УДК 004.65:004.75:004.5:004.91

ПОРТАЛ ЯДЕРНЫХ ЗНАНИЙ BELNET – СЕМАНТИЧЕСКИЙ ПОРТАЛ

С.Н. Сытова, В.В.Гавриловец, А.П. Дунец, А.Н. Коваленко, С.В.Черепица
Институт ядерных проблем Белгосуниверситета, Минск, Беларусь,
e-mail: sytova@inp.bsu.by

Статья посвящена становлению активно развивающегося в настоящее время портала ядерных знаний BelNET https://belnet.bsu.by/ как семантического портала. В настоящее время в Республике Беларусь формируется система управления ядерными знаниями и портал BelNET призван сыграть в ней одну из ключевых ролей [1].

Необходимость использования семантических технологий в системе управления контентом портала BelNET вызвана тем, что ранее разработчики столкнулись с отрицательным опытом при ручной систематизации (определении разделов и подразделов портала, в которых должна быть размещена запись) создаваемых на портале BelNET новых записей. В результате этого многие разделы таксономии (иерархической структуры портала), которая оказалась очень большой и сложной, до сих пор остаются пустыми либо слабо заполненными. Однако, очевидно, многие ресурсы могли бы быть размещены в нескольких разделах и стать более доступными для читателей.

Для развития семантических технологий на портале BelNET были предложены оригинальные алгоритмы автоматической систематизации – размещения записей контента в таксономии портала на основе семантических технологий с использованием реализованного ранее инструмента полнотекстового поиска. Данные алгоритмы позволяют, во-первых, автоматически формировать группу дескрипторов – ключевых слов данного ресурса, а также автоматически определять один или несколько разделов для размещения на портале создаваемого информационного ресурса (записи). Тщательное автоматическое определение ключевых слов необходимо для повышения видимости и позиций показа материалов портала в поисковой выдаче поисковых интернет-систем.

Создатель новой записи (нового материала для контента портала) начинает ее размещение в своем личном кабинете на портале. Далее производится автоматический анализ текста и пользователю предлагается утвердить набор ключевых слов и положение создаваемой записи в таксономии портала (один или несколько разделов и подразделов). Автор

подтверждает эти данные либо дополнительно предлагает свои варианты, после чего происходит окончательная систематизация ресурса на портале.

Для реализации указанных семантических алгоритмов необходимо разработать несколько тематических глоссариев, объединенных в тезаурус.

Глоссарий – перечень терминов предметной области (ПрО) с их определениями. В состав глоссария, как правило, включаются термины, которые часто используются в узкой предметной области. Более широкий и системный перечень терминов и понятий ПрО называется тезаурусом. Глоссарий, как правило, составляется на основе ограниченного набора текстов ПрО и предназначен для решения некоторой частной задачи. Составление тезауруса – серьезная системная работа, которая требует значительных ресурсов и привлечения экспертов ПрО.

В списке проблем, которые приходится решать при составлении тезауруса, можно указать следующие:

1. Полнота покрытия понятийного поля ПрО – разработчики должны охватить всю область знаний, не упустив ни одной части.

2. Верификация определений – сверка и согласование формулировок определений экспертами.

3. Структурирование терминов – организация понятийного поля в иерархическую, древовидную или другую систему взаимосвязей с учетом семантики ПрО.

4. Установка границ понятийного поля ПрО.

Отметим, что нередко разработчики тезауруса начинают ощущать себя энциклопедистами в стиле Дидро и Д'Аламбера или пытаются превзойти Википедию. Необходимо учитывать, что полезный тезаурус – это терминология конкретной ПрО, собираемая и систематизируемая для решения конкретных практических задач. Попытка «объять необъятное» может ухудшить практическую полезность результатов работы и привести к неоправданным трудозатратам.

Для тестирования использовался один оригинальный белорусский глоссарий объемом 525 терминов в области ядерной и радиационной безопасности, созданный разработчиками портала BelNET. Как было отмечено выше, семантические алгоритмы рекомендуют подключить в дальнейшем несколько других необходимых глоссариев, объединенных в тезаурус.

При реализации предложенных семантических алгоритмов разработчики столкнулись с проблемой терминологии – «смысл / семантика». Был проведен анализ статьи «Атом» и выявлены следующие ключевые слова: конечное состояние, процесс, работа, система, элементы.

Со словом «процесс» получилась следующая семантическая коллизия. В нашем глоссарии значение термина «процесс» выглядит так: «*Процесс* – последовательность действий или операций, в особенности ряд последовательных стадий изготовления продукта или некоторых других

операций. Ряд взаимосвязанных или взаимодействующих операций, которые преобразуют вкладываемые ресурсы в конечные результаты».

В тексте статьи «Атом» есть фраза: «Издревле ученых, прежде всего химиков, волновали вопросы: как устроено вещество, можно ли бесконечно дробить его на все более мелкие части? Идея о том, что этот процесс не бесконечен, возникла у древнегреческих и древнеиндийских философов». Очевидно, что в данном случае проблему различия семантики в тексте и семантики термина «процесс» из глоссария может решить только автор ресурса. Интересно, что ключевое слово «атом» (Рис. 2) в тексте алгоритмом определено не было. Это связано с тем, что данный термин попросту отсутствует в глоссарии.

На Рис. 3 приведены результаты нагрузочного испытания на компьютере с процессором Intel i5-7200U (тактовая частота 2,5 ГГц) реализованного алгоритма для выяснения поведения системы на текстах различного объема. Объем оперативной памяти практически не влиял на результат, так как требуемый объем ОЗУ не превысил 50 Мб.. В качестве проверочных текстов использовались реальные документы различного объема и форматов.

Число символов	Число лемм в тексте	Число найденных терминов	Время обработки, сек
15 438	1 208	6	0.03
24 921	1 951	15	0.05
182 764	17 669	43	0.45
4 870 803	165 478	113	45.50

Рис. 3. Результаты испытаний работы системы под нагрузкой

Оказалось, что даже для очень больших документов (в рассмотренном примере самый большой документ содержал около 300 страниц и почти 5 миллионов символов) время работы находится в приемлемых рамках – 45 секунд. Типичные документы – статьи и т.п. объемом несколько страниц обрабатываются менее секунды. Поскольку от работы данных алгоритмов не требуется мгновенного результата, работа может быть выполнена в фоновом режиме, а размещение одной записью таких огромных документов зачастую не очень оптимально с точки зрения читабельности и наглядности материала. Понятно, что гораздо лучше использовать разбиение материала на главы и разделы. Таким образом, можно констатировать, что результаты реализации алгоритма являются очень хорошими.

Список литературы

1. Сытова, С. Н. Система управления ядерными знаниями в Республике Беларусь / С. Н. Сытова // Журнал Белорусского государственного университета. Физика. – 2022, № 2. – С. 87–98.