Раздел I. Анализ данных и управление знаниями

УДК 002.53:004.89 DOI 10.18522/2311-3103-2016-7-518

Ю.А. Кравченко

ЗАДАЧИ СЕМАНТИЧЕСКОГО ПОИСКА, КЛАССИФИКАЦИИ, СТРУКТУРИЗАЦИИ И ИНТЕГРАЦИИ ИНФОРМАЦИИ В КОНТЕКСТЕ ПРОБЛЕМ УПРАВЛЕНИЯ ЗНАНИЯМИ*

Статья посвящена постановке задач и разработке способов решения задач семантического поиска, классификации, структуризации и интеграции знаний применительно к проблеме управления знаниями. Управление информационными потоками рассматривается как совокупность процессов систематического приобретения, синтеза, обмена и использования знаний. Целью работы является модификация методов решения задач семантического поиска и структуризации применительно к проблеме управления знаниями. Научная новизна представлена модифицированным методом семантического поиска знаний, графовой моделью структуризации знаний на основе оценки наличия существенных признаков и абстрактной моделью представления семантической сети с многоуровневой декомпозицией междисциплинарных связей различных предметных областей. Проблема автоматизации управления знаниями, как непрерывного процесса проверки знаний для выявления закономерностей с целью создания и удовлетворения спроса на новые знания, напрямую связана с решением задач семантического поиска, классификации, структуризации и интеграции информации. Возможности современных информационных систем ограничены эффективным решением задач хранения и передачи информации. Одной из основных научных задач в сфере информационных технологий на сегодняшний день является разработка механизмов анализа и обработки информации в гетерогенных источниках с целью наделения информационных систем способностями логического анализа информации и генерации выводов, которые станут основой выполнения процедур накопления и обработки знаний. Так как, по мнению ученых, решением актуальной проблемы информационного переполнения станет переход от хранения и обработки данных к накоплению и обработке знаний. Несмотря на выраженную специфику предметных областей, онтологию необходимо строить как цепочку взаимосвязанных процессов, что позволит получить интегрированный характер интеллектуальной системы управления знаниями. На этапе идентификации областей знаний необходимо в первую очередь определить множество исследуемых характеристик. Далее требуется выбрать источники априорной информации и приступить к формированию базы знаний и хранилищ данных, которые впоследствии позволят задать отношения между категориями знаний. В качестве источников знаний легче всего подсоединяются базы данных оперативных информационных систем через механизм создания информационных хранилищ. Интеграция знаний из различных источников может проводиться на основе онтологии, требования к разработке которой будут находиться в заранее сформированных спецификациях.

Семантический поиск; онтологии; классификация; структуризация; интеграция; системы управления знаниями; информационные процессы; поддержка принятия решений.

* Работа выполнена при финансовой поддержке РФФИ (проект № 14-07-00841).

Yu.A. Kravchenko

INFORMATION'S SEMANTIC SEARCH, CLASSIFICATION, STRUCTURING AND INTEGRATION OBJECTIVES IN THE KNOWLEDGE MANAGEMENT CONTEXT PROBLEMS

The article is devoted to setting objectives and developing ways to solve problems of knowledge semantic search, classification, structuring and integration in relation to knowledge management issue. Information management is viewed as a set of processes systematic acquisition, synthesis and sharing of knowledge. The aim is to modify methods of solving problems of semantic search and structuring in relation to knowledge management issue. Scientific novelty is represented by the modified method of semantic knowledge discovery, graph model of structuring knowledge based on an assessment of material characteristics and abstract representation of a semantic network model with multi-level decomposition of different subject areas' interdisciplinary connections. Knowledge management automation problem as continuous knowledge verification process to identify patterns in order to create and meet demand for new knowledge, directly related to the solution of information's semantic search, classification, structuring and integration problems. The possibilities of modern information systems are limited to an effective solution of information's storage and transmission problems. One of the major scientific challenges in the field of information technology today is the development of information analysis and processing mechanisms in heterogeneous sources with the aim of empowering information systems logical analysis of information capabilities and generate conclusions that will form the basis of knowledge accumulation and processing execution procedures. Since, according to scientists, the decision of information overflow actual problem will shift from data storage and processing to the knowledge accumulation and processing. Despite the pronounced specificity of subject areas, ontology should be built as a chain of interrelated processes that will provide integrated nature of intellectual knowledge management system. At the stage of identifying the areas of expertise necessary to first define a set of study characteristics. Next, you need to choose a priori information sources and begin to form a knowledge base and data warehouse, which later will set the relationship between the categories of knowledge. As a source of knowledge operational information systems' database is most easily connected through a mechanism for creating data warehouses. Integrating knowledge from different sources may be based on the ontology requirements for the development of which will be a pre-formed sheets.

Semantic search; ontology; classification; structuring; integration; knowledge management systems; information processes; decision support.

Введение. Проблема информационного переполнения возникает в среде с интенсивным обменом и информационными потоками. Одним из следствий тенденции к децентрализации информационных ресурсов является постепенное усиление информационной активности субъектов информационного поиска и обмена. В этих условиях сеть гетерогенных информационных ресурсов, задействованная в обмене информацией, неизбежно проходит через "точку роста", в которой привлекательность результатов накопления и обработки знаний падает, т.к. пользователи начинают испытывать трудности информационного переполнения.

Будем утверждать, что системы управления знаниями являются развитием концепции информационных систем (достигших определенных технологических высот в решении задач эффективного хранения, обработки и предоставления информации пользователю по регламенту или формализованному запросу) и предполагают решение задач увеличения объема и повышения уровня использования имеющихся знаний, а также генерации идей для создания новых знаний.

Гипотезой данной разработки является предположение о том, что семантическое идентифицирование ключевой информации может быть эффективно проведено на основе построения семантического дерева таксономии понятий, как систематизации сложноорганизованных областей действительности и знания, имеющих иерархическое строение, с целью определения и упорядочивания терминов и

их синонимов. Сложность проблемы построения семантического дерева таксономии состоит в определении эффективных способов реализации процедур семантического поиска, классификации, структуризации и интеграции информации. Данная проблема является актуальной для любой инженерно-исследовательской деятельности, направленной на извлечение, накопление, анализ и конкретизацию имеющихся научных знаний применительно к определенной инженерной задаче.

1. Модифицированный метод семантического поиска знаний. В наиболее обобщенном понимании *информационный поиск* – сложная проблема, состоящая из двух подзадач:

1) сопоставления представления пользователя о нужных ему знаниях с содержанием доступных распределенных неоднородных источников знаний;

2) построения на основе этого сопоставления информационного объекта (ИО) с конечным набором свойств, значения которых извлекаются из этих источников.

Частными случаями такой задачи можно считать распознавание образов, понимание речи или изображений, семантический анализ, перевод естественно-языковых текстов, исследование и композицию сервисов. Такая задача является сложноформализуемой. На практике целесообразно рассматривать относительно упрощенные подзадачи семантического поиска, в которых заранее накладывается ряд ограничений на потребности пользователя и на те информационные объекты, которые должны быть сгенерированы.

На основе анализа существующих информационно-поисковых систем можно выделить основные тенденции усовершенствования систем информационного поиска [1–3]:

- от формального – к семантическому;
- от унифицированного – к персонифицированному;
- от индивидуального – к совместному;
- от закрытого – к управляемому.

Семантический поиск – это информационный поиск, в котором такое сопоставление и построение информационного объекта выполняются на семантическом уровне, т.е. с использованием знаний. Семантический поиск базируется на достижениях в области искусственного интеллекта, в частности – общей теории представления и обработки знаний, распознавания образов и логического вывода, методах математической статистики и социопсихологии.

Основное отличие семантического поиска от традиционного – использование *знаний* об объекте поиска, источниках и предметной области поиска, а также возможность обнаружения не данных, а знаний.

Информацию, представленную в структурированном виде (онтологии, метаописания, семантически размеченные файлы т.п.), обрабатывать проще, но в таком виде представлено относительно мало информации. Поэтому поиск среди таких источников часто не позволяет находить нужные сведения.

Следует отметить, что предметом поиска в большинстве случаев является не какой-то конкретный документ, а некие сведения о каком-либо объекте – реальном либо виртуальном. При этом у пользователя присутствует часть информации об этом объекте, о его свойствах и структуре, а недостающие сведения он рассчитывает получить в результате поиска.

Информационный объект – это модель какого-либо реального или виртуального объекта предметной области (предмета, существа, события, процесса и т.д.) в информационном пространстве, которая определяет структуру, атрибуты, ограничения целостности и поведение этого объекта. При семантическом поиске пользователь может указать, экземпляром какого класса является тот информационный объект (или группа информационных объектов), информация о котором ему нужна.

Как основу для представления структуры ИО можно использовать классы соответствующей онтологии, а из источника знаний извлекать сведения для создания экземпляров ИО. При семантическом поиске ИО может иметь заранее заданную сложную структуру, формализованную в виде класса соответствующей онтологии. Примеры ИО – организация, учебное заведение, человек-эксперт, Web-сервис, аналитическая модель бизнес-процесса. Наличие таксономии ИО позволяет пользователю более формально указать, какую именно информацию ему нужно найти и какими сведениями о ней он обладает. При этом возникает проблема поиска онтологии, которая отображает структуру ИО, знания о которых необходимы пользователю.

В данной разработке будем использовать следующие обозначения компонентов онтологии: онтология O представляет собой знаковую систему $O = <P,V,R,C>$, где P – множество понятий (классов); V – множество экземпляров понятий; R – множество предикатов – типов отношений; C – множество отношений, которые задают следующие виды связи между сущностями:

1. Частичный порядок на множествах P и R, задающий отношения *is-a* – «подкласс-суперкласс» – «выше-ниже».

2. Отношение между понятиями, которое представляет собой триплет вида $<p_1 – r_1 – p_2>$, где $p_1, p_2 \in P$; $r_1 \in R$.

3. Отношение между экземплярами, которое представляет собой триплет вида $<v_1 – r_1 – v_2>$, где $v_1, v_2 \in V$; $r_1 \in R$.

4. Отношение между предикатами, которое представляет собой триплет вида $<r_1 – r_i – r_2>$, где $r_1, r_2, r_i \in R$.

Оценкой близости между элементом знания и запросом будем считать числовое значение, которое выражает степень сходства между ними, оценка близости называется оценкой семантической близости, если и только если она определена на основе семантики документов и запросов [4–9].

Идея модифицированного метода семантического поиска заключается в описании поисковых запросов в виде набора триплетов. Пусть имеется запрос q, состоящий из набора триплетов $T(q)$. В таком случае результатом поиска в источнике знаний будет набор элементов знания $E=\{e_i \mid i \in [1,k]\}$, где k – количество элементов знания e_i, являющихся результатом поиска. Причем, семантические метаданные набора элементов знания $T(e)$ должны удовлетворять следующему условию семантической близости $sim(T(q),T(e))$ с описанием запроса $T(q)$: $sim(e,q)=sim(T(q),T(e))>\varepsilon$, где $sim(e,q)$ близость запроса q и элемента знания e, а ε – установленное пороговое значение релевантности. Результаты поиска ранжируются по значениям их семантической близости к запросу.

Семантический поиск позволяет отбирать близкие по общему контексту документы, даже если они принадлежат к разным предметным областям. Тем самым можно накапливать или извлекать, обновлять и обобщать знания, рассредоточенные в различных предметных областях на основе выявления и использования их междисциплинарных отношений.

2. Проблема классификации в контексте задач управления знаниями. Представим математическую постановку задачи классификации знаний. Пусть X – множество описаний элементов знаний, Y – множество наименований классов. Существует неизвестная целевая зависимость – отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$.

Представим вероятностную постановку задачи классификации знаний, которая считается более общей. Предполагается, что множество пар «элемент знания, класс» $X \times Y$ является вероятностным пространством с неизвестной веро-

ятностной мерой P. Имеется конечная обучающая выборка наблюдений $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, сгенерированная согласно вероятностной мере P. Требуется построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$.

Рассмотрим проблему классификации в контексте предшествующих и последующих задач циклического сценария управления знаниями (рис. 1).

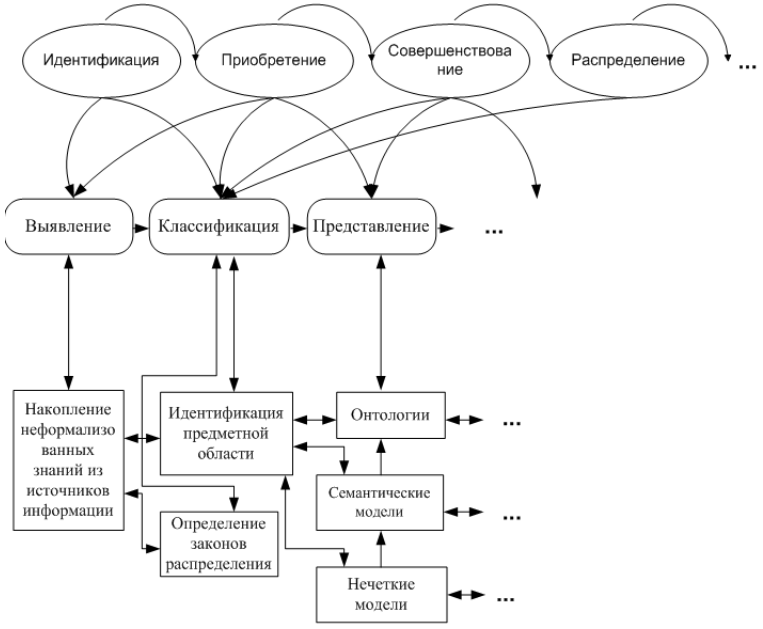

Рис. 1. «Окружение» проблемы классификации в контексте задач управления знаниями

Процессы управления знаниями базируются на следующих основных принципах: *идентификации; приобретения; совершенствования; распределения; использования и сохранения.* При этом, непосредственно связанными с задачей классификации при управлении знаниями являются принципы: *идентификации; приобретения; совершенствования и распределения.* Самой же задаче *классификации* знаний предшествует задача их *выявления*, которая отвечает за извлечение неформализованных знаний из разнородных источников информации, а последующей является – задача *представления*, т.е. *формализации* знаний на основе создания онтологии, семантических, нечетких и других моделей.

Приведем основные виды классификации знаний:

Классификация знаний по Махлуп:

- практическое знание;
- интеллектуальное знание (включая, научное, гуманитарное и культурное);
- бытовое знание;
- духовное знание;
- нежелательное знание.

Классификация знаний по К. Вииг:

- идеалистические знания (знания о цели, мировоззрении, системе понятий);
- систематические знания (знание систем, схем, методов; используются для глубокого анализа причин, формулировки новых подходов, генерирования альтернатив и принятия решений);

- практические знания (умение принимать решение);
- автоматические знания (применяются в ситуациях, не требующих логических рассуждений).

Классификация знаний на личностном уровне:

- знания как результат познания (знание «что» – владение основами предметной области);
- специальная квалификация (знание «как» – практическое выполнение, требующее больше знаний, чем можно извлечь из источников);
- системное понимание (знание «почему» – знание причин, следствий, симптомов);
- самомотивируемое творчество (понимание «зачем» – желание и мотивация успеха).

Классификация организационных знаний по Ф. Блеклеру:

- интеллектуальное знание, зависящее от навыков абстрактного мышления и познавательных способностей (знание «что»);
- воплощенное знание, ориентированное на действие и, как правило, лишь частично явное (знание «как»);
- запечатленное в культуре знание, относящееся к процессу достижения общего понимания;
- встроенное знание, содержащееся в системных процедурах;
- закодированное знание, передаваемое через знаки и символы;

Классификация организационного знания Н. Тонака и Х. Такеучи:

- явное (систематизированное) знание. Может выражаться в словах и числах и легко может передаваться и обмениваться в виде точных данных, научных формул, упорядоченных процедур или универсальных принципов;
- скрытое (несистематизированное и неформализованное) знание – это нечто трудновыявляемое и трудновыражаемое, является личным, обусловленным конкретным контекстом, а также слабоформализуемым и передаваемым другим людям. К ним относятся: озарение, интуиция и предчувствия [5, 10–17].

Говоря об *идентификации* знаний, будем понимать первый этап определения системы междисциплинарных связей между элементами знаний используемых предметных областей. Основной целью данного этапа является выявление тех знаний, которые можно рассматривать как уже приобретенные. В первую очередь имеющиеся данные делят на явные и неявные (неформализованные). Перевод неявных знаний в явные происходит посредством процедуры *экстернализации*, а обработка и интеграция элементов формализованного явного знания проводится на основе операции *комбинирования*.

Следующими моментами идентификации знаний являются определение степеней *детализации* знаний и их *значимости*. Определение значимости знаний напрямую связано с процессом *приобретения* знаний, так как задает основные направления поиска знаний в областях значимых тематик предметных областей. *Приобретение* знаний – выбор источников получения знаний; оценка полезности и отбор знаний; обеспечение соответствия между притоком знания и потребности в нем. *Приобретение* знаний реализуется с помощью двух функций: получения информации извне и ее систематизации.

Совершенствование знаний происходит в процессе семантического поиска в неоднородных распределенных источниках знаний и энциклопедических справочных системах на основе онтологических, case- моделей и моделей успешных прецедентов поисковых запросов. Совершенствование знаний можно рассматривать

как подзадачу *создания* новых знаний – обеспечения условий для творчества, генерации идей, обмена и интеграции знаний. *Распределение* полученных и формализованных знаний происходит на основе поискового запроса и модели пользовательских предпочтений.

3. Разработка модели структуризации знаний на основе оценки наличия существенных признаков. Сложность проблемы накопления и обработки знаний состоит в оценке наличия или отсутствия существенных признаков системности в структуре элементов знания из различных предметных областей и определении отношений между ними. Гипотеза разработки основана на предположении, что данная проблема имеет комплексный характер, и на одном из этапов может быть решена при помощи структуризации знаний.

Структуризация знаний – процесс деления элементов знаний на устойчивые группы и подгруппы. Опишем основные принципы структуризации знаний:

1. Знания должны быть поделены на группы и подгруппы в соответствии с определенными системно значимыми признаками.
2. Сформированные группы и подгруппы должны быть логично связаны и выстроены в необходимом порядке (по важности, по времени и т.п.).

Лавинообразный рост используемых объемов знаний значительно увеличил энтропию как меру информационного хаоса. Размеры сохраняемой и копируемой информации уже давно на порядки превышают когнитивные возможности человека. Находясь в режиме такого информационного обмена, человек подвергается угрозе информационной перегрузки. Поэтому перед каждым субъектом, включенным в процесс накопления и обработки знаний, стоят две актуальные задачи:

1) исключить информационный шум из анализируемых источников;
2) противостоять манипуляциям через информацию.

Вариантом решения данных задач может стать использование структуризации знаний [6, 7, 18–20].

Введем ряд основных определений необходимых для дальнейшего рассмотрения выбранной проблемы. Отображение в мозгу человека основных признаков объектов и событий назовем мышлением. Коренное, наиболее важное свойство объекта или события (процесса), без которого теряется смысл, будем считать существенным (системно значимым) признаком. Под знанием в общем смысле будем понимать субъективное отражение реальности в форме понятий, определений и отношений между ними [7, 21, 22]. Определим гипотезу, согласно которой понятие имеет характер всеобщего отображения существенных признаков объектов и событий реального мира, тогда как представление – всего лишь наглядный образ объекта или события, складывающийся из несущественных признаков и имеющий индивидуальный характер. Определением понятия будем считать логическое действие, позволяющее раскрыть содержание понятия. Определение содержит в себе только существенные (системные) признаки, которые отделяют одно понятие от другого.

Система знаний строится на основе задания отношений в совокупности элементов знаний одной предметной области, или на более высоком (междисциплинарном) уровне – между различными предметными областями.

Сформулируем постановку задачи структуризации знаний. Предположим, что все системно значимые признаки элементов знания из определенной предметной области можно разбить на m классов. Тогда можно сформировать множество необходимых признаков системной значимости $F = \{F_1 \cup F_2 \cup ... \cup F_m\}$.

$$F_1 = \{f_{11}, f_{12}, ..., f_{1(i-1)}, f_{1i}\},$$

где $f_{11}, f_{12}, ..., f_{1(i-1)}, f_{1i}$ – элементы множества F_1, задающие 1-ый класс системно значимых признаков для элементов знания некоторой предметной области;

$$F_2 = \{f_{21}, f_{22}, ..., f_{2(j-1)}, f_{2j}\},$$

где f_{21}, f_{22},…,$f_{2(j-1)}$, f_{2j} – элементы множества F_2, задающие 2-ой класс системно значимых признаков для элементов знания некоторой предметной области;

$$F_m = \{f_{m1}, f_{m2}, \ldots, f_{m(k-1)}, f_{mk}\},$$

где f_{m1}, f_{m2}, …,$f_{m(k-1)}$, f_{mk} – элементы множества F_3, задающие m класс системно значимых признаков для элементов знания некоторой предметной области.

Подобным образом задаем для каждого анализируемого элемента знания q_z $(z=1\ldots n)$ множество имеющихся у него системно значимых признаков $Q_z = \{Q_{11} \cup Q_{12} \cup \ldots \cup Q_{nm}\}$,

где $Q_{11} \subset F_1, Q_{12} \subset F_2, Q_{nm} \subset F_m$.

Тогда выражение определения соответствия элемента знания системно значимым требования предметной области можно представить в виде:

$$M_o = Q_z \cap F.$$

А целевая функция в таком случае примет вид:

$$M_o = F.$$

В иных случаях можно будет делать вывод о не полном соответствии элемента требованиям, выдвигаемым к системно значимым знаниям. Что в дальнейшем позволит либо предпринять действия по изменению элемента знания, либо – исключить его полностью из результата поиска [8, 9].

Для формального представления модели описанной постановки задачи будем использовать двудольный ориентированный граф:

$$G=<P,F,I,L>,$$

где P – множество элементов знания; F– множество необходимых признаков системной значимости; I – дуги, указывающие на характеристики, присущие конкретному элементу знания; L – дуги, указывающие на желаемые характеристики системной значимости для соответствующей предметной области.

Проиллюстрируем данную модель структуризации знаний на абстрактном примере (рис. 2).

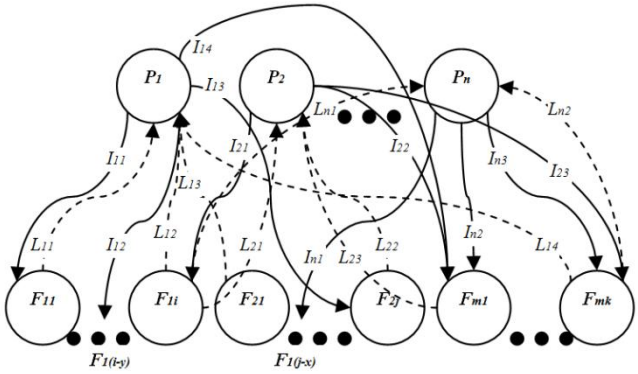

Рис. 2. Графовая модель структуризации знаний на основе оценки наличия существенных признаков

Данный пример можно пояснить следующим образом:

1) элемент знания P_1 обладает признаками: $\{F_{11}, F_{1(i-y)}, F_{2j}, F_{m1}\}$, (дуги: I_{11}, I_{12}, I_{13}, I_{14}), где y – неизвестное целочисленное значение. При этом, существенными для рассматриваемой предметной области являются признаки: $\{F_{11}, F_{1i}, F_{21}, F_{mk}\}$, (дуги: L_{11}, L_{12}, L_{13}, L_{14}). Тогда выражение определения соот-

ветствия элемента знания системно значимым требованиям предметной области примет вид: $M_1 = I_1 \cap L_1 = F_{11,}$ а цель поиска или доработки – $S_1 = L_1 \setminus M_1 = \{F_{1i}, F_{21}, F_{mk}\}$;

2) элемент знания P_2 обладает признаками: $\{F_{1i}, F_{m1}, F_{mk}\}$ (дуги: I_{21}, I_{22}, I_{23}). При этом, существенными для рассматриваемой предметной области являются признаки: $\{F_{1i}, F_{2j}, F_{m1}\}$, (дуги: L_{21}, L_{22}, L_{23}). Тогда выражение определения соответствия элемента знания системно значимым требованиям предметной области примет вид: $M_2 = I_2 \cap L_2 = \{F_{1i}, F_{m1}\}$, а цель поиска или доработки – $S_2 = L_2 \setminus M_2 = F_{2j}$;

3) элемент знания P_n обладает признаками: $\{F_{1(j-x)}, F_{m1}, F_{mk}\}$ (дуги: I_{n1}, I_{n2}, I_{n3}). При этом, существенными для рассматриваемой предметной области являются признаки: $\{F_{1i}, F_{mk}\}$, (дуги: L_{n1}, L_{n2}). Тогда выражение определения соответствия элемента знания системно значимым требованиям предметной области примет вид: $M_n = I_n \cap L_n = F_{mk}$, а цель поиска или доработки – $S_n = L_n \setminus M_n = F_{1i}$.

Проблема накопления знаний предметных областей и на междисциплинарном уровне является в настоящее время сложной и трудоемкой. Неконтролируемый рост окружающих человека информационных потоков ведет к неоправданно высоким затратам, а чаще – к нехватке ресурсов на обработку входного потока информации, которая только после надлежащей переработки может стать знанием. Одним из этапов решения данной проблемы является задача структуризации знаний, однозначно устанавливающая отношения на наборах понятий и определений.

Основной идеей структуризации знаний на основе оценки наличия существенных признаков является реализация возможности построения многоуровневой сети связанных между собой понятий. Все понятия связываются между собой через определения, причем вышележащие понятия могут быть определены, если заданы отношения на понятиях, лежащих на более низком уровне.

4. Постановка задачи интеграции знаний. Задача интеграции знаний связана с целым рядом подзадач разработки: баз знаний, содержащих блоки правил принятия решений и прецедентов; множеств объектных, онтологических, нечетких, семантических и аналитических моделей, реализующих процессы принятия решений; модулей выбора моделей и формирования решений на основе базы знаний, математического и имитационного моделирования.

Для создания мультидисциплинарных знаний необходима системная интеграция уже разработанных онтологий различных предметных областей. Сформулируем постановку задачи системной интеграции множества онтологий:

$$O^{\cup} = \bigcup_{i} O_i, i = \overline{1, N},$$

где O_i – онтограф; i – номер предметной области; N – количество предметных областей.

Под объединением будем понимать концептуальную системную интеграцию исходных онтологических графов (ОГ) и их взаимосвязи. Объем знаний V в предметных областях будем оценивать через параметры их формально-онтологических представлений. В частности, при представлении онтографом без учета типов отношений и сложности функций интерпретации величина V примет значение числа вершин ОГ. При простой древовидной структуре выражение определения объема знаний в предметной области примет следующий вид:

$$V = \sum_{i} \sum_{p} \sum_{d} O_i \cdot Z_{p,d},$$

где $Z_{p,d}$ – степень инцидентности вершины под номером d; $p = \overline{1, P}$ – количество уровней ОГ; $d = \overline{1, D_p}$ – номер вершины на p-ом уровне ОГ.

Учёт типов отношений и сложность функций интерпретации приводит к ОГ со взвешенными вершинами и ребрами. Выражение определения объема знаний в этом случае примет вид:

$$V = \sum_{i} \sum_{p} O_i \cdot \left(\gamma_d + \sum_{j} \delta_{d,j} \right),$$

где γ_d и $\delta_{d,j}$ – значения весовых функций отношений и интерпретации.

Основные результаты проведенного исследования предполагают выполнение дальнейших работ по разработке формальной методологии проектирования онтологии предметной области, алгоритмов и процедур системной интеграции знаний с применением методов и алгоритмов биоинспирированного поиска.

5. Экспериментальные исследования. Семантические сети являются одной из моделей представления знаний. Будем использовать семантические сети в качестве структуры, которая будет хранить базу распределенных неоднородных источников знаний, разделенную по выбранным атрибутам на объекты, с четко заданной иерархией отношений. Использование семантических сетей в системах управления знаниями позволяет расширить потенциал информационных систем не только в области извлечения знаний, но и в области их обработки за счет задания системы отношений на множествах элементов знания.

Перед созданием семантической сети, будем проводить декомпозицию знаний ресурса на некоторое количество структурных единиц с целью повышения эффективности процедур обработки отношений между ними. Исследователь имеет возможность задать произвольное количество единиц декомпозиции, оптимальным для подобных информационных моделей считается количество структурных единиц равное шести [2]. Такая организация сети позволяет выделить уровни детализации исследуемого материала. Каждый узел сети имеет привязку к определенному объекту знаний (рис. 3).

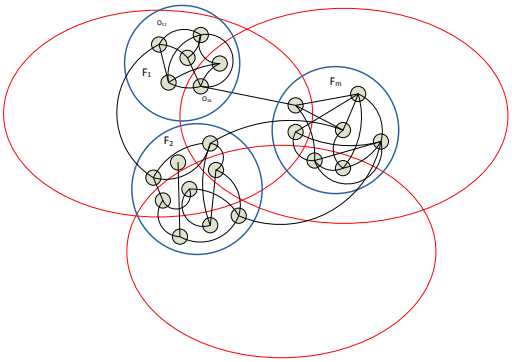

Рис. 3. Абстрактная модель семантической сети с несколькими уровнями декомпозиции

Первому уровню принадлежит только один понятийный узел сети. Он является корнем сети и обобщает все ее содержимое, являясь мета-знанием. Каждый новый уровень раскрывает содержание понятие предыдущего уровня, тем самым раскрывая для исследователя всю предметную область и обнаруживая наличие междисциплинарных связей с другими узлами сети.

Сеть включает в себя множество предметных областей $F = \{F_1, F_2, F_3 \dots F_z\}$, где $z = \overline{1,m}$, m – integer constant (количество предметных областей), причем, $F_z = \{O_{z1}, O_{z2}, O_{z3}, \dots, O_{zk_z}\}$, где k_z – integer constant (количество элементов (объек-

тов) знаний в F_z), причем, в случае задания отношений между объектами одной предметной области получим, что $\forall i,j \to i = \overline{1,k_z};\ j = \overline{1,k_z}; i \neq j\ \exists\ C_z[i][j] = n$, где n – количество связей между O_{zi} и O_{zj}, z – integer constant (номер исследуемой предметной области). В случае задания всех, в т.ч. междисциплинарных, отношений между объектами предметных областей получим, что $\forall a,b \to a = \overline{1,\sum_{z=1}^{m} k_z}; b = \overline{1,\sum_{z=1}^{m} k_z}; a \neq b\ \exists\ C_{interdisc}[a][b] = p$, где p – количество всех, в т.ч. междисциплинарных, связей между объектами сети.

Заключение. Данная работа развивает методы решения задач семантического поиска, классификации, структуризации и интеграции знаний применительно к проблеме управления знаниями. Данная проблема является актуальной для любой инженерно-исследовательской деятельности, направленной на извлечение, накопление, анализ и конкретизацию имеющихся научных знаний применительно к определенной инженерной задаче.

Управление информационными потоками рассматривается как совокупность процессов систематического приобретения, синтеза, обмена и использования знаний. Оценка достоверности, дающая исследователю аргументы и доказательства, стоящие за этими утверждениями, основана на использовании в управлении знаниями метаданных и метаутверждений в отличие от технологий управления информацией.

Конкретные научные результаты проявляются в систематизированной совокупности действий, которые необходимо предпринять, для решения задачи изучения и поиска отношений в распределенных источниках знаний с целью увеличения вероятности продуцирования нового знания.

В статье представлен модифицированный метод семантического поиска знаний, использующий онтологию в виде знаковой системы и поисковые запросы в виде наборов триплетов, позволяющих определить семантическую близость метаданных. Разработана графовая модель структуризации знаний на основе оценки наличия существенных признаков, позволяющая строить многоуровневую сеть связанных между собой понятий. Для проведения экспериментальных исследований предложена многоуровневая абстрактная модель семантической сети предметных областей. Данная модель позволяет на основе способа силовой релаксации определить «корень сети», являющийся «центром притяжения» и мета-знанием, максимально раскрывающим наличие междисциплинарных отношений между элементами знания из различных предметных областей.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Amerland D.* Google Semantic Search: Search Engine Optimization (SEO) Techniques That Gets Your Company More Traffic, Increases Brand Impact and Amplifies Your Online Presence. – Que Publishing, 2013. – 230 p.
2. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Decision Support Systems for Knowledge Management // Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P.123-130.
3. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Development of Distributed Information Systems: Ontological Approach // Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P. 113-122.
4. *Нгуен Б.Н., Тузовский А.Ф.* Обзор подходов семантического поиска // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – Т. 2, № 2. – С. 234-237.

5. *Guarino N., Welty C.A.* Towards a Metotodology for Ontology - Model Engineering // Proceeding of the ECOOP-2000 Workshop on Model Engineering (eds. by Bezivin J. and Ernst J.). – 2000. – Mode of access: http://www.metamodel.com/IWME00/articles/guarino.pdf.
6. *Gangeni A., Pisanelli D.M., Steve G.* An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies // Data & Knowledge Engineering. – 1999. – Vol. 31. – P. 183-220.
7. *Федоров Д. Ю.* Применение структуризации знаний для обеспечения информационной безопасности личности // Национальная безопасность и стратегическое планирование. – 2013. – № 2. – С. 40-43.
8. *Бова В.В.* Концептуальная модель представления знаний при построении интеллектуальных информационных систем // Известия ЮФУ. Технические науки. – 2014. – № 7 (156). – С. 109-117.
9. *Kravchenko Y.A., Kureichik V.V.* Knowledge management based on multi-agent simulation in informational systems // Conference proceedings. 8th IEEE International Conference "Application of Information and Communication Technologies – AICT 2014". – 15-17 October 2014, Astana, Kazakhstan. – P. 264-267.
10. *Тузовский А.Ф., Чириков С.В., Ямпольский В.З.* Системы управления знаниями (методы и технологии) / под общ. ред. В.З. Ямпольского. – Томск: Изд-во НТЛ, 2005. – 260 с.
11. *Курейчик В.М., Кажаров А.А.* Использование шаблонных решений в муравьиных алгоритмах // Известия ЮФУ. Технические науки. – 2013. – № 7 (144). – С. 11-17.
12. *Gladkov, L.A., Gladkova, N.V., Legebokov, A.A.* Organization of knowledge management based on hybrid intelligent methods // Advances in Intelligent Systems and Computing. – 2015. – Vol. 349. – P. 107-112.
13. *Dukkardt, A.N., Lezhebokov, A.A., Zaporozhets, D.* Informational system to support the design process of complex equipment based on the mechanism of manipulation and management for three-dimensional objects models // Advances in Intelligent Systems and Computing. – 2015. – Vol. 347. – P. 59-66.
14. *Курейчик В.М.* Особенности построения систем поддержки принятия решений // Известия ЮФУ. Технические науки. – 2012. – № 7 (132). – С. 92-98.
15. *Курейчик В.В., Родзин С.И.* О правилах представления решений в эволюционных алгоритмах // Известия ЮФУ. Технические науки. – 2010. – № 7 (108). – С. 13-21.
16. *Qing He, Xiu-Rong Zhao, Ping Luo, Zhong-Zhi Shi*. Combination methodologies of multi-agent hyper surface classifiers: design and implementation issues // Second international workshop, AIS-ADM 2007, Proceedings. – Springer Berlin Heidelberg, 2007. – P. 100-113.
17. *A.De Nicola, Missikoff M., Navigli R.* A software engineering approach to ontology building // Informatioh systems. – 2009. – Vol. 34. – P. 258-275.
18. *Guarino N., Oberle D., Staab S.* What is an Ontology // Handbook on Ontologies. – Springer, 2009. – P. 1-17.
19. *Yang X.-S.* A new metaheuristic sat-inspired algorithm // Nature Inspired Cooperative Strategies for Optimization (NISCO'2010), Berlin: Springer, 2010. – Vol. 284. – P. 65-74.
20. *Sarraipa J., et al.* Semantic Enrichment of Standard-based Electronic Catalogues // 13th IFAC Symposium on Information Control Problems in Manufacturing, 2009.
21. *Kerschberg L., Kim W., Scime A.* Personalizable semantic taxonomy-based search agent. USA: George Mason Intellectual Properties, INC (Fairfax, VA). – 2006.
22. *Kerschberg L., Jeong H., Kim W.* Emergent Semantic in Knowledge Sifter: An Evolutionary Search Agent based on Semantic Web Services // Journal on Data Semantics. – VI. LNCS. Vol. 4090. – Springer, Heidelberg, 2006. – P. 187-209.

REFERENCES

1. *Amerland D.* Google Semantic Search: Search Engine Optimization (SEO) Techniques That Gets Your Company More Traffic, Increases Brand Impact and Amplifies Your Online Presence. Que Publishing, 2013б 230 p.
2. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Decision Support Systems for Knowledge Management, *Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015).* Vol. 3. Springer International Publishing AG Switzerland, 2015, pp.123-130.

3. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Development of Distributed Information Systems: Ontological Approach, *Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015)*. Vol. 3. Springer International Publishing AG Switzerland, 2015, pp. 113-122.
4. *Nguen B.N., Tuzovskiy A.F.* Obzor podkhodov semanticheskogo poiska [Overview of the approaches for semantic search], *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki* [Reports of Tomsk state University of control systems and Radioelectronics], 2010, Vol. 2, No. 2, pp. 234-237.
5. *Guarino N., Welty C.A.* Towards a Metotodology for Ontology - Model Engineering, *Proceeding of the ECOOP-2000 Workshop on Model Engineering (eds. by Bezivin J. and Ernst J.)*, 2000. Mode of access: http://www.metamodel.com/IWME00/articles/guarino.pdf.
6. *Gangeni A., Pisanelli D.M., Steve G.* An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies, *Data & Knowledge Engineering*, 1999, Vol. 31, pp. 183-220.
7. *Fedorov D.Yu.* Primenenie strukturizatsii znaniy dlya obespecheniya informatsionnoy bezopasnosti lichnosti [The use of structuring knowledge to ensure personal information security], *Natsional'naya bezopasnost' i strategicheskoe planirovanie* [National Security and Strategic Planning], 2013, No. 2, pp. 40-43.
8. *Bova V.V.* Kontseptual'naya model' predstavleniya znaniy pri postroenii intellektu-al'nykh informatsionnykh sistem [Conceptual model of knowledge representation in the constructing intelligent information systems], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2014, No. 7 (156), pp. 109-117.
9. *Kravchenko Y.A., Kureichik V.V.* Knowledge management based on multi-agent simulation in informational systems, *Conference proceedings. 8th IEEE International Conference "Application of Information and Communication Technologies – AICT 2014". – 15-17 October 2014, Astana, Kazakhstan*, pp. 264-267.
10. *Tuzovskiy A.F., Chirikov S.V., Yampol'skiy V.Z.* Sistemy upravleniya znaniyami (metody i tekhnologii) [The knowledge management system (methods and techniques)], under ed. of V.Z. Yampol'skogo. Tomsk: Izd-vo NTL, 2005, 260 p.
11. *Kureychik V.M., Kazharov A.A.* Ispol'zovanie shablonnykh resheniy v murav'inykh algoritmakh [Template using for ant colony algorithms], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2013, No. 7 (144), pp. 11-17.
12. *Gladkov, L.A., Gladkova, N.V., Legebokov, A.A.* Organization of knowledge management based on hybrid intelligent methods, *Advances in Intelligent Systems and Computing*, 2015, Vol. 349, pp. 107-112.
13. *Dukkardt, A.N., Lezhebokov, A.A., Zaporozhets, D.* Informational system to support the design process of complex equipment based on the mechanism of manipulation and management for three-dimensional objects models, *Advances in Intelligent Systems and Computing*, 2015, Vol. 347, pp. 59-66.
14. *Kureychik V.M.* Osobennosti postroeniya sistem podderzhki prinyatiya resheniy [Features of decision making support system design], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2012, No. 7 (132), pp. 92-98.
15. *Kureychik V.V., Rodzin S.I.* O pravilakh predstavleniya resheniy v evolyutsionnykh algoritmakh [On the rules for the submission decisions in evolutionary algorithm], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2010, No. 7 (108), pp. 13-21.
16. *Qing He, Xiu-Rong Zhao, Ping Luo, Zhong-Zhi Shi*. Combination methodologies of multi-agent hyper surface classifiers: design and implementation issues, *Second international workshop, AIS-ADM 2007, Proceedings*. Springer Berlin Heidelberg, 2007, pp. 100-113.
17. *A.De Nicola, Missikoff M., Navigli R.* A software engineering approach to ontology building, *Informatioh systems*, 2009, Vol. 34, pp. 258-275.
18. *Guarino N., Oberle D., Staab S.* What is an Ontology, *Handbook on Ontologies*. Springer, 2009, pp. 1-17.
19. *Yang X.-S.* A new metaheuristic sat-inspired algorithm, *Nature Inspired Cooperative Strategies for Optimization (NISCO'2010), Berlin: Springer, 2010,* Vol. 284, pp. 65-74.
20. *Sarraipa J., et al.* Semantic Enrichment of Standard-based Electronic Catalogues, *13th IFAC Symposium on Information Control Problems in Manufacturing, 2009.*

21. *Kerschberg L., Kim W., Scime A.* Personalizable semantic taxonomy-based search agent. USA: George Mason Intellectual Properties, INC (Fairfax, VA), 2006.
22. *Kerschberg L., Jeong H., Kim W.* Emergent Semantic in Knowledge Sifter: An Evolutionary Search Agent based on Semantic Web Services, *Journal on Data Semantics*. VI. LNCS. Vol. 4090. Springer, Heidelberg, 2006, pp. 187-209.

Статью рекомендовал к опубликованию д.т.н., профессор М.М. Ошхунов.

Кравченко Юрий Алексеевич – Южный федеральный университет; e-mail: yakravchenko@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Kravchenko Yury Alekseevich – Southern Federal University; e-mail: yakravchenko@sfedu.ru; 44, Nekrasovskiy lane, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; associate professor.

УДК 002.53 DOI 10.18522/2311-3103-2016-7-1828

С.Б. Картиев, В.М. Курейчик

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМА РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ДЛЯ ОСУЩЕСТВЛЕНИЯ ВОПРОСНО-ОТВЕТНОГО ПОИСКА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ ПРОГНОЗИРОВАНИЯ*

Данная статья посвящена проблемам построения модуля вопросно-ответного поиска неструктурированной информации в информационно-аналитической системе прогнозирования относительно коллекции исходных состояний сложной технической системы. Современные информационно-поисковые системы основаны на принципах поиска по ключевым словам. Данный вид поиска предоставляет на выходе коллекцию веб-страниц, которая по вероятности может содержать нужный материал для пользователя. В статье предлагается подход приведения задачи кластеризации к оптимизационной задаче и ее решения с использованием метаэвристических методов. Дано введение в традиционные методы кластеризации, приведены их преимущества и недостатки. Кластеризация является частным случаем обучением без учителя. Отсутствие учителя предусматривает то что в системе нет эксперта, который может присваивать документам классы. Приведено описание основной модели вычислений и средств хранения данных, применяемых в разработанной системе. Предложен подход к построению подобных модулей и их математическое обеспечение, которое является решением некоторых проблем обработки естественного языка. Новизна работы заключается в использовании модифицированного генетического алгоритма для решения задачи кластеризации текстовых документов, который позволяет параллельно анализировать ряд наилучших решений.. Это позволяет повысить качество подсистемы поиска информационно-аналитической системы (ИАС) прогнозирования. Подсистема поиска ИАС применяется для извлечения информации для прогнозирования из коллекции исходных состояний сложной технической системе. Произведена разработка модифицированного генетического алгоритма кластеризации. Приведена программная реализация модуля информационного поиска ИАС прогнозирования с использованием разработанного алгоритма на языке Java для решения задачи кластеризации и применение библиотеки OpenNLP для обработки естественного языка. Также определено место разработанного модуля в системе диагностирования сложных технических систем по поддержанию работоспособности программной системы. Проведены тестовые испытания подобной системы на последней версии копии сайта Wikipedia.org. Эксперименты показали уменьшение времени выполнения алгоритма и улучшение качества полученных результатов.

Генетический алгоритм; кластеризация; информационный поис; прогнозирование.

* Работа выполнена за счет полного финансирования по ГЗ 01/ОПНИ и гранта РФФИ № 15-07-05523.