Семантическая библиотека как средство построения пространства знаний научной предметной области

О.М. Атаева[1], В.А. Серебряков[1]

[1] Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, Москва, ул. Вавилова, 40

Аннотация. В работе рассматривается информационная система, предназначенная для представления предметной области, связанной с наукой и ее особенности. Выделены общие концепции для формального описания такой предметной области в базе знаний семантической библиотеки. Особенность этих областей заключается в том, что структура данных подвержена частым изменениям. Поэтому средство организации знаний, в качестве которого выступает семантическая библиотека, должно быть достаточно универсальным и не требовать глубоких технических познаний. В работе приведено описание функциональности системы и ее использования.

Ключевые слова: семантическая библиотека, онтология, представление знаний

Semantic library LibMeta as a tool for building the knowledge space of a scientific subject area

O.M. Ataeva, V.A. Serebryakov[1]

[1] Dorodnicyn Computing Center FRC CSC of RAS

Abstract. The paper considers an information system designed to represent a subject area related to science and its features. Highlighted general concepts for formal description of such a subject area in the knowledge base of the semantic library. The peculiarity of these areas is that the data structure is subject to frequent changes. Therefore, the means of organizing knowledge, which is a semantic library, should be sufficiently universal and not require deep technical knowledge. The paper describes the functionality of the system and its use.

Keywords: semantic library, ontology, knowledge representation

Введение

Вопросами семантической организации знаний занимались различные исследователи с древнейших времен. Специализированные по конкретным областям библиотеки используют обычно свои классификаторы для систематизации своих ресурсов. Такой подход обеспечивает более детальный анализ содержания документов и соотношение смысловых понятий содержимого библиотеки с определенным направлением специализированной области знания

Накопленные данные стали доступны широкому кругу пользователей через сеть, удовлетворяя *информационные потребности* которых, функциональность цифровых библиотек становится все разнообразней.

В фокусе предлагаемой работы предметные области, связанные с наукой и их особенности. Выделены общие концепции для их формальных описаний в базе знаний. Особенность этих областей заключается в том, что структура данных подвержена частым изменениям [1 – 4]. Основной акцент сделан на представлении обобщенной модели научной предметной области и ее особенностях, реализациях в поисковых системах и отличий от классических подходов к поиску информации в научных массивах данных.

Главной задачей создания и описания обобщенного представления научных знаний для некоторой области является помощь экспертам в организации знаний и предоставления доступа к ней [5 – 9]. При этом средство организации знаний должно быть достаточно универсальным и не требовать глубоких технических познаний.

Была поставлена задача создания такой информационной системы, которая могла бы учитывать все разнообразие различных типов ресурсов научной предметной области, которые могут в ней храниться и при этом поддерживать ее терминологическое описание. Одна из основных решаемых задач в контексте системы – это обеспечение возможности интегрирования данных из источников поддерживающих семантическое описание модели данных. Фактически такая система должна представлять собой конструктор для создания цифровой библиотеки любой направленности и с адаптируемой моделью контента хранимых данных. Адаптируемая модель данных позволяет описывать произвольную модель данных контента библиотеки в рамках фиксированной в терминах тезауруса предметной области.

На данный момент реализован и готов к использованию дистрибутив семантической библиотеки. Далее представлено описание основных идей модели данных и подсистем, которые представлены в дистрибутиве информационной системы.

О модели данных

В описании информационной модели семантической библиотеки были введены понятия для описания содержимого библиотеки для некоторой предметной области [10 – 13]. Эти понятия позволяют сконструировать описание любых типов информационных ресурсов для этой области. При этом, согласно определению, информационные объекты, являющиеся непосредственно содержимым библиотеки, имеют распределенную природу, что означает, что данные могут поступать из различных источников и агрегировать информацию об информационном объекте из различных источников, непосредственно сохраняя данные в самой библиотеке или сохраняя ссылки на идентичные объекты в источниках данных.

Для описания ресурсов, составляющих контент конкретной предметной области, используются понятия, общие для любой из них. То есть набор понятий, формирующих описание контента библиотеки, должен быть настолько универсальным, чтобы мог адаптироваться под нужды конкретной области.

Контент библиотеки тесно связан с тезаурусом, который поддерживает родственные связи различных типов, как между концептами, так и между концептами и информационными объектами. Это позволяет реализовать гибкий настраиваемый поиск, результатом которого будет сбалансированный список объектов по предметной области. На основе одного и того же тезауруса определяются коллекции самых разнообразных типов ресурсов. Такой подход чрезвычайно полезен для создания раздельных пользовательских коллекций

Фактически, понятия делятся на три категории: первая включает определения понятий контента семантической библиотеки, вторая категория относится к определению понятий необходимых для поддержки терминов в тезаурусе предметной области и третья включает определения, необходимые для определения процессов интеграции контента этих ресурсов [14 – 23]. На основе этих определений описываются основные процессы такие, как, например, интегрирование данных из разных источников, категоризация/классификация, отображение разных моделей данных источников на заданную предметную область, построение классов эквивалентности и т.д.

Архитектура

Рассмотрим формальное описание системы, определяющее ее цели, функции, внешне видимые свойства и интерфейсы. Оно также включает описание компонентов системы и их отношений наряду с принципами, управляющими ее дизайном, функционированием и возможным последующим развитием. Это описание включает программные

подсистемы, визуализированные свойства этих подсистем, отношения между подсистемами и ограничения на их использование. При этом каждая подсистема может состоять из нескольких уровней абстракции, и каждый уровень может иметь свою архитектуру. Ниже приведен список подсистем:

- подсистема описания контента информационной системы;
- подсистема управления тезаурусом;
- подсистема поддержки коллекций;
- подсистема автоматизированной обработки и представления данных;
- подсистема реализации задач интеграции данных из источников LOD;
- подсистема поддержки пользователей LibMeta;
- подсистема поддержки микротезауруса пользователя;
- рекомендательная подсистема.

Каждая из этих подсистем отвечает за определенную функциональность и использует определенное подмножество понятий из информационной модели.

Основная функциональность LibMeta

Основная функциональность LibMeta:

- создание/просмотр/редактирование информационных ресурсов и их структуры;
- создание/просмотр/редактирование информационных объектов и их структуры;
- подключение источников данных;
- загрузка данных из подключенных источников данных, в дальнейшем становящихся частью контента библиотеки;
- создание/просмотр/редактирование структуры тезауруса поддерживаемой предметной области;
- создание/просмотр/редактирование понятий тезауруса
- пакетная загрузка данных составляющих контент библиотеки;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным информационным объектам системы;
- атрибутный/семантический/полнотекстовый поиск по источникам данных;
- создание/просмотр/редактирование коллекций информационных объектов;
- формирование онтологии предметной области по описанию структуры информационных ресурсов и тезауруса;

- предоставление данных составляющих контент системы в машиночитаемом формате;
- выделение связей между информационными объектами и понятиями тезауруса;
- поддержка семантических меток или *фолксономии* [24 – 26] для описания тематической направленности информационных объектов;
- создание/просмотр/редактирование области интересов пользователя;
- создание рекомендательной системы:
    a. на основе описания интересов пользователя;
    b. на основе рассматриваемого тезауруса предметной области;
- поддержка микротезаурусов пользователей на основе тезауруса предметной области.

Функциональность LibMeta, доступная для всех публичных пользователей:

- просмотр информационных ресурсов и их структуры;
- просмотр информационных объектов и их структуры;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным ресурсам системы;
- атрибутный и семантический поиск по источникам данных;
- просмотр общедоступных коллекций информационных объектов.

С точки зрения авторизированного пользователя, семантическая библиотека обеспечивает ему дополнительно следующую функциональность:

- определение своего микротезауруса как расширение некоторого узла определенного в системе основного терминологического тезауруса. Также обеспечивается поддержка создания так называемых *аннотационных онтологий* или *онтологий пользователей* (фолксономии), которые представляют собой коллективный словарь пользователей, составленный в результате процесса проставления семантических меток ими для ресурсов;
- определение собственных коллекций информационных объектов;
- организация совместных тематических коллекций для групп пользователей;
- атрибутный и семантический поиск по источникам данных с возможностью сохранения результатов поиска;
- пользователь в роли администратора системы имеет доступ ко всей вышеопределенной функциональности и может воспользоваться дополнительной, доступной только ему функциональностью:
    a. может по запросу пользователей расширять описания типов ресурсов или создавать новые;
    b. может по запросу пользователей включать их объекты ресурсов в общедоступный список объектов;

c. для групп пользователей делать доступными возможности редактирования определенных типов ресурсов или таксономий;
d. редактировать группы и роли пользователей и набор доступных им операций;
e. осуществлять редактирование и настройку основного терминологического тезауруса и его связей.

Примеры

1. «Обыкновенные дифференциальные уравнения»

В качестве примера реализации семантической библиотеки была использована предметная область обыкновенных дифференциальных уравнений. На основе предоставленной семантической библиотекой функциональности было выполнено конструирование библиотеки для этой области. В качестве тезауруса использован тезаурус ОДУ, разработанный коллективом специалистов в этой области [27].

В данной задаче в качестве внешнего источника данных была использована система MathNet, для выявления дополнительных связей между информационными объектами, а именно персонами и публикациями. В качестве источника данных было использовано внутреннее RDF хранилище с данными из MathNet.

2. «Математическая энциклопедия»

Другим примером реализации семантической библиотеки является электронная версия советской математической энциклопедии 1978 года. Это справочное издание по всем разделам математики, основу которого составляют статьи, посвященные важнейшим направлениям математики. Принцип расположения статей в энциклопедии — алфавитный. В ней широко используется система ссылок на другие статьи. В данной реализации было рассмотрено все богатство связей между ее понятиями. Также для ее расширения использовалась англоязычная версия этой энциклопедии, а именно использовались коды MSC проставленные для понятий в англоязычной версии и текстовое представление формул в виде TEX нотаций [28]. В качестве внешнего источника данных была использована система DBPedia, для выявления дополнительных связей.

3. Семантическая библиотека «Задачи математической физики»

Сейчас ведется разработка семантической библиотеки в предметной области задач математической физики. Так как область уравнений математической физики, куда входят уравнения в частных производных, как предметная область, включает в себя необъятное количество материала, в тезаурусе ограничиваются вопросами определения терминологии для ***идентификации физических*** процессов, как основы

для математических моделей и для уравнений в частных производных с примерами из уравнений смешанного типа [29].

На данный момент все три приведенных примера объединены в один ресурс.

Выводы

Представлено описание прототипа информационной системы для реализации функциональности семантической библиотеки для некоторой предметной области. Таким образом, эксперты предметной области получают возможность реализации главной задачи библиотеки – *семантического/интеллектуального* конструирования научного пространства знаний для некоторой предметной области. То есть наделение его семантикой за счет выделения явно интеллектуально значимых связей, поддержкой семантической разметки. Основным инструментов конструирования является, конечно, онтология предметной области, которая позволяет осмысленно структурировать и обеспечить связность между ресурсами, которые включены в научное пространство знаний предметной области и использование унифицированной терминологической поддержки в виде тезауруса этой предметной области. Для реализации функций открытости научного пространства знаний были реализованы возможности интеграции других источников данных и возможности связывания с их данными. Предоставление функциональности для совместной работы над развитием пространства научного знания, повышает эффективность проводимых в нем исследований и расширяет возможности по его поддержке в актуальном состоянии, несмотря на лавинообразный рост информации последние десятилетия.

Литература

1.Леонова Ю. В., Федотов А. М. Создание прототипа системы управления информационными ресурсами //Вестник Восточно-Казахстанского гос. Техн. Университета и журнала Вычислительные технологии ИВТ СО РАН.–CITech-2018, Усть-Каменогорск, Казахстан. – 2018. – С. 47-56.

2.Кулагин М. В., Лопатенко А. С. Научные информационные системы и электронные библиотеки. Потребность в интеграции //Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – 2001.

3.Шокин Ю. И., Федотов А. М., Барахнин В. Б. Проблемы поиска информации. – 2010.

4.Börner K. et al. VIVO: A semantic approach to scholarly networking and discovery //Synthesis lectures on the Semantic Web: theory and technology. – 2012. – Т. 7. – №. 1. – С. 1-178.
5.Нгок Н. Б., Тузовский А. Ф. Обзор подходов семантического поиска //Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – №. 2-2 (22).
6.Апанович З. В., Винокуров П. С., Кислицина Т. А. Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data //Труды. – 2011. – С. 113-120.
7.Оробинская Е. А., Дорошенко А. Ю. Использование онтологий для автоматической обработки текстов на естественном языке. – 2011.
8.Добров Б. В., Лукашевич Н. В. Тезаурус РуТез как ресурс для решения задач информационного поиска //Труды Всероссийской Конференции Знания-Онтологии-Теории (ЗОНТ-09), Новосибирск. – 2009. – Т. 10.
9.Ngonga Ngomo A. C. et al. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language //Proceedings of the 22nd international conference on World Wide Web. – ACM, 2013. – С. 977-988.
10.Серебряков В.А., Атаева О.М. Основные понятия формальной модели семантических библиотек и формализация процессов интеграции в ней // Программные продукты и системы. 2015. № 4. С. 180-187.
11.О. М. Атаева, В. А. Серебряков Персональная открытая семантическая цифровая библиотека LibMeta. Конструирование контента. Интеграция с источниками LOD // Информ. и её примен., 11:2 (2017), 85–100.
12.Атаева О.М. Информационная модель семантической библиотеки LibMeta // Программные продукты и системы. 2016. № 4. С. 36-44.
13.Атаева О. М., Серебряков В. А. Онтология цифровой семантической библиотеки LibMeta //Информатика и её применения. – 2018. – Т. 12. – С.
14.Ломов П.А., Шишаев М.Г. Интеграция онтологий с использованием тезауруса для осуществления семантического поиска // Информационные технологии и вычислительные системы. - 2009. - № 3. -С. 49-59.
15.Katsis Y., Papakonstantinou Y. View-based data integration //Encyclopedia of Database Systems. – 2009. – С. 3332-3339.
16.Xu L., Embley D. W. Combining the Best of Global-as-View and Local-as-View for Data Integration //ISTA. – 2004. – Т. 48. – С. 123-36.
17.Когаловский М. Р. Методы интеграции данных в информационных системах //Институт проблем рынка РАН. – 2010. – Т. 74.
18.Карабач А. Е. Системы интеграции информации на основе семантических технологий //Наука, техника и образование. – 2014. – №. 2 (2).

19.Lenzerini M. Data integration: A theoretical perspective //Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. – ACM, 2002. – С. 233-246.
20.Calvanese D., De Giacomo G., Lenzerini M. Ontology of Integration and Integration of Ontologies //Description Logics. – 2001. – Т. 49. – №. 10-19. – С. 30.
21.Noy N. F. Semantic integration: a survey of ontology-based approaches //ACM Sigmod Record. – 2004. – Т. 33. – №. 4. – С. 65-70.
22.Zhao L., Ichise R. Ontology integration for linked data //Journal on Data Semantics. – 2014. – Т. 3. – №. 4. – С. 237-254.
23.Ле Хоай, Тузовский, А.Ф.: Разработка семантических электронных библиотек на основе онтологических моделей. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года, сс. 143- 151 (2013)
24. Noruzi A. Folksonomies:(un) controlled vocabulary? //KO KNOWLEDGE ORGANIZATION. – 2006. – Т. 33. – №. 4. – С. 199-203.
25. Vander Wal T. Folksonomy. – 2007.
26. Gruber T. Ontology of folksonomy: A mash-up of apples and oranges //International Journal on Semantic Web and Information Systems (IJSWIS). – 2007. – Т. 3. – №. 1. – С. 1-11.
27.Муромский А. А., Тучкова Н. П. О тезаурусе для предметной области" Обыкновенные дифференциальные уравнения". – Вычисл. центр им. АА Дородницына РАН, 2004.
28.Ataeva O., Serebryakov V., Sinelnikova E.. Thesaurus and Ontology Building for Semantic Library Based on Mathematical Encyclopedia // Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019). p.148-157
29.Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Mathematical Physics Branches: Identifying Mixed Type Equations // Lobachevskij Journal of Mathematics. 2019. Vol. 40. N. 7. P. 876–886. DOI: 10.1134/S1995080219070047

References

1.Leonova YU. V., Fedotov A. M. Sozdanie prototipa sistemy upravleniya informacionnymi resursami //Vestnik Vostochno-Kazahstanskogo gos. Tekhn. Universiteta i zhurnala Vychislitel'nye tekhnologii IVT SO RAN.– CITech-2018, Ust'-Kamenogorsk, Kazahstan. – 2018. – S. 47-56.
2.Kulagin M. V., Lopatenko A. S. Nauchnye informacionnye sistemy i elektronnye biblioteki. Potrebnost' v integracii //Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii. – 2001.

3.SHokin YU. I., Fedotov A. M., Barahnin V. B. Problemy poiska informacii. – 2010.
4.Börner K. et al. VIVO: A semantic approach to scholarly networking and discovery //Synthesis lectures on the Semantic Web: theory and technology. – 2012. – T. 7. – №. 1. – C. 1-178.
5.Ngok N. B., Tuzovskij A. F. Obzor podhodov semanticheskogo poiska //Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki. – 2010. – №. 2-2 (22).
6.Apanovich Z. V., Vinokurov P. S., Kislicina T. A. Sredstva vizual'nogo analiza informacionnogo napolneniya portalov, vhodyashchih v oblako Linked Open Data //Trudy. – 2011. – S. 113-120.
7.Orobinskaya E. A., Doroshenko A. YU. Ispol'zovanie ontologij dlya avtomaticheskoj obrabotki tekstov na estestvennom yazyke. – 2011.
8.Dobrov B. V., Lukashevich N. V. Tezaurus RuTez kak resurs dlya resheniya zadach informacionnogo poiska //Trudy Vserossijskoj Konferencii Znaniya-Ontologii-Teorii (ZONT-09), Novosibirsk. – 2009. – T. 10.
9.Ngonga Ngomo A. C. et al. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language //Proceedings of the 22nd international conference on World Wide Web. – ACM, 2013. – C. 977-988.
10.Serebryakov V.A., Ataeva O.M. Osnovnye ponyatiya formal'noj modeli semanticheskih bibliotek i formalizaciya processov integracii v nej // Programmnye produkty i sistemy. 2015. № 4. S. 180-187.
11.O. M. Ataeva, V. A. Serebryakov Personal'naya otkrytaya semanticheskaya cifrovaya biblioteka LibMeta. Konstruirovanie kontenta. Integraciya s istochnikami LOD // Inform. i eyo primen., 11:2 (2017), 85–100.
12.Ataeva O.M. Informacionnaya model' semanticheskoj biblioteki LibMeta // Programmnye produkty i sistemy. 2016. № 4. S. 36-44.
13.Ataeva O. M., Serebryakov V. A. Ontologiya cifrovoj semanticheskoj biblioteki LibMeta //Informatika i eyo primeneniya. – 2018. – T. 12. – S.
14.Lomov P.A., SHishaev M.G. Integraciya ontologij s ispol'zovaniem tezaurusa dlya osushchestvleniya semanticheskogo poiska // Informacionnye tekhnologii i vychislitel'nye sistemy. - 2009. - № 3. -S. 49-59.
15.Katsis Y., Papakonstantinou Y. View-based data integration //Encyclopedia of Database Systems. – 2009. – C. 3332-3339.
16.Xu L., Embley D. W. Combining the Best of Global-as-View and Local-as-View for Data Integration //ISTA. – 2004. – T. 48. – C. 123-36.
17.Kogalovskij M. R. Metody integracii dannyh v informacionnyh sistemah //Institut problem rynka RAN. – 2010. – T. 74.
18.Karabach A. E. Sistemy integracii informacii na osnove semanticheskih tekhnologij //Nauka, tekhnika i obrazovanie. – 2014. – №. 2 (2).

19.Lenzerini M. Data integration: A theoretical perspective //Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. – ACM, 2002. – С. 233-246.
20.Calvanese D., De Giacomo G., Lenzerini M. Ontology of Integration and Integration of Ontologies //Description Logics. – 2001. – Т. 49. – №. 10-19. – С. 30.
21.Noy N. F. Semantic integration: a survey of ontology-based approaches //ACM Sigmod Record. – 2004. – Т. 33. – №. 4. – С. 65-70.
22.Zhao L., Ichise R. Ontology integration for linked data //Journal on Data Semantics. – 2014. – Т. 3. – №. 4. – С. 237-254.
23.Le Hoaj, Tuzovskij, A.F.: Razrabotka semanticheskih elektronnyh bibliotek na osnove ontologicheskih modelej. Trudy XV Vseros. nauch. konf. «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii» – RCDL’2013, g. YAroslavl', 14 – 17 oktyabrya 2013 goda, ss. 143- 151 (2013)
24.Noruzi A. Folksonomies:(un) controlled vocabulary? //KO KNOWLEDGE ORGANIZATION. – 2006. – Т. 33. – №. 4. – С. 199-203.
25.Vander Wal T. Folksonomy. – 2007.
26.Gruber T. Ontology of folksonomy: A mash-up of apples and oranges //International Journal on Semantic Web and Information Systems (IJSWIS). – 2007. – Т. 3. – №. 1. – С. 1-11.
27. Muromskij A. A., Tuchkova N. P. O tezauruse dlya predmetnoj oblasti" Obyknovennye differencial'nye uravneniya". – Vychisl. centr im. AA Dorodnicyna RAN, 2004.
28.Ataeva O., Serebryakov V., Sinelnikova E.. Thesaurus and Ontology Building for Semantic Library Based on Mathematical Encyclopedia // Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019). p.148-157
29.Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Mathematical Physics Branches: Identifying Mixed Type Equations // Lobachevskij Journal of Mathematics. 2019. Vol. 40. N. 7. P. 876–886. DOI: 10.1134/S1995080219070047