УДК 001(06)

ОНТОЛОГИЧЕСКИЕ МОДЕЛИ ПОИСКА ЭКСПЕРТОВ В СИСТЕМАХ УПРАВЛЕНИЯ ЗНАНИЯМИ НАУЧНЫХ ОРГАНИЗАЦИЙ

Л.А. Панкова, В.А. Пронина, К.В. Крюков

Предложены и исследованы онтологические модели поиска экспертов по тематической близости к запросу, представленному текстовым документом, на основе моделей текстового поиска по образцу. Исследованы модели поиска в целях выбора наиболее адекватных из них с точки зрения эксперта.

Ключевые слова: система управления знаниями, модель поиска, онтология, семантическая близость, профиль специалиста, тематическая близость.

ВВЕДЕНИЕ

Корпоративные системы управления знаниями становятся важнейшим инструментом управления организациями, в том числе и научными. Для создания таких систем наиболее перспективно применение семантических технологий, использующих формальное описание областей знаний, разрабатываемых в научной организации, — онтологии областей знаний. Исходя из своего предназначения, система управления знаниями научной организации поддерживает процессы принятия управленческих решений в различных сервисах. В данной работе в рамках онтологического подхода рассматривается задача выбора специалистов из сотрудников научной организации для оценки документа, подлежащего экспертизе, или для исполнения проекта.

При поиске экспертов/исполнителей традиционно используется формальное описание специалиста, содержащее набор взаимосвязанных свойств — так называемая обобщенная модель специалиста. Она включает в себя идентификационные атрибуты (ФИО, дату рождения, место работы, контактную информацию, образование), профессиональные достижения (показатель результативности научной деятельности, тематический индекс цитирования, места в конкурсах, грамоты, медали и т. п.), различные характеристики личности, компетенции в различных областях знаний.

Задача автоматизации поиска экспертов для проведения конкретной экспертизы стала предметом научных исследований и областью практической деятельности. Во многих работах рассматривается задача поиска экспертов по названным в запросе компетенциям, а компетенции специалистов формируются по метаописаниям их деятельности на основании организационных документов, самооценки, взаимооценки, аккредитации, информации из различного рода социальных сетей (см. книгу [1], в разделе 4 которой дан обзор современных подходов к поиску экспертов). Поиск производится по совпадению ключевых слов запроса и компетенций. В работах, основанных на так называемом документно-ориентированном подходе, по запросу находится коллекция документов и для каждого документа этой коллекции рассматривается, кто связан с этим документом. Исходный запрос можно обогатить, выбирая термины из документов, которые уже считаются релевантными запросу.

В работе [2] уровень компетентности специалиста в области знаний предлагается определять путем проведения квалификационного аудита. В работе [3] предлагается метод описания компетенций специалиста по терминам онтологии области знаний из названий его публикаций, паспорте ВАК его специальности, названиям научно-исследовательских проектов, которые он выполняет и т. п. Эксперты отбираются на основе пересечения терминов запроса и компетенций специалиста.

Задача формирования команды для выполнения проекта рассматривается во многих работах. В книге [4] даны формальные постановки задачи выбора команды в виде задач дискретной оптимизации.

В статье [5] задача выбора команды сводится к обобщенной многокритериальной задаче о назначениях. В работах [4, 5] характеристики специалистов считаются заданными. В статье [6] поиск кандидатов для выполнения проекта осуществляется по прецедентам: потенциальными исполнителями считаются исполнители близких проектов, т. е. характеристики специалистов не рассматриваются.

На наш взгляд, самый важный критерий при выборе экспертов или исполнителей нового проекта состоит в тематической близости к документу, подлежащему экспертизе, или проекту. В связи с этим сначала следует определить множество потенциальных экспертов/исполнителей по тематической близости, а затем воспользоваться методами сужения этого множества до множества специалистов, привлекаемых к участию в разрешении поставленной проблемы. Данная работа посвящена поиску потенциальных экспертов или исполнителей нового проекта, близких по тематике к текстовому документу, подлежащему экспертизе, или проекту, представленному текстовым документом.

Подход к решению задачи выбора экспертов/исполнителей по тематической близости, предлагаемый в настоящей работе, заключается в следующем.

- Ключевая проблема состоит в описании знаний специалиста. Общепризнанным считается «объективное» описание знаний с помощью онтологической модели предметной области знаний.
- Свидетельствами знаний специалиста выступают текстовые документы, автором которых он является (публикации, доклады, отчеты, проекты и др.). Свидетельства знаний представляются своим онтологическим описанием, а знания специалиста — на основе совокупности онтологических описаний свидетельств знаний. Запросный документ (документ, подлежащий экспертизе, или новый проект) также представляется своим онтологическим описанием.
- Подобный способ описания специалиста и запросного документа позволяет свести задачу поиска экспертов/исполнителей по тематической близости к текстовому поиску по образцу [7], где образцом служит запросный документ, с помощью формальных методов определения семантической близости онтологических описаний текстовых документов. В связи с этим проанализированы и классифицированы существующие модели текстового поиска по образцу (см. далее § 2).
- Модель текстового поиска (модели документа и модели семантической близости) может быть выбрана с учетом человеческого фактора. Для получения большей адекватности с точки зрения пользователя некоторые параметры модели документа (например, значимость ключевых слов, порог частоты встречаемости) и модели близости (например, значимость составляющих меры, выбор типа агрегации) могут быть заданы пользователем.

1. ОНТОЛОГИЯ И МЕРЫ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ

Онтология — это способ представления знаний о предметной области с помощью конечного множества *понятий* предметной области и *отношений* между ними. Основные преимущества использования онтологий — это единая модель данных, формальная семантика, логический вывод. Для представления онтологических знаний служит модель данных *RDF* (Resource Description Framework)[1]. Онтология как *RDF*-модель данных — это совокупность утверждений, *RDF*-троек — триплетов «субъект—предикат—объект». Триплет — это экземпляр некоторого бинарного отношения, представляется в виде графа, где субъект и объект — вершины, а предикат — дуга, которая эти вершины соединяет. Например, «Курс—Читается—Преподаватель».

Онтологии различают по способу применения. Онтологии для текстового поиска часто называют *лингвистическими*. Наиболее известные лингвистические онтологии: WordNet, EuroNet, Mikrokosmos, Sensus, РуТез. Особенность лингвистических онтологий заключается в том, что они привязаны к семантике грамматических единиц. В лингвистических онтологиях областей знаний источником понятия служит термин — слово либо словосочетание. Понятия существуют как абстрактные сущности независимо от терминов, которые их выражают, при этом может быть несколько возможных вариантов лексического представления понятия в тексте, которые рассматриваются как синонимы. Среди таких синонимов выделяется *дескриптор* — термин, который рассматривается как основной способ ссылки на понятие в рамках онтологии. Другие термины из синонимического ряда употребляются как вспомогательные элементы, текстовые входы, помогающие найти подходящие дескрипторы. Набор дескрипторов со своими синонимами, которые называют *синсетами*, образует словарь, позволяющий ставить в соответствие понятиям онтологии их языковые выражения в текстах. Например, в онтологии «Когнитивный анализ ситуаций» могут быть определены следующие синсеты (дескрипторы подчеркнуты): Модель (когнитивная карта, ориентированный граф, понятийная

[1] Стандарт, разработанный по инициативе World Wide Web Consortium W3C.

структура предметной области), Вес (значение, значение силы влияния), Вершина (фактор, узел), Дуга (причинно-следственная связь, связь, ориентированное ребро, ориентированная связь). Кроме того, лингвистические онтологии могут снабжаться словарями шаблонов лексических выражений, соответствующих триплетам. Так, триплету «Метод—Разработан—Модель» соответствует выражение в тексте: «разработан метод формирования понятийной структуры предметной области».

В системах текстового поиска приходится иметь дело с лингвистическими средствами обработки естественного языка, которые позволяют выделить из документов слова, словосочетания или фразы, соотнесенные с понятиями — носителями содержания документов, выявить зависимости между ними — триплеты. Последнее особенно трудно: шаблоны, соответствующие отношениям, зависят от лексики предметной области; аргументы отношений могут оказаться далеко друг от друга в тексте, между ними может быть другое отношение или совсем не быть отношений [8].

В коллекции научных текстов понятия — это научные термины, которые, как правило, однозначны, не связаны с контекстом, и отношения между ними определяются стандартным набором слов (например, «анализирует», «разрабатывает», «исследует», «доказывает» и т. д.). Таким образом, в коллекциях научных текстов многие лингвистические проблемы решаются проще, в том числе проблема разрешения многозначности.

Онтологический подход в текстовом поиске позволяет оперировать не простыми словами, как в традиционном поиске, а их смыслом, описанным в онтологии, и оценивать близость содержания текстовых документов.

Содержание текстовых документов описывается в терминах заданной онтологии: в тексте распознаются термины понятий и связи между ними (точнее, термины сокращаются до *токенов* — частей терминов, остающихся после отсечения окончаний). Документ представляется набором элементов онтологии с весами (концептуальное индексирование [8]). При этом все синонимы сводятся к одному понятию, многозначные слова отнесены к разным понятиям, связи между понятиями и соответствующими словами описаны и могут быть использованы при анализе текста. Близость двух документов определяется через близость между элементами онтологии.

В зависимости от того, как описаны понятия в онтологии, возникают различные меры близости между понятиями: меры, основанные на взаимном положении в онтологических иерархиях, на свойствах (отношениях и атрибутах). В работе [9] дан обзор мер семантической близости понятий. Меры семантической близости отношений рассматриваются в работах [10, 11].

Во многих работах, например [12, 13], описываются результаты экспериментов по сравнению различных мер семантической близости понятий. Адекватность мер оценивается близостью к оценкам пользователей. Эксперименты показывают, что качество мер зависит от конкретной коллекции документов, от конкретной онтологии, от пользователей и др. Таким образом, «золотого стандарта» меры близости не существует.

2. ОБЗОР СУЩЕСТВУЮЩИХ МОДЕЛЕЙ ТЕКСТОВОГО ПОИСКА ПО ОБРАЗЦУ

Текстовый поиск по образцу — это поиск в коллекциях текстовых документов по запросу, заданному текстовым документом. Модель текстового поиска по образцу включает модель документа (как документа коллекции, так и документа запроса) и модель близости (релевантности) документов (документа запроса и документа коллекции).

Модель документа (обозначим ее *MD*), или поисковый образ документа, — это формальное представление содержания текстового документа в некотором поисковом пространстве. Координатами поискового пространства могут быть *ключевые слова*, значимые для рассматриваемой коллекции документов или элементы онтологии рассматриваемой предметной области — *термины-понятия* или триплеты отношений терминов-понятий. Процесс отображения документа в поисковое пространство называется *индексированием* и заключается в присвоении каждому документу некоторого *индекса* в поисковом пространстве. Индекс текстового документа есть вектор значений координат в выбранном поисковом пространстве, равных частотам встречаемости соответствующих координатам токенов в документе. Веса, характеризующие значимость понятий в документе, могут быть пересчитаны с учетом значений координат индексов документов.

Индексирование может быть автоматическим с применением средств автоматической обработки текстов и автоматизированным с применением интеллектуального интерфейса. При автоматизированном режиме индексирования по визуально представленной онтологии лицо, выполняющее индексирование, может выбирать понятия, отражающие семантику соответствующего результата деятельности.

Тип координат поискового пространства (ключевые слова, понятия, триплеты) определяет тип модели документа: *пословный*, *концептуальный*, *реляционный* соответственно. Модель документа определяет модель близости. Если близость послов-

ных моделей основана на совпадении слов, то онтологические модели документа (концептуальная и реляционная) позволяют более адекватно определить близость между документами благодаря семантической информации.

Модель документа может быть основана как на информации, содержащейся в самом документе, так и на информации, полученной из ссылок на документ.

Пословная модель. Традиционно используется пословная модель документа — *BOW* (bag of words) — вектор весов ключевых слов [14]. Классическим подходом к определению ключевых слов считается статистический подход, основанный на мере $TF \cdot IDF$ (*TF* — term frequency, *IDF* — inverse document frequency)[2], которая используется для оценки веса слова в документе, являющегося частью коллекции документов. Вес слова пропорционален частоте его употребления в документе и обратно пропорционален частоте его употребления во всей коллекции документов. В качестве весов ключевых слов в *BOW* используют также различные модификации *TF-IDF*, например, с учетом их положения в тексте документа — заглавии, аннотации и т. д. [14]. Близость текстовых документов, представленных *BOW*, вычисляется как косинус угла между векторами весов ключевых слов.

Таким образом, поиск сходства документов в этом случае основан на совпадении ключевых слов. Например, два документа, модели которых представлены множествами ключевых слов {змея, пустыня} и {гадюка, Сахара} соответственно, имеют нулевую близость по косинусной мере, хотя на самом деле ключевые слова (и документы) связаны семантически.

Концептуальная модель. Документ представляется вектором весов понятий — *BOC* (bag of concepts). При определении веса понятия в *BOC* учитываются частоты встречаемости (*TF*) терминов, соответствующих этому понятию в документе, и «онтологический» вес этого понятия — так называемое *информационное содержание* понятия *IC* (information content) в онтологии (аналог *IDF*). Информационное содержание понятия *c* в онтологии в работе [15] определяется так: $IC(c) = -log(P(c))$, где $P(c)$ — частота встречаемости понятия и его подпонятий в коллекции (чем абстрактнее понятие, тем меньше величина *IC*). В работе [16] в качестве информационного содержания понятия предлагается использовать так называемое *внутреннее информационное содержание* (intrinsic information content), основанное только на иерархической структуре самой онтологии (таксономии):

$$IC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log K},$$

где *hypo* (*c*) — число прямых потомков понятия *c*, *K* — общее число вершин иерархии.

В простейшем случае вес понятия определяется нормированной суммой весов терминов соответствующего синсета в документе.

В другом подходе для определения весов понятий в *BOC* используется семантическая близость терминов [17, 18]: строится взвешенный граф, вершины которого — понятия документа, вес ребра — значение семантической близости инцидентных вершин. В работе [17] определение весов основано на использовании *плотности* и *информативности* тематически связанных подграфов (так называемых сообществ[3]). Считается, что вес понятия в документе тем больше, чем больше в документе связанных с ним понятий:

$$w_{d,\,c} = 1 + \ln\left(f_{d,\,c}\left(1 + \sum_{c_i \in R(c)} f_{d,\,c_i} S(c, c_i)\right)\right),$$

где $f_{d,\,c}$ — частота встречаемости понятия *c* в документе *d*, $R(c)$ — множество понятий в документе, связанных с *c*, $S(c, c_i)$ — мера семантической близости вершин-понятий *c* и c_i [18]. Для вычисления близости между документами, представленными моделями *BOC*, в работе [19] предлагается трехэтапная модель близости:

— вычисляются близости между каждой парой понятий сравниваемых документов;

— определяются близости между каждым понятием одного документа и множеством понятий другого документа — так называемые частные близости;

— частные близости агрегируются в близость между двумя множествами понятий из моделей *BOC* документов, т. е. между документами.

Частные близости предлагается определять как максимальное значение из близостей между понятием первого документа и понятиями второго документа. Разные модели агрегации (аддитивная,

[2] *TF* — частота встречаемости токена в документе, *IDF* — величина, обратная частоте встречаемости токена во всех документах коллекции.

[3] Сообщества — наиболее массивные и сильно связанные подграфы (как правило, соотносятся с главными темами документа). Плотность сообщества — сумма весов ребер, соединяющих вершины этого сообщества. Информативность сообщества — это сумма весов терминов (TF-IDF), входящих в это сообщество, деленная на число терминов сообщества. Вес сообщества вычисляется как плотность сообщества, умноженная на его информативность.

сигмоидальная, различные средние и др.) обеспечивают разные вклады частных близостей в общую близость между документами и могут выбираться пользователем.

Во многих работах для более адекватной оценки близости документов рассматривается расширение концептуальной модели документа семантически близкими понятиями. Так, модели *BOC* сравниваемых документов расширяются понятиями, близкими по онтологии к понятиям из *BOC* (например, гиперонимами[4] или гипонимами[5]), с весами, пропорциональными весам исходных понятий [20]. На каждом шаге итерации в качестве близости между двумя расширенными *BOC* используется среднее значение косинусных мер, полученных на всех предыдущих итерациях.

Для учета отношений между понятиями из *BOC* сравниваемых документов определяются соответствующие пути между понятиями из двух *BOC* в семантической сети онтологии [20]. В итерационном процессе прохождения этих путей пересчитываются веса исходных понятий с учетом весов добавленных дуг. Близость между исходными *BOC* вычисляется агрегацией новых весов. В другом подходе к учету отношений между понятиями предлагается рассматривать двудольный граф с множествами вершин-понятий из *BOC* сравниваемых документов и с дугами, взвешенными минимальными длинами путей между вершинами из двух *BOC* в онтологии. Близость между двумя документами вычисляется агрегацией весов дуг оптимального паросочетания этого двудольного графа [21].

В работе [22] предлагается представлять документ набором понятий, извлеченных из гиперссылок на документ, считая, что содержание документа лучше характеризуется понятиями из гиперссылок в документах, ссылающихся на него. Вес понятия задается нормированным числом гиперссылок, помеченных этим понятием. Близость двух документов предлагается вычислять так:

$$S(A, B) = \frac{1}{2}\left[\left(\frac{1}{K}\sum_{i=1}^{|A|}\max_{j\in[1,|B|]}(\lambda_{i,j}S(k_i, h_j))\right) + \left(\frac{1}{H}\sum_{i=1}^{|B|}\max_{j\in[1,|A|]}(\mu_{i,j}S(k_i, h_j))\right)\right],$$

где $A = \{(k_i, w_i)\}$ и $B = \{(h_j, v_j)\}$ — *BOC* документов A и B, k_i и h_j — веса понятий w_i и v_j, $\lambda_{i,j} = \frac{w_i + v_j}{2\max(w_i, v_j)}$, $K = \sum_{i=1}^{|A|}\lambda_{i,x(i)}$, $x(i) = x|(\lambda_{i,x}S(k_i, h_x) = \max_{j\in[1,|B|]}(\lambda_{i,x}S(k_i, h_j))$.

Подобным образом определяются $\mu_{i,j}$ и H. Здесь $S(k_i, h_j)$ — мера семантической близости между двумя понятиями [23]. Предложенная мера фактически представляет собой одну из разновидностей трехэтапной модели с агрегирующей аддитивной функцией. Среднее арифметическое вычисляется из-за несимметричности меры [23].

Реляционная модель. В подходе, основанном на триплетах, при создании модели документа из текста извлекаются триплеты (по шаблонам отношений). Модель документа *BOT* (bag of triplets) есть вектор весов триплетов [11, 24, 25].

Для вычисления весов триплетов в *BOT* предлагается использовать меру *TF/IDF*, рассматривая каждый триплет как отдельное «слово», а для вычисления близости — косинусную меру [24].

В работе [25] близость двух триплетов вычисляется как среднее арифметическое близости субъектов, объектов и предикатов триплетов. Близость двух *BOT* вычисляется как сумма близостей между всеми парами триплетов, входящими в поисковые образы.

В работах [26, 27] близость двух *RDF*-графов вычисляется как взвешенная сумма концептуальной близости и близости по отношениям. Концептуальная близость оценивается долей совпадающих понятий с учетом весов, близость по отношениям — суммой долей общих дуг отдельно по каждому типу отношений с учетом весов типов отношений.

В работе [11] предлагается вычислять семантическую близость документов, представленных *BOT*, на основе соответствия *RDF*-графов документов. Проблема определения близости графов в такой формулировке лежит в рамках известной задачи проверки изоморфизма графов [28], которая является NP-полной. Однако учет специфики *RDF*-графов позволяет избежать проблемы NP-полноты. Приведен полиномиальный алгоритм поиска наилучшего отображения, при котором сумма близостей соответствующих вершин и дуг максимальна[6].

[4] Более общими понятиями (предками в таксономической иерархии).

[5] Менее общими понятиями (потомками в таксономической иерархии).

[6] Заметим, что в такой постановке задачу нахождения близости *RDF*-графов нельзя считать решенной. Остаются вопросы, связанные с существованием циклов, определением наилучшего отображения и др.

3. ПОИСК ЭКСПЕРТОВ/ИСПОЛНИТЕЛЕЙ ПО ТЕМАТИЧЕСКОЙ БЛИЗОСТИ

При поиске экспертов/рецензентов для оценки научных документов, а также исполнителей нового проекта сначала следует определить множество потенциальных кандидатов, знания которых наиболее близки тематике научного документа/проекта.

Пусть $O = \{O_1, ..., O_n\}$ — множество онтологий направлений (областей знаний), разрабатываемых в организации. Будем считать, что свидетельствами знаний специалиста служат результаты его деятельности, опубликованные в документальных источниках (текстовые документы в электронном виде): научные статьи, доклады, проекты, отчеты и т. д., автором которых он является.

Текстовые документы представляются совокупностью своих моделей в поисковых пространствах онтологий направлений, разрабатываемых в организации, — профилем документа.

Знания сотрудника представляются *профилем специалиста***,** создаваемым на основе профилей свидетельств его знаний. Выбор кандидатов основывается на сопоставлении профилей текстового документа-запроса и специалиста с учетом их семантической близости.

Профиль документа есть вектор (упорядоченная последовательность элементов)

$$PD = (MD_1, ..., MD_n),$$

где MD_i — модель документа в i-й области знаний O_i, n — число областей знаний (направлений), разрабатываемых в организации.

Профиль специалиста есть вектор

$$PS = \{MS_1, ..., MS_n\} = \{(MD_{11}, ..., MD_{1m_1}), ..., (MD_{n1}, ..., MD_{nm_n})\},$$

где MD_{ij} — модель j-го документа в i-й области знаний, MS_i — модель специалиста в i-й области знаний, $MS_i = (MD_{i1}, ..., MD_{im_i})$, m_i — число документов (свидетельств знаний) специалиста в i-й области знаний.

Профиль специалиста в i-й области знаний MS_i определяется по множеству профилей текстовых документов в i-й области знаний — свидетельств знаний специалиста и представляет собой вектор весов элементов онтологии — модель специалиста. В простейшем случае, профиль специалиста в i-й области знаний есть индекс документа, представляющего собой объединение свидетельств знаний этого специалиста в i-й области знаний. Возможны различные способы вычисления весов элементов онтологии в профиле специалиста, в том числе с учетом значимости свидетельств знаний.

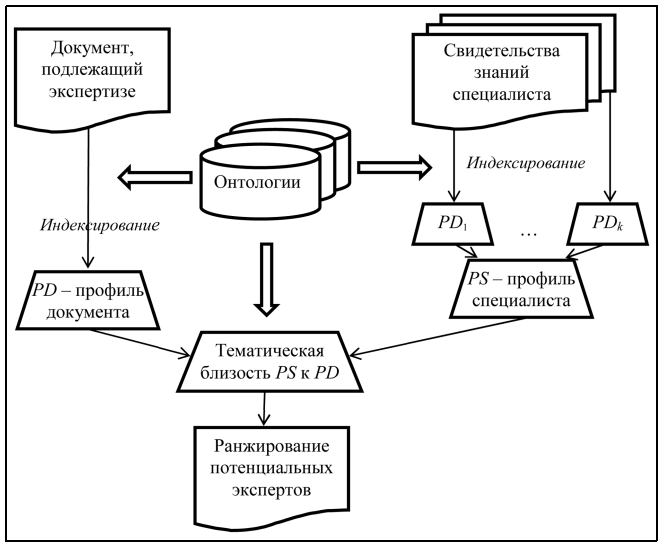

Рис. 1. Схема выбора экспертов по тематической близости документу, подлежащему экспертизе

Тематическая близость специалиста документу, подлежащему экспертизе (образцу), есть вектор $TS(PS/PD) = S_1, ..., S_n$, где S_i — семантическая близость профиля специалиста к образцу по i-й области знаний.

Семантическая близость S_i профиля специалиста к образцу по i-й области вычисляется по модели близости, определяемой выбранным типом модели документа (BOC, BOT).

Список кандидатов может определяться выбором порога (или порогов по каждой области знаний с учетом значимости) семантической близости. При ранжировании кандидатов в качестве критериев тематической близости специалиста к образцу (CTS) могут использоваться различные агрегации близостей S_i по каждой области знаний с весами, определяющими важность тематики для образца, например простейшие:

- тематическое соответствие профиля специалиста образцу

$$\bar{S}(PS/PD) = \frac{1}{n}\sum_{i=1}^{n} S_i;$$

- универсальность по широте охвата специалистом тематик образца (стандартное отклонение близостей)

$$U(PS/PD) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{S}(PS/PD) - S_i)^2}.$$

Потенциальные эксперты ранжируются по выбранному критерию тематической близости специалиста документу, подлежащему экспертизе.

На рис. 1 показана схема выбора кандидатов в эксперты.

4. УЧЕТ ПРЕДПОЧТЕНИЙ ПОЛЬЗОВАТЕЛЯ

Важный фактор, определяющий эффективность любого информационного поиска, — это человеческий фактор, который зачастую либо игнорируется, либо его значение во многом недооценивается. А именно, не учитывается тот факт, что во многом поиск определяется слабо формализуемыми и нечеткими условиями, в значительной степени зависящими от опыта и предпочтений человека.

Предпочтения пользователя могут учитываться при выборе модели близости, а именно при:

— выборе меры близости между понятиями онтологии — определение значимости составляющих: таксономической, реляционной, атрибутивной;

— определении весов (значимости) ключевых слов, понятий, отношений, триплетов, выборе порога частоты встречаемости;

— выборе типа функции близости между документами.

В работе [15] предлагается способ интерактивного взаимодействия пользователя и компьютера для трехэтапной модели близости. Пользователь настраивает агрегирующую функцию при предоставлении ему результатов запроса в каждой итерации, показывающих, как отображаются его предпочтения на их ранжировании.

Меры могут настраиваться также на базе машинного обучения [29].

5. НАСТРОЙКА ПАРАМЕТРОВ МОДЕЛИ ПОИСКА

В рамках разработки семантической поисковой системы для поддержки принятия решений в портале научной организации исследуются модели поиска — модели (профили) специалиста/документа и модели релевантности (близости) — в целях выбора наиболее адекватных с точки зрения эксперта. Рассматриваются итеративные процедуры настройки параметров модели поиска, удовлетворяющие предпочтениям эксперта для одной области знаний (направления, разрабатываемого в организации).

Для представления профилей документа и специалиста выбрана концептуальная модель (*ВОС*) и способ формирования профиля специалиста в рассматриваемой области знаний как профиль документа, представляющего собой объединение свидетельств знаний специалиста в этой области (суммарного документа).

Будем считать, что заданы:

— онтология области знаний;

— коллекция документов (свидетельств знаний специалистов) и их индексы, в которых координаты равны частотам встречаемости понятий онтологии в документах;

— список специалистов научной организации (авторов документов) и индексы соответствующих суммарных документов.

Для формирования профилей документов и специалистов на основе их индексов применяется процедура выбора метода вычисления значимости понятия для отображения содержания документов, т. е. весов понятий в профилях, с участием эксперта (например, автора). Эксперт выбирает метод вычисления весов понятий из заданного списка методов. В случае учета места встречаемости понятия в тексте документа настраиваются веса (значимости) мест встречаемости. Кроме того, экспертом выбирается порог веса понятия, ниже которого вес считается равным нулю. Для каждого документа поисковая система ранжирует понятия по весам, вычисленным по выбранному методу, а эксперт выбирает метод подсчета весов понятий, наиболее соответствующий его предпочтениям. По набранной статистике по документам репрезентативной выборки из коллекции будет выбран метод подсчета весов понятий в профилях документов и специалистов.

Для вычисления близости профилей документа и специалиста используется трехэтапная модель:

— вычисляются близости между каждой парой понятий из сравниваемых профилей;

— определяются близости между каждым понятием из профиля документа и профилем специалиста — частные близости;

— частные близости агрегируются для вычисления близости между двумя профилями.

В рамках этой модели применяется итеративная процедура взаимодействия эксперта с системой, обеспечивающая настройку параметров, удовлетворяющих мнению эксперта в ранжировании специалистов по близости к документу-запросу. Эксперт выбирает следующие параметры (из заданных списков):

- меру близости между понятиями онтологии (меры, основанные на: иерархических отношениях, неиерархических отношениях, на атрибутах; гибридные меры);
- метод вычисления частной близости (степень, в которую возводится значение семантической близости между понятиями; тип агрегации близостей между понятиями — максимум, среднее арифметическое);
- метод вычисления близости между профилями (степень, в которую возводится значение се-

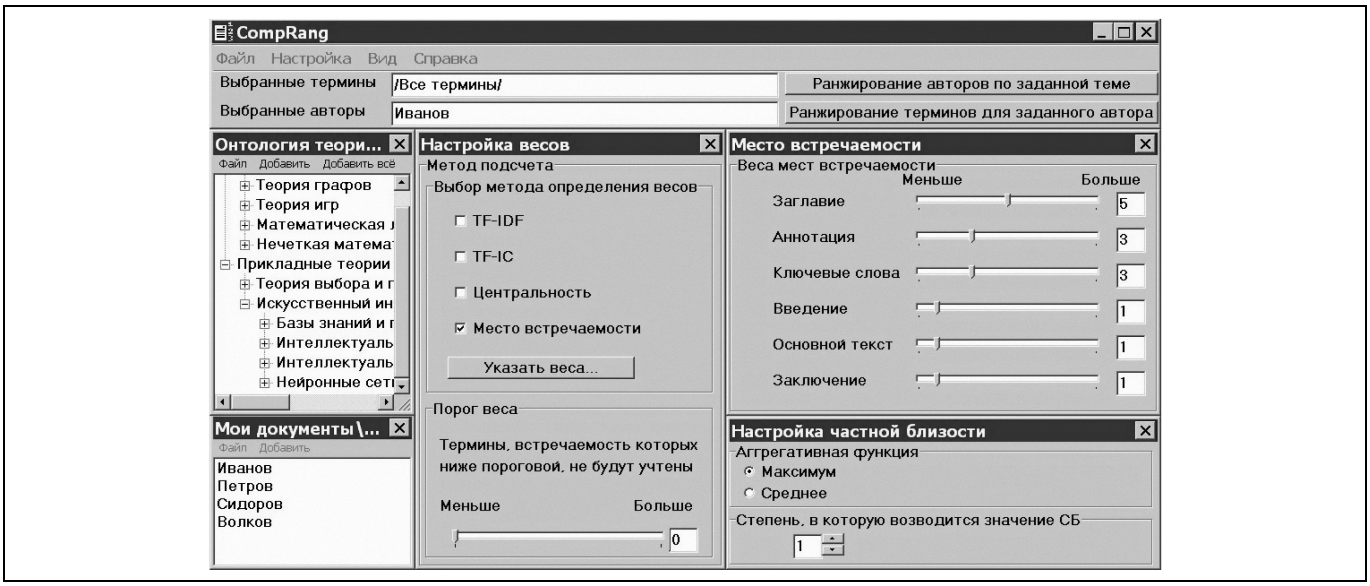

Рис. 2. Интерфейс выбора метода расчета весов и частной близости

мантической близости между частными близостями; тип агрегации частных близостей).

Система ранжирует специалистов по близости их профилей к профилю заданного документа-запроса в соответствии с выбранными параметрами. В процессе итеративной процедуры настройки эксперт определяет параметры, соответствующие его предпочтениям в ранжировании специалистов по близости данному документу-запросу. По набранной статистике по запросам-документам и специалистам будут определены параметры трехэтапной модели близости профилей специалистов к запросу-документу.

На рис. 2 представлен интерфейс выбора метода расчета весов и частной близости: главное окно, окно онтологии, окно специалистов, окно выбора метода расчета весов понятий, окно задания весов мест встречаемости (заглавие, аннотация и др.), окно настройки частной близости.

ЗАКЛЮЧЕНИЕ

Предложены онтологические модели поиска экспертов/исполнителей по тематической близости к запросу, представленному текстовым документом, на основе моделей текстового поиска по образцу. Знания специалиста представляются метаописанием, которое формируется на основе совокупности метаописаний свидетельств знаний. Подобный способ описания специалистов позволяет решать задачи поиска экспертов/исполнителей по тематической близости с помощью формальных методов определения семантической близости метаописаний текстовых документов. Разработана семантическая поисковая система для поддержки принятия решений при выборе экспертов/исполнителей. Исследованы онтологические модели поиска — модели специалиста/документа и модель релевантности (близости) — в целях выбора наиболее адекватных с точки зрения эксперта.

ЛИТЕРАТУРА

1. *Сетевая* экспертиза / Под ред. РАН Д.А. Новикова, А.Н. Райкова / Д.А. Губанов и др. — М.: Эгвес, 2010. — 168 с.
2. *Тузовский А.Ф.* Создание и использование базы знаний профилей компетентности специалистов организации // Изв. Томского политехн. ун-та. — 2007. — № 2.— Т. 310, № 2. — С. 186—189.
3. *Гладун А.Я.*, *Рогушина Ю.В.* Использование онтологических знаний и тезаурусов для объективного профилирования специалистов // Штучний інтелект. — 2006. — № 3. — С. 379—390.
4. *Новиков Д.А.* Математические модели формирования и функционирования команд. — М.: Физматлит, 2008. — 184 с.
5. *Ларичев О.И.*, *Стернин М.Ю.* Человеко-машинные методы решения многокритериальной задачи о назначениях // Автоматика и телемеханика. — 1998. — № 7. — С. 135—156.
6. *Верещак И.А.* Разработка метода семантического описания проектных работ и подбора подходящих исполнителей // Системи обробки інформації. — 2009. — Вип. 6 (80). — С. 147—149.
7. *Российский* семинар по оценке методов информационного поиска, 2009, поиск похожих документов по документу-образцу или фрагменту текста. — URL: http://romip.ru/ru/2009/tracks/mixed-feedback.html (дата обращения: 7.04.2011).
8. *Онтологии* и тезаурусы. Учебное пособие / В.Д. Соловьев и др. — Казань, М., 2006. — 157 с.

9. *Меры* семантической близости в онтологии / К.В. Крюков и др. // Проблемы управления. — 2010. — № 5. — С. 2—14.
10. *Maedche A.*, *Staab S.* Measuring Similarity between Ontologies // Lecture Notes in Artificial Intelligence. — 2002. — N 2473. A. — P. 251—263.
11. *An Approach* for Semantic Search by Matching RDF Graphs / Heaping Zhu and others // Proc. FLAIRS Conference. — 2002. — P. 450—454.
12. *Nguyen Hoa A.*, *Eng B.* New Semantic Similarity Techniques of Concepts applied in the Biomedical Domain and WORDNET: Thesis Presented to the Faculty of the University of Houston Clear Lake in Partial Fulfillment of the Requirements for the Degree Master of Science the University of Houston-Clear Lake. — 2006.
13. *Крижановский А.А.* Оценка результатов поиска семантически близких слов в Википедии // Тр. СПИИРАН. — 2007. — Вып. 5. — С. 113—116.
14. *Сафронов А.В.* HeadHunter на РОМИП-2008 // Тр. российского семинара по оценке методов информационного поиска (РОМИП) 2007—2008. — СПб., 2008. — С. 33—42.
15. *Resnik P.* Using information content to evaluate semantic similarity in ontology // Proc. of the 14th Int'l Joint Conference on Artificial Intelligence, 1995. — P. 448—453.
16. *Seco N.*, *Veale T.*, *Hayes J.* An intrinsic information content metric for semantic similarity in WordNet / Proceedings of the 16th European Conference on Artificial Intelligence. — Valencia, 2004. — P. 1089—1090.
17. *Гринева М.*, *Гринев М.*, *Лизоркин Д.* Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов // Тр. Ин-та системного программирования РАН. — URL: http://citforum.ru/database/articles/kw_extraction/ (дата обращения: 7.04.2011).
18. *Болдаков А.* Поиск в коллекциях текстовых документов на основе методов семантического анализа, использующих универсальные гипертекстовые энциклопедии // Московская секция ACM SIGMOD. — 2008. — URL: http://synthesis.ipi.ac.ru/sigmod/seminar/boldakov08.ppt (дата обращения: 7.04.2011).
19. *User* Centered and Ontology Based Information Retrieval System for Life Sciences / Sylvie Ranwez and others // Proc. of the 3rd Intern. / Workshop on Semantic Web Applications and Tools for the Life Sciences. — Berlin, 2010. — URL: http://arxiv.org/ftp/arxiv/papers/1012/1012.1617.pdf (дата обращения: 7.04.2011).
20. *Thiagarajan Rajesh*, *Geetha Manjunath*, *and Markus Stumptner.* Computing Semantic Similarity Using Ontologies // Intern. Semantic Web Conference (ISWC). — Karlsruhe, 2008. — URL: http://www.hpl.hp.com/techreports/2008/HPL-2008-87.pdf (дата обращения: 7.04.2011).
21. *Harold W. Kuhn.* The Hungarian Method for the Assignment Problem. Naval // Naval Research Logistic Quarterly. — 1955. — P. 83—97.
22. *Maria Halkidi*, et al. THESUS: OrganizingWeb document collections based on link semantics // The VLDB Journal. — 2003. — Vol. 12, N 4. — P. 320—332.
23. *Wu Z.*, *Palmer M.* Verb semantics and lexical selection / 32nd Annual Meeting of the Association for Computational Linguistics. — 1994. — P. 133—138.
24. *Рабчевский Е.А.* Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска // Тр. 11-й Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2009). — Петрозаводск, 2009. — С. 69—77.
25. *Черний А.В.*, *Тузовский А.Ф.* Развитие информационной системы организации с использованием семантических технологий / Знания-Онтологии-Теории 2009 (ЗОНТ-09). — URL: http://www.math.nsc.ru/conference/zont09/reports/08Cherniy-Tuzovskiy.pdf (дата обращения: 7.04.2011).
26. *Богатырев М.Ю.* , *Латов В.Е.* , *Столбовская И.А.* Применение концептуальных графов в системах поддержки электронных библиотек // Тр. 9-й Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007. — Переславль-Залесский, Россия, 2007. — Т. 2. — С. 104—110.
27. *Карпенко А.П.* Оценка релевантности документов онтологической базы знаний // Электронное науч.-техн. издание «Наука и образование». — URL: http://technomag.edu.ru/doc/157379.html (дата обращения: 7.04.2011).
28. *Гэри М.*, *Джонсон Д.* Вычислительные машины и труднорешаемые задачи. — М.: Мир, 1982. — 194 с.
29. *Irena Spas, Goran Nenadi, Kostas Manios, and Sophia Ananiadou.* Supervised Learning of Term Similarities // Lecture Notes in Computer Science. — 2002. — N 2412. — P. 429—434.

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Панкова Людмила Александровна — канд. техн. наук, ст. науч. сотрудник, ☎ (495) 334-92-49, ✉ pankova@ipu.ru

Пронина Валерия Александровна — канд. техн. наук, ст. науч. сотрудник, ☎ (495) 334-92-49, ✉ pankova@ipu.ru,

Крюков Кирилл Вячеславович — ст. математик, ☎ (495) 334-76-39, ✉ kryukovkirill@yandex.ru,

Институт проблем управления им. В.А. Трапезникова РАН, г. Москва.