

УДК 002.53:004.89

Метод поиска информации на основе онтологии

Ахмадеева И.Р. (Институт систем информатики СО РАН)

В статье предлагается метод поиска информации на основе онтологии научной деятельности. Для поиска используются глобальные поисковые системы, которым отправляются автоматически сгенерированные поисковые запросы, включающие названия сущностей онтологии и термины тезауруса. Поисковые запросы формируются таким образом, чтобы найти как можно больше научных ресурсов, релевантных определенной области знаний. При этом результаты поиска, не содержащие информации о научной деятельности, отфильтровываются с использованием онтологии.

Ключевые слова: *онтология, тезаурус, информационный поиск, поисковый запрос*

1. Введение

Несмотря на то, что в настоящее время накоплены огромные объемы информации по различным областям научных знаний, причем значительная ее часть представлена в сети Интернет, ученые пока не имеют удобного содержательного доступа ко всем интересующим их знаниям и данным, в той области, в которой они проводят исследования.

Для решения этой проблемы была предложена концепция и архитектура тематического интеллектуального научного интернет-ресурса (ИНИР) [1], обеспечивающего доступ к систематизированным научным знаниям и информационным ресурсам определенной области знаний и к средствам их интеллектуальной обработки и анализа. Информация в ИНИР доступна ученому в виде сети знаний и данных, как наиболее естественной и удобной форме подачи информации для человека.

Основу системы знаний ИНИР составляет онтология [6], 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, интегрируемых информационных ресурсах и методах и средствах их обработки и анализа. Эта задача довольно трудоемкая и решить ее можно только за счет автоматизации сбора и накопления релевантной информации из сети Интернет.

В данной работе анализируются различные модели информационного поиска, делается обзор исследований, изучающих поведение пользователей в процессе поиска информации, а также предлагается подход к автоматизации поиска ресурсов в Интернете для сбора информации о научной деятельности в определенной области знаний.

Благодарности. Работа поддержана Российским Фондом Фундаментальных Исследований (грант №16-07-00569).

2. Особенности поведения пользователей в процессе поиска информации

Большинство информационно-поисковых систем (ИПС) работают в соответствии с традиционной моделью информационного поиска: Google, Yandex, Bing, Yahoo!. Такие системы лучше всего подходят для задач поиска фактической информации, когда известна цель поиска. Традиционная модель информационного поиска состоит из четырех компонентов:

- коллекция документов;
- индекс документов (обычно инвертированный для быстрого поиска);
- информационная потребность пользователя;
- поисковый запрос, сформулированный пользователем.

В такой модели предполагается, что пользователь может выразить свою информационную потребность в виде списка ключевых слов. Обычно это можно сделать при решении задач поиска фактов, навигации, ответов на вопросы. В работе [9] такой поиск называется простым. Ему противопоставляется разведывающий поиск (exploratory search), который используется в задачах получения, интерпретации, интеграции, анализа, синтеза знаний и т.д. Для разведывающего поиска характерны следующие особенности [14]:

- пользователь не знаком с областью, которая его интересует (т.е. он хочет получить некоторую информацию об этой области, чтобы понять, как достичь своей цели),
- пользователь не уверен в способах достижения своей цели,
- пользователь не уверен в своей цели, не знает, что ищет.

В работе [3] была предложена метафора «сбора ягод», которая ближе к реальному поведению людей ищущих информацию, чем традиционная модель информационного поиска. В данной модели поиск рассматривается как итеративный процесс, в котором пользователь в самом начале знает совсем немного об интересующей его области и постепенно, получив новую крупную информацию, переформулирует запрос с учетом нового

знания. И таким образом шаг за шагом в процессе эволюционирующего поиска пользователь «собирает» информацию. Причем в процессе такого поиска может измениться как направление поиска, так и сама цель.

Многие исследователи изучали поведение пользователей в процессе поиска информации. В работе [15] выделяются две стратегии поведения пользователей: исследователи и навигаторы. Для навигаторов характерна последовательность в поведении, а для исследователей – изменчивость. Авторы предполагают, что навигаторы решают простую задачу нахождения фактов, а исследователи – более сложную задачу определения смысла.

В работе [2] анализируется поведение пользователей, когда перед ними стоит сложная задача поиска, и выделяются следующие особенности:

- пользователи формулируют больше поисковых запросов,
- они чаще используют специальные операторы поисковых запросов,
- они тратят больше времени на странице с результатами запроса,
- они формулируют самый длинный запрос в середине поисковой сессии.

Часто в исследованиях опытных пользователей сравнивают с новичками и выделяют следующие отличия: опытные пользователи склонны тратить меньше времени на поисковые задачи [11], реже переформулируют запросы [7], используют более длинные запросы [7], используют более систематическую стратегию уточнения запроса [4].

Таким образом, решение сложной задачи поиска является итеративным процессом, в процессе которого пользователь уточняет свой поисковый запрос в зависимости от полученных результатов. Эта идея использовалась при разработке автоматизированной системы поиска информации о научной деятельности в определенной области знаний.

3. Подход к автоматизации поиска 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 статье предлагается метод поиска информационных ресурсов, который использует глобальные поисковые системы (метапоиск [10]), онтологию и тезаурус для генерации поискового запроса и оценки релевантности найденных информационных ресурсов тематике ИНИР. Онтологии и тезаурусы часто используются в задачах информационного поиска [12,13,16]. Обычно они помогают расширить поисковый запрос (например, синонимами), который запросил пользователь [16]. В данной же работе

онтология и тезаурус используются для автоматической генерации поисковых запросов. Понятиям онтологии сопоставляются термины тезауруса, с помощью которых их можно выразить на естественном языке. Соответствующие термины тезауруса используются при построении поисковых запросов и оценке релевантности.

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 запроса в текстовом виде. В данном случае информационная потребность подсистемы сбора информации состоит в необходимости получить информацию о научной деятельности в определенной области знаний, еще не представленную в контенте ИНИР.

Процесс поиска в сети Интернет научных ресурсов включает следующие этапы:

- Генерацию поисковых запросов;
- Отправку поисковых запросов ИПС;
- Выбор из списка результатов релевантных ресурсов, и сохранение их в базе данных;
- Анализ ссылок в релевантных ресурсах;
- Уточнение поисковых запросов на основе полученных результатов;

Можно выделить несколько видов информационных потребностей, возникающих в процессе поиска научных интернет-ресурсов.

Во-первых, необходимо искать уже известные факты. Такая необходимость возникает в случаях, когда необходимо проверить на корректность только что найденную информацию либо информацию, уже содержащуюся в контенте ИНИР. Кроме этого, такие ресурсы могут ссылаться на другие ресурсы, потенциально содержащие релевантную информацию. Корректность найденной информации предлагается проверять аналогично работе [5] по количеству найденных результатов на соответствующий запрос.

Во-вторых, информация о некоторых экземплярах онтологии может быть неполной. Например, значения атрибутов экземпляра понятия или его связи с экземплярами других понятий могут отсутствовать в контенте ИНИР.

В-третьих, не все экземпляры классов, описанных в онтологии, представлены в контенте ИНИР. В данном случае требуется найти информацию о таких экземплярах.

Для каждого из этих случаев строятся свои наборы шаблонов генерации запросов. 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.

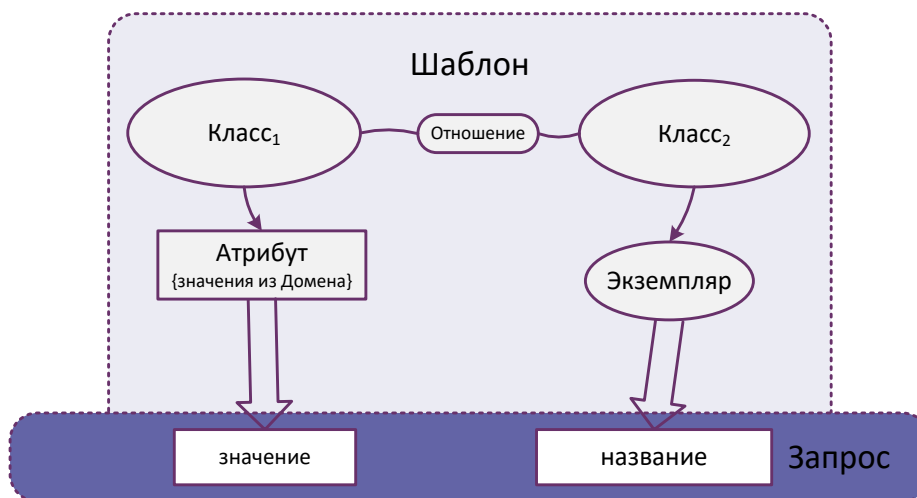


Рис. 1. Пример шаблона поискового запроса.

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. Причем, информация о всех возможных значениях атрибута хранится в онтологии и может быть использована для поиска экземпляров этого класса.

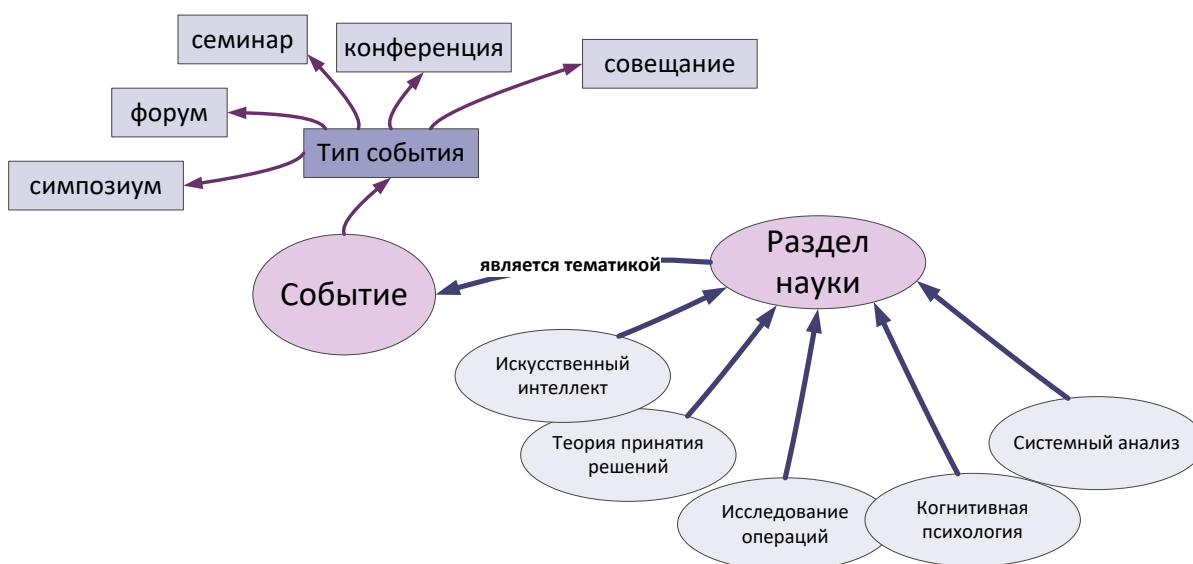


Рис. 2. Фрагмент онтологии, соответствующий шаблону на рисунке 1.

В примере таким атрибутом является *Тип события*, который может принимать одно из пяти возможных значений. Таким образом, *Класс₁* в шаблоне соответствует классу *Событие* онтологии, а *Класс₂* 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 поисковые запросы, представленные в таблице 1.

Таблица 1. Пример сгенерированных поисковых запросов

конференция Системный анализ
конференция Когнитивная психология
конференция Искусственный интеллект
конференция Теория принятия решений
конференция Исследование операций
семинар Системный анализ
семинар Когнитивная психология
...

Чтобы получить множество ссылок на веб-страницы, релевантные построенному поисковому запросу, используется свободная метапоисковая система с открытым исходным кодом Searx¹, которая соответствует спецификациям OpenSearch². Система Searx позволяет выполнять поиск на различных языках и с помощью различных поисковых систем, сгруппированных по категориям.

Далее для каждого найденного документа (веб-страницы) оценивается его релевантность, на основании которой принимается решение о сохранении ее в базу и дальнейшем анализе. Релевантность оценивается в два этапа:

- В первую очередь нужно понять, посвящен ли найденный документ научной деятельности в определенной области знаний. Для этого сначала вычисляется релевантность данного документа онтологии, лежащей в основе ИНИР.
- Затем нужно определить конкретный класс (классы) онтологии, которому посвящен документ.

Для оценки релевантности документа классу онтологии используется векторная модель [8], в которой вектор документа включает абсолютные частоты встречаемости терминов

¹ <http://asciimoo.github.io/searx/>

² <http://www.opensearch.org/Specifications/OpenSearch/1.1>

(слов) за исключением стоп-слов, т.е. слов, не несущих смысловой нагрузки (предлогов, общеупотребимых слов и т.п.). Для учета встречаемости терминов в разных морфологических формах используются их основы (стемминг).

Векторное представление класса онтологии строится по формуле (1) на основе описания этого класса в онтологии: его атрибутов, связей с другими классами и соответствующих терминов тезауруса. Вспомогательный вектор \vec{c}_{attr} включает :

- абсолютные частоты встречаемости терминов, входящих в название этого класса, и его атрибутов;
- абсолютные частоты терминов, входящих в допустимые значения доменных атрибутов.
- абсолютные частоты терминов тезауруса, связанных с этим классом.

$$\vec{c} = \vec{c}_{attr} + \gamma \sum \vec{c}'_{attr} \quad (1)$$

Где \vec{c} – вектор класса онтологии, \vec{c}_{attr} – вспомогательный вектор, учитывающий описание класса без его связей, γ – коэффициент, показывающий вклад связанных (отношением в онтологии) классов в векторное представление данного класса, \vec{c}'_{attr} – вспомогательные вектора классов онтологии, связанных с данным классом.

Значение релевантности вычисляется с помощью косинусной меры между векторами класса онтологии и документа по формуле (2).

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

Где \vec{c} – вектор класса онтологии, \vec{d} – вектор документа, n – длина векторов (число учитываемых терминов), а i – позиция термина в векторе.

Релевантность документа онтологии вычисляется аналогичным образом: вектор документа строится также, как и в предыдущем случае, а вот вектор онтологии строится аналогично вектору класса, только теперь учитываются все сущности онтологии. Классом онтологии, которому посвящен документ, считается класс с максимальным значением $similarity_c$.

Для уточнения запроса предполагается выделять ключевые слова из найденных документов с помощью статистических методов и в зависимости от значения релевантности добавлять эти слова в поисковый запрос вместе с различными операторами на языке поисковых запросов, которые поддерживает метапоисковая система *Searx*. Так, если

документ оказался не релевантным онтологии, в поисковый запрос добавляются ключевые слова из этого документа с оператором «-», исключающим результаты, содержащие эти слова.

Общий алгоритм поиска научных ресурсов, релевантных определенной области знаний, представлен на рисунке 3. Поиск запускается в соответствии с установленным расписанием. В начале каждого сеанса на основе *шаблонов поисковых запросов* генерируются запросы и добавляются в *очередь поисковых запросов*. Каждый поисковый запрос из очереди отправляется в метапоисковую систему *Searx*, после чего найденные с ее помощью ссылки на страницы добавляются в *очередь страниц на обработку*.

Затем для каждой страницы вычисляется релевантность и если она выше определенного порога, то ссылка на страницу сохраняется для дальнейшей обработки. Далее анализируются ссылки на этой странице и добавляются в *очередь страниц на обработку*. Кроме этого для каждой страницы уточняется поисковый запрос (в зависимости от значения релевантности) и добавляется в *очередь поисковых запросов*.

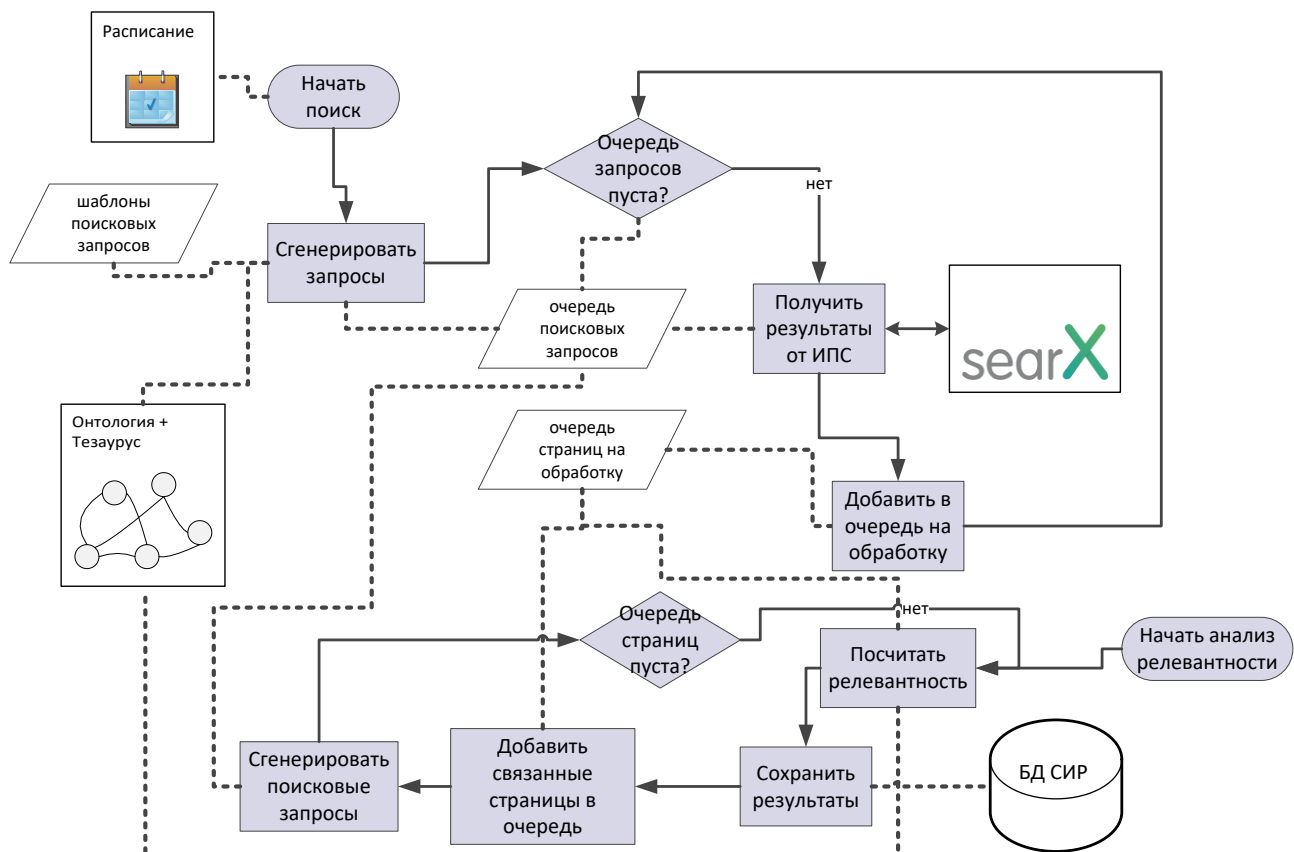


Рис. 3. Алгоритма поиска релевантных ресурсов.

4. Заключение

В статье предлагается метод поиска научных интернет-ресурсов в определенной области знаний, использующий 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 и термины тезауруса.

В ходе дальнейшей работы предполагается доработать метод уточнения поисковых запросов и применить предложенный подход для пополнения онтологии ресурса по поддержке принятия решений в слабоформализованных областях.

Список литературы

1. Загоруйко Ю. А., Загоруйко Г. Б., Боровикова О. И. Технология создания тематических интеллектуальных научных Интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. 2016. Т. 7, № 2. С. 51-60.
2. 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. New York: ACM, 2010. P. 35-44
3. Bates M. J. The design of browsing and berrypicking techniques for the online search interface // Online review. 1989. Vo.13(5). P. 407-424.
4. Fields B., Keith S., Blandford A. Designing for expert information finding strategies //People and computers XVIII—Design for life. London: Springer, 2005. P. 89-102.
5. Geleijnse G., Korst J. H. M. Automatic Ontology Population by Googling // Proceedings of the Seventeenth Belgium-Netherlands Conference on Artificial Intelligence. Koninklijke Vlaamse Academie van Belie voor Wetenschappen en Kunsten, 2005. P. 120-126.
6. Guarino N. Formal Ontology in Information Systems // Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, June 6–8, 1998 / Ed. N. Guarino. Amsterdam: IOS Press, 1998. P. 3-15.
7. Hölscher C., Strube G. Web search behavior of Internet experts and newbies //Computer networks. 2000. Vol. 33. P. 337-346.
8. Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. Online edition. Cambridge University Press. 2009. 544 pp.
9. Marchionini G. Exploratory search: from finding to understanding //Communications of the ACM. 2006. Vol. 49(4). P. 41-46.
10. Meng W., Yu C., Liu K. L. Building Efficient and Effective Metasearch Engines // ACM Computing Surveys (CSUR). 2002. Vol. 34. No. 1. P. 48–89.

11. Saito H., Miwa K. A cognitive study of information seeking processes in the WWW: the effects of searcher's knowledge and experience // Proceedings of the Second International Conference on Web Information Systems Engineering. IEEE, 2001. Vol. 1. P. 321-327.
12. Vallet D., Fernández M., Castells P. An ontology-based information retrieval model //European Semantic Web Conference. – Springer, Berlin, Heidelberg, 2005. – С. 455-470.
13. 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.
14. White R. W. et al. Supporting exploratory search, introduction, special issue, communications of the ACM //Communications of the ACM. 2006. Vol. 49(4). P. 36-39.
15. White R. W., Drucker S. M. Investigating behavioral variability in web search //Proceedings of the 16th international conference on World Wide Web. ACM, 2007. P. 21-30.
16. Xiong C., Callan J. Query expansion with Freebase //Proceedings of the 2015 International Conference on The Theory of Information Retrieval. – ACM, 2015. – С. 111-120.