8Г

УДК 681.3

А.Я. Гладун
Международный научно-учебный центр информационных технологий и систем НАНУ и МОНУ, г. Киев, Украина, glanat@yahoo.com
Ю.В. Рогушина
Институт программных систем НАНУ, г. Киев, Украина, _jjj_@ukr.net

Применение тезауруса предметной области для повышения релевантности поиска в Интернете

Для того чтобы повысить релевантность поиска информации в Интернете, предлагается использовать знания пользователя о ПрО, которая его интересует, представленные в виде онтологии. На основе множества терминов онтологии ПрО строится тезаурус пользователя, который используется для оценки того, насколько интересен этот ИР пользователю.

Проблемы поиска информации в Интернет

В настоящее время основные направления развития информационных технологий (ИТ) связаны с созданием информационных систем, основанных на знаниях соответствующих предметных областей (ПрО). Большинство людей могут считаться экспертами в определенных ПрО, отражающих, например, их профессиональную или научную деятельность, другие интересы.

Одна из наиболее часто встречающихся задач в ИТ – поиск информации (в Интернет, локальной сети, на отдельном компьютере), представленной в различных формах (текст, графика, аудиоинформация, мультимедиа и т.д.). Пользователю доступно большое количество информации, которую он должен отфильтровывать и искать релевантную информацию. Механизмы поиска типа Google и Yahoo пытаются облегчать эту проблему, индексируя в значительной степени неструктуризированную и неуправляемую информацию в Интернете.

При этом пользователь, как правило, не всегда является специалистом в области ИТ и вследствие этого может применять только наиболее простые и интуитивно понятные средства формирования запросов. Так, большинство пользователей, обращающихся к информационно-поисковым системам (ИПС) Интернета, используют только простые запросы, состоящие из 2 – 3 слов, не используют логические операторы и прочие механизмы расширенного поиска. Вследствие этого они получают в результате выполнения такого запроса большое количество информационных ресурсов (ИР), релевантных запросу, но не отвечающих реальной информационной потребности пользователя.

Запрос пользователя представляет собой некоторый образ (описание) информации, доступ к которой он хочет получить. Такой запрос может, например, содержать ключевые слова, связанные логическими операторами; документ-образец; тип документа и его тему по классификатору; списки рекомендованных или

запрещенных пользователем информационных источников; ограничения на время или объем поиска; объем, время создания, язык искомого документа [1].

Релевантность – формальное соответствие информации, выдаваемой ИПС, запросу, введенному пользователем. Однако для пользователя важнее другой параметр оценки качества функционирования ИПС – *пертинентность*, т.е. соотношение объема полезной для него информации к общему объему полученной информации и этот параметр часто имеет решающее значение. При этом следует учитывать, что формальный запрос к системе является попыткой пользователя формализовать свою информационную потребность и, к сожалению, не всегда точно отражает последнюю (либо вследствие низкой выразительной мощности языка создания запросов к ИПС, либо из-за низкой квалификации пользователя). Неумение большинства пользователей правильно формулировать запросы и получать приемлемые объемы ИР после отклика приводит к снижению эффективности использования ИР Интернета и требует разработки новых интеллектуальных программных средств, позволяющих облегчить пользователю взаимодействие с ИПС.

При поиске информации пользователь выражает свою информационную потребность посредством ключевых слов часто неточно. Но, учитывая предысторию запросов, приоритеты для различных источников информации и другие характеристики искомых документов (язык, размер, наличие терминов онтологии и т.д.), а также знания пользователя о той ПрО, к которой должен относиться искомый ИР, можно более точно (по сравнению с существующими ИПС) прогнозировать, удовлетворит ли пользователя предложенный ему документ. Кроме того, полезно учитывать, был ли ранее предложен пользователю данный документ и какую оценку дал ему пользователь.

Описание ПрО, интересующей пользователя, через онтологии

Знания пользователя о конкретной ПрО, которая его интересует, необходимо представить в некоторой форме, пригодной для автоматической обработки. При этом важно достигнуть интероперабельности знаний, т.е. знания, полученные при создании одной ИС, должны быть пригодны при работе других ИС. Онтологии являются именно такой формой представления знаний.

Онтология – соглашение об общем использовании понятий, которое содержит средства представления предметных знаний и договоренности о методах соображений. Она может рассматриваться как определенное описание взгляда на мир в конкретной сфере интересов, который состоит из набора терминов и правил использования этих терминов, которые ограничивают их значение в рамках конкретной ПрО [2]. Онтологии позволяют формализовать знания пользователей о той ПрО, которая их интересует. При этом такие знания становятся доступны другим пользователям и могут применяться в других ИС. Онтологии, созданные для поиска информации, могут потом использоваться и для решения других задач, стоящих перед пользователем (например, при выборе товаров е-коммерции, нахождении подходящего курса дистанционного обучения).

Интерес к использованию онтологического представления знаний привел к разработке ряда формальных языков описания онтологий. К наиболее известным языкам относятся OIL и OWL, а также существует программное обеспечение (как коммерческое, так и свободно распространяемое), предназначенное для обработки и

анализа знания, представленного на этих языках, – Ontolingua [3], Protégé [4], DOE [5], OntoEdit, OilEd и т.д.

В частности, Protégé – локальная, свободно распространяемая Java-программа, предназначенная для построения (создания, редактирования и просмотра) онтологий. На основе сформированной онтологии Protege генерирует формы получения знаний для введения экземпляров классов и подклассов. Инструмент имеет удобный графический интерфейс. Он поддерживает использование языка OWL и разрешает генерировать html-документы, которые отображают структуру онтологий.

Использование тезаурусов для обработки информации на семантическом уровне

Термин «*тезаурус*» был применен впервые в XIII веке учителем Данте, флорентийцем Б. Латини, как название энциклопедии. Слово «thesaurus» означает сокровище, богатство, запас. Согласно «Современному словарю иностранных слов»: *тезаурус* – 1) словарь, в котором максимально полно представлены все слова языка с исчерпывающим перечнем примеров их употребления в текстах; в полном объеме осуществим лишь для мертвых языков; 2) идеографический словарь, в котором показаны семантические отношения (синонимические, родовидовые и др.) между лексическими единицами; 3) в информатике – полный систематизированный набор данных о какой-либо области знаний, позволяющий человеку или вычислительной машине в ней ориентироваться. Тезаурус (согласно третьему определению) можно рассматривать как частный случай онтологии. Очевидно, что можно говорить о тезаурусе человечества как о сумме накопленных им знаний. Можно исследовать как тезаурусы отдельных специалистов, так и тезаурусы областей знания.

По онтологии ПрО строится тезаурус для определения количества информации в ИР на семантическом уровне как критерий оценки его пертинентности. При этом дескрипторами тезауруса являются элементы множества терминов онтологии.

Для измерения количества информации в ИР на *семантическом* уровне используют тезаурус, чтобы связать семантические свойства информации с возможностью пользователя воспринимать сообщение, которое поступило. Тезаурус – это совокупность терминов, которые применяет пользователь ИС.

Предельные случаи, если количество семантической информации в сообщении равняется нулю: 1) пользователь вообще не понимает информации; 2) пользователь все знает, а та информация, которая поступает, ему не нужна. Примером первого предельного случая может быть текст на неизвестном пользователю языке, а второго – таблица умножения для студента. Максимальное количество семантической информации пользователь приобретает при согласовании его тезауруса с содержанием ИР, если информация понятна пользователю и несет сведения, отсутствующие в его тезаурусе.

Большинство существующих информационно-поисковых систем имеют развитые средства контекстного поиска документов с учетом морфологической информации о словах. Однако в настоящее время очень незначительное число информационных систем предоставляют возможность тематического поиска, например поиска с использованием тезауруса.

Толковые словари в электронном виде, используемые для описания терминологии какой-либо отрасли знаний в автоматизированных системах поиска информации, получили название информационно-поисковых тезаурусов. Он задает систему

семантических, смысловых связей между понятиями. Каждое понятие в тезаурусе может объясняться через набор других понятий, что приводит к появлению семантического поля. Фактически тезаурус пользователя – потребителя информации – это вербализованная совокупность его представлений об исследуемой ПрО.

Основной целью разработки информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования.

Разработка тезауруса для автоматической оценки семантического количества информации в ИР характеризуется прежде всего необходимостью описания значительно большего количества терминов (слов и словосочетаний), встречающихся в текстах данной ПрО. Тезаурус должен включать не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывать широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня.

В результате сопоставления контента ИР с тезаурусом пользователя создается понятийный индекс ИР, в котором указывается, какие дескрипторы тезауруса обнаружены.

Если пользователю предлагают ИР, в котором нет ни одного знакомого ему термина, то он не извлекает из ИР никакой информации. Если же все сведения, содержащиеся в ИР (термины и связи между ними), уже известны пользователю, то никаких изменений в его знаниях тоже не происходит и, таким образом, семантическое количество информации такого ИР также равно нулю.

Чем больше тезаурусы пользователя и ИР, тем больше вероятность того, что они пересекутся. Между семантическим количеством информации в ИР и тезаурусом пользователя существует нелинейная зависимость, вид которой зависит от специфики и широты ПрО, от функции распределения терминов (рис. 1). Вариант А характерен, например, для компьютерной техники, а вариант Б – для математики. Общими в них являются две точки: если тезаурус пользователя – пустое множество, то семантическая ценность любого ИР равна нулю; если тезаурус пользователя покрывает всю ПрО, то семантическая ценность любого ИР этой ПрО также равна нулю.

Чем шире тезаурус пользователя, тем сложнее ему найти ИР, удовлетворяющие его информационные потребности. Этот эффект можно интерпретировать как закон нарастающей трудности в достижении полной информированности.

Именно эта двойственность природы тезауруса отражает одну из объективных предпосылок возникновения смежных научных дисциплин – интеграции наук: по мере углубления познания в процессе развития отдельной конкретной науки, все более детального расчленения ее предмета на частные направления информация, приносимая их изучением, утрачивает свою ценность. На этом этапе возникает объективная потребность расширения объекта познания, объединения нескольких научных направлений, на стыке которых научные исследования снова обретают свою ценность.

Иными словами, для того чтобы процесс расширения тезауруса имел достаточно высокий стимул, необходимо, чтобы тезаурус исследователя был постоянно выше того уровня, который необходим для адекватного восприятия поступающей информации. Интересна технология его расширения для этой цели, это освоение основных понятий и идей смежных отраслей знания.

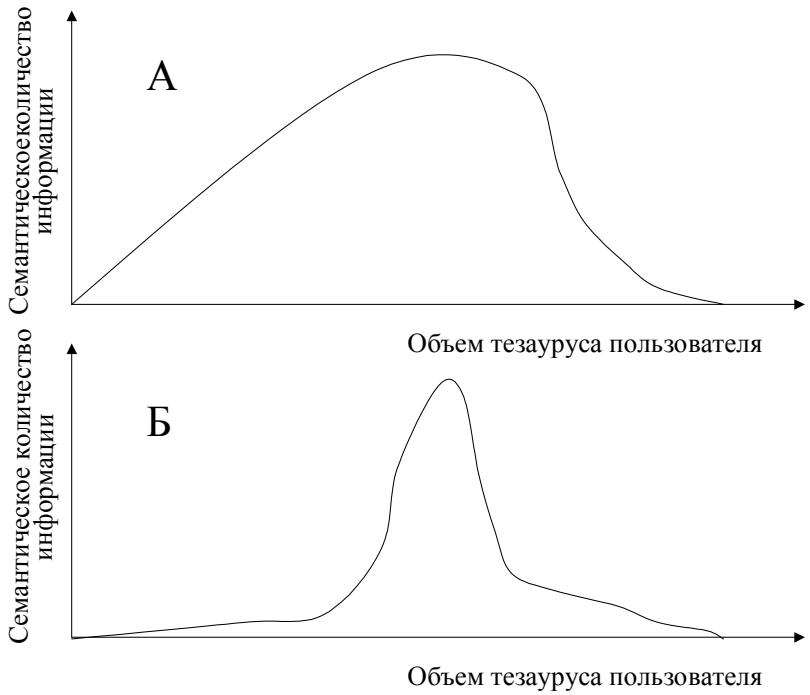

Рисунок 1 – График зависимости между тезаурусом пользователя и семантическим количеством информации ИР

Объекты внешнего мира и отношения между ними, отражаясь мозгом человека, образуют его *тезаурус плана содержания*. Вербализованная часть плана содержания (слова, поставленные в соответствие элементам плана содержания – информационным единицам – узлам и дугам – отношениям) составляет *тезаурус плана выражения.* Планы содержания и планы выражения не обязательно идентичны, поскольку «слово не покрывает понятия». Однако исследовать тезаурус плана выражения существенно легче, чем структуру нейронных ансамблей – физических (физиологических) носителей информации плана содержания.

Постановка задачи

Для того чтобы повысить релевантность поиска информации в Интернете, предлагается использовать знания пользователя о ПрО, которая его интересует, представленные в виде онтологии. На основе множества терминов онтологии ПрО строится тезаурус пользователя, который используется для оценки того, насколько интересен этот ИР пользователю.

Алгоритм определения оценки соответствия ИР информационным потребностям пользователя

1. На первом этапе пользователь должен создать онтологию, в которой содержатся основные термины ПрО и связи между ними, и сохранить ее. Предлагается использовать для этого редактор онтологий Protégé (рис. 2).

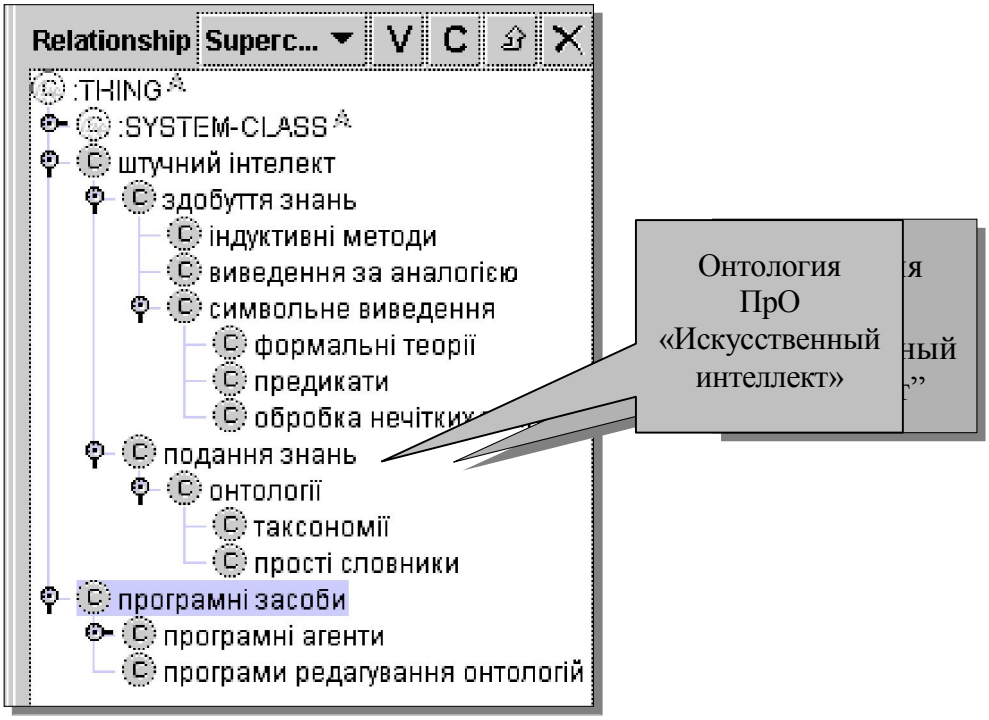

Рисунок 2 – Представление онтологии ПрО «Искусственный интеллект» в Protégé

2. На следующем этапе пользователь отмечает в онтологии нужные ему подклассы и сохраняет информацию о них в формате html или xml (рис. 3).

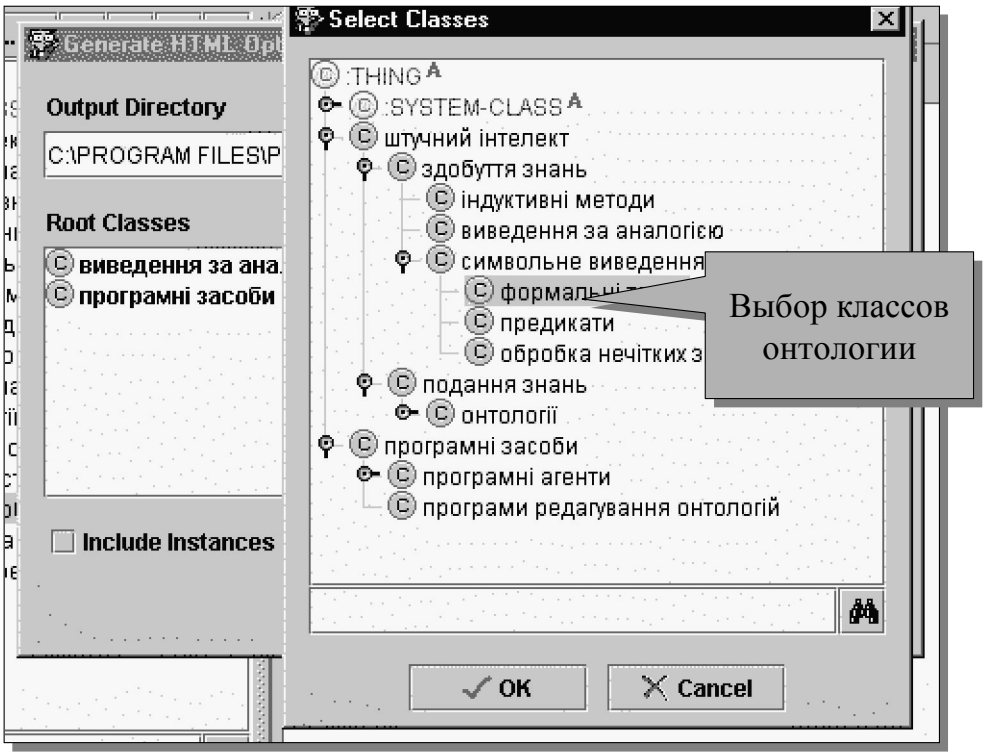

Рисунок 3 – Выбор терминов онтологии для формирования тезауруса ПрО

3. По сформированному на втором этапе файлу строится тезаурус ПрО.

Альтернативным способом построения тезауруса является непосредственный ввод терминов в соответствующем окне программы OntologySearch.

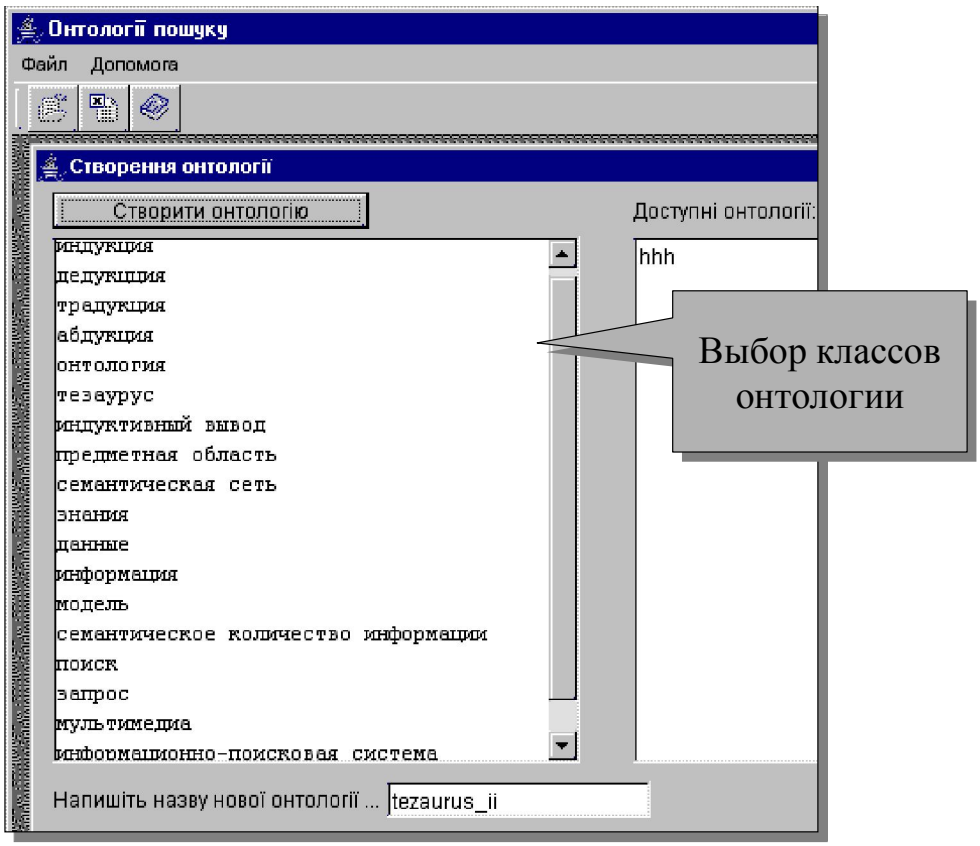

Рисунок 4 – Окно редактирования тезауруса в OntologySearch

Затем этот тезаурус используется для фильтрации результатов запроса пользователя к внешней информационно-поисковой системе (ИПС) Интернета. При тестировании OntologySearch в качестве внешней ИПС использовался Google. В ответ на запрос Z пользователя ИПС формирует множество ссылок на ИР, каждая из которых содержит адрес ИР и его краткое описание. $f_{ИПС}(Z) = \{ИР_i = \langle a_i, abstr_i \rangle, i = \overline{1,n}\}$. Затем для всех дескрипторов тезауруса $\forall t \in Tezaurus$ производится проверка на их наличие в описании каждого из ИР, входящих в множество $f_{ИПС}(Z)$. При вхождении термина оценка ИР изменяется (при вхождении в список положительного рейтинга – увеличивается, отрицательного – уменьшается). Пользователю ИР предлагаются в порядке, зависящем от полученной ресурсом оценки.

Программная реализация

Программа *OntologySearch* написана на языке программирования Java. Программа состоит из трех основных частей: поисковый модуль, модуль создания онтологий и модуль подключения онтологий.

С помощью модуля создания онтологий создается тезаурус пользователя для ПрО, в которых будет проводиться поиск. Для этого пользователю нужно ввести

список терминов, которые важно учитывать при поиске, или импортировать его из другой программы. Информация хранится в xml-файле и подключается к поисковому модулю в процессе поиска и затем для определения релевантности результатов в соответствии с ПрО, заданной пользователем.

Модуль подключения тезаурусов позволяет подключить ранее созданные онтологии к поисковому модулю. Причем можно подключить онтологии как для повышения, так и для понижения рейтинга поисковых результатов и для уточнения соответствия результатов поиска ИР.

Если в найденных результатах встречаются термины из *положительного* тезауруса (тезауруса, подключенного к поисковому модулю для того, чтобы указать, что пользователь интересуется этой ПрО), то соответственному результату приписывается больший рейтинг и он продвигается вверх, к началу списка результатов. Но если в результатах встречаются термины из отрицательного тезауруса (тезауруса, подключенного к поисковому модулю для того, чтобы указать, что пользователь не интересуется этой ПрО), то соответственному результату приписывается меньший рейтинг и он продвигается вниз, к концу списка результатов [6].

Модуль управления поиском (для взаимодействия с поисковой машиной) отвечает непосредственно за поиск и сортировку результатов в соответствии с подключенными онтологиями. С помощью этого модуля можно просмотреть список найденных и отсортированных результатов, а также перейти по гиперссылке к интересующему результату и подробно с ним ознакомиться.

Перспективы дальнейших исследований

При создании онтологии необходимо явно указать основные понятия ПрО и связи между ними. К сожалению, большинству пользователей достаточно сложно это сделать (даже имея соответствующие знания и применяя их в своей деятельности). На первом этапе формирования онтологии пользователь может выбрать одно из следующих решений:

– самостоятельно построить с помощью одного из редакторов онтологий онтологическое описание области его информационных интересов;
– найти (например, в Интернете) какую-либо онтологию, представленную на языке OWL, которую описывает ПрО, близкую к области его информационных интересов;
– сформировать множество понятий ПрО, которое содержит наиболее характерные слова и словосочетания, встречающиеся в интересующих его ИР.

В любом случае в процессе работы пользователь может вносить изменения и дополнения в сформированную ранее онтологию. Говоря о системе, условно выделяют некоторую группу наиболее сильно связанных между собой обратными и прямыми связями элементов. При этом связи данной группы элементов с внешним миром считаются несущественными. Важно определить, какие именно связи между элементами являются существенными (и их, следовательно, необходимо включить в систему), а какие можно не учитывать при построении модели. Если несколько параметров связаны друг с другом, то следует учитывать только один из них. В связи с тем, что не все связи между терминами онтологии могут быть очевидны пользователю, он может воспользоваться для их нахождения методами индуктивного вывода.

Существуют независимые направления развития подобных систем: ID3, ACLS, CART – и коммерчески, и академически успешные. Наиболее интересным в связи со спецификой проводимой работы оказался алгоритм ID3 [7], который специально

разработан для извлечения ценной информации из больших объемов слабоструктурированных данных, так как при работе этого алгоритма время вычислений зависит линейно от числа введенных примеров, числа атрибутов, используемых для описания примеров, и числа узлов в строящемся дереве решений. Это качество отличает его от таких известных алгоритмов построения деревьев решений, как INDUCE, SPROUTER, ROTH-P, в которых усилия, требующиеся для решения задачи, резко возрастают вместе со сложностью задачи.

Если методы, подобные МГУА, предназначены для нахождения закономерностей по набору количественных измерений параметров и полученному по ним результату, то методы, подобные ID3 и его вариациям (C4.5, ID4), предназначены для обобщения опыта экспериментов, параметры и результаты которых описаны через качественные оценки (лингвистические переменные). В большинстве случаев между их значениями невозможно установить даже относительное упорядочение (например, различные симптомы и диагнозы пациентов). К таким задачам относится и задача поиска информации в Интернете. Например, такой существенный параметр ИР, как язык, не может быть описан количественно.

Алгоритм ID3 принадлежит к невозрастающим алгоритмам, то есть при добавлении к набору классифицированных примеров определенного количества новых нужно обрабатывать снова как старые, так и новые примеры. Если алгоритм возрастающий, текущее определение понятия пересматривается, если необходимо, для каждого нового примера.

К невозрастающим алгоритмам относятся также THOTH и UNDUCE. Как примеры возрастающих алгоритмов можно привести Candidate Elimination Algorithm, STAGGER, COBWER, PockerAlgorithm.

Предлагается использовать ID3m [8] – модификацию ID3 для произвольного (конечного) количества решений. Он принадлежит к невозрастающим алгоритмам. Этот алгоритм обладает следующими свойствами.

Свойство 1. ID3m строит дерево решений, которое является оптимальным в том смысле, что время, затрачиваемое пользователем для получения результата консультации, будет в среднем лучше, чем при консультации с любым другим деревом решений, дающим тот же результат.

Свойство 2. ID3m является оптимальным в смысле средней длины пути к решению в дереве на распознаваемом подобными алгоритмами классе задач.

В данном случае примерами обучающей выборки являются ИР, ранее полученные пользователем в результате запросов к ИПС. Параметрами, по которым они описываются, являются свойства ИР (язык, время создания, размер, формат, право доступа и т.д.), а также термины онтологии пользователя. Значения, соответствующие терминам онтологии: «Термин отсутствует в ИР», «Термин встречается в ИР редко», «Термин встречается в ИР часто». В качестве результата используется оценка, данная пользователю, найденному ИР (качественная оценка, имеющая два и более значений).

На вход алгоритма поступает обучающая выборка – набор классифицированных (получивших одну из возможных оценок) примеров одинаковой размерности.

Если обучающая выборка содержит примеры, в которых все значения атрибутов одинаковы, а решения различны, то введенная информация недостаточна для построения классификационного правила.

Если множество примеров пустое, то можно произвольно связать его с любым решением.

Если все примеры относятся к одному классу, строится один лист дерева решений, связанный с этим классом. В противном случае необходимо выбрать один из атрибутов и разделить множество атрибутов на подмножества в зависимости от значения этого атрибута и применить алгоритм к каждому из полученных подмножеств.

На каждом шаге работы алгоритма вычисляется, какой атрибут несет наибольшее количество информации о результате (т.е. выяснение значения которого позволяет максимально снизить энтропию).

$$C(A_m) = \sum_i \sum_j \frac{C(A_m = a_{mi}, R = R_j)}{T(A_m)} = \max_s C(A_s) = \max_s \sum_i \sum_j \frac{C(A_s = a_{si}, R = R_j)}{T(A_m)},$$

где – C(X,Y) – количество информации

$$C(X,Y) = \sum_i \sum_j p(X = x, Y = y) * \log p(X = x, Y = y),$$

– p(X = x,Y = y) – возможность общего наступления событий X = x и Y = y,
– $T(A_m)$ – стоимость получения значения A_m.

В результате работы алгоритма ID3m формируется дерево решений, в котором каждый лист связан с одним из решений, каждый узел характеризуется именем одного из атрибутов, а выходящие из такого узла ветви – значениями этого атрибута.

Такое дерево решений позволяет ИПС по параметрам вновь найденного ИР прогнозировать, как именно оценит его пользователь, и предлагать пользователю в первую очередь те ИР, которые соответствуют его индивидуальным предпочтениям.

Так как точные значения вероятностей событий из обучающей выборки неизвестны, то они аппроксимируются на основе рассматриваемого множества примеров, т.е. P(A = a[i]]) – это отношение количества примеров из рассматриваемого набора Q, для которых выполняется данное равенство, к общему количеству примеров.

$$P(A[m] = a[m][i], R = r[j]) = \frac{N(A[m] = a[m][i], R = r[j])}{N(R = r[j])},$$

где функция N(q) – количество примеров в подмножестве X, которое рассматривается на этом шаге, для которых выполняется условие q.

Еще один вариант индуктивного обобщения опыта взаимодействия пользователя с ИПС – внесение уточнений в онтологию ПрО [9]. В этом случае результирующим параметром обучающей выборки является тот термин, связи которого с другими терминами онтологии пользователь хочет уточнить, а примерами обучающей выборки – только те ИР, которые удовлетворяют информационным потребностям пользователя. Те термины онтологии, которые вошли в построенное по такой обучающей выборке дерево решений и связаны с результатом «Термин встречается в ИР часто» ветвями, также связанными со значениями «Термин встречается в ИР часто», должны быть и в пользовательской онтологии связаны с этим термином (семантику связи должен определить пользователь).

Получив дерево решений, пользователь может увидеть эти связи и с помощью редактора онтологий зафиксировать их.

Таким образом, методы индуктивного обобщения представляют средства автоматизации создания онтологий ПрО и повышения релевантности поиска.

Выводы

Предложенный в работе подход к поиску информации в Интернете основывается на использовании знаний пользователя о ПрО, характеризующей его информационные потребности. Пользователь может явно указывать интересующие его термины и получать те информационные ресурсы, которые соответствуют его запросу и содержат также и эти термины. Система OntologySearch ориентирована на пользователя с относительно стабильными информационными потребностями, не являющегося специалистом в области информационных технологий. Ее использование позволяет пользователю избежать рутинной работы по фильтрации результатов обращения к ИПС.

Кроме того, следует отметить, что в настоящее время в области разработки и реализации интеллектуальных систем сложилось следующее положение: с одной стороны, квалификация коллективов разработчиков, как правило, достаточно высока, с другой стороны, одной из сложнейших проблем, препятствующих широкому внедрению ИС, является недостаточное знание системными аналитиками и программмистами предметных областей, в рамках которых готовятся проекты [9], [10].

Литература

1. Рогушина Ю.В. Использование онтологического описания предметной области для повышения релевантности информационного поиска // Проблемы программирования. – 2003. – № 4. – С. 54-64.
2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001.
3. Farquhar A., Fikes R., Rice J. The Ontolingua server: A tool for collaborative ontology construction // International Journal of Human-Computer Studies. – 1997. – № 46(6). – P. 707-728.
4. Musen M. Domain Ontologies in Software Engineering: Use of Protege with the EON Architecture // Methods of Inform. in Medicine. – 1998. – P. 540-550.
5. DOE – The Differential Ontology Editor // http://opales.ina.fr/public/.
6. Noy N., Musen M. The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping // Stanford Medical Informatics. – Stanford Univ. – 2003.
7. Quinlan J.R. Discovery rules from large collections of examples: a case study // Expert Systems in the Microelectronic Age. – Edinburg, 1979. – P. 87-102.
8. Рогушина Ю.В. Применение методов индуктивного вывода для создания прикладных экспертных систем // Разработка и использование информационных технологий в системах управления. – Киев: Ин-т кибернетики им. В.М. Глушкова НАН Украины, 1993. – С. 122-128.
9. Гладун А.Я, Рогушина Ю.В., Штонда В.М. Аналіз онтологічних моделей предметних областей як засіб інтелектуалізації пошуку в Інтернеті // Труды Межд. научной конф. DPMSI'2005. – Киев: КГУ им. Шевченко. – 2005. – С. 202-203.
10. Гладун А.Я., Рогушина Ю.В., Штонда В.М. Використання онтологічних моделей предметних областей для інтелектуалізації пошуку в гетерогенному інформаційному просторі // Материалы Межд. конф. по математическому моделированию МММ'2005. – Феодосія. – 2005. – С. 145-151.

А.Я. Гладун, Ю.В. Рогушина
Застосування тезауруса предметної області для підвищення релевантності пошуку в Інтернеті
Для підвищення релевантності пошуку інформації в Інтернеті, пропонується використати знання користувача, представлені у вигляді онтології про предметну область, яка його цікавить. На основі набору онтологічних термінів предметної області будується тезаурус користувача, який використовується для оцінки того, наскільки цей ІР є цікавим користувачеві.

A.J. Gladun, U.V. Roguschina
The Use of the Thesaurus of Object Srhere to Increase the Relevance of Search in the Internet
In order to promote relevance of information retrieval in the Internet the user knowledges about domain that interests him are used and represented as ontology is suggested. On the basis of set of domain ontology terms the user thesaurus that is used for estimation of informational resources is generated.

Статья поступила в редакцию 18.07.2005.