УДК 001(06)

МЕРЫ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ В ОНТОЛОГИИ

К.В. Крюков, Л.А. Панкова, В.А. Пронина, В.С. Суховеров, Л.Б. Шипилина

Представлены обзор и классификация существующих мер семантической близости для онтологических термов — понятий, отношений и экземпляров — и онтологий. В основу классификации мер положены характеристики онтологических термов — положения в иерархических структурах, отношения и их типы, атрибуты понятий и экземпляров. Дан обзор способов оценки мер близости.

Ключевые слова: онтология, понятие, отношение, атрибут, экземпляр, семантическая близость.

ВВЕДЕНИЕ

Онтология — это спецификация концептуализации предметной области (ПО). Онтологический подход дает целостный, системный обзор ПО и позволяет сделать знания доступными и повторно используемыми. Онтологии — неотъемлемый компонент Semantic Web.

Онтология состоит из организованных в иерархию понятий ПО, отношений между понятиями и атрибутов понятий, а также аксиом и правил вывода. Понятия представляют множества экземпляров. Выражения, используемые в онтологии для обозначения понятий, экземпляров и отношений, будем называть *онтологическими термами*. Каждая онтология отражает взгляд некоторого сообщества на ПО и сконструирована для специфических задач, решаемых в рамках этой онтологии.

Под *семантической близостью* подразумевается сходство пар объектов, а не семантическая связанность[1] объектов, когда объекты различной природы имеют некоторую связь, например *публикация — автор*. Семантическая близость объектов включает в себя множество аспектов сходства, поэтому выбор тех или иных критериев для оценки близости в каждом отдельном случае представляет собой непростую задачу, зависящую от целей исследования. Онтологический подход позволяет нам оперировать не простыми словами (*plain words*), а их смыслом (*senses of words*), т. е. словами, смысл которых определен онтологией. Меры близости онтологических термов используют различные семантические характеристики сравниваемых термов — их свойства (атрибуты и отношения с другими термами), взаимное положение в онтологических иерархиях.

Данный обзор рассматривает онтологические меры семантической близости, предполагающие однозначную интерпретацию термов для одной онтологии. Для кросс-онтологических мер при разных лексиконах онтологий используется еще и лексическая близость термов.

Онтологический подход обеспечивает новый уровень в решении задач *поиска* и *интеграции информации* (объединение информации из разных источников).

Запрос пользователя, как правило, не полностью отражает его интерес, так как пользователь, с одной стороны, не знает всех терминов и структур данных, заложенных в систему, с другой — не всегда точно выражает, что он ищет. Использование семантической близости дает возможность расширять запросы и ранжировать результаты запросов. Другими словами, объект *c* может быть представлен как размытое (нечеткое) множество, включающее в себя, кроме этого объекта, семантически близкие объекты со значением семантической близости выше заданного порога.

При интеграции информации, например, при операциях над онтологиями, использование мер близости позволяет автоматически находить семантически близкие понятия, принадлежащие к разным системам концептуализации.

Ключевой момент в решении задач поиска и интеграции информации состоит в разработке количественных оценок семантической близости. В работе представлен обзор и классификация методов, которые используют знания, заложенные в

[1] Связанность (*relatedness*) — более общее понятие, чем близость [1].

концептуализации ПО — онтологии, для оценок семантической близости понятий, отношений и онтологий.

1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Онтология есть кортеж $<L := L^C \cup L^P \cup L^A \cup L^{VA}$, $C, P, A, F, G, H^C, J, H^P, I>$, где L — лексикон, т. е. множество терминов понятий (L^C), отношений (L^P), атрибутов (L^A), значений атрибутов (L^{VA}); C — множество понятий; P: $C \times C$ — множество отношений между понятиями; A: $C \times L^{VA}$ — множество атрибутов понятий; F: $L^C \to C$ — функция связи лексикона с понятиями; G: $L^P \to P$ — функция связи лексикона с отношениями; J: $L^A \to A$ — функция связи лексикона с атрибутами; H^C: $C \times C$ — таксономическая иерархия классов; H^P: $P \times P$ — иерархия отношений, I — множество экземпляров (экземпляр — понятие единичного объема).

Для отношения $p(X, Y)$ X будем называть *доменом* (*domain*) — множество, для которого допустимо использование отношения, Y — *диапазоном* (*range*) — область допустимых значений отношения. Например, для отношения «иметь публикацию» множество авторов является доменом, множество публикаций — диапазоном.

Семантическая близость онтологических термов x, y определяется с помощью функции $S(x, y) \in [0, 1]$.

2. МОДЕЛЬ ТВЕРСКИ

Традиционно объект представляется точкой в n-мерном пространстве свойств (геометрическая модель). Близость двух объектов является функцией от расстояния между соответствующими точками пространства. Расстояние между точками i и j рассчитывается по формуле:

$$D(i, j) = \left[\sum_{k=1}^{n} |X_{ik} - X_{jk}|^r \right]^{1/r},$$

где n — размерность пространства, X_{ik} — значение координаты k для объекта i, r — параметр, позволяющий использовать различные пространственные метрики. При $r = 2$ используется евклидова мера расстояния. Каждая ось интерпретируется как свойство. Качественная характеристика объекта может быть приведена к количественной одной из процедур приведения лингвистических шкал к цифровым.

В геометрической модели расстояние $d(a, b)$ между объектами a и b должно удовлетворять следующим аксиомам:

$d(a, b) \geqslant d(a, a) = 0$ (минимальность);

$d(a, b) = d(b, a)$ (симметрия);

$d(a, b) + d(b, c) \geqslant d(a, c)$ (неравенство треугольника).

Недостаток геометрической модели заключается в том, что субъективная оценка близости не всегда удовлетворяет аксиомам геометрической меры близости.

В основу многих онтологических мер близости положен теоретико-множественный подход Тверски [2], определяющий меру близости двух объектов через сопоставление свойств (*feature matching*), как одинаковых, так и различных. Если мера близости $S(a, b)$ между объектами a и b является функцией трех аргументов $A \cap B$, $A - B$, $B - A$, где A и B — множества свойств этих объектов, и удовлетворяет аксиомам монотонности, независимости, разрешимости и инвариантности [2], то она определяется формулой:

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) + \beta f(B - A),$$

где f — некоторая функция; θ, α, и $\beta \geqslant 0$ — веса для общих и различных свойств объектов. Веса позволяют определить ассиметричную меру близости. Эта модель называется *контрастной моделью* (*contrast model*) близости объектов. В развитие модели Тверски была предложена *нормализованная модель* (*ratio model*):

$$S(a, b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}.$$

В большинстве методов вычисления мер близости используется *ratio model*, а в качестве функции f — мощность множества-аргумента.

Модель Тверски более подходит для мер близости в онтологии, чем геометрическая, так как в геометрической модели при большом числе координат наглядность представления существенно уменьшается. Кроме того, свойства геометрической модели не всегда соответствуют содержательным представлениям о семантической близости. Так, при сравнении варианта с прототипом (или дочернего понятия с родительским) вариант более подобен прототипу, чем прототип варианту. Например, портрет больше похож на человека, изображенного на нем, чем наоборот [2], т. е. аксиома симметрии не соблюдается. При попарном сравнении России, Кубы и Ямайки получаем, что Ямайка подобна Кубе (географическая близость), Куба подобна России (политическая близость (в 1970-е гг.)), но Ямайка вовсе не подобна России, т. е. в данном случае отношение близости не транзитивно, и неравенство треугольника не выполняется.

Кроме того, геометрическая модель близости не «реагирует» на добавление общих свойств к объектам, что должно бы увеличивать их близость [3].

3. МЕРЫ БЛИЗОСТИ ОНТОЛОГИЧЕСКИХ ТЕРМОВ

3.1. Меры близости понятий

3.1.1. Меры, основанные на иерархических структурах

Близость двух понятий оценивается по положению вершин, соответствующих этим понятиям в иерархических онтологических структурах — главным образом в таксономической иерархии (иерархии, основанной на отношениях $IS - A$).

Простейшая мера близости такого рода основана на *длине кратчайшего пути*, измеряемого числом вершин (или ребер) в пути между двумя соответствующими вершинами таксономии [4], с учетом глубины таксономической иерархии[2] [5] — чем меньше длина пути между вершинами, тем они ближе:

$$S(c_1, c_2) = \log \frac{2N}{d(c_1, c_2)},$$

где N — глубина дерева, $d(c_1, c_2)$ — длина кратчайшего пути между вершинами. Рассмотрим, например, фрагмент таксономического дерева из *Medicine Subject Heading* онтологии (рис. 1).

Длина пути между вершинами *G01* и *G03* равна 3 (число вершин). Длина пути между вершинами *G01.273* и *G01.550* тоже 3. Однако интуитивно близость первой пары меньше, чем второй, которая принадлежит более низкому уровню и разделяет больше информации, чем первая пара. Чем ниже уровень пары вершин, тем они семантически ближе по сравнению с парами более высокого уровня. По мере спуска уровень семантической детализации увеличивается. Значит, глубина вершин должна тоже приниматься во внимание.

В работе [6] предложена мера близости, учитывающая только глубины вершин понятий:

$$S(c_1, c_2) = \frac{2 \times N(LCS)}{N(c_1) + N(c_2)}, \qquad (1)$$

где $N(LCS)$ — глубина наименьшей общей родовой вершины — ближайшего общего родителя (*least common subsumer — LCS*), $N(c_1)$ и $N(c_2)$ — глубины вершин. Например, на рис. 1 LCS(*G01.273*, *G01.550*) = *G01* и LCS(*G01*, *G03*) = *G*.

В работе [7] предложена мера близости, учитывающая два параметра: расположение вершин в таксономии (длину кратчайшего пути между вершинами) и глубину *LCS*-вершины — с учетом их весов *a* и *b*. Были исследованы линейные и нелинейные комбинации этих параметров, и наибольшая корреляция с экспертными оценками получена при применении формулы:

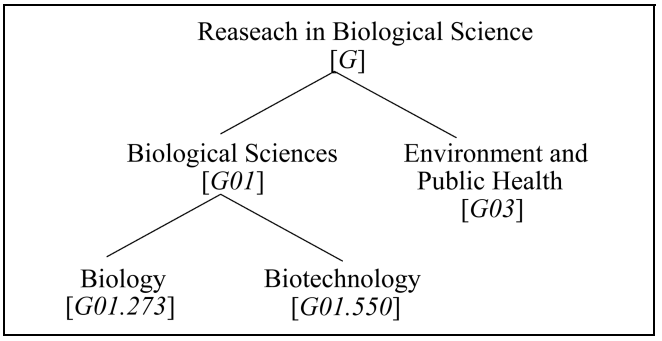

Рис. 1. Фрагмент *MeSH*-онтологии

$$S(c_1, c_2) = e^{-ad(c_1, c_2)} \frac{(e^{bN(LCS)} - e^{-bN(LCS)})}{(e^{bN(LCS)} + e^{-bN(LCS)})},$$

где d и N — длина кратчайшего пути между вершинами и глубина *LCS*-вершины, соответственно.

При оценке семантической близости понятий предлагается ограничивать конфигурацию пути: длину пути и количество перегибов [8]. Предполагается, что два понятия семантически близки, если соединены достаточно коротким путем с малым числом перегибов:

$$S(c_1, c_2) = \frac{k_0 - d(c_1, c_2) - kh}{k_0},$$

где h — число перегибов на протяжении пути, k_0 и k — константы. Например, при $k_0 = 8$, $k = 1$ максимальная длина пути ограничивается пятью шагами.

Этот подход развивается в работах [9, 10], рассматриваются либо пути, состоящие из совокупности иерархических отношений, направленных в одну сторону (например, последовательность отношений от потомка к предку), либо включающие ровно один перегиб, т. е. изменение направления движения. Рассматриваются перегибы двух видов: перегиб-сверху, например, сначала несколько отношений от видовых понятий к родовым, затем несколько отношений от родовых понятий к видовым, и перегиб-снизу — наоборот.

Для измерения близости используется *семантическое расстояние SemDist* [1], инверсное семантической близости: чем больше семантическое расстояние, тем меньше семантическая близость. Вводится понятие *общей специфичности* двух вершин $CSpec(c_1, c_2) = N - N(LCS(c_1, c_2))$, где N — глубина таксономического дерева. Чем меньше специфичность двух вершин, тем больше их близость. Семантическое расстояние является функцией двух параметров — длины кратчайшего пути между

[2] Глубина вершины — путь от вершины до корня (top-вершины), глубина таксономической иерархии — это максимальная глубина по всем вершинам.

вершинами и общей специфичности двух вершин с их весами, причем при совпадении сравниваемых вершин (единичная длина пути) семантическое расстояние нулевое:

$$SemDist(c_1, c_2) =$$
$$= \log((d(c_1, c_2) - 1)^{\alpha}(CSpec(c_1, c_2))^{\beta} + k), \quad (2)$$

где $\alpha > 0$, $\beta > 0$; $k \geqslant 1$ — константа (обеспечивает нелинейность и положительность *SemDist*), $d(c_1, c_2)$ — длина кратчайшего пути между двумя вершинами, выражаемая числом вершин.

При определении семантического расстояния в таксономическом дереве онтологии предлагается учитывать гранулированность (*granularity*)[3] кластеров [1]. Кластером называется категориальное[4] поддерево таксономического дерева. Кластер, имеющий наибольшую глубину, объявляется главным (*primary*), и остальные (вторичные) шкалируются относительно главного. При вычислении кросскластерного расстояния между вершинами, находящимися в разных кластерах, *LCS* двух вершин есть глобальный корень дерева. Тогда:

$$CSpec(c_1, c_2) = CSpec_{primary} = N_{primary} - 1,$$
$$\text{и } d(c_1, c_2) = d(c_{1,}\ root) + d(c_{2,}\ root) - 1,$$

где $N_{primary}$ — глубина главного кластера. Предлагается шкалировать путь во вторичном кластере величиной $(2N_1 - 1)/(2N_2 - 1)$, где N_1 и N_2 — глубины кластеров, $(2N_1 - 1)$ и $(2N_2 - 1)$ — максимальная длина пути между вершинами в главном и вторичном кластерах. Длина пути между вершинами с учетом гранулированности кластеров вычисляется как

$$d(c_1, c_2) = N(c_1) + \frac{2N_1 - 1}{2N_2 - 1} N(c_2) - 1.$$

В работе [11] вводится *информационное содержание IC* (*information content*) понятия, которое определяется как частота встречаемости понятия и его подпонятий в стандартном корпусе текстов и трактуется как значение вероятности $P(c)$. Если c_2 — родитель для c_1, то $P(c_1) \leqslant P(c_2)$. Значение $P(c)$ для корня иерархии равно 1. В теории информации $IC(c) = -\log P(c)$. Чем более абстрактно понятие, тем меньше значение *IC*. В мерах семантической близости используется таксономическая иерархия в онтологии как первичный источник информации и *IC* — как вторичный.

В качестве информационного содержания понятия предлагается использовать так называемое «внутреннее» информационное содержание (*intrinsic information content*), основанное только на иерархической структуре [12]:

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

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

В работе [11] близость между двумя понятиями оценивается по информационному содержанию *IC* ближайшего по таксономической иерархии общего понятия (родителя сравниваемых понятий): $S(c_1, c_2) = IC(LCS(c_1, c_2))$. В случае множественного наследования для сравниваемых понятий учитывается множество $\mathrm{Sup}(c_1, c_2)$ ближайших общих родителей (*LCS*):

$$S(c_1, c_2) = \max_{c \in \sup(c_1, c_2)} [IC(c)].$$

В работе [13] учитывается не только *IC* ближайшего общего родителя, но и *IC* самих понятий — их исходное местоположение:

$$S(c_1, c_2) = IC(c_1) - 2IC(LCS(c_1, c_2)).$$

Мера близости из работы [14] подобна мере близости из работы [6] — см. формулу (1), но вместо глубины вершины используется ее *IC* — «взвешенная» глубина:

$$S(c_1, c_2) = \frac{2IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)}.$$

Общая специфичность двух концептов *CSpec* [1] вычисляется с использованием *IC*:

$$CSpec(c_1, c_2) = ICmax - IC(LCS(c_1, c_2)),$$

где *ICmax* — максимальное значение *IC* для всех понятий онтологии, а семантическое расстояние *SemDist* вычисляется по формуле (2).

Для вычисления семантической близости понятий вводится множество вершин [15], которое содержит все вышележащие вершины (суперпонятия) по таксономической иерархии H^C по отношению к заданной вершине-понятию и заданную вершину, — так называемая «верхняя котопия» (*upwards cotopy — UC*):

$$UC(c_i, H^C) := \{c_j \in C | H^C(c_i, c_j) \vee (c_i = c_j)\}.$$

Таксономическая мера близости понятий определяется как отношение числа общих суперпонятий обеих вершин к числу всех суперпонятий обеих вершин:

$$S(c_1, c_2) = \frac{\left|UC(c_1, H^C) \cap UC(c_2, H^C)\right|}{\left|UC(c_1, H^C) \cup UC(c_2, H^C)\right|}. \quad (3)$$

При вычислении семантической близости понятий *ислам* и *христианство* по фрагменту онто-

[3] Чем больше вершин в кластере, тем больше гранулированность кластера.

[4] Вершины-категории — первый после корня уровень таксономического дерева.

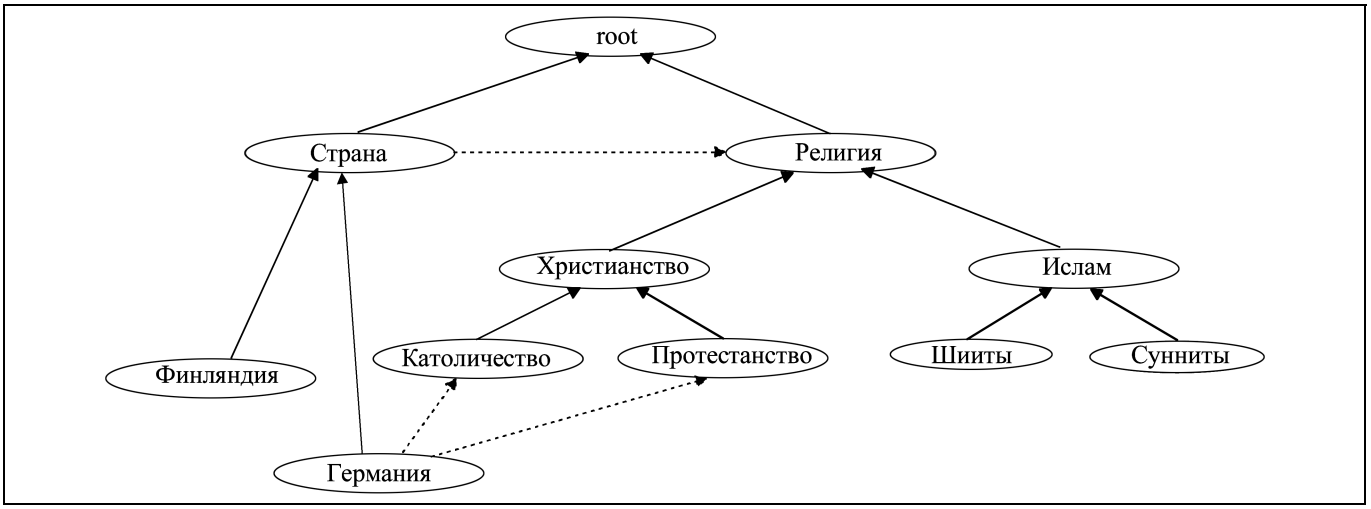

Рис. 2. Пример онтологии (пунктирными стрелками показано отношение «*верить*»)

логии, изображенной на рис. 2, по формуле (3) получаем:

UC(*христианство*) = {*христианство*, *религия*, *root*};

UC(*ислам*) = {*ислам*, *религия*, *root*};

S(*ислам*, *христианство*) = 1/2.

Недостаток большинства мер, основанных на структуре графа онтологии, состоит в симметричности (экспертные оценки показывают, что мера близости не всегда симметрична). Кроме того, эти меры независимы от контекста и чувствительны к структуре иерархии.

В работе [16] предлагается асимметричная мера семантической близости. В зависимости от направления прохождения ребрам придается разный вес, так как потомок более подобен родителю, чем наоборот.

Пусть $d = \{c_1, ..., c_n\}$ — путь в вершинах между вершинами c_1 и c_n, $s(d)$ ($g(d)$) — число непосредственных ребер от дочерней вершины до родительской (от родительской до дочерней) в этом пути, σ — вес ребра в направлении от дочерней вершины до родительской, γ — от родительской до дочерней. Тогда:

$$S(c_1, c_2) = \max_{j = 1, ..., m} \{\sigma^{s(d^j)} \gamma^{g(d^j)}\},$$

где $d^1, ..., d^m$ — пути в вершинах между вершинами c_1 и c_n.

3.1.2. Меры, использующие неиерархические отношения

Оценка близости понятий, использующая неиерархические («горизонтальные») отношения, опирается на предположение, что если два понятия имеют одно и то же отношение с третьим, то они ближе, чем два понятия, которые имеют это же отношение с различными понятиями, т. е. близость двух понятий зависит от близости понятий, с которыми они имеют отношения. Таким образом, мера близости вычисляется рекурсивно.

В работе [15] предлагается мера близости между экземплярами[5], основанная на неиерархических отношениях.

Пусть сравниваются два экземпляра i_1 и i_2. Пусть P_1 (P_2) — множество отношений, для которых i_1 (i_2) является элементом домена или диапазона. Множество $P_{CO} = P_1 \cap P_2$ будем использовать для реляционной меры близости экземпляров i_1 и i_2. В свою очередь, $P_{CO} = P_{CO-I} \cup P_{CO-O}$, где P_{CO-I} — входящие отношения (i_1 и i_2 — диапазоны), P_{CO-O} — выходящие отношения (i_1 и i_2 — домены). Множества входящих и выходящих отношений пар i_1 и i_2 — $P_{CO-I}(i_1, i_2) = P_{CO-I}(i_1) \cap P_{CO-I}(i_2)$ и $P_{CO-O}(i_1, i_2) = P_{CO-O}(i_1) \cap P_{CO-O}(i_2)$, соответственно.

Для i_1 и каждого отношения p_n из P_{CO-I} (P_{CO-O}) существует множество AS ассоциированных экземпляров i_x, таких что $p_n(i_x, i_1)$ ($p_n(i_1, i_x)$).

Для p_n из P_{CO-I}: $AS(p_n, i_1) = \{i_x | p_n(i_x, i_1)\}$.

Для p_n из P_{CO-O}: $AS(p_n, i_1) = \{i_x | p_n(i_1, i_x)\}$.

Сравнение экземпляров относительно отношения p_n будет сводиться к сравнению множеств $AS(p_n, i_1)$ и $AS(p_n, i_2)$.

[5] Экземпляр — понятие единичного объема (висячая вершина таксономической иерархии).

Близость экземпляров относительно одного отношения $S(i_1, i_2, p)$ вычисляется по формуле:

$$S(i_1, i_2, p) = \begin{cases} \dfrac{1}{|AS(p, i_1)|} \sum\limits_{a \in AS(p, i_1)} \max\limits_{b \in AS(p, i_2)} \{S(a, b)\}, \\ \quad \text{если } |AS(p, i_1)| \geq |AS(p, i_2)|, \\ \dfrac{1}{|AS(p, i_2)|} \sum\limits_{a \in AS(p, i_2)} \max\limits_{b \in AS(p, i_1)} \{S(a, b)\}, \\ \quad \text{если } |AS(p, i_1)| < |AS(p, i_2)|. \end{cases}$$

Реляционная мера близости $S(i_1, i_2)$ с учетом всех отношений из $P_{CO} = P_1 \cap P_2$ вычисляется по формуле:

$$S(i_1, i_2) = \frac{1}{|P_{CO-I}| + |P_{CO-O}|} \times$$
$$\times \left(\sum_{p \in P_{CO-I}} S(i_1, i_2, p) + \sum_{p \in P_{CO-O}} S(i_1, i_2, p) \right).$$

Задается максимальная глубина рекурсии. При ее достижении близость экземпляров оценивается, например, по таксономической мере.

Вычислим реляционную меру близости между вершинами *Финляндия* и *Германия* для отношения *верить* (см. рис. 2) при глубине рекурсии 1:

AS (*верить*, *Германия*) = {*Католичество*, *Протестантство*};

AS (*верить*, *Финляндия*) = {*Протестантство*};

S(*Протестантство*, *Протестантство*) = *1*;

AS (*верить*, *Протестантство*) = {*Финляндия*, *Германия*};

AS (*верить*, *Католичество*) = {*Германия*};

S(*Католичество*, *Протестантство*) = 0,75.

Следовательно, *S*(*Финляндия*, *Германия*) = (1 + + 0,75)/2 = 0,87.

3.1.3. Меры, учитывающие значения атрибутов

Атрибутивная мера близости основана на близости значений общих атрибутов понятий. Атрибуты можно рассматривать как отношения, диапазоны которых — литералы, числа, строки и другие типы данных.

В работе [15] предлагается мера близости между экземплярами, учитывающая значения атрибутов.

Пусть A — множество атрибутов; $A(i)$ — множество атрибутов экземпляра i; $A_{CO} = A(i_1) \cap A(i_2)$ — множество общих атрибутов экземпляров i_1 и i_2; $AS(a, i) = \{l_x | a(i, l_x)\}$ — множество значений атрибута a для экземпляра i; $LS(l_1, l_2, a) \in [0, 1]$ — близость значений l_1, l_2 атрибута a, причем минимальная близость при сравнении атрибутов равна нулю.

В качестве меры близости для строковых данных можно использовать пронормированное редакторское расстояние[6] [17], для чисел — инверсию разности, пронормированную максимальным значением атрибута. Тогда близость двух экземпляров i_1 и i_2 в отношении одного атрибута $S(i_1, i_2, a)$ может быть вычислена так [15]:

$$S(i_1, i_2, a) =$$
$$= \begin{cases} \dfrac{1}{|AS(a, i_1)|} \sum\limits_{l \in AS(a, i_1)} \max\limits_{m \in AS(a, i_2)} \{LS(l, m, a)\}, \\ \quad \text{если } |AS(a, i_1)| \geq |AS(a, i_2)|, \\ \dfrac{1}{|AS(a, i_2)|} \sum\limits_{l \in AS(a, i_2)} \max\limits_{m \in AS(a, i_1)} \{LS(l, m, a)\}, \\ \quad \text{если } |AS(a, i_1)| < |AS(a, i_2)|. \end{cases}$$

Атрибутивная мера близости экземпляров $S(i_1, i_2)$ с учетом всех атрибутов из $A_{CO} = A_1 \cap A_2$ вычисляется по формуле:

$$S(i_1, i_2) = \frac{1}{|A_{CO}|} \sum_{a \in A_{CO}} S(i_1, i_2, a).$$

3.1.4. Гибридные меры

Гибридные меры являются свертками перечисленных мер близости понятий. Чем полнее будут учитываться характеристики двух сущностей с разных точек зрения, тем более качественную меру близости можно получить. В связи с этим наиболее перспективными представляются именно гибридные меры, сочетающие несколько подходов.

Чаще всего в гибридных мерах используется аддитивная свертка:

$$S(c_1, c_2) = \sum_{i=1}^{n} w_i S^i(c_1, c_2),$$

где S^i — мера близости по определенному критерию, вес w_i определяет относительную важность критерия, сумма весов равна 1, n — число критериев. Распространенная модификация аддитивной свертки основана на использовании сигмоидальной функции:

$$S(c_1, c_2) = \sum_{i=1}^{n} w_i \text{sig}(S^i(c_1, c_2)),$$

где $\text{sig}(x) = 1/(1 + e^{-ax})$, $a > 0$.

Сигмоидальная функция позволяет повысить веса мер с большими значениями и практически пренебречь мерами с малыми значениями.

[6] Редакторское расстояние (*ed*) между строками измеряется в минимальном числе символов, которые надо удалить или добавить при переходе от одной строки к другой. Например, *ed*(«*TopHotel*», «*Top Hotel*») = 1.

Веса могут определяться интерактивно экспертами и (или) пользователями, а также автоматически с помощью обучаемой нейронной сети [18] или генетического алгоритма [19].

В работе [20] гибридная мера оценивает близость понятий, комбинируя *ratio*-модель Тверски с мерой, основанной на иерархических структурах. Мера близости представляет собой аддитивную свертку трех составляющих по разным типам свойств: части (компоненты понятий); функции (функциональные возможности); атрибуты (остальные свойства):

Для вычисления значения меры по каждому типу используется *ratio*-модель:

$$S(c_1, c_2) = w_p S^p(c_1, c_2) + w_f S^f(c_1, c_2) + w_a S^a(c_1, c_2).$$

$$S^t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1, c_2)|C_1 - C_2| + (1 - \alpha(c_1, c_2))|C_2 - C_1|},$$

где S^p, S^f и S^a — составляющие близости по частям, функциям и атрибутам, $t = \{p, f, a\}$, c_1 и c_2 — сравниваемые понятия, C_1 и C_2 — множества различимых свойств данного типа сравниваемых понятий, α определяется через длины путей между вершинами, соответствующими сравниваемым понятиям, и ближайшим общим родителем (LCS) в иерархии:

$$\alpha(c_1, c_2) = \begin{cases} \dfrac{d(c_1, LCS)}{d(c_1, c_2)}, \text{ если } d(c_1, LCS) \le d(c_2, LCS), \\ 1 - \dfrac{d(c_1, LCS)}{d(c_1, c_2)}, \text{ если } d(c_1, LCS) > d(c_2, LCS). \end{cases}$$

Из определения α следует несимметричность меры близости. При сравнении подкласса и надкласса (варианта и прототипа) $\alpha = 0$, т. е. не учитываются свойства варианта, которыми не обладает прототип, а при сравнении прототипа и варианта они учитываются ($\alpha = 1$). Таким образом, вариант ближе к прототипу, чем прототип к варианту, что согласуется с интуицией.

Гибридная мера, предложенная в работе [15], содержит оценку близости экземпляров, состоящую из трех частей — таксономической, реляционной и атрибутивной (см. п. 3.1.1 — 3.1.3), представленную аддитивной сверткой:

$$S(i_1, i_2) = tS^t(i_1, i_2) + pS^p(i_1, i_2) + aS^a(i_1, i_2),$$

где S^t, S^p и S^a — таксономическая, реляционная и атрибутивная составляющие близости, соответственно, t, p и a — веса.

3.2. Меры близости отношений

До сих пор рассматривались меры близости между понятиями. Рассмотрим меры близости между отношениями.

В работе [21] близость двух бинарных отношений p_1 и p_2 вычисляется как среднее геометрическое (а не арифметическое среднее) близости понятий доменов и диапазонов этих отношений. Это основано на интуиции: если близость по одному из компонентов достигает значения 0, то и общая близость двух предикатов также достигает значения 0:

$$S(p_1, p_2) := \sqrt{S(d(p_1), d(p_1))S(r(p_1), r(p_2))},$$

где $d(p)$ — понятие домена отношения p, $r(p)$ — диапазон отношения p, $S(c_1, c_2)$ — близость понятий в таксономической иерархии.

В работе [22] близость двух n-арных отношений p_1 и p_2 вычисляется как среднее геометрическое близости понятий соответствующих аргументов сравниваемых отношений. В измерении близости двух n-арных отношений учитываются как близости пар соответствующих аргументов сравниваемых отношений, так и близость самих отношений по иерархии отношений H^R, измеряемая аналогично таксономической близости понятий в иерархии понятий H^C, см. формулу (3):

$$S(p_1, p_2, H^R) = \frac{\left|(UC(p_1, H^R) \cap UC(p_2, H^R))\right|}{\left|(UC(p_1, H^R) \cup UC(p_2, H^R))\right|},$$

т. е. отношение числа одинаковых надотношений (всех предков в иерархии отношений) обеих вершин к числу всех надотношений обеих вершин, и с учетом аргументов i, j:

$$S(p_1(i_1, ..., i_n), p_2(i_1, ..., i_n)) = \sqrt[n+1]{S(p_1, p_2, H^R)S(i_1, j_1)...S(i_n, j_n)}.$$

3.3. Кросс-онтологические меры

Трудности сравнения разных онтологий ПО (различных концептуализаций одной и той же ПО) заключаются, с одной стороны, в различии используемых лексиконов термов, с другой — в различных путях концептуализации и ее представления.

3.3.1. Меры близости понятий

Отображение онтологии O_1 на онтологию O_2 означает попытку найти для каждого из концептов онтологии O_1 подобный ему концепт в онтологии O_2, т. е. возникает задача поиска наилучших кандидатов для установления соответствий онтологий.

В работе [1] таксономии двух онтологий связываются через «мосты» («якори») — вершины, соот-

ветствующие эквивалентным понятиям, которые определяются с использованием синсетов (множеств синонимов) из онтологий *MeSH* и *WordNet*.

Например, вершины a_9 и b_2 из разных онтологий (рис. 3, *а*) сливаются в одну вершину-мост (рис. 3, *б*).

Параметры меры — длина кратчайшего пути между двумя вершинами, соответствующими сравниваемым понятиям, и общая специфичность вершин — рассчитываются с учетом введенных мостов. Ближайшим общим родителем (*LCS*) для сравниваемых понятий из разных онтологий O_1 и O_2 является ближайший общий родитель первого элемента сравниваемой пары и вершины-моста:

$$LCS(c_{1i}, c_{2j}) = LCS(c_{1i}, Bridge).$$

Путь между двумя сравниваемыми вершинами проходит через вершину-мост и через две онтологии с разными глубинами таксономии. Часть длины пути во вторичной онтологии «шкалируется» длиной пути в первичной онтологии:

$$d(c_{1i}, c_{2j}) = d(c_{1i}, Bridge) + \frac{2N_1 - 1}{2N_2 - 1} d(c_{2j}, Bridge) - 1,$$

где N_1 и N_2 — глубина соответствующих таксономий, $d(c_{1i}, Bridge)$ и $d(c_{2j}, Bridge)$ — длины путей от каждой вершины до вершины-моста, вычитание 1, так как вершина-мост считается дважды.

Общая специфичность *CSpec* двух вершин и семантическое расстояние *SemDist* при наличии одной вершины-моста рассчитываются по формулам:

$$CSpec(c_{1i}, c_{2j}) = N_1 - N(LCS(c_{1i}, Bridge)),$$

$$SemDist(c_{1i}, c_{2j}) =$$

$$= \log((d(c_{1i}, c_{2j}) - 1)^{\alpha}(CSpec(c_{1i}, c_{2j})^{\beta} + k).$$

Две онтологии могут иметь много пар эквивалентных вершин, образующих вершины-мосты.

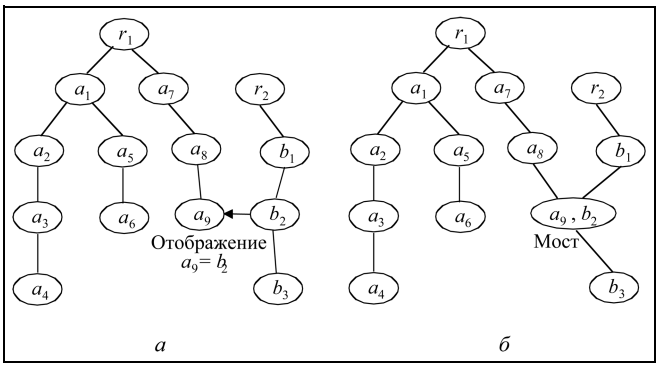

Рис. 3. Связь двух онтологических фрагментов: слияние двух вершин из разных онтологий (*а*) в одну вершину-мост (*б*)

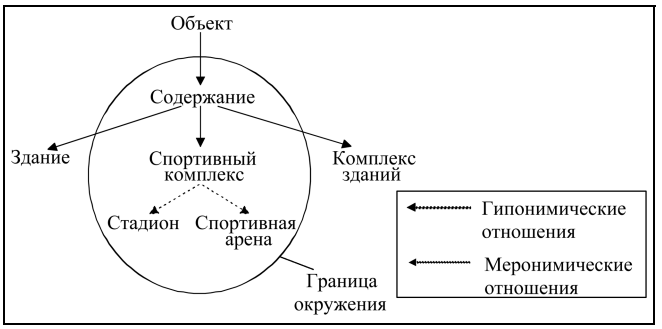

Рис. 4. Множество семантического соседства вершины *Спортивный комплекс* радиуса 1

Для нескольких мостов, связывающих две онтологии:

$$SemDist(c_{1i}, c_{2j}) = \min_{k = 1, ..., n} \{SemDist_k(c_{1i}, c_{2j})\},$$

где n — число вершин-мостов. Таким образом, семантическое расстояние между двумя сравниваемыми вершинами (понятиями в разных онтологиях) определяется как минимальное из множества расстояний по всем вершинам-мостам.

Для вычисления кросс-онтологической меры таксономии двух онтологий связываются через вводимый корень обеих иерархий [20]. Близость понятий в двух онтологиях вычисляется с учетом лексической близости[7] термов, соответствующих сравниваемым вершинам, семантической близости соседних (в заданном радиусе окрестности вершины в иерархии) вершин, а также близости различимых свойств понятий, соответствующих сравниваемым вершинам.

Множество семантического соседства вершины радиуса r в объединенной онтологической иерархии, основанной на таксономических (*IS-A*) и меронимических (*PART-OF*) отношениях — множество вершин-понятий, расстояние (в ребрах) до которых от заданной вершины меньше r (рис. 4):

$$N(c, r) = \{c_i | d(c, c_i) \leqslant r\}.$$

Сравнение семантических соседей сравниваемых объектов основано на том, что чем ближе объекты, с которыми сравниваемые объекты связаны таксономическими и меронимическими отношениями, тем сравниваемые объекты ближе.

Мера близости между объектами в разных онтологиях представляет собой взвешенную сумму значений близости для каждой составляющей: лек-

[7] Основывается, например, на редакторском расстоянии $ed(l_i, l_j)$ между двумя терминами l_i, l_j (сколько символов надо добавить, удалить или изменить, чтобы сделать из одной лексемы другую): $LS(l_i, l_j) = \max\left(0, \frac{\min(|l_i|, |l_j|) - ed(l_i, l_j)}{\min(|l_i|, |l_j|)}\right) \in [0, 1]$.

сическое соответствие, соответствие свойств, соответствие семантического соседства. Веса определяют относительную важность каждой составляющей:

$$S(c_{1i}, c_{2j}) = w_l S^l(c_{1i}, c_{2j}) + w_u S^u(c_{1i}, c_{2j}) + w_n S^n(c_{1i}, c_{2j}),$$

где c_{1i} и c_{2j} — сравниваемые вершины-понятия из онтологий O_1 и O_2, соответственно S^l, S^u и S^n — составляющие близости по лексикону, свойствам и семантическому соседству, w — веса составляющих близости объектов в интервале [0, 1], их сумма равна единице.

Для вычисления значения по каждой составляющей меры близости используется *ratio*-модель [2], адаптированная для каждого типа близости:

$$S^t(c_{1i}, c_{2j}) = \frac{|C_{1i} \cap C_{2j}|}{|C_{1i} \cap C_{2j}| + \alpha(c_{1i}, c_{2j})|C_{1i} - C_{2j}| + (1 - \alpha(c_{1i}, c_{2j}))|C_{2j} - C_{1i}|},$$

где $t = \{l, u\}$, c_{1i} и c_{2j} — сравниваемые понятия; C_{1i} и C_{2j} — множества слов в названиях вершин (или в множестве их синонимов) при $t = l$, множества различимых свойств сравниваемых вершин-понятий при $t = u$.

Близость по семантическому соседству S^n радиуса r между вершинами c_{1i} и c_{2j} из онтологий O_1 и O_2 есть функция от так называемого аппроксимирующего пересечения ($\cap n$) между множествами семантического соседства для c_{1i} и c_{2j}, которое характеризует семантическую близость окрестности первой вершины по отношению к окрестности второй:

$$c_{1i} \cap_n c_{2j} = \left[\sum_{k \le n} \max_{l \le m} S(c_{1i}^k, c_{2j}^l)\right] - \varphi S(c_{1i}, c_{2j}),$$

$$\varphi = \begin{cases} 1, \text{ если } S(c_{1i}, c_{2j}) = \max\limits_{l \le m} S(c_{1i}, c_{2j}^l), \\ 0 \text{ в противном случае,} \end{cases}$$

где n и m — число вершин в сравниваемых окрестностях.

Близость по семантическому соседству определяется по формуле:

$$S^n(c_{1i}, c_{2j}) = [c_{1i} \cap_n c_{2j}]/[c_{1i} \cap_n c_{2j} + \alpha(c_{1i}, c_{2j}) \times \delta(c_{1i}, c_{1i} \cap_n c_{2j}, r) + (1 - \alpha(c_{1i}, c_{2j})\delta(c_{1i}, c_{1i} \cap_n c_{2j}, r)],$$

$$\delta(c_{1i}, c_{1i} \cap_n c_{2j}, r) = \begin{cases} |N(c_{1i}, r)| - c_{1i} \cap_n c_{2j}, \text{ если } |N(c_{1i}, r)| < c_{1i} \cap_n c_{2j}, \\ 0 \text{ в противном случае.} \end{cases}$$

Коэффициент α для *ratio*-модели [2] — функция от глубин вершин-понятий — кратчайшего расстояния от данной вершины до введенного корня, который является единственным общим надклассом для вершин-понятий из разных онтологий:

$$\alpha(c_{1i}, c_{2j}) = \begin{cases} \dfrac{N(c_{1i})}{N(c_{1i}) + N(c_{2j})}, N(c_{1i}) \le N(c_{2j}), \\ 1 - \dfrac{N(c_{1i})}{N(c_{1i}) + N(c_{2j})}, N(c_{1i}) > N(c_{2j}). \end{cases}$$

Существенную роль в этой мере играет различие путей формализации и ее представления в онтологиях сравниваемых понятий (например, различная классификация различимых свойств или несовпадение множества свойств подобных понятий). Лексика определяет все три составляющие меры близости, поэтому при использовании этой меры существенно соответствие лексиконов онтологий сравниваемых понятий.

В работе [23] мера близости между термами разных онтологий разбивается на элементарные критерии, для свертки которых используются аддитивная или сигмоидальная функции:

- лексическая близость[8];
- отношение «*SameClassAs*» и «*SameIndividualAs*»;
- близость свойств — атрибутов и отношений;
- близость доменов и диапазонов для отношений;
- близость родительских понятий;
- близость родительских атрибутов;
- близость дочерних понятий;
- близость дочерних атрибутов;
- близость понятий одного уровня;
- близость экземпляров понятий;
- близость ограничений и правил;
- другие.

Расчет близости между понятиями в разных онтологиях представляет собой итерационный процесс, поскольку многие из рассмотренных критериев близости двух понятий основываются на близости других сущностей (понятий, свойств, экземпляров). На первой итерации используются критерии близости, которые не основываются на других критериях, т. е. критерии, основанные на лексике и отношениях «*SameClassAs*» и «*SameIndividualAs*».

3.3.2. Меры близости онтологий

В работе [21] рассматриваются методы измерения близости между онтологиями на двух уровнях — вербальном и концептуальном. На вербальном уровне сравниваются лексиконы двух онтологий, на концептуальном — сравниваются таксономии понятий и других отношений двух он-

[8] Здесь лексическая близость термов основывается на редакторском расстоянии между терминами.

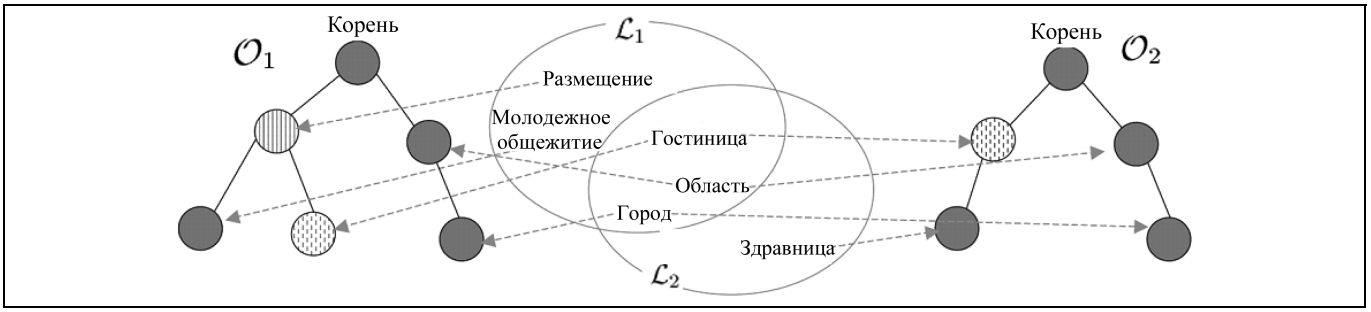

Рис. 5. Пример двух онтологий O_1 и O_2

тологий. Под лексиконом подразумевается множество терминов понятий, отношений и атрибутов онтологии $L := L^C \cup L^P \cup L^A$. Лексикон и понятия (отношения) связаны через функцию $F(G)$, которая ставит термины в соответствие понятиям (отношениям) в онтологии.

Лексическая близость между двумя онтологиями вычисляется с использованием лексической близости между терминами:

$$S(L_1, L_2) = \frac{1}{|L_1|} \sum_{l_i \in L_1} \max_{l_j \in L_2} LS(l_i, l_j),$$

где L_1, L_2 — лексиконы сравниваемых онтологий.

Концептуальная близость онтологий оценивается с двух сторон — близость таксономий и близость отношений.

Для вычисления близости таксономий используется множество вершин, которое содержит все выше- и нижележащие вершины по таксономической иерархии H^C по отношению к заданной вершине-понятию, — семантическая котопия (*SC — semantic cotopy*):

$$SC(c_i, H^C) := \{c_j \in C | H^C(c_i, c_j) \vee H^C(c_i, c_j)\}.$$

Для множества понятий:

$$SC(\{c_1, ..., c_n), H^C) := \bigcup_{i := 1, ..., n} SC(c_i, H^C).$$

Если термин l используется в обеих онтологиях O_1 и O_2 с таксономиями H_1^C и H_2^C, то таксономическая близость онтологий относительно этого термина определяется по формуле:

$$S'(l, O_1, O_2) :=$$

$$:= \frac{\left|F_1^{-1}(SC(F_1(\{l\}), H_1^C)) \cap F_2^{-1}(SC(F_2(\{l\}), H_2^C))\right|}{\left|F_1^{-1}(SC(F_1(\{l\}), H_1^C)) \cup F_2^{-1}(SC(F_2(\{l\}), H_2^C))\right|}.$$

Чем больше одинаковых терминов в семантических котопиях понятий, названных этим термином в обеих онтологиях, тем больше таксономическая близость онтологий относительно этого термина.

Если термин есть только в одной онтологии, а во второй отсутствует, то сравниваются термины семантических котопий понятия, соответствующего термину l, и всех понятий во второй онтологии, и в качестве таксономической близости относительно этого термина берется максимум:

$$S''(l, O_1, O_2) :=$$

$$:= \max_{c \in C_2} \frac{\left|F_1^{-1}(SC(F(\{l\}), H_1^C)) \cap F_2^{-1}(SC(c, H_2^C))\right|}{\left|F_1^{-1}(SC(F(\{l\}), H_1^C)) \cup F_2^{-1}(SC(c, H_2^C))\right|}.$$

Такая операция делается для каждого термина в первой онтологии, после чего вычисляется таксономическая близость для двух онтологий как среднее значение по лексикону понятий первой онтологии:

$$S_t(O_1, O_2) = \frac{1}{|L_1^C|} \sum_{l_i \in L_1^C} S(l, O_1, O_2),$$

где

$$S(l, O_1, O_2) = \begin{cases} S'(l, O_1, O_2), \text{ если } l \in L_2^C, \\ S''(l, O_1, O_2), \text{ если } l \notin L_2^C. \end{cases}$$

Рассмотрим пример. На рис. 5 представлены две онтологии O_1 и O_2, которым соответствуют частично пересекающиеся лексиконы L_1 и L_2. Таксономическая близость S' («гостиница, H_1, H_2) определяется следующим образом:

$F_1^{-1}(SC(F_1(\{\text{«гостиница»}\}), H_1^C)) = \{\text{«гостиница», «размещение»}\}$ и

$F_2^{-1}(SC(F_2(\{\text{«гостиница»}\}), H_2^C)) = \{\text{«здравница», «гостиница»}\}$.

Следовательно, $S'(\text{«гостиница»}, H_1^C, H_2^C) = 1/3$.

Для термина «размещение», которая есть только в лексиконе L_1^C:

$F_1^{-1}(SC(F_1(\{\text{«размещение»}\}), H_1^C)) = \{\text{«молодежное общежитие», «размещение», «гостиница»}\}$.

В онтологии O_2 нет понятия, соответствующего «размещению», поэтому берется то, которое даст наилучший результат. В данном случае это «гостиница»:

$F_2^{-1}(SC(F_2(\{\text{«гостиница»}\}))) = \{\text{«здравница», «гостиница»}\}$.

Таким образом: $S''(\text{«размещение»}, H_1^C, H_2^C) = 1/4$.

Для оценки близости отношений используется понятие верхней котопии — все предки понятия c_i. Вводится таксономическая близость понятий с учетом лексиконов[9]:

$$S^l(c_1, O_1, c_2, O_2) := \frac{\left|F_1^{-1}(UC(c_1, H_1^C)) \cap F_2^{-1}(UC(c_2, H_2^C))\right|}{\left|F_1^{-1}(UC(c_1, H_1^C)) \cup F_2^{-1}(UC(c_2, H_2^C))\right|}.$$

Близость отношений из разных онтологий вычисляется как среднее геометрическое близостей понятий для доменов и для диапазонов:

$$S'(p_1, O_1, p_2, O_2) := \sqrt{S^l(d(p_1), O_1, d(p_2), O_2) \cdot S^l(r(p_1), O_1, r(p_2), O_2)},$$

где $d(p)$ — домен отношения p, $r(p)$ — диапазон отношения p.

Для отношений, присутствующих в обеих онтологиях:

$$S''(l, O_1, O_2) := \frac{1}{\left|G_1(\{l\})\right|} \sum_{R_1 \in G_1(\{l\})} \max_{R_2 \in G_2(\{l\})} \{S'(R_1, O_1, R_2, O_2)\},$$

так как в онтологиях может быть несколько отношений с одинаковыми названиями.

Если отношение присутствует только в одной онтологии O_1, то:

$$S'''(l, O_1, O_2) := \frac{1}{\left|G_1(\{l\})\right|} \sum_{R_1 \in G_1(\{l\})} \max_{p_2 \in P_2} \{S'(p_1, O_1, p_2, O_2)\}.$$

Тогда реляционная близость между онтологиями будет вычисляться по формуле:

$$S^p(O_1, O_2) := \frac{1}{\left|L_1^p\right|} \sum_{l_i \in L_1^p} S(l, O_1, O_2)\},$$

где

$$S(l, O_1, O_2) := \begin{cases} S''(l, O_1, O_2), \text{ если } l \in L_2^p, \\ S'''(l, O_1, O_2), \text{ если } l \notin L_2^p. \end{cases}$$

Предлагается подход к семантическому ранжированию ответов на запрос к Web-порталу [22]. Задача ранжирования может быть сведена к сравнению пар баз знаний[10] — каждого результата запроса (QKB_i) и портала (KB). Один ответ — результат запроса, преобразованный в предложения *F-Logic*, интерпретируется как база знаний. Базы знаний запроса, результата и портала имеют один лексикон и одни понятия, поэтому сравниваются только отношения. Ранжирование производится по значению близости баз знаний результатов запроса к базе знаний портала, причем понятие близости между двумя базами знаний сводится к определению близости отношений:

$$S(QKB_i, KB) = \frac{1}{\left|P_Q\right|} \sum_{p_j \in P_Q} \max_{p_i \in P} S(p_j, p_i),$$

где P_Q — множество отношений базы знаний результата запроса QKB_i, P — множество отношений базы знаний портала, $S(p_j, p_i)$ — близость двух n-арных отношений p_j и p_i.

4. СПОСОБЫ ОЦЕНКИ МЕР БЛИЗОСТИ ПОНЯТИЙ

Отсутствие «золотого стандарта» меры семантической близости — хорошо известная проблема. Усилия многих исследователей направлены на оценку и сравнение мер семантической близости.

В работе [24] выделяются три подхода к оценке мер семантической близости.

Первый подход — теоретическая проверка желательных математических свойств предлагаемой меры [2, 14]. Например, является ли мера симметричной или асимметричной, гладкой, имеет ли особые точки и т. д. Такой анализ может служить грубым фильтром при оценке меры.

Второй подход — сравнение с экспертными оценками [1, 20, 24]. Насколько экспертные оценки считаются корректными, настолько этот подход дает правильную оценку мере. Главная помеха — трудность проектирования психолингвистического эксперимента и обоснования его результатов для получения достоверных оценок.

Третий подход — оценка эффективности мер в рамках конкретного приложения [24].

В рамках второго подхода рассматриваются разные способы проведения экспериментов для получения экспертных оценок. Можно выделить три

[9] В работе [15] введена таксономическая близость понятий как отношение числа совпадающих понятий к общему числу понятий в верхних котопиях, а здесь таксономическая близость понятий вводится как отношение числа совпадающих терминов понятий к общему числу терминов понятий в верхних котопиях.

[10] Считается, что база знаний и онтология — одно и то же.

способа получения информации о семантической близости терминов от экспертов:

— для множества пар понятий по заданной шкале оценивается семантическая близость;

— понятия из выбранного множества ранжируются по семантической близости с заданным понятием;

— из заданного множества понятий, достаточно близких к некоторому понятию, выбирается ближайшее.

Мера семантической близости оценивается степенью совпадения с экспертными оценками, выраженной значением корреляции или в процентах совпадения.

Для автоматизированной оценки мер близости существуют электронные ресурсы:

— онтологии (например, *WordNet*, *MeSH*);

— корпуса текстов (например, *PubMed — http://www.ncbi.nlm.nih.gov/pubmed/*);

— наборы пар понятий (*Data Set*) экспертными оценками семантической близости (например, [25, 26]);

— наборы понятий, семантически близких к фиксированному понятию с выбранным ближайшим (например, *TOEFL_Synonym_Questions — http://www.aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions*);

— программные системы (например, *WordNet:similarity — http://wn-similarity.sourceforge.net/*).

Система *WordNet:similarity* позволяет оценить заданную меру (из существующего списка или пользовательскую), используя онтологию *WordNet*, а также корпуса текстов, которые можно выбрать самому.

Результаты экспериментов показали, что оценка мер семантической близости зависит от контекста сравнения понятий, при учете которого корреляция увеличивалась [20]. Также замечено, что специалисты разного профиля в одной ПО могут давать разные оценки семантической близости. Например, в работе [1] семантическую близость оценивали практикующие врачи и эксперты в области медицины, и оценки различались. Кроме того, оценки зависят от корпуса текстов, а также от его размера.

Во многих работах оцениваются результаты сравнения ряда мер семантической близости. Так, была проведена серия экспериментов [1], в которых сравнивались различные меры, основанные на иерархических онтологических структурах (см. п. 3.1.1). Результаты, наиболее близкие к экспертным оценкам, дала мера самого автора, а также мера из работы [13]. Среди мер, не использующих информационный контент, лучше всего показал себя метод из работы [5]. В исследовании [27] среди иерархических мер также лучшей оказалась мера из работы [5].

Однако общепринятая методология оценки мер семантической близости отсутствует. К настоящему времени каких-либо рекомендаций по использованию той или иной меры для конкретных задач не прослеживается.

5. О РЕКОМЕНДАЦИЯХ ПО ПРИМЕНЕНИЮ МЕР БЛИЗОСТИ

Как уже было сказано, меры семантической близости используются в широком спектре задач. Эффективность применения той или иной меры, по нашему мнению, зависит как от задачи, так и от пользователя — автора запросов. Этот вопрос не рассматривается в известной авторам данного обзора литературе и ожидает своего исследования. Например, факторами, определяющими выбор меры семантической близости, применяемой системой для ответа на запрос, могут быть предпочтения пользователя:

- выбор критериев, составляющих меру близости, как то: таксономические отношения между понятиями и (или) конкретный набор отношений и (или) атрибутов;
- выбор значимости критериев — коэффициентов важности в гибридной аддитивной мере близости.

Интерактивный интерфейс при задании запроса поможет пользователю определить свои предпочтения при выборе меры семантической близости.

ЗАКЛЮЧЕНИЕ

В работе дан обзор и классификация существующих мер семантической близости. Рассмотрена семантическая близость между онтологическими термами — понятиями, отношениями и экземплярами — и между онтологиями. В основу классификации мер положены характеристики термов — свойства, отношения и их типы, атрибуты понятий — и структура онтологии.

При построении мер близости рассматриваются разные подходы:

- теоретико-множественный подход — пересечение одинаковых и различных свойств;
- информационный подход — статистика на стандартных корпусах текстов (частота встречаемости терминов) или на таксономической иерархии;
- структурный подход — характеристики онтологических структур (длина пути, глубина иерархии и т. д.).

При кросс-онтологическом подходе важный момент состоит в соответствии лексиконов. Для установления соответствия понятий рассматриваются «окрестности» понятий в онтологических структурах.

Гибридные меры используют различные комбинации мер близости понятий. Наиболее эффективными представляются именно гибридные меры, сочетающие несколько критериев, так как чем полнее будут учитываться характеристики двух сущностей с разных точек зрения, тем более качественную меру близости можно получить. Чаще всего в качестве гибридной меры используется аддитивная свертка мер. Использование сигмоидальной функции в гибридных мерах позволяет повысить веса мер, имеющих большие значения, и практически пренебречь мерами с малыми значениями. Веса определяются на основе суждения экспертов и (или) обучающих алгоритмов. Дан обзор способов оценки мер близости.

ЛИТЕРАТУРА

1. *Nguyen H.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, December 2006.
2. *Tversky A.* Features of similarity // Psychological Review. — 1977. — 84(4). — P. 327—352.
3. *Goldstone R.L., Son J.* Similarity // In Cambridge Handbook of Thinking and Reasoning / K. Holyrak & R. Morrison (Eds.) — Cambridge: Cambridge University Press. 2005. — P. 13—36. — URL: http://cognitivn.psych.indiana.edu/rgoldsto/pdfs/similarity2004.pdf.
4. *Development* and Application of a Metric on Semantic Net / Rada R., et al. // IEEE Trans. on Systems, Man and Cybernetics. — 1989. — 19(1). — P. 17—30.
5. *Leacock C., Chodorow M.* Combining local context and WordNet similarity for word sense identification // WordNet: An electronic lexical database / Fellbaum C. (ed.). — Cambridge, MA: MIT press, 1998. — P. 265—283.
6. *Wu Z., Palmer M.* Verb semantics and lexical selection // 32nd Annual Meeting of the Association for Computational Linguistics, 1994. — P. 133—138.
7. *Li Y., Bandar Z.A., McLean D.* An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources // IEEE Trans. on Knowledge and Data Engineering, 2003. — 15(4). — P. 871—882.
8. *Hirst G., St-Onge D.* Lexical Chains as representation of context for the detection and correction malapropisms // WordNet: An electronic lexical database and some of its applications / C. Fellbaum (ed.). — Cambrige, MA: The MIT Press, 1997.
9. *Лукашевич Н.В., Добров Б.В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. семинара «Диалог'2002». — М., 2002. — Т. 2. — С. 338—346.
10. *Лукашевич Н.В., Добров Б.В.* Разрешение лексической многозначности на основе тезауруса предметной области // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.). — М., 2007. — С. 400—406.
11. *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.
12. *Seco N., Veale T., Hayes J.* An intrinsic information content metric for semantic similarity in WordNet // Proc. of the 16th European Conference on Artificial Intelligence,Valencia, Spain, 23—27 August 2004. — P. 1089—1090.
13. *Jiang J. and Conrath D.* Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy // Intern. Conf. on Computational Linguistics (ROCLING X), Taiwan, 1997. — P. 19—35.
14. *Lin D.* An information-theoretic definition of similarity // Proc. of the Int'l Conference on Machine Learning, 1998.
15. *Maedche A., Zacharias V.* Clustering Ontology-Based Metadata in the Semantic Web / Proceedings PKDD-2002, LNAI 2431, 2002. — P. 348—360.
16. *Bulskov H., Knappe R., Andreasen T.* On Measuring Similarity for Conceptual Querying / FQAS 2002, LNAI 2522, 2002. — P. 100—111.
17. *Levenshtein I.V.* Binary Codes capable of correcting deletions, insertions, and reversals // Cybernetics and Control Theory. — 1966. — 10(8). — P. 707—710.
18. *Ehrig M., Sure Y.* An Ontology Mapping — An Integrated Approach / The semantic web: Research and applications. — Berlin: Springer, 2004. — P. 3—13.
19. *Supervised* Learning of Term Similarities / I. Spasi´c, G. Nenadi´c, K. Manios, and Ananiadou S. // IDEAL 2002, H. Yin et al. (eds.), LNCS 2412, 2002. — P. 429—434.
20. *Rodríguez M.A.* Assessing Semantic Similarity among Spatial Entity Classes / A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (in Spatial Information Science and Engineering), The Graduate School University of Maine May, 2000.
21. *Maedche A., Staab S.* Measuring Similarity between Ontologies / EKAW 2002, A. Gomez-Perez and V.R. Benjamins (eds.), LNAI 2473, 2002. — P. 251—263.
22. *A framework* for developing semantic portals / N. Stojanovic, et al. // Proc. of the first international ACM conferences on knowledge capture K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada, http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/sealkcap2.pdf.
23. *Карпенко А.П., Сухарь Р.С.* Методы отображения онтологий. Обзор / Наука и образование: электронное научно-техническое издание, 2009, 1 (январь). — URL: http://elibrary.ru/item.asp?id=11991555.
24. *Budanitsky A., Hirst G.* EvaluatingWordNet-based Measures of Lexical Semantic Relatedness // Computational Linguistics. — 2006. — Vol. 32, N 1. — P. 13—47.
25. *Placing* Search in Context: The Concept Revisited / L. Finkelstein, et al. // ACM Trans. on Information Systems. — January 2002. — 20(1). — P. 116 — 131, http://www.cs.technion.ac.il/~gabr/papers/tois_context.pdf.
26. *Miller G.A., Charles W.G.* Contextual Correlates of Semantic Similarity // Language and Cognitive processes. — 1991. — N 6(1).
27. *Крижановский А.А.* Оценка результатов поиска семантически близких слов в Википедии / Тр. СПИИРАН. — СПб., 2007. — Вып. 5. — С. 113—116.

Статья представлена к публикации членом редколлегии В.Г. Лебедевым.

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

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) 334-92-39, ✉ pron@ipu.ru,

Суховеров Виктор Степанович — канд. техн. наук,
ст. науч. сотрудник, ☎(495) 334-76-39, ✉ suhoverv@ipu.ru,

Шипилина Любовь Борисовна — канд. техн. наук,
ст. науч. сотрудник, ☎(495) 334-76-39, ✉ lubship@ipu.ru,

Институт проблем управления им. В.А. Трапезникова РАН.