# СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СИСТЕМЫ TEXTERRA

**Турдаков Д. Ю.** (turdakov@ispras.ru),
**Андрианов И. А.** (ivan.andrianov@ispras.ru),
**Астраханцев Н. А.** (astrakhantsev@ispras.ru),
**Майоров В. Д.** (vmayorov@ispras.ru),
**Недумов Я. Р.** (yaroslav.nedumov@ispras.ru),
**Сысоев А. А.** (sysoev@ispras.ru),
**Федоренко Д. Г.** (fedorenko@ispras.ru)

Институт системного программирования РАН,
Москва, Россия

**Ключевые слова:** семантический анализ текстов, Википедия, базы знаний, семантические онтологии, Викификация

# SEMANTIC ANALYSIS OF TEXTS USING TEXTERRA SYSTEM

**Turdakov D. Y.** (turdakov@ispras.ru),
**Andrianov I. A.** (ivan.andrianov@ispras.ru),
**Astrakhantsev N. A.** (astrakhantsev@ispras.ru),
**Mayorov V. D.** (vmayorov@ispras.ru),
**Nedumov Y. R.** (yaroslav.nedumov@ispras.ru),
**Sysoev A. A.** (sysoev@ispras.ru),
**Fedorenko D. G.** (fedorenko@ispras.ru)

Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

Texterra delivers a scalable solution for text processing based on novel methods which exploit knowledge extracted from the Web and collections of domain-specific documents. The paper describes the process of semantic model construction within Texterra. The system first detects compound terms and annotates each term with a meaning by assigning an appropriate concept from the knowledge base using disambiguation algorithm. After that, the key concepts are extracted by detecting the most relevant concepts to the text. Information from Wikipedia is used as a basis for automatic knowledge base construction. In addition, Texterra provides tools for

extending knowledge base with information extracted from websites and collections of domain-specific documents.

Evaluation results include term extraction, word sense disambiguation, and key concept extraction from Russian and English corpora. Current technology level allows using Texterra for improving quality of different applications, such as semantic search, recommender systems, and social network analysis. Demonstrations and API are available at https://api.ispras.ru.

**Keywords:** semantic analysis, Wikipedia, knowledge bases, semantic ontologies, Wikification

## 1. Введение

Применение баз знаний, или онтологий, показало свою эффективность во многих приложениях, связанных с обработкой естественного языка, таких как извлечение информации, вопросно-ответные системы и информационный поиск. Использование баз знаний позволяет осуществить переход от отдельных слов к выражаемым ими понятиям, что, в свою очередь, сокращает влияние разреженности языка и многозначности лексических единиц [Biemann 2005].

Идея использования баз знаний для семантического анализа текстов лежит в основе проекта Texterra. Texterra представляет собой технологию для многоязычного анализа текстовых документов, которая основана на использовании знаний, извлекаемых из Веб-ресурсов и коллекций документов. Данная технология позволяет добиться высокой точности анализа при низких затратах на обучение и настройку.

Система Texterra предоставляет широкий набор инструментов для решения задач обработки текстов, включающий в себя как стандартные методы, например определение частей речи, так и оригинальные методы, основанные на использовании базы знаний. Кроме того, Texterra включает в себя инструменты для обработки неформальных пользовательских текстов, таких как сообщения социальных сетей [Korshunov 2013].

Большая часть функциональности системы Texterra доступна через открытый программный интерфейс[1], который уменьшает расходы на интеграцию системы обработки текстов в пользовательские проекты. Таким образом, проект Texterra может представлять интерес для широкого круга разработчиков и исследователей, которым необходимы инструменты для обработки текста.

В данной статье описывается процесс построения семантической модели для произвольного текста на естественном языке с помощью системы Texterra. Семантическая модель текста состоит из трех частей:

1. Граф концептов, описывающих значения терминов текста. Каждому концепту назначается вес, демонстрирующий важность данного концепта для описания текста.

[1] https://api.ispras.ru

2. Термины текста и связи между терминами и концептами. Каждый термин связан только с одним концептом. Такой концепт является значением термина. При этом концепт может быть связан с несколькими терминами. Такие термины являются синонимами.
3. Связи между найденными концептами и другими концептами базы знаний. Эти связи позволяют перейти от работы с одним документам к коллекциям документов.

Пример семантической модели текста представлен на рисунке 1.

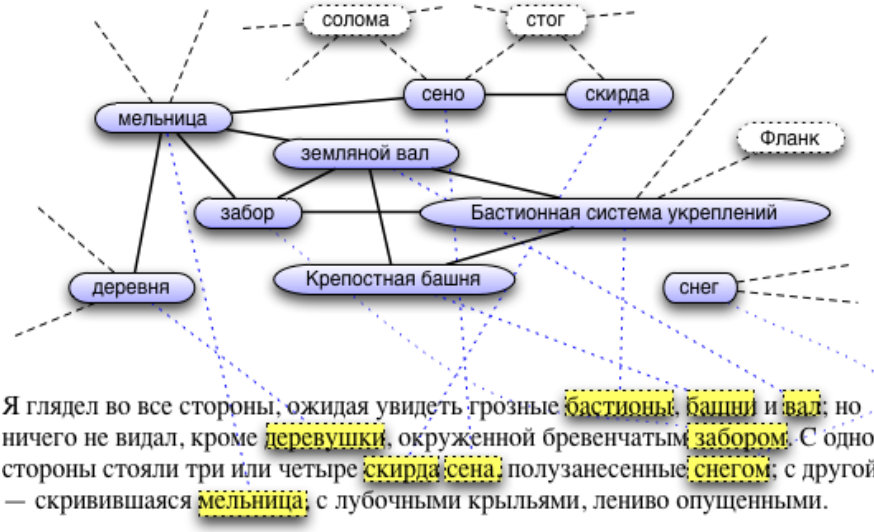


**Рис. 1.** Семантическая модель текста. Показаны связи между концептами с семантической близостью более 0,15

Предложенная модель является достаточно общей для решения широкого круга задач, связанных с обработкой текста. При этом для построения модели используются автоматические методы с высокой точностью и скоростью работы, что делает такой подход привлекательным для практического применения.

Следующий раздел посвящен базе знаний, необходимой для построения семантической модели. Процесс построения модели представлен в третьем разделе. В четвертом разделе приведены результаты экспериментального тестирования. В разделе 5 представлен обзор наиболее близких альтернативных подходов. В заключении описаны преимущества и возможные приложения системы Texterra.

## 2. База знаний Texterra

Для построения семантической модели текстов, необходима база знаний, содержащая концепты и их текстовые представления. База знаний системы

Texterra состоит из двух уровней: на уровне онтологии находится граф концептов, на уровне представления находятся термины, отражающие концепты в естественных языках (рис. 2). Узлами графа концептов являются непосредственно концепты, ребрами — связи между ними. Каждый концепт может иметь одно или несколько текстовых представлений, являющихся синонимами. В свою очередь, каждое текстовое представление может быть представлением одного или более концептов, отражая явление многозначности.

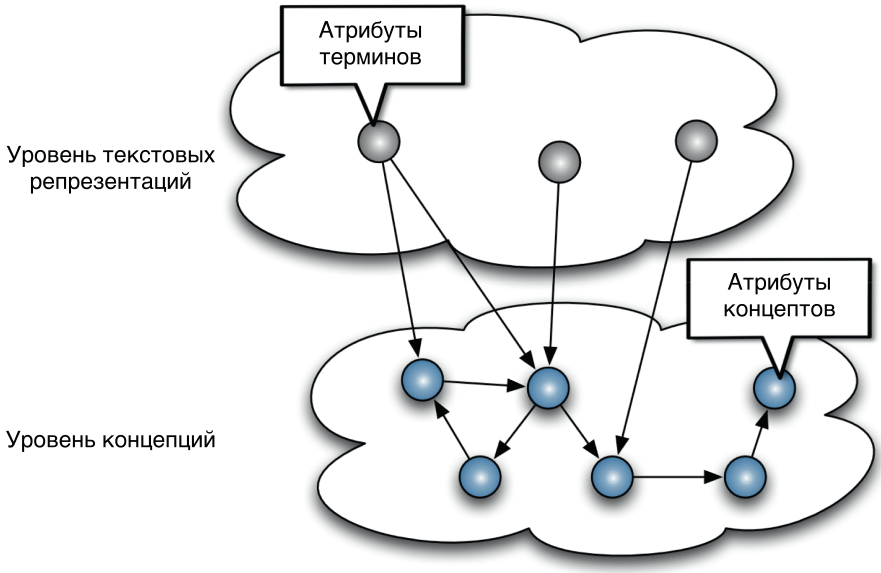


**Рис. 2.** Двухуровневая модель базы знаний Texterra

Подсистема управления базой знаний Texterra предоставляет интерфейс, состоящий из трех групп функций:

1. Функции перемещения по графу концептов позволяют перемещаться по ребрам графа и находить концепты, тем или иным образом связанные с данным, а также возвращать значения атрибутов концептов.
2. Функции подсчета семантической близости позволяют вычислять семантическую близость между отдельными концептами или группами концептов.
3. Функции привязки к онтологии позволяют находить для заданных терминов связанные с ними концепты и характеристики этих связей.

Для построения базы знаний системы Textrra разработаны два инструмента [Turdakov 2014]. Первый из них позволяет строить базу знаний на основе Википедии и других ресурсов работающих на технологии MediaWiki, второй — на основе произвольной коллекции текстов, относящийся к одной тематике.

Основой базы знаний системы Texterra служит информация, извлекаемая автоматическими методами из открытой Интернет-энциклопедии Википедии. Каждая неслужебная статья Википедии, описывающая некоторое понятие реального мира, преобразуется в концепт; каждое название статьи и названия страниц-перенаправлений преобразуются в соответствующие концептам термины. Кроме того, в словарь терминов добавляется текст гиперссылок, при условии, что это текст встретился в гиперссылках на соответствующую статью не менее пяти раз. Например, статья «Российская академия наук» преобразуется в концепт, которому соответствуют термины «Российская академия наук», «РАН» и другие термины-синонимы.

Для покрытия узкоспециализированных предметных областей, разрабатывается инструмент для обогащения базы знаний на основе анализа коллекций предметно-ориентированных текстов. На основе алгоритмов машинного обучения и эвристических правил этот инструмент позволяет извлекать термины предметной области, соответствующие им концепты и связи между концептами, расширяя, таким образом, существующую базу знаний на предметную область.

Подсистема управления базой знаний Texterra позволяет эффективно вычислять специальный тип отношений между концептами — семантическую близость. Это отношение определено для любой пары концептов или групп концептов и выражается некоторым числом от 0 до 1: чем ближе значение функции к 1, тем больше общего между концептами. Семантическая близость вычисляется на основе гиперссылок между статьями с помощью меры Дайса (нормализованное число общих соседей, т.е. статей, имеющих гиперссылки с одной на другую) [Turdakov 2008]. Оптимизация работы с именно таким типом отношений связана с тем, что, с одной стороны, семантическая близость позволяет решать задачи обработки текстов, а с другой стороны, существенно снижает затраты на построение базы знаний, за счет снижения требований к формальности онтологии.

Таким образом, с помощью системы Texterra можно строить семантические модели для текстов, как общей тематики, так и более специализированных. При этом, система не накладывает жестких требований к сложности онтологии и может работать как с формальными онтологиями, содержащими такие типы связей между концептами как IS-A, PART-OF и другими, так и с простыми семантическими сетями.

## 3. Автоматическое построение семантической модели

Построение семантической модели текста с помощью базы знаний является одной из основных задач, решаемых системой Texterra. Процесс построения семантической модели состоит из трех этапов: распознавание терминов, определение значений терминов и извлечение ключевых концептов.

На первом этапе из текста выделяются термины, присутствующие в базе знаний системы Texterra. Далее для каждого найденного термина запускается

алгоритм разрешения лексической многозначности, выбирающий для каждого найденного термина наиболее подходящее значение — концепт базы знаний. Заключительным этапом построения семантической модели является извлечение ключевых концептов, позволяющее получить сжатое высокоуровневое представление текста в виде ключевых концептов. Этапы разрешения лексической многозначности и выделения ключевых концептов могут работать без модификации для текстов на любом языке, для которого есть соответствующая языковая версия Википедии. Далее описываются детали каждого этапа.

Наиболее эффективный способ определения терминов — это поиск текстовых строк из словаря базы знаний. Термины могут состоять из нескольких слов, но при этом они не должны пересекаться. Если для некоторой предметной области полнота базы знаний недостаточна, ее можно расширить новыми терминами и концептами с помощью инструмента, кратко описанного в предыдущем разделе. Для поиска терминов используется жадный алгоритм. При этом на основе алгоритма определения частей речи фильтруются термины, не являющиеся именными фразами.

Для определения значений терминов используется классификатор на основе метода максимальной энтропии. Следуя работе Милна [Milne 2008] мы используем следующие признаки: вероятность появления термина в ссылке; частота употребления концепта; семантическая близость к контексту; качество контекста.

Значения первых двух признаков вычисляются заранее на основе анализа источника базы знаний и сохраняются для дальнейшего использования. В случае использования в качестве источника Википедии, для аппроксимации вероятности появления термина в ссылке подсчитывается частота его появления в виде заголовка ссылки (caption) в тексте статей Википедии. Для определения частоты употребления концепта, во всех случаях употребления термина в качестве заголовка ссылки агрегируются концепты на которые эта ссылка указывала.

Например, такая разметка «[[Компьютерная платформа|платформа]]» рассматривается как термин «платформа», являющийся ссылкой на концепт «Компьютерная платформа». Проанализировав все случаи употребления слова «платформа» мы сможем оценить его вероятность быть ссылкой как отношение количества раз, когда это слово являлось заголовком ссылки к общему количеству употреблений. А проанализировав все случаи, когда слово «платформа» являлось ссылкой, мы сможем оценить общеупотребимость каждого концепта, подсчитав сколько раз слово платформа являлось ссылкой на концепт «Компьютерная платформа», а сколько раз на концепт «Нефтяная платформа» и т. д.

Для других источников знаний, содержащих гиперссылки, эти признаки вычисляются аналогичным образом. В случае использования инструмента для автоматического расширения базы знаний на основе текстовых документов, эти признаки вычисляются из внутренних параметров алгоритмов, основываясь на предположении, что коллекция документов содержит только тексты из определенной предметной области.

Два других признака позволяют учесть специфику конкретного текста. Для этого в документе находятся все однозначные термины (для них исключена ошибка разрешения лексической многозначности) и соответствующие

им концепты объединяются вместе. Признак "качество контекста" вычисляется для всего документа целиком как взвешенная средняя семантическая близость между концептами (sim) контекста и вероятностью быть ссылкой у соответствующих им терминов (informativeness).

$$\frac{1}{3}\sum_{i}\left(2 * sim(c_i, other) + informativeness(t_i)\right)$$

где $sim(c_i, other) = \sum_{j:\forall j, i \neq j} sim\left(c_i, c_j\right)$, $t_i$ — i-ый однозначный термин, $c_i$ — его значение.

Признак «семантическая близость к контексту» вычисляется с помощью базы знаний как сумма семантических близостей между концептом кандидатом и всеми концептами контекста.

Используемый классификатор для каждого концепта возвращает степень уверенности в том, что данный концепт является значением термина. Для каждого термина выбирается концепт, для которого степень уверенности максимальна. Кроме того, когда требуется повысить точность результатов за счет полноты, мы дополнительно используем отсечение по порогу, чтобы не возвращать часть неверных значений терминов. Оптимальный порог заранее подбирается на настроечной (скрытой) коллекции документов таким образом, что на данной коллекции достигается оптимальное значение F-меры с коэффициентом бета равным 0.3, позволяющим отдать приоритет точности.

На вход алгоритму выделения ключевых концептов приходят все концепты, выбранные в результате разрешения лексической многозначности. Для каждого концепта определяется вес: произведение среднего числа слов в соответствующих концепту терминах документа на их количество. Концепты с наибольшим весом считаются ключевыми. Выбор количества ключевых терминов или порога, выше которого термины считаются ключевыми, зависит от задачи и предоставляется пользователю системы.

## 4. Экспериментальное тестирование

Для тестирования подзадач семантического анализа использовалось 5 англоязычных коллекций документов, состоящих из текстов различных предметных областей, первые три из которых были разработаны авторами. Первая коллекция, обозначаемая **MODIS-texts**, состоит преимущественно из технических текстов, связанных с информационными системами и обработкой данных; размер данной коллекции — 131 документ. Коллекция **BoardGames** состоит из 35 текстов, относящихся к единственной предметной области — «Настольные игры». Тексты из коллекции **Tweets** характеризуются малой длиной, обилием неформальных терминов и различной тематической направленностью; данная коллекция состоит из 100 документов. Коллекции **AQUAINT** (50 новостей различных тематик [Milne 2008]) и **Wikipedia** (100 случайно выбранных

статей ресурса Википедия) пригодны только для тестирования определения значений терминов, поскольку в них отсутствует разметка большинства терминов и ключевых концептов.

Результаты тестирования алгоритмов распознавания терминов, определения их значений и извлечения ключевых концептов представлены в таблицах 1 и 2. Стоит отметить, что определение значений терминов тестируется только для корректно определенных терминов.

**Таблица 1.** Результаты тестирования методов для английского языка

| | **Распознавание терминов** | | | **Определение значений терминов** | **Извлечение ключевых концептов (топ-5 наиболее вероятных)** | | |
|---|---|---|---|---|---|---|---|
| | **Точность** | **Полнота** | **$F_1$-мера** | **Достоверность** | **Точность** | **Полнота** | **$F_1$-мера** |
| MODIS-texts | 58% | 77% | 67% | 76% | 31% | 39% | 34% |
| Board Games | 66% | 78% | 71% | 62% | 27% | 19% | 22% |
| Tweets | 48% | 72% | 85% | 73% | 42% | 31% | 35% |
| AQUAINT | — | — | — | 87% | — | — | — |
| Wikipedia | — | — | — | 92% | — | — | — |

**Таблица 2.** Результаты тестирования определения значений терминов для английского языка с учетом порогов для достижения высокой точности

| | **Точность** | **Полнота** | **$F_1$-мера** |
|---|---|---|---|
| MODIS-texts | 94% | 64% | 76% |
| Board Games | 94% | 53% | 68% |
| Tweets | 94% | 37% | 53% |
| AQUAINT | 95% | 74% | 84% |
| Wikipedia | 94% | 44% | 60% |

Для тестирования качества обработки текстов на русском языке мы вручную разметили семь новостных статей, содержащих в общей сложности 555 терминов. Кроме того мы дополнительно разметили ключевыми концептами еще 25 новостных статей, получив в общей сложности 196 ключевых концептов.

Этих данных достаточно для оценки качества работы алгоритмов системы Texterra, дальнейшее увеличение тренировочной выборки не приводит к существенным изменениям качества работы. Результаты тестирования представлены в Таблицах 3, 4 и 5.

**Таблица 3.** Результаты тестирование метода распознавания терминов для русского языка

| Распознавание терминов | | | Определение значений терминов |
|---|---|---|---|
| **Точность** | **Полнота** | **$F_1$-мера** | **Достоверность** |
| 70% | 83% | 76% | 88% |

**Таблица 4.** Результаты тестирования определения значений терминов для русского языка с учетом порогов для достижения высокой точности

| Точность | Полнота | $F_1$-мера |
|---|---|---|
| 91% | 46% | 61% |

Для задачи выделения ключевых концептов тесты проводились в зависимости от настройки количества возвращаемых ключевых концептов.

**Таблица 5.** Результаты тестирования метода выделения ключевых концептов для русского языка

| | 1 | 3 | 5 | 6 | 7 | 9 | 15 |
|---|---|---|---|---|---|---|---|
| Точность | 75,0% | 61,5% | 45,6% | 43,8% | 39,7% | 34,4% | 23,9% |
| Полнота | 12,2% | 30,1% | 37,2% | 42,9% | 45,4% | 50,5% | 58,2% |
| $F_1$-мера | 21,1% | 40,4% | 41,0% | 43,3% | 42,4% | 40,9% | 33,9% |

Для тестирования скорости работы использовался коллекция **MODIS-texts**, состоящая из 131-го текстового документа на английском языке (суммарный объём — 190КБ; ~242 слова на документ). Для обеспечения нагрузки систем использовался пул из 10-ти параллельно работающих потоков. Результаты представлены в таблице 6.

**Таблица 6.** Сравнительное тестирование скорости работы системы Texterra и DBpedia Spotlight для задач определения терминов и их значений

| Решаемая задача | Система | КБ/с | Слов/с | Терминов/с |
|---|---|---|---|---|
| Определение терминов | Texterra | 94 | 15 722 | 2 472 |
| | DBpedia Spotlight | 34 | 5 679 | 327 |
| Определение значений терминов | Texterra | 82 | 13 684 | 1 521 |
| | DBpedia Spotlight | 35 | 5 824 | 333 |

Как видно из представленных рисунков, скорость работы системы Texterra в несколько раз превышает скорость работы аналогичной системы DBpedia Spotlight [Mendes 2011]. Кроме того, можно заметить, что скорость работы системы Texterra при определении терминов выше, чем при определении

их значений — это объясняется тем, что первая задача в системе Texterra является составной частью второй.

## 5. Обзор области

Процесс автоматического связывания терминов, встретившихся в тексте, с концептами Википедии, описывающими данные термины, в литературе имеет название «викификация». Викификация является разновидностью задачи разрешения лексической многозначности, в которой в качестве источника знаний выступают статьи Википедии.

В работе Милна [Milne 2008] описывается один из первых значимых подходов к викификации, эффективно использующий информацию о частоте употребления концептов на страницах Википедии и ссылочных связях между ними. Для выбора подходящих концептов используется классификатор, основанный на распределении концептов для каждого термина, семантической близости концептов к окружающему контексту и показателе качества данного контекста. В роли контекста используется набор однозначных терминов, встретившихся в тексте.

Работа Ратинова [Ratinov 2011] разбивает задачу викификации на два последовательных этапа: локальный и глобальный. На первом шаге каждый термин викифицируется независимо от других терминов; для этого вычисляются меры близости между актуальным контекстом и текстовым описанием концепта на соответствующей странице Википедии. На втором этапе полученные концепты используются в качестве контекста, и процедура викификации повторяется: для каждого термина концепт выбирается на основе семантической близости к вычисленному контексту, что позволяется получить более точные результаты по сравнению с локальным этапом.

Задача викификации также может решаться путем анализа статистики совместной встречаемости концептов. Так, в работе [Cai 2013] процесс викификации представлен как поиск наиболее вероятной последовательности концептов, вычисляемой путем максимизации встречаемости каждой возможной пары концептов данной последовательности. Также в данной работе предпринята попытка автоматического улучшения разметки Википедии с целью уточнения информации о связях между концептами.

Принципы работы системы Texterra схожи с идеями, предложенными в работе Милна, но имеют и ряд важных отличий. В частности, для вычисления семантической близости концептов используется подход, предложенный в работе [Turdakov 2008], учитывающий типы ссылок. Это также позволяет использовать более сложные типы связей между концептам. К другим отличиям относятся способ извлечения терминов и метод определения значений терминов, включая алгоритмы классификации и повышения точности результатов. Кроме того, насколько известно авторам, на момент написания статьи система Texterra являлась единственной системой, использующей данный подход для обработки русскоязычных текстов.

## 6. Заключение

В рамках проекта Texterra была создана технология, позволяющая решать широкий класс задач, связанных с обработкой текстовых данных. В отличие от многих существующих систем обработки текстов, Texterra предоставляет возможность перехода от работы с отдельными словами и терминами к работе с их значениям, что позволяет увеличить точность решения многих прикладных задач [Turdakov 2014]. При этом особое внимание при разработке технологии уделялось производительности системы — на данный момент Texterra является одним из самых быстрых решений в данной области.

Важным преимуществом технологии Texterra являются низкие затраты на внедрение и поддержание системы за счет автоматизации процесса построения и обновления базы знаний. В качестве основной базы знаний используется информация, автоматически извлекаемая из Википедии. Далее эта база знаний расширяется информацией из других Веб-ресурсов и за счет анализа текстовых документов. Такой подход позволяет применять разработанные методы не только к заранее определенной предметной области, но и быстро адаптировать технологию к новым задачам и языкам.

Texterra может применяться для решения различных задач, требующих обработки текстов. Например, использование Texterra позволяет перейти от классического информационного поиска по ключевым словам к семантическому поиску по значениям слов. Кроме того, наличие базы знаний, позволяющей оценивать близость между понятиями, помогает решать и другие задачи из областей информационного поиска и анализа данных, включая: расширение запросов с целью увеличения полноты поиска; построение фасетных поисковых интерфейсов[Grineva 2011]; создание рекомендательных систем на основе сравнения описаний рекомендуемых объектов; анализ текстовых сообщений пользователей социальных сетей и форумов, например, с целью определение значений неизвестных демографических атрибутов пользователей [Korshunov 2013]; разработку вопросно-ответных систем, систем автоматического реферирования, диалоговых систем и др.

## Литература

1. *Astrakhantsev N. A., Turdakov D. Yu.* (2013), Automatic construction and enrichment of informal ontologies: A survey. Programming and Computer Software 39(1): 34–42
2. *Biemann C.* (2005), “Ontology Learning from Text: A Survey of Methods”, LDV-Forum, vol. 20, pp. 75–93.
3. *Cai Z., Zhao K., Zhu K. Q., and Wang H.* (2013), Wikification via link co-occurrence. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13).
4. *Grineva M., Grinev M., Lizorkin D., Boldakov A., Turdakov D., Sysoev A., Kiyko A.* (2011), Blognoon: Exploring a Topic in the Blogosphere. In proceedings of the 20th International conference companion on World wide web, Hyderabad, India, pp. 213–216

5. *Korshunov A.* (2013), Problems and methods for attribute detection of social network users [Zadachi i metody opredelenja atributov pol'zovateley social'nyh setej]. In proceedings of "Fifteenth conference on Digital Libraries: Advanced Methods and Technologies".
6. *Mendes P. N., Jakob M., García-Silva A., Bizer C.* (2011), DBpedia Spotlight: Shedding Light on the Web of Documents. In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011). Graz, Austria, September 2011.
7. *Milne D. and Witten I. H.* (2008), Learning to link with Wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08). ACM, New York, NY, USA
8. *Ratinov L., Roth D., Downey D., Anderson M.* (2011). Local and Global Algorithms for Disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — Volume 1 (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 1375–1384.
9. *Turdakov D., Velikhov P.* (2008), Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In proceedings of the Fifth Spring Young Researchers Colloquium on Databases and Information Systems, SYRCoDIS'2008
10. *Turdakov D., Astrakhantsev N., Nedumov Y., Sysoev A., Andrianov I., Mayorov V., Fedorenko D., Korshunov A., Kuznetsov S.* (2014). Texterra: A Framework for Text Analysis [Texterra: infrastruktura dlja analiza tekstov]. In proceedings of the Institute for System Programming of RAS, volume 26, 2014, Issue 1.