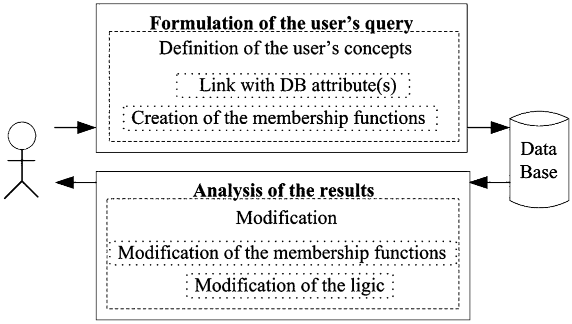


Fig. 2 The structure of an adaptive semantic layer

2.1 *Definition of the User's Concepts*

The work of user begins with a linguistic description of the objects he wants to find in the data base. If the system does not recognise or know this linguistic description, control is transmitted to the program block for the construction of membership functions.[1] If, on the other hand, the system recognises the description, it will retrieve the membership function, associated with it. The user can then, in case of disagreement, edit the membership function, and a new membership function will now be associated with this user, reflecting his "view" (interpretation) of the description. The membership function editor is based on the principle of cognitive graphics and does not require a specialised knowledge of computers.

We can mark out two situations: metric ($U \subseteq R^1$) and non-metric ($U \not\subseteq R^1$) univerums. For both cases we have well-defined methods which were tested in a number of applications of fuzzy models. We can provide the following continuation of example 1:

Example 2

(a) Metric universum: $U =$ [\$10,000, \$50,000]; user's concept $A =$ "not-very-expensive car" is presented in Fig. 3. We can also use fuzzy clustering methods (for instance, Fuzzy C-Means) for building membership functions (Fig. 3).

(b) Non-metric universum: U is a set of all cars' models from the data base; user's concept B = "sport cars for city" is presented in Fig. 4. Using the same way, we can define "cars for hunting", "cars for farmers", "cars for young girls", etc.

Here in right column we have all the cars; green box is collection of cars definitely belongs to user's concept; red box is collection of cars definitely not belongs to user's concept; yellow box is a set of cars which partially belongs to user's concept. We start from empty boxes (all models are in the right column) and split all the models to these three boxes. We can order elements from yellow box

1 Following [15], we associate semantics of the terms (words) with membership functions.

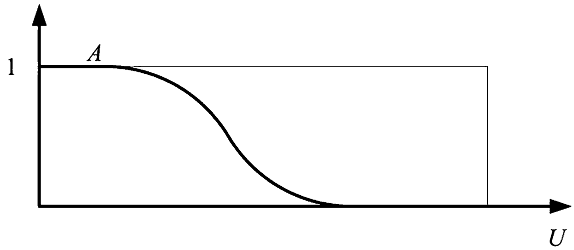


Fig. 3 Example of semantic of user's concept for metric universum

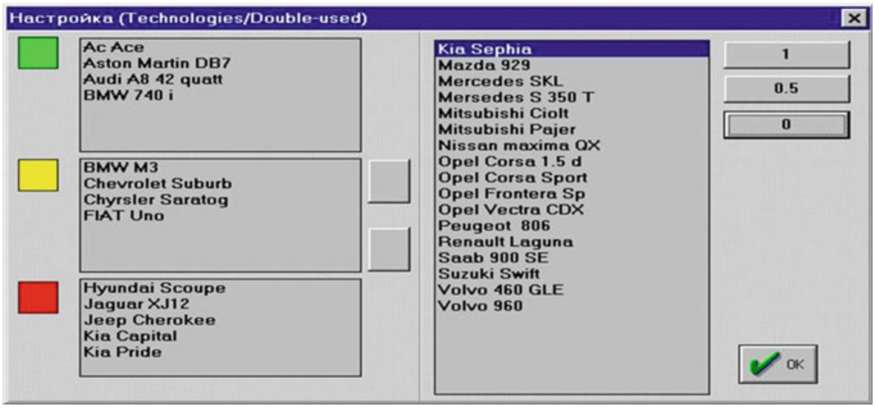


Fig. 4 Example of semantic of user's concept for non-metric universum

according to belongings to user's concept and define membership function as linear one.

After the formulation of the linguistic description, and the association of the membership functions with this description, an information search in the data base can be effected.

2.2 The Information Retrieval Algorithm

The information retrieval algorithm consists in calculating for each record in data base the degree of satisfaction to the formulated request: from 1 (total satisfaction) to 0 (total non-satisfaction). The result of the search is an ordering of the records in the data base on the basis of the degree of satisfaction to the request.

Notations:

- i $(1 \leq i \leq N)$—index of data base attributes;
- U_i—domain of attribute i;

- $A^i = \{A^i_1, \ldots, A^i_{ni}\}$—user's concepts defined on i-attribute ($ni \geq 0$);
- $a^i_{nk}(u_i) = \mu_{A^i_{nk}}(u_i)$—membership function of nk–concept of i-attribute ($nk \leq ni$, $u_i \in U_i$);
- $Q = \langle A^1_{k1} \circ A^2_{k2} \circ \cdots \circ A^N_{kN} \rangle$, where $ki \leq ni$, $A^i_{ki} \in A^i \cup \emptyset (1 \leq i \leq N)$; $\circ \in$ {and, or, not}—is user's concepts based database query.

Algorithm:

1. $r = 0$ (r—index of records in database; $r \leq R$);
2. $r = r + 1$;
3. $i = 0$ (i—index of record's attribute in database);
4. $i = i + 1$;
5. if $A^i(.) \in Q$ then calculate $a^i_{nk}(u_i)$;
6. if $i < N$ then goto 4;
7. calculate $\mu_Q(r) = a^1_{k1} \bullet a^2_{k2} \bullet \cdots \bullet a^N_{kN}$, where $\bullet$ is: t-norm if "o" in Q is "and"; t-conorm if "o" in Q is "or"; $1 - a^i_{nk}(u_i)$ if "o" in Q is "not";
8. if $r < R$ then goto 2.

Result:

$\mu_Q(1)$, …, $\mu_Q(R)$—degree of belonging of each records from database to user's query.

As different classes of users can have different membership functions, the results of the search for the same query can be different for different users, or classes of users. This allows us to have different "views" on the same data base.

2.3 Adjustment of User's Concepts Based on Search Results

It is obvious enough that different users (classes of users) can have different formalization of the concepts (different membership functions). For example, concept "expensive" for student and for businessman can be different. How can we make our interface "personalized"?

In general terms, if we allow using uncertainty at the point of "entry" of the system, we have to provide tools for manipulation of uncertainty at the "output".

We can propose two ways to adjust or tune the interface. First way is adjustment of membership function, second one is tuning of t-notms and t-conorms. The following example can explain this idea.

Example 3

(a) Adjustment of the membership functions is shown in Fig. 5.
    Here *more*, *less*—directions of modification; *a bit*, *not so far*, *a far*,…—volume ("power") of modificators.
    This approach is described in [9].

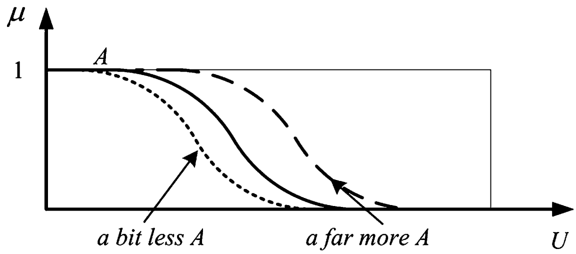


Fig. 5 Adjustment of the semantic of user's concept *A*

(b) Adjustment of the logic (t-norms and t-conorms).
We can use parametric representation of t-norms and t-conorms like

$$T_\lambda(a,b) = \frac{\mu_a \times \mu_b}{\lambda + (1-\lambda)(\mu_a + \mu_b - \mu_a \times \mu_b)}$$

and use genetic algorithms for choosing the best value of λ.

This approach is described in detail in [6].

3 Optimization of Semantic Layer

It is assumed that the person describes the properties of real objects in the form of linguistic values. The subjective degree of convenience of such a description depends on the selection and the composition of such linguistic values. Let us explain this on a model example.

Example 4 Let it be required to evaluate the height of a man. Let us consider two extreme situations.

Situation 1 It is permitted to use only two values: "small" and "high".

Situation 2 It is permitted to use many values: "very small", "not very high",…, "not small and not high",…, "very high".

Situation 1 is inconvenient. In fact, for many men both the permitted values may be unsuitable and, in describing them, we select between two "bad" values.

Situation 2 is also inconvenient. In fact, in describing height of men, several of the permitted values may be suitable. We again experience a problem but now due to the fact that we are forced to select between two or more "good" values. Could a set of linguistic values be optimal in this case?

In SN, one object may be described by different persons. Therefore it is desirable to have assurances that the different participants of the SN describe one and the same object in the most "uniform" way.

On the basis of the above we may formulate the first problem as follows:

Problem 1 Is it possible, taking into account certain features of the man's perception of objects of the real world and their description, to formulate a rule for selection of the optimum set of values of characteristics on the basis of which these objects may be described? Two optimality criteria are possible:

Criterion 1. We regard as optimum those sets of values through whose use man experiences the minimum uncertainty in describing objects.

Criterion 2. If the object is described by a certain number of user's, then we regard as optimum those sets of values which provide the minimum degree of divergence of the descriptions.

This problem is studied in Sect. 3.1. It is shown that we can formulate a method of selecting the optimum set of values of qualitative indications. Moreover, it is shown that such a method is stable, i.e., the natural small errors that may occur in constructing the membership functions do not have a significant influence on the selection of the optimum set of values. The sets which are optimal according to criteria 1 and 2 coincide.

What gives us the optimal set of values of qualitative attributes for information retrieval in a SN (see Fig. 1)? In this connection the following problem arises.

Problem 2 Is it possible to define the indices of quality of information retrieval in fuzzy (linguistic) databases and to formulate a rule for the selection of such a set of linguistic values, use of which would provide the maximum indices of quality of information retrieval?

This problem is studied in Sect. 3.2. It is shown that it is possible to introduce indices of the quality of information retrieval in fuzzy (linguistic) databases and to formalize them. It is shown that it is possible to formulate a method of selecting the optimum set of values of qualitative indications which provides the maximum quality indices of information retrieval. Moreover, it is shown that such a method is also stable.

3.1 Description of Objects for Social Networks

The model of an estimating of real object's properties by a person as the procedure of measuring in Fuzzy Linguistic Scale (FLS) has been analyzed at first time in [11] and described in details in [7]. The set of scale values of some FLS is a collection of fuzzy sets defined on the same universum.

Let us consider t fuzzy variables with the names $a_1, a_2, \ldots, a_t$, specified in one universal set (Fig. 6). We shall call such set the scale values set of a FLS.

Let us introduce a system of limitations for the membership functions of the fuzzy variables comprising s_t. For the sake of simplicity, we shall designate the membership function a_j as μ_j. We shall consider that:

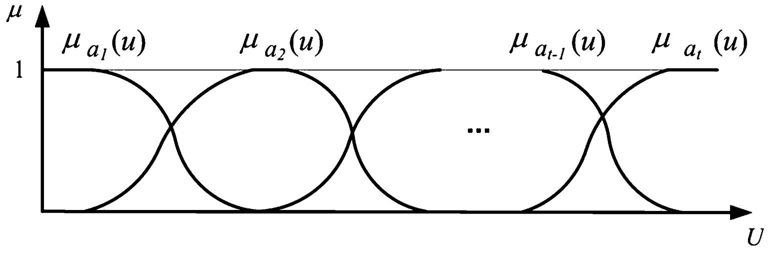


Fig. 6 The scale values set of a FLS

1. $\forall\mu_j(1 \le j \le t)\ \exists\ U_j^1 \ne \emptyset$, where $U_j^1 = \{u \in U : \mu_j(u) = 1\}$, U_j^1 is an interval or a point;
2. $\forall j(1 \le j \le t)\ \mu_j$ does not decrease on the left of U_j^1 and does not increase on the right of U_j^1 (since, according to 1, U_j^1 is an interval or a point, the concepts "on the left" and "on the right" are determined unambiguously).
   Requirements 1 and 2 are quite natural for membership functions of concepts forming a scale values set of a FLS. In fact, the first one signifies that, for any concept used in the universal set, there exists at least one object which is standard for the given concept. If there are many such standards, they are positioned in a series and are not "scattered" around the universe. The second requirement signifies that, if the objects are "similar" in the metrics sense in a universal set, they are also "similar" in the sense of FLS.
   Henceforth, we shall need to use the characteristic functions as well as the membership functions, and so we shall need to fulfil the following technical condition:
3. $\forall j(1 \le j \le t)\ \mu_j$ has not more than two points of discontinuity of the first kind.
   For simplicity, let us designate the requirements 1–3 as L.
   Let us also introduce a system of limitations for the sets of membership functions of fuzzy variables comprising s_t. Thus, we may consider that:
4. $\forall u \in U\ \exists\, j(1 \le j \le t) : \mu_j(u) > 0$;
5. $\forall u \in U \sum_{j=1}^{t} \mu_j(u) = 1$.

Requirements 4 and 5 also have quite a natural interpretation. Requirement 4, designated the *completeness* requirement, signifies that for any object from the universal set there exists at least one concept of FLS to which it may belong. This means that in our scale values set there are no "holes". Requirement 5, designated the *orthogonally* requirement, signifies that we do not permit the use of semantically similar concepts or synonyms, and we require sufficient distinction of the concepts used. Note also that this requirements is often fulfilled or not fulfilled depending on the method used for constructing the membership functions of the concepts forming the scale values set of a FLS [12].

For simplicity, we shall designate requirements 4 and 5 as G.

We shall term the FLS with scale values set consisting of fuzzy variables, the membership functions of which satisfy the requirements 1–3, and their populations the requirements 4 and 5, a *complete orthogonal FLS* and denote it *G(L)*.

As can be seen from example 4, the different FLS have a different degree of internal uncertainty. Is it possible to measure this degree of uncertainty? For complete orthogonal FLS the answer to this question is yes.

To prove this fact and derive a corresponding formula, we need to introduce a series of additional concepts.

Let there be a certain population of t membership functions $s_t \in G(L)$. Let $s_t = \{\mu_1, \mu_2, \ldots, \mu_t\}$. Let us designate the population of t characteristic functions $\hat{s}_t = \{h_1, h_2, \ldots, h_t\}$ as *the most similar population of characteristic functions*, if $\forall j (1 \le j \le t)$

$$h_i(u) = \begin{cases} 1, & \text{if } \mu_i(u) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

It is not difficult to see that, if the complete orthogonal FLS consists not of membership functions but of characteristic functions, then no uncertainty will arise when describing objects in it. The expert unambiguously chooses the term a_j, if the object is in the corresponding region of the universal set. Some experts describe one and the same object with one and the same term. This situation may be illustrated as follows. Let us assume that we have scales of a certain accuracy and we have the opportunity to weigh a certain material. Moreover, we have agreed that, if the weight of the material falls within a certain range, it belongs to one of the categories. Then we shall have the situation accurately described. The problem lies in the fact that for our task there are no such scales nor do we have the opportunity to weigh on them the objects of interest to us.

However we can assume that, of the two FLS, the one having the least uncertainty will be that which is most "similar" to the space consisting of the populations of characteristic functions. In mathematics, distance can be a degree of similarity. Is it possible to introduce distance among FLS? For complete orthogonal FLS it is possible.

First of all, note that the set of functions L is a subset of integrable functions on an interval, so we can enter the distance in L, for example,

$$\rho(f, g) = \int_U |f(u) - g(u)| du, f \in L, g \in L.$$

Lemma 1 Let $s_t \in G(L)$, $s'_t \in G(L)$; $s_t = \{\mu_1(u), \mu_2(u), \ldots, \mu_t(u)\}$, $s'_t = \{\mu'_1(u), \mu'_2(u), \ldots, \mu'_t(u)\}$; $\rho(f, g)$—some distance in L. Then $d(s_t, s'_t) = \sum_{j=1}^{t} \rho\left(\mu_j, \mu'_j\right)$ is a distance in *G(L)*.

The semantic statements formulated by us above may be formalized as follows.

Let $s_t \in G(L)$. For the measure of uncertainty of s_t we shall take the value of the functional $\xi(s_t)$, determined by the elements of *G(L)* and assuming the values in [0,1] $\left(\text{i.e. } \xi(s_t) : G(L) \to [0,1]\right)$, satisfying the following conditions (axioms):

A1. $\xi(s_t) = 0$, if s_t is a set of characteristic functions;

A2. Let $s_t, s'_{t'} \in G(L)$, t and t' may be equal or not equal to each other. Then $\xi(s_t) \leq \xi\left(s'_{t'}\right)$, if $d(s_t, \hat{s}_t) \leq d\left(s'_{t'}, \hat{s}'_{t'}\right)$.

(Let us recall that $\hat{s}_t$ is the set of characteristic functions determined by (1) closest to s_t.).

Do such functional exist? The answer to this question is given by the following Theorem [12].

Theorem 1 *(Theorem of existence)*. Let $s_t \in G(L)$. Then the functional

$$\xi(s_t) = \frac{1}{|U|} \int\limits_U f\left(\mu_{i_1^*}(u) - \mu_{i_2^*}(u)\right) du, \tag{2}$$

where

$$\mu_{i_1^*}(u) = \max_{1 \leq j \leq t} \mu_j(u),\ \mu_{i_2^*}(u) = \max_{1 \leq j \leq t, j \neq i_1^*} \mu_j(u), \tag{3}$$

f satisfies the following conditions:

F1. f(0) = 1, f (1) = 0;

F2. f does not increase—is a measure of uncertainty of st, i.e. satisfies the axioms A1 and A2.

There are many functional satisfying the conditions of Theorem 1. They are described in sufficient detail in [12]. The simplest of them is the functional in which the function *f* is linear. It is not difficult to see that conditions F1 and F2 are satisfied by the sole linear function *f*(*x*) = 1 − *x*. Substituting it in (2), we obtain the following simplest measure of uncertainty of the complete orthogonal FLS:

$$\xi(s_t) = \frac{1}{|U|} \int\limits_U \left(1 - \left(\mu_{i_1^*}(u) - \mu_{i_2^*}(u)\right)\right) du, \tag{4}$$

where $\mu_{i_1^*}(u)$, $\mu_{i_2^*}(u)$ are determined by the relations (3).

We can provide the following interpretation of (4). We consider the process of description by person of a real objects. We do not have any uncertainty in the process of a linguistic description of an object which possessing a "physical" significance of the attribute u_1 (Fig. 7).

We attribute it to term a_1 with total reliance. We can to repeat these statement about an object which have "physical" significance of attribute u_5. We choose the term a_3 for its description without fluctuations. We begin to test the difficulties of choosing of a suitable linguistic significance in the description of an object,

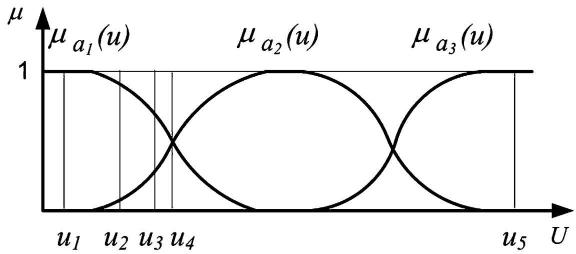


Fig. 7 Interpretation of degree of fuzziness of a FLS

possessing the physical significance of attribute u_2. These difficulties grow (u_3) and reach the maximal significance for an objects, possessing the physical significance of attribute u_4: for such objects both linguistic significances (a_1 and a_2) are equal. If we consider the significance of the integrand function

$$\eta(s_t u) = 1 - \left(\mu_{i_1^*}(u) - \mu_{i_2^*}(u)\right)$$

in these points, we can say, that

$$0 = \eta(s_t, u_5) = \eta(s_t, u_1) < \eta(s_t, u_2) < \eta(s_t, u_3) < \eta(s_t, u_4) = 1$$

Thus, the value of the integral (4) is possible to be interpret as an average human doubts degree while describing some real object.

It is also proved that the functional has natural and good properties for fuzziness degree. In particular the following Theorems then hold [12].

Let us define the following subset of function set L:

$\bar{L}$ is a set of functions from L, which are part-linear and linear on $\bar{U} = \{u \in U : \forall j(1 \le j \le t)\, 0 < \mu_j(u) < 1\}$;

$\hat{L}$ is a set of functions from L, which are part-linear on U (including $\bar{U}$).

Theorem 2 Let $s_t \in G(\bar{L})$. Then $\xi(s_t) = \frac{d}{2|U|}$, where d = $|\bar{U}|$.

Theorem 3 Let $s_t \in G(\hat{L})$. Then $\xi(s_t) = c\frac{d}{|U|}$, where $c < 1$, c = Const.

Adjustment of the membership functions (Sect. 2.3) means theirs displacements to left or right according to directions of modification (*more* or *less*). What will be with our measure of uncertainty after these displacements?

Let g is some biunique function, which is defined on U. This function is induced transformation of some FLS $s_t \in G(L)$ on universum U to FLS $g(s_t)$ on universum U', where $U' = g(U) = \{u' : u' = g(u), u \in U\}$. The above induction is defined by following way: $g(s_t)$ is a set of membership functions $\{\mu_1'(u'), \mu_2'(u'), \ldots, \mu_t'(u')\}$, where $\mu_j'(u') = \mu_j'(g(u)) = \mu_j(g^{-1}(u')) = \mu_j(u), 1 \le j \le t$. The following example illustrates this definition.

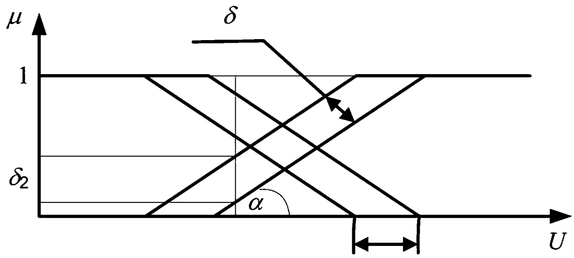


Fig. 8 Presentation of $G^{\delta}(L)$

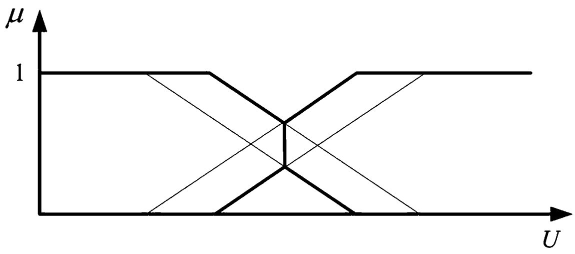


Fig. 9 FLS with minimal degrees of fuzziness

Example 5 Let $s_t \in G(L)$, U is universum s_t, and g is expansion (compression) of universum U. In this case, $g(s_t)$ is a set of functions produced from s_t by the same expansion (compression).

The following Theorem is hold [12].

Theorem 4 Let $s_t \in G(L)$, U is universum s_t, g is some linear function defined on U, and $\xi(s_t) \neq 0$. Then $\xi(s_t) = \xi(g(s_t))$.

Inasmuch as displacement is a subcase of linear function, our measure of uncertainty will be not change during adjustment of the membership functions.

Finally, we present the results of the analysis of our model, when the membership functions which are members of the given collection of fuzzy sets, are not given with absolute precision, but with some maximal inaccuracy δ (Fig. 8). Let us call this particular situation the δ-model and denote it by $G^{\delta}(L)$.

In this situation we can calculate the top $(\overline{\xi}(s_t))$ and the bottom $(\underline{\xi}(s_t))$ valuations of the degree of fuzziness.

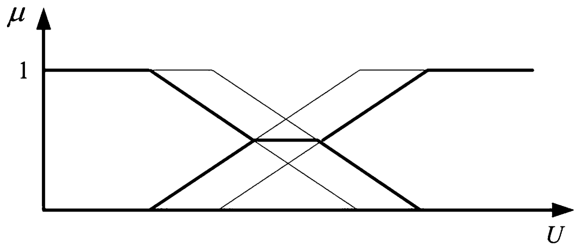
FLS with minimal and maximal degrees of fuzziness for simple case $t = 2$ is shown on Figs. 9 and 10 correspondingly.

The following Theorem is hold [12].

Theorem 5 Let $s_t \in G^{\delta}(\bar{L})$. Then $\underline{\xi}(s_t) = \frac{d(1-\delta_2)^2}{2|U|}$, $\overline{\xi}(s_t) = \frac{d(1+2\delta_2)}{2|U|}$, where $d = |\bar{U}|$.

By comparing the results of the Theorem 2 and 5, we see that for small significances δ, the main laws of our model are preserved. Therefore, we can use our technique of estimation of the degree of fuzziness in practical tasks, since we have shown it to be stable.

Fig. 10 FLS with maximal degrees of fuzziness



Based on these results we can propose the following method for selection of the optimum set of values of characteristics:

1. All the “reasonable” sets of linguistic values are formulated;
2. Each of such sets is represented in the form of $G(L)$;
3. For each set the measure of uncertainty (4) is calculated;
4. As the optimum set minimizing both the uncertainty in the description of objects and the degree of divergence of opinions of users we select the one, the uncertainty of which is minimal.

Following this method, we may describe objects with minimum possible uncertainty, i.e. guarantee that different users will describe the objects for SN in the most possible unified manner (see Criterion 2 in Problem 1). It means that the number of situations when one real object has more than one image in SN, or different real objects have the same image in SN, will be minimal. Accordingly, we will have a maximal possible adequacy of the SN as a model of real world from this point of view. Stability of the measure of uncertainty (Theorem 5) allows us to use this method in practical applications. We also will have an optimal set of values of attributes after adjustments (Sect. 2.3) we need for personalization (see Theorem 4).

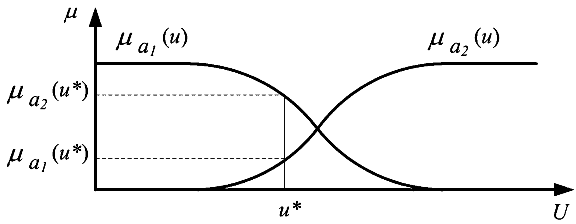
3.2 Modeling of Information Retrieval in Social Networks

SN is an information model of the real world (Fig. 1). The quality of this model is expressed, in particular, through parameters of the information retrieval. If the database containing the linguistic descriptions of objects of a subject area allows to carry out qualitative and effective search of the relevant information then the system will work also qualitatively and effectively.

As well as in Sect. 3.1, we shall consider that the set of the linguistic meanings can be submitted as $G(L)$.

In our study of the process of information searches in data bases whose objects have a linguistic description, we introduced the concepts of loss of information ($\Pi_X(U)$) and of information noise ($H_X(U)$). These concepts apply to information searches in these data bases, whose attributes have a set of significances X, which

Fig. 11 Simple case $t = 2$



are modelled by the fuzzy sets in s_t. The meaning of these concepts can informally be described as follows [8]. While interacting with the system, a user formulates his query for objects satisfying certain linguistic characteristics, and gets an answer according to his search request. If he knows the real (not the linguistic) values of the characteristics, he would probably delete some of the objects returned by the system (information noise), and he would probably add some others from the data base, not returned by the system (information losses). Information noise and information losses have their origin in the fuzziness of the linguistic descriptions of the characteristics.

These concepts can be formalized as follows.

Let's consider the case $t = 2$ (Fig. 11). Let's fix the number $u^* \varepsilon U$ and introduce following denotes:

- $N(u^*)$ is the number of objects, the descriptions of which are stored in the data base, that possess a real (physical, not linguistic) significance equal to $N(u^*)$;
- N^{user}—the number of users of the system.

Then

- $N_{a_1}(u^*) = \mu_{a_1}(u^*)N(u^*)$—the number of data base descriptions, which have real meaning of some characteristic equal $N(u^*)$ and is described by source of information as a_1;
- $N_{a_2}(u^*) = \mu_{a_2}(u^*)N(u^*)$—the number of the objects, which are described as a_2;
- $N_{a_1}^{user}(u^*) = \mu_{a_1}(u^*)N^{user}$— the number of the system's users who believe that $N(u^*)$ is a_1;
- $N_{a_2}^{user}(u^*) = \mu_{a_2}(u^*)N^{user}$—the number of the users who believe that $N(u^*)$ is a_2.

That's why under the request "To find all objects which have a meaning of an attribute, equal a_1" (let's designate it as $\langle I(O) = a_1 \rangle$) the user gets $N_{a_1}(u^*)$ descriptions of objects with real meaning of search characteristic is equal to $N(u^*)$. Under these circumstances $N_{a_1}^{user}(u^*)$ users do not get $N_{a_2}(u^*)$ object descriptions (they carry loses). It goes about descriptions of objects which have the meaning of characteristic equal $N(u^*)$, but described by sources as a_2. By analogy the rest $N_{a_2}^{user}(u^*)$ users get noise ("unnecessary" descriptions in the volume of given $N_{a_1}(u^*)$ descriptions).

Average individual loses for users in the point $N(u^*)$ under the request are equal

$$\pi_{a_1}(u^*) = \frac{1}{N^{user}} N_{a_1}^{user}(u^*) \times N_{a_2}(u^*) = \mu_{a_1}(u^*)\mu_{a_2}(u^*)N(u^*) \tag{5}$$

By analogy average individual noises in the point $N(u^*)$

$$h_{a_1}(u^*) = \frac{1}{N^{user}} N_{a_2}^{user}(u^*) \times N_{a_1}(u^*) = \mu_{a_1}(u^*)\mu_{a_2}(u^*)N(u^*) \tag{6}$$

Average individual information loses and noises, given under analyzed request ($\Pi_{a_1}(U)$ and $\mathrm{H}_{a_1}(U)$ accordingly) are naturally defined as

$$\Pi_{a_1}(U) = \frac{1}{|U|}\int\limits_U \pi_{a_1}(u)du,\ \mathrm{H}_{a_1}(U) = \frac{1}{|U|}\int\limits_U h_{a_1}(u)du$$

It's obvious that

$$\Pi_{a_1}(u^*) = \mathrm{H}_{a_1}(u^*) = \frac{1}{|U|}\int\limits_U \mu_{a_1}(u)\mu_{a_2}(u)N(u)du \tag{7}$$

By analogy for the request $\langle I(O) = a_2\rangle$ or from symmetry considerations we can get that in this case average loses and noises are equal ($\Pi_{a_2}(U) = \mathrm{H}_{a_2}(U)$) too and are equal the right part of (7). Under information loses and noises appearing during some actions with characteristic which has the set of significance $X = \{a_1, a_2\}$ $\big((\Pi_X(U)$ and $\mathrm{H}_X(U)\big)$ and we naturally understand

$$\Pi_X(U) = p_1\Pi_{a_1}(U) + p_2\Pi_{a_2}(U), \mathrm{H}_X(U) = p_1\mathrm{H}_{a_1}(U) + p_2\mathrm{H}_{a_2}(U),$$

where $p_i(i = 1, 2)$—the probability of some request offering in some i—meaning of the characteristic.

It's obvious that as $p_1 + p_2 = 1$, then

$$\Pi_X(U) = \mathrm{H}_X(U) = \frac{1}{|U|}\int\limits_U \mu_{a_1}(u)\mu_{a_2}(u)N(u)du \tag{8}$$

Let's consider general case: t—meanings of the retrieval attribute. We can generalize the formula (8) in case of t meanings of the retrieval attribute the following way [9]:

$$\Pi_X(U) = \mathrm{H}_X(U) = \frac{1}{|U|}\sum_{j=1}^{t-1}(p_j + p_{j+1})\int\limits_U \mu_{a_j}(u)\mu_{a_{j+1}}(u)N(u)du, \tag{9}$$

where $X = \{a_1, \ldots, a_t\}$, $p_i(i = 1, 2, \ldots, t)$—the probability of some request offering in some i—meaning of the characteristic.

Theorem 6 Let $s_t \in G\ (\bar{L})$, *N(u)* = *N* = *Const* and $p_j = \frac{1}{t}$ ($j = 1, \ldots, t$). Then $\Pi_X(U) = \mathrm{H}_X(U) = \frac{ND}{3t|U|}$, where $D = |\bar{U}|$.

Corollary 1 Let the restrictions of the Theorem 6 are true. Then $\Pi_X(U) = \mathrm{H}_X(U) = \frac{2N}{3t}\xi(s_t)$.

For proof of the Corollary is enough to compare Theorems 2 and 6.

We can generalize Corollary 1 for $s_t \in G(L)$. The following Theorem is hold.

Theorem 7 Let $s_t \in G(L)$, *N(u)* = *N* = *Const* and $p_j = \frac{1}{t}$ ($j = 1, \ldots, t$). Then $\Pi_X(U) = \mathrm{H}_X(U) = \frac{c}{t}\xi(s_t)$, where *c* is a constant with depends from *N* only.

The proof of Theorems 6 and 7 are given in [9].

This Theorems showing that the volumes of losses of the information and of information noise arising by search of the information in a SN are coordinated with a degree of uncertainty of the description of objects. It means that describing objects by an optimum way (with minimization of degree of uncertainty) we provide also optimum search of the information in SN.

By analogue with Sect. 3.1, we can construct the top $\left(\overline{\Pi}_X(U)\overline{\mathrm{H}}_X(U),)\right)$ and bottom $(\underline{\Pi}_X(U),\ \underline{\mathrm{H}}_X(U))$ valuations of the $\Pi_X(U)$ and $\mathrm{H}_X(U)$.

The following Theorems and Corollaries are hold [9].

Theorem 8 Let $X = \{a_1, \ldots, a_t\}$, $s_t \in G^{\delta}(\bar{L})$, *N(u)* = *N* = *Const* and $p_j = \frac{1}{t}$ ($j = 1, \ldots, t$). Then

$$\underline{\Pi}_X(U) = \underline{\mathrm{H}}_X(U) = \frac{ND(1-\delta_2)^3}{3t|U|}, \tag{10}$$

where $D = |\bar{U}|$.

Corollary 2 Let the restrictions of the Theorem 8 are true. Then

$$\underline{\Pi}_X(U) = \underline{\mathrm{H}}_X(U) = \frac{2N}{3t}(1-\delta_2)\underline{\xi}(s_t) \tag{11}$$

Theorem 9 Let $X = \{a_1, \ldots, a_t\}$, $s_t \in G^{\delta}(\bar{L})$, *N(u)* = *N* = *Const* and $p_j = \frac{1}{t}$ ($j = 1, \ldots, t$). Then

$$\overline{\Pi}_X(U) = \overline{\mathrm{H}}_X(U) = \frac{ND(1-\delta_2)^3}{3t|U|} + \frac{2ND\delta_2}{t|U|}, \tag{12}$$

where $D = |\bar{U}|$.

Corollary 3 Let the restrictions of the theorem 9 are true. Then

$$\overline{\Pi}_X(U) = \overline{\mathrm{H}}_X(U) = \frac{2N}{t(1+2\delta)}\left[\frac{(1-\delta_2)^3}{3} + 2\delta_2\right]\overline{\xi}(s_t) \tag{13}$$

By comparing the results of the Theorem 6, 8 and 9 or the Corollary 1, 2, and 3, we see that for small significances δ, the main laws of our model of information retrieval are preserved. Therefore, we can use our technique of estimation of the degree of uncertainty and our model of information retrieval in fuzzy (linguistic) data bases in practical tasks, since we have shown it to be stable.

4 Conclusion Remarks

There is no SN without social beings. They are different, because they are human beings. It means that we have to take into consideration human's perception of objects of the real world and manner of their description for SN.

Here we have focused on two important from our point of view issues:

- How we can make our SN more personal, i.e. different for different users?
- How we can make our SN more optimal, i.e. more adequate as a model of a real world we would like to operate in?

This chapter describes only general properties of SN as a model of real world. I do hope that these ideas and results will allow building a maximal comfortable SN for the participants.

Taking an opportunity, I would like to express my thanks to Professor Witold Pedrycz and Professor Shy-Ming Chen for the great idea to prepare this book, valuable remarks, and help in preparation the text.

References

1. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation. (May 2011)
2. IBM. Bringing Big Data to the Enterprise. http://www-01.ibm.com/software/data/bigdata/. (2011)
3. Microsoft Big Data. http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/big-data.aspx. (2012)
4. Oracle and Big Data. Big Data for the Enterprise. http://www.oracle.com/us/technologies/big-data/index.html. (2012)
5. Ryjov, A., Belenki, A., Hooper, R., Pouchkarev, V., Fattah, A., Zadeh, L.A.: Development of an Intelligent System for Monitoring and Evaluation of Peaceful Nuclear Activities (DISNA). IAEA, STR-310, p. 122. Vienna (1998)

6. Ryjov, A., Feodorova, M.: Genetic algorithms in selection of adequate aggregation operators for information monitoring systems. In: Proceedings of V Russian conference on Neurocomputers and its Applications, Moscow, pp. 267–270, February 1999
7. Ryjov, A.: Fuzzy linguistic scales: definition. Properties and applications. In: Reznik, L., Kreinovich, V. (eds.) Soft Computing in Measurement and Information Acquisition, pp. 23–38. Springer, New York (2003)
8. Ryjov, A.: Modeling and optimization of information retrieval for Perception-Based Information. In: Zanzotto, F.; Tsumoto, S.; Taatgen, N.; Yao, Y.Y (eds.) Proceedings of Brain Informatics. International Conference, BI 2012, (Dec. 2012) DOI =http://link.springer.com/chapter/10.1007/978-3-642-35139-6_14 (2012)
9. Ryjov, A.: Models of Information Retrieval in Fuzzy Environment, p. 96. Publishing house of center of applied research, department of mechanics and mathematics, Moscow University Publishing, Moscow (2004)
10. Ryjov, A.: On application of a linguistic modeling approach in information collection for future evaluation. In: Book of Extended Synopses, International Seminar on Integrated Information Systems, Vienna, IAEA-SR-212, pp. 30–34 (2000)
11. Ryjov, A.: The degree of fuzziness of fuzzy descriptions. In: Krushinsky, L.V., Yablonsky, S.V., Lupanov, O.B. (eds.) Mathematical Cybernetics and its Application to Biology, pp. 60–77. Moscow University Publishing, Moscow (1987)
12. Ryjov, A.: The Principles of Fuzzy Set Theory and Measurement of Fuzziness, p. 116. Dialog-MSU, Moscow (1998)
13. Zadeh, L.A.: Computing with words—principal concepts and ideas. Stud. Fuzziness Soft. Comput. Springer, ISBN **277**, 978-3-642-27472-5 (2012)
14. Zadeh, L.A.: From computing with numbers to computing with words—from manipulation of measurements to manipulation of perceptions. Int. J. Appl. Math. Comput. Sci. **12**(3), 307–324 (2002)
15. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning. Part 1, 2, 3. Inform.Sci. **8**, 199–249; **8**, 301–357; **9**, 43–80 (1975)

Springer

http://www.springer.com/978-3-319-02992-4

Social Networks: A Framework of Computational Intelligence
Pedrycz, W.; Chen, S.-M. (Eds.)
2014, XI, 440 p. 168 illus., Hardcover
ISBN: 978-3-319-02992-4