УДК 519.767

АНАЛИЗ МЕТОДОВ СЕМАНТИЧЕСКОГО ПОИСКА ИНФОРМАЦИОННЫХ РЕСУРСОВ

***Валентин Викторович Нечаев**, д-р физ.-мат. наук (ВМАКК),*
канд. техн. наук, проф., зав. кафедрой
E-mail: nechaev@mirea.ru
***Всеволод Михайлович Трофименко**, аспирант кафедры*
Интеллектуальных технологий и систем
E-mail: trofsev@mail.ru
Московский государственный технический университет радиотехники, электроники и автоматики
https://www.mirea.ru

Описывается методика смыслового отбора пертинентных информационных ресурсов на основе бионического подхода и метода онтологий. В ее основу положены методы семантического анализа лексических единиц, методы статистического анализа и технологии интеллектуального анализа данных. Рассматриваются биологические особенности мыслительной деятельности головного мозга. Работа ориентирована на реализацию проблемы автоматизации процесса повышения компетенции эксперта в предметно-ориентированной области знаний.

Ключевые слова: смысловой отбор информационных ресурсов, бионический подход, пертинентные информационные ресурсы, семантический поиск информации, показатель релевантности информационного ресурса, показатель пертинентности информационного ресурса.

Введение

Проблема поиска и извлечения необходимой информации возникла задолго до появления интернета [1]. Актуальность проблемы целенаправленного поиска информации, соответствующей запросам и потребностям пользователя, т.е. релевантной и пертинентной информации, в последние годы существенно возросла. Многократное увеличение информационных потоков, циркулирующих между пользователями[1], динамичное развитие информационных ресурсов приводит к тому, что пользователь уже не в состоянии самостоятельно, без технической поддержки, находить требуемую информацию. Интенсивный рост информационных потоков также свидетельствует о необходимости постоянного совершенствования методов, приемов и технологий обработки данных. Следовательно, для обеспечения быстрого поиска необходимой информации пользователю необходимо применять все более совершенные навигационные сервисы и приемы поиска соответствующих его потребностям информационных ресурсов.


В.В. Нечаев


В.М. Трофименко

Релевантность и пертинентность информационных ресурсов

На современном этапе развития информационных технологий в качестве одного из основных инструментальных средств проведения навигации и поиска в сети интернет необходимых информационных ресурсов являются поисковые системы. Их работа основана на использовании ключевых слов. Под ключевыми словами подразумеваются наиболее важные термины, которые загружаются в поисковую систему в качестве аргумента поиска. Поисковая система составляет и хранит предметный указатель сети Интернет и при формировании запроса пользователем находит в нем заданные ключевые слова. Результатом поиска является список ресурсов Интернета, подлежащих детальному рассмотрению пользователем.

[1] По оценкам, объемы созданных во всем мире данных уже измеряются в зеттабайтах.

При этом современные поисковые системы работают таким образом, что при определении смысла запроса система ограничивается только лишь введенным пользователем словом или словосочетанием. Однако в информационном поиске не учитываются семантические аспекты понятий, включенных в запрос. При этом следует иметь в виду, что главной задачей при работе поисковой системы является нахождение и выдача пользователю максимально *релевантных* результатов [2].

Релевантность – это показатель степени соответствия найденного документа (ресурса), сформированному пользователем запросу. Показатель релевантности определяется путем сравнения образа поискового запроса с поисковым образом документа по определенному алгоритму.

Так как пользователь формирует запрос на естественном языке, то релевантная запросу информация может не соответствовать информационной потребности пользователя. Из-за того, что одно и то же понятие может быть многозначным, существует большая вероятность того, что поисковая система не правильно «поймет» что от нее требуется найти и выдаст пользователю хотя и релевантные данные, но не удовлетворяющие задуманному пользователем смыслу. Релевантных запросу данных в лучшем случае несколько сотен, а в худшем случае миллионы, из которых пользователю приходится выбирать наиболее подходящие по смыслу. Отсюда и возникает необходимость использования иного, более «тонкого» критерия, благодаря которому можно оценить соответствует ли результат поиска заданному смыслу запроса. Такой критерий называется *пертинентностью.*

Пертинентность (в информационном поиске) − соответствие полученной информации информационной потребности пользователя. Пертинентность характеризуется степенью соответствия между тем, что необходимо получить пользователю и непосредственными результатами поиска, т.е. определяется как отношение объема полезной для пользователя информации к общему объему полученной информации, найденной поисковой системой.

На данный момент достижение высокой степени пертинентности информационного ресурса – основное поле конкурентной борьбы современных поисковых систем. Требуемой степени пертиненстности можно добиться, используя в поисковой системе в качестве ключевых слов – понятия, учитывающие такие характеристики как словоизменение, полисемию (одно слово имеют несколько значений), синонимию, омонимию. При этом необходимо выявлять смысловые связи между понятиями и формировать семантическую сеть понятий.

По мнению Ф. де Соссюра[1][3], языковые знаки состоят из двух компонентов: означающего и означаемого. Означающее – это звуковой или графический образ знака, а означаемое – соответствующее ему понятие [3]. Согласно [4], наиболее устойчивыми единицами смысла являются понятия. «*Центральной процедурой любых систем автоматической смысловой обработки текстов должна быть процедура семантико-синтаксического концептуального (понятийного) анализа. Она должна быть реализована, прежде всего, как процедура фразеологического концептуального анализа на основе мощных словарей наименований понятий»* [4].

Таким образом, для пертинентного поиска, как и для поисковых систем сети Интернет, необходимо предварительно составить словарь из слов, встречающихся в документах, в котором при каждом слове будет храниться список документов, из которых взято данное слово. Таким образом, при создании системы, которая на запрос пользователя выдавала бы пертинентный результат, предварительно должен быть составлен не просто словарь из слов, а словарь, состоящий из понятий, в котором они связаны в рамках некоторой семантической сети. Первой при разработке подобной системы является задача извлечения и представления понятий. Под понятием в данной работе подразумевается некоторое знание, сформированное при анализе текстовой информации. Очевидно, что знание необходимо рассматривать в динамике [5]. В семантической сети знание должно рассматриваться как системная категория, выполняющая определенную функцию формирования идеального результата на основе наследст-

[1] Ф. де Соссюр является одним из основоположников современной науки семиотики.

венного и эволюционно накопленного опыта. Таким образом, в системном знании должен фиксироваться опыт о прошлом, настоящем и возможном будущем системы и ее среды [5].

Для решения задачи получения из данных знаний используют различные типы интеллектуальных систем. Кроме информационно-поисковых к подобным системам относятся системы поддержки принятия решений и экспертные. Системы, предназначенные для оперирования знаниями (сбора, хранения, поиска и выдачи знаний) называют базами знаний. На основе баз знаний создаются системы, основанные на знаниях. Системы, основанные на знаниях, могут быть как монодисциплинарными, так и мультидисциплинарными.

Понятия, отражающие предметы реального мира – первичные понятия, находятся в сложной взаимосвязи. Исходя из того, что характеристики понятий, описывающих внешний мир, весьма многообразны, следует, что при целевом описании необходимо выделять наиболее существенные характеристики такой предметной области. С учётом вышесказанного, в рамках данной работы рассматривается возможность построения предметно-ориентированной системы, основанной на знаниях.

Формирование модели предметной области

Основные трудности при конструировании концептуальной модели предметной области связаны с невозможностью работы в компьютерной среде на естественном языке. Прежде чем начать анализ информации необходимо ее представить в пригодном для компьютерного распознавания виде. На первом этапе необходимо выделить нужную информацию, на втором выделенная информация определенным образом абстрагируется и формализуется. Другими словами, возникает необходимость выделить понятийную модель предметной области.

Для дальнейшего, уже формального описания предметной области введем определение понятийной модели предметной области. Понятийная модель предметной области – это совокупность понятий (концептов, терминов) и отношений между ними, которым соответствуют сущности из реального мира. Такая модель реализуется в форме ориентированного помеченного графа[1] $G_d = <V_d, E_d, M_d, L_d>$, у которого любая метка также является вершиной: $L_d \subset V_d$. Каждая вершина граф-модели является понятием, каждая дуга из вершины v_1 в вершину v_2 с меткой l описывает отношение l понятия v_1 к понятию v_2. Таким образом, любое отношение является понятием. Для построения понятийной модели предметной области необходимо решить следующие задачи: выделить понятия предметной области, установить отношения между этими понятиями. Подобного представления информации можно добиться с помощью онтологического описания.

Представление предметной области на основе онтологического описания

В настоящее время в области искусственного интеллекта разработан ряд средств представления знании. К наиболее эффективным из них, по мнению многих разработчиков программного обеспечения, относится онтология.

Под онтологией подразумевается система понятий (концептов, сущностей, классов), отношений между ними и правил операций над ними в определенной предметной области [6].

В качестве концептов в онтологических описаниях должна присутствовать большая часть понятий, характеризующих возможные запросы на поиск решений и фигурирующие в метаданных документов хранилищ данных. В общем виде структура онтологии представляет собой набор элементов четырех категорий: понятия, отношения, аксиомы и отдельные экземпляры [6].

Лексемы в проблемно-ориентированных текстах

Онтологии строятся в форме тезаурусов, содержащих терминосистему и категорийно-понятийный аппарат предметной области. Под терминосистемой понимается система-

[1] Ориентированный помеченный граф – это четверка $G = <V, E, M, L>$, где V – множество вершин, E – множество дуг $e = (v_1, v_2)$ из вершины v_1 в вершину v_2; M: $E \leftarrow L$ – функция разметки дуг, которая каждой дуге сопоставляет элемент из множества меток L.

тизированная совокупность терминов.

В данной работе предполагается, что знания для создания терминосистемы будут извлекаться из текстовых документов, касающихся определенной предметной области. Поэтому необходимо, чтобы до формирования тезауруса был проведен анализ существующих текстов на предмет выявления среди них таких, которые относятся к рассматриваемой предметной области. После первичного отбора текстовых документов должен происходить их терминологический анализ, целью которого является первичное выделение «*терминов-кандидатов*». Термины-кандидаты – это слова, которые извлекаются непосредственно текста путем первичной обработки текста.

Основные этапы терминологического анализа:

1. Выделение предложения из текста, определение границ слов предложения.

2. Выполнение лингвистической обработки. Выделенные лексические единицы (слова и словосочетания) приводятся к именительному падежу единственного числа и накапливаются в терминосистеме. Если выделенное словосочетание имело уже вхождение в терминосистему, то увеличивается счетчик количества его появления.

3. Определение частей речи слов и определение «стоп-слов». Слова, имеющие часть речи, отличную от прилагательных, причастий, порядковых числительных, выступающих в роли определений, существительных, номинирующих объект, процесс, состояние и т.д., а также предлогов и союзов, связывающих между собой номинативные группы, определяются как «стоп-слова». «Стоп-слова» назначаются границами фрагментов предложения.

4. Выполнение синтаксического анализа. К каждому фрагменту предложения последовательно применяется следующий набор правил:

4.1. В фрагменте предложения определяются анафорические местоимения[1] типа «он» и восстанавливаются анафорические конструкции до полных (на основании предыдущих предложений или других фрагментов анализируемого предложения).

4.2. В фрагменте находятся союзы «и» и «или» и устанавливается сочинительная связь между ними и примыкающими к ним справа словом.

4.3. Устанавливаются связи между существительными и согласованными определениями, а также устанавливаются связи между однородными определениями. Применяется правило установления связи предлога с существительным.

4.4. В фрагменте определяются предлоги и находятся существительные, к которым они относятся. Согласование предлога с существительным позволяет снять омонимию[2].

4.5. Устанавливаются подчинительные связи между существительными:

а) родительный падеж без предлога (генетивные цепочки);

б) творительный падеж без предлога;

в) предложное управление во всех косвенных падежах.

4.6. Восстанавливаются анафорические конструкции типа «этот».

При реализации терминологического анализа можно воспользоваться системой порождения грамматических парсеров[3] AGFL. AGFL (affix grammars over a finite lattice) – свободно распространяемое программное обеспечение для решения задач автоматической обработки текстов на естественном языке, использующее формализм AGFL [7]. Система AGFL позволяет генерировать эффективные парсеры для анализа морфологических и синтаксических структур естественных языков, при этом формат последовательности парсеров на выходе можно задавать при помощи трансдукций в правилах формализма. Помимо парсеров, система AGFL дает возможность подключить лексические базы данных большого объема. В конечном итоге, система AGFL позволяет создавать парсеры на основании лингвистических описаний, которые легко создаются и видоизменяются.

Для составления терминосистемы предметной области из выделенных терминов

[1] Анафорическая связь – это отношение, возникающее между словами в фразе или в тексте, когда смысл одного слова или выражения содержит отсылку к другому при отсутствии синтаксической связи между ними.

[2] Омонимия – это звуковое совпадение разных языковых единиц, значения которых не связаны друг с другом.

[3] Парсер – это программа или часть программы, выполняющая синтаксический анализ.

должно происходить автоматическое выделение наиболее важных терминов. Для этого необходимо назначить вес каждой лексической единице. При этом используются как статистические, так и семантические методы выделения ключевых слов из текстов. Комбинируя методы статистического и семантического анализа текстового документа можно выделить наиболее важные лексические единицы для составляемого тезауруса.

Во всех созданных человеком текстах можно выделить статистические закономерности. Для статистического анализа можно воспользоваться законами Ципфа (или Ципфа–Мандельброта) и выводами, которые можно сделать из этих законов. Ципф предположил, что слова с большим количеством букв встречаются в тексте реже коротких слов. Основываясь на этом постулате, он вывел два универсальных закона:

1. Если измерить количество вхождений каждого слова в текст и взять только одно значение из каждой группы, имеющей одинаковую частоту, расположить частоты по мере их убывания и пронумеровать (порядковый номер частоты называется рангом частоты), то наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними – 2 и т.д. Вероятность встретить произвольно выбранное слово будет равна отношению количества вхождений этого слова к общему числу слов в тексте.

2. Если построить график, отложив по одной оси (оси X) частоту вхождения слова, а по другой (оси Y) – количество слов в данной частоте, то получившаяся кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов.

Исследования показывают, что наиболее значимые лексические единицы лежат в средней части описанного Ципфом графика. Лексемы, которые попадаются слишком часто, в основном оказываются предлогами, местоимениями, а в английском языке – артиклями и т.п. Редко встречающиеся лексемы в большинстве случаев не имеют решающего смыслового значения. Сделать выделение наиболее значимых лексических единиц качественнее помогает предварительное исключение из исследуемого текста некоторых слов, которые априори не могут являться значимыми и поэтому являются «шумом».

Во многих современных поисковых системах используются законы Ципфа в преобразованном виде. Например, для оценки весов терминов можно воспользоваться статистической мерой TF-IDF (term frequency – inverse document frequency). Алгоритм статистического анализа, использующего меру TF-IDF следующий.

1. Выполняется вычисление частоты конкретной лексической единицы TF (term frequency). При этом оценивается важность лексической единицы t_i в пределах конкретного документа.

$$TF = \frac{n_i}{\sum_1^k n_k},$$

где n_i – число вхождений слова в документ;
$\sum_1^k n_k$ – общее число слов в данном документе.

2. Выполняется вычисление обратной частоты документа IDF (inverse document frequency). При этом оценивается инверсия частоты, с которой некоторая лексическая единица встречается в документах коллекции.

$$IDF = log\frac{|D|}{|(t_i \in d_i)|},$$

где $|D|$ – количество документов в корпусе,
$|t_i \in d_i|$ – количество документов,
в которых встречается t_i (когда $n_i \neq 0$).

3. Вычисление меры TF-IDF (веса лексической единицы).

$$TF - IDF = TF \cdot IDF.$$

Таким образом, последовательно анализируя каждую лексическую единицу текстового документа, можно сделать вывод о том, какой вклад она внесла в анализируемый документ, а при повторной встрече данной лексической единицы в других текстах можно сделать вывод о ее весе в рассматриваемых документах. Однако нельзя ограничиваться только статистической обработкой. Необходимо также применить метод семантического анализа текстовой информации, так как семантический анализ позволяет не только распо-

знать наиболее важные термины в тексте, но и классифицировать термины по семантическим признакам с учетом синонимических и гипонемических (общее – частное) классов.

Семантический анализ – процесс выявления смыслового содержания слов и словосочетаний в предложении. Тематику любого понятия или текста можно представить комбинацией ассоциирующихся с ним базовых семантических категорий, число которых уже гораздо меньше, чем число слов. Таким образом, семантическое описание использует вместо слов укрупненные понятия – категории, каждая из которых характеризуется своим набором терминов. Поэтому семантическое представление содержания текстов сопровождается существенным сжатием информации. Сжатие информации при переходе от лексического к семантическому описанию документов происходит за счет использования некоторого знания о структуре языка. Прежде всего, изучив выделенный экспертом документ, должна формироваться сеть основных (наиболее значимых) понятий, содержащихся в анализируемом тексте. Такая сеть служит представлением смысла текста и основой для всех видов дальнейшего анализа.

Семантическая сеть понятий – это множество терминов из текстов – слов и словосочетаний, связанных между собой по смыслу. В сеть включены не все термины текста, а лишь наиболее значимые, несущие основную смысловую нагрузку. Аналогичным образом должны представляться и смысловые связи между понятиями текстов – отражаются лишь наиболее явно выраженные из них. Поэтому, с одной стороны сеть достаточно полно описывает смысл текстов, а с другой – позволяет отбросить несущественную информацию и представить содержание в сжатом виде, так называемым «смысловым портретом». При этом каждое понятие, повторявшееся в различных местах текстов множество раз, оказывается представлено в единственном узле сети. В этом узле также собирается разбросанная информация, касающаяся понятия – формируется список предложений, в которых оно употреблялось. А различные формы слов приводятся к общей грамматической форме для отображения в один элемент сети. Аналогичным образом собирается информация по смысловым связям каждого понятия – в виде списка всех связанных с ним в тексте понятий, дополненного предложениями, в которых отражаются данные связи.

Таким образом, возникает возможность сразу увидеть всю информацию по каждому понятию. В результате, передвигаясь по смысловым связям от понятия к понятию, появляется возможность находить и исследовать лишь интересующие места текстов, не затрудняя себя просмотром всей попавшейся на пути информации.

Каждый элемент семантической сети понятий – это понятие, характеризующееся числовой оценкой – смысловым весом. Связи между парами понятий, в свою очередь, также характеризуются весами. Эти оценки позволяют сравнить относительный вклад различных понятий и их связей в семантику текста. Согласно ГОСТ 7.25-2001 «Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления» можно установить следующие связи между лексическими единицами: род − вид; часть − целое; причина − следствие; сырье − продукт; административная иерархия; процесс – объект; функциональное сходство; процесс − субъект; свойство − носитель свойства; антонимия.

Данные отношения могут быть разделены на два класса: *иерархические и ассоциативные*. *Родовидовая связь* устанавливается между двумя дескрипторами, если объем понятия нижестоящей лексической единицы входит в объем понятия вышестоящей лексической единицы. Связь «*часть − целое»* устанавливается между двумя лексическими единицами в том случае, если нижестоящая лексема определяет компонент объекта, обозначаемого вышестоящей лексемой. Если для одной лексической единицы можно указать более одной непосредственно вышестоящей лексической единицы, то в иерархических отношениях должны быть установлены связи со всеми лексическими единицами. При установлении иерархических отношений должны быть указаны связи со всеми нижестоящими лек-

сическими единицами независимо от аспекта деления. Аспект деления может быть указан в примечании при ссылке.

Ассоциативное отношение является объединением отношений, не входящих в иерархические отношения или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения «род − вид». В целях обеспечения ведения информационно-поискового тезауруса и индексирования документов для каждой ссылки, указывающей связь заглавной лексической единицы с другой лексической единицей, в другой лексической единице должна быть обратная ссылка. Если нецелесообразно использовать обратную ссылку при поиске информации, то следует применять технологическую обратную ссылку «сравни», обеспечивающую ведение информационно-поискового тезауруса. В термосистеме указывают все синонимы заглавной лексической единицы. Для лексических единиц, иерархически связанных отношением «часть − целое», допустимо давать иерархическую ссылку только от вышестоящего к нижестоящему или наоборот. Построение некоторого *семантического поля терминов* дает возможность применить принцип ассоциативно связанных цепочек при поиске информации.

В памяти головного мозга человека информация в целях компактной упаковки хранится в виде ассоциативных структур, а в целях экономии памяти пересекающиеся элементы и связи не дублируются, а имеют один и тот же след [8]. В этом случае информация сильно сжимается, и возникает эффект легкого вытаскивания по цепочке ассоциативно связанных структур. Воспроизведенный один элемент облегчает воспроизведение следующего. Таким образом, в процессе формирования модели окружающего мира ассоциативные нейронные цепочки пересекаясь, переходят из одной структуры головного мозга человека в другие. То есть в этом смысле они взаимно связаны, а попадание в определенный участок какой-либо структуры зависит от того, в каком месте мы находились до этого.

После выполнения терминологического, семантического и частотного анализа лексических единиц для оценки правильности выделенных терминов из текста эксперт предметной области может воспользоваться предметным указателем документа (при его наличии), который принимается за список терминов, выделенных автором. После проведения успешной оценки с помощью методов математической статистики производиться подсчет объема выделенной терминов. При этом вычисляется минимальный объем выборки по формуле:

$$N = \frac{Z}{\sigma^2 f},$$

где N – объем выборки;
f – относительная частота (не менее 3,19);
Z –коэффициент доверия;
σ – относительная ошибка.

Если количество терминов не обеспечивает достаточную надежность результатов, эксперту должен направляться запрос на предоставление системе дополнительных текстов с целью дальнейшего анализа и расширения сформированного тезауруса. Если терминов было выделено достаточно и проведена экспертная оценка, возможно сделать вывод об успешно сформированном тезаурусе, который в дальнейшем разрешается использовать при построении онтологии предметно-ориентированной области знаний.

Иерархия онтологии

Наряду с указанными элементами онтологии в нее также входят так называемые «экземпляры» или «инстанции». Экземпляры (инстанции) – это отдельные представители класса сущностей или явлений, т.е. конкретные элементы какого-либо понятия (например, экземпляром понятия «Компьютерная техника» будет «Системный блок») [6]. Составляющие онтологии подчиняются своеобразной иерархии. На нижнем уровне этой иерархической лестницы находятся экземпляры, конкретные индивиды, выше идут понятия, т.е. категории. На уровень выше располагаются отношения между этими понятиями, а обобщающей и связующей является ступень правил или аксиом. В рамках данной работы для

отнесения экземпляров к тому или иному понятию предлагается использовать метод k-ближайших соседей [9].

Метод k-ближайших соседей

Человек, сталкиваясь с новой задачей, использует свой жизненный опыт, вспоминает аналогичные ситуации, которые когда-то с ним происходили. О свойствах нового объекта можно судить, полагаясь на похожие знакомые наблюдения. Смысловое сходство объектов лежит в основе алгоритма k-ближайших соседей (k-nearest neighbor algorithm – KNN). Такой алгоритм способен выделять среди всех наблюдений k известных объектов (k-ближайших соседей), похожих на новый неизвестный ранее объект [9]. На основе классов ближайших соседей выносится решение касательно нового объекта. Важной задачей данного алгоритма является подбор коэффициента k – количество записей, которые будут считаться похожими. Отметим, что алгоритм k-ближайших соседей широко применяется в интеллектуальном анализе данных.

В методе k-ближайших соседей для классификации нового наблюдения X_{N+1} проводится упорядочивание исходные элементов выборки по какой-либо метрике. При этом определяется не один ближайший сосед, а группа наблюдений, наиболее близких к X_{N+1}, причем каждый из соседей имеет равный вес. Число соседей k является настраиваемым параметром метода на стадии обучения, или задаваемым экспертом. Решение об отнесении X_{N+1} к классу Q_k, где k = (1, …, K) принимается путем голосования его k-ближайших соседей с помощью простого подсчета голосов. Если более половины k-ближайших соседей принадлежат классу Q_k, то X_{N+1} также относится к этому классу. Таким образом, при использовании правила k-ближайших соседей строится гиперсфера объема V с центром в точке, соответствующей новому (нераспознанному) наблюдению X_{N+1}. Это наблюдение относится к тому классу, к которому принадлежит большинство элементов, оказавшихся внутри гиперсферы. Алгоритм k-ближайших соседей устойчив к аномальным выбросам, так как вероятность попадания такой записи в число k-ближайших соседей мала. Если же это произошло, то влияние на голосование (особенно взвешенное) (при k > 2), скорее всего, будет незначительным, а, следовательно, малым будет и влияние на итог классификации.

Аксиоматика для выделенных терминов

В рамках искусственного интеллекта можно описать онтологию программы, определив множество объектов, связав их с описаниями, а также введя формальные аксиомы, которые ограничивают интерпретацию и совместное употребление этих терминов [6]. Аксиомы онтологии представлены в форме утверждений и правил, объединение которых вместе определяет конкретную проблемно-ориентированную область. Аксиомы задают условия, определяющие взаимосвязь между понятиями и их отношений и выражают очевидные утверждения, связывающие понятия и отношения. Под аксиомой подразумевается утверждение (правило), вводимое в онтологию в готовом виде, из которого могут быть выведены другие аксиоматические утверждения (правила). Благодаря аксиомам становится возможным выразить информацию, которая не может быть представлена в онтологии посредством построения иерархии понятий и установки различных отношений между понятиями. Ниже приведены примеры аксиом:

1. «Если x является мужем y, то x и y женаты» или (∀x, y) муж (x, y) → женат (x, y);
2. «Если x человек, то количество биологических родителей равно 2» или ∀x ∈ человек → кол-во биологические родители = 2;
3. «Если x – это свойство агрегатного состояния вещества, то оно может выражаться только одним из трех состояний (твердое, жидкое или газообразное)» или ∀x ∈ агрегатное состояние вещества → агрегатное состояние (h, w, g).

Из приведенных примеров следует, что на основании аксиом посредством метода k-ближайших соседей онтология может автоматически добавлять новую информацию о понятиях, экземплярах и отношениях между ними. Аксиомы могут представлять собой понятийные или числовые ограничения, накладываемые на какие-либо отношения, делающие возможным выведение умозаключений [6]. В заключении настоящего раздела отме-

тим: авторами проведен анализ и описание процессов построения различных категорий предметно-ориентированной онтологии. Структура онтологии представляется набором элементов – понятий, отношений, аксиом и отдельных экземпляров.

Заключение

Таким образом, в статье рассмотрены основные аспекты построения предметно-ориентированной онтологии, позволяющие создавать такие онтологии в автоматическом режиме. При формировании онтологии формируется семантическая сеть понятий, которую в дальнейшем можно использовать в поисковых системах сети Интернет для выдачи пользователю пертинентных запросу результатов. По ходу изложения приведены практические рекомендации и описания методов формирования новых понятий и отношений. При построении онтологии используются косвенные аналогии подобного рода процессов, реализуемых головным мозгом человека, т.е. бионический подход.

Литература:

1. *Михеев А.С.* Когнитивная система экстрагирования концептуальных знаний из научно-технических текстов: автореф. / науч. рук. В.В. Нечаев. М.: МИРЭА, 1990. 23 с.

2. *Ашманов И., Иванов А*. Оптимизация и продвижение сайтов в поисковых системах. СПб.: Питер, 2008. 400 с.

3. *Соссюр Ф., де*. Курс общей лингвистики / под ред. и с примеч. Р.И. Шор. 3-е изд., стер. М.: КомКнига, 2006. 256 с.

4. *Белоногов Г.Г.* Теоретические проблемы информатики. Т. 2. Семантические проблемы информатики / под общ. ред. К.И. Курбакова. М.: РЭА им. Г.В. Плеханова. 2008. 342 с.

5. *Шемакин Ю.И., Ломако Е.И.* Основы системантики. М.: Финансы и статистика, 2009. 401 с.

6. *Константинова Н.С.* Онтологии, как системы хранения знаний / Н.С. Константинова, О.А. Митрофанова. СПГУ, 2006. URL: http://window.edu.ru/window catalog/files/r58795/68352e2-st08.pdf

7. *Азарова И.В.* Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции. Диалог. 2003 (Протвино, 11–16 июня 2003 г.). М., 2003. С. 51–55.

8. *Синицын Е.* Теория творчества. Структурный анализ мышления. Теория интегрированного обучения. Новосибирск: НГАХА, 2001. 440 с.

9. *Kozma L.* K-algorithm Nearest Neighbours A, Helsinki University of Technology, 2008. URL: http://www.lkozma.net/knn2.pdf.

Analisis of the methods of semantic search of information resources

***Valentin Viktorovich Nechayev**, Dr. of physical and mathematical Sciences (UMACC), the candidate technology. Sciences, Professor, head of the Department, Moscow state technical University of Radioengineering, electronics and automation*

***C.M. Trofimenko**, graduate student, Moscow state technical University of Radioengineering, electronics and automation*

This article describes the technique of semantic pertinence selection of information resources on the basis bionic approach and ontologies. It is based on the method of semantic analysis of lexical units, methods of statistical analysis and data mining technology. This article describes the biological characteristics of mental activity of the human brain. The article focused on the implementation problems of automation process of improving the competence of an expert in a subject-oriented field of knowledge.

Keywords: semantic pertinence selection of information resources bionic approach, pertinent information resources, semantic information resources, indicator of the relevance, indicator of the pertinence.