УДК 002.53:004.89

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИИ ДЛЯ АВТОМАТИЗАЦИИ ПОИСКА НАУЧНОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Ахмадеева Ирина Равильевна
Аспирант, Институт систем информатики им. А.П. Ершова,
630090, Российская Федерация, г. Новосибирск, проспект Академика Лаврентьева, 6,
e-mail: i.r.akhmadeeva@iis.nsk.su

Аннотация. В работе предлагается подход к поиску информации на основе онтологии научной деятельности в определенной области знаний. Для поиска используются глобальные поисковые системы, которым отправляются сгенерированные поисковые запросы, включающие названия сущностей онтологии. Запросы в процессе поиска изменяются в зависимости найденных релевантных результатов. Результаты, не содержащие информации о научной деятельности, отфильтровываются с использованием онтологии.

Ключевые слова: онтология, информационный поиск, генерация поисковых запросов, извлечение ключевых слов.

Цитирование: Ахмадеева И.Р. Использование онтологии для автоматизации поиска научной информации в сети Интернет // Информационные и математические технологии в науке и управлении. 2018. № 4 (12). С. 42–49. DOI: 10.25729/2413-0133-2018-4-04

Введение. Для решения проблемы удобного содержательного доступа ученых ко всем интересующим их знаниям и данным была предложена концепция и архитектура тематического интеллектуального научного интернет-ресурса (ИНИР) [2], обеспечивающего доступ к систематизированным научным знаниям и информационным ресурсам определенной области знаний и к средствам их интеллектуальной обработки и анализа.

Основу системы знаний ИНИР составляет онтология [8], которая вводит формальные описания понятий некоторой области знаний, типов информационных ресурсов и методов их интеллектуальной обработки в виде классов и отношений между ними.

Важным этапом построения ИНИР является наполнение его контента актуальной информацией о реальных объектах моделируемой области знаний. Эта задача довольно трудоемкая, а автоматизация в первую очередь требует большой коллекции документов, содержащих информацию об интересующей области знаний. В данной работе предлагается использовать информацию, представленную в онтологии, для автоматизации процесса поиска и накопления релевантных документов из сети Интернет.

Если посмотреть на то, как человек ищет информацию в сети Интернет [4, 5, 13], то можно отметить, что это итеративный процесс, на каждом шаге которого пользователь под влиянием найденной информации может изменить (уточнить) поисковый запрос. При этом результатом поиска не обязательно являются документы, найденные на последней итерации. Обычно пользователь «собирает» информацию по крупицам в течение всего процесса поиска, поэтому для описания поведения пользователя в задаче поиска информации была предложена метафора «сбора ягод» [5]. Эта идея использовалась при разработке

автоматизированной системы поиска информации о научной деятельности в определенной области знаний.

Статья организована следующим образом. В первом разделе статьи приведен краткий обзор исследований, в которых используются онтологии для улучшения процесса поиска информации. Во втором разделе предлагается алгоритм для поиска научных ресурсов, релевантных онтологии. В третьем разделе приводятся результаты эксперимента поиска научных ресурсов в области поддержки принятия решений в слабоформализированных областях.

1. Обзор. Онтологии часто используются в задачах информационного поиска [11, 12, 14]. Обычно они помогают расширить сформулированный пользователем поисковый запрос (например, синонимами) [14]. В данной же работе онтология используются для автоматической генерации поисковых запросов.

В работе [7] предлагается метод пополнения онтологии, с помощью извлечения информации из произвольных веб сайтов. Для поиска релевантных веб сайтов авторы генерируют поисковые запросы для ИПС Google. Для этого они используют экземпляры и отношения онтологии. Сам поисковый запрос генерируется на основании шаблонов, специфичных для каждого отношения, которые должен сформулировать эксперт.

В работе [12] используются Связанные Данные [9], чтобы улучшить традиционный поиск по ключевым словам. Для повышения точности поиска авторы предлагают использовать модель векторного пространства с улучшенным методом вычисления весовых коэффициентов на основе семантических связей между сущностями в документе. В данной работе используется похожий способ вычисления релевантности, однако оценивается релевантность документа онтологии, а не релевантность документа поисковому запросу.

2. Алгоритм поиска научной информации. Чтобы анализировать каждый сайт в Интернете, учитывая, что Интернет постоянно растет и изменяется, нужно иметь огромные вычислительные мощности. Это могут себе позволить немногие компании, поэтому необходимо разрабатывать методы поиска, которые позволяют выбирать из всех ресурсов небольшое подмножество для последующего анализа.

В данной работе предлагается использовать существующие информационно-поисковые системы (ИПС) вместо написания своей. Преимущество использования глобальных ИПС заключается в том, что они индексируют весь Интернет. С другой стороны, базируясь на традиционной модели информационного поиска, они требуют формулирования информационной потребности в виде списка ключевых слов.

Предлагаемый подход к поиску научных ресурсов включает следующие шаги: 1) сгенерировать начальное множество поисковых запросов; 2) отправить поисковые запросы ИПС; 3) оценить релевантность полученных результатов; 4) уточнить поисковые запросы; 5) повторить.

Начальное множество поисковых запросов генерируется с помощью шаблонов на основе онтологии. Чтобы получить множество документов, релевантных построенному поисковому запросу, используется свободная метапоисковая система с открытым исходным кодом Searx [3]. Система Searx позволяет выполнять поиск на различных языках и с помощью различных ИПС, сгруппированных по категориям.

Значения релевантности вычисляются на основе названий понятий и экземпляров онтологии. Чтобы в рабочем множестве не оказалось много похожих поисковых запросов,

расширяются только те запросы, которые позволяют находить новые документы (документы, ранее не известные системе).

2.1. Генерация начального множества поисковых запросов. Начальное множество поисковых запросов строится по шаблонам. Шаблоны поисковых запросов представляют собой набор элементов, ограничений и правил построения запроса. Элементами шаблона могут быть классы, экземпляры, отношения и атрибуты онтологии. Пример шаблона показан на рисунке 1.

При построении поискового запроса по шаблону каждый его элемент связывается с конкретным понятием в онтологии (с учетом ограничений, заданных в шаблоне), после чего формируется список ключевых слов по правилам, указанным в шаблоне.

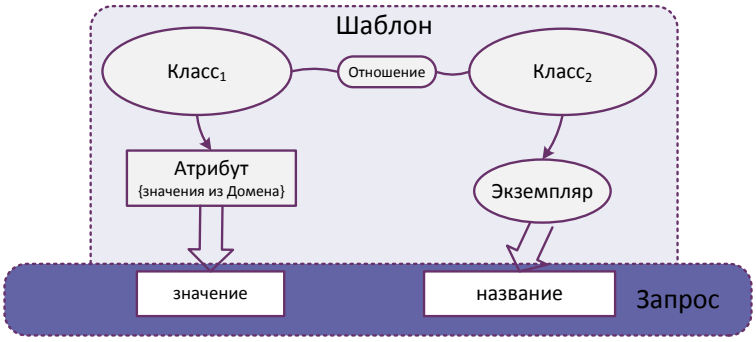

Рис. 1. Пример шаблона поискового запроса

Например, шаблону, представленному на рисунке 1, соответствует фрагмент онтологии, изображенный на рисунке 2, поскольку он удовлетворяет его ограничениям: два класса, связанные отношением, у одного из которых атрибут должен иметь значения из *Домена*, т.е. у такого атрибута ограничена область допустимых значений.

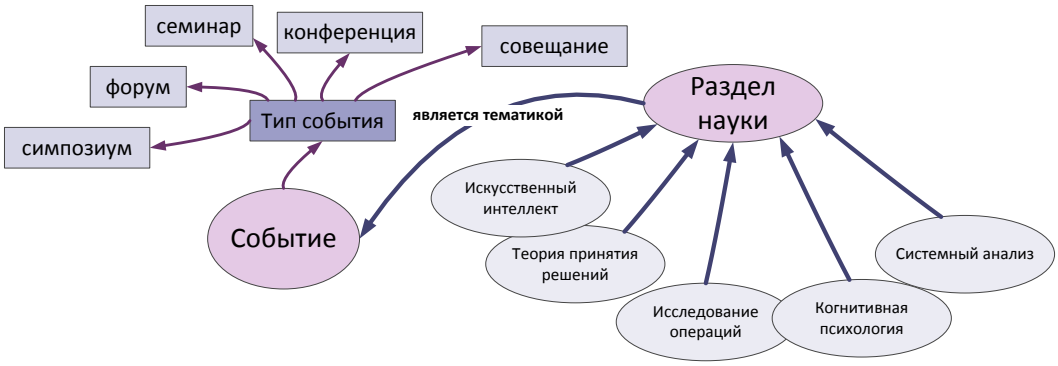

Рис. 2. Фрагмент онтологии, соответствующий шаблону на рисунке 1

В примере таким атрибутом является *Тип события*, который может принимать одно из пяти возможных значений. Таким образом, *Класс*$_1$ в шаблоне соответствует классу *Событие* онтологии, а *Класс*$_2$ классу *Раздел науки*.

Согласно шаблону поискового запроса, приведенному на рисунке 1, в поисковый запрос попадает значение доменного атрибута первого класса и название экземпляра второго класса. Тогда для фрагмента онтологии, представленного на рисунке 2, можно построить следующие поисковые запросы: "конференция Системный анализ", "конференция Искусственный интеллект", "семинар Системный анализ" и др.

2.2. Изменение поискового запроса. Для каждого поискового запроса, строится множество ключевых слов, полученных на основе результатов поиска по этому запросу. Каждому ключевому слову сопоставляется вес – значение релевантности страницы, на которой было найдено это ключевое слово. Затем эти веса нормируются и с соответствующей вероятностью выбираются слова для уточнения поискового запроса.

Поисковый запрос уточняется только в том случае, если доля новых страниц (ранее не известных системе) среди всех найденных с помощью этого поискового запроса результатов превышает определенный порог. Если же в результатах поиска по запросу слишком много страниц, которые были найдены с помощью других поисковых запросов, то такой поисковый запрос считается не перспективным и больше не уточняется.

Для извлечения ключевых слов используется алгоритм TextRank [10]. Для этого сначала строится граф документа, вершинами которого являются нормализованные слова из документа. Между двумя вершинами существует ребро, если соответствующие слова следуют подряд в тексте документа, причем вес этого ребра зависит от того, как часто эти слова стоят рядом. После чего, для нахождения наиболее важных слов, к полученному графу применяется алгоритм PageRank [6].

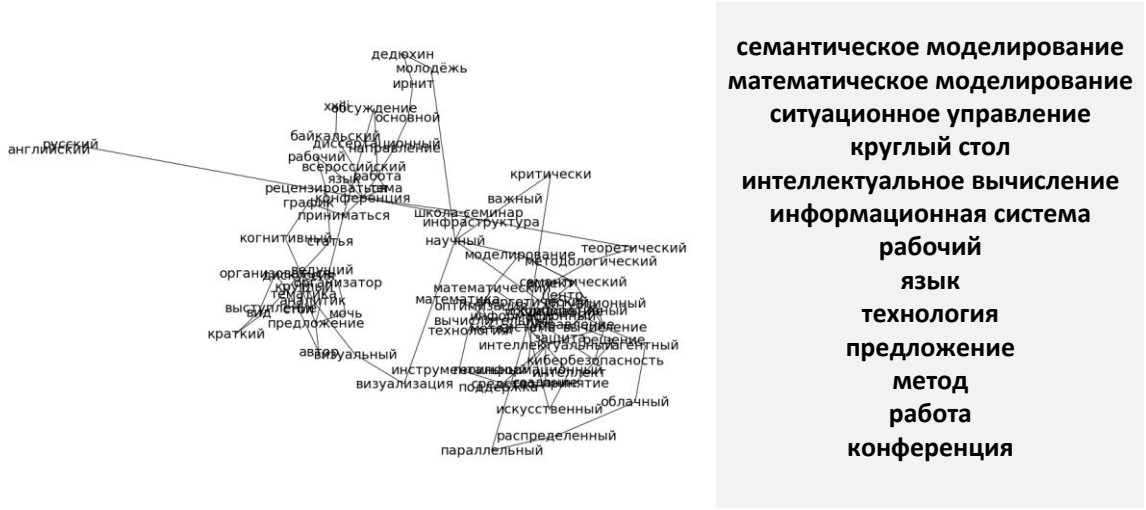

Рис. 3. Извлечение ключевых терминов со страницы конференции ИМТ

Затем, из двадцати наиболее значимых слов оставляются только существительные и прилагательные, которые группируются в словосочетания в зависимости от веса, соединяющего их ребра. Таким образом, получается множество ключевых словосочетаний вида «прилагательное + существительное» и ключевых слов (существительных, которые не вошли ни в одно словосочетание). Пример выделения ключевых слов со страницы конференции ИМТ-2018 изображен на рисунке 3. Здесь, слева представлен построенный по основному тексту страницы граф, а справа – полученное множество ключевых слов и словосочетаний.

2.3. Оценка релевантности. Для оценки релевантности документа и онтологии строятся два бинарных вектора: вектор документа $\vec{d}$ и вектор онтологии $\vec{o}$. Для построения вектора онтологии используются классы, отношения и экземпляры классов онтологии. Для всех перечисленных элементов онтологии значение соответствующей координаты вектора считается равным единице.

Значения координат вектора документа, соответствующих классам и экземплярам онтологии, заполняются на основе вхождения их текстовых меток в документ (с учетом морфологического изменения). Для этого автоматически строятся правила для библиотеки Yargy [1]. Считается, что отношение входит в документ, если в него входят оба его аргумента. Пример построения вектора страницы конференции ИМТ-2018 представлен на рисунке 4.

Рис. 4. Пример построения вектора документа (страница конференции ИМТ)

Далее, для вычисления самого значения релевантности используется косинусная мера (1). Пороговое значение подбиралось экспериментальным путем на основе точности поиска.

$$similarity_c = \frac{\vec{o} \cdot \vec{d}}{\|\vec{o}\| \cdot \|\vec{d}\|} = \frac{\sum_{i=1}^{n} o_i \times d_i}{\sqrt{\sum_{i=1}^{n} o_i^2} \times \sqrt{\sum_{i=1}^{n} d_i^2}} \quad (1)$$

3. Экспериментальная часть. Оценка качества поиска научных ресурсов проводилась на основе онтологии по поддержке принятия решений в слабоформализованных областях. В ходе экспериментальной проверки был сгенерирован 221 поисковый запрос и найдено 903 релевантных страниц. Начальное множество запросов было сформировано из названий экземпляров класса «Раздел науки»: “теория принятия решений”, “искусственный интеллект”, “системный анализ”, “когнитивная психология”, “исследование операций”, “онтологический инжиниринг”.

Точность отдельно оценивалась для трех значений порога релевантности: 0.1, 0.15 и 0.2, по выборке в 100 документов. Полученные значения представлены на рисунке 5.

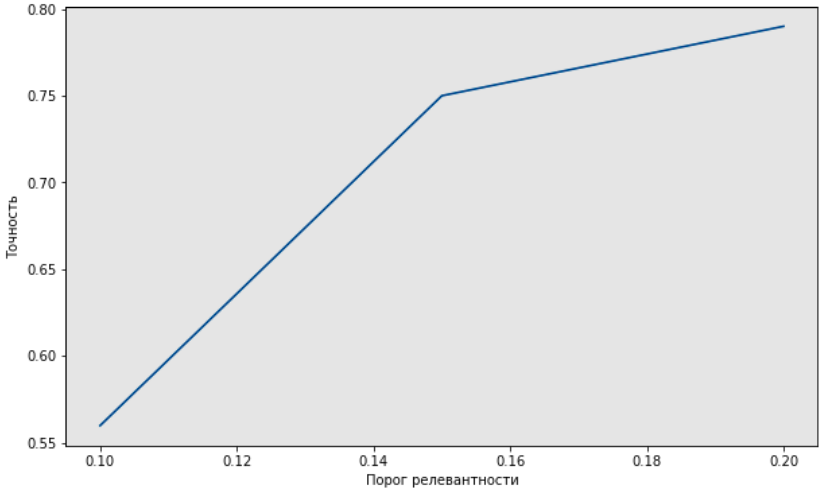

Рис. 5. Точность на выборке при различных значениях порога релевантности

Анализ полученных результатов показал два основных источника ошибок. Во-первых, построение характеристического вектора документа (см. раздел 2.3) не учитывает длину этого документа, только вхождение определенных словосочетаний в документ, что влечет не всегда адекватную оценку значения релевантности для очень длинных документов, в которых, например, названия классов онтологии встречаются достаточно далеко друг от друга и не связаны между собой.

Во-вторых, в онтологии встречаются общие понятия, не имеющие прямого отношения к теории принятия решений, которые искажают оценку релевантности. Примером таких понятий могут быть географические места, в которых расположены научные заведения или проходят конференции.

Заключение. В данной работе предлагается итеративный подход к поиску научной информации в определенной области знаний, который использует онтологию для построения поисковых запросов и оценки релевантности найденных ресурсов. Для поиска используются глобальные поисковые системы, которым отправляются сгенерированные поисковые запросы, включающие названия сущностей онтологии. Уточнение запросов для следующей итерации поиска происходит с помощью выделения ключевых слов. Экспериментальная оценка показала работоспособность предложенного подхода.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 16-07-00569).

СПИСОК ЛИТЕРАТУРЫ

1. Документация библиотеки Yargy. Режим доступа: http://yargy.readthedocs.io/ru/latest/ (дата обращения 11.07.2018).
2. Загорулько Ю.А., Загорулько Г.Б., Боровикова О.И. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. 2016. № 2. С. 51-60.
3. Метапоисковая система Searx. Режим доступа: http://asciimoo.github.io/searx/ (дата обращения 11.07.2018).

4. Aula A., Khan R. M., Guan, Z. How does search behavior change as search becomes more difficult? // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. 2010. Pp. 35–44.
5. Bates M. J. The design of browsing and berrypicking: Techniques for the online search interface // Online Information Review. 1989. № 13(5). Pp. 407–424.
6. Brin S., Page L. The anatomy of a large-scale hypertextual web search engine // Computer networks and ISDN systems. 1998. Т. 30. №. 1-7. Pp. 107–117.
7. Geleijnse G., Korst J. H. M. Automatic Ontology Population by Googling // Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC). Belgum: Koninklijke Vlaamse Academie van Belie voor Wetenschappen en Kunsten. 2005. Pp. 120–126.
8. Guarino N. Formal Ontology in Information Systems // Proceedings of FOIS'98. Amsterdam: IOS Press. 1998. Pp. 3–15.
9. Linked Data - Connect Distributed Data across the Web. Режим доступа: http://linkeddata.org/ (last accessed 11.07.2018).
10. Mihalcea R., Tarau P. Textrank: Bringing order into text // Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
11. Vallet D., Fernández M., Castells P. An ontology-based information retrieval model //European Semantic Web Conference. Springer, Berlin, Heidelberg. 2005. Pp. 455–470.
12. Waitelonis J., Exeler C., Sack H. Linked data enabled generalized vector space model to improve document retrieval //Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC). CEUR-WS. – 2015. – Т. 1486.
13. White R.W., Roth R.A. Exploratory Search: Beyond the Query-Response Paradigm // Synthesis lectures on information concepts, retrieval, and services / CA: Morgan and Claypool. 2009. № 1(1). Pp. 1–98.
14. Xiong C., Callan J. Query expansion with Freebase // Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ACM. 2015. Pp. 111–120.

UDK 002.53:004.89

USING ONTOLOGY FOR AUTOMATIZATION OF SCIENTIFIC INFORMATION RETRIEVAL IN THE INTERNET

Irina R. Akhmadeeva
Graduate student, A.P. Ershov Institute of Informatics Systems
6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia, e-mail: i.r.akhmadeeva@iis.nsk.su

Abstract. The paper suggests an approach to information retrieval based on the ontology of scientific activity in a certain area of knowledge. This method makes use of general-purpose search engines to retrieve links to relevant Internet resources using the search queries generated on the basis of the ontology concepts. In the search process queries change depending on the number of retrieved results found. Search results that do not contain information about scientific activity are filtered using ontology.

Keywords: ontology, information retrieval, search query generation, keywords extraction.

References

1. Dokumentatsiya biblioteki Yargy [Documentation of the Yargy library]. Available at: http://yargy.readthedocs.io/ru/latest/, accessed 11.07.2018. (in Russian)
2. Zagorulko Yu.A., Zagorulko G.B., Borovikova O.I. Tehnologija sozdanija tematicheskih intellektual'nyh nauchnyh internet-resursov, bazirujushhajasja na ontologii [Technology for building subject-based intelligent scientific internet resources based on ontology] // Programmnaja inzhenerija = Software Engineering. 2016. no. 2. Pp. 51–60. (in Russian)
3. Metapoiskovaya sistema Searx [Metasearch engine Searx]. Available at: http://asciimoo.github.io/searx/ , accessed 11.07.2018. (in Russian)
4. Aula A., Khan R. M., Guan, Z. How does search behavior change as search becomes more difficult? // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. 2010. Pp. 35–44.
5. Bates M. J. The design of browsing and berrypicking: Techniques for the online search interface // Online Information Review. 1989. № 13(5). Pp. 407–424.
6. Brin S., Page L. The anatomy of a large-scale hypertextual web search engine // Computer networks and ISDN systems. 1998. Т. 30. №. 1-7. Pp. 107–117.
7. Geleijnse G., Korst J. H. M. Automatic Ontology Population by Googling // Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC). Belgum: Koninklijke Vlaamse Academie van Belie voor Wetenschappen en Kunsten. 2005. Pp. 120–126.
8. Guarino N. Formal Ontology in Information Systems // Proceedings of FOIS'98. Amsterdam: IOS Press. 1998. Pp. 3–15.
9. Linked Data - Connect Distributed Data across the Web. Режим доступа: http://linkeddata.org/ (last accessed 11.07.2018).
10. Mihalcea R., Tarau P. Textrank: Bringing order into text // Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
11. Vallet D., Fernández M., Castells P. An ontology-based information retrieval model //European Semantic Web Conference. Springer, Berlin, Heidelberg. 2005. Pp. 455–470.
12. Waitelonis J., Exeler C., Sack H. Linked data enabled generalized vector space model to improve document retrieval //Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC). CEUR-WS. – 2015. – Т. 1486.
13. White R.W., Roth R.A. Exploratory Search: Beyond the Query-Response Paradigm // Synthesis lectures on information concepts, retrieval, and services / CA: Morgan and Claypool. 2009. № 1(1). Pp. 1–98.
14. Xiong C., Callan J. Query expansion with Freebase // Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ACM. 2015. Pp. 111–120.