World Applied Sciences Journal 24 (Information Technologies in Modern Industry, Education & Society): 12-17, 2013
ISSN 1818-4952
© IDOSI Publications, 2013
DOI: 10.5829/idosi.wasj.2013.24.itmies.80003

An Agent-Based Approach for Resolving Non-Factors in the Problem of Semantic Comparison of Essence Notions in Applied Program Solution

Vladislav Leonodovich Berdnik and Alla Viktorovna Zaboleeva-Zotova

Volgograd State Technical University, Volgograd, Russia

Submitted: Aug 5, 2013; **Accepted:** Sep 15, 2013; **Published:** Sep 25, 2013

Abstract: Automation of semantic comparison of symbol names is an imminent task of practical application in automation accounting systems of companies. It is basically always possible to get an electronic copy of incoming document by, but subsequent incorporation into a company's accounting system may be difficult due to different ways of naming an accounted item. An analysis of semantic comparison issues was given in previous papers. Key sources of knowledge in our problem are PC operators. The number of questions to PC operator must be substantially less than questions during direct comparison of symbol denomination over a long period of system operation. Competition environment as a labyrinth (graph) out of semantic elements (rooms) and mutual links between them (passages) is offered to clarify the application of multi-agent approach in resolving non-factors to build-up intellectual user interface on the basis of intuition. Architecture of intellectual agent and intuition modeling inference engine by the method of "semantic spot" are reported here.

Key words: Semantic comparison of symbol names · Non-factors · An agent-based approach for building intellectual interface · Modeling intuition

INTRODUCTION

Comparing character strings is the process of determining equivalence signs (coreference reference identity) for later use in different systems of automation of production, trade and accounting. For example, S-multitude of symbol lines (symbol of names), D-multitude of documents d. Each document:

$$d = \left\langle e, S^d \right\rangle, \quad (1)$$

where e – denotation, corresponding to the essence of the real world (commodity, physical person, service, department in organization etc.);

S^d – submultitude of lines S, for which denotation e is known.

Therefore

$$S = S^i \cup \left(\bigcup S^d \right), \quad (2)$$

where S^i – submultitude of lines S, for which the denotation is not determined.

Authors of the article previously proposed models and methods, as well as program solution on automation of semantic comparison of symbol names in collection of documents [1].

Determination of denotation e for lines S^i is an imminent task of practical application in automation accounting systems of companies. In non-automated mode, the problem solution (input into the system of waybills, clients' requests, making price lists, updating electronic reference books etc.) requires much effort from PC operator. It is basically always possible to get an electronic copy of incoming document by e-mail or by scanning a hard copy, but subsequent incorporation into a company's accounting system may be difficult due to different ways of naming an accounted item (different lines of S^i denotation *e)*. Automation of semantic comparison of symbol names will make it possible to significantly cut down long-term company's costs, but requires a higher volume of input data at the stage of implementation. [2] Example for symbol names:

"Spring Bottled Water,-Non-aerated 19 L Discount": As is known [3], intuition is direct perception of truth without logical analysis; it is based on imagination, empathy and previous experience. In our problem, intuition is a combination of hypotheses, bases on

Corresponding Author: Berdnik, Volgograd State Technical University, Lenin avenue, 28, 400005, Volgograd, Russia.

principles of system true knowledge, as well as on inter-relations inside the current and actually existing complex of non-factors.

Resolution of another non-factor is adding new knowledge about the term, its semantics, semantic inter links. Each new term has a new semantic meaning which cannot be derived on rules of knowledge base. In view of previous experiments, conducted by the authors and above considerations, the use of any logical device is not deemed efficient. Intuition operates latent knowledge and associations. It is known [4] that the words of one sentence are associated. In our case the role of a sentence is played by a line in symbol denomination, the multitude of terms reflects the principle of item or marketing structure. In symbol denominations the terms have concealed associations between themselves, which can be extrapolated to new symbol denominations. According to the authors the best intuition modeling is the one described above by the method of "semantic spot". [5]

MATERIALS AND METHOD

Methods of semantic comparison described before [2] are based on splitting multitude D to submultitudes D^r on the basis of type-aspect relations, so that each type of substances r $\in$ has its specific set of basic inherent signs. For each type of substantial signs (for example, color) we assume a semantic field-aggregate of semantic signs, which mutually separate substance types one from another within the type [6] (for example, white, black, silvery). Correlation between line S^i and substance type is an easy task as the type sign is stated clearly in a line. This structure of semantic signs will hereinafter be named as model frame. Example:

"Spring Bottled Water, - Non-aerated 19 L Discount": "Bottled water" - sign of substance type, non-substantial sign-"discount", all other signs are substantial.

A double-stage determination of semantic equivalency makes it possible to avoid building models of each document d and to reduce the volume of knowledge basis approximately by two exponents by aggregating situational semantics.

Let us take as example non-factor in the problem on semantic comparison of symbol names. For example, there is a document describing item "Motherboard 3PE-A". At some moment of time the manufacturer of the mother board modifies the item and comes up with "Motherboard 3PE-A Green". At the level of the program system knowledge models it is not known which semantics is implied by term "Green" as it is encountered for the first time. Maybe it is addition of a new energy saving system to the item, or different color of building paint. It is possible that this item will completely supersede the previous one, or both items will be on sale. The factor of incomplete knowledge results in indiscrete relations between symbol denomination and document denotation. A more detailed analysis of semantic comparison issues was given in work [7].

The purpose of semantic comparison is to specify the chain of non-factors [8] by cutting down the missing information about semantics of terms, to resolve the whole spectrum of non-factors, for example, between "Motherboard 3PE-A" and "Motherboard 3PE-A Green" and to select a singular value by system operator.

Work [2] deals also with methods of determining semantic tolerance based on TF-IDF metrics calculation [9]. Intersection of all terms S^i and terms in all lines S^d makes it possible, using term frequency in automatic mode, to build multitude of fuzzy semantic equivalency [10].

$$A\left(s^i\right)=\left\{d,\mu(d,s^i)\right\},\quad \forall d\in D,\quad \mu\left(d,s^i\right)=[0..1] \tag{3}$$

Subsequent application of semantic comparison methods helps to reduce the power of fuzzy set (i.e. specify the relation).

The automation efficiency of semantic comparison depends upon the quality of knowledge basis, its completeness, consistency, clarity, etc. A qualitative index of automation efficiency is a value inversely proportional to power of multitude D^i.

$$D^i\left(s^i\right)=\left\{d\middle|\mu\left(d,s^i\right)>0\right\},\quad \forall s^i\in S^i \tag{4}$$

Key sources of knowledge in our problem are PC operators. Apart from asking operators, we can obtain hypotheses for non-factors from the structure of documents D collection, from texts in Internet etc. The fewer questions there are from the system to the operator, the higher is efficiency of problem automation. The number of questions to PC operator must be substantially less than questions during direct comparison of symbol denomination over a long period of system operation.

We require such an automation system, which with limited knowledge received from PC users, will fix not less than the preset number of pairs $\left\langle s^i,d\right\rangle$ in one interactive session with a user and minimize average, over a long operation period, power of multitude

$$D^i\left(s^j\right)=\left\{d\middle|\mu\left(d,s^j\right)>0\right\},\quad \forall s^j\in S^j\ , \tag{5}$$

where S^j – most demanded in future by PC operators symbol denomination. By $D^i\left(s^j\right)$ we assume active in production process (for example, in commerce,-list of sold commodities), documents $\forall d\in D^i$ within a limited time, preceding the current period (for example, 3 months).

Let Us Describe Classes of Non-factors in the Problem:
Incompleteness:

- Term semantics not established.
- Semantic field for substance type not established.
- Type sign not specified in document d.
- Model frame for the document not approved by the user.
- Missing value by default for substantial signs of model frames and available line S^d in which this sign is not established (the sign is present by default).
- No conventional coded denomination is available in the document.

Fuzziness:

- The existing mechanism of semantic comparison does not provide for unambiguous identification of a document.
- The existing mechanism of essence type identification does not provide for unambiguous identification of essence type.

Inconsistency:

- The same line is found in different documents.
- Line s is related, depending upon different rules, to different documents.
- Line s is related, depending upon different rules, to different essence types.
- Lines of the same essence type have semantic signs with different values of one semantic field.
- Line S^d in document d, which has a semantic sign, contrary to the semantic sign by default.
- Analysis of line term pre-supposition shows semantic signs, which are clearly contrary to the semantic signs, assumed in the line.

Vagueness:

- Line S^i relates definitely based on all available knowledge to some document d, but contains new semantic signs (see the example above).
- Line S^d may possibly contain author's errors (line written with errors).

Normally when two persons have a discussion each question contains an intuitive premise. The interlocutors could ask more questions replenishing each detail of their ignorance. But each party has its own vision of the situation and this vision is a hypothesis, answering a multitude of primitive questions. In the beginning of discussion the interlocutors casually verify their opinions, updating this way the hypothesis. Then the interlocutors with their questions and answers mutually supplement their perception of the discussion subject. This way of gaining information is more efficient than a list of simple questions; it significantly minimizes addresses to an interlocutor and is more comfortable. A verified view point of interlocutors makes it possible to extrapolate answers to new questions. [11] This concept makes a basis of the intellectual interface under development. The program system must always find some general principle, which will speed up determination of semantic equivalency for a big number of pairs $\left\langle s^i,d\right\rangle$.

As is typical of any hypothesis, its trustworthiness is not finally known. It is good to have during discussion several different hypotheses, out of which we choose a most truthful. Besides, we deal with a multitude of compared pairs $\left\langle s^i,d\right\rangle$. In this presentation of the problem we can see competition between multitude of hypotheses for each of compared pairs $\left\langle s^i,d\right\rangle$. The limiting resource of the problem is a right to put a question to PC user. Competition environment is a labyrinth (graph) out of semantic elements (rooms) and mutual links between them (passages). Non-factors in this space are concealed passages and rooms, as well as false links (passages).

This non-traditional presentation of the problem is offered to clarify the application of multi-agent approach in resolving non-factors to build-up intellectual user interface on the basis of intuition.

Agents, as is known, are program actors in the problem sphere, which have mutual obligations, established in the process of dialogue, they conduct negotiations and coordinate transition of information [12].

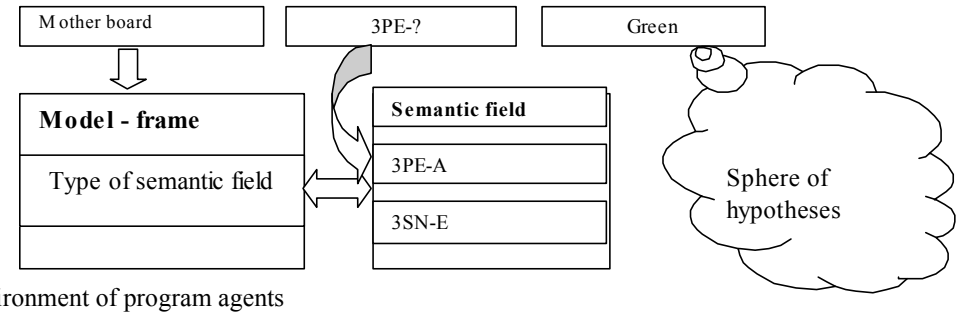

Fig. 1: Environment of program agents

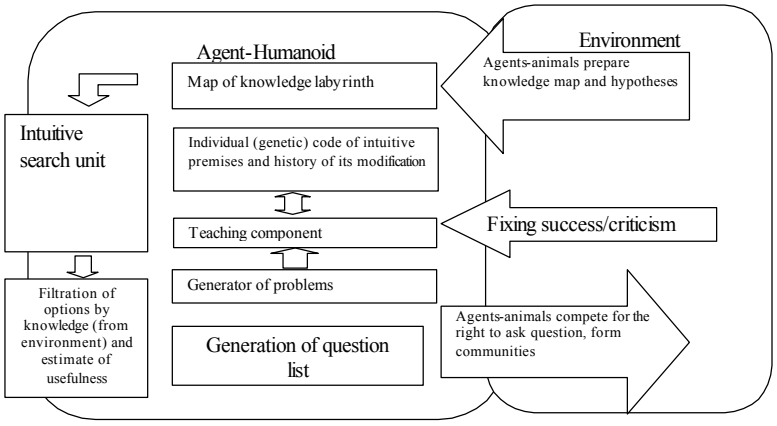

Fig. 2: Architecture of intellectual agent.

In the multi-agent environment under development we will create for each d^j a virtual agent-humanoid, which, depending on the goal, can create reactive agents-animals for solution of a local problem. [13] The obtained information supplements the knowledge basis and the agent informs other agents. If agents fulfill their purpose, they get an extra right to error. Agents are considered friendly, if non-factor resolution of one agent can with a certain degree of probability, resolve non-factor of another agent, even if they are created by different agents-humanoids. The power of such group rises in the agent, which is critical for resolution of the whole chain of non-factors.

A separate group of agents are agents-critics. These agents-animals are created if there are contradictions in semantic rules of the knowledge base.

Main Part:

Architecture of Agents: Agent-humanoid is created for each $d \in D^j$. During creation it is given relevant selection

$$S^r(d) = \{s | \mu(d,s) > L\}, \quad \forall s \in S^i, \tag{6}$$

where L-threshold value, assumed by the user in the policy of intellectual system, μ-is calculated by method TF-IDF.

Taking into account that $d = \langle e, S^d \rangle$, he agent has multitude S^d – lines, for which semantic equivalency is assumed $\mu(d^j, s^d) = 1, \quad \forall s^d \in S^d$. As the line is a multitude of terms, it seems likely $\mu(d, s^d) = 1$, if

$$\bigcup \left(s^r / s^d \right) = \varnothing, \quad \forall s^d \in S^d, \tag{7}$$

where S^d - lines of document d, S^r - multitude of terms on compared line, proceeding from extinct syntactical structure of symbol denomination.

Agents Can Have the Following Individual Targets [14]:

- Competition for a right to put a question to a user.
- Submittal of hypotheses.
- Denial of rivals' hypotheses.
- Joint participation in proving an advantageous hypothesis

The choice of target for agent humanoid is done on the graph by the method of semantic spot [5]. The graph if formed first by agents-animals. Graph tops are terms, semantic fields, documents, types of essence.

Graph curves denote the weight, calculated as the quantity of associations between the tops. Proportion term-term is based on the frequency of term occurrence in one symbol denomination. Proportion term-document is determined as a number of lines in a document, containing the term, etc.

Then the graph weights are rated so that the total weight of all outcoming tops would be equal to 1. Therefore, the weight of graph rib is in interval [0..1]. In analysis S^r - many terms of compared line are at the graph top. For the top found, excitation is formed = 1. Then the excitation spreads by waves along the graph. When crossing a rib, the excitation value is multiplied by rib weight. Excitations reaching one top by different routes are summed. The tops which received highest secondary excitation are assumed by agent as hypothesis. For example, if the frame model is not established, we assume for hypothesis the most excited top of essence type, for terms with unknown semantics we make hypotheses about their semantic fields, based on relevant excited tops, related to essence type, etc.

A hypothesis, obtained this way, is verified with announcement board. If some other hypothesis comes up at another agent, they exchange addresses. If a similar hypothesis is not found on the announcement board, it is placed on the board.

For most excited tops-documents, agents are searched for. If such agents are found, exchange of addresses is done with them.

Agents have scored most addresses as compared to associated agents, are declared as leaders. Agents relating to more than one leader, decide which leader they will cooperate with from now on. A leader which does not have no-leaders, forfeits its announcement and joins a leader.

Each group chooses a question to PC user. If the hypothesis has been proven right, each element of the group scores an extra point. The obtained knowledge is entered in the knowledge base. If the hypothesis is wrong a point is removed from all agents of the group and, unless the agent has points in reserve, it dies. Surviving agents memorize hypotheses of dead agents as wrong ones.

CONCLUSIONS

Normally, agent's behavior is based on knowledge base and an inference engine. [14] Intelligent agents might observable environments to achieve their goals. Possibility to build an agent behavior by intuitive are reported here.

A program was written based on the method described. Now statistical data are being accumulated on efficiency of this solution.

ACKNOWLEDGMENTS

The work was prepared with support of grant No. 13-07-00461 from the Russian Fund of Fundamental Research.

REFERENCES

1. Berdnik, V.L. and A.V. Zaboleeva-Zotova, 2007. Problem of identifying the essence by low-structured text. Journal “News of Volgograd State Technical University”, series "Pending problems of control, computing and informatics in technical systems", 2(2): 26-28.
2. Berdnik, V.L., A.V. Zaboleeva-Zotova and Yu.A. Orlova, 2012. Semantic analysis of symbol denominations in collection of documents. Volgograd State Technical University Press, pp: 124.
3. Intuition (psychology). Date Views 12.07.2013 http://en.wikipedia.org/wiki/Intuition_%28psychology%29.
4. Chanyshev, O.G., 2011. Associative fields of dominants and text analysis. In the Proceedings of the 2011 All-Russian conference with foreign participation “Knowledge-Ontology-Theory” (ZONT-11), Novosibirsk, pp: 126-135.
5. Berdnik, V.L. and A.V. Zaboleeva-Zotova, 2012. Model “Semantic spot” in complicated formalized problems of information intellectual processing. Journal “News of South Federal Territory. Technical science”, 1: 116-121.
6. Kobozeva, I.M., 2007. Linguistic semantics. KomKniga Press, Moscow, pp: 352.
7. Berdnik, V.L. and A.V. Zaboleeva-Zotova, 2007. Semantic analysis of predicative substance names for essence identification. Journal “News of Volgograd State Technical University”, series "Pending problems of control, computing and informatics in technical systems”, 9(3): 43-46.
8. Narinyani, A.S., 1994. Non-factors and knowledge engineering: from petty formalization to natural pragmatism. In the Proceedings of the 1994 National Conference on Artificial Intellect, Rjibinsk, pp: 9-18.

9. Wu, H.C., R.WP. Luk, K.F. Wong and K.L. Kwok, 2008. Interpreting tf–idf term weights as making relevance decisions. Journal "ACM Transactions on Information Systems", 26(3): 1-37.
10. Tarassov, V.B., 2002. General approaches to the modelling of soft estimates and beliefs in strategic decision engineering. In the Proceedings of the 2002 IEEE International Conference of Artificial Intelligence Systems (ICAIS 2002), pp: 45-49.
11. Eysenck, M.W., 2009. Fundamentals of psychology. Psychology Press, New York, pp: 580.
12. Foundation for intelligent physical agents, FIPA 99 Specification, FIPA Developers Guide Preliminary. Date Views 01.01.2013 http://www.fipa.org/specs/ fipa00021/PC00021.pdf.
13. Braspenning, P., 1997. Animal-like and Humanoid Agents and Corresponding Multi-Agent Systems. In the Proceedings of the 1997 International Workshop "Distributed Artificial Intelligence and Multi-Agent Systems", (DAIMAS'97), St. Petersburg, Russia, pp: 64-77.
14. Russel, S. and P. Norwig, 2010. Artificial Intelligence: A Modern Approach (3d edition). Prentice Hall Press, pp: 1152.