УДК 004.65:378.4

АНАЛИЗ РЕПОЗИТОРИЯ УНИВЕРСИТЕТА С ИСПОЛЬЗОВАНИЕМ ГРАФОВОЙ БАЗЫ ДАННЫХ

В.А. Прытков
Декан факультета компьютерных систем и сетей БГУИР, кандидат технических наук, доцент

И.И. Пилецкий
Доцент кафедры информатики БГУИР, кандидат технических наук, доцент

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: prytkov@bsuir.by

Аннотация. Анализ университетского репозитория позволяет определить ключевые направления исследований университета, найти экспертов в предметной области, оценить динамику исследований в различных предметных областях. Однако выполнить этот анализ возможно только с использованием специализированных средств. В работе на примере репозитория БГУИР демонстрируются возможности такого анализа с использованием графовой базы данных Neo4j.

Ключевые слова: репозиторий, автор, публикация, NoSQL, графовые базы данных, Neo4j

Введение. На настоящий момент многие организации, занимающиеся научной или образовательной деятельностью, имеют свои электронные репозитории с публикациями сотрудников. Это не только дает возможность оперативно находить требуемую публикацию, но и позволяет организации занять более высокое место в различных рейтингах. Однако помимо этого репозиторий является и источником различной аналитической информации: по каким направлениям организация проводит исследования, в каких из них исследования наиболее интенсивны, какие новые направления осваиваются, кто является экспертом в различных направлениях и т.д. В то же время провести такой анализ вручную, просматривая публикации, не представляется возможным – их количество исчисляется десятками тысяч.

Различные средства автоматизации позволяют значительно облегчить процесс получения первичной информации, выполнив рутинные операции парсинга и скрапинга веб-страниц. Однако дальнейшая обработка и анализ таких слабоструктурированных данных является нетривиальной задачей.

Белорусский государственный университет информатики и радиоэлектроники создал свой репозиторий несколько лет назад. В нем хранятся научные статьи, методические материалы, учебные пособия, материалы конференций, которые для удобства сгруппированы в несколько коллекций. Мы решили посмотреть, какую информацию к размышлению может дать анализ репозитория БГУИР.

Сбор первичных данных. В ходе экспериментов выяснилось, что репозиторий БГУИР является просто идеальным источником для получения первичной информации. Во-первых, это свободный доступ к материалам даже без регистрации; во-вторых, очень простой URL для каждой из публикаций; в-третьих, очень хорошая структура разметки страницы, позволяющая без труда выделить необходимые данные.

URL репозитория БГУИР состоит из постоянной и переменной частей адреса, причем

переменная часть является уникальным идентификатором публикации. В итоге URL имеет вид: https://libeldoc.bsuir.by/handle/123456789/“идентификатор_материала”, например: https://libeldoc.bsuir.by/handle/123456789/1361. Такая простая структура позволила краулеру просмотреть все ресурсы репозитория простым перебором чисел. Обнаружилось, что имеются неиспользуемые диапазоны, кроме того, ряд страниц содержит не публикации непосредственно, а различные списки, с иной структурой разметки. На момент проведения исследования (январь 2018 года) репозиторий содержал 15796 материалов.

Разметка страницы содержит теги метаописания для поисковых систем, которые включают авторов (отдельный тег для каждого), ключевые слова (отдельный тег для каждого), название, полное библиографическое описание, дату публикации, ссылку на саму страницу, ссылку на страницу с публикацией в виде pdf файла, аннотацию, издателя и др. Примеры тегов метаописания для авторов, названия и ключевых слов приведены ниже:

```
<meta name="citation_author" content="Бевз, Е. Г." />
<meta name="citation_author" content="Прытков, В. А." />
<meta name="DC.title" content="Алгоритмы сегментации ..." xml:lang="ru_RU" />
<meta name="DC.subject" content="сегментация" xml:lang="ru_RU" />
<meta name="DC.subject" content="текстура" xml:lang="ru_RU" />
```

Парсинг такой страницы не представляет сложности, более того, поскольку ключевые слова уже выделены, для простого анализа нет необходимости применять к основному тексту статьи или аннотации алгоритмы обработки естественного языка с целью извлечения ключевых слов (основной темы) материала.

Модель данных. Для проведения эксперимента было решено ограничиться определением соавторов, основных направлений исследований автора, а также авторов, обладающих наибольшей экспертизой в предметных областях.

Для определения таких данных необходимо знать для каждого материала только его авторов и ключевые слова. Соответствующая модель представлена на рис. 1.

Модель очень проста, она включает всего три вида сущностей (вершин): публикации, авторы, темы (ключевые слова), и два вида связей: между публикацией и автором (связь авторства), между публикацией и ключевым словом (тематика публикации). Несмотря на эту простоту, она позволяет успешно решить поставленные задачи.

Рисунок 1. Модель, используемая для эксперимента

Модель хорошо описывается графом, соответственно, наиболее подходящим решением для работы с такими данными будут являться графовые базы данных. Из доступных решений единственный наиболее подходящий вариант – Neo4j, которая не только является свободно распространяемой, но и достаточно широко используется для решения прикладных задач.

В результате построения графа на основе данных репозитория нами был получен граф, содержащий 13133 автора, 27941 тематику работ, 35465 связей авторства, 62899 связей между публикациями и тематиками.

Эксперимент. Графовая база данных Neo4j поддерживает встроенный язык запросов к данным Cypher, синтаксис которого адаптирован под графовую модель, но вместе с тем подо-

бен стандартному SQL. Ниже приведен запрос, который выбирает все публикации и соответствующие им ключевые слова для одного из авторов данной статьи:

```
match (a:Author {Name: "Прытков, В. А."})-->(p:Paper)-->(t:Term) return a,p,t
```

На рисунке представлены результаты выполнения этого запроса, также приведены результаты выполнения запроса на поиск всех соавторов данного автора:

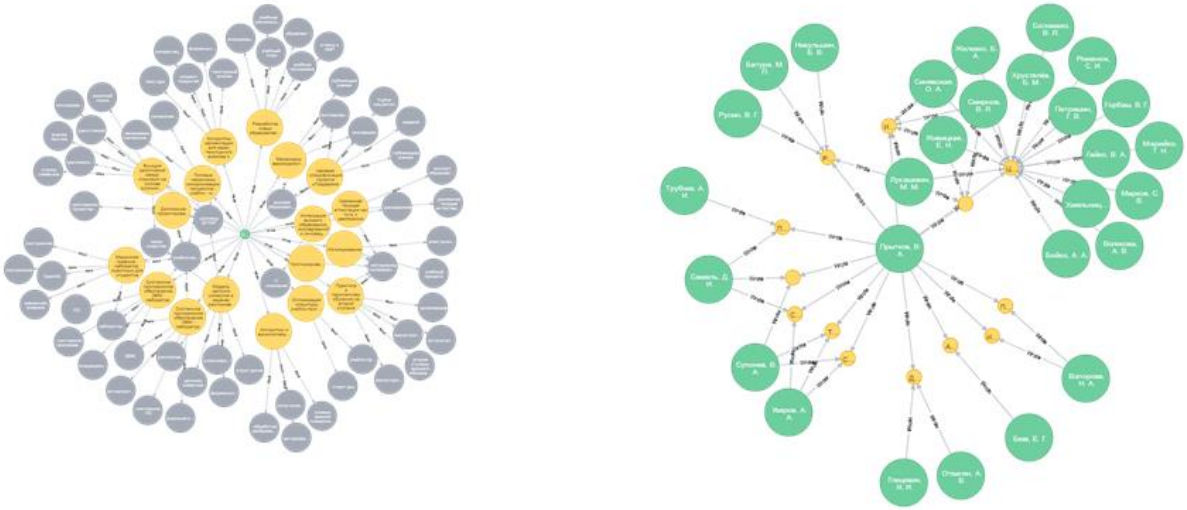

Рисунок 2. Пример выборки данных: слева - все публикации автора и соответствующие им ключевые слова; справа - все соавторы автора

Анализ показал, что в университетском репозитории всего у 27 авторов размещено более 50 публикаций, а максимальное количество размещенных публикаций одного автора — 107.

Однако для более сложных запросов из всего множества ключевых слов необходимо исключить «плохие слова», т. е. те, которые по сути не являются ключевыми. Так, если вывести первые 500 записей репозитория, то обнаружим, что все публикации разделяются на два основных подмножества, сгруппированных по наличию ключевого слова «материалы конференций» или «учебно-методические пособия».

Несмотря на то, что такие ключевые слова могут быть полезны для поисковых систем, анализ они явно затрудняют. Для упрощения эксперимента было принято решение считать «плохими словами» такие, которые встречаются у более чем 100 публикаций. Список таких слов включил 18 позиций, которые в дальнейшем исключались из результатов анализа.

Для определения, в каких предметных областях автор является экспертом, определялись такие темы (ключевые слова), которые присутствуют в качестве ключевых слов в нескольких публикациях автора. При этом «плохие слова» исключались. Соответствующий Cypher запрос уже более сложен, в нем используется потоковая передача результатов выполнения предыдущего запроса в качестве входных данных к следующему:

```
match (nt:Term)<-[r]-(p:Paper)
with nt, count(r) as cnt where cnt>100
with collect(nt.Term) as nnt
match (a:Author {Name:"Прытков, В. А."})-->(p:Paper)-[r]->(t:Term)
where not (t.Term in nnt)
with t,count(r) as cnt where cnt>1
with collect(t.Term) as nnt
match (a:Author {Name:"Прытков, В. А."})-->(p:Paper)-[r]->(t:Term)
where t.Term in nnt
```

```
return a,p,t
```

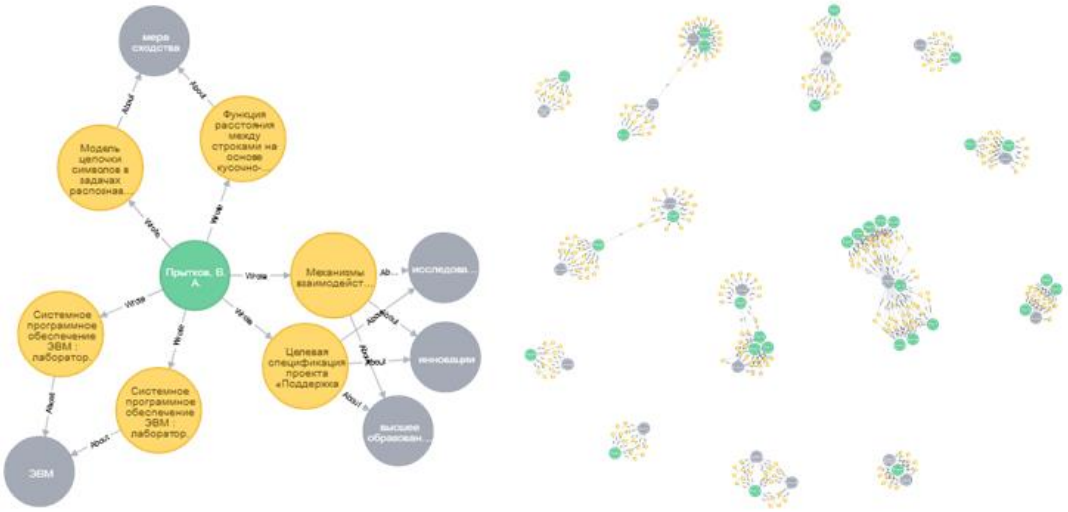

Рисунок 3. Пример выборки данных: слева — определение тем, в которых автор является экспертом; справа — темы, в которых у сотрудников университета наибольший экспертный уровень

Результат выполнения запроса приведен на рис. 3. Из рисунка видно, что у данного автора только пять тем, которые указаны в качестве ключевых слов несколько раз и в которых он, предположительно, может являться экспертом. Если сделать выборку по всему репозиторию, считая за экспертов тех авторов, у которых не менее 10 публикаций по одной теме, получим, что таких экспертов всего 31, из которых 9 — эксперты в английском языке! Очевидно, что качество и количество размещенных в репозитории материалов не в полной мере соответствует экспертизе университета.

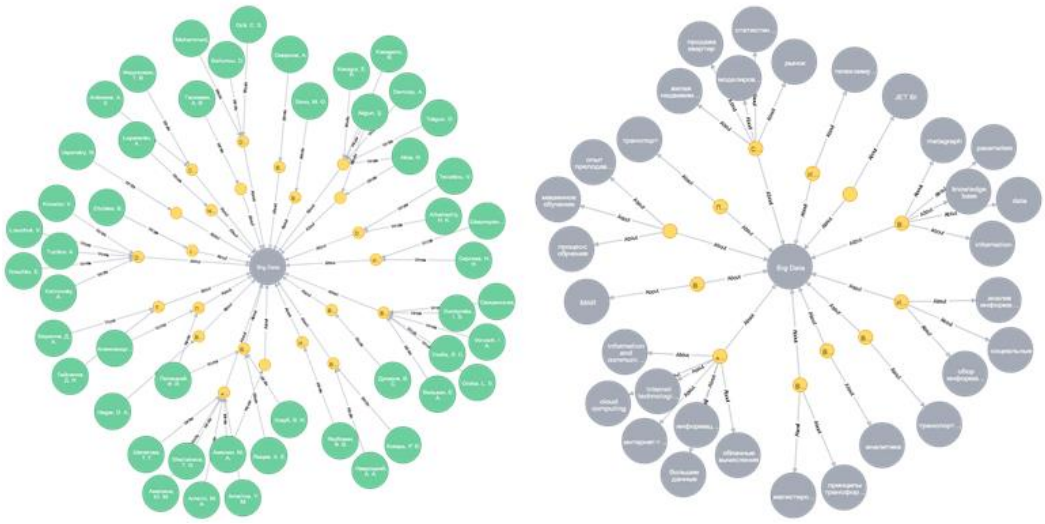

Рисунок 4. Пример выборки данных: слева — авторы, в чьих публикациях встречено ключевое слово «Big Data»; справа — ключевые слова, которые используются совместно с «Big Data

Если оценивать, кто является экспертом в конкретной предметной области, например, «Big Data», не представляет труда определить авторов соответствующих публикаций (рис. 4,

слева). Однако зачастую в качестве ключевых слов указываются более детализированные темы, а обобщающая тема включается в перечень ключевых слов далеко не всегда. На рис. 4, справа представлены ключевые слова, которые использовались совместно с «Big Data» и которые, предположительно, могут быть использованы и без «Big Data», хотя публикация может относиться к этой теме.

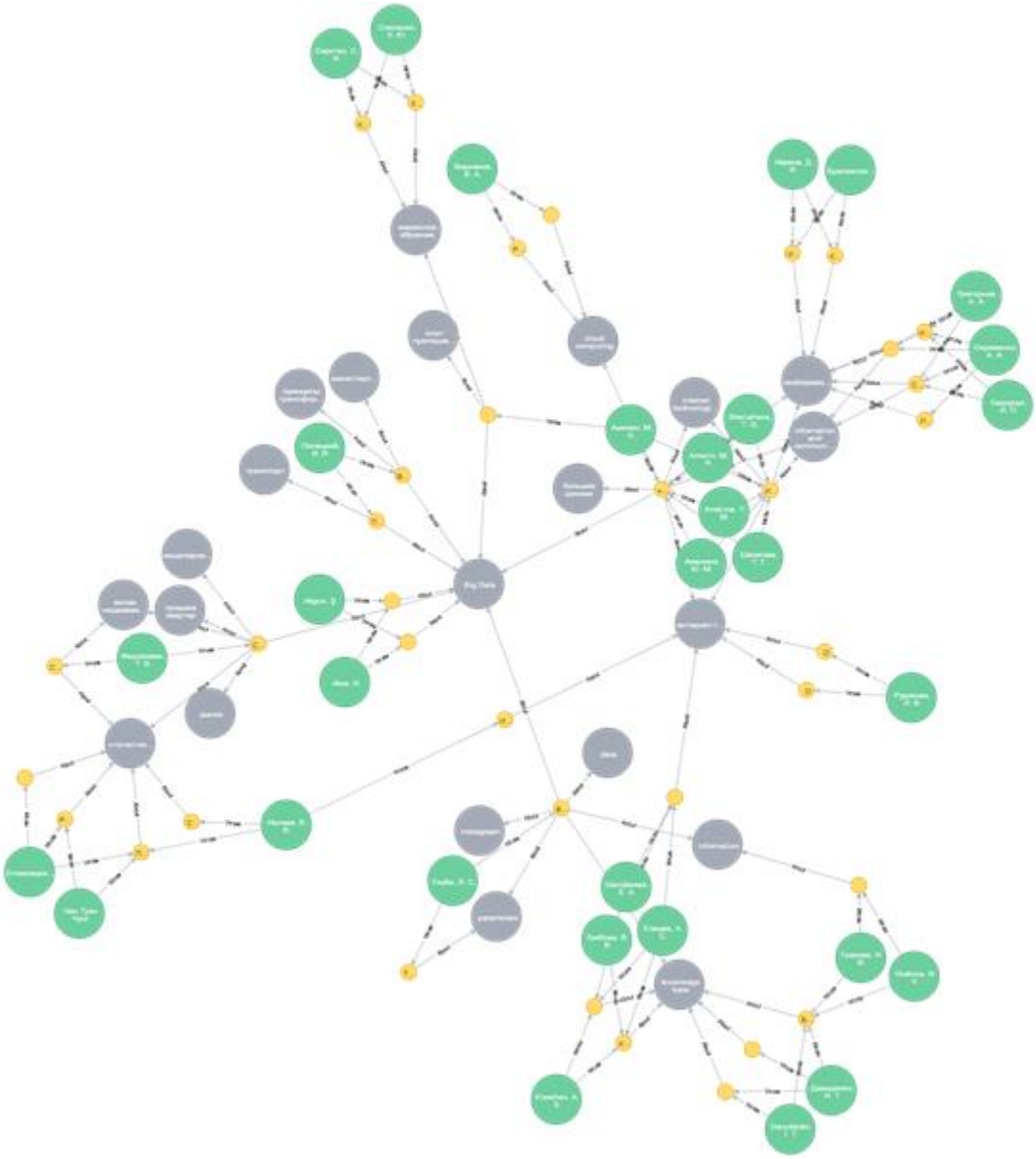

Рисунок 5. Результат выполнения запроса на определение экспертов в области «Big Data» с учетом совместно используемых ключевых слов

Была обнаружена 31 тема, которая указывалась в качестве ключевого слова совместно с «Big Data». Для оценки, кто является экспертом в конкретной предметной области, было решено оценивать не только экспертизу именно с данным ключевым словом, но и экспертизу в областях, которые указаны в качестве ключевых слов совместно с оцениваемой предметной областью. Например, если для направления «Big Data» совместно с этим ключевым словом используется и ключевое слово «Cloud Computing», то экспертами будут считаться также и те, у кого несколько публикаций по теме «Cloud Computing». Соответствующий Cypher запрос

выглядит следующим образом:

```
match (nt:Term)<-[r]-(p:Paper)
with nt, count(r) as cnt where cnt>20
with collect(nt.Term) as nnt
match (nt:Term {Term: "Big Data"})<--(p:Paper)-->(t:Term)
where not (t.Term in nnt)
with collect(t.Term)+["Big Data"] as nnt
match (a:Author)-->(p:Paper)-[*1..3]->(t:Term) where t.Term in nnt
with distinct nnt,a,p
match (a:Author)-[r]->(p:Paper)
with nnt,a, count(r) as cnt where cnt>1
with collect(a.Name) as aat,nnt
match (a:Author)-->(p:Paper)-[*1..3]->(t:Term)
where t.Term in nnt and a.Name in aat
return a,p,t
```

Результат выполнения запроса на определение экспертов для направления «Big Data» представлен на рисунке 5:

Видно, что авторов, у которых термин «Big Data» указан в качестве ключевого, всего 49, причем это авторы всего 23 публикаций. А авторов, у которых «Big Data» или смежные ключевые слова встречаются несколько раз, т. е. те, кто, предположительно, наиболее глубоко разбираются в данной теме, всего 31. И это несмотря на то, что конференция «Big Data» проводится в БГУИР, начиная с 2015 года! Как выяснилось, причина этого в том, что публикации конференции «Big Data» не имели в своей структуре информации о ключевых словах.

Заключение. Даже уже такой достаточно простой анализ на основе несложной графовой модели позволят сделать определенные выводы. Так, можно рекомендовать:

– профессорско-преподавательскому составу вуза размещать в репозитории и свои более ранние работы, это будет способствовать более точному позиционированию экспертизы вуза для внешнего наблюдателя;

– разработчикам серверной части системы проанализировать формат описания на предмет разделения ключевых слов, являющихся темой публикации, и ключевых слов для поисковых систем;

– в структуру документов публикаций включать ключевые слова;

– в перечень ключевых слов публикации включать и одно-два обобщающих ключевых слова, которые определяют глобальное направление публикации.

Для построения системы подобного типа с более совершенной моделью анализа требуется:

– определение ключевых тем публикаций не только по списку ключевых слов, но и по названию, аннотации, тексту публикации;

– более совершенный скрапинг исходных страниц, включающий определенную нормализацию слов - удаление лишних символов, преобразование некоторых символов латинского алфавита к русскому и т. д.;

– использование алгоритмов определения синонимов как для тем, так и для авторов, а в отдельных случаях — и для названий публикаций;

– более сложный анализ потребует использовать такие сведения о публикациях, как дату публикаций, издателей, цитирование, может потребовать использования более сложной графовой модели.

Список литературы

[1]. Фаулер, М., Садаладж, П. Дж. NoSQL: новая методология разработки нереляционных баз данных. : Пер. с англ. - М.: ООО "И.Д. Вильяме", 2013. - 192 с.

[2]. Робинсон Я., Вебер Д., Эифрем Э. Графовые базы данных: новые возможности для работы со связанными данными / пер. с англ. Р. Н. Рагимова; науч. Ред. А. Н. Кисилев. – 2-е изд. – М.: ДМК Пресс, 2016. – 256 с. - ISBN 978-5-97060-201-0

ANALYSIS OF THE UNIVERSITY REPOSITORY USING A GRAPH DATABASE

V.A. PRYTKOV, PhD
Dean of the Faculty of Computer Systems and Networks of the BSUIR, Associate Professor

I.I. PILETSKI, PhD
Associate Professor of Informatics Department of the BSUIR

N.A. VOLARAVA, PhD
Head of the Chair of Informatics of the BSUIR, Associate Professor

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: prytkov@bsuir.by

Abstract. The analysis of the university repository allows us to determine the key areas of university research, to find experts in the subject area, to assess the dynamics of research in various subject areas. However, this analysis is possible only with the use of specialized tools. The possibilities of such analysis demonstrates by example of the repository of the BSUIR using the Neo4j graph database.

Key words. repository, author, publication, NoSQL, graph databases, Neo4j.