НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2018 Том 18 № 4 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2018 Vol. 18 No 4 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.822

ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА НАУЧНОГО И ОБРАЗОВАТЕЛЬНОГО КОНТЕНТА НА ОСНОВЕ СВЯЗАННЫХ ОТКРЫТЫХ ДАННЫХ

И.А. Радченко[a], М.А. Навроцкий[a], О.В. Герасин[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: iradche@gmail.com

Информация о статье
Поступила в редакцию 29.03.18, принята к печати 15.05.18
doi: 10.17586/2226-1494-2018-18-4-686-689
Язык статьи – русский

Ссылка для цитирования: Радченко И.А., Навроцкий М.А., Герасин О.В. Информационно-поисковая система научного и образовательного контента на основе связанных открытых данных // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 4. С. 686–689. doi: 10.17586/2226-1494-2018-18-4-686-689

Аннотация
Приведен краткий обзор результатов разработки прототипа информационно-поисковой системы научного и образовательного контента на основе связанных открытых данных. Авторы предложили решение проблемы сбора информации о публикациях из децентрализованных источников. Представлен принцип работы системы, кратко описана архитектура прототипа системы. Особенностью разработанного прототипа является предоставление возможности поиска научных публикаций с учетом профессиональных интересов пользователя. Данная работа направлена на стимуляцию и увеличение популярности такого направления, как связанные открытые данные в академической среде.

Ключевые слова
семантический поиск, семантический веб, связанные открытые данные, SPARQL, RDF

INFORMATION RETRIEVAL SYSTEM OF SCIENTIFIC AND EDUCATIONAL CONTENT BASED ON LINKED OPEN DATA

I.A. Radchenko[a], M.A. Navrotskiy[a], O.V. Gerasin[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: iradche@gmail.com

Article info
Received 29.03.18, accepted 15.05.18
doi: 10.17586/2226-1494-2018-18-4-686-689
Article in Russian

For citation: Radchenko I.A., Navrotskiy M.A., Gerasin O.V. Information retrieval system of scientific and educational content based on linked open data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 4, pp. 686–689 (in Russian). doi: 10.17586/2226-1494-2018-18-4-686-689

Abstract
The paper briefly provides the results of developing an information retrieval system prototype for scientific and educational content based on linked open data. The authors proposed a solution for the problem of information acquisition on publications from decentralized sources. The paper also presents the system functioning principle and shortly describes the architecture of the system prototype. A feature of the developed prototype is the provision of the ability to search for scientific publications taking into account the professional interests of the end user. The aim of this work is spreading a word and increasing of the linked open data popularity in the academia.

Keywords
semantic search, semantic web, LOD, linked open data, SPARQL, RDF

С развитием таких направлений, как машинное обучение, связанные данные и большие данные, открытый доступ к данным приобретает все большую ценность [1]. Публикация связанных открытых данных предоставляет обширные возможности для дальнейшей машинной обработки в силу того, что связанные открытые данные предоставляются в структурированной форме, снабженной семантической компонентой [2]. Но поиск среди таких данных для конечного пользователя обычно затруднителен, так как это требует от него знания синтаксиса языка запросов SPARQL. Отсутствие инструмента централизованного поиска по нескольким источникам данных также препятствует широкому распространению связанных открытых данных среди пользователей Интернета.

Для облегчения поиска среди децентрализованных источников данных был разработан прототип системы информационного поиска, которая осуществляет поиск среди связанных открытых данных, а также скрывает от конечного пользователя сложность запросов. Прототип системы централизованно делает удаленные вызовы к точкам доступа, а по запросу пользователя предоставляет сохраненные данные прошлого запроса.

Связанные данные [3, 4] – это метод публикации структурированных данных, связанных между собой. Он основан на стандартах консорциума W3C: HTTP, RDF, URI и др. При публикации данных указывается унифицированный идентификатор ресурсов для однозначной идентификации объектов, предоставляются метаданные в соответствии с открытым стандартом RDF[1]. Для обнаружения большего количества информации сюда также включаются ссылки на связанные сущности. Используя данные, представленные по модели RDF, можно логически выводить новые факты, а сама модель является гибкой и легко расширяемой [5].

Когда связанные данные публикуются в открытом доступе и соответствуют требованиям, предъявляемым к открытым данным, они называются связанными открытыми данными (Linked Open Data, LOD) [6]. Научное сообщество, бесспорно, играет ключевую роль в развитии этого направления [7]. Например, многие зарубежные университеты интегрировались в международный образовательный процесс на основе связанных открытых данных путем предоставления точек доступа к связанным открытым данным (в рамках европейского проекта Linked Universities) [8].

Некоторая часть точек доступа проекта Linked Universities имеет общие предметные области (например, публикации), однако, для того, чтобы получить результаты определенного поискового запроса, нужно выполнить запрос поочередно к каждой из этих точек доступа. Для этого необходим инструмент централизованного поиска по этим источникам, предоставляющий общую страницу результатов.

После проведения анализа существующих академических онлайновых платформ (Google Scholar[2], Semantic Scholar[3]) было решено, что отличительной особенностью разрабатываемой системы должна быть возможность учета профессиональных интересов пользователя при отображении результатов поискового запроса. Для этого в прототипе системы была реализована возможность формирования отображения результатов поиска с учетом ключевых слов, представленных в профиле пользователя и отображающих его профессиональные интересы.

Разработанный прототип системы[4], архитектура которого представлена на рисунке, позволяет получать по ключевому слову информацию о публикациях из внешних независимых источников.

Для создания прототипа системы использовался язык Java 8 и веб-фреймворк Java Server Faces (JSF), в качестве системы управления базами данных в прототипе используется MongoDB 3.6.

Принцип его работы заключается в следующем.

1. Конечный пользователь вводит ключевые слова, по которым хочет найти научные публикации, и выбирает предпочитаемые источники данных – открытые точки доступа, а также устанавливает флаг – будут ли результаты для запроса браться из кеша.
2. Если конечный пользователь установил флаг, то система делает запрос к базе данных – был ли такой запрос ранее. Если в базе данных есть соответствующая запись, то результаты для поискового запроса берутся из нее.
3. Если запись не была найдена или конечный пользователь не пожелал брать кешированные результаты, то система делает удаленные SPARQL-запросы к выбранным пользователем точкам доступа с помощью Apache Jena[5] – фреймворка с открытым программным кодом для языка Java, созданным для работы с семантическим вебом. После этого система осуществляет запись результатов запроса в базу данных (кеширование).
4. После получения результатов запроса среди них выбираются те, которые могут быть интересны пользователю, т.е. название или описание которых содержат ключевые слова – интересы пользователя из его профиля.

[1] RDF – Semantic Web Standards. URL: https://www.w3.org/RDF/ (дата обращения 14.05.2018).
[2] Академия Google. URL: https://scholar.google.ru/ (дата обращения 14.05.2018).
[3] Semantic Scholar. URL: https://www.semanticscholar.org/ (дата обращения 14.05.2018).
[4] A Prototype of Semantic Search for LOD: https://github.com/LODIFMO/semantic_search (дата обращения 14.05.2018).
[5] Apache Jena. URL: https://jena.apache.org/ (дата обращения 10.05.2018).

Таким образом, разработанный прототип системы работает со связанными открытыми данными высокого качества (предоставленными разными университетами) и учитывает профессиональные интересы пользователя при формировании и отображении результатов запроса.

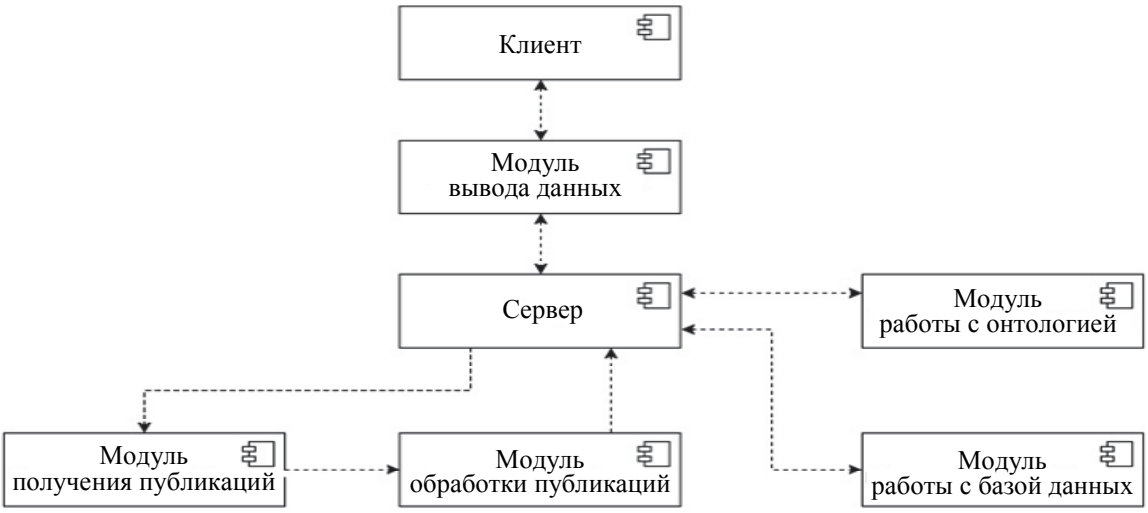

Рисунок. Архитектура прототипа информационно-поисковой системы

Для развертывания прототипа системы на локальных машинах был создан Docker-образ[1]. Для централизованного доступа к прототипу системы и базе данных было развернуто приложение на платформе Heroku[2]. Для предоставления возможности осуществления GET-запросов на получение наборов данных по ключевому слову был реализован программный интерфейс REST API.

Подводя итог, отметим, что в данной работе представлены результаты разработки прототипа системы, предназначенной для поиска публикаций по связанным открытым данным, предоставляемыми университетами. Данная работа направлена на стимуляцию и увеличение популярности такого направления, как связанные открытые данные.

Разработанный программный интерфейс REST API прототипа позволяет другим системам запрашивать результаты запросов пользователей для их последующей обработки.

В дальнейшем авторы планируют расширить систему, добавив в нее автоматическое определение интересов конечного пользователя на основе анализа его запросов.

Литература

1. Wu X., Zhu X., Wu G., Ding W. Data mining with big data // IEEE Transactions on Knowledge and Data Engineering. 2014. V. 26. N 1. P. 97–107.
2. Радченко И.А. Использование открытых данных в научных исследованиях // Информационное общество. 2013. № 1-2. С. 93–101.
3. Bizer C., Heath T., Berners-Lee T. Linked data - the story so far // International Journal on Semantic Web and Information Systems. 2009. V. 5. N 3. P. 1–22. doi: 10.4018/jswis.2009081901
4. Data – W3C. Linked Data [Электронный ресурс]. Режим доступа: https://www.w3.org/standards/semanticweb/data, свободный. Яз. англ. (дата обращения 10.05.2018).
5. Cyganiak R., Stenzhorn H., Delbru R., Decker S., Tummarello G. Semantic sitemaps: efficient and flexible access to datasets on the semantic web // Lecture Notes in Computer Science. 2008. P. 690–704. doi: 10.1007/978-3-540-68234-9_50
6. Bizer C., Heath T. Linked data: evolving the web into a global data space // Synthesis Lectures on the Semantic Web. 2011. V. 1. N 1. P. 1–136. doi: 10.2200/s00334ed1v01y201102wbe001
7. Муромцев Д.И., Леманн Й., Семерханов И.А., Навроцкий М.А., Ермилов И.С. Исследование актуальных способов публикации открытых научных данных в сети // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 6. С. 1081–1087. doi:

References

1. Wu X., Zhu X., Wu G., Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, vol. 26, no. 1, pp. 97–107.
2. Radchenko I.A. The use of open data in scientific studies. *Informatsionnoe Obshchestvo*, 2013, no. 1-2, pp. 93–101. (in Russian)
3. Bizer C., Heath T., Berners-Lee T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 2009, vol. 5, no. 3, pp. 1–22. doi: 10.4018/jswis.2009081901
4. Data – W3C. Linked Data. Available at: https://www.w3.org/standards/semanticweb/data (accessed 10.05.2018).
5. Cyganiak R., Stenzhorn H., Delbru R., Decker S., Tummarello G. Semantic sitemaps: efficient and flexible access to datasets on the semantic web. *Lecture Notes in Computer Science*, 2008, pp. 690–704. doi: 10.1007/978-3-540-68234-9_50
6. Bizer C., Heath T. Linked data: evolving the web into a global data space. *Synthesis Lectures on the Semantic Web*, 2011, vol. 1, no. 1, pp. 1–136. doi: 10.2200/s00334ed1v01y201102wbe001
7. Mouromtsev D.I., Lehmann J., Semerkhanov I.A., Navrotskiy M.A., Ermilov I.S. Study of current approaches for Web publishing of open scientific data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*,

[1] Docker Hub. URL: https://hub.docker.com/r/ovger/semanticsearchtest/ (дата обращения 14.05.2018).
[2] Semantic Search. URL: https://semanticsearchtest.herokuapp.com/ (дата обращения 14.05.2018).

10.17586/2226-1494-2015-15-6-1081-1087
8. Zablith F., Fernandez M., Rowe M. The OU linked open data: production and consumption // Lecture Notes in Computer Science. 2012. P. 35–49. doi: 10.1007/978-3-642-25953-1_4

Авторы

Радченко Ирина Алексеевна – кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56724746000, ORCID ID: 0000-0001-8658-4083, iradche@gmail.com

Навроцкий Михаил Александрович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57190969016, ORCID ID: 0000-0003-2323-8196, m.navrotskiy@gmail.com

Герасин Олег Владимирович – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0002-7336-5137, ovgerasin@gmail.com

2015, vol. 15, no. 6, pp. 1081–1087. doi: 10.17586/2226-1494-2015-15-6-1081-1087
8. Zablith F., Fernandez M., Rowe M. The OU linked open data: production and consumption. *Lecture Notes in Computer Science*, 2012, pp. 35–49. doi: 10.1007/978-3-642-25953-1_4

Authors

Irina A. Radchenko – PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56724746000, ORCID ID: 0000-0001-8658-4083, iradche@gmail.com

Mikhail A. Navrotskiy – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57190969016, ORCID ID: 0000-0003-2323-8196, m.navrotskiy@gmail.com

Oleg V. Gerasin – student, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0002-7336-5137, ovgerasin@gmail.com