Семантические технологии для семантических приложений. Часть 2. Модели сравнительной семантики текстов*

В. И. Городецкий[I], О. Н. Тушканова[II]

[I] TRA Robotics Ltd., г. Санкт-Петербург, Россия
[II] Санкт-Петербургский институт информатики и автоматизации Российской академии наук, г. Санкт-Петербург, Россия

Аннотация. В статье обсуждаются базовые аспекты современного понимания семантических вычислений, семантических технологий и приложений в области обработки больших данных, представленных текстами на естественном языке, выполняемой в интересах извлечения знаний для принятия решений. Рассмотрены базовые компоненты семантических технологий, к которым относятся онтологии и модели их использования, семантические ресурсы, которые содержат знания о семантике слов естественного языка и средства ее уточнения, а также семантическая компонента технологии, которая используется для формального описания смысла сущностей естественного языка и численной оценки их попарной семантической близости. Основное внимание уделяется моделям последней компоненты технологии, которые важны для решения задач семантической кластеризации и классификации текстов и различных их приложений. Обсуждаются и сравниваются различные типы мер семантической близости сущностей естественного языка в контексте задач семантических вычислений и анализируются проблемы, которые сдерживают практическое использование семантических технологий.

Ключевые слова: семантические технологии, семантические вычисления, семантический ресурс, семантическая связанность, семантическая близость.

DOI 10.14357/20718594190105

Введение

Одной из наиболее заметных тенденций современного искусственного интеллекта (ИИ) является значительное повышение практического интереса исследователей и разработчиков приложений к моделям и методам работы с данными знаниями в терминах естественного языка (ЕЯ). В первую очередь, этот интерес обусловлен потребностями со стороны технологий обработки больших данных, представленных на ЕЯ, удельный вес которых в общем объеме различных типов данных составляет порядка 90%. С другой стороны, к настоящему времени накоплен достаточный научный потенциал теоретических знаний и алгоритмов формализации семантики ЕЯ-текстов. Этот потенциал включает в себя, прежде всего, достижения в области получения, представления и использования знаний на основе инжиниринга онтологий, накопленные семантические ресурсы, которые связывают сущности ЕЯ (слова, понятия, паттерны текста и т.п.) и их смысл. В частности, это веб-ресурсы, лавинообразный рост объема и разнообразия которых превратил

[1]Работа выполнена при поддержке программы президиума РАН №26 "Фундаментальные основы алгоритмов и программного обеспечения для перспективных сверхвысокопроизводительных вычислений".

✉ *Тушканова Ольга Николаевна. E-mail: tushkanova.on@gmail.com*

их практически в универсальный источник знаний о смысле сущностей ЕЯ, а также методы и алгоритмы установления и формального описания смысла сущностей ЕЯ и численной оценки их попарной семантической близости.

В Части 1 было показано на примерах, что семантические технологии уже достигли такого уровня зрелости, который позволяет использовать их для создания индустриальных приложений. Описаны доступные семантические ресурсы и предложенные варианты семантических примитивов текстов (СПТ), описывать смысл текстов в векторных пространствах признаков.

Основное внимание в Части 2 уделяется моделям и мерам численной оценки попарной семантической близости сущностей ЕЯ. Эти модели играют первостепенную роль в большинстве задач семантической обработки больших данных, представленных выборками текстов на ЕЯ, в частности, в задачах кластеризации текстов, машинного обучения алгоритмов и др.

В Разделе 1 обсуждаются и сравниваются различные типы мер семантической близости сущностей ЕЯ в контексте семантических вычислений, которые ранее традиционно развивались области компьютерной лингвистики. В Разделе 2 рассматриваются возможности веб-ресурсов и поисковых машин по выявлению и формальному описанию семантики ЕЯ-сущностей и мер их семантической близости. В Разделе 3 анализируются проблемы, сдерживающие практическое использование семантических технологий для разработки семантических приложений.

1. Семантическая связанность и семантическая близость сущностей естественного языка

Как уже упоминалось, в семантических приложениях наиболее часто решаются задачи типа кластеризации множества текстов и смысловой классификации новых текстов при заданном множестве классов. Во всех этих задачах алгоритмы обучения строятся, главным образом, на различных свойствах, которые в той или иной форме характеризуют связи между признаками объектов. В задачах семантических вычислений роль признаков играют СПТ, а их семантические связи описываются с помощью численной меры, называемой семантической связанностью.

Под *семантической связанностью* пары СПТ авторы работы [1] понимают некоторую меру, которая численно оценивает их семантическое сходство (смысловую близость) на основании *анализа множества отношений*, заданных для этой пары. Функция, аргументами которой является пара СПТ и множество отношений на них, а результатом является числовая оценка семантической связанности или семантической близости этой пары, называется мерой их *семантической близости* (*сходства*). Обычно эта функция выбирается так, чтобы ее значения лежали в интервале [*–1, 1*] или [*0, 1*], где значение *1* означает смысловую эквивалентность ЕЯ-сущностей.

Заметим, что понятие семантической связанности является более общим, чем понятие семантического сходства. Очевидно, что чем полнее и точнее представлены отношения на паре сущностей ЕЯ, тем надежнее может быть построена оценка их семантического сходства. Это еще раз подчеркивает важность качества семантических ресурсов, используемых в семантических технологиях. От вида функции, задающей меру семантического сходства, зависит то, насколько полно используются потенциальные возможности семантического ресурса.

Одна из принятых классификаций, а также достаточно полное описание соответствующих мер дано в работе [2]. В ней выделены четыре типа мер:

1) топологические, измеряющие длину пути между понятиями в онтологии (англ. *path based*);

2) использующие информационное содержании понятий (англ. *information content-based*);

3) гибридные меры, объединяющие идеи мер (1) и (2);

4) меры на основе признаков (англ. *feature-based*).

Однако эта классификация является неполной. Существуют и другие идеи для построения мер семантической близости. Например, они могут быть основаны на наличии общих слов в сравниваемых текстах, могут использовать обогащение контекста сравниваемых сущностей ЕЯ за счет веб-ресурсов, обнаруженных той или иной поисковой машиной. Важным классификационным признаком является тип используемого семантического ресурса, что априори характеризует потенциальную точность оценки семантической близости пары ЕЯ-сущностей.

Хотя предложено много различных мер сходства СПТ, все они построены на основе небольшого количества идей и различаются деталями реализации и сложностью алгоритмов их вычисления. Далее рассматриваются только некоторые представители их основных групп.

Топологические меры близости (англ. *topological similarity*) измеряют длину пути ("семантическое расстояние") между понятиями в общей таксономии понятий онтологии. Для их вычисления нужно иметь иерархию понятий онтологии, в которой представлены оба сравниваемые понятия. Меры этого типа могут работать со всеми разновидностями семантических ресурсов ЕЯ. Примерами являются мера Лекокка *lch* [3], Ву и Палмера *wup* [4] и Ли *li* [1, 5]. Для иллюстрации приведена формула для меры *wup*:

$$sim_{wup}(c_1.c_2) = \frac{2 \times deep(lso(c_1, c_2))}{len((c_1, c_2) + 2 \times deep(lso(c_1, c_2))}, \quad (1)$$

где $len(c_1, c_2)$ – расстояние между понятиями c_1 и c_2 в онтологии (число соединяющих их ребер); $2 \times deep_{max}$ – удвоенное значение максимальной глубины отношения; $lso(c_1, c_2)$ – глубина онтологии от корня до минимального общего элемента пары понятий c_1 и c_2 в онтологии.

Аналогично строится и мера *li*, правило вычисления которой можно найти в [1, 5].

Меры, использующие оценку информационного содержания сравниваемых понятий (англ. *information content, IC*) [6–8]. Для вычисления мер этой группы, как правило, необходима выборка документов, содержащих сравниваемые понятия, т.к. эти меры используют частоту встречаемости обоих понятий в выборке. К этой группе относят меры Резника *res* [6], Лина *lin* [7], расстояние Джанга *jang* [8] и другие. Поясним суть таких мер на примере меры *res*.

Информационное содержание $IC(c)$ некоторого понятия c онтологии на выборке документов D, для которого оценка вероятности в этой выборке равна $p(c)$, вычисляется по формуле:

$$IC(c) = -log(p(c)). \quad (2)$$

Для подсчета оценки эмпирической вероятности понятия c используется набор слов текста, который включает все слова, соответствующие понятиям-потомкам понятия c.

Мера Резника *res* для двух понятий вычисляется как значение информационного содержания, сосчитанное для их минимального общего родительского понятия c в иерархической структуре онтологии (англ. *Most Common Specific Abstraction*, *MCSA*):

$$sim_{res}(c_1, c_2) = min_{c \in S(c_1, c_2)} IC(c), \quad (3)$$

где $S(c_1, c_2)$ – множество понятий, которые являются родительскими для понятий c_1 и c_2.

Мера Лина *lin* и расстояние Джанга *jang* являются модификациями меры *res* [9]. Важно отметить, что все эти меры используют статистические и семантические свойства сравниваемых понятий.

Гибридные методы объединяют идеи описанных выше двух вариантов мер, а именно оценивают и длину пути между понятиями, и их информационное содержание. Примером может служить мера *wpath* (от англ. *weighted path length* – взвешенная длина пути) [10]:

$$sim_{wpath}(c_1, c_2) = \frac{1}{1 + lenght\ (c_1, c_2) \cdot k^{IC(c)}}, \quad (4)$$

где $k \in (0, 1]$.

Меры семантической близости, которые строятся с помощью признаков. Одна из таких мер, известная под названием ESA (от англ. *Explicit Semantic Analysis*), описана в [11]. Эта мера была первой, в которой в качестве ресурса для установления смысла терминов и последующей попарной оценки их близости использовалась Википедия. Для вычисления значения этой меры близости пары СПТ используется *векторное представление* смысла примитива или текста в целом с последующей оценкой близости векторов, например, с помощью косинусной меры сходства.

Алгоритм вычисления меры ESA состоит в следующем. Сначала каждому слову сравниваемой пары ставится в соответствие вектор понятий Википедии, примером которых это слово является. Для каждого такого понятия выбирается текст соответствующей словарной статьи Википедии, а множество этих текстов рассматривается далее в качестве обучающего множества текстов, для каждого из которых вычисляется значение меры *TF-IDF* слов сравниваемой пары. Последующий алгоритм, который называется семантическим интерпретатором, просматривает итеративно все слова текста, анализирует входы со стороны инвертированного индекса и вычисляет для каждого найденного понятия величину меры *TF-IDF*. В итоге каж-

дому слову текста ставится в соответствие взвешенный вектор понятий. Эти величины рассматриваются далее в качестве весов соответствующих координат сравниваемых слов в пространстве понятий Википедии. Семантическая близость пары слов оценивается по той или иной мере близости в векторном пространстве, например, с помощью традиционной косинусной меры или другой меры сходства. Мера ESA далее обобщается на оценку близости текстов, которая также строится в векторном пространстве слов, извлеченных из текстов.

Заметим, что в алгоритмах семантической кластеризации и классификации различие в признаках, описывающих сущности ЕЯ, является не менее важным, чем их сходство. По этой причине полезными являются меры, которые учитывают оба эти фактора. Существуют различные варианты таких мер, например, хорошо известная мера Жаккара. Для оценки семантической близости сущностей ЕЯ и текстов чаще используется мера Тверски [12]. Эта мера игнорирует взаимное положение сравниваемых понятий в онтологии и их информационное содержание, а также не использует взвешивание признаков. Этим она представляется удобной в случаях, когда некоторому тексту ставится в соответствие, например, множество понятий онтологии, представленных в нем. В нормализованном варианте эта мера имеет вид:

$$Sim_{tvsk}(c_1, c_2) = \\ = \frac{|P(c_1) \cap P(c_2)|}{|P(c_1) \cap P(c_2)| + \propto |P(c_1) \setminus P(c_2)| + (1-\propto)|P(c_2) \setminus P(c_1)|}, \quad (5)$$

где символ | * | обозначает мощность множества, указанного в вертикальных скобках, и $\propto \in [0, 1]$. В этой формуле общие признаки пары объектов ЕЯ или текстов повышают их семантическую близость, в то время как различные признаки эту близость уменьшают.

Далее кратко описываются меры, выпадающие из классификации, данной в работе [2].

Мера, использующая сравнение перекрытия текстов, изначально была предложена для решения WSD-проблемы для слов текста [13] путем привлечения различных машиночитаемых словарей. Однако позднее, когда был создан ресурс WordNet, идея работы [13] была использована для создания полноценной меры семантической близости, которая называется в настоящее время по имени ее автора мерой Леска (англ. *lesk*). Алгоритм вычисления этой меры состоит в том, что сначала для каждого слова сравниваемой пары решается WSD-проблема с использованием коротких текстов (англ. *glosses*), которые описывают их различные смыслы в WordNet. Для каждого из этих текстов подсчитывается число общих слов в пересечении с текущим контекстом слова (некоторым числом его ближайших слов-соседей в тексте), и в качестве решения выбирается тот смысл, которому отвечает наибольшее число слов в его пересечении с контекстом. По существу, в этом и состоит вся идея первого шага, на котором решается WSD-проблема. Далее такая же задача решается для пары сравниваемых слов с найденным смыслом путем пересечения соответствующих текстов, описывающих их в WordNet. Мера близости *lesk* пары слов в заданном контексте полагается тем большей, чем больше число общих слов в этом пересечении.

Но в такой реализации алгоритма вычисления меры *lesk* она оказывается очень чувствительной к замене слов на синонимы или слова, близкие по смыслу. Эта замена может радикально изменить оценки близости, поэтому предложено много модификаций алгоритма вычисления меры *lesk*, повышающих его устойчивость. Описание и сравнение модификаций алгоритмов, реализующих базовую идею меры *lesk*, можно найти, например, в работе [14]. Они, главным образом, касаются расширения множества сравниваемых текстов, поскольку тексты WordNet являются достаточно короткими, а потому может так случиться, что даже для слов, близких по смыслу, их описания в WordNet не будут иметь общих слов. Эти расширения могут быть получены разными способами, например, за счет привлечения текстов WordNet, относящихся к понятиям, которые связаны с анализируемым понятием. При этом ресурсы веб могут оказаться полезными для улучшения алгоритмизации меры *lesk*. Эти возможности будут освещены ниже.

Модели лексических цепей и меры семантической близости. Напомним, что лексической цепью называют последовательность связанных слов, выбранных из текста в порядке их появления в нем, и бинарных отношений, заданных на них. Мотивация их использования для представления семантики текста основана на том, что близко расположенные слова текста (вместе их называют контекстом) связаны общей темой, и

потому они могут быть хорошей подсказкой о смысле текста в целом. По этой причине задачу построения лексических цепей можно рассматривать как проблему обнаружения семантически значимых агрегированных СПТ.

Лексические цепи, веденные еще в 1991 г. в работе [15], использовались многими исследователями семантики ЕЯ-текстов, например, для задачи сегментации текстов, для автоматической генерации ссылок в гипертексте и др.

Опишем один из примеров построения и использования лексических цепей для представления семантики текстов и построения меры их семантической близости, предложенный в относительно недавней работе [16]. Этот пример достаточно полно поясняет суть модели лексических цепей и их потенциальные возможности. В этой работе на множестве лексических цепей строится *семантическое ядро текста*, которое представляет собой множество взвешенных лексических цепей с однозначно определенной семантикой. Авторы полагают, что каждая лексическая цепь является смысловой компонентой текста, представляющей одну из его тем. С другой стороны, множество взвешенных лексических цепей используют как множество агрегированных семантических признаков текста. Значения весов, поставленных в соответствие лексическим цепям, используются как меры их информативности в прикладных задачах машинного обучения, а также как управляющие параметры, позволяющие пользователю изменять общее число лексических цепей в семантическом ядре путем вариации значения пороговой функции их выбора.

Авторы [16] утверждают, что их подход позволяет описать семантику множества текстов значительно меньшим числом семантически однозначных агрегатов, чем это удается при использовании понятий онтологии, но построение семантического ядра является вычислительно трудоемкой процедурой. Действительно, алгоритму построения семантического ядра предшествует относительно трудоемкая WSD-процедура [16], подробности которой описаны, например, в [17]. Из лексической базы WordNet авторы работы [16] используют только понятия онтологии типа существительных совместно с модифицированной структурной мерой близости Ву-Палмера [4]. В дополнение к этой семантической мере смысл понятий уточняется путем анализа структурных связей в тексте и их сравнения со связями, свойственными понятиям различных синсетов одного и того же слова, т.е. омонимов словаря WordNet, причем для каждого понятия используется не более трех его синсетов (толкований, разных по смыслу).

Далее для понятий с установленным смыслом лексические цепи строятся так:

1. Из текста извлекаются лексические цепи, которые строятся с учетом четырех отношений семантической связанности между парами понятий, последовательно встречающимися в тексте, а именно отношений *совпадения* (англ. *identity*), *синонимии* (англ. *synonymy*), *более общее* (англ. *hyperonymy*) или *более частное* (англ. *hyponymy*) и *часть-целое* (англ. *meronymy*).

2. Далее текст документа представляется графом, узлы которого ставятся в соответствие понятиям, которые встречаются в построенных лексических цепях, а ребра – отношениям на множестве этих понятий, которыми они связны в лексических цепях. Для этого графа в два этапа вычисляются веса вершин и ребер. Сначала для них вычисляются “веса в графе” как сумма частот встречаемости в документе каждого из примеров рассматриваемого понятия. Для каждого понятия вычисляется его “вес в лексической цепи“, который определяется как сумма его “веса в графе” и суммы “весов в графе” инцидентных ему вершин, умноженных на веса ребер. Вес каждого ребра зависит от типа отношения, которое оно задает. Напомним, что среди отношений в графе могут присутствовать отношения *совпадения*, *синонимии*, *более общее – более частное* и *часть-целое*, каждому из которых пользователь присваивает вес по убыванию их значений в соответствии с указанным порядком их следования.

3. Веса понятий лексической цепи складываются, и результат рассматривается как вес лексической цепи в целом. Если этот вес превышает порог, определенный пользователем, то эта цепь включается в ядро семантических признаков текста.

Формальные детали описанных вычислений могут быть найдены в работе [16].

2. Меры близости, использующие обогащение контекста и веб-ресурсы

В последнее десятилетие активно развиваются методы описания семантики сущностей ЕЯ, а также меры их семантической близости, которые для повышения точности выявления,

представления и использования семантики ЕЯ-сущностей используют *расширение контекста* за счет привлечения дополнительных источников и, прежде всего, привлечения веб-ресурсов.

Хотя подобные идеи были высказаны еще в первых работах по семантическому поиску информации [18], современная их алгоритмизация в задачах установления семантики текстов и их семантической близости была предложена в [19]. Заметим, что эти достаточно очевидные идеи оказались весьма продуктивными, причем особенно в приложениях, в которых сравниваемые тексты слишком коротки, чтобы можно было однозначно установить их семантику, решая WSD-проблему известными методами, и, соответственно, оценивать их семантическую близость с другими, тоже короткими текстами. Рассмотрим один из таких подходов [19].

Меры, использующие расширение контекста за счет ресурсов веб. Основная идея подхода к установлению семантики короткого текста и его сходства с аналогичными текстами состоит в следующем. Использование короткого текста в качестве запроса поисковой машины веб позволяет извлечь много веб-документов. Поскольку результат веб-поиска – это множество документов, по смыслу близких к запросу, то их рассматривают как расширение контекста запроса для установления смысла его слов. Иначе, возвращаемые документы используются для того, чтобы построить вектор контекста для слов запроса, которые часто встречаются в документах вместе со словами запроса. Так устанавливается семантика запроса. И уже к этому вектору контекста можно применять косинусную меру близости пары текстов, что дает более робастный результат.

В конкретной реализации, описываемой в [19], возвращаемые документы упорядочиваются по релевантности (по рангу документа, выдаваемому поисковой машиной), и общее число наиболее релевантных документов ограничивается некоторой величиной *n*, например, *n*=200. Далее смысл короткого текста отождествляется со смыслом всех отобранных *n* документов, и последующая задача состоит в том, чтобы этот смысл извлечь и представить в компактной форме. В качестве такой формы представления смысла отобранных документов в [19] используется нормированное выборочное среднее значений *TF-IDF*-меры, сосчитанное для множества наиболее весомых терминов, извлеченных из *n* документов. Эта форма описания семантики запроса представляется вектором *QE(x)* (от англ. *Query Expansion*), который формально находится так:

1. Текст x используется в качестве запроса к веб, выполняемого поисковой машиной S. Пусть $R(x) = \{d_1, \dots, d_n\}$ – множество документов, возвращаемых поисковой машиной по запросу x.

2. Из каждого документа множества $R(x) = \{d_1, \dots, d_n\}$ извлекаются слова, для которых вычисляются значения меры *TF-IDF*. Далее для каждого документа d_i строится "контекстный" вектор значений V_i, компонентами которого являются отобранные слова документа со значениями меры *TF-IDF*. Общее число таких слов для каждого документа – не более m.

3. Строится вектор $C(x)$ – центроид всех нормализованных векторов V_i:

$$C(x) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{V_i}{\sqrt{\sum_{j=1}^{m}\left(\bar{V}_{ij}^2\right)}}. \qquad (6)$$

4. Далее строится вектор $QE(x)$ как нормализованный вектор $C(x)$ размерности m:

$$QE(x) = \frac{C(x)}{\sqrt{\sum_{i=1}^{m}(c_i)^2}}. \qquad (7)$$

Заметим, что нормализованный вектор $QE(x)$ зависит от использованной поисковой машины S, однако для упрощения записи в векторе $QE(x)$ эта зависимость не отражена. В принципе, возможно также использование другой весовой функции вместо *TF-IDF*. Авторы брали $m = 50$, а общее число документов *n* выбиралось равным 200, 500 и 1000 в разных экспериментах.

Аналогичным образом строится расширение $QE(y)$ запроса y, которое интерпретируется как компактное описание его смысла. Тогда *ядерная мера семантической близости пары запросов* (коротких текстов) x и y строится как скалярное произведение векторов $QE(x)$ и $QE(y)$:

$$K(x,y) = \big(QE(x), QE(y)\big) = \sum_{i=1}^{m} QE_i(x)\, QE_i(y). \qquad (8)$$

Построенное ядро $K(x,y)$ авторы [19] называют ядром семантического сходства, *kbss* (от англ. *Kernel-based Semantic Similarity*). Далее полагается, что каждая компонента $QE(x)$ представляет некоторую тему, отраженную в тексте, и множество этих тем в нем упо-

рядочено по их важности. Аналогично представлено и компонентами расширения $QE(\mathrm{y})$.

В качестве возможного способа снижения объема вычислений расширения $QE(x)$ авторы предлагают вместо документов использовать только короткий текст, который поисковая машина выдает вместе с каждым найденным документом, длиной не более 1000 слов.

Отметим ряд особенностей этой модели семантики короткого текста. Во-первых, модель описывает семантику текста в целом, а не его отдельных слов. Во-вторых, экспериментально установлено, что для семантически близких текстов величина меры близости всегда более 0.5, а для различных по смыслу она всегда имеет значение менее 0.3. В-третьих, при увеличении числа учитываемых документов робастность оценки семантики короткого текста и значения его семантической близости с другими тестами увеличивается. Важно также отметить, что даже в том случае, когда оба запроса не имеют вообще общих слов, но смысл их близок, итоговое значение меры близости никогда не становится равным нулю, как это бывает в других мерах.

Google-семантика и семантическая близость. Мотивация авторов данного подхода к извлечению семантики текстов и к оценке их семантической близости достаточно проста [20]. Неструктурированное множество веб-документов есть практически универсальная выборка текстов, а потому она может быть использована для выявления их семантики и семантической близости наравне с другими множествами документов, на которые опираются другие варианты решения таких задач. По мнению авторов подхода, слова, близкие по смыслу, должны приводить к похожим результатам поиска в веб. Поэтому уже даже число общих документов, которые найдены поисковой машиной, например, Google-машиной, для двух слов, может использоваться в качестве аргумента функции, задающей меру близости, поскольку, чем больше относительное число таких общих страниц для пары слов или текстов, тем они должны быть ближе по смыслу.

Заметим, что с первого взгляда может показаться, что эта мотивация возвращает проблему к статистическим моделям оценки семантики, однако это не вполне так. И причина этого состоит в том, что множество веб-документов, индексированных для мощной поисковой машины типа Google, не является случайной выборкой из некоторой генеральной совокупности. В отличие от этого, она представляет собой распределенное хранилище знаний человечества, которое по своей сути не имеет статистического характера, а использование статистических моделей формализации семантики следует рассматривать только как вариант аппроксимации семантики веб-данных с помощью статистических моделей. Однако этот подход не всегда способен различать смысл омонимов, хотя в некоторых задачах соответствующие ошибки могут быть малыми.

В варианте, предложенном в [20], метод использует поисковую машину Google, хотя это справедливо и для других поисковых машин веб. Особенность и специфика подхода состоит в том, что в нем СПТ не используются, как это требуется в других мерах и алгоритмах их вычисления, рассмотренных ранее, которые опираются на СПТ, их контекст и семантические ресурсы. В данном подходе исходной информацией является только сам объект (точнее, его “имя” на ЕЯ), а информация о нем отыскивается в огромной базе веб-ресурсов. Результатом поиска является оценка частоты слова в веб-ресурсах и оценка семантической близости пары текстовых объектов. Соответствующая семантика называется *Google-семантикой*, если в качестве поисковой машины используется Google-машина.

Такая мера семантической близости имеет формальное обоснование. Оно опирается на понятие Колмогоровской сложности (КС) цепочки символов и ее свойства [21]. Абстрактно она определяется как длина в битах самой короткой программы, которая могла бы ее сгенерировать. Авторы [20] вводят понятие *информационного расстояния* $E(x,y)$ между произвольными цепочками x и y, которое определяется как КС-оценка бинарной программы, для которой x является входной цепочкой, а y – выходной и наоборот, для входной цепочки y эта программа выдает цепочку x. Авторы строго показывают, что

$$E(x,y) = K(x,y) - min\{K(x), K(y)\}, \qquad (9)$$

где $K(x,y)$ – КС цепочки xy (конкатенации цепочек x и y); $K(x)$, $K(y)$ – КС цепочек x и y, соответственно.

Оказывается также, что величина $E(x,y)$ обладает свойствами метрики (неотрицательна,

симметрична относительно аргументов и удовлетворяет неравенству треугольника). Поэтому $E(x,y)$ формально может использоваться в качестве метрики для оценки семантического расстояния между парой цепочек. Авторы показывают, что ее значение минимально на множестве других вариантов определения расстояний для пары цепочек x и y.

Но, к сожалению, величина $E(x,y)$ непригодна для оценки сходства или различия строк, поскольку на ее значение влияет также и длина строк. По этой причине авторы вводят нормализованное расстояние *NID* (от англ. *Normalized Information Dista*nce), которое уже оказывается инвариантным к длине строк и изменяется в диапазоне $[0, 1]$:

$$NID(x,y) = \frac{K(x,r) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}. \qquad (10)$$

Однако эта величина является невычислимой, поэтому предлагается использовать ее аппроксимацию. В формуле (10) можно использовать частотные оценки величин ассоциативных связей, которые могут быть вычислены с помощью поискового механизма Google на множестве всех индексированных веб-страниц. Их обоснование формулируется таким образом: "*Веб содержит настолько большой и разнообразный корпус документов, а Google является настолько мощным поисковым механизмом, что относительное число возвращаемых страниц аппроксимирует истинное общественное использование слов и фраз*" [20, с. 4 – 5]. Другими словами, вывод работы состоит в том, что оценки ассоциаций возвращаемых страниц усредняют семантическую информацию, представленную в веб, и потому они могут использоваться для автоматической генерации мер семантической близости слов и фраз запроса.

Кроме полисемии, причинами ошибок в оценке семантической близости могут быть неточности в нахождении числа возвращаемых страниц и динамика числа индексированных страниц веб.

Итоговое выражение для метрики, которую авторы называют *нормализованным Google-расстоянием* (англ. *normalized Google distance, NGD*) имеет вид:

$$NGD(x,y) = \frac{G(x,y) - \min\{G(x),G(y)\}}{\max\{G(x),G(y)\}} = \\ = \frac{\max\{\log(f(x)logf(y)\} - logf(x,y)}{LogN - \min\{logf(x), logf(y)\}}. \qquad (11)$$

где $f(x)$ и $f(y)$ – число страниц, возвращаемых Google по запросам x и y, которые содержат x и y, соответственно; $f(x,y)$ – число страниц, возвращаемых по запросу xy, которые содержат одновременно обе цепочки.

Метрика NGD аппроксимирует сверху нормализованное информационное расстояние, введенное ими с помощью понятия КС.

Позднее были предложены аналогичные подходы с использованием Google-семантики совместно с другими веб-ресурсами. Некоторые из таких вариантов кратко описываются ниже.

Авторы работы [22] отмечают, что семантическое сходство, основанное на *NGD*-мере, которая базируется на оценках априорных вероятностей совместного появления пары слов во всем множестве веб-документов, не зависит от их семантики в конкретном контексте. Для уточнения смысла сравниваемых слов авторы предлагают дополнительно использовать реальный контекст из документов. Самый естественный вариант обогащения Google-семантики – это привлечь семантические ресурсы типа WordNet, Вики-словарей и Википедии для интерпретации контекста, который сопровождает поисковые запросы в возвращаемых веб-страницах. Эту роль могут играть лексико-синтаксические паттерны из коротких текстов, возвращаемых поисковой машиной Google по запросу для сравниваемой пары слов.

В работе [22] семантическая близость пары слов вычисляется как оценка вероятности случайного события "два слова запроса являются синонимами" с помощью бинарного SVM-классификатора, обученного на ограниченной выборке, которая состоит из пар слов-синонимов и «не-синонимов», формируемых вручную на основании синсетов WordNet. В результате из коротких текстов, содержащих оба слова, извлекаются типовые лексико-синтаксические паттерны, состоящие из слов и символов пунктуации и скобок, на основе которых формируется множество признаков. Авторы экспериментально показывают, что описанная мера близости, оказывается гораздо ближе к оценкам человека по сравнению с вариантами, не использующими контекста.

Однако описанный подход вычислительно сложен и предполагает предварительное формирование обучающей выборки и SVM-обучение. Поэтому его следует рассматривать не как вычислительный алгоритм поиска се-

мантической близости сущностей ЕЯ, а как один из аргументов в пользу интеграции Google-семантики и других доступных семантических ресурсов ЕЯ.

Еще один алгоритм оценки семантической близости СПТ, в котором используется Google-семантика в комбинации с привлечением семантических ресурсов, предложен в работе [23]. Подход рассматривается как составная часть одного из важнейших семантических приложений, а именно задачи инжиниринга онтологии на основании текстов, относящихся к приложению. Этот алгоритм ориентирован на решение задачи семантической кластеризации терминов текстов, но в нем явно выделяется оригинальный подход к оценке семантической близости терминов.

Алгоритм реализуется в два прохода. На первом для СПТ, извлеченных из текстов, строится иерархическое дерево кластеризации по *NGD*-мере. При этом число и состав кластеров, формирующих понятия онтологии, определяется алгоритмом автоматически. Для описания первого прохода авторы используют метафору муравьиного алгоритмов, в котором муравьи растаскивают СПТ по разным кластерам и метят пути перемещения СПТ феромоном, чтобы можно было при необходимости для каждого из них восстановить путь обратного перемещения. Элементы матрицы парных *NGD*-мер близости $S(t_x, t_y)$ между СПТ вычисляются по формуле:

$$S(t_x, t_y) = 1 - NGD(t_x, t_y)\alpha, \qquad (12)$$

где α – положительная масштабирующая константа.

Однако мера (12), которая не позволяет учитывать синонимию и не способна различать омонимы, недостаточна точна. Но ошибки такого рода, характерные для *NGD*-меры близости СПТ, недопустимы при инжиниринге онтологий. Как следствие таких ошибок *NGD*-меры может оказаться, что некоторые кластеры содержат или единичные экземпляры СПТ, или имеют слишком малую мощность. Присутствие таких элементов в построенном бинарном дереве рассматривается авторами как ошибки кластеризации или выбросы (англ. *outlie*r), и для отобранного множества СПТ значение близости уточняется на втором проходе с помощью n^oW-меры, которая измеряет кратчайший взвешенный путь между понятиями (статьями) в структуре категорий Википедии. На этом проходе для каждого изолированного кластера выполняется движение обратно по пути, помеченному феромоном, с оценкой вероятности оставить СПТ в текущем узле или переместить его в другой. Решение принимается по среднему значению меры близости изолированного СПТ по отношению к другим СПТ кластера.

Этот метод оценки семантической близости СПТ в процессах семантической кластеризации обладает следующими особенностями: использует семантический ресурс Википедии для уточнения *NGD*-меры близости; хорошо приспособлен для параллельной реализации с использованием многоагентной технологии; специально рассчитан на автоматизацию процессов построения онтологии; позволяет автоматически определять число кластеров, что немаловажно в технологии семантической кластеризации.

3. Обсуждение компонент семантических технологий

Алгоритмические основы семантических технологий, которые активно развиваются в последние два десятилетия, к настоящему времени достигли уровня зрелости, необходимого для их практического использования в широком ряде приложений. Некоторые убедительные примеры уже разработанных семантических приложений описаны в книге [24]. В ближайшее время следует ожидать еще большей активизации исследований в этой области.

В фокусе настоящей работы находятся методы, алгоритмы и семантические ресурсы, составляющие методическую и алгоритмическую основу и информационный базис для выявления и формального представления семантики ЕЯ-текстов и вычисления семантической близости различных его сущностей. В работе эта компонента названа семантической компонентой технологии. Ее специфика состоит в том, что именно она выделяет класс семантических технологий и вычислений из множества всех других интеллектуальных информационных технологий.

Информационная компонента семантического приложения включает в себя, прежде всего, онтологию как базу знаний и данных, которые должны быть доступны через соответствующие интерфейсы, как пользователю, так и

программе. При необходимости эта компонента может обогащаться разнообразными веб-ресурсами и знаниями, извлекаемыми из больших данных.

Методы и алгоритмы создания семантической компоненты имеют длительную историю исследований и разработок, и к настоящему времени они достигли того уровня зрелости, который необходим для их практического использования в семантических приложениях.

Дальнейшее повышение уровня адекватности семантической компоненты следует искать, прежде всего, в создании более богатых и лучше структурированных семантических ресурсов, поскольку именно они определяют потенциальные возможности по полноте и точности выявления и представления семантики сущностей ЕЯ и их семантической близости. Сейчас бесспорным лидером в области таких ресурсов являются разнообразные веб-ресурсы, в частности Вики-словари, а также другие богатые и хорошо структурированные ресурсы Linked Data. Достоинство этих ресурсов состоит также в том, что они постоянно развиваются и обогащаются.

Каждый алгоритм выявления семантики ЕЯ, каждая мера семантической близости является неточной, и не может быть точной, поскольку само понятие смысла далеко от точного понимания. Но каждая мера схватывает какой-то специфический аспект семантики ЕЯ и семантической близости. Поэтому при создании семантических приложений разумно привлекать, где это возможно, одновременно сразу несколько вариантов выявления и представления семантики и мер семантической близости, и в окончательном решении использовать методы объединения решений, разработанные в машинном обучении [25].

Важно заметить, что статистические меры только аппроксимируют семантику СПТ. Эта аппроксимация будет более точной при использовании большего объема данных. Поэтому не случайно, что все большей популярностью пользуются методы и алгоритмы, которые интегрируют веб-ресурсы для обогащения выборки текстов.

Сдерживающим фактором развития моделей семантики сущностей ЕЯ является отсутствие необходимого разнообразия тестовых множеств с известной семантикой (англ. *benchmarks*). Анализ десятков работ, посвященных выявлению и представлению смысла СПТ, показывает, что для тестирования предлагаемых методов и алгоритмов авторы, как правило, не используют сырые данные известных тестовых множеств, а обычно “вычищают” их вручную, чтобы избежать проблем, которые могут существенно ухудшить результаты их экспериментов. Поэтому многие “хорошо зарекомендовавшие” себя методы не обладают нужной устойчивостью, что снижает их ценность. Это препятствует также объективному сравнению решений.

Заключение

Развитие семантических технологий и их использование для повышения интеллектуальности современных программных приложений и систем в настоящее время рассматриваются как важные и перспективные направления исследований и разработок в области ИИ. В работе обсуждаются различные аспекты современного понимания, а также состояния исследований и разработок в области семантических вычислений, семантических технологий и семантических приложений, и возможности их практического использования уже в настоящее время.

Зрелость любого семантического приложения определяется качеством реализации трех его базовых компонент, а именно зрелостью онтологии как интегратора знаний и данных, и ее доступностью для понимания как человеком, так и программой, качеством и адекватностью представления семантической компоненты приложения, а также возможностями используемых информационных ресурсов типа веб-ресурсов и больших данных как источников знаний.

В работе показано, что потенциальные возможности выявления семантики ЕЯ-сущностей и оценки их семантической близости, а также адекватного формального представления семантики для последующего практического применения, зависят от доступных семантических ресурсов. Поэтому отмечается важность дальнейших работ по расширению тезауруса и обогащения структурной компоненты Википедии - лидера среди других семантических ресурсов, как по полноте понятийной структуры, так и по богатству множества представленных в ней отношений.

В работе не анализируется проблема практического эффективного использования онтологий совместно с семантической компонентой

приложения. Эта задача относится в настоящее время к числу наиболее серьезных вызовов в области семантических технологий. Однако она представляет собой самостоятельную важную проблему, и потому является одной из целей будущих исследований и разработок в области семантических технологий и семантических вычислений.

Литература

1. Meng L., Huang R., Gu J. A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6 (1), 2013, pp. 1-12.
2. Feng Y., Bagheri E., Ensan F., Jovanovic J. The state of the art in semantic relatedness: a framework for comparison. Knowledge Engineering Review, 2017, pp. 1-30.
3. Leacock C., Chodorow M. Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database, 1998, vol. 49, no. 2, pp. 265-283.
4. Wu Z., Palmer M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133-138.
5. Li Y., Bandar Z., Mclean D. An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 2003, vol. 15, no. 4, pp. 871-882.
6. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 448-453.
7. Lin D. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296-304.
8. Jiang J. J., Conrath D.W. Semantic similarity based on corpus statistics and lexical taxonomy. Computational Linguistics, 1997, vol. cmp-lg/970, no. Rocling X, p. 15.
9. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов, Труды ИСП РАН, 2017, том 29, выпуск 2, с. 161-200.
10. Zhu G., Iglesias C.A. Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering 29.1, 2017, pp. 72-85.
11. Gabrilovich E., Markovitch, S. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07),Sangal, R., Mehta, H. & Bagga, R. K. (eds). Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.
12. Tversky A. Features of Similarity. Psycological Review, 1977, 84(4), pp. 327-352.
13. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86), DeBuys, V. (ed.). ACM, 1986, pp. 24–26.
14. Vasilescu F., Langlais P., Lapalme G. Evaluating Variants of the Lesk Approach for Disambiguating Words. Proceedings of The Fourth International Conference on Language Resources and Evaluation (LREC 2004), Portugal, 2004, pp. 633-636.
15. Morris J., G. Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991, vol. 17, 1, pp. 21-43.
16. Wei T., Lu Y., Chang H., Zhou Q., Bao X. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications 2015, 42, pp. 2264-2275.
17. Ткач С.С. Применение лексических цепочек для разрешения лексической многозначности на основе Русского Викисловаря. Магистерская диссертация. Петрозаводский Государственный университет. Петрозаводск, 2016. 60 с.
18. Mitra M., Singhal A., Buckley C. Improving automatic query expansion. In Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 206-214.
19. Sahami M., Heilman T.D. A web-based kernel function for measuring the similarity of short text snippets.In Proceedings of the 15th International Conference on World Wide Web (WWW '06), ACM, 2006, pp. 377-386.
20. Cilibrasi R.L., Vitanyi P. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 2007, 19(3), pp. 370-383.
21. Верещагин Н.К. Успенский В.А., Шень А. Колмогоровская сложность и случайность. М.:Издательство МЦНМО, 2013, 575 с.
22. Bollegala D., Yutaka Matsuo Y., Ishizuka M. WebSim: a Web-based Semantic Similarity Measure. The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007, pp.1-4.
23. Wong W., Liu W., Bennamoun M. Tree-traversing ant algorithm for term clustering based on featureless similarities. Data Mining and Knowledge Discovery, 2007, 15 (3), pp. 349-381.
24. Bartussek W., Bense H., Hoppe T., Humm B.G., Reibold A., Schade U., Siegel M., Walsh P. Introduction to Semantic Applications. In Thomas Hoppe, Bernhard Humm, Anatol Reibold (Eds.). Semantic Applications. Methodology, Technology, Corporate Use. Springer-Verlag GmbH Germany, part of Springer Nature 2018.
25. Городецкий В.И., Серебряков С.В. Методы и алгоритмы коллективного распознавания// Автоматика и телемеханика, 2008, № 11, с. 3-40.

Semantic technologies for semantic applications. Part 2. Models of comparative text semantics

V. I. Gorodetsky[I], O. N. Tushkanova[II]

[I]TRA Robotics Ltd., St. Petersburg, Russia
[II]Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, St. Petersburg, Russia

Abstract. The both parts of the paper discuss the basic aspects of semantic computing, semantic technologies and semantic applications applied to NL-texts big data processing for knowledge extracting and decision-making. The basic components of the corresponding systems and technologies are reviewed, which include ontologies and semantic models of their use, semantic resources, and semantic component. The semantic resources contain knowledges about the words semantics and means for refinement of this semantics. The semantic component of the technology is used to formally describe the meaning of NL-entities and numerically evaluate their pairwise semantic similarity. The main focus of this part is on numerical models of pairwise semantic similarity of NL-entities. These models are important for solving tasks of text semantic clustering and classification and their various applications. Various types of semantic relatedness and semantic similarity measures for NL-entities in the context of semantic computing tasks are discussed and compared. Problems that constrain the practical use of semantic technologies for the development of semantic applications are analyzed.
Keywords: semantic technology, semantic computing, semantic resource, comparative semantics, semantic relatedness, semantic similarity.

DOI 10.14357/20718594190105

References

1. Meng L., Huang R., Gu J. A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6 (1), 2013, pp. 1-12.
2. Feng Y., Bagheri E., Ensan F., Jovanovic J. The state of the art in semantic relatedness: a framework for comparison. Knowledge Engineering Review, 2017, pp. 1-30.
3. Leacock C., Chodorow M. Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database, 1998, vol. 49, no. 2, pp. 265-283.
4. Wu Z., Palmer M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133-138.
5. Li Y., Bandar Z., Mclean D. An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 2003, vol. 15, 4, pp. 871-882.
6. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 448-453.
7. Lin D. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296-304.
8. Jiang J. J., Conrath D.W. Semantic similarity based on corpus statistics and lexical taxonomy. Computational Linguistics, 1997, vol. cmp-lg/970, no. Rocling X, p. 15.
9. Parkhomenko P.A., Grigor'yev A.A., Astrakhantsev N.A. Obzor i eksperimental'noye sravneniye metodov klasterizatsii tekstov [Review and experimental comparison of methods of text clustering]. Trudy ISP RAN [ISP RAS Proceddings]. 2017, 29 (2), pp. 161-200.
10. Zhu G., Iglesias C.A. Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering 29.1, 2017, pp. 72-85.
11. Gabrilovich E., Markovitch, S. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07), Sangal, R., Mehta, H. & Bagga, R. K. (eds). Morgan Kaufmann Publishers Inc., 2007, pp. 1606-1611.
12. Tversky A. Features of Similarity. Psycological Review, 84 (4), 1977. P. 327-352.
13. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86), DeBuys, V. (ed.). ACM, 1986, pp. 24–26.
14. Vasilescu F., Langlais P., Lapalme G. Evaluating Variants of the Lesk Approach for Disambiguating Words. Proceedings of The Fourth International Conference on Language Resources and Evaluation (LREC 2004), Portugal, 2004, pp. 633-636.
15. Morris J., G. Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991, vol. 17, 1, pp. 21–43.
16. Wei T., Lu Y., Chang H., Zhou Q., Bao X. A semantic approach for text clustering using WordNet and lexical

chains. Expert Systems with Applications, 2015, 42, pp. 2264–2275.
17. Tkach S.S. Primeneniye leksicheskikh tsepochek dlya razresheniya leksicheskoy mnogoznachnosti na osnove Russkogo Viki-slovarya [Application of lexical chains for solving lexical polysemy based on the Russian Wiki Dictionary]. Masters thesis. Petrozavodsk State University, Petrozavodsk, 2016, 60 p.
18. Mitra M., Singhal A., Buckley C. Improving automatic query expansion. In Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 206-214.
19. Sahami M., Heilman T.D. A web-based kernel function for measuring the similarity of short text snippets.In Proceedings of the 15th International Conference on World Wide Web (WWW ’06), ACM, 2006, pp. 377-386.
20. Cilibrasi R.L., Vitanyi P. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 2007, 19 (3), pp. 370-383.
21. Vereshchagin N.K. Uspenskiy V.A., Shen' A. Kolmogorovskaya slozhnost' i sluchaynost' [Kolmogorov complexity and randomness]. Moscow, MTSNMO, 2013, 575 p.
22. Bollegala D., Matsuo Y., Ishizuka M. WebSim: a Web-based Semantic Similarity Measure. The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007, pp. 1-4.
23. Wong W., Liu W., Bennamoun M. Tree-traversing ant algorithm for term clustering based on featureless similarities. Data Mining and Knowledge Discovery 15 (3), pp. 349-381.
24. Bartussek W., Bense H., Hoppe T., Humm B.G., Reibold A., Schade U., Siegel M., Walsh P. Introduction to Semantic Applications. In Thomas Hoppe, Bernhard Humm, Anatol Reibold (Eds.). Semantic Applications. Methodology, Technology, Corporate Use. Springer-Verlag GmbH Germany, part of Springer Nature 2018.
25. Gorodetskiy V.I., Serebryakov S.V. Metody i algoritmy kollektivnogo raspoznavaniya [Methods and algorithms of collective recognition]. Avtomatika i telemekhanika [Automation and Remote Control], 2008, vol. 69 (11), pp. 3-40.