УДК 004.62 (004.9) **DOI: 10.24151/1561-5405-2018-23-3-308-312**

Модель оценки эффективности информационного поиска в рамках концепции Семантического Web

В.В. Слюсарь

Национальный исследовательский университет «МИЭТ», г. Москва, Россия

vslyusar@mail.ru

В настоящее время повышение эффективности поиска информации имеет особое значение в связи с реализацией концепции Семантического Web. Большинство современных информационно-поисковых систем, ориентированных на оперирование знаниями, недостаточно эффективны с точки зрения реализации семантических связей. В работе предложена модель оценки эффективности поиска информации на основе частот совместной встречаемости понятий в документах с учетом семантических корреляций между информационными единицами. Показано, что использование методов онтологического проектирования и онтологического моделирования для документо-ориентированных баз знаний позволит привести уровень представления знаний и их систематизации к уровню естественного языка. Предложенная модель существенно повышает уровень достоверности поиска информации, что актуально для больших объемов данных и их интеллектуальной обработки.

Ключевые слова: поиск информации; семантическая сеть; оценка эффективности поиска

Для цитирования: Слюсарь В.В. Модель оценки эффективности информационного поиска в рамках концепции Семантического Web // Изв. вузов. Электроника. – 2018. – Т. 23. – № 3. – С. 308–312. DOI: 10.24151/1561-5405-2018-23-3-308-312

Model for Evaluation of Information Retrieval Effectiveness within Semantic Web Concept

V.V. Sliusar

National Research University of Electronic Technology, Moscow, Russia

vslyusar@mail.ru

Abstract: Nowadays, the task of increasing the efficiency of the information retrieval is of particular importance because of the implementation semantic Web concept. In this study a model for estimating the information retrieval efficiency based on the frequencies of the joint occurrence of concepts in documents has been proposed, taking into account the semantic correlations between the information units. It has been shown that the use of the ontological design methods and ontological modeling the document-based knowledge bases will permit to bring the level of the knowledge representation and systematization to the level of natural language. The proposed model significantly improves the level of the information retrieval reliability, which is especially important for processing of large amounts of data and for their intelligent processing.

Keywords: information retrieval; semantic web; information retrieval efficiency evaluation

For citation: Sliusar V.V. Model for evaluation of information retrieval effectiveness within semantic web concept. *Proc. Univ. Electronics*, 2018, vol. 23, no. 3, pp. 308–312. DOI: 10.24151/1561-5405-2018-23-3-308-312

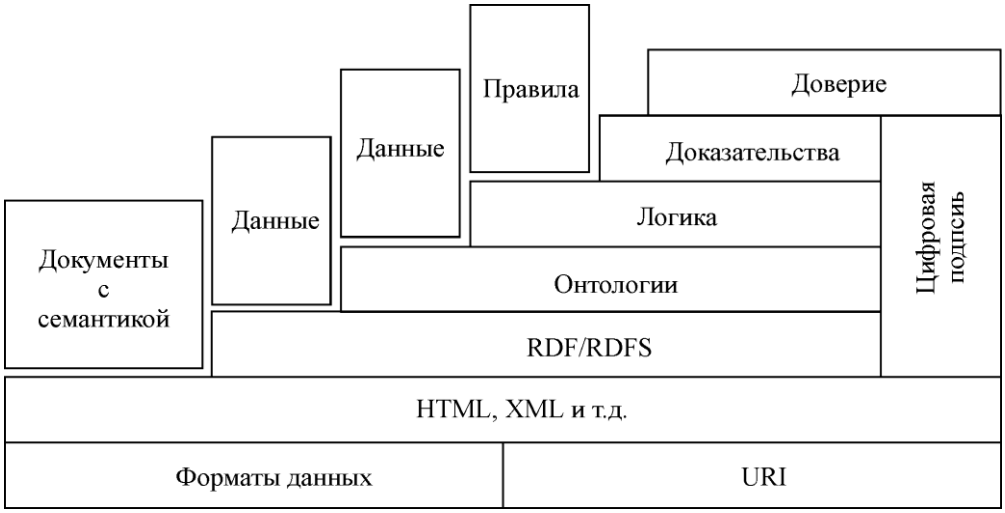
Одна из ключевых проблем, возникающая при построении поисковых систем и оценке их эффективности, состоит в необходимости выбора способа представления информации. Особое значение проблема приобретает в условиях реализации концепции Семантического Web, или Web 3.0.

Большинство современных информационно-поисковых систем, ориентированных на оперирование знаниями, недостаточно эффективны с точки зрения реализации семантических связей. Вызвано это слабым уровнем интеллектуальности поиска фрагментов в базе знаний и неудовлетворительной полнотой и точностью нахождения релевантных документов. Один из способов устранения данного недостатка – использование современных наработок в области Семантического Web.

Семантический Web – новая концепция развития сети Интернет, принятая и продвигаемая W3C (World Wide Web Consortium), разрабатывающей и внедряющей технологические стандарты для Всемирной паутины. Концепция предполагает расширение существующей сети Интернет для обеспечения точно определенного значения информационных единиц, что позволит повысить эффективность человеко-машинного взаимодействия. Применение Семантического Web направлено на повышение эффективности решения следующих проблем:

- расширение навигации в информационном web-пространстве и реализация многомерного поиска;
- семантическая интероперабельность порталов и других источников и хранилищ информации (данные из разных источников и разных форматов могут быть интегрированы в одном приложении);
- реструктуризация информации в порталах, описание содержимого и взаимосвязей web-сайтов, страниц, библиотек [1].

Семантический Web представляет собой иерархическую структуру, включающую несколько слоев моделей и языков описания информации. Расширенный вариант иерархической структуры Семантического Web, представленный в работе [2], приведен на рисунке.



Полная структура уровней Семантического Web [2]
Full semantic Web levels structure [2]

Логический вывод используется для обеспечения связности и корректности информации, для получения новых данных. Доказательства отслеживают и объясняют шаги логического вывода. Заслуживающий доверия Семантический Web (обозначенный на рисунке как «Доверие») – средства, выполняющие аутентификацию, проверку достоверности информации, надежности сервисов и агентов.

Такие современные требования к моделям представления и обработки знаний, как универсальность, открытость и возможность динамического реконфигурирования структуры базы знаний (БЗ), возможность отражения структурных отношений объектов, использование многоуровневых иерархических представлений, в полной мере сохраняются при реализации концепции Семантического Web.

Эффективность поиска информации в БЗ можно оценить с помощью следующих критериев:

- коэффициента полноты:

$$K_{\text{п}} = \lim_{k\to m}\frac{1}{k}\sum_{i=1}^{k}\frac{\left|D_i \cap D_i^0\right|}{\left|D_i\right|};$$

- коэффициента точности:

$$K_{\text{т}} = \lim_{k\to m}\frac{1}{k}\sum_{i=1}^{k}\frac{\left|D_i \cap D_i^0\right|}{\left|D_i^0\right|};$$

- коэффициента шума:

$$K_{\text{ш}} = \lim_{k\to m}\frac{1}{k}\sum_{i=1}^{k}\frac{\left|\frac{D_i}{D_i^0}\right|}{\left|D_i^0\right|},$$

где D_i – изначальное количество документов, релевантных i-му запросу; $D^0{}_i$ – полученное в результате поиска количество релевантных документов; m – достаточно большое число, обеспечивающее требуемую достоверность результата эксперимента по определению.

Традиционно в задачах анализа и полнотекстового поиска используются упрощенные векторные модели, представляющие собой описание предметной области в виде набора составляющих слов и не всегда обеспечивающие необходимое качество поиска информации в БЗ [3].

При реализации концепции Семантического Web модель автоматизированного анализа текста документа должна учитывать корреляции появления тех или иных терминов и определений в тексте, вызываемые семантическими связями между ними [4].

Обозначим набор терминов как одномерную матрицу $Q = (q_i)$ размера N, где

$$q_i = \begin{cases} 1, & \text{если } i\text{-е понятие сети входит в набор;} \\ 0, & \text{если } i\text{-е понятие сети не входит в набор.} \end{cases}$$

Семантический фрагмент может быть определен как набор входящих в него понятий $Q(t) = (q_j(t))$, где $t = 1, ..., M$ – порядковый номер фрагмента [5]. Если провести семантический анализ документа и определить группы понятий, появляющиеся в одном текстовом фрагменте (например, предложении), то оценка релевантности поискового образа текста запросу пользователя может быть определена как

$$R_{ij} = \text{Ч}_{\text{с.в}} / \text{Ч}_{\text{в}},$$

где $\text{Ч}_{\text{с.в}} = \sum_{t=1}^{M} \frac{q_i(t) q_j(t)}{\sum_{j=1}^{N} (q_j(t) - 1))}$ – частота совместной встречаемости понятий в фрагментах текста, нормированная с учетом количества понятий в каждом фрагменте; $\text{Ч}_{\text{в}} = \sum_{t=1}^{M} q_i(t)$ – частота встречаемости понятия в тексте.

Задача становится актуальной при необходимости обработки значительных объемов данных и применения технологий интеллектуальной обработки данных [6, 7].

Таким образом, можно говорить о существенном повышении уровня знаний и приближении к уровню «Доверие» в иерархической структуре, представленной на рисунке, с помощью предложенной оценки. При дальнейшем построении модели знаний на основе применения предложенной оценки в рамках концепции Семантического Web можно существенно сжать предметную область, что исключает «мнимые» понятия, ошибочно входящие в рассматриваемую предметную область.

Дальнейшее развитие систем управления знаниями – расширение использования методов онтологического проектирования и онтологического моделирования для документо-ориентированных БЗ, что позволит привести уровень представления знаний и их систематизации к уровню естественного языка. Такое представление знаний, в свою очередь, обеспечит высокое качество обмена и управления информацией и знаниями в условиях повсеместного распространения семантических метаязыков, описывающих содержание документов, хранящихся в БЗ для автоматизированного поиска, обмена и выдачи информации по запросам пользователей и информационных систем.

Литература

1. Среда описания ресурса RDF: понятия и абстрактный синтаксис. – URL.: https://www.w3.org/2007/03/rdf_concepts_ru/ (дата обращения: 25.12.2017).

2. ***Антониоу Г., Грос П., Хармелен Ф., Хоекстра Р.*** Семантический веб. – М.: ДМК Пресс, 2016. – 240 с.

3. ***Mizarro St.*** Relevance: the whole history // J. of the American Society for Information Science. – 1997. – Vol. 48. – Iss. 9. – P. 810–832.

4. ***Гасанов Э. Э., Кудрявцев В.Б.*** Интеллектуальные системы. Теория хранения и поиска информации. – 2-е изд., испр. и доп. – М.: Юрайт, 2016. – 289 с.

5. ***Баин А.М., Слюсарь В.В., Со Тант.*** Методика автоматизированного анализа документированной информации в системах поддержки принятия решений // Изв. вузов. Электроника. – 2008. – №3. – С. 81–84.

6. ***Слюсарь В.В., Николаев О.В., Высочкин А.В.*** Особенности онтологического проектирования в задачах интеллектуальной обработки данных // Оборонный комплекс – научно-техническому прогрессу России. – 2016. – №4. – С. 22–26.

7. ***Маннинг К., Рагхаван П., Шютце Х.*** Введение в информационный поиск. – М.: Вильямс, 2014. – 528 с.

Поступило 19.02.2018 г.; принято к публикации 27.02.2018 г.

Слюсарь Валентин Викторович – кандидат технических наук, доцент кафедры информатики и программного обеспечения вычислительных систем Национального исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, д. 1), vslyusar@mail.ru

References

1. *Sreda opisaniya resursa RDF: ponyatiya i abstraktnyj sintaksis* [Research Description Framework (RDF): basics and syntax]. Available at: https://www.w3.org/2007/03/rdf_concepts_ru/ (accessed: 25.12.2017). (In Russian).

2. Antoniou G., Gros P., van Harmelen F., Hoekstra R. *Semanticheskij veb* [A Semantic Web Primer]. Moscow, DMK Press Publ., 2016. 240 p. (In Russian).

3. Stefano Mizarro. Relevance: the whole history. *Journal of the American society for information science*, September, 1997, vol. 48, iss. 9, pp. 810–832.

4. Gasanov E.E., Kudryavcev V.B. *Intellektual'nye sistemy. Teoriya hraneniya i poiska informacii. 2-e izd., ispr. i dop*. [Intellectual systems. Theory of information storage and retrieval. 2nd ed.]. Moscow, Yurajt Publ., 2016. 289 p. (In Russian).

5. Bain A.M., Slyusar' V.V., So Tant. Metodika avtomatizirovannogo analiza dokumentirovannoj informacii v sistemah podderzhki prinyatiya reshenij [Documentary Information Automated Analytical Processing Methods in Decision Support Systems]. *Izvestiya vysshih uchebnyh zavedenij. Elektronika = Proceedings of Universities. Electronics*, 2008, no. 3, pp. 81–84. (In Russian).

6. Slyusar' V.V., Nikolaev O.V., Vysochkin A.V. Osobennosti ontologicheskogo proektirovaniya v zadachah intellektual'noj obrabotki dannyh [Applying ontological engineering for the intellectual data analysis tasks]. *Oboronnyj kompleks – nauchno-tekhnicheskomu progressu Rossii = Defense complex for the scientific and technical progress of Russia*, 2016, no. 4, pp. 22–26. (In Russian).

7. Manning K., Raghavan P., Shyutce H. *Vvedenie v informacionnyj poisk* [Introduction to Information Retrieval]. Moscow, Vil'yams Publ., 2014. 528 p. (In Russian).

Submitted 19.02.2018; Accepted 27.02.2018.

Information about the author:

Valentin V. Sliusar – Cand. Sci. (Tech.), Assoc. Prof. of the Computer Science Department, National Research University of Electronic Technology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), vslyusar@mail.ru