

Проблемы использования данных из облака LOD для обогащения контента научных баз данных и знаний

© З. В. Апанович

Институт систем информатики им. А.П. Ершова СО РАН,
Новосибирск

apanovich@iis.nsk.su

© А. Г. Марчук

mag@iis.nsk.su

Аннотация

В данной работе описаны проблемы, возникающие в процессе использования данных из облака LOD для обогащения контента научных баз данных и знаний и подходы к их решению. Эксперименты выполнялись при помощи набора инструментов, разработанного для упрощения анализа данных из разных наборов. В качестве тестовых примеров использовались данные открытого Архива СО РАН, и его онтология ОНС, а также различные наборы библиографических данных, структурированные при помощи онтологии AKT Reference ontology.

Введение

В связи с бурно развивающимся направлением Semantic Web и его новой ветвью LOD (Связанные Открытые Данные) в Интернете становятся доступными большие объемы информации, посвященной различным научным направлениям. Облако LOD содержит в настоящий момент более 28 миллиардов троек RDF. С одной стороны, эти данные могут быть использованы для обогащения имеющихся семантических баз данных, с другой стороны, имеющиеся базы данных могут быть также полезны для уточнения информации, хранящейся в облаке LOD.

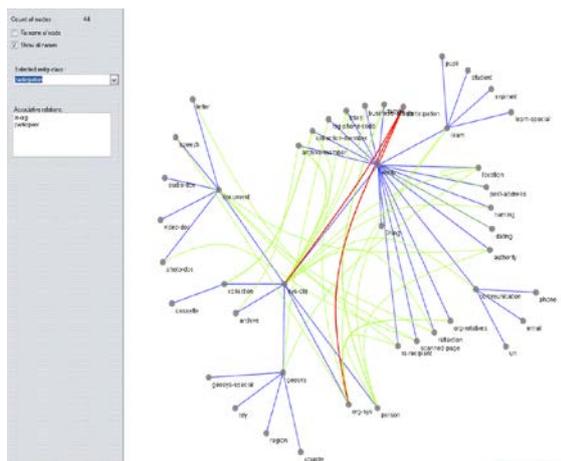
В работе [18] предложена четырехшаговая стратегия интеграции Связанных Данных в приложения. Помимо проблем, специфических для конкретного приложения, требуется решить проблему доступа к связанным данным (1), проблему нормализации словарей (2), установления идентичности сущностей (3) и фильтрации данных

(4). Способы решения этих проблем варьируют в диапазоне от полностью ручных до автоматизированных [4, 8, 10, 13, 14, 19]. При этом многие проблемы, такие, как проблема обработки данных большого объема, проблема установления соответствия между онтологиями, а также проблема объединения данных из разных наборов «еще находятся в детском состоянии» [19]. С другой стороны, проблема (1) может быть решена при помощи запросов SPARQL 1.1 [8]. Проблема (2) может быть решена как при помощи сложных запросов SPARQL 1.1. Проблема (3) может быть частично решена при помощи таких полуавтоматических инструментов, как SILK [10] или LIMES [13] совместно с использованием запросов SPARQL. Наконец, проблема (4) также может быть решена при помощи запросов SPARQL 1.1. Поскольку практически каждая из проблем может быть решена при помощи подходящего набора SPARQL-запросов, мы расширили разработанную ранее программу визуализации онтологий средствами построения SPARQL-запросов и генерации результатов как в текстовом виде, так и в виде графа. SPARQL-запросы могут быть сгенерированы также на основе нашей визуализации одной или двух онтологий. В качестве тестовых примеров использовались онтология ОНС и данные открытого Архива СО РАН [20], а также AKT Reference ontology [1], при помощи которой структурированы различные наборы библиографических данных. В работе сравнивается их структура, и обсуждается стратегия установления связей между наборами данных, описанных при помощи этих онтологий.

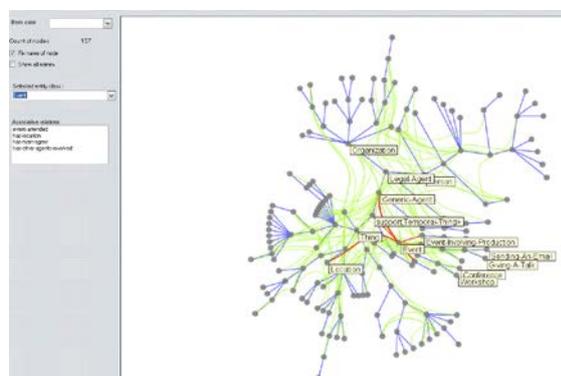
1 Визуализация онтологий для исследования семантических систем

В настоящий момент в ИСИ СО РАН выполняется проект, направленный на интеграцию баз данных, разработанных в ИСИ СО РАН с данными мирового сообщества. С этой целью изучаются базы данных облака Linked Open Data [4] и выясняются возможности интеграции с ними систем, разработанных в ИСИ СО РАН [20], в

частности, научной информации из открытого архива и фотоархива СО РАН [http://soran1957.iis.nsk.su/pa2/Home/Portrait?id=c_do1000663]. Ее основное наполнение составляют документы, посвященные различным событиям СО РАН, начиная с 1957 года. В базе имеется также структурированная информация о людях, отраженных в документах, научных организациях, и важнейших событиях в жизни СО РАН, в частности, о научных конференциях. Структура Открытого архива организована при помощи Онтологии Неспецифических Сущностей (ОНС), описанной в OWL-формате и содержащей 44 класса.



(a)



(б)

Рис. 1. (а) Классы и отношения онтологии ОНС, (б) классы и отношения АКТ reference ontology.

На Рис. 1(а) показано изображение онтология ОНС открытого архива СО РАН, построенное нашей программой визуализации [21]. Прямолинейные ребра изображают таксономию, задаваемую отношениями класс-подкласс. Криволинейные ребра изображают отношения типа *owl:ObjectProperty*. При выборе одного из классов в поле “Selected entity class” на панели визуализации графа высвечиваются все отношения, описанные в онтологии как *owl:ObjectProperty*, а в нижнем поле выдается список отношений этого класса

(*owl:DatatypeProperty* и *owl:ObjectProperty*). При выборе элемента этого списка соответствующие ребра высвечиваются в окне визуализации. Это свойство визуализации весьма существенно для понимания незнакомой онтологии. Например, при выборе в онтологии такого класса как *onc:participation* высвечиваются ребра, соединяющие классы *onc:person*, *onc:participation* и *onc:org-sys*. Эта визуализация демонстрирует специфическую особенность онтологии ОНС, состоящую в том, что многие сущности, обычно описываемые как отношения, в данной онтологии описаны как классы, компенсируя отсутствие атрибутов у отношений в формате RDF.

Для сравнения, на Рис. 1(б) показаны классы и отношения АКТ reference ontology [1], которая в облаке LOD используется для описания многих библиографических порталов, таких как DBLP, Citeseer, ACM, IEEE и др. Часть данных этих порталов представлены в облаке Open Linked Data[5, 7]. Она содержит 157 классов.

Попытки установления соответствия между этими онтологиями при помощи одной из лучших программ выравнивания AgreementMaker [6] оказались неудачными как из-за существенных лексических и структурных различий между рассматриваемыми онтологиями, так и из-за специфических особенностей онтологии ОНС, обсуждаемых ниже. Единственное очевидное соответствие наблюдается между классами *onc:person* и *akt:Person*. Остальные связи гораздо менее очевидны. Рассмотрим, например, класс *onc:participation*. В онтологии открытого архива этот класс используется для описания фактов работы персон в различных организациях, а также фактов участия в различных мероприятиях, например, конференциях. Этот класс связан отношением *onc:participant* с классом *onc:person* и отношением *onc:in-org* с классом *onc:org-sys*. Класс *onc:org-sys* используется как для описания различных организаций, так и для описания мероприятий, например, конференций. В АКТ Reference ontology эти же самые факты могут быть описаны несколькими способами. Это может быть отношение *akt:works-for* между экземпляром класса *akt:Employee* и экземпляром класса *akt:Organization*, отношение *akt:has-affiliation* между экземпляром класса *akt:Person* и экземпляром класса *akt:Organization*, а также отношения *akt:has-main-agent*, *akt:has-other-agents-involved* между экземплярами классов *akt:Event* и *akt:Generic-Agent*. Из-за наличия в онтологии ОНС таких классов как *onc:participation*, при установлении соответствия между онтологиями возникает систематическая потребность в генерации экземпляров классов, которых до этого не было ни в одной из онтологий.

2 Эксперименты по выравниванию онтологий

Рассмотрим для определенности случай генерации экземпляра класса *онс:participation* по отношению *акт:has-affiliation* между экземпляром класса *акт:Employee* и экземпляром класса *акт:Organization*. Для пополнения открытого архива информацией о местах работы из одного библиографических порталов нам потребуется сначала установить соответствие между экземплярами классов *акт:Person* и *онс:person*, *акт:Organization* и *онс:org-sys*, а затем для каждого факта наличия отношения *акт:has-affiliation* между экземплярами классов *акт:Person* и *акт:Organization* потребуется сгенерировать новый экземпляр класса *онс:participation*, а также связать его отношением *онс:in-org* с соответствующим экземпляром класса *онс:org-sys*, и отношением *онс:participant* с соответствующим экземпляром класса *онс:person*. При обратной трансляции нам потребуется генерировать отношения *акт:works-for* соответствующие экземплярам класса *онс:participation*.

Поскольку в онтологии ОНС имеется достаточно много классов, аналогичных классу *онс:participation*, систематически возникает необходимость устанавливать соответствие между различными группами классов и отношений этих двух онтологий. А именно, необходимо установить соответствие между группой вида "Class 1-relation1-Class2" онтологии АКТ Reference ontology и одной или несколькими группами вида "Class3- relation2-Class4-relation3-Class5" онтологии ОНС. При этом между объектами классов Class1 и Class3 следует установить связи типа *owl:sameAs* также как и для классов Class 2 и Class5. Помимо этого, необходимо сгенерировать экземпляр класса Class4 для каждой тройки <Class1:instance1, relation1, Class2:instance2>. Такая трансляция может быть осуществлена при помощи запроса SPARQL 1.1. Упрощенная версия этого запроса имеет следующий вид:

```
PREFIX iis:<http://iis.nsk.su#>
PREFIX akt:<http://www.aktors.org/ontology/portal#>
PREFIX
aks:<http://www.aktors.org/ontology/support#>
CONSTRUCT {
  _:p a iis:Class4.
  _:p iis:relation2 ?instance1.
  _:p iis:relation3 ?instance2.
}
WHERE {
  ?instance1 akt:relation1 ?instance2.
  ?instance1 a akt:Class1.
  ?instance2 a akt:Class2.
}
```

Для упрощения задачи нами разработана программа, которая позволяет генерировать SPARQL-запросы на основе визуализации онтологий. Пример установления такого соответствия показан на Рис. 2. Сначала в интерактивном режиме устанавливается соответствие между двумя наборами классов и отношений, а затем автоматически генерируется шаблон SPARQL-запроса, осуществляющий трансляцию данных.

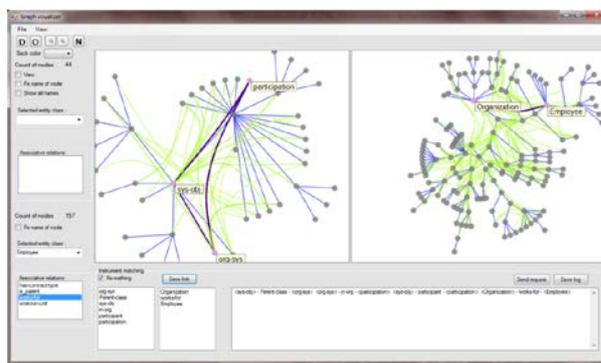


Рис. 2 Интерактивное установление соответствия между классами и отношениями двух онтологий.

3 Проблема установления идентичности сущностей

Как уже было сказано выше, существенным моментом обогащения одной базы знаний при помощи другой является этап установления идентичности сущностей в наборах данных LOD и данных открытого архива, то есть, генерация отношений вида *owl:sameAs*. Рассмотрим следующий пример. В Открытом архиве имеется экземпляр класса *онс:person*, описывающий бывшего директора ИСИ СО РАН В.Е. Котова:

```
<person rdf:about="piu_200809052136">
  <name xml:lang="ru">Котов Вадим
Евгеньевич</name>
  <name xml:lang="en">Kotov, Vadim
Yevgenievich</name>
  <from-date>1938-07-23</from-date>
  <sex>m</sex>
</person>
```

Также, в Открытом архиве СО РАН имеется достаточно подробная информация о его местах работы как в различных организациях СО РАН, так и в США. При этом отсутствует информация о его публикациях. С другой стороны, достаточно много информации о публикациях В.Е. Котова содержится в различных наборах данных облака LOD таких как: *actm.rkbexplorer.com*, *dblp.rkbexplorer.com*, *citeseer.rkbexplorer.com*. Но в этих наборах данных нет информации о местах работы В.Е. Котова, присутствующей в Открытом архиве. Для взаимовыгодного обмена данными надо, прежде

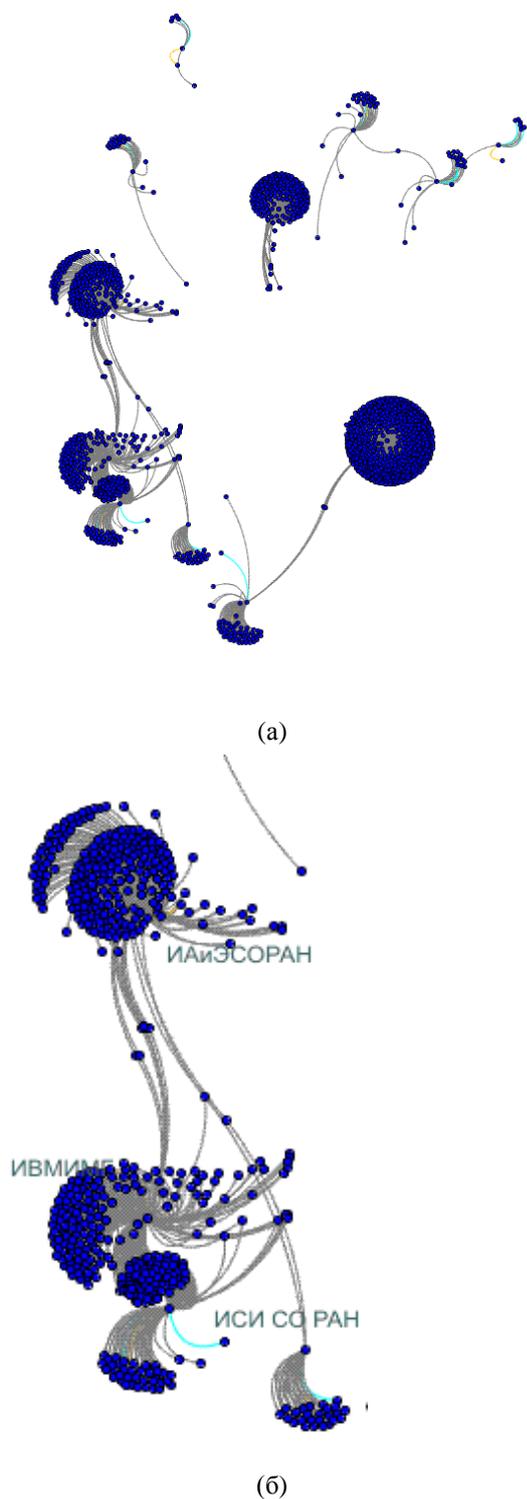


Рис. 4. Сети, выдаваемые в результате запросов к семантической системе открытого архива СО РАН.

Граф состоит из нескольких компонент связности. Люди сгруппированы вокруг организаций, в которых они работают или работали. Следует отметить, что поскольку указанный запрос не использовал фильтрацию по дате работы, некоторых людей ребра связывают с несколькими организациями. Так на Рис. 4(б) показан фрагмент изображения с Рис. 4(а), на котором видно, что

часть коллег связано с ИСИ СО РАН, часть - с ИВМ и МГ, а часть – с обеими организациями. Эта промежуточная часть достаточно велика, поскольку ИСИ СО РАН был создан на базе одного из отделов ИВМ и МГ. Помимо просмотра данных из исследуемых баз данных эта компонента дает нам возможность визуализации и исследования сетей цитирования и соавторства и их кластеризацию, что важно для данного приложения.

Заключение

В данной работе рассмотрены проблемы обогащения научных баз знаний при помощи контента библиографических порталов из облака LOD и подходы к их решению. Сравняются онтология ОНС и АКТ Reference ontology и соответствие между наборами данных, основанными на этих онтологиях, устанавливается при помощи SPARQL-запросов, которые могут быть сгенерированы на основе визуализации онтологий.

Эксперименты показали, что для выравнивания онтологий недостаточно установления простых соответствий и могут потребоваться более сложные шаблоны. Также продемонстрировано, что обычные инструменты, применяемые для установления идентичности сущностей на основе метрик сходства, не позволяют различать авторов публикаций, являющихся тезками или даже однофамильцами. Для решения этой проблемы планируется использовать информацию о временных границах мест работы персон, описанных в Открытом Архиве СО РАН, а также методы изучения сетей самоцитирования.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 11-07-00388).

Литература

- [1] АКТ ontology description: <http://www.aktors.org/ontology>.
- [2] Alani, H. TGVizTab: An Ontology Visualization Extension for Protege. // Proceedings of Knowledge Capture (K-Cap'03), Workshop on Visualization Information in Knowledge Engineering, Sanibel Island, Florida, USA. 2003.
- [3] Apanovich Z. V., Vinokurov P. S. An extension of a visualization component of ontology based portals with visual analytics facilities. // Bulletin of NCC .— Issue 31.— 2010.— pp. 17-28.
- [4] Bizer, C., Heath, T., Berners-Lee, T. Linked Data - The Story So Far. // Int. J. Semantic Web Inf. Syst., 5 (3). 2009. P. 1-22
- [5] CiteSeer dataset : <http://citeseer.rkbexplorer.com/>.
- [6] Cruz I. F., Stroe C., Caimi F., Fabiani A., Pesquita C., Couto F. M., Palmonari M. Using AgreementMaker to Align Ontologies for OAEI

2011. http://ceur-ws.org/Vol-814/oaiei11_paper1.pdf
- [7] DBLP dataset: <http://dblp.rkbexplorer.com/>.
- [8] Erling O. How Virtuoso uses Relational Technology in its RDF Triple Store and SPARQL implementation. <http://virtuoso.openlinksw.com/whitepapers/SPARQL%20RDF%20Store%20using%20SQL-ORDBMS.html>
- [9] Fruchterman T. M. J., Reingold E. M. Graph Drawing by Force-Directed Placement//Software - Practice and Experience, 1991, Vol. 21, N11, P. 1129-1164.
- [10] Isele R., Jentzsch A., Bizer Ch. Silk Server - Adding missing Links while consuming Linked Data// 1st International Workshop on Consuming Linked Data (COLLD 2010), Shanghai, November 2010.
- [11] Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E. (2007). Ontology Visualization Methods - a Survey. ACM Computing Surveys, 39(4).
- [12] B. Kernighan and S. Lin, An efficient heuristic procedure for partitioning graphs, Bell System Technical Journal, 49 (1970), pp. 291- 307.
- [13] Ngomo A.-C. N., Auer S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. //IJCAI 2011: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011 pp. 2312-2317 .
- [14] Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H. and Tummarello, G.(2008). Sindice.com: a document-oriented lookup index for open linked data.// Int. J. Metadata, Semantics and Ontologies, Vol. 3, No. 1, pp. 37–52 (2008)
- [15] Pietriga, E. IsaViz. <http://www.w3.org/2001/11/IsaViz> .
- [16] Sintek, M. Ontoviz tab: Visualizing Protégé ontologies. 2003 <http://protegewiki.stanford.edu/wiki/OntoViz>.
- [17] Storey, M.-A. D. , Muller, H. A. Manipulating and documenting software structures using shrimp views. // Proc. of the Intl. Conf. on Software Mainten. — 1995.
- [18] Schultz A. et al. How to integrate LINKED DATA into your application //Semantic technology & Business Conference, San Francisco, June 5, 2012. <http://mes-semantic.com/wp-content/uploads/2012/09/Becker-et-al-LDIF-SemTechSanFrancisco.pdf>.
- [19] Tramp S., Williams H., Eck K., Creating Knowledge out of Interlinked Data: The LOD2 Tool Stack <http://lod2.eu/Event/ESWC2012-Tutorial.html>.
- [20] Марчук А.Г., Марчук П.А. Особенности построения цифровых библиотек со связанным контекстом //Труды RCDL'2010- Двенадцатая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» Казань, Казанский университет , 2010. — С. 19-23.
- [21] Апанович З.В., Винокуров П.С., Кислицина Т.А. Гибкая подсистема визуализации онтологии и информационного наполнения порталов знаний на протяжении их жизненного цикла // Труды RCDL'2010 - Двенадцатая Всероссийская научная конференция "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" Казань, Казанский университет, 2010.— С. 265-272.

Problems of using the LOD cloud datasets to enrich the content of scientific data and knowledge bases

Zinaida V. Apanovich, Alexander. G. Marchuk

This paper describes some problems arising during the use of the LOD cloud datasets to enrich the content of scientific knowledge bases and approaches to their solution. The experiments are carried out with the help of a toolkit intended to simplify analysis and integration of data from different datasets. The dataset of the Open Archive of the Russian Academy of Sciences, based on the ONS ontology, as well as various bibliographic datasets , structured by AKT Reference ontology, are used as test examples.