ISSN 1512-1232

UDC: 004.62

CONCEPT PATTERN FORMATION IN SEMANTIC SEARCH PROBLEMS

M.Khachidze[1], M.Tsintsadze[2], M.Archuadze[3], G.Besiashvili[4]
[1,2,3,4] Department of Computer Sciences, Iv.Javakhishvili Tbilisi State University, Georgia
[1]manana.khachidze@tsu.ge
[2]magda.tsintsadze@tsu.ge
[3] maia.archuadze@tsu.ge
[4] gela.besiashvili@tsu.ge

Abstract
In accordance of information flow, the importance of semantic search rises as well. In the article the method of concept pattern formation is presented. Our method in fact represents the Analytic Heuristic method that might be used in semantic search problems successfully. We also have to note that the application of the given method in combination of other conceptual searching methods is encouraged.

Keywords: *Concept Pattern, Semantic Search, Information Retrieval*

1. INTRODUCTION

Nowadays the development and application of information searching algorithms are quite actual, especially when in giant information flow the appropriate queried information is needed to be found. We have two main issues to solve: the given result has to be precise as possible (in a semantic aspect) and it should be given in timely manner (as soon as possible).

From the semantic point of view to receive the precise respond, the query should be presented in such a formal way that its context stays untouched. As the search language plays quite important role, its peculiarity should be taken in to account.

Research hold from 60-ies of last century is still actual nowadays, thus any new or modified method counts in concept application [1] new horizons.

The concept formation [2], Character Recognition and object classification Analytic Heuristic methods are quite interesting. The application of this method in knowledgebase formation for diverse expert system is quite successful [3, 4, 5, 6].

Concept base information retrival is an alternative way based on real word human identification. Here the learning process, the identification of objects and knowledge saving are fulfilled with help of concepts [7]. According to this method the search query is presented in frames of consepts unlike the keywords (BOW method), thus it is less depended on specific terms (keywords).

One of the successful examples of knowledge base representation in Mathematics, using concepts is given in Lenat paper in details [8].

We are for using concepts in text processing as well as in speech. Nowadays the following methods of concept base information search are known: ESA (Explicit Semantic Analysis) [9, 10], WordNet [11], LSA (Latent Semantic Analysis) [12], WikiRelate [13] methods.

From many methods studied by authors, those that partially handle the problem are mainly based on statistical analysis of Data. The method that differs called Explicit Semantic Analysis (ESA) [9, 10]. From our point of view ESA is semantic more precisely representing undefined size natural texts. This method is based on concept formation from Wikipedia resource analysis.

In the presented paper we deal with concept formation method that we have produced on the basis of two method synthesis: the method of Explicit Semantic Analysis [9, 10] and the Analytic Heuristic method [2] , call it The Hybrid Method of Analytic Heuristics (HAH Method). Our method might be used in cooperation of other information searching tools.

ISSN 1512-1232

2. CONCEPT PATTERN FORMATION METHOD

2.1 The method description

Let the $C=\{C_1, C_2, \dots, C_n\}$ class of objects is given. It consists of finite number or nonequivalent objects. Each object C_i, i=1, … , n is described by finite number of properties $A=\{A_1, A_2, \dots, A_m\}$ along with evaluation to which class this object belongs to: $C^+ \subset C$ or $C^- \subset C$. Each property A_j, j=1,…,*m may receive value* $b_{jk} \subset B, k=1,2,\dots,n_j$.

The sorted set of values $A_1, A_2, \dots, A_m$, in case of C_i object definition, call the ”trajectory”. Each object C_i *might be presented with help of appropriate “trajectories”:*

$$C_i=\{b_1(i), b_2(i), \dots, b_m(i)\}, b_j(i) \in B,\ j=1, 2, \dots, m$$

After observation of objects from subclasses, subject should define the notion that is appropriate of C^+ and C^- subclasses. The method of Analytic Heuristics gives possibility to construct pattern appropriate of C^+ and C^- on the basis of evaluated objects.

Pattern formation process consists of the following stages:

I. Property binarization based on its “value set”

For the most general binarisation method the set splitting method might be used. According to this method set is split into two fulfilling each other sets: “is” and “is not”. In this case appropriate notations will be C_i have property (A_k,k = 1, 2, … ,m.) or C_i not have property ($\overline{A}_k$,k = 1, 2, … ,m.).

II. Re-Coding of Properties

Let us introduce the numeric Al-set:

In this case instead of property set $A=\{A_1, A_2, \dots, A_m\}$ we will have : $N_A=\{1, 2, \dots, m\}$,

Instead of value set $B=\{b_{11}, b_{12}, \dots, b_{m\,n_m}\}$ we have: $\underline{N}_B=\left\{\overset{\vee}{1}, \overset{\vee}{2}, \dots, \overset{\vee}{n}\right\}$, where

$$\overset{\vee}{k}=\begin{cases} k \\ \overline{k}, \end{cases} \quad k=1, 2, \dots, n;$$

Instead of „trajectory“ $C_i=\{b_1(i),\ b_2(i), \dots, b_m(i)\}$ we will have $\underline{N}_{C_i}=\left\{\overset{\vee}{\alpha}_1(i), \overset{\vee}{\alpha}_2(i), \dots, \overset{\vee}{\alpha}_m(i)\right\}$, where $\overset{\vee}{\alpha}_j(i) \in \underline{N}_B$, j = 1, 2, … ,m .

III. Orthonormal Binary State Vector Construction

Let’s introduce V matrix with the following dimension: $n \times m$ ($2n = 2^m$). The columns of this matrix represent state orthonormal vectors(the filters) ψ_i , i = 1, 2, … ,m , that is produced via $\underline{N}_B$ elements (Table 1).

ISSN 1512-1232

IV. Filtration Operation - for Each C_i trajectory the orthonormal filter should be applied:

Each trajectory $C_i = \left\{ \overset{\vee}{\alpha}_1(i), \overset{\vee}{\alpha}_2(i), \ldots, \overset{\vee}{\alpha}_m(i) \right\}$ equals to conjunctive product of state orthonormal vectors

$$\varphi(C_i) = \left(\overset{\vee}{\psi}_1 \overset{\vee}{\psi}_2, \ \ldots, \overset{\vee}{\psi}_m \right)_i , \ i = 1, 2, \ldots, n$$

where $\overset{\vee}{\psi}_j = \psi_j$, if the *j*-th element of trajectory belongs to ψ_j vector as „არის“ :

$e_j , j = 1, 2, \ldots, m$

and $\overset{\vee}{\psi}_j = \overline{\psi}_j$, if the *j*-th element of trajectory belongs to ψ_j vector as „არ არის“ :

$\overline{e}_j$, $j = 1, 2, \ldots, m$.

V. The operation of Disjunctive Superposition.

$$\varphi_+ = \bigcup_{C_i \in C^+} \varphi(C_i) \quad \text{in the case of} \quad C^+ \quad \text{pattern} \quad (1)$$

$$\varphi_- = \bigcup_{C_i \in C^-} \varphi(C_i) \quad \text{in the case of} \quad C^- \quad \text{pattern} \quad (2)$$

If

- the number of objects in case of C^+ and C^- subclasses is enough big,
- the objects are non-identical and are widely representing the appropriate subclass,
- the enough number of properties and their value set are defined correctly and the binarization of these sets are made successfully

then patterns φ_+ and φ_- are containing the full information on C^+ and C^-, and are in no opposition to each other.

In case of great *n* and *m* it is impossible to describe pattern with crisp logic formulation, thus the next stage : pattern simplification is needed.

VI. Conditional transition operation on Boolean variables

If in (1) and (2) we replace each ψ_i vector with x_i, and each $\overline{\psi}_i$ vector with $\overline{x}_i$, then the Functionals φ_+ and φ_- will receive the full disjunctive normal form [13, 14]:

$$\varphi_+ = \bigvee_{I_+(\sigma_1 \sigma_2 \ldots \sigma_m)} x_1^{\sigma_1} x_2^{\sigma_2} \ldots x_m^{\sigma_m}$$

$$\varphi_- = \bigvee_{I_-(\sigma_1 \sigma_2 \ldots \sigma_m)} x_1^{\sigma_1} x_2^{\sigma_2} \ldots x_m^{\sigma_m}$$

$$\text{where} \quad \sigma_i = \begin{cases} 1 & if \quad x_i \\ 0 & if \quad \overline{x}_i \end{cases} \quad i = 1, 2, \ldots, m ,$$

$I_+(\sigma_1 \sigma_2 \ldots \sigma_m)$ – collection set , appropriate of C^+ subclass trajectories;

$I_-(\sigma_1 \sigma_2 \ldots \sigma_m)$ – collection set, appropriate of subclass trajectories;

after the binarization of these normal disjunctive forms the pattern binary form is being received [10, 36]:

$$K_+ = f^+\left(\xi_1^{\sigma_1}, \ \xi_2^{\sigma_2} \ldots \xi_l^{\sigma_l}\right) = \vee \xi_1^{\sigma_1} \ \xi_2^{\sigma_2} \ldots \xi_l^{\sigma_l} ,$$

ISSN 1512-1232

where $l < m$ and $\xi_1^{\sigma_1}, \xi_2^{\sigma_2}, \ldots, \xi_l^{\sigma_l}$ are used to reselect those variables $x_i^{\sigma_i}, x_j^{\sigma_j}, \ldots, x_k^{\sigma_k}$ that stayed after minimization of φ_+ full disanctive normal form (analogious form is received for φ_-). Other variables $\xi_{l+1}, \ldots, \xi_m$ are less important as they have no impact on on object evaluation result.

The pattern binary form K_+ contains some important values of properties and it describes exclusiveness that is typical to C^+ subclass finite collection of objects. for a quite large(great) number of *n and m, the pattern* K_+ contains those rules (knowledge in general case) that was used by evaluator who spitted the set of objects into C^+ and C^- subclasses, thus with help of binary pattern K_+, the evaluation of elements, excluded from pattern formation , is possible: enough condition for new object to belong to C^+, is that new trajectories variables $\xi_1, \ldots, \xi_l$ should have values fixed at least in one of the implicants of pattern K_+ , variables $\xi_{l+1}, \ldots, \xi_m$ have the arbitrary values. note that binary pattern is easily presented as a productive rule.

2.2 The Hybrid Method of Analytic Heuristics (HAH Method)

Let us introduce our HAH method and first define the main concepts used in this method:

The set of properties

For the set of properties we are setting the $\mathcal{A}$ set of words (of apropreate language). Each of such $\mathcal{A}$ sets might be presented as union of eigt to ten subsets that consists of words in appropriate A_i parts of speech (lexical class: Noun, Pronoun, Adjective, Verb, Adverb, Preposition, Conjunction, subordinator, coordinator, Interjection). The number of subsets depends on certain language specification (for Georgian Language we have 10). Each $A_i, (i = \overline{1,10})$, is set of $a_{i,j}, j = \overline{1,N_i}$ where the N_i stands for the number of words in appropriate A_i part of speach.

Let us present one of x "defining text" for C „concept" in the following way: $T_C^x = \{w_{i,j}\}, i = \overline{1,10}$,$j = \overline{1,M}$, , where $w_{i,j}$ presents all different words in this text. M is the number of such words. Sure there is other text that presents the definition of the same C "concept" : C^y. All such sets are the subsets of initial $\mathcal{A}$ set. We may generelize this set to Al-set so that the usage of Analytic Heuristics method be possible.

As we mentioned above the Explicit Semantic Analysis (ESA) method is using Wikiwhorehouse [9]. Mainly the definding text T_c^{wik} for each concept C^{wik} is used. Each of these texts are presented as weighted collections: using TF-IDF scheme [16]. The semantic transformer iterates the text words, then takes the appropriate inverted index, unites it into the concept vector that defines the text.

According to [10] let us present *C concept descriptive text with help of* w_i *words set* : $T_c^{wik} = \{w_i\}$, i=1,…,M^{wik}(M^{wik} is the number of words in text defining this concept at Wikiwhorehouse), it has the appropriate vector TF-IDF(v_i^{wik}) , where each v_i^{wik} weight is appropriate of w_i *word*.

The following term weighting scheme is successfully used to define the appropriate term weight:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

ISSN 1512-1232

also the vector $\langle k_i^{wik} \rangle$ is created, where k_i^{wik} represents w_i word's inverted index of C concept appropreate text from Wikiwhorehouse. For *C* concept appropreate T_C^{wik} text the following V_C^{wik} weight vector is defined as:

$$\sum_{w_i \in T_C^{wik}} v_i^{wik} \cdot k_i^{wik}$$

Let describe the C concept for heuristics method in frames of [2] form. Suppose that each *C* is generally described with $w_1^{wik}, w_2^{wik}, \dots, w_N^{wik}$ "words". The number of words in description depends on researcher (the concept developer) point of view. Different methods of definition for this number might be used, all of them are based on full text length, number of different words and etc. The presence of each w_i^{wik} word in description is defined by V_j^{wik} weights vector that is appropriate of C concepts T_j^{wik} words. Besides, the presence of the word in description depend on part of speech it belongs to (to which A_i subset of $\mathcal{A}$ it belongs to).

Generally conjunction is always out from description. Great majority of researchers are more fond of nouns, or combinations: "adjective"+"noun" or "noun"+"verb". For definition of word combination weights vectors are not enough, we will not go for details of this case as it represents the area of our future research.

Let us generelize concept formation area and use other wiki based worehouses such as:

- AllRefer.com,
- bartleby.com,
- britannica.com,
- infoplease.com,
- encyclopedia.com,
- techweb.com/encyclopedia,
- libraryspot.com/encyclopedias.htm#science, education.yahoo.com/reference/encyclopedia).

In all such warehouses the appropriate T_C^x text of C concept is presented, thus ESA method of concept formation might be applied.

In this warehouse the concept appropriate of C concept presents as well. thus the given method might be used for it and then it maight be described with $w_1^x, w_2^x, \dots, w_L^x$ "words" (L is the number of words conteined in the worehouse). If we repeat the proccedure for each warehouse we mentioned, we will receive several (may be different) descriptions of the same C concept. for each description text T_C^x of the concept C , we will have the appropriate weight's vector V_C^x

Table 1

Warehouse	c_j concept definition
Wikipedia	$w_1^{wik}, w_2^{wik}, \dots, w_N^{wik}$
x	$w_1^x, w_2^x, \dots, w_L^x$
…	…
y	$w_1^y, w_2^y, \dots, w_K^y$

It is clear that in each "different description vectors" , that are defining the C concept, some of contained words are the same. Lets unite these words and receive warehouse words common set: $W = (w_1, w_2, \dots, w_{max})$, max – is the maximum number of different words in all warehouses, call

ISSN 1512-1232

it N. According to our notations every describing vector of C concept might be presented with help of same N size vector with elements $\widetilde{w_i}$, *i=1, …N*

$$\widetilde{w_i} = \begin{cases} w_i \ \ presents\ in\ \ C\ cioncept\ description; \\ \overline{w}_i\ does\ not\ present\ in\ C\ concept\ description. \end{cases}$$

Table 1 might be rewritten in the following unified view:

Table2

warehouse	$\mathbf{w_1}$	$\mathbf{w_2}$	...	$\mathbf{w_i}$	...	$\mathbf{w_N}$
R^1	$\widetilde{w}_{1,1}$	$\widetilde{w}_{2,1}$	…	$\widetilde{w}_{i,1}$	…	$\widetilde{w}_{N,1}$
R^2	$\widetilde{w}_{1,2}$	$\widetilde{w}_{2,2}$	…	$\widetilde{w}_{i,2}$	…	$\widetilde{w}_{N,2}$
…	…	…	…	…	…	…
R^k	$\widetilde{w}_{1,k}$	$\widetilde{w}_{2,k}$	…	$\widetilde{w}_{i,k}$	…	$\widetilde{w}_{N,k}$
…	…	…	…	…	…	…
R^m	$\widetilde{w}_{1,m}$	$\widetilde{w}_{2,m}$	…	$\widetilde{w}_{i,m}$	…	$\widetilde{w}_{N,m}$

Where

$$\widetilde{w_{i,k}} = \begin{cases} w_i\ presents\ in\ C\ concept\ description\ in\ R^k\ warehouse; \\ \overline{w}_i\ does\ not\ presents\ in\ C\ concept\ description\ in\ R^k\ warehouse. \end{cases}$$

As every description of C concept is finite the set $W = \{w_1,\ w_2, \dots, w_N\}$ is finite as well. We may present this set as an Al-set without any restrictions [2], and then define all appropriate operations on such type of sets.

According to analytic heuristics method definitions and notations, we are able to say that C concept is same as the object and each $\mathbf{w_i}$ might be used as A_i properties that are defining the object. From the all above mentioned we are now able to use this method completely. So, each realization of C concept is presented as an ordinary implicant, and to receive the pattern of concept, the minimization of normal disjunctive form is all that we need.

Example: Let have 5 different descriptions of some C concept, where 4 different words are presented:

Table 3

warehouse	*C* concept Implicant
R^1	$w_1 \& w_2 \& \overline{w}_3 \& w_4$
R^2	$w_1 \& \overline{w}_2 \& \overline{w}_3 \& w_4$
R^3	$w_1 \& w_2 \& \overline{w}_3 \& w_4$
R^4	$w_1 \& w_2 \& w_3 \& w_4$
R^5	$\overline{w}_1 \& w_2 \& w_3 \& w_4$

Now write these realizations in normal disjunctive form:

$$(w_1 \& w_2 \& \overline{w}_3 \& w_4) \vee (w_1 \& \overline{w}_2 \& \overline{w}_3 \& w_4) \vee (w_1 \& w_2 \& \overline{w}_3 \& w_4) \vee (w_1 \& \overline{w}_2 \& \overline{w}_3 \& w_4) \vee (w_1 \& w_2 \& \overline{w}_3 \& w_4)$$

ISSN 1512-1232

Let minimize the form, we will receive the generalized description of C concept based on every text in all whorehouses:

$$w_1 \& \bar{w}_3 \& w_4$$

Sure in a real case we will have much more wors in a concept description, but we are able to select all high weight words based on vector received by Explicit Semantic Analysis (ESA) method. the number of words might also be depended on ratio of text word number to different word number in the same text. more words presented in concept description will lead to more semantically adequate result, but from other hand many word in description might lead to case when concept will be useless in information retrieval.

3. THE METHOD TESTING

To avaluate the method its testing was performed. The testing stages were as follows: 1. Concept Formation; 2. Retrival according to formed concept.

As the concept in fact represents an implicant (disjunction, conjunction), the Boolean search algorithm might be used for method testing [16].

From the various 5 concepts, 70 text in total had been selected from above mentioned warehouses . For each concept we selected ten highest weighted words and on the basis of these words appropriate implicant for each descriptive text was created. from 10 to 16 different descriptive texts has been processed using above mentioned method and the appropriate concept was formed. As a result we received five different descriptions in a normal disjunctive form.

The fulfilled information retrieval based on our method in 300 different text contained warehouse, for each different concept, gave the following results:

- each concept was described from 42 to 65 texts;
- the search preciseness was between 0.81 to 0.92;

Note: For the received concepts information retrieval was made separately.

CONCLUSION

Method testing showed that the proposed method of concept description is describing its semantic meaning in more general way.

The method gives the opportunity for generalized semantical structure formation on the basis of concept's descriptive nonstructured metadata. This structure stands for one of the main components at information retrieval.

Nowadays we are working to increase the number of concept formation base texts and to define the optimal number of descriptive words. We believe this will lead us to optimization of our information retrival algorithm.

ISSN 1512-1232

REFERENCES

1. Chavchanidze, V. (1974) , "Towards the General Theory of Conceptual Systems: (A New Point of View)", Kybernetes, Vol. 3 Iss: 1, pp.17 – 25.
2. Chavhcanidze, V. (1970), "Heuristic Analysis of Artificial Intelligence in the Formation of Concepts, Pattern Recognition and Classification of Objects", p.20 Institute of Cybernetes, Georgian Academy of Sciences,dep. 2080-70, Tbilisi.
3. Khachidze M. (1998), "Artinformatic Knowledge and Some Ways of Its Presentation".Bulletin of Georgian Academy of Scienes, vol. 165, no. 6, pp. 60-65.
4. Kvinikhidze K.S., Chavchanidze V.V. (1976). "Applocation of Conceptual Approach to Describe the Evolution of Protein Structure".Reporp of the 8-th International Congress on Cybernetics, Namur, French.
5. Mikeladze M., Khachidze M. (2000), "Modified Conceptual-Probabilistic Method of Formation Rules for Medical Diagnostic Expert Systems", Proceedings of the XIV International Symposium "Large System Control", Tbilisi, 2000, pp. 162-163.
6. Khachidze M., Archuadze M., Besiashvili M., The Method of Concept Formation for Semantic Search. 7th International Conference on APPLICATION of INFORMATION and COMMUNICATION TECHNOLOGIES, 23-25 October 2013, Baku, Azerbaijan.
7. Salton G., Buckley C. (1988), "Term-weighting approaches in automatic text re trieval", Information Processing and Management, vol. 24 (5),pp. 513–523
8. Davis R. And Lenat D. (1982), Knowledge-Based Systems in Artificial Intelligence. McGraw-Hill Advanced Computer Science Series.
9. Egozi O., Markovitch S. and GabrilovichE. (2011), "Concept-Based Information Retrieval using Explicit Semantic Analysis", ACM Transactions on Information Systems, Vol. 29, No. 2, Article 8, Publication date: April 2011.
10. Gabrilovich, E. and Markovitch, S. (2007), "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In 20th International Joint Conference on Artificial Intelligence (IJCAI'07) procedings of international conference in Hyderabad, India, January 6-12, 2007, Morgan Kaufmann Publishers, pp. 1606–1611.
11. Budanitsky A. and Hirst G. (2006), "Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness", Computational Linguistics,Vol. 32 Issue 1, pp. 13-47.
12. Deerwester S., Dumais S., Furnas G., Landauer T. and Harshman R (1990), "Indexing by latent Semantic Analysis", Journal of the American Society for Information Science, Vol. 41 Num 6, pp. 391–407.
13. Strube M., Ponzetto S. P. (2006), "Wikirelate! Computing semantic relatedness using Wikipedia", Proceedings of the 21st National Conference on Artificial Intelligence, Vol. 2, AAAI Press, Boston, MA, pp.1419-1424.
14. Чавчанидзе В.В. К началам теории принятия концептуальных решений в системе искусственного интеллекта. Сообщения АН ГССР, т.70, №2, 1973.
15. Чавчанидзе В.В. К проблеме распознавания образов и об универсальной природе концептуальной интеллектуальной активности. Материалы коллоквиума по "концептуальному системному анализу естественных и искусственных систем". (Медицина, наука, техника), Батуми, 24-28 апреля, 1973.
16. Manning D., Raghavan P., Schütze H. (2008), " Boolean Retrieval", in Introduction to Information Retrieval,Cambridge University Press, pp. 1-18.

Article received: 2014-05-13