UDC 004.832.2 : 004.853

PERTINENCE EVALUATION SYSTEM ARCHITECTURE ON A BASIS OF LEARNING ONTOLOGY WITH PLANNING IN A CERTAIN DOMAIN

D. H. Dosyn

Lviv Polytechnic National University

E-mail: dmytro.h.dosyn@lpnu.ua

The method of text document pertinence estimation is proposed. It is based on agent approach, expected value of perfect information analysis, hierarchical task network structure of a knowledge base and automated planning algorithms. Use of Markov decision process approach allows us to estimate expected utility of the strategy built in the framework of agent knowledge base with the aim to evaluate gain of expected utility caused by account of information extracted from the text document. For this purposes the text document is considered as a message with a two-part structure which should help us to supplement information contained in this document by relevant context information from the knowledge base.

Keywords: *pertinence evaluation, ontology learning, automated planning, hierarchical task network, expected value of perfect information.*

АРХІТЕКТУРА СИСТЕМИ ОЦІНЮВАННЯ ПЕРТИНЕНТНОСТІ, ЩО БАЗУЄТЬСЯ НА НАВЧАННІ ОНТОЛОГІЇ ПЛАНУВАННЯ У ВИБРАНІЙ ПРЕДМЕТНІЙ ОБЛАСТІ

Д. Г. Досин

Національний університет “Львівська політехніка”

Запропоновано метод оцінювання пертинентності текстових документів, який базується на агентному підході, аналізі очікуваної величини досконалої інформації, структурі бази знань у форматі ієрархічної мережі задач та алгоритмів автоматичного планування. Марківська модель прийняття рішень дає змогу обчислювати очікувану корисність стратегії, побудованої засобами бази знань агента, для оцінки приросту корисності, зумовленого врахуванням інформації, отриманої з текстового документа. Для цього текстовий документ розглянуто як повідомлення з двохелементною структурою, що допомагає доповнювати інформацію, яка міститься в цьому документі, відповідною контекстною інформацією з бази знань.

Ключові слова: *оцінювання пертинентності, навчання онтології, автоматичне планування, ієрархічна мережа задач, очікувана величина досконалої інформації.*

Introduction. An effective information search in the early 21st century has become a decisive factor in the success of any business that requires a competent resolution. The creation and rapid development of the Internet has put on the agenda a high pressing problem of finding information in the virtually infinite amount of data available through the global computer network. Thus a concept of relevant information as information connected with a search query, pertinent information as information that meets the information needs of the user of the information search system, as well as the concept of a query as a compactly formulated topic of search appeared. According to definition given by Claude Shannon, information is a value reversed to the degree of uncertainty (entropy) of the system. Now the concept of information received its extended interpretation, close to the meaning of the concept of knowledge according to which information as knowledge is considered within the framework of the “agent–environment” model and is characterized by the subjective utilitarian value. This means that although any system with uncertainty objectively contains information, only a part of it can be used to make a decision by some agent and only that pertinent part has significant value for this agent.

Modern information systems, which are intended for searching the information, are mainly based on using keywords and, in best cases, on the history of system-user interactions, i.e. on feedback based on search results. Keyword-based information interest representation is too primitive to be exact enough. It needs multiple cycling approach of search – select – keyword update – search. Therefore it is useless without personal participation of a search service user. An automation of this process makes the problem of creation of more comprehensive model of information needs of a user of a search service very important and actual.

Related work. Use of ontology for enhancing information search was widely studied in the last years. A comprehensive review of the recent works in a field of ontology learning was published in [1]. All these works are important contribution to the general task of pertinence evaluation. For example, in paper [2] metadata and ontology application for raising the quality of pertinent information resources retrieval is considered. But only qualitative modelling is proposed without mentioning any planning techniques that made numerical efficiency estimation there of using information almost impossible. Other work [3] includes both ontology building and planning process for rescue operations in possible metro emergency states. But also in this case no one numerical estimation of the efficiency of a plan application is proposed.

In paper [4] a general framework of unconventional emergency handling based on ontology is presented. The relevant information is extracted from the historical cases and domain knowledge. The key advantage of the proposed approach consist in using HTN (Hierarchical Task Network) structure of a domain knowledge which is unavoidable condition to estimate the task solving gain and connected expenses.

Other works are related to effectiveness of ontology usage for some specific domain problems, for example, by using ontologies as an intermediate layer while forming queries to databases [5–7]. In such systems, queries are formulated using concepts and relations from ontology [6]. After that they are rewritten as SQL queries. The effectiveness of the result is evaluated using well known measures such as query complexity. The query optimization methods reduce complexity by using dependences between concepts in ontology [7], such as semantic index [5].

More relevant work, although not the newest one in the field of analytic knowledge based systems design was published by Eric Bouillet et al [8]. In the paper a domain-independent, general purpose knowledge engineering and planning framework is described. Such framework, as expected, should support the construction of planning domains and problems on a basis of OWL ontologies. A planning process could be integrated with reasoning according to formalism of description logic. The authors have developed a planner that can construct plans based on domains expressed in OWL, giving a goal description. The planner is based on the Stream Processing Planning Language (SPPL) model, derived from PDDL. It can generate optimal plans according to an additive quality metric, where each action is associated with a quality vector and a cost of execution. A planner integrates DL reasoning with planning by using a two-phase planning approach where it performs DL reasoning in an offline manner and builds plans online without doing any reasoning. The planner uses a subset of DL called DLP (Description Logic Programs, which has polynomial time complexity and can be evaluated using a set of logic rules.

Another more advanced approach to the problem of creation of a framework for pertinence measurement consists in encapsulation of a planner in an ontology structure and estimation of the so called expected utility of agent plan (strategy) after each insertion of new data to the ontology. In [9] it was considered the agent planning task in terms of OWL and the possibility of mutual conversion HTN and OWL. On that basis it is possible to build domain ontology structure as an optimal plan (optimal strategy) of an intellectual agent – owner of the ontology.

Methods of formalization of actions and measures that person can or should perform are investigated in the dissertation of Michael Carl Tschantz, "Formalizing and Enforcing Purpose Restrictions", published in 2012 at Carnegie Mellon University [10]. Such formalization is achieved by applying the methodology of automatic planning, or more precisely, approach to the problem of planning based on POMDP model.

Significant contribution to the study of the information value was made by R. L. Stratonovich [11, 12]. Speaking of the Ukrainian researchers, approaches to numerical analysis of new knowledge were considered in the papers of Tatiana Bochuli [13] and a number of other scholars. Russian researchers also studied the problem of pertinence estimation [14, 15].

Formulation of the problem. How to measure the expected utility for an intelligent agent? This can be done within the framework of the theory of automatic planning – a section of the general theory of artificial intelligence, designed to build a model of rational behavior of the agent in various, including unfavorable conditions – insufficient information, uncertainty or absence of the end goal, limited resources, etc. – in their various combinations. Because of a variety of the intellectual agents nature there is no universal model of their rational behavior but some of the model approximations may be useful for practical application. These may include, for example, so called "Partly Observable Markov Decision Process (POMDP). In this model approach it is proposed in a running state not to consider the background of the decision-maker, i.e. the success or failure of all previous decisions and states in which the agent was. Only possible future conditions, possible solutions in these states and likely results of such decisions are taken into consideration. In such case the information about the states, both running and future, possible actions in that states and predictive estimates of their results should be stored and accumulated in some form of knowledge of the agent in a certain form (in a certain format). The choice of the form (knowledge base architecture) and the format for representing knowledge about states and actions (decisions) is a particular actual serious problem. Let us note that without having a knowledge base that includes so elaborated optimal strategy of behavior the agent cannot evaluate the usefulness of new knowledge – the pertinence of a new portion of information. Pertinence in contrast to relevance characterizes the usefulness of the information provided to the client by the information search service. No any pertinent information is relevant, and no all relevant information is pertinent to the client in view of the ultimate goal of his information retrieval.

The purpose of this paper is to formulate the ways of constructing tools for evaluating the pertinence of information in a given subject area. Despite of many proposed means and approaches on the market the task remains unsolved yet [16]. We have two main causes for this: first of all, usefulness criteria are substituted by similarity (relevance) and secondly – adequate model of user information needs (also methods and means for it fast creation) is still absent. Thus the mentioned main task of pertinence evaluation is divided at least onto two subtasks:

1) elaboration of numerical criteria of usefulness (utility) of text document;

2) development of methods and means for modeling of user information needs.

Expected utility estimation. Depending on the field of application there exist different ways to estimate an optimal strategy. For example, an agent in the field of materials sciences (corrosion protection domain) should be able to estimate the actions and conditions for selecting the necessary actions. It is easier to do with states in which structures of metal are already presented. It is more difficult to estimate future states. The meta-knowledge which remains in the ontology of materials science is used for estimation. Therefore, at first, we will consider the estimation of past states, then actions, and finally, their combinations which lead to a new state.

Let $v(St(i))$ is evaluation of state $St(i)$. State of aim “Goal” is defined by necessity of some set of attributes X_W to reach some values of $z(x, Goal), \forall x \in X_W$. Any state $St(i)$ is given by its set of attributes Y_i, which are taking the values $z(y, St(i))$, $\forall y \in Y_i$.

To estimate the state $St(i)$ set of attributes and their Ψ values of state $St(i)$ need to be reflected into set of attributes and values of state *Goal*. Obviously this reflection uses a base of knowledge of software, namely, additional modulus of ontology O Semantic Web Rule Language (SWRL): $\Psi : Y_i \xrightarrow{O} X$. Then the estimation of state $v(St(i))$ is calculated

$$v\left(St\left(i\right)\right) = d\left(St\left(i\right), Goal\right) = \sum_{x \in X_W} \varphi\left(z\left(\psi\left(y\right), St\left(i\right)\right), z\left(x, Goal\right)\right),$$

where φ is metric, the choice of which is depending on the type of attributes of the problem (quantitative, qualitative, mixed).

Effectiveness of action a_{ij}^k is the meaning of function which depends on an assessment of the transition to a new state $v(St(j))$, the consumption of resources g_{ij}^k and the gain $f(Z_j)$: $U\left(a_{ij}^k\right) = \delta\left(St(j), g_{ij}^k, f(Z_j)\right)$, where Z_j is the meaning of a token in the state $St(j)$. Then the task could be reduced to the task of dynamic programming:

$$\begin{cases} U = \sum\limits_{i=0}^{N-1} U\left(a_{ij}^k\right) \to \max, \\ r \geq r_e, \\ \sum\limits_{i=0}^{N-1} g_{ij}^k \leq G. \end{cases}$$

Estimated change of maximum expected utility (MEU) could be used as a measure of a new data inserted in a knowledge base.

Agent approach. An intellectual agent is an open information system by the definition, therefore for analysis of its own development progress it compares internal (i.e. exclusively controlled) and external resources. Agent interprets information about possibility of transformation of external resources into internal as knowledge. Most effective transformation strategy is selected by consequent MEU estimation for each possible action execution (decision making).

Let expected utility function for action A be:

$$EU(A \mid E) = \sum P(Result_k(A) \mid Do(A), E) \cdot U(Result_k(A))$$

where $Result_k(A)$ – one of k possible results of an action A, $U(Result_k(A)$ – usefulness of such result (state).

If a new knowledge appears, an agent re-estimates the utility function (UF) for possible actions and compares it to previous values. Its gain (in dimension of resources) is proposed to be considered as a resulting measure of knowledge pertinence.

Our approach to knowledge representation in the form of semantic network (conceptual graph) starts with the assumption that any possible generalization appears in a knowledge base as a separate concept. Therefore if some generalization has common characteristics or methods they can be realized by means of properties and event handlers of corresponding summarizing concept. Such kind of generalization corresponds to a container of organized algorithms with a resulting expected utility, evaluated in terms of resources specific for the domain. Each included sub-algorithm recursively evaluates its own part of common expected utility (Fig. 1).

Rational agent by the definition follows to its own knowledge base algorithms, set of which could be interpret as its optimal strategy according to MEU with aim to achieve its purpose of functioning.

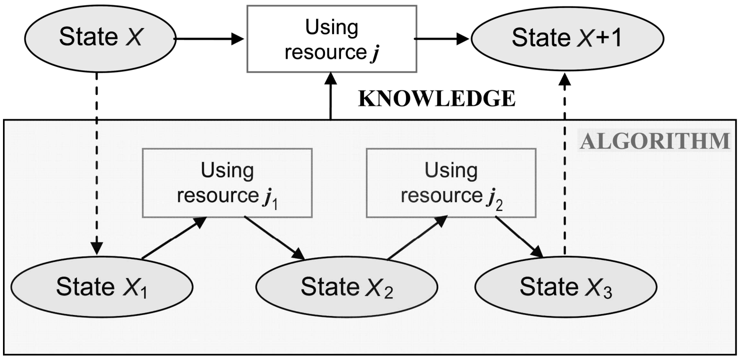

Fig. 1. Knowledge as a combination of algorithms (strategy).

If some text document describes new algorithm and agent can recognize it their consideration could produce nonzero UF which will be estimated as a difference between gain of the strategy which is known for an agent and gain of the new strategy.

Consideration of a text document as a message. According to declared agent approach an agent receives natural language text (NLT) document in the form of a message. The message is formed by the agent as well. The structure of the message is focused on perception by other agent, therefore contains at least two parts (Fig. 2): ascertaining part, by which the addressee estimate relevancy of the message (1) and defines its context (2), and constructive part – potentially new knowledge in a recognized context (3).

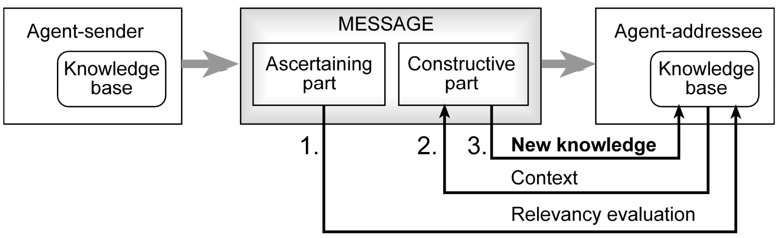

Fig. 2. New knowledge evaluation in a message with a purpose
of selective knowledge base completion.

Each algorithm or strategy existing in the knowledge base is represented there as a separate concept with its own value of expected utility in the form of object property. Agent activity consists in decision making to realize step-by-step, which provide MEU behavior. If a message received is recognized as a new knowledge, new concept in knowledge base is created and expected utility is taken from that message or context the same time. Not every algorithm in NLT message possesses an explicit value of UF. Therefore an important part of the method of knowledge pertinence evaluation consist in inference of message context and re-estimation of the appropriate expected utility of the message.

Recognized new knowledge consists of two parts: common with prototype knowledge part and difference part which UF must be evaluated or described explicitly (Fig. 3):

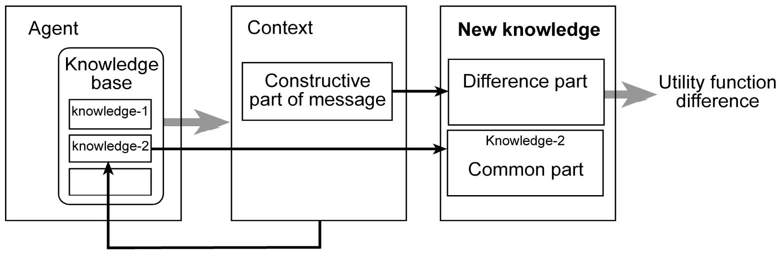

Fig. 3. Utility function evaluation for a new knowledge as a difference between utility function of the prototype knowledge and utility function estimated from the context of proposed in a message new algorithm.

There is an algorithm of text message analysis aimed at extracting and evaluating new knowledge pertinence:

1. Each sentence of the text must be divided into semantic pairs of recognized words.
2. For significant kinds of semantic links pairs of words are consequently connected into a conceptual graph of sentence and so is done for all text under analysis.
3. According to recognized words all vertices of the built conceptual graph obtain its weight from intellectual agent ontology.
4. Obtained weighted conceptual graph is completed by closely related concepts from ontology to form relevant context of the message, which is represented by such graph.
5. Updated graph is reduced to special pattern subgraph small enough to estimate the expected utility of new algorithm. Such special pattern sounds like: “Using A we obtain the possibility B”, where A – is a new algorithm of decision making and B – is an explicitly expressed value of needed resources.

Therefore the analysis of each message could be explicitly expressed by such consequence of procedures. When agent receives a message he first of all estimates its importance using this algorithm and then decides to update own knowledge base by information from the message.

Knowledge pertinence evaluation. All needed procedures for evaluating the knowledge are contained in the analyzed message. It depends on the message context which is explicitly expressed in the agent ontology and the optimal strategy for reaching its goal. The evaluation method is based on the Expected Value of Perfect Information

$$\mathrm{EVPI} = \mathrm{EV|PI} - \mathrm{EMV},$$

where EMV is the probability weighted sum of possible payoffs per each alternative; $\mathrm{EMV} = \max_i \sum_j p_j R_{ij}$, $\sum_j p_j R_{ij}$ – is the expected payoff for action i; EV|PI is the expected or average return if we priori have the “perfect” (i.e. new) information for the best i choice:

$$\mathrm{EV} \mid \mathrm{PI} = \sum_j p_j (\max_i R_{ij}) .$$

To estimate new EVPI knowledge we must have the EMV value for each solving approach for each task from HTN of our ontology. To obtain them all we have to create and solve the appropriate POMDP task:

$$\mathrm{EMV}_i \equiv U(S_i) = R(S_i) + \gamma \cdot \max_{A_{ik}} \sum_k P(S_i, A_{ik}, S_j) U(S_j) .$$

The last equation describes the reward for taking the action that gives the highest expected return. The additional information decreases the model uncertainty and the expected common reward (utility) is not less than without such information.

If we use POMDP algorithms, such as value and policy iteration for particular domain model, then we may evaluate the expected utility for both cases: with new information and without it.

CONCLUSIONS

In this paper the methodological principles of estimation of information pertinence using expected value of perfect information are described, the basic concepts of involving the automated planning methods are formulated, the features and advantages of partly observable Markov decision process model are outlined.

The main difficulty connected with this approach realization consists in the absence of effective tools and even technics of ontology learning from the text. Moreover, such technics depend on the possibility to distinguish a useful part of information inside the text document to add it to the knowledge base and its ontology. To solve this problem the smart data integration approach is proposed. This approach is based on the selective goal driven ontology learning. The automated planning paradigm in a combination with a value of the perfect information is used for evaluating the knowledge correspondence with the learning goal for the data integration domain.

1. *A survey* of ontology learning techniques and applications / Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood and Hafiza Mahnoor Abbasi // Database. – 2018. – **Vol. 2018**. – article ID bay101; doi:10.1093/database/bay101
2. *Approach* for Forming the Bionic Ontology / A. Sigov, V. Baranyuk, V. Nechaev, O. Smirnova, A. Melikhov // Procedia Computer Science. – 2017. – **Vol. 103**. – P. 495–498.
3. *Hanbin Luo, Xiaofan Peng, Botao Zhong.* Application of Ontology in Emergency Plan Management of Metro Operation // Procedia Engineering. – 2016. – **Vol. 164**. – P. 158–165.
4. *Unconventional* Emergencies Management Based on Domain Knowledge / Mengyan Feng, Aihua Li, Chuanliang Jia, Zhidong Liu // Procedia Computer Science. – 2016. – **Vol. 91**. – P. 268–275.
5. *Calvaneze D.* Optimizing ontology-based data access. – Free University of Bozen-Bolzano: KRDB Research Centre for Knowledge and Data, 2013 [Електронний ресурс]. – Режим доступу: https://goo.gl/F6Nyro
6. *Gottlob G., Orsi G., Pieris A.* Ontological queries: Rewriting and optimization Data Engineering. – 2011 [Електронний ресурс]. – Режим доступу: https://goo.gl/QHHcFa
7. *Li Y., Heflin J.* Query optimization for ontology-based information integration // 19th Int. Conf. on Information and Knowledge Management (CIKM 10). – 2010. – P. 1369–1372.
8. *A Knowledge* Engineering and Planning Framework based on OWL Ontologies / Eric Bouillet, Mark Feblowitz, Zhen Liu, Anand Ranganathan, Anton Riabov. – 2014 [Електронний ресурс]. – Режим доступу: https://goo.gl/Z5uCTF
9. *Freitas A., Schmidt D., Meneguzzi F., Vieira R., Bordini R. H.* Using Ontologies as Semantic Representations of Hierarchical Task Network Planning Domains. – 2014 [Електронний ресурс]. – Режим доступу: http://www.meneguzzi.eu/felipe/pubs/emas-semantic-representations-2014.pdf
10. *Tschantz Michael Carl.* Formalizing and Enforcing Purpose Restrictions: Dissertation. – Pittsburg: School of Computer Science Carnegie Mellon University, 2012. – 134 p.
11. *Стратонович Р. Л.* Про цінність інформації // Изв. АН СРСР: Технічна кібернетика. – 1965. – № 5. – С. 25–38.
12. *Стратонович Р. Л.* Теория информации. – М.: Сов. радио, 1975. – 424 с.
13. *Бочуля Т.* Вартість інформації в обліковому вимірі: реалії теорії та практики // Бухгалтерський облік і аудит. – 2013. – № 10. – С. 28–32.
14. *Савотченко С. Е.* Показатели оценки качества пертинентности результатов автоматизированного поиска в информационных системах // Науч. ведомости Белгород. гос. ун-та. Сер.: Экономика. Информатика. – 2016. – **38**, № 9 (230). – P. 135–138.
15. *Галимов А. А*. Разработка онтологической модели публикаций // Кибернетика и программирование. – 2015. – № 2. – С. 98–106.
16. *Nappi M., Ricciardi S., Tistarelli M.* Context awareness in biometric systems and methods: State of the art and future scenarios // Image and Vision Computing. – 2018. – **76**. – P. 27–37.

Received 24.10.2018