UDC 004.82

Данило Сергеєв

ІНТЕГРАЦІЙНА МОДЕЛЬ ПРЕДСТАВЛЕННЯ ЗНАНЬ ДЛЯ СЕМАНТИЧНОГО WEB

INTEGRATION MODEL FOR KNOWLEDGE REPRESENTATION FOR SEMANTIC WEB

Ця стаття представляє інтеграційну модель представлення знань у природно-мовних базах знань, яка дозволяє досягти гарної обчислювальної складності операцій додавання та пошуку даних у базі знань. Така модель може бути використана як проміжна структура даних між спеціалізованими системами обробки тексту та знання-орієнтованими системами в екосистемі Semantic Web.

Ключові слова: квант знань, база знань, обробка природної мови, інтеграційний підхід.

Рис.: 2.Табл.: -. Бібл.: 11.

This paper presents an integration model for knowledge representation in natural language knowledge bases that allows to achieve good computational complexity of operations of adding and searching data in the knowledge base. The model can be used as an intermediate data structure between specialized text-based and knowledge-based systems in Semantic Web ecosystem.

Key words: quantum of knowledge, knowledge base, natural language processing, integration approach.

Fig.: 2. Tabl.: -. Bibl.: 11.

Relevance of research topic. Creating of Semantic Web and instruments based on it has been a well-established area of research for at least last two decades. It remains an active topic, worked on as part of the “Web 3.0” by the W3C committee [1] as well as by independent researchers around the world. Despite this fact, there are still unresolved problems with fully implementing it [2], many of which are linked closely to the data model used in the underlying knowledge base.

Formulation of the problem. Recent research in the field shows promising success in solving some of the underlying problems, namely mining existing Semantic Web for domain-specific tasks [3], fine-tuning user tools in order to retrieve more precise results [4] and creating knowledge bases both for raw natural language data [5] and structured knowledge [6]. However, there still remains the problem of coordinating different parts of the system to achieve cohesive structure of knowledge in it that would allow to coordinate processing of natural language data and structured knowledge. This paper proposes a new model for knowledge representation for natural language knowledge base to be used as a part of Semantic Web ecosystem.

Analysis of recent research and publications. The Semantic Web technologies rely on various data models to store semantically interlinked knowledge, which together can be generalized as knowledge bases, most prominently ontologies, frame networks and semantic networks [7]. On the other hand, data structures for storing text information are mostly based on language-specific syntactical structure, such as parse trees, grammars and n-grams [8]. Both groups of models work with

fundamentally similar types of data, from words to texts in natural language models and from concepts to rules systems in knowledge bases, but they typically are not well-suited to handle the other type.

Both knowledge bases and natural language data structures are used for various high-level tasks, but their functions can be typically brought down to the basic I/O operations of "read-write". Taking into account specifics of aforementioned systems, these operations represent two major functions of a knowledge base:

- reading, or searching and reading, knowledge according to a specific query;
- writing, or adding new knowledge into an existing knowledge base.

Both these functions require additional operations to be executed inside the system, and both natural language and structured knowledge bases introduce additional complexity to these operations.

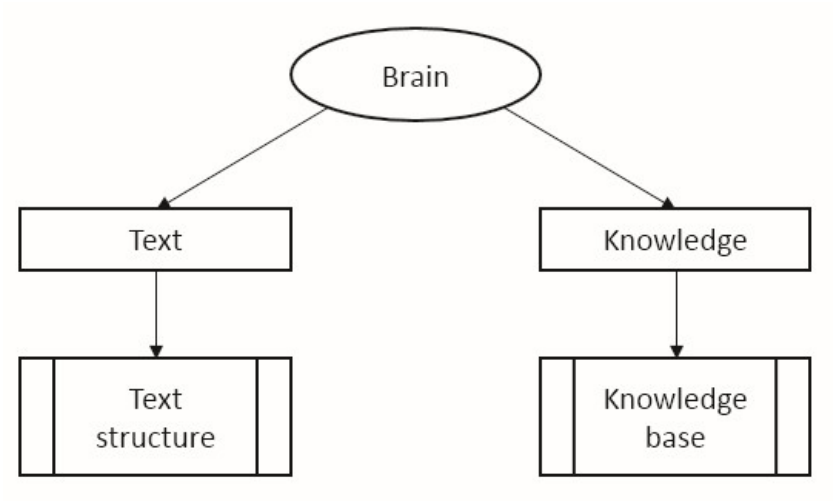

Fig. 1. Data flow in natural language and structured knowledge bases.

Selection of unexplored parts of the general problem. Natural language knowledge bases rely heavily on parsing syntactic structure of the incoming text. This allows for easier addition of new information, as the knowledge structure of incoming text is by definition same as that of the knowledge base, but searching data over such base requires traversing all text-based data structures, which makes applying advanced searching algorithms very hard or outright impossible.

Structured knowledge bases, on the other hand, provide rigid data structure that allows for effective searching, but require any incoming data to be adapted and possibly trimmed in accordance with this structure, making expanding the knowledge base an algorithmically and computationally complex task.

This allows existing knowledge models to be efficient in certain tasks that demand good performance from only certain part of their functions, but effectively prevents any of the models to be used as an average ubiquitous solution.

Presentation of the main material. In this work, a new model of combined knowledge representation is presented that allows to combine best parts of existing models while avoiding their downsides. The model is based on the ideas of integrated approach to modelling of human speech activity, described in [9–11], most notably the proposed basic semantico-syntactic structure, or BSSS, that represents a single "situation", or quantum of knowledge in human brain.

Given that a single BSSS represents a single unit of knowledge, closely tied to corresponding sensory memories, and any fragment of text can be decomposed into several BSSS and relations between them, it can be used as an intermediate structure between knowledge bases of natural language and of structured knowledge.

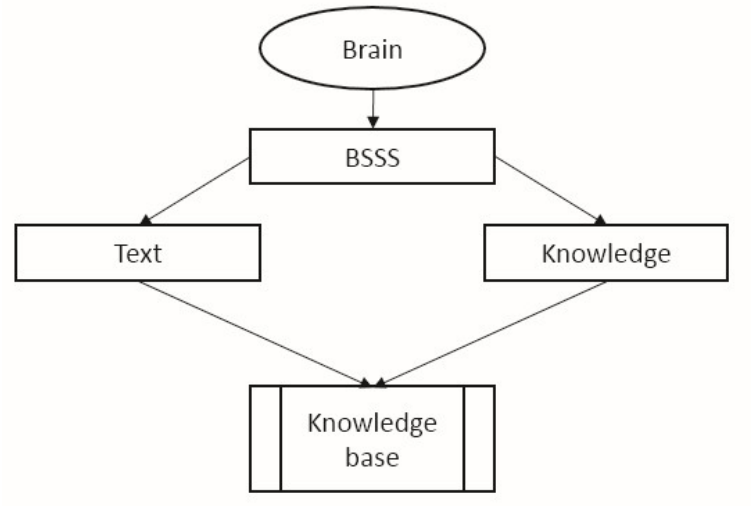

Fig. 2. Data flow in integration model knowledge base.

This model allows to consolidate the knowledge itself in a single, universal structure, filled with BSSS's and relations between them, that provides certain benefits compared to other approaches.

First, like in structured knowledge bases, data is stored in structured, formally well-defined form, which allows for easier processing and automatic manipulation of this data.

Second, like natural language knowledge bases, a model like this provides structural links between the contents of the knowledge base and structure of the text linked to these contents, which preserves easiness of adding new knowledge to the knowledge base from natural language text.

Third, unlike the other approaches, in integrated knowledge base new knowledge is added to an existing structure without altering it – that is, complexity of expanding the base remains relatively constant regardless of its size and doesn't require mandatory rebuilding of previously added data.

It is also worth noting that an integrated knowledge base can be linked to separate natural language and structured knowledge bases, thus allowing to use language-specific language processing tools and domain-specific logic and resolution engine to enhance capabilities of the whole system.

Conclusions. Future work will involve improving building links between natural language constructs, especially multi-word and artificial ones like abbreviations, and complex semantic concepts. More research is also required to create a linguistic processor capable of working with the data model presented in this paper.

References

1. W3.org Semantic web - w3c / W3.org. — 2012.
2. Benjamins V. R. Six challenges for the semantic web / V. R. Benjamins, J. Contreras, O. Corcho, A. Gómez-Pérez // Challenges. — 2004. — Vol. 1, No. i. — P. 1–15.

3. Yong-gui W. Research on semantic web mining / W. Yong-gui, J. Zhen. — 2010. — V1-67-V1-70 p. — ISBN 978-1-4244-7164-5.

4. Ceolin D. Subjective logic extensions for the semantic web / D. Ceolin, A. Nottamkandath, W. Fokkink. — 2012. — 27–38 p.

5. Kaufmann E. Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases / E. Kaufmann, A. Bernstein // Web Semantics: Science, Services and Agents on the World Wide Web. — 2010. — Vol. 8, No. 4. — P. 377–393.

6. Auer S. Semantic wiki representations for building an enterprise knowledge base / S. Auer, B. Jungmann, F. Schönefeld // Reasoning Web. — 2007. — P. 330–333.

7. Brachman R. Knowledge representation and reasoning / R. Brachman, H. Levesque. — 2004. — ISBN 1558609326.

8. Stavrianou A. Overview and semantic issues of text mining / A. Stavrianou, P. Andritsos, N. Nicoloyannis // ACM SIGMOD Record. — 2007. — Vol. 36, No. 3. — P. 23.

9. Kyslenko Y. I. Back to basics of speech activity / Y. I. Kyslenko // Biologically Inspired Cognitive Architectures. — 2014. — Vol. 8. — P. 46–68.

10. Kyslenko Y. Cognitive architecture of speech activity and modelling thereof / Y. Kyslenko, D. Sergeiev // Biologically Inspired Cognitive Architectures. — 2015. — Vol. 12.

11. Кисленко Ю. І. Архітектура мови / Ю. І. Кисленко. — Київ : ІЗМН, 1998.

ДОВІДКА ПРО АВТОРІВ

Сергеєв Данило Сергйіович – асистент, кафедра технічної кібернетики, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

DanyloSergeiev – assistant professor, Department of Technical Cybernetics, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: d.sergeiev@gmail.com

Д. С. Сергеєв

ІНТЕГРАЦІЙНА МОЕДЕЛЬ ПРЕДСТАВЛЕННЯ ЗНАНЬ ДЛЯ СЕМАНТИЧНОГО WEB

Актуальність теми дослідження. Створення Semantic Web та інструментів для нього, є добре розвиненою сферою досліджень, як в рамках розробок W3C , так і серед також незалежних дослідників по всьому світу. Незважаючи на це, все ще залишаються невирішеними проблеми з повним його впровадженням, значна частина яких тісно пов'язана зі структурою моделі даних, яка при цьому використовується.

Постановка проблеми. Відсутність моделі представлення знань, яка може однаково якісно працювати з природно-мовними даними та структурованими знаннями.

Аналіз останніх досліджень і публікацій. Недавні дослідження демонструють успіх у вирішенні багатьох задач SemanticWeb, зокрема уточнення запитів для певної галузі пошуку, розробка окремих баз знань для природно-мовних та структурованих знань, взаємодія з зовнішніми системами тощо. При цьому залишається невирішеною проблема координації між різними базами знань в рамках системи.

Постановка завдання. Завданням є розробкацілісної моделі знань, що дозволить координувати обробку природно-мовних даних та структурованих знань.

Викладення основного матеріалу. Представлено модель знань для природно-мовних баз знань, що дозволяє створити базу знань, яка поєднує переваги систем представлення текстових знань та структурованих баз знань і при цьому не наслідує їх недоліки.

Висновки. Розроблена модель дозволяє спростити автоматичну обробку даних у природно-мовних базах знань, зберігаючи повноту знань, зберігаючи постійну складність операції додавання нових знань.

Ключові слова: квант знань, база знань, обробка природної мови, інтеграційний підхід.