Раздел II. Искусственный интеллект и нечеткие системы

УДК 002.53:004.89

В.В. Бова, Д.В. Лещанов

СЕМАНТИЧЕСКИЙ ПОИСК ЗНАНИЙ В СРЕДЕ ФУНКЦИОНИРОВАНИЯ МЕЖДИСЦИПЛИНАРНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ ОНТОЛОГИЧЕСКОГО ПОДХОДА*

Одной из основных функций современных междисциплинарных информационных систем является семантический поиск элементов знаний из распределенных источников. Основная проблема в области поиска и обработки знаний заключается в постоянно растущей сложности их идентифицирования и структуризации с целью представления в виде, доступном для понимания и дальнейшего использования. Для ее решения предложен метод формирования многоуровневой онтологической структуры, заданной в виде семантической сети для решения задач поиска и оценки близости элементов знаний в онтологиях междисциплинарных функциональных областей. Разработанная на его основе семантическая модель поиска позволит наглядно и компактно представить структуру семантических отношений между функциональными областями источников знаний с устойчивыми междисциплинарными связями. Для описания связей между элементами знаний в распределенных информационных массивах предлагается использовать их семантические метаописания, представленные в терминах онтологий и терминах поискового запроса. Рассмотрен процесс оптимизации модели поиска и оценки семантически близких элементов знаний на основе кластеризации семантических сетей, представленных графовыми моделями соответствующих уровней: онтологий функциональных предметных областей, поисковых образов и семантических метаописаний терминов словаря онтологии. Семантические метаописания терминов (документов) рассматриваются как набор понятий и отношений (наборы триплетов) в единой модели представления онтологических знаний. Методика оценки релевантности (семантической близости) основана на оценке близости объектов знаний в семантической сети документов и семантической сети запроса. Для анализа разработанного метода была проведена серия вычислительных экспериментов. Полученные данные подтвердили теоретическую значимость и перспективность применения данного метода.

Информационные системы семантического поиска; системы управления знаниями; семантическая сеть; онтологии; семантические метаописания; модель кластеризации объектов знаний; семантическая близость.

V.V. Bova, D.V. Leshchanov

THE SEMANTIC SEARCH OF KNOWLEDGE IN THE ENVIRONMENT OF OPERATION OF INTERDISCIPLINARY INFORMATION SYSTEMS BASED ON ONTOLOGICAL APPROACH

One of the main functions of modern interdisciplinary information systems is the semantic search of knowledge elements from distributed sources. The main problem in the search and processing of knowledge lies in the ever increasing complexity of their identification and structuring with the aim of presenting it in a form that is accessible to understanding and further use. For its

* Работа выполнена при финансовой поддержке РФФИ (проект № 17–07–00449).

solution, a method is proposed for the formation of a multilevel ontological structure for the search and evaluation of the closeness of knowledge elements in ontologies of various functional domains, defined as a semantic network. The semantic model of search developed on its basis will allow to visually and compactly present the structure of semantic relations between functional areas of sources of knowledge with stable interdisciplinary connections. To describe the links between knowledge elements in distributed information arrays, it is suggested to use their semantic meta-descriptions, presented in terms of ontologies and search query terms. The method of optimization of the search model and evaluation of semantically close knowledge elements based on the clustering of semantic networks represented by graph models of the corresponding levels is considered: ontologies of functional subject areas, search images and semantic meta-descriptions of the terms of the ontology dictionary. Semantic meta-descriptions of terms (documents) are considered as a set of concepts and relationships (sets of triplets) in a single model of representation of ontological knowledge. The method for assessing the relevance (semantic proximity) is based on the evaluation of the proximity of knowledge objects in the semantic network of documents and the semantic query network. To analyze the developed method, a series of computational experiments was carried out. The obtained data confirmed the theoretical significance and the prospects of the application of this method.

Information systems of semantic search; knowledge management system; semantic network; ontology; semantic meta-descriptions; clustering objects of knowledge; semantic similarity.

Введение. Интенсивный рост объемов различного вида информационных ресурсов, представленных в распределенных источниках и многообразные потребности общества и личности в их оперативной семантической обработке и эффективной организации поиска знаний, являются важной проблемой в задачах создания междисциплинарных информационных систем (МИС) семантического поиска [1–3]. Актуальность проведения исследования вызвана необходимостью развития подходов и методов анализа сложной (структурной) распределенной информации для задач разработки новых методов и моделей поиска и обработки семантической информации с учетом релевантности по отношению к запросам пользователей.

Одной из основных функций современных МИС является семантический поиск знаний имеющих распределенный и, как следствие, гетерогенный характер представления. Онтологический подход обеспечивает новый уровень в решении задач структуризации, поиска и интеграции данных и знаний из распределенных источников с целью представления в виде, доступном для понимания и дальнейшего использования [4–7].

В статье предлагается метод формирования многоуровневой онтологической структуры, представленной в виде семантической сети для описания междисциплинарных связей между объектами знаний, поиска и оценки близости элементов знаний в онтологиях различных функциональных областей. Модификацией метода является создание семантических метаданных для построения модели поиска с использованием релевантных критериев, отражающих основные семантики описываемых элементов знаний соответствующей проблемной области. Для этого предлагается модель оптимизации поиска и оценки семантически близких элементов в базах знаний МИС на основе кластеризации семантических сетей, представленных графовыми моделями соответствующих уровней: онтологий функциональных предметных областей, поисковых образов и семантических метаописаний терминов словаря онтологии МИС. Семантические метаописания терминов (документов) рассматриваются как набор понятий и отношений (наборы триплетов) в единой модели представления онтологических знаний [1, 8–10]. Методика оценки релевантности (семантической близости) документа основана на оценке близости в некоторой метрике семантической сети этого документа и семантической сети запроса.

Предложенный метод на основе семантической сети позволит интегрировать в модель поиска и оценки близости специализированные онтологии различных функциональных областей знаний и поисковых запросов, наглядно и компактно представить структуру семантических отношений в онтологической структуре.

1. Формулировка проблемы исследования и постановка задачи. Онтологический подход дает целостный, системный обзор предметной области (ПрО) и позволяет сделать знания доступными и повторно используемыми [4–6, 11]. В статье онтологическая структура интерпретируется как система соглашений о некоторой ПрО для достижения заданных целей и предполагает комплексное представление о соответствующей ПрО, включающее формальные и декларативные (описательные) компоненты, словарь терминов, накладываемые на них ограничения целостности, а также логические утверждения, ограничивающие интерпретацию терминов и их отношения друг с другом. В контексте рассматриваемой проблемы, выражения, используемые в онтологической структуре для обозначения понятий будем называть онтологическими терминами, а их семантические метаописания – документом, представляющий особый вид описаний знаний, включающий концептуальное (аннотированное) изложение содержания и смысла информации об объекте. Для построения семантической сети онтологии предлагается использовать метод создания метаописаний, представляющих собой наборы простых высказываний вида «субъект *(s)*– предикат *(p)* – объект *(o)*», которые также называются триплетами *(t)* и отражают основные семантики описываемых элементов знаний [9]. Семантические метаописания являются ценными источниками информации для выполнения поиска и с их применением возможно значительное улучшение функциональности поисковых механизмов в МИС [8]. Оценкой близости между элементами знаний и запросом является числовое значение, которое выражает степень сходства между ними; оценка близости называется оценкой семантической близости, если и только если она определена на основе семантики метаописаний и запросов [3, 9].

Введем следующие условные обозначения компонентов семантической сети онтологии МИС: онтология O представляет собой знаковую систему $O = <C, TG, E, T, R>$, где C – множество понятий (онтологических терминов), для каждого из которых определена его роль (тематическая группа – словарь терминов) $tg(c_i) \in TG$; $TG = \{tg_k\}$ - множество ролей понятий C; $E=(E_c, E_r, E_{tg})$, $E_C= \{e_C\}$, $E_r= \{e_r\}$, $E_{tg}= \{e_{tg}\}$ – значения мер важности множеств экземпляров понятий C, отношений R и ролей TG соответственно; T – множество предикатов – типов отношений; R – множество отношений, которые задают следующие виды связи между сущностями: таксономические, атрибутивные, квантифицирующие, логические и др. [4–6, 12].

Введем следующие правила и ограничения.

1. На основе онтологии ПрО O для каждого термина ci множества концептов $C = \{ci\}$ созданы семантические метаописания $m(c_i) = \{t_1, t_2, \ldots, t_{n(i)}\}$, где $n(i)$ – количество триплетов в логическом представлении онтологического термина c_i; t_i – *RDF*-триплеты – кортежи вида $<s_i, p_i, o_i>$, где s_i и o_i включены в объединение C_i и E_i, а p_i включен в R.

2. Каждый запрос q, данный пользователем из множества запросов Q, также состоит из множества триплетов $q = \{t_1, t_2, \ldots, t_{n(q)}\}$, где $n(q)$ – количество триплетов, содержащихся в запросе q.

3. Весовая функция w определяет значимость любого триплета $t \in T$ (T – множество возможных триплетов) при описании терминов c_i и запроса q: $0 \le w(t, c_i) \le 1$, где $t \in T, c_i \in C, 0 \le w(t, q) \le 1$, где $t \in T, \ q \in Q$.

Для каждого запроса q требуется определить подмножество *RES* множества концептов *C*, которое состоит из релевантных понятий для заданного запроса q – результирующее множество. *Ci* считается релевантным заданному запросу q, если и только если оценка семантической близости между ними превышает некоторую пороговую величину семантического сходства. При этом для вычисления близости между терминами и запросом используются их семантические метаописания.

2. Метод построения семантической сети. Семантические сети являются универсальной структурной моделью формализации онтологической структуры для задач накопления, структуризации, поиска и обработки знаний [1, 4]. Базовым функциональным элементом семантической сети служит графовая структура [13], состоящая из узлов и связывающих их дуг. Каждый узел представляет некоторое понятие, а дуга – отношение между парами понятий. Каждая такая пара отношений определяет некоторое утверждение, являющееся функциональным элементом семантической сети. Узлы помечаются именами соответствующего отношения [14].

Рассмотрим модель семантической сети *SS(O)* и представим ее в виде взвешенного неориентированного графа *G(O). SS(O)* имеет многоуровневую структуру (рис. 1), которая включает в себя три типа слоев:

1. Слой образов поисковых запросов – семантическая сеть *S(Q)* поисковых образов запросов *Q* представляет собой *k* семантических сетей S_i^Q, формализованных в виде графов G_i^Q; $i \in [1: k]$. G_i^Q может быть представлен как однородной сетью с отношениями одного типа, так и неоднородной сетью, которая объединяет в себе несколько типов отношений [4-6].

2. Онтологический слой – семантическая сеть *S(O)* онтологий S_i^O различных функциональных ПрО, формализованных в виде графов G_i^O; $i \in [1: k]$. Онтология представлена ассоциативной сетью понятий, связанных между собой ассоциациями – отношениями семантической близости, тип которых определен.

3. Концептуальный слой – соответствует узлам графа G_i^O – концептам множества $C(O) = \{c_i, i \in [1:n^o]\}$, а ребра $r(O)=\{r_j, j=1,2,\ldots\}$ – четким бинарным отношениям между ними, каждое из которых принадлежит U_p, $p \in [1:m^o]$.

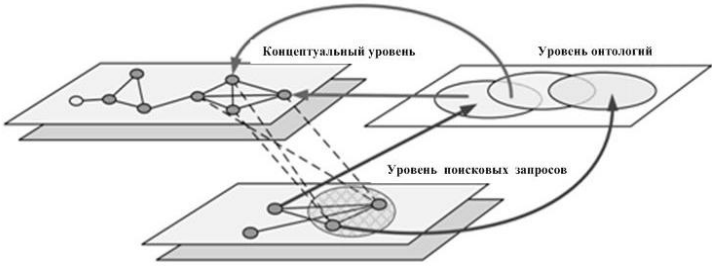

Рис. 1. Структура семантической сети

Для семантической сети *SS(O)* определяются весовые коэффициенты w_i^O, $i \in [1:n^o]$ узлов графа *G(O),* формализующие важность соответствующих концептов в сети. Для каждого из ребер (c_i, c_j), $i,j \in [1:n^o]$, $i \neq j$ графа *G(O)* полагается также заданным вектор весов $\{v_{i,j,p}^O, p \in [1:m^o]\}$, здесь v_p^O – априори заданный вес отношений U_p в онтологии *S(O).* Метод определения комбинированной метрики близости, определения весов концептов и отношений предложен в работах [2, 9. 12]. Аналогичным образом задается семантическая сеть *S(D)* документа *D* в виде взвешенного графа *G(D).*

3. Модель кластеризации семантических сетей. Кластеризация – объединение в группы схожих объектов – является одной из фундаментальных задач в области анализа данных и Data Mining [15–17]. Применение метода кластерного анализа в общем виде сводится к следующим этапам [16].

1. Отбор выборки объектов для кластеризации.

2. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.

3. Вычисление значений меры сходства между объектами.

4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).

5. Представление результатов анализа.

Кластеризация семантических сетей *S(O), S(D)* и *S(Q)* по тематическим группам основана на выявлении признако-фактовых множеств, соответствующих ролям *E* концептов онтологии (ПрО, объект, свойство, действие и т.д.) [18]. Построенные семантические сети кластеризируются по тематическим группам tg_k, $k \in [1:K]$ и образуют внутренне-устойчивые множества вершин графовых моделей. Выделенные группы разбивают множество *C(O)* на *k* непересекающихся кластеров H_i^O. Множество концептов, принадлежащих H_i^O обозначим C_i^O, для которого справедливо выражение:

$$C(O) = \bigcap_{i=1}^{k} C_i^O. \quad (1)$$

Число концептов в кластере обозначим n_i^O. Тогда $\sum_{i=1}^{k} n_i^O = n^O$. Аналогично группы tg_k, разбивают множество концептов C^D документа *D* на *k* ролевых кластеров H_i^d, концепты которых образуют множества C_i^d с числом концептов в них n_i^d:

$$C^D = \bigcap_{i=1}^{k} C_i^d; \sum_{i=1}^{k} n_i^d = n^d. \quad (2)$$

Кластерам H_i^O, H_i^d ставим в соответствие их семантические сети S_i^O, S_i^d и графы G_i^O, G_i^d; i∈[1:k]. Обозначим вес $w_{i,p}^O$ – вес узла $c_{i,p}$ графа G_i^O, $v_{i,p,q}^O$ – вес ребра графа G_i^O, связывающего его узлы $c_{i,p}$, $c_{i,q}$; p, q ∈ [1: n_j^O], p≠q. Аналогичные обозначения $w_{i,p}^d$, $w_{i,p,q}^d$ введем для графа G_i^d.

Поисковый образ запроса *Q* представляет собой *k* семантических сетей S_i^Q, разбитых на *k* ролевых кластеров H_i^Q и формализованных в виде графов G_i^Q; $i \in [1:k]$ по аналогии построения графов G_i^O, G_i^d в соответствии с (1), (2).

Модельный пример кластеризации графа G_i^O представлен на рис. 2.

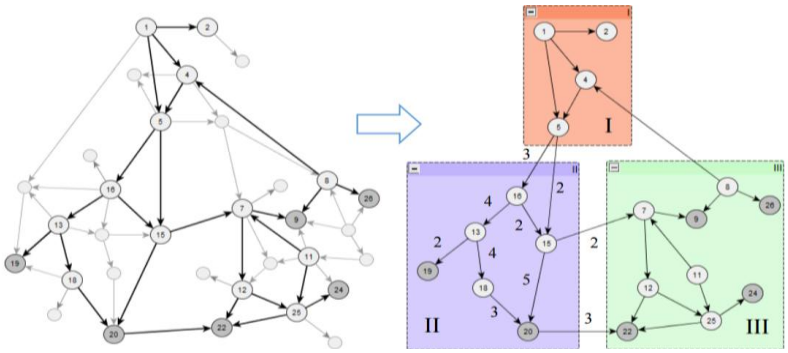

Рис. 2. Модельный пример кластеризации графа

Для решения задачи кластеризации предлагается использовать эвристические алгоритмы, в частности на основе поведения пчелиной колонии, предложенные ведущими учеными в области исследования и разработки методов биоинспириро-

ванного поиска [19–23]. Основу поведения пчелиного роя составляет самоорганизация, обеспечивающая достижение общих целей роя на основе низкоуровневого взаимодействия. Основная идея парадигмы пчелиной колонии заключается в использовании двухуровневой стратегии поиска [21]. На первом уровне с помощью пчел разведчиков формируется множество перспективных областей (источников), на втором уровне с помощью пчел фуражиров осуществляется исследования окрестностей данных областей (источников). Цель пчелиной колонии найти источник, содержащий максимальное количество нектара.

В алгоритмах рассматриваемой задачи кластеризации каждое решение представляется в виде точки (позиции) в пространстве поиска [24–26]. Найденное количество нектара представляет собой значение целевой функции в этой точке. Решение представляет комбинацию уникальных компонент (вершин и ребер графа поиска решений), выбираемых, как правило, из конечного набора конкурирующих между собой компонент. Значения целевой функции определяется комбинациями, выбранными агентами [24]. Целью является поиск оптимальной комбинации компонент.

4. Оценка релевантности модели. Оценку релевантности кластеров тематических групп документов в работе [18] предлагается производить на основании определения близости семантических сетей S_i^D поискового образа D и семантических сетей S_i^Q запроса Q – мер близости соответствующих графов G_i^d, G_i^Q и обозначим их как

$$rez_i = rez(S_i^d, S_i^q). \quad (3)$$

Меру близости семантических метаописаний $m(c_i)$ концептов множества C_i^O определим как:

$$l(c_{i,p}, c_{i,q}) = l_{i,p,q} = min\,(v_{i,p,\alpha}^O + v_{i,p,\beta}^O + \ldots + v_{i,\gamma,q}^O), \quad (4)$$

где минимум берется по всем возможным цепям.

В C_i^D для концепта $c_{i,p} \in C_i^Q$ и $c_{i,p} \notin C_i^d$ необходимо найти $c_{i,q}$, расстояние которого до $c_{i,p}$ будет $l_{i,p,q} = l_{i,p,q}^1$. Включим полученный концепт в $H_{1,i}^Q$. Указанные действия выполняются для всех концептов множества C_i^Q не принадлежащих C_i^D. Результирующий кластер $H_{1,i}^Q$ представляет совокупность $m(c_i) \in C_i^D$ и не принадлежащих C_i^Q, но находящийся ближе всего в соответствии с (3) к этому множеству. Мощность кластера $H_{1,i}^Q$ равна $n_{1,i}^Q$.

Аналогично определим кластер $H_{2,i}^Q$, который является совокупностью концептов множества C_i^D, не принадлежащих C_i^Q и $H_{1,i}^Q$ и находящийся ближе всего в соответствии с (4) к $D_{1,i}^Q$. Мощность кластера $H_{2,i}^Q$ равна $n_{2,i}^Q$. Взаимосвязи кластеров представлены на рис. 3.

Для каждого из $c_{i,p} \in C_i^Q$ и $c_{i,p} \notin C_i^d$ и концептов $c_{i,q} \in H_{k,i}^Q$ определим функцию как:

$$f_{i.p,q}^k(w_{i,p}^d, l_{i,p,q}^k) = f_{i.p,q}^k = \lambda_1 \frac{w_{i,p}^d}{l_{i,p,q}^k}. \quad (5)$$

Функция $f_{i.p,q}^k$ является положительно возрастающей относительно $w_{i,p}^d$ и убывающей относительно $l_{i,p,q}^k$, а также формализует уменьшение весов концептов из кластера $H_{k,i}^Q$ по мере их удаления от кластера H_i^Q.

Оценка релевантности документа производится на основании поисковых образов D и запроса Q в соответствии с (3) и определяется следующим выражением:

$$\text{REZ(D,Q)} = \text{REZ}\,(rez_1^Q, rez_2^Q, \ldots, rez_k^Q, \mu^D) = \mu^D \sum_{i=1}^k \lambda_i\, rez_k^Q, \quad (6)$$

где D – поисковый образ документа, заданный семантическими сетями S_i^D, формализованный в виде графов G_i^D; поисковый образ запроса, заданный семантическими сетями S_i^Q, формализованный в виде графов G_i^Q; REZ – неотрицательная положительная функция своих аргументов; rez – мера близости соответствующих аргументам семантических сетей; λ – положительный скалярный вещественный множитель, определяющий относительный вес аддитивной свертки; μ^D– нормированная взвешенная сумма мер значимости $\mu_1^d, \mu_2^d, \ldots, \mu_k^d$ документа D;

$$\mu^D = \sum_{i=1}^{k} \lambda_i \mu_i^d . \quad (7)$$

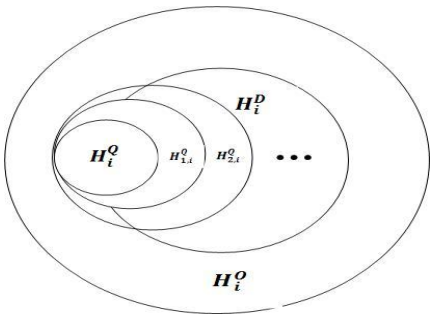

Рис. 4. Взаимосвязи кластеров H_i^Q, $H_{1,i}^Q$, $H_{2,i}^Q$, …, H_i^D, H_i^O

Предложенная в работе методика оценки семантической близости обладает высокой точностью поиска релевантных образов элементов знаний, но имеет высокую вычислительную сложность. Задача определения релевантности документов является многокритериальной, поэтому представляет интерес исследование возможностей оптимизации модели поиска семантически близких элементов знаний.

5. Оптимизация модели поиска. Анализ публикаций [3, 9] показывает, что для оптимизации модели поиска широко используются два метода повышения скорости обработки запросов: фильтрация коллекции документов с помощью инвертированного индекса и применение статических оценок близости. Эти методы могут быть применены отдельно или совместно в зависимости от доступных ресурсов вычислительных систем.

Первый способ оптимизации выполнения запроса заключается в определении нечеткого подмножества Dq множества документов D, документы из которого могут быть релевантными запросу q. Затем сравнения выполняются только в данном множестве Dq. Данное множество Dq называется контекстным множеством запроса q [9]. В данной работе предлагается определение контекстного множества Dq как объединение списков релевантных документов следующим образом:

$$D_q = \bigcup_{t \in q} I_t, \quad (8)$$

где I_t – список релевантных документов триплета t множества D. На основе индекса I список I_t определяется следующим образом:

$$I_t = \{\mu_{I(t)}(d) / d \mid d \in D\}, \ \mu_{I(t)}(d) = \mu_I(d,t). \quad (9)$$

С использованием контекстного множества Dq имеется следующее оптимизированное определение множества результатов запроса q:

$$REZ = \{\mu_{REZ}(d) / d \mid d \in {}^{\alpha}Dq\}, \quad (10)$$

где ${}^{\alpha}Dq$ – четкое множество, полученное в результате операции α-срезки над нечетким множеством Dq. При этом ускорение получается за счет того, что размер множества ${}^{\alpha}Dq$ меньше, чем размер коллекции документов D. Однако снижается полнота результатов в связи с возможными ошибками фильтрации. Скорость вы-

полнения запроса может быть увеличена за счет использования статических оценок близости между элементами метаописаний [3–5, 11]. В предлагаемой модели поиска данная идея применима для вычисления элементарных оценок близости и близости между триплетами.

6. Экспериментальные исследования. С целью определения эффективности использования предложенного в работе метода были проведены вычислительные эксперименты. Целью эксперимента являлась проверка способа оптимизации при вычислении оценки семантической близости на основе времени, потраченного на ее выполнение. Тестовые вычисления оценки семантической близости между компонентами триплетов выполнялись в 5 базах данных без применения способа оптимизации и с его применением.

Для проведения исследований использовалась база знаний компании Prometheus Research, которая предоставляет предприятиям и организациям, проводящим экспериментальные исследования, средства управления web-ориентированными базами данных. Результаты тестирования представлены в табл. 1.

Таблица 1

Сравнительный анализ результатов тестирования

Базы данных	Количество триплетов	Время без оптимизации (сек)	Время с оптимизацией (сек)
Словарь терминов_I	160322	21,1	1,05E-4
Словарь терминов_II	314987	39,8	1,21E-4
Словарь терминов_III	468989	51	1,29E-4
Словарь терминов_IV	631052	71,4	1,33E-4
Словарь терминов_V	1252327	208	1,18E-4

Время без оптимизации в данном эксперименте включает время загрузки графа *G(O)* и время вычисления близости. А время с оптимизацией означает только время вычисления близости. Из табл. 1 видно, что возрастание числа семантических триплетов незначительно влияет на время вычисления близости при использовании предложенного способа оптимизации, что доказывает его эффективность.

Для оценки качества информационного поиска по предложенной методу использовались критерии полноты *R* (отношение количества найденных при поиске релевантных значений к общему количеству значений, релевантных запросу) и точности *P* (отношение количества попавших в результат значений, релевантных запросу, к общему количеству выбранных значений) [9, 18].

Результаты экспериментальных исследований для определения влияния выразительности предложенной модели и HITS (Hyperlink Induced Topic Search) метода [28] на полноту и точность поиска представлены на рис. 4.

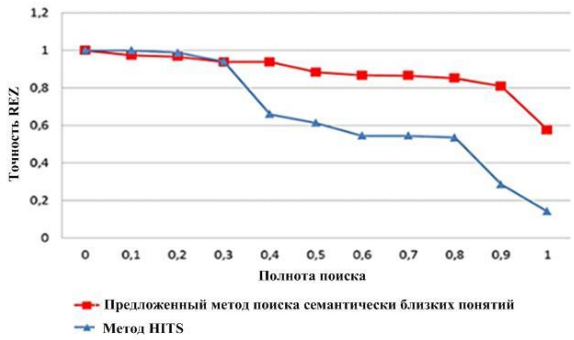

Рис. 4. Графики зависимости точности поиска от полноты

Использование представленной модели информационного поиска с использованием онтологической структуры базы знаний и семантических метаданных дает лучшие показатели полноты и точности поиска для различных категорий пользователей. Причиной снижения формальной точности при увеличении полноты могут быть неточные метаданные компонентов триплетов или нечеткое формирование запроса из-за неточного понимания предметной области пользователем.

Заключение. В работе предложен метод формирования семантической сети для описания связей между метаописаниями элементов знаний МИС и модель поиска и оценки семантически близких элементов в базах знаний МИС на основе кластеризации семантических сетей: метаданных онтологии различных функциональных ПрО, поисковых образов и документов. В описанной теоретической модели семантического поиска элементарными единицами для составления поисковых запросов и метаописаний документов являются триплеты. Предложенная в работе методика оценки семантической близости обладает высокой точностью поиска релевантных элементов знаний, но имеет высокую вычислительную сложность. Задача определения релевантности документов является многокритериальной, поэтому рассмотрен процесс оптимизации обработки запроса на основе фильтрации коллекции документов и вычисления близости между компонентами триплетов с помощью инвертированного индекса. Экспериментальные исследования по критериям полноты и точности поиска продемонстрированы на тестовых задачах и подтвердили теоретическую значимость и перспективность применения предложенного метода. Областью применения полученных результатов является широкий спектр МИС, где необходим семантический поиск и анализ знаний из распределенных источников (медицинская информатика, биоинформатика, социология, оптимизация логистических процессов и др.).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Castano S., Ferrara A., Montanelli S., Racca G.* Semantic information interoperability in open networked systems // Proceedings of the International Conference «SNW». – 2004. – P. 215-230.
2. *Kravchenko Y.A., Kursitys I.O., Bova V.V.* Models for Supporting of Problem-Oriented Knowledge Search and Processing // Proceedings of the First International Scientific Conference «Intelligent Information Technologies for Industry». – 2016. – Vol. 1. – P. 287-297.
3. *Kravchenko Y.A., Kuliev E.V., Kursitys I.O.* Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems // Proceeding of the 10th IEEE International Conference on «Application of Information and Communication Technologies». – 2016. – P. 136-141.
4. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Development of distributed information systems: Ontological approach // Advances in Intelligent Systems and Computing. – 2015. – Vol. 349. – P. 113-122.
5. *Вагин В.Н., Михайлов И.С.* Разработка метода интеграции информационных систем на основе метамоделирования и онтологии предметной области // Программные продукты и системы. – 2008. – С. 22-26.
6. *Бова В.В., Лещанов Д.В.* О вопросе интеграции ресурсов знаний на основе анализа и синтеза онтологий // Информатика, вычислительная техника и инженерное образование. – 2014. – № 3 (18). – С. 14-22.
7. *Бова В.В.* Концептуальная модель представления знаний при построении интеллектуальных информационных систем // Известия ЮФУ. Технические науки. – 2014. – № 7 (156). – С. 109-117.
8. *Нгуен Б.Н., Тузовский А.Ф.* Обзор подходов семантического поиска // Известия Томского государственного университета систем управления и радиоэлектроники. – 2010. – № 2. – С. 234-237.
9. *Нгуен Б.Н., Тузовский А.Ф.* Модель информационного поиска на основе семантических метаописаний // Управление большими системами. – 2013. – № 41. – С. 51-92.

10. *Бова В.В., Лещанов Д.В., Кравченко Д.Ю., Новиков А.А.* Компьютерная онтология: задачи и методология построения // Информатика, вычислительная техника и инженерное образование. – 2014. –№ 4 (19). – С. 18-24.
11. *Крюков К.В., Панкова Л.А., Шипилина Л.Б.* Меры семантической близости в онтологиях // Проблемы управления. – 2010. – № 2. – С. 2-14.
12. *Бова В.В., Заруба Д.В., Курейчик В.В.* Эволюционный подход к решению задачи интеграции онтологий // Известия ЮФУ. Технические науки. – 2015. – № 6 (167). – С. 41-56.
13. *Zhu H., Zhong J., Li J., Yu Y.* An approach for semantic search by matching RDF graphs // Proceedings LAIRS Conference. – 2012. – P. 450-454.
14. *Гладков Л.А., Курейчик В.В., Курейчик В.М.* Дискретная математика: Теория графов. – Таганрог: Изд-во ТТИ ЮФУ. 2010. – 162 с.
15. *Запорожец Д.Ю., Кравченко Ю.А., Лежебоков А.А.* Способы интеллектуального анализа данных в сложных системах // Известия КБНЦ РАН. – 2013. – № 3. – С. 52-56.
16. *Бериков В. С., Лбов Г. С.* Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». – 2008. – 26 с.
17. *Шевченко И.В., Минашкин А.О., Осипчук Л.Н.* Эвристический метод кластеризации в метрическом пространстве признаков // Новые технологии. – 2009. – № 4 (26). – С. 101-106.
18. *Карпенко А.Н.* Оценка релевантности документов онтологической базы знаний // Наука и образование. – 2010. – № 9. – С. 1-26.
19. *Курейчик В.М., Каланчук С.А.* Обзор и состояние проблемы роевых методов оптимизации // Информатика, вычислительная техника и инженерное образование. – 2016. – № 1 (25). – С. 1-13.
20. *Курейчик В.М., Кажаров А.А.* Использование роевого интеллекта в решении NP-трудных задач // Известия ЮФУ. Технические науки. – 2011. – № 7 (120). – С. 30-36.
21. *Лебедев Б.К., Лебедев О.Б., Лебедева Е.М.* Разбиение на классы методом альтернативной коллективной адаптации // Известия ЮФУ. Технические науки. – 2016. – № 7 (180). – С. 89-101.
22. *Карпенко А.П.* Популяционные алгоритмы глобальной поисковой оптимизации. Обзор новых и малоизвестных алгоритмов // Информационные технологии. – 2012. – № 7. – С. 1-32.
23. *Rodzin S., Rodzina L.* Theory of bioinspired search for optimal solutions and its application for the processing of problem-oriented knowledge // Proceeding of the 8th IEEE International Conference «Application of Information and Communication Technologies». – 2014. – P. 142-147.
24. *Курейчик В.В., Полупанова Е.Е.* Эволюционная оптимизация на основе алгоритма колонии пчел // Известия ЮФУ. Технические науки. – 2009. – № 12 (101). – С. 41-46.
25. *Кулиев Э.В., Лежебоков А.А., Дуккардт А.Н.* Подход к исследованию окрестностей в роевых алгоритмах для решения оптимизационных задач // Известия ЮФУ. Технические науки. – 2014. – № 7 (156). – С. 15-25.
26. *Semenova A.V., Kureychik V.M.* Multi-objective particle swarm optimization for ontology alignment // Proceeding of the 10th International Conference on «Application of Information and Communication Technologies». – 2016. – P. 141-148.
27. *Bova V.V., Kureichik V.V., Zaruba D.V.* Data and knowledge classification in intelligence informational systems by the evolutionary method // Proceeding of the 6th International Conference «Cloud System and Big Data Engineering (Confluence)». – 2016. – P. 6-11.
28. *Mizzaro S., Robertson S.* HITS hits TREC - exploring IR evaluation results with network analysis // SIGIR 2007. ACM, 2007. – P. 479-486.

REFERENCES

1. *Castano S., Ferrara A., Montanelli S., Racca G.* Semantic information interoperability in open networked systems, *Proceedings of the International Conference «SNW»,* 2004, pp. 215-230.
2. *Kravchenko Y.A., Kursitys I.O., Bova V.V.* Models for Supporting of Problem-Oriented Knowledge Search and Processing, *Proceedings of the First International Scientific Conference «Intelligent Information Technologies for Industry»,* 2016, Vol. 1, pp. 287-297.
3. *Kravchenko Y.A., Kuliev E.V., Kursitys I.O.* Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems, *Proceeding of the 10th IEEE International Conference on «Application of Information and Communication Technologies»,* 2016, pp. 136-141.

4. *Bova V.V., Kravchenko Y.A., Kureichik V.V.* Development of distributed information systems: Ontological approach, *Advances in Intelligent Systems and Computing*, 2015, Vol. 349, pp. 113-122.
5. *Vagin V.N., Mikhaylov I.S.* Razrabotka metoda integratsii informatsionnykh sistem na osnove metamodelirovaniya i ontologii predmetnoy oblasti [Development of the method of integration of information systems based on metamodelling and ontology of the subject domain], *Programmnye produkty i sistemy* [Software products and systems], 2008, pp. 22-26.
6. *Bova V.V., Leshchanov D.V.* O voprose integratsii resursov znaniy na osnove analiza i sinteza ontologiy [On the issue of integrating knowledge resources on the basis of analysis and synthesis of ontologies], *Informatika, vychislitel'naya tekhnika i inzhenernoye obrazovaniye* [Informatics, Computer Science and Engineering Education], 2014, No. 3 (18), pp. 14-22.
7. *Bova V.V.* Kontseptual'naya model' predstavleniya znaniy pri postroenii intellektual'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.
8. *Nguen B.N., Tuzovskiy A.F.* Obzor podkhodov semanticheskogo poiska [An overview of semantic search approaches], *Izvestiya Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki* [Izvestiya Tomsk State University of Control Systems and Radioelectronics], 2010, No. 2, pp. 234-237.
9. *Nguen B.N., Tuzovskiy A.F.* Model' informatsionnogo poiska na osnove semanticheskikh metaopisaniy [Model of information retrieval based on semantic meta descriptions], *Upravlenie bol'shimi sistemami* [Managing large systems], 2013, No. 41, pp. 51-92.
10. *Bova V.V., Leshchanov D.V., Kravchenko D.Yu., Novikov A.A.* Komp'yuternaya ontologiya: zadachi i metodologiya postroeniya [Computer ontology: objectives and methodology], *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie* [Information, Computing and Engineering Education], 2014, No.4 (19), pp. 18-24.
11. *Kryukov K.V., Pankova L.A., Shipilina L.B.* Mery semanticheskoy blizosti v ontologiyakh [Measures of semantic closeness in the ontology], *Problemy upravleniya* [Problems of Management], 2010, No. 2, pp. 2-14.
12. *Bova V.V., Zaruba D.V., Kureychik V.V.* Evolyutsionnyy podkhod k resheniyu zadachi integratsii ontologiy [The evolutionary approach for ontologies integration problem], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2015, No. 6 (167), pp. 41-56.
13. *.Zhu H., Zhong J., Li J., Yu Y.* An approach for semantic search by matching RDF graphs, *Proceedings LAIRS Conference,* 2012, pp. 450-454.
14. *Gladkov L.A., Kureychik V.V., Kureychik V.M.* Diskretnaya matematika: Teoriya grafov [Discrete mathematics: Graph theory]. Taganrog: Izd-vo TTI YuFU. 2010, 162 p.
15. *Zaporozhets D.Yu., Kravchenko Yu.A., Lezhebokov A.A.* Sposoby intellektual'nogo analiza dannykh v slozhnykh sistemakh [Methods data mining in complex systems], *Izvestiya KBNTs RAN* [izvestiya of Kabardino-Balkar Scientific Centre of the RAS], 2013, No. 3, pp. 52-56.
16. *Berikov V.S., Lbov G.S.* Sovremennye tendentsii v klasternom analize [Modern trends in cluster analysis], *Vserossiyskiy konkursnyy otbor obzorno-analiticheskikh statey po prioritetnomu napravleniyu «Informatsionno-telekommunikatsionnye sistemy»* [All-Russian competitive selection of survey and analytical articles on priority direction "Information-telecommunication systems"], 2008, 26 p.
17. *Shevchenko I.V., Minashkin A.O., Osipchuk L.N.* Evristicheskiy metod klasterizatsii v metricheskom prostranstve priznakov [A heuristic method of clustering in a metric space of attributes], *Novye tekhnologii* [New Technology], 2009, No. 4 (26), pp. 101-106.
18. *Karpenko A.N.* Otsenka relevantnosti dokumentov ontologicheskoy bazy znaniy [Assessing document relevance in ontology knowledge base], *Nauka i obrazovanie* [Science and Education], 2010, No. 9, pp. 1-26.
19. *Kureychik V.M., Kalanchuk S.A.* Obzor i sostoyanie problemy roevykh metodov optimizatsii [Overview and status of the problem of swarm optimization methods], *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie* [Computer science, computer engineering and engineering educatio], 2016, No. 1 (25), pp. 1-13.
20. *Kureychik V.M., Kazharov A.A.* Ispol'zovanie roevogo intellekta v reshenii NP-trudnykh zadach [Swarm intelligence using for NP-tasks solving], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2011, No. 7 (120), pp. 30-36.

21. *Lebedev B.K., Lebedev O.B., Lebedeva E.M.* Razbienie na klassy metodom al'ternativnoy kollektivnoy adaptatsii [Partition a class method alternative collective adaptation], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2016, No. 7 (180), pp. 89-101.
22. *Karpenko A.P.* Populyatsionnye algoritmy global'noy poiskovoy optimizatsii. Obzor novykh i maloizvestnykh algoritmov [Population algorithms of global search engine optimization. Overview of new and little-known algorithms], *Informatsionnye tekhnologii* [Information Technology], 2012, No. 7, pp. 1-32.
23. *Rodzin S., Rodzina L.* Theory of bioinspired search for optimal solutions and its application for the processing of problem-oriented knowledge, *Proceeding of the 8th IEEE International Conference «Application of Information and Communication Technologies»,* 2014, pp. 142-147.
24. *Kureychik V.V., Polupanova E.E.* Evolyutsionnaya optimizatsiya na osnove algoritma kolonii pchel [Artificial bee colony algorithm of evolutionary optimization], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2009, No. 12 (101), pp. 41-46.
25. *Kuliev E.V., Lezhebokov A.A., Dukkardt A.N.* Podkhod k issledovaniyu okrestnostey v roevykh algoritmakh dlya resheniya optimizatsionnykh zadach [Approach to research environs in swarms algorithm for solution of optimizing problems], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2014, No. 7 (156), pp. 15-25.
26. *Semenova A.V., Kureychik V.M.* Multi-objective particle swarm optimization for ontology alignment, *Proceeding of the 10th International Conference on «Application of Information and Communication Technologies»,* 2016, pp. 141-148.
27. *Bova V.V., Kureichik V.V., Zaruba D.V.* Data and knowledge classification in intelligence informational systems by the evolutionary method, *Proceeding of the 6th International Conference «Cloud System and Big Data Engineering (Confluence)»,* 2016, pp. 6-11.
28. *Mizzaro S., Robertson S.* HITS hits TREC - exploring IR evaluation results with network analysis, *SIGIR 2007. ACM, 2007,* pp. 479-486.

Статью рекомендовала к опубликованию д.т.н., профессор Л.С. Лисицына.

Бова Виктория Викторовна – Южный федеральный университет; e-mail: vvbova@yandex.ru; 347928, г. Таганрог, Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Лещанов Дмитрий Валерьевич – e-mail: leshok.dimkaa@yandex.ru; кафедра систем автоматизированного проектирования; студент.

Bova Victoria Victorovna – Southern Federal University; e-mail: vvbova@yandex.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; associate professor.

Leshchanov Dmitriy Valeryevich – e-mail: leshok.dimkaa@yandex.ru; the department of computer aided design; student.

УДК 002.53:004.89

Ю.А. Кравченко, А.Н. Нацкевич

МОДЕЛЬ РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ДАННЫХ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ БУСТИНГА АЛГОРИТМОВ АДАПТИВНОГО ПОВЕДЕНИЯ МУРАВЬИНОЙ КОЛОНИИ И К-СРЕДНИХ*

Рассмотрена разработка модели решения задачи кластеризации. Приведена постановка задачи. Рассматриваются классические (k-means) и современные (метод ядра, метод ансамблей, аффинное распределение) алгоритмы решения задачи кластеризации, выделяются их достоинства и недостатки. Аналитический обзор методов кластеризации показывает, что для

* Работа выполнена при финансовой поддержке РФФИ (проект № 17–07–00446).