Семантические вычисления, большие данные, кибербезопасность

В.И. Городецкий, TRA Robotics Ltd.

Ключевые слова

- Семантика естественного языка
- **Семантические вычисления**
- **Семантические приложения**
- **Семантические технологии**
- Семантические ресурсы
- Семантические примитивы
- Модели ЕЯ-семантики
- Модели сравнительной семантики ЕЯ
- **Онтология приложения**
- **Использование онтологий**
- **Теория и факты онтологии**
- **Альтернативная семантика дескриптивной логики**
- **Кибербезопасность**

Содержание

1. Введение: Договоримся о терминах
2. Семантические приложения
3. Три кита семантических технологий:
   - Онтологии
   - Семантические ресурсы
   - Модели семантики естественного языка
4. Онтологии:
   - Понимание, создание, представление, использование
   - Альтернативная семантика дескриптивной логики и эффективное использование онтологий
5. Семантические ресурсы
6. Модели семантики естественного языка
   - Семантические примитивы для представления семантики
   - Семантика текстов и сравнительная семантика: Типы моделей
7. Заключение

Содержание

1. Введение: Договоримся о терминах

Введение: Договоримся о терминах

Базис семантических вычислений - это работа с *семантикой* естественного языка (*ЕЯ*).

Семантика ЕЯ как научное направление имеет целью выявить и представить, формально *отношения* между *сущностями ЕЯ* - словами, понятиями, группами слов (терминами), паттернами текстов и текстов в целом - и *их смыслом*, т.е. тем значением, которое эти сущности несут для человека-***носителя ЕЯ***.

Все слова и понятия ЕЯ имеют неточный, как правило, – многозначный смысл. В таких случаях человек определяет *семантику термина* с помощью ***контекста***, т.е. с помощью других слов текста, связанных с ним синтаксически и семантически.

На *формальном уровне* явное *описание*/представление *семантики* терминов выполняется с помощью ***онтологии***. Она устанавливает *единое,* одинаковое *понимание* всеми участниками процесса вычислений (пользователями и компьютерными программами) смысла ЕЯ-сущности (понятия, например).

Под термином "*семантические вычисления*" далее понимается комплекс методологий, методов, алгоритмов и реализующих их программ обработки данных, и, прежде всего, - текстовых данных, в которых ***ЕЯ-смысл*** исходных, промежуточных и конечных *данных вычислений* однозначно задается, и в процессе вычислений корректно поддерживается и однозначно интерпретируется *всеми участниками* процесса вычислений.

Введение: Договоримся о терминах

Под ***семантическими приложениями*** понимаются приложения, в которых компоненты формальных моделей приложения, e.g. понятия, их атрибуты и отношения между понятиями представлены в *терминах понятий и отношений ЕЯ,* а процессы преобразования входных и промежуточных данных приложения выполняются с помощью *семантических вычислений*.

“Семантическое приложение есть программное приложение, которое явно или неявно использует семантику предметной области” [1].

Под ***семантическими технологиями*** понимается комплекс методологий, алгоритмов и инструментальных средств их поддержки, которые используются для *проектирования и программной реализации семантических приложений*.

Онтология приложения– это *ключевое* понятие семантических технологий. *Онтология приложения* – это *спецификация* разделяемой (shared – всеми одинаково понимаемой) *концептуальной модели* приложения. Другими словами – это декларативная *мета модель* данных и знаний приложения.

[1] W. Bartussek, H. Bense, T. Hoppe, B.G. Humm, A. Reibold, U. Schade, M. Siegel, and P. Walsh. Introduction to Semantic Applications. In T. Hoppe, B. Humm, et al (Eds.). Semantic Applications. Methodology, Technology, Corporate Use. Springer, 2018.

2. Семантические приложения

2. Семантические приложения

Предполагается, что в *семантическом приложении* пользователь общается с программой *на ЕЯ*, программа общается с пользователем *на ЕЯ*, а интерфейсы построены так, что ***перевод данных*** с человеко-понятного языка на машино-понятный и обратно, где это требуется, выполняется *автоматически*.

В работе [1] приводятся примеры более десяти реализованных семантических приложений. В них ***пользователь и программа*** семантического приложения **общаются** между собой и с данными, главным образом, через ***онтологию***.

В настоящее время семантические приложения рассматриваются, прежде всего, в задачах *семантической интеграции данных*. Особенно актуальна эта функция при хранении и обработке больших данных, в том числе, больших данных, представленных текстами на ЕЯ и/или слабо структурированных данных типа XML, JSON и др., которым свойственна неполнота или отсутствие схем данных.

Роль онтологий как *семантических метамоделей данных* и *семантических вычислений* становится *ключевой* при проектировании, наполнении и поддержке “*озер данных*”, где семантические вычисления позволят обеспечивают прозрачность и лучшее понимание данных.

Примеры семантических приложений

Система *сбора текстовых документов о медицинских устройствах*, удовлетворяющих законодательству с проверкой их актуальности и погружение документов в индексированный репозиторий;

Семантическая система управления контентом для веб-портала Leipzig Health Atlas; он содержит только верифицированный контент по теме здравоохранения.

Семантическое приложение, предназначенное для *создания архивов культурного наследия* с использованием словарей различных культурных проектов.

Системы *аннотирование контента* тегами, связанными с экземплярами понятия онтологии; они "*индексирую*т" *семантику*, скрытую в данных, упрощая и ускоряя поиск информации. По сути, здесь выполняется *обогащение* данных с помощью разметки в интересах *поиска и интеграции* данных, т.к. извлекаемые ключевые слова формируют *дополнительный источник знаний*, которые далее используются в приложениях, например, для тематического моделирования, кластеризации и др.

Семантический поиск, который использует такие модели семантики как таксономии, тезаурусы, онтологии или графы знаний.

Семантическое приложение, которое помогает пользователю сформулировать запрос к веб более точно, предлагая ему несколько возможных вариантов.

Примеры семантических приложений

Во всех приведенных примерах созданных семантических приложений используется только *одна семантическая технология – технология интеграции* распределенных знаний и данных.

И это отражает реально *слабое состояние исследований* в области *семантических интерфейсов* при работе с онтологиями в различных случаях *использования* .

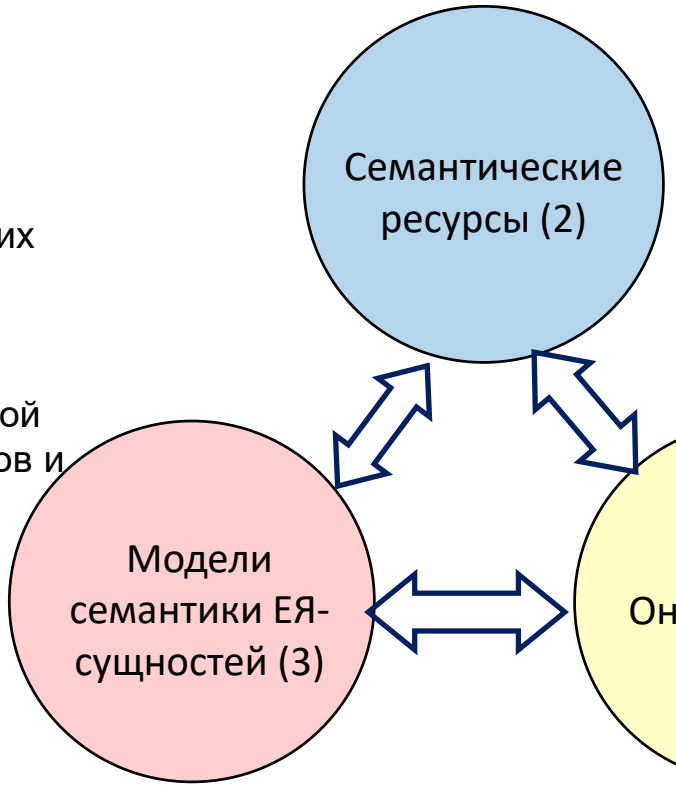
Далее будут рассмотрены подходы, которые частично решает данную актуальную проблему.

3. “Три кита” семантических технологий

“Три кита” семантических технологий

Алгоритмическое и информационное ядро семантических вычислений составляют:

- *онтологии* и эффективные алгоритмы их использования,
- *семантические ресурсы* (источники знаний о семантике естественного языка),
- *семантическая компонента*, включающая в себя модели семантики текстов и модели их сравнительной семантики.



-Ресурсы, построенные на лингвистических принципах;
- Вики-словари для разных языков
- Ресурсы, использующие совместно разные типы знаний
- Веб-ресурсы +поисковые машины веб

- Модели семантических примитивов
- Модели семантики текстов
- Модели сравнительной семантики примитивов и текстов

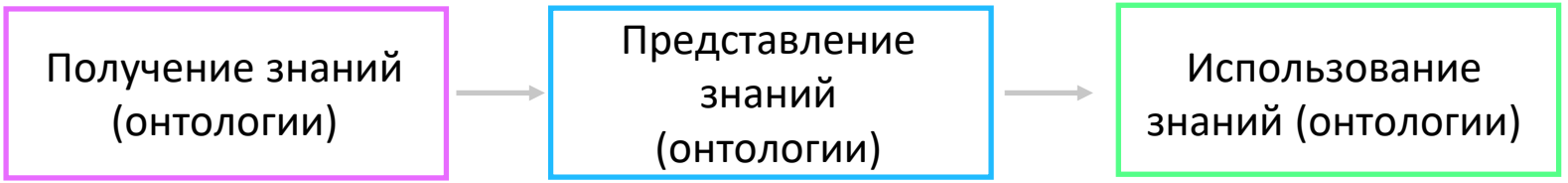
Получение, представление, *использование* онтологии

4. Онтологии

Онтологии приложения

Люди и компьютеры говорят *на разных языках,* и онтология призвана играть роль *посредника* между ними, т.е. она должна быть понятна им обоим.

Онтология приложения *– это формальная спецификация концептуальной модели знаний и данных приложения, но возможно, не всех знаний приложения.*



Получение, представление и использование знаний и данных – основные разделы и основные задачи ИИ и потому основные этапы жизненного цикла онтологий.

Использование онтологии – это множество use cases онтологии в приложениях. Их состав зависит от приложения.

Но сейчас *на практике случаи использования ограничиваются запросами на данные, обладающие какими-то свойствами.*

Онтология приложения

Подчеркнем, что здесь и далее речь идет об *онтологии приложения*

- Понятия, представляемые в модели концептуального уровня, независимо от того, представлены ли они неформально или формально, являются ***понятиями естественного языка*** (ЕЯ).

- Понятиям ***онтологии приложения*** соответствуют реальные или виртуальные объекты (классы) того или иного *уровня обобщения*. Каждый объект описывается конкретной *структурой данных* и множеством *атрибутов*, а также представляется ***непустым множеством примеров***.

- Семантика термина ЕЯ и, в том числе, *семантика понятия не может быть описана точно*, для нее могут быть определены только некоторые границы. В ***онтологии приложения*** семантика понятия определяется ***множеством его примеров***. При этом чем больше примеров понятия содержится в онтологии, тем точнее определена его семантика, т.к. каждый новый пример понятия уточняет (расширяет) его смысл.

- Онтология приложения может быть *онтологией* (мета-моделью) *данных*.

Для формального описания онтологий используются языки ***дескриптивных логик***, но разработчики и программисты *редко о них вспоминают*.

Дескриптивная логика: Язык описания формальной модели онтологии приложения

- ***Дескриптивные логики*** (ДЛ) – это фрагменты ИП-1, которые используют только одноместные предикаты, называемые *концептами,* и двухместные предикаты, называемые *ролями*, что делает ДЛ разрешимыми фрагментами ИП1 . Содержательно, *концепты* ставятся в соответствие *понятиям* онтологии, а *роли* – *бинарным отношениям* на их множестве .

- База знаний, формализованная ***в языке ДЛ***, состоит из двух компонент:

Tbox (терминология, *схема* базы знаний) содержит множество *атомарных концептов* ***A***=$\{A_1, A_2, \dots, A_n\}$, множество *атомарных ролей* $\boldsymbol{R} = \{r_1, r_2, \dots, r_m\}$ и ряд *синтаксических* правил, с помощью которых *индуктивно* задаются другие (новые) концепты и новые роли (отношения) базы знаний. Не вдаваясь в детали *синтаксиса* этих правил, отметим, что в правилах порождения новых понятий и отношений используются символы ⊤ и ⊥, называемые "*истина*" и "*ложь*", *операторы* (связки) ⊓ (пересечение), ⊔ (объединение), ¬ (дополнение), и *кванторы* всеобщности ∀ и существования ∃.

Abox - это множество *экземпляров* понятий и ролей, называемых *фактами*. *Синтаксис Abox* включает в себя утверждения двух видов:

- *a: A* ("*a* является экземпляром понятия *A*"), и
- *aRb* ("экземпляры понятий *a* и *b* связаны ролью *R*").

Булева семантика дескриптивной логики

Эта семантика *традиционно* используется для формального описания онтологий в терминах ДЛ. В ней любой *формуле* f из множества формул $\mathcal{F}$ (одноместных и двухместных формул из множества *Tbox*) ДЛ ставится в соответствие *оценка* (*интерпретация*)

$$i\,(f)\colon \mathcal{F} \to \{0,\,1\}\,.$$

В ней символ “0” интерпретируется как “*ложь*”, а “1” – как “*истина*”. Множество формул $\mathcal{F}$ строится индуктивно с помощью операций множества $\{\sqcap,\ \sqcup,\ \neg\}$, символов констант $\{\top,\ \bot\}$ и кванторов $\forall, \exists$, булева семантика которых задается традиционным образом.

Интерпретации **э**тих *формул* вычисляются на основе заданных интерпретаций *атомарных формул* (атомарных концептов и атомарных ролей ДЛ) и хорошо известных таблиц истинности для указанных *логических операторов* и *кванторов*, а также на основе интерпретаций *логических констант* $i\,(\top) = 1$ и $i(\bot) = 0$ для символов $\top$, $\bot$.

Булева семантика ДЛ позволяет исследовать формальные свойства множеств концептов схемы *Tbox* и алгоритмические проблемы их установления, ***не опираясь*** на конкретное содержание ***множеств примеров*** из *Abox* базы знаний.

Алгоритмические проблемы ДЛ, решаемые в терминах булевой семантики ДЛ

Совместная выполнимость атомарных концептов и ролей (терминологии) базы знаний, представленных в Tbox, - это проблема установления непротиворечивости множества атомарных концептов и ролей Tbox, т.е. проверка отсутствия ошибок в их задании.

Выполнимость (выводимость) произвольного концепта, представленного в виде некоторой формулы ДЛ, например, формулирующей запрос к базе знаний, т.е. проверка логического следования такой формулы из атомарных концептов базы знаний, представленных в Tbox.

Выполнимость отношения вложения пары атомарных концептов $C \sqsubseteq D$ в схеме Tbox (отношения частное – общее).

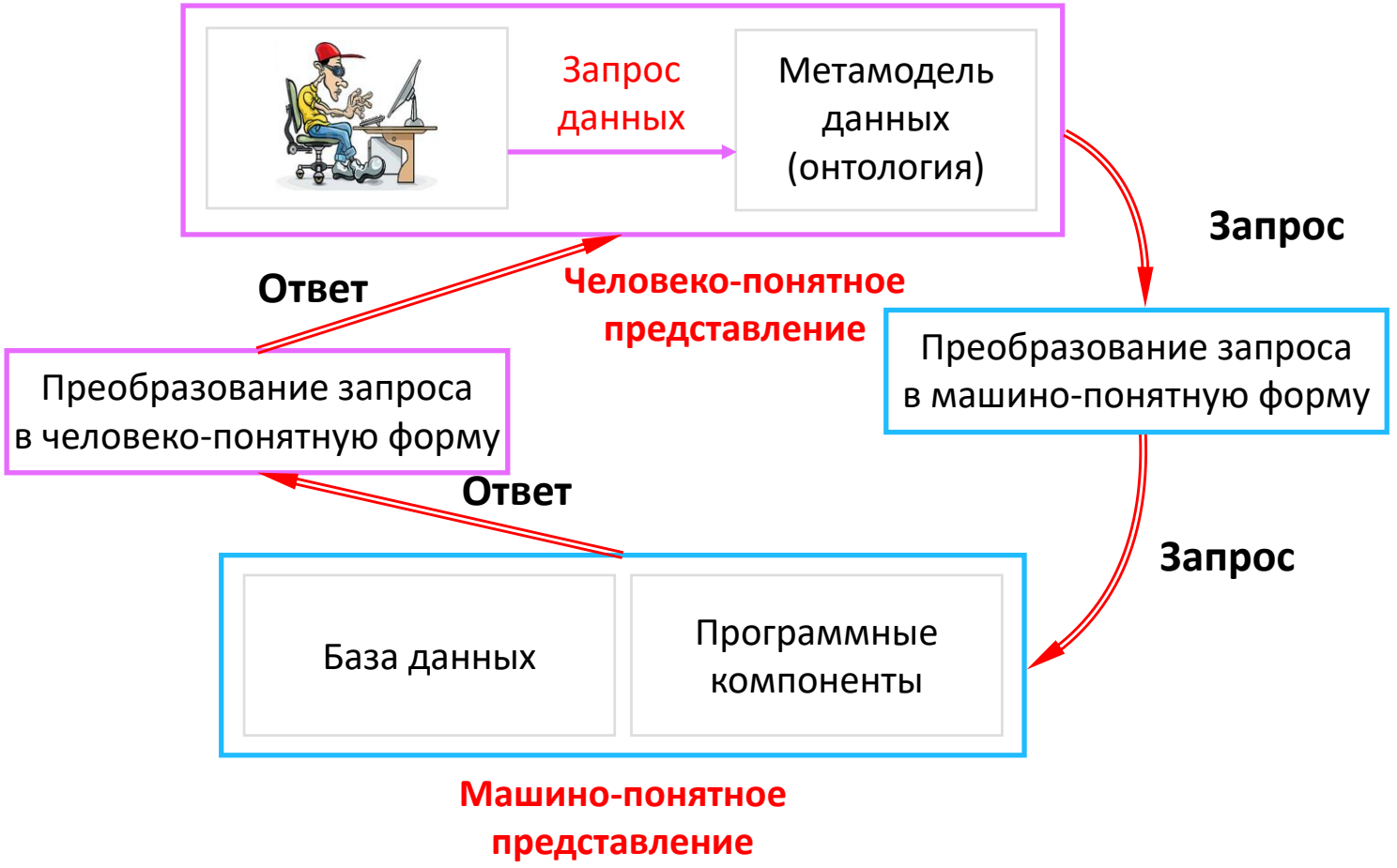
Построение классификации для всех атомарных концептов Tbox на основе проверки выполнимости отношения вложения $A \sqsubseteq B$ для любой пары концептов A и B из Tbox; эта задача имеет целью построение таксономии концептов Tbox, и проверку избыточность множества атомарных концептов.

По существу, все эти алгоритмические проблемы сводятся к первой проблеме-проблеме совместной выполнимости атомарных концептов и ролей.

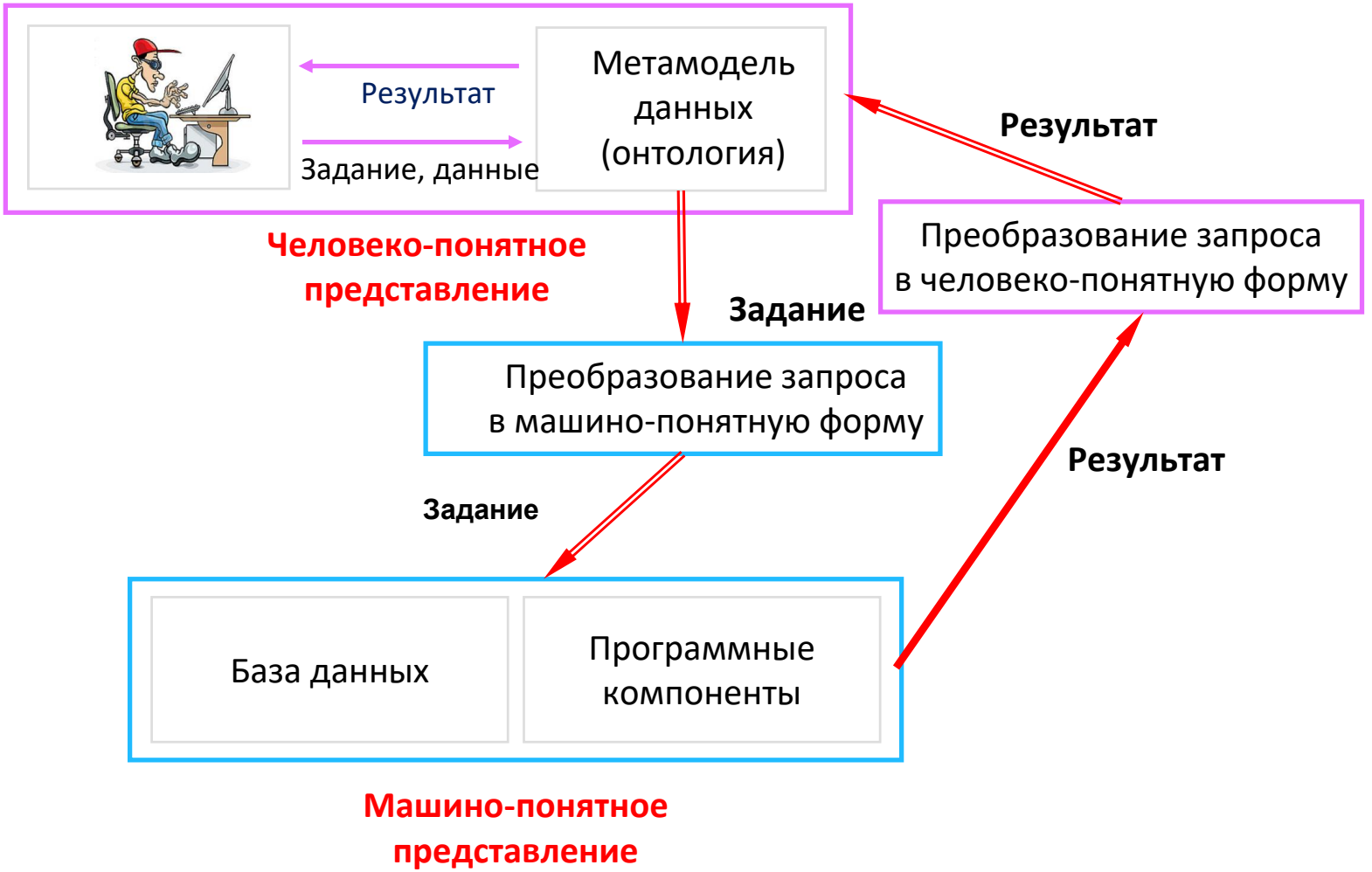
Булева семантика дескриптивной логики

ДЛ с булевой семантикой решает задачи установления свойств онтологии (типа теорем существования), но *не поддерживает базовые случаи ее использования*

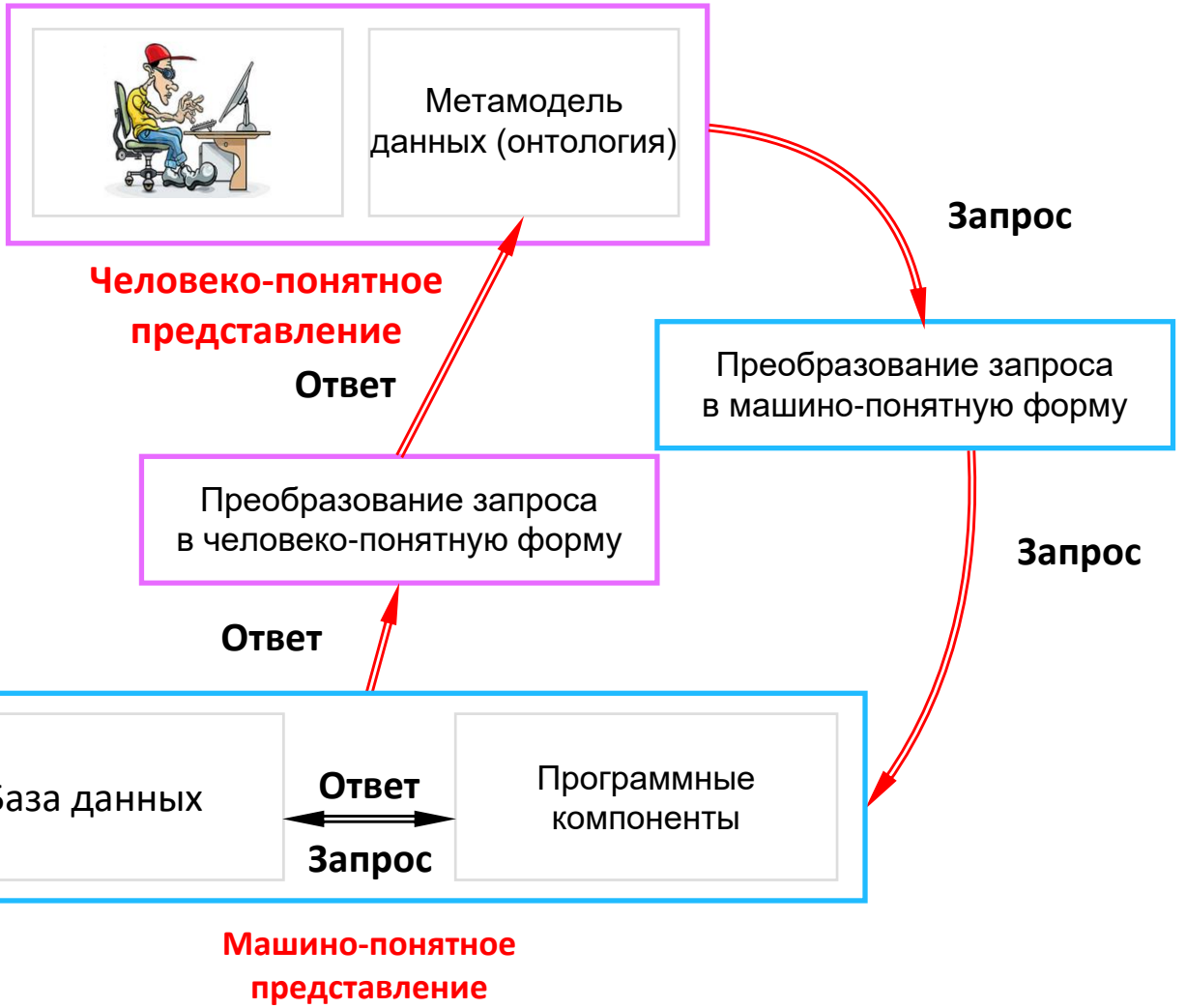
Случай использования (use case) 1: Запрос пользователя на данные



Случай использования (use case) 2: Исполнение приложения



Случай использования (use case) 3: Запросы на данные от программ



Предикатная (модельная) семантика дескриптивной логики

Предикатная семантика (интерпретация) понятий и отношений ДЛ формально записывается в виде пары:

$$I = (\Delta^I, \cdot^I),$$

где Δ^I—непустое множество (*домен* данной *интерпретации*),

$\cdot^I$—*интерпретация*, в которой на место символа "$\cdot$" подставляется или имя концепта *A*, или имя роли *r*.

Предполагается, что Δ^I есть *универсальное множество* для рассматриваемой интерпретации, т.е. *множество всех экземпляров концептов онтологии и всех экземпляров ролей* (отношений).

Тогда интерпретация (Δ^I, A^I) есть непустое подмножество A^I множества Δ^I ($A^I \subseteq \Delta^I$), которое *содержит множество экземпляров этого концепта* в онтологии. Другими словами, $A^I \subseteq \Delta^I$ — это *множество истинности* для предиката ***A***..

Аналогичным образом задается предикатная семантика для ролей *Tbox:*

Если A^I и B^I есть интерпретации концептов *A* и *B* и *r*(*A*, *B*) есть бинарное отношение на этой паре концептов, то интерпретация отношения r^I задается непустым подмножеством пар $< a_i, b_j >$ декартова произведения $A^I \times B^I \subseteq \Delta^I \times \Delta^I$, для которых отношение *r*(*A*, *B*) выполнено.

Предикатная (модельная) семантика дескриптивной логики

В этой интерпретации множество операторов булевой семантики *отображается* в множество *теоретико-множественных* операторов:

$$\{\sqcap, \sqcup, \neg, \top, \bot\} \rightarrow \{\cap, \cup, -, \emptyset, \Delta\},$$

Предикатная интерпретация представляется *удобной при формировании запросов к базе знаний*, метамодель которой представлена схемой *Tbox,* если для всех атомарных концептов и атомарных ролей заданы их множества истинности как подмножества универсального множества примеров Δ (для концептов) и их декартова произведения $\Delta \times \Delta$ (для ролей).

Эта интерпретация *не рассматривалась ранее применительно к ДЛ*, поскольку сама ***ДЛ создавалась для рассуждений с объектами Tbox***, но не с базой знаний приложений, в которой для всех понятий имеются множества примеров.

Пояснения требует *вычисление предикатных интерпретаций* концептов и ролей, которые построены с **использованием *кванторов* $\forall$ и $\exists$ *ДЛ***. Алгоритмы вычисления этих интерпретаций несколько сложнее, но для конечных *множеств примеров и ролей они вычисляются с помощью* переборных алгоритмов.

Предикатная семантика дескриптивной логики и теория моделей

Описанный подход к *установлению связи между синтаксисом и семантикой* в формальных языках, в нашем случае – между синтаксисом атомарных концептов и других правильно построенных концептов и ролей, заданных *синтаксисом* ДЛ, и их *предикатной семантикой* хорошо известен, а его корректность следует из ***теории моделей*** [Мальцев-1965].

В этой теории **предикатная семантика** любой атомарной и другой правильно построенной *формулы*, которая *истинна в заданной интерпретации* $I = (\Delta^I, \cdot^I)$, называется *множеством ее (формулы) моделей*.

Таким образом, *предикатная семантика* любого правильно построенного концепта или роли ДЛ *строится однозначно* при использовании соответствия $\{\sqcap, \sqcup, \neg, \top, \bot\} \rightarrow \{\cap, \cup, -, \emptyset, \Delta\}$

операторов и констант ДЛ и теоретико-множественных операций и констант, используемых при вычислении предикатных интерпретаций формул ДЛ.

В терминах предикатной интерпретации *формула выводима*, если она имеет хотя бы *один пример в Abox* базы знаний, тождественно *ложная формула* не *имеет в нем примеров*, а для *общезначимой* формулы *все примеры Abox* являются и примерами формулы.

Представление онтологии с булевой семантикой

Tbox – множество атомарных объектов онтологии (концептов и ролей)

$P_1^{(4)}$	Tetris	Loss function	$P_1^{(7)}$	…	$P_1^{(3)}$	Learning	Machine learning	…

Запросы на данные, представленные в терминах онтологии (ЕЯ-запросы

Система поиска ответов на запросы

SQL-база данных (как правило)

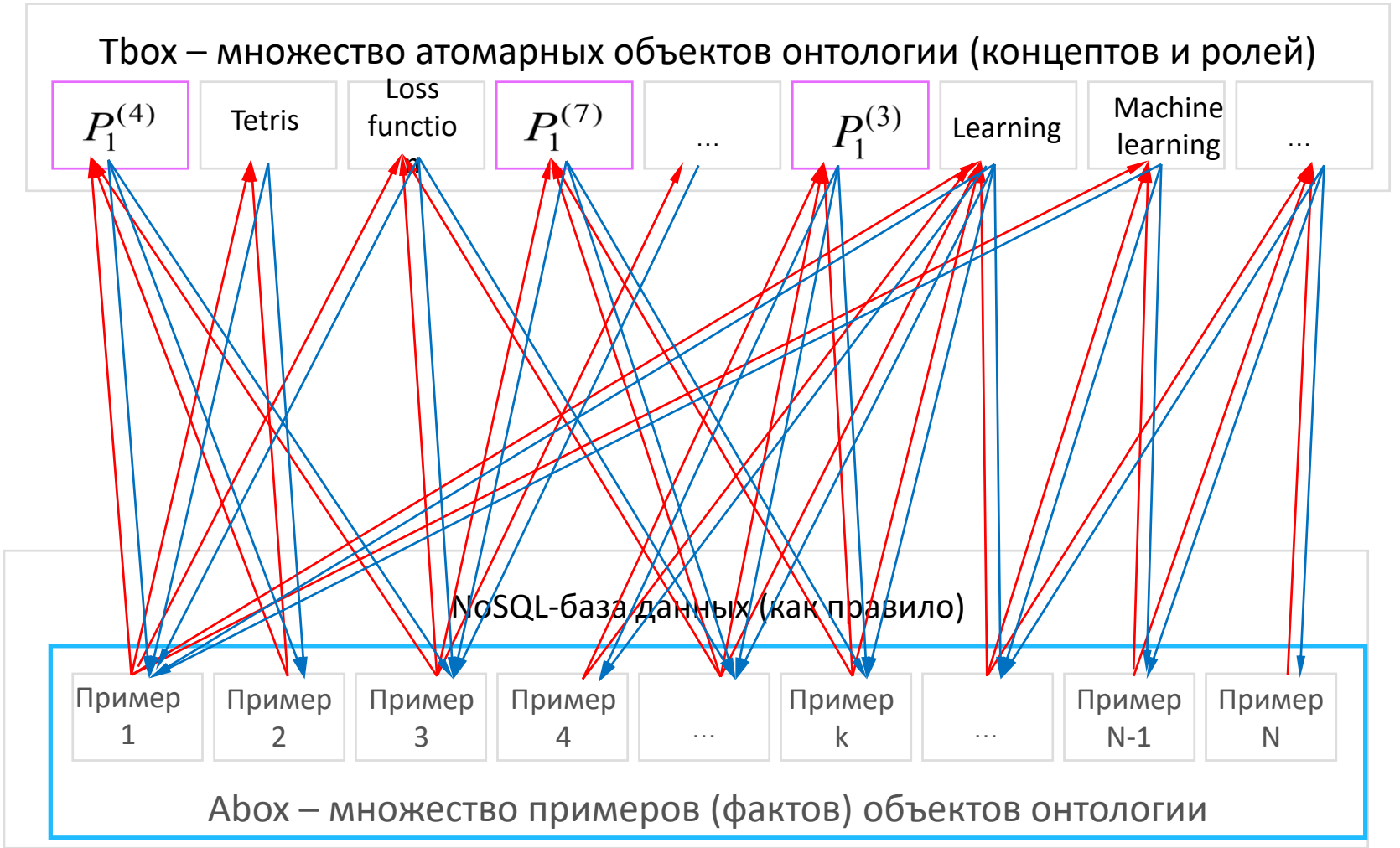
Пример 1	Пример 2	Пример 3	Пример 4	…	Пример k	…	Пример N-1	Пример N

Abox – множество примеров (фактов) объектов онтологии

Представление онтологии с булевой семантикой

В представлении онтологии приложения, которая содержит и *Tbox,* и *Abo*x , булева семантика оказывается не при чём: она не влияет на представление и никак на нем не отражается

А что нового в представлении онтологии с предикатной семантикой?



В представлении онтологии приложения, которая содержит и *Tbox*, и *Abox* , предикатная семантика может быть представлена в явном виде *двунаправленными связями* между объектами *Tbox*, и Abox, которые принципиально *меняют* возможную *технологию отработки* различных случаев использования онтологии

А что нового в представлении онтологии с предикатной семантикой?

Tbox – множество атомарных объектов онтологии (концептов и ролей)

$P_1^{(4)}$	Tetris	Loss function	$P_1^{(7)}$	…	$P_1^{(3)}$	Learning	Machine learning	…

Запросы на данные, представленные в терминах онтологии (ЕЯ-запросы)

Модель двунаправленных связей “Атомарные объекты Tbox” “Примеры Abox”

NoSQL-база данных (как правило)

Пример 1	Пример 2	Пример 3	Пример 4	…	Пример k	…	Пример N-1	Пример N

Abox – множество примеров (фактов) объектов онтологии

5. Семантические ресурсы

Семантические ресурсы

Потенциальное качество работы сематических приложений определяется, в первую очередь и главным образом, качеством использованных семантических ресурсов –источников знаний о семантике сущностей ЕЯ.

Качество семантических ресурсов, т.е. их потенциальные возможности по установлению семантики ЕЯ-текстов, определяются двумя главными факторами:

(1) *мощностью* представленной в нем *лексики* (слов, понятий) и *множеством сематических отношений* (синонимов, антонимов, омонимов и др.) , заданных на его *лексических единицах*, примеров их использования, а также дополнительной информацией о каждом из них, имеющейся в ресурсе явно или неявно;

(2) *структуризацией* его компонент, а также наличием и *эффективностью поисковых инструментов* для поиска ЕЯ-сущностей словаря, связей между ними, и другой информации семантического и синтаксического характера, которая в ресурсе присутствует явно или неявно

Семантические ресурсы

1. *Ресурсы, построенные на лингвистических принципах* (*linguistically constructed*)

Примерами являются семантические словари типа WordNet в английском, Европейском (немецком и др.) и русском вариантах. Они содержат понятия, множество их синонимов и примеров понятия, а также ряд семантических отношения для понятий, например, антонимов, паронимов (*live* и *leave)*, гипонимов, гиперонимов, дополнительную информацию в виде текстовых описаний понятий и др.. Их называют также *лексическими базами данных*

2. Вики-словари для различных языков (англ. Wikitionary),

Они представляют собой *многофункциональные словари и тезаурусы*. Словарные статьи в них содержат информацию о *синтаксических* и *морфологических* свойствах слова, информацию о его значении (аналогично ресурсам, построенным на лингвистических принципах), а также различные *родственные слова*. К ресурсам такого же уровня информативности относят *предметно-ориентированные онтологии*.

Ресурсы Вики-словарей и онтологий с их информационными и структурными отношениями *намного богаче* чисто лингвистических ресурсов типа WordNet.

Семантические ресурсы

3. Ресурсы, совместно использующие разные типы знаний (*collaborative resources*).

Примером является *Wikipedia* с ее инструментами типа *Dbpedia* и ее варианты на различных языках, включая русский. Этот ресурс содержит *словарные статьи* (как и в Вики-словарях) для категорий понятий и их синонимов, *ссылки* на *более общие* и *более специализированные* понятия, тексты, уточняющие семантику понятия, а также ряд других информационных и структурных отношений. Заметим, что *словарный запас* Wikipedia *намного богаче*, чем у ранее названных ресурсов, и, что еще важнее, он постоянно пополняется.

4. Ресурсы Веб в целом совместно с множеством поисковых машин (Google и др.)

Это наиболее богатые ресурсы и они весьма разнообразны. Например, они содержат Википедию и все множество веб-страниц (сейчас их более 10 млрд.), ресурсы Семантического веба, например данные *Linked Open Data*, представленные в облаке. Использование *поисковых машин* дает возможность получить в *явной форме* знания, которые представлены в веб-ресурсах явно или неявно, а это очень много. Но поисковым машинам доступны только те знания , которые содержатся в веб-ресурсах, *индексированных* в используемой поисковой машине. Обычно они активно используют *средства собственных онтологий* и другие семантические средства поиска, скрытые от пользователей.

.

6. Модели семантики естественного языка

VSM-модель семантики текста онтологии приложения

Семантическими примитивами (СП) называют некоторые компоненты, или сущности ЕЯ, например, слова, понятия, фрагменты текста или/и результаты трансформаций фрагментов текста или даже текста в целом.

Семантика ЕЯ-текста обычно описывается множеством *СП*, *смысл* каждого из которых должен быть определен *однозначно* (с точностью до синонимов). Иногда им приписываются значения *весовых коэффициентов*, которые задают относительную *важность* примитивов в формировании смыла текста. Основная задача –определить смысл СП (*Word Sense Disambiguation*, *WSD*)

Множество СП обычно рассматриваются в качестве признаков, описывающих семантику текста в пространстве этих примитивов. Тогда модель семантики текста описывается формально точкой в соответствующем векторном пространстве, а саму модель называют моделью векторного пространства (Vector Space Model, VSM).

Типы семантических примитивов

N-граммы - последовательности из N слов (обычно от двух до четырех), извлеченных из текста (BOW). Каждой N-грамме ставится в соответствие величина некоторой меры, оценивающей ее относительную значимость в формирование смысла всего текста, например ContBow, TF-IDF, BM25 и др. Обычно число N выбирается в пределах от двух до четырех.

Термины (terms) – это современная модификация N-грамм. Они чаще формируются из понятий или строятся как более сложные агрегаты (Астрахнцев-2014, Ermolaev 2017), частности, латентные факторы, лексические цепи (последовательности слов текста с бинарными отношениями на словах).

Все они строятся с использованием того или иного *семантического ресурса* и с решением WSD-задачи.

V. Ermolaev, *et al*. Terminological Saturation in Retrospective Text Document Collections: Cross-Evaluation of Automated Term Extraction Tools. Техн. отчет, 2017.

Н. Астрханцев. Диссертация, 2014.

Сравнительная семантика: Зачем это нужно?

В семантических приложениях наиболее часто решаются задачи типа кластеризации множества текстов и классификации новых текстов при заданном множестве классов.

Во всех этих задачах алгоритмы машинного обучения строятся, главным образом, на поиске *сходства* и *различий* в свойствах текстов. Они оцениваются с помощью функций, которые в той или иной форме характеризуют *связи между семантическими примитивами* (признаками) модели семантики текстов.

В семантических технологиях машинного обучения эти функции базируются на понятии *семантической связанности* пары СП, под которой понимают некоторую функцию (меру), которая *численно оценивает близость их семантики* (смысла) с помощью *анализа отношений*, заданных на паре СП.

Функция, аргументами которой является пара СП 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 полную смысловую эквивалентность ЕЯ-сущностей.

Типы мер близости

1. *Топологические меры близости*. Они измеряют длину пути (“семантическое расстояние”) между понятиями в общей таксономии понятий онтологии. Для их вычисления нужно иметь в распоряжении *иерархию понятий онтологии*, в которой представлены сравниваемые понятия. Такие меры могут использовать все виды семантических ресурсов ЕЯ. Примерами являются мера Лекокка *lch* [26], мера Ву и Палмера *wup* [27] и мера Ли *li* [25, 28].

2. *Меры, использующие оценку информационного содержания* сравниваемых понятий (англ. *information content, IC*). Для мер этой группы, как правило, *нужно иметь выборку документов*, которые содержат сравниваемые понятия, поскольку в них используется частота встречаемости каждого понятия в выборке. К этой группе относят меры Резника *res*, Лина *lin* и расстояние Джанга *jang* и др. Например, *мера res* для двух понятий вычисляется как *значение информационного содержания*, сосчитанное для их минимального общего родительского понятия *c* в иерархической структуре онтологии (*Most Common Specific Abstraction*, *MCSA*):

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

где $$IC(c) = -log(p(c))$$

Меры обоих типов могут работать со всеми теми *семантическими ресурсами*, которые содержат *онтологию*

Типы мер близости

3. *Меры которые описываются вектором признаков*. Для вычисления этой меры семантической близости пары СП используется *векторное представление* описания *слова* и *текста* в целом с последующим их сравнением, например, с помощью стандартной косинусной меры сходства.

Пример – мера *ECA* (*Explicit Semantic Analysis*). Сначала каждому слову пары ставится в соответствие *вектор понятий Википедии*, примером которых это слово является. Каждое такое понятие Википедии представляется в ней текстом словарной статьи, а множество этих текстов рассматривается далее в качестве обучающего множества, с помощью которого вычисляется значение *TF-IDF* понятий сравниваемой пары по множеству этих текстов. В итоге каждому слову текста ставится в соответствие взвешенный вектор понятий и по ним вычисляется мера близости слов в пространстве понятий Википедии. Для пар текстов получаются векторы, сравниваемые по косинусной мере сходства.

4. *Гибридные методы* объединяют идеи описанных выше подходов, 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)

Типы мер близости

5. *Меры семантической близости для лексических цепей*. Лексическая цепь описывается последовательным словами текста и бинарными отношениями на них. Каждая такая цепь – это некоторый граф, представляющий одну из смысловых компонент текста, одну из его тем. Множество взвешенных лексических цепей используют как *множество агрегированных семантических признаков текста*. Существуют разные варианты построения цепи, отвечающие разным вариантам вычисления меры близости.

Типы мер близости

6. *Меры, основанные на расширении контекста коротких текстов за счет семантических ресурсов.* Основная идея этой группы подходов к установлению *семантики короткого текста* и его сходства с другими аналогичными текстами состоит в том, что этот текст используется в качестве *запроса поисковой машины* веб, что позволяет извлечь много веб-документов. Поскольку результат веб-поиска – это множество документов по смыслу близких к запросу, то их можно рассматривать как *расширение контекста запроса* для установления смысла его терминов. Таким образом, 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 может применяться косинусная мера близости пары текстов, которая должна в итоге дать более робастный результат.

7. *Меры близости, использующие Google-семантику*. Мотивация подхода к достаточно проста: неструктурированное множество веб-документов есть практически универсальная выборка текстов, а потому она может быть использована для поиска их семантики и семантической близости так же, как и *другие множества документов*, на которые опираются, например, все описанные выше варианты решения таких задач.

Типы мер близости

По мнению авторов, слова, близкие по смыслу, должны приводить к похожим результатам поиска в веб. Поэтому даже число документов, содержащих оба слова, которые найдены поисковой машиной (e.g., Google-машиной) для двух слов, может использоваться в качестве аргумента функции, описывающей меру их близости. поскольку, в соответствии с мотивацией авторов, чем больше *относительное число таких общих страниц для пары слов* и/или для пары текстов, тем они *ближе по смыслу*. Итоговое выражение для метрики, которую называют *нормализованным Google-расстоянием* (normalized Google distance, NGD) имеет следующий вид:

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

В этой формуле $f(x)$ и $f(y)$ есть число страниц, возвращаемых Google по запросам *x* и *y*, которые содержат x и y, соответственно, а $f(x,y)$ – это число страниц, возвращаемых Google по запросу xy, в которые содержат обе цепочки. $NGD(x,y)$ –семантическая близость имеет иной смысл, чем это характерно для других методов. Формально она *не различает термины с разным смыслом*.

Позднее были предложены аналогичные подходы, которые используют Google-семантику *совместно с другими веб-ресурсами.*

7. Заключение

Заключение

1. *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*.
2. Показано, что *зрелость семантической технологии* определяется уровнем интеграции и качеством технологий реализации *трех его базовых компонент*, а именно *онтологии* как интегратора знаний и данных, качеством и полнотой используемых *семантических ресу*рсов, особенно веб-ресурсов, а также качеством модели *формализации ЕЯ-семантики текстовых*, слабо структурированных и структурированных данных.
3. В работе показаны *преимущества использования формальной модели онтологии в терминах дескриптивной логики с предикатной (модельной) семантикой*, которая способна качественно улучшить эффективность использования онтологии, которую следует рассматривать как центральное звено в триаде базовых компонент семантических технологий, семантических вычислений и семантических приложений.

Спасибо!

Вопросы?