# Использование онтологического подхода к разработке каталога пользовательских предпочтений

© Б.Г. Циркин

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

bgdotmail@gmail.com

**Аннотация**

Описан подход к организации прототипа системы каталогизации пользовательских предпочтений, представленного в виде интернет-портала. Этот портал решает задачу упорядочивания введённой пользователями информации на основе ключевых слов, которые её характеризуют. Предполагается, что данные на портале представлены в виде гиперссылок на некоторые информационные объекты, снабженные набором тегов. При этом тегом считается не только визуальное представление, но и его окружение (контекст). Такой портал имеет некоторую базовую онтологию и позволяет каждому пользователю уточнять её, создавая на ее основе свою онтологию. Рассматривается алгоритм работы интеллектуального агента, который на основе уже внесенных пользователем данных (как объектов, так и набора характеризующих их тегов) будет предлагать пополнение в рамках каждой заданной предметной области из данных, которые внесли другие пользователи.

## 1 Обзор проблемной области

С течением времени особую остроту приобретает проблема поиска релевантной информации. Любой процесс ныне сложно представить без упорядоченной системы знаний, несмотря на то, что накопленных данных становится всё больше. В связи с этим особую остроту приобретает задача создания системы взаимоупорядоченных и взаимовлияющих друг на друга элементов, отражающих видение мира на языке данной области познания.

Это происходит из-за того, что в представлении данных на ресурсах очень часто отсутствует система, они слабо структурированы по интернет-сайтам, электронным библиотекам, архивам, что существенно ограничивает к ним доступ. Более того, по историческим, техническим и другим причинам тематически связанные данные сохраняются в разных форматах под управлением различных систем хранения и обработки данных.

Такое положение дел приводит к тому, что разнообразные коллекции, базы персоналий и публикаций, даже расположенные на одном физическом сервере, зачастую имеют различные логические входы и представляют собой разрозненные автономные информационных ресурсы.

Картина значительно усугубляется следующими факторами: неудобочитаемость; дублирование и избыточность информации; неопределенность, неоднозначность и многозначность в определениях понятий; встречающиеся противоречия.

Отсутствие связанности информационных ресурсов и унифицированного доступа к ним приводят к неполноте рассмотрения и учета существующих данных и знаний при решении возникающих задач. Возможность получения необходимой информации также недостаточно высока из-за отсутствия содержательного доступа к накопленным информационным ресурсам и знаниям.

Потребность в поиске, выявлении необходимых данных и организации к ним доступа невероятно велика. Существует огромное количество различных подходов к их организации с целью облегчения различной деятельности пользователей интернета (прежде всего поисковой). Невозможно представить себе пользователя интернета, который не использует поисковые машины, автоматически индексирующие содержание веб-сайтов, такие, как *Alta Vista*, *Google*, *Excite* и др. Они обеспечивают поиск по любому слову из текста, обнаруженного на сайте. Кроме того, некоторую распространённость получили также и интернет-справочники *Yahoo*, *LookSmart*, *About* и другие. Они представляют подход, предполагающий вовлечение человеческого интеллекта в процессы отбора и аннотирования вебсайтов. Такие каталоги составляются вручную и поэтому требуют огромного времени как для создания, так и для сопровождения. Кроме этого, могут существовать различия между критериями класси-

**Труды 12[ii] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010**

фикации понятий автора и пользователей, а также в их восприятии.

В данной работе рассматривается проблема каталогизации уже найденных ресурсов, что должно позволить ускорить повторный поиск данных. В идеальной ситуации система должна обеспечивать быстрый и информативный переход к релевантной информации.

## 2 Обзор существующих подходов

### 2.1 Подход сетевых библиотек

Одним из подходов к решению поставленной задачи организации данных является применение традиционных принципов каталогизации для описания материалов интернета и предоставление к ним доступа через онлайн-каталоги библиотек.

Данное направление активно развивалось и до сих пор развивается как в библиотеках США, так и в российском сегменте интернета (см. eLIBRARY.RU). Многие библиотеки являются инициаторами и участницами интересных проектов в этой новой области каталогизации [1].

Существует также и альтернативный подход – расширение стандартного HTML специальными семантическими тэгами для внесения знаний прямо в страницы. Такие документы несут информацию о взаимосвязях понятий и их семантических атрибутах в HTML-подобном формате, то есть не требуют внутреннего языка представления знаний.

Это решение воплощено в рамках стандарта языка XML. Этот язык предназначается для разметки синтаксической структуры документов, облегчающий использование таких документов в качестве сообщений при общении множества агентов.

Для аннотаций документов с помощью XML разработан формат описания ресурсов RDF. Метаинформация, определяемая этим форматом, нередко размещается как некоторый блок внутри каждой страницы (аннотация каждого элемента страницы непосредственно в тексте исходного документа невозможна, что приводит к их повторению с дополнительной метаинформацией). Этот способ влечет за собой многократное увеличение объемов информации.

Предлагалось и создание всемирной базы данных документов, которые, в свою очередь, могли бы включать в себя части объектов интернета и других документов этой базы данных с аннотациями к ним, написанными на специальном языке. Однако развития эта идея не получила, прежде всего, потому, что потребовался бы гигантский даже по нынешним временам объем предполагаемого хранилища.

### 2.2 Социальные сети закладок

Существует большое количество попыток построения каталогов ссылок, но их использование затруднено. Предпринятые на данный момент попытки (http://del.icio.us, http://socialpage.ru, http://linkomatic.ru) страдают слабостью взаимосвязей между блоками целого. Эти ресурсы концентрируют внимание на наборе ключевых выражений (или тегов), заданных пользователями [2], что позволяет строить облако понятий (*совокупность тегов, характеризующих некоторый информационный объект*), связанных с тем или иным объектом. В некоторых случаях теги можно собирать в группы, что, безусловно, упрощает решение части задач.

Однако данный подход содержит в себе немало проблем. Прежде всего, не учитывается морфология, что затрудняет обнаружение связей между сходными объектами. Кроме того, теги, являющиеся многозначными понятиями, затрудняют навигацию, приводя к более широкому, чем ожидается, результату. В результате такие ресурсы используются в основном лишь как средство хранения и, гораздо реже, обмена данными между пользователями интернета.

### 2.3 Другие подходы

Существуют также подходы к решению поставленной задачи, которые позволяют задавать информационному объекту требуемые атрибуты сверх уже зафиксированных. Однако в подобной системе такие атрибуты оказывают влияние лишь на представление пользовательских данных в рамках одной конкретной страницы (или набора страниц), но не на их каталогизацию. Интеллектуальное пополнение на данный момент не использует данную технологию.

Кроме того, с участием автора разработан [3] настраиваемый портал научных знаний, однако этот портал не имеет личной онтологии, что уменьшает его гибкость.

## 3 Подход к решению

### 3.1 Общая схема

В качестве первой итерации построения систематизированной базы знаний взята система элементарного библиотечного каталога с тематическим рубрикатором (упорядоченная система «карточек», соответствующих конкретным информационным объектам). При этом различные варианты тегов, т. е. *ключевых выражений, описывающих некоторый информационный объект*, будут составлять набор данных для заполнения своеобразных «карточек».

Такие «карточки» обеспечат прозрачный и понятный переход к данным, прежде всего, другим ссылкам, релевантным уже выбранной ссылке. Иными словами, ключевые выражения, характеризующие конкретные информационные объекты, позволят осуществлять переход между присутствующими в системе данными, соответствующими этим информационным объектам в рамках информационного значения тега [4]. Это поможет пониманию не только содержимого информационного объекта, но и его непосредственного окружения

через совокупность ключевых выражений, которые описывают этот объект и подобные ему объекты.

Для обеспечения единообразного представления данных и учета связанности информации в рамках единой тематики необходима единая концептуальная схема информационного содержания портала – адекватная информационная модель портала, которая должна решать следующие основные задачи:

- организацию предметного каталога;
- наличие прозрачной системы взаимосвязей между тегами; организация «личной онтологии» (такая онтология, вероятно, будет отличаться от общей онтологии и использоваться как ее дополнение); примерное соотношение между ними показано на рис. 1;
- выбор способа пополнения онтологии; это будут динамическое объединение или интеграция этих двух онтологий, возможно, также потребуется отображение одной онтологии в другую.

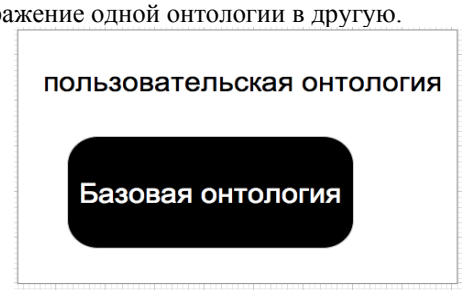


Рис. 1. Соотношение базовой и пользовательской онтологий

В результате построения онтологии на ключевых словах вероятны упрощение пользовательской навигации, а также более четкое очерчивание проблемной области. Данный подход позволит достаточно просто облегчить решение второй и третьей задач.

Скорее всего, в рамках решения первой задачи потребуется построение мета-онтологии, вернее, решением первой задачи и будет являться такая онтология.

Использование подхода «социальной сети» позволит решить как проблему наполнения данными, так и задачу расширения как мета-онтологии, так и онтологии в целом.

Важными, но не первоочередными проблемами являются:

- выбор подходящих критериев «схожести» для понятий;
- проблема самообучаемости системы или интеллектуального ассистента в построении связей между ссылками и возможном добавлении тегов к ним;
- решение задачи влияния «жаргона» на релевантность поиска в силу превалирования над первоначальным значением.

### 3.2 Ранжирование пользовательских онтологий

Решение поставленной выше задачи представляется неполным без эффективного взаимодействия между пользователями каталога. Каждый из них должен иметь возможность не только хранить некоторый набор аннотаций: набора ключевых понятий в рамках пользовательской онтологии, а также связей, соответствующих соотношениям между понятиями с точки зрения онтологии в отношении каждого объекта, включенного в каталог.

Представляется естественным для простоты рассмотрения принять в качестве такого объекта гиперссылку, которая, очевидно, может быть основой соответствия объекта и его представления в каталоге.

Подобный каталог, естественно, требует решения задачи пополнения данными каталога каждого конкретного пользователя, основываясь на его представлении мира (онтологии), ведь представление знаний уже достаточно давно рассматривается как удачный инструмент для организации и обмена информацией. Из-за сложности и достаточно высокой временной стоимости построения онтологий довольно частым видится следующий образ поведения типичного пользователя и системы:

- оформление сферы интересов через построение или уточнение (с точки зрения пользователя) существующей онтологии;
- обнаружение определенных системой корреляций между частями пользовательской онтологии и частями личных онтологий иных пользователей системы;
- анализ пользователем степени полезности предлагаемых данных с возможным ранжированием по релевантности/полноте/доступности изложения и т. п.;
- анализ системой полученных на предыдущем шаге от пользователя оценок и ранжирование в связи с полученными результатами пользователей по степени схожести взглядов в конкретной предметной области.

В результате крайне необходимо иметь способ измерения различия между онтологиями схожих предметных областей, ведь ранжирование всегда лежало в самом сердце получения информации. Это стало особенно заметно с расширением интернета, когда Google стал использовать PageRank, основанный на анализе гиперссылок. Однако сложно представить себе настолько идеальный случай, что пользовательские онтологии связаны между собой в части, отличающейся от базовой онтологии, а такая ситуация делает PageRank или его аналоги бесполезными.

В данном случае необходимо, чтобы система могла измерить различие между онтологиями различных пользователей и в случае достаточно небольших различий в какой-то части выделить связи и данные для пополнения.

Для определения полноты онтологии O относительно онтологии O' (рис. 2) можно использовать методику, предложенную в [5]. Результатом может являться набор ключевых слов или гиперссылок для пополнения, а методика, предложенная в [6, 7], позволяет проводить сравнения онтологий как на син-

таксическом (с использованием меры редактирования Дамерау – Левенштейна), так и на семантическом уровнях.

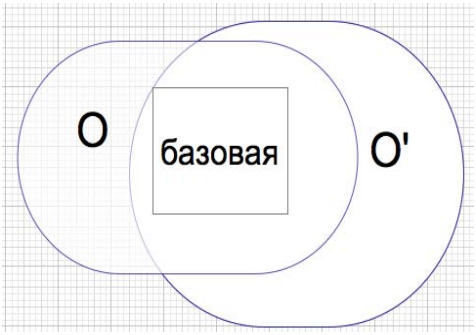


Рис. 2. Соотношение онтологий

Благодаря указанным методикам становится возможным использование следующего алгоритма:

- используя расстояние Дамерау – Левенштейна, обнаруживаем похожие объекты в онтологиях пользователей;
- учитывая полученные данные, производим сравнение онтологий в рамках выделенных предметных областей; в итоге получаем онтологии пользователей, ранжированные в порядке убывания схожести с онтологией выбранного пользователя;
- несколько первых из полученных онтологий проверяются на полноту относительно зафиксированной онтологии, а зафиксированная онтология проверяется на полноту относительно них.

В итоге получаем набор онтологических данных для пополнения. Кроме того, полученные на начальном шаге данные могут быть использованы в дальнейшем при вводе новых ключевых слов для контроля или уточнения введенных пользователем данных (например, для проверки правописания или уточнения контекста вводимого слова или выражения).

## 4 Заключение

Выбранное решение использовано для разработки прототипа интеллектуальной системы каталогизации пользовательских закладок.

Данная система предполагает расширение введённых пользователем данных путём обнаружения соответствий с данными, уже присутствующими в системе. При этом расширение касается не только информационных объектов, но и описывающих их тегов. Таким образом, происходит расширение контекста, в котором информационный объект может быть рассмотрен в терминах пользователя.

На данный момент главной нерешённой проблемой является проблема выбора адекватной схемы визуализации набора тегов, с учётом омонимов и жаргона. Возможно, одним из наиболее удачных решений будет создание специальной «карточки» с указателем на понятие для каждого многозначного случая.

## Литература

[1] Еременко Т.В. Каталогизация ресурсов Интернета: Опыт библиотек США// Электронные ресурсы в библиотеках. – http://www.eril.ru/magazine/archive/2004/1/theme/yeryomenko.php.

[2] Mika P. Ontologies are us: a unified model of social networks and semantics// 4th Int. Semantic Web Conf. – http://www.cs.vu.nl/~pmika/research/papers/ISWC-folksonomy.pdf.

[3] Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А. Циркин Б.Г., Холюшкин Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Труды 10-й национальной конф. по искусственному интеллекту.

[4] Halpin H., Robu V., Shepherd H. The complex dynamics of collaborative tagging// 16th Int. World Wide Web Conf.

[5] Кучеренко Е.И., Павлов Д.А. О проблемах выявления неполноты и избыточности в онтологических пространствах объектов исследования. – http://shcherbak.net/protivorechivost/ (блог PhD С. Щербака).

[6] Maedche A., Staab S. Comparing ontologies – similarity measures and a comparison study// Technical Report 408, University of Karlsruhe.

[7] Alani H., Brewster C. Metrics for ranking ontologies// 15th Int. Conf. for World Wide Web, 2006. – Edinburgh, UK.

## Using ontologies for implementation of the catalogue of user predilections

B.G. Tsirkin

The article presents the way of organization of intellectual information system as social bookmaking system based on ontologies. The suggested solution is compared with traditional library and with existing social bookmaking systems.

This system is based on global ontology, which can be extended by each user. In addition there should be interaction between user ontologies and data provided by them.

It is assumed, that data on the portal is represented by hyperlinks, which are pointing to any information object. Each information object has a collection of tags (or key expressions), and both a visual presentation of tag itself and its context can be used as a tag.

Algorithm of data-mining for user ontology and hyperlinks is also one of the themes of this work.

The paper also speaks about main problems and tasks, which are presenting the most difficult part of the implementation and portal organization.