OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ГИБРИДНЫЙ ПОДХОД К ВЫЯВЛЕНИЮ КОМПЛЕКСНЫХ ОБЪЕКТОВ В ОБЛАСТИ НАУЧНО-ТЕХНИЧЕСКОГО ПРОГНОЗИРОВАНИЯ: ПРИНЦИП «ЧЕРНОГО ЯЩИКА»

Ефименко И.В.

Центр информационно-аналитических систем ИСИЭЗ НИУ ВШЭ
г. Москва, Россия
iefimenko@hse.ru

В работе обсуждаются методы и алгоритмы идентификации в текстах объектов высокого уровня концептуальной сложности, значимых для области научно-технического прогнозирования. Приводятся результаты применения представленных методов, которые используются, в частности, при создании программного комплекса «Интерактивная дорожная карта с обратной связью».
Ключевые слова: дорожная карта; научно-техническое прогнозирование; технологический тренд; гибридный подход; лингво-статистический метод обработки текстов; идентификация объектов в текстах.

Библиотека БГУИР

Введение

Настоящая статья посвящена методам и алгоритмам идентификации в текстах объектов, значимых для области научно-технического прогнозирования. Представленные методы и алгоритмы применяются, в частности, при создании программного комплекса «Интерактивная дорожная карта с обратной связью».

Дорожная карта (ДК) — форма наглядного визуального представления многоуровневой системы стратегического развития предметной области в рамках единой временной шкалы, которая содержит показатели экономической эффективности перспективных технологий и продуктов, обладающих высоким потенциалом спроса и привлекательными потребительскими свойствами. ДК иллюстрирует взаимосвязи между основными технологиями, определяющими развитие предметной области, существующими и перспективными продуктами и их характеристиками, динамикой российского и мирового рынков в контексте научно-технологического развития (далее – НТР).

ДК формируется в виде набора слоев. Форма представления, состав элементов ДК отличаются в зависимости от требований заказчика и предметной области картирования. Информационное наполнение элементов и взаимосвязи между ними определяются (рассчитываются) базовыми алгоритмами расчета элементов или алгоритмами консолидации экспертных сведений. Для разработки каждой ДК привлекается от десятков до нескольких сотен экспертов предметной области.

В качестве слоев дорожной карты в общем случае выступают: 1) научно-технологическое развитие; 2) технологии; 3) продукты; 4) рынки. В некоторых случаях при разработке дорожной карты формируется дополнительный слой: 5) глобальный контекст развития.

Для всех слоев используется один и тот же горизонт планирования (от 20 лет). Как правило, на одном слое ДК размещается 5-7 элементов.

1. Принцип «черного ящика»

В целях реализации описываемых методов и алгоритмов был предложен подход, в основе которого лежит принцип «черного ящика». Суть подхода состоит в следующем.

Предметная область научно-технического прогнозирования (далее – НТП) характеризуется высокой степенью сложности – и для случаев, когда речь идет об исследовании той или иной предметной области с применением экспертных процедур, т.е. вручную и с использованием специальных знаний, и, тем более, для сценариев, основанных на применении средств автоматизации. Информационные объекты, относящиеся к сфере НТП (элементы дорожных карт, технологические тренды и др.), плохо поддаются автоматической идентификации в документах в силу:

- Высокого уровня концептуальной сложности, комплексной природы;
- Низкого уровня формализации (так, например, фактически отсутствуют однозначные и общепринятые определения ряда явлений; для значительного числа объектов в области НТП представляется затруднительным дать определения, которые можно считать формальными);
- Высокой степени зависимости от специфики конкретной предметной области (нанотехнологии, медицина, транспорт, энергетика, информационные технологии и т.п.) и, как следствие, необходимости применения специальных (профессиональных) знаний, обеспечивающих возможность корректной интерпретации информации.

При этом однотипные объекты даже в далеких друг от друга предметных областях обладают родственными чертами, позволяющими отнести соответствующие явления к одному и тому же концепту сферы НТП. Таким образом, внешнее поведение однотипных объектов для различных предметных областей существенно более универсально, чем сами объекты, и можно говорить о «признаках присутствия» объекта НТП, по которым его следует идентифицировать. Сами же объекты на этапе анализа могут оставаться для наблюдателя «черным ящиком».

Аналогичный принцип проявляется также в следующем: потребностей и проблем, на решение которых нацелен процесс НТР, в каждой предметной области существенно меньше, чем способов их решения. На глобальном уровне можно говорить о незначительном по объему перечне ценностей, на обеспечение которых направлено развитие во всех предметных областях (стабильность общества, здоровье человека и др.). Также универсальными являются параметры, изменение значений которых (в сторону повышения или понижения, в зависимости от специфики параметра) является целью разработки и внедрения новых технологий: стоимость (положительным эффектом считается снижение стоимости производства и использования), эффективность (новые технологии призваны, среди прочего, повышать эффективность чего-либо), размер (в зависимости от ситуации желательно уменьшение или увеличение размеров тех или иных объектов), безопасность (может восприниматься и как глобальная ценность, и как параметр, на увеличение значения которого должно быть направлено НТР в любой области) и т.п.

Методы выявления объектов НТП, разработанные с учетом отмеченных выше особенностей, позволяют избежать таких проблем, как существенная зависимость от предметной области и др., и сформировать универсальный подход к обработке текстов в целях информационно-аналитической поддержки процессов научно-технического прогнозирования.

В результате выполненных работ была сформирована система онтологических моделей для предметной области НТП, специфицирующая как ключевые типы объектов для предметной области дорожного картирования и НТП в целом, отношения между такими объектами и их атрибуты, так и сопутствующие явления и процессы, которые задают систему индикаторов для объектов НТП в концепции «черного ящика».

2. Общий алгоритм. Модификация статистических методов с целью разработки гибридного подхода

С учетом постоянного развития науки и техники, появления новых направлений исследований, технологий, продуктов, рынков и т.п., для обработки документов с целью выявления объектов НТП представляется целесообразным использовать статистические методы. В качестве основы могут быть использованы классические методы анализа n-грамм с применением метрик частоты (term frequency, TF) и обратной частоты (inverse document frequency, IDF). Однако опыт показывает, что для сложных задач и комплексных, плохо формализуемых предметных областей применение классических статистических методов часто приводит к означиванию ложных групп слов. Такого рода эффект не удается преодолеть ни с использованием стоп-слов (списки стоп-слов должны быть настолько большими и постоянно пополняться и модифицироваться вручную с учетом специфики коллекции – жанровой, по предметной области и т.п., – что теряется смысл автоматизации процессов анализа), ни с помощью введения метрики IDF. Цель IDF состоит в уменьшении веса широкоупотребительных слов, однако в ряде случаев речь идет о словах и словосочетаниях, не являющихся в общем смысле широкоупотребительными. Их частота в коллекции такова, что использование меры TF-IDF приводит к попаданию их в число значимых словосочетаний, однако такого рода словосочетания не являются желательными результатами. В качестве примера можно привести термины, обладающие специфичной для предметной области, но слишком широкой семантикой (например, термины типа architecture, warehouse и т.п. для предметной области «Информационные технологии»).

С целью уменьшения уровня шума целесообразно ввести в общий алгоритм анализа документов следующие этапы:

- Морфологический анализ документа. При этом обеспечивается означивание слов по частям речи и морфологическим формам;
- Синтаксический анализ. Выполняется на уровне отдельных элементов, без полномасштабного построения лингвистических деревьев. В частности, обозначаются однословные и многословные именные группы;

Библиотека БГУИР

- Выбор из документов наиболее значимых фрагментов. Настоящая процедура реализуется следующим образом. На основании индикаторов объектов НТП, заданных в онтологиях, обеспечивается выбор предложений и абзацев для последующего анализа, при этом наличие в выбранном фрагменте отдельного индикатора или их совокупности является гарантом релевантности данного фрагмента для задачи идентификации объекта НТП.

Предусматривается обработка документов на английском и на русском языках.

В результате обеспечивается:

- Анализ отдельных слов и многословных словосочетаний, гарантированно являющихся именными группами, т.е. потенциальными терминами, и отсеивание до проведения анализа биграмм и триграмм, не являющихся таковыми. Кроме того, сборка словосочетаний на основе лингвистических признаков позволяет не задавать ограничений на их количество. Таким образом, обеспечивается обработка терминов любой длины;
- Отсеивание менее значимых с точки зрения задачи выявления элементов дорожных карт и других объектов НТП частей текста и выполнение анализа на наиболее релевантных фрагментах, что позволяет включать в выдачу только словосочетания, потенциально относящиеся к наименованию объекта НТП или его атрибутам.

Для полученных на выходе словосочетаний выполняется статистический анализ. В ряде случаев, например, для идентификации технологических трендов, необходим учет временного аспекта, динамики появления терминов. В этом случае формируются отдельные коллекции, структурированные по периодам времени, например, по годам. Для каждой коллекции формируются результаты, которые затем подвергаются обработке с целью выявления динамики употребления терминов.

Таким образом, предлагаемый алгоритм сочетает в себе черты лингвистического и статистического подходов, его общая схема включает следующие шаги:

- Морфологический анализ текста. При этом для английского языка используется модуль морфологического анализа платформы с открытым кодом (open source) GATE – General Architecture for Text Engineering, – разработанной в Шеффилдском университете, Великобритания. Для русского языка используется модуль морфологического анализа, разработанный компанией Яндекс;
- Синтаксический анализ, основной целью которого является сборка именных групп. При этом используются модули GATE, доработанные в Центре информационно-аналитических систем Института статистических исследований и экономики знаний (ЦИАС ИСИЭЗ) НИУ ВШЭ;
- Выбор релевантных фрагментов документа с использованием системы индикаторов, т.е. фрагментов, с наибольшей вероятностью содержащие информацию об объектах НТП (разработка ЦИАС ИСИЭЗ с использованием редактора онтологий Protégé – для спецификации моделей, платформы GATE – для генерации специальных аннотаций, среды IntelliJIDEA – для создания системы индикаторов и правил их идентификации);
- Выбор именных групп на заданных фрагментах документа;
- Статистический анализ именных групп (разработка ЦИАС ИСИЭЗ с использованием среды IntelliJIDEA);
- Выдача результатов, обеспечивающая возможность анализа динамики (разработка ЦИАС ИСИЭЗ);
- Представление результатов для интерпретации экспертам в предметной области.

В последующих пунктах более детально описаны некоторые из представленных выше шагов алгоритма.

Сведения о реализации алгоритма представлены в статье [Хорошевский, 2013].

3. Индикаторы объектов НТП

3.1. Система эвристик

Как уже указывалось выше, цель поиска в тексте индикаторов состоит в выявлении в документах объектов НТП по «внешним проявлениям». Индикаторы позволяют также выбрать наиболее релевантные фрагменты документов коллекции, на которых затем может выполняться анализ терминов, в т.ч., статистический. Целесообразно ввести параметр веса для отдельных индикаторов и их сочетаний. Таким образом, становится возможным применение подхода на предобработанных, искусственно сформированных коллекциях документов – в зависимости от веса самого индикатора или задающего его лингвистического шаблона идентифицируются:

- Непосредственно информационные объекты из контекста индикатора;
- Наиболее релевантные фрагменты текста (предложения, абзацы).

При этом применяется система эвристик о наличии в анализируемом контексте объекта НТП или, в случае комплексного объекта, его отдельных составляющих.

Все эвристики могут быть сгруппированы по следующим уровням (рисунок 1):

- Внешний контекст (экстралингвистические признаки);
- Метаинформация на уровне коллекций документов;
- Данные о структуре документа;

Библиотека БГУИР

- Лингвистические маркеры в теле документа.

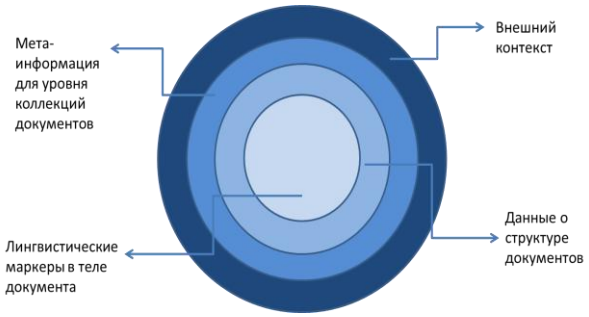

Рисунок 1 – Схема системы эвристик для выявления технологических трендов

3.2. Лингвистические маркеры в теле документа

Ключевым типом индикаторов для объектов НТП являются лингвистические маркеры в теле документа. Такого рода маркеры формируют семантические поля, каждое из которых соответствует определенному фрагменту системы онтологических моделей (например, для онтологии «Исследования и разработки» фиксируются стадии научно-исследовательского и разработческого процессов, основные типы получаемых результатов и т.п.).

Некоторые примеры семантических полей, формирующих группы индикаторов, представлены ниже.

4. Лингвистические шкалы

Одним их ключевых индикаторов релевантности фрагмента документа предметной области НТП является наличие в тексте упоминания движения того или иного параметра по лингвистической шкале, т.е. по шкале оценок типа «хорошо/плохо», «быстро/медленно», «сильно/слабо», «много/мало» и др. Это обусловлено тем, что при описании инновации, новой технологии, технологического тренда и т.п., обязательно фиксируются характеристики объектов, состояние которых изменяется в результате такого развития, указываются параметры, значение которых удалось улучшить, и т.п.

В результате анализа были выделены два основных супертипа объектов, в отношении которых предполагается проводить анализ оценочной информации в текстах, основанный на использовании лингвистических шкал:

- Потребности (человека, общества);
- Собственно объекты НТР (направления исследований и разработок, технологии, продукты, рынки и т.п.), т.е. объекты, которые создаются или трансформируются в результате НТР.

Более детально соответствующая проблематика освещена в работе [Ефименко, 2012].

Примеры, приведенные на рисунке 2, показывают, что лингвистические шкалы используются в текстах для концептов различных типов, в частности:

- Решаемые «общечеловеческие» проблемы, потребности;
- Ресурсы, ценности;
- Проблемы научного и технологического характера;
- Технологические параметры.

Все они, так или иначе, сводятся к характеристикам двух основных супертипов, представленных выше (потребности человека и общества; объекты НТР).

При этом, например, для технологических параметров выстраивается следующая модель: собственно технологические параметры (пример выделен на рисунке голубым), внешние характеристики, на достижение которых направлена технология (пример выделен зеленым), характеристики, апеллирующие к ценностям, к потребностям, ради удовлетворения которых создается технология (пример выделен красным).

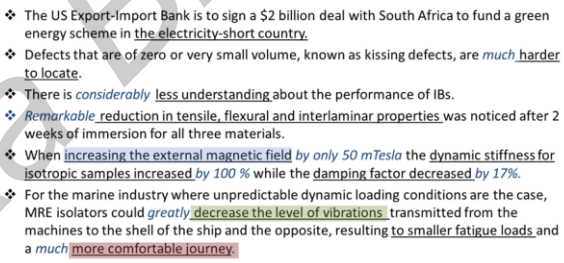
- The US Export-Import Bank is to sign a $2 billion deal with South Africa to fund a green energy scheme in the electricity-short country.
- Defects that are of zero or very small volume, known as kissing defects, are *much* harder to locate.
- There is *considerably* less understanding about the performance of IBs.
- *Remarkable* reduction in tensile, flexural and interlaminar properties was noticed after 2 weeks of immersion for all three materials.
- When increasing the external magnetic field *by only 50 mTesla* the dynamic stiffness for isotropic samples increased *by 100 %* while the damping factor decreased *by 17%*.
- For the marine industry where unpredictable dynamic loading conditions are the case, MRE isolators could *greatly* decrease the level of vibrations transmitted from the machines to the shell of the ship and the opposite, resulting to smaller fatigue loads and a *much* more comfortable journey.

Рисунок 2 – Примеры лингвистических шкал в текстах научных статей

На рисунке также выделены слова и словосочетания (синим шрифтом), задающие оценку степени изменения, которые также могут быть использованы в качестве индикаторов (в сочетании с индикаторами, фиксирующими наличие движения по лингвистической шкале, или самостоятельно).

В зависимости от параметра, значение которого изменяется по лингвистической шкале, а также от нюансов семантики самого лингвистического маркера, движение по шкале в каждом из направлений может интерпретироваться и как положительный, и как отрицательный результат (снижение стоимости vs. снижение эффективности; impairment или depletion vs. remission), при этом интерес для анализа представляет каждый из вариантов. В общем случае, в первой категории (положительный результат) имеет место фиксация нового решения – продукта, технологии и т.п. Во втором варианте (отрицательный результат) можно говорить о возникновении новой потребности или (технологической) проблемы.

Использование лингвистических шкал может обеспечить выявление технологических трендов на

Библиотека БГУИР

различной стадии их развития, в т.ч., на начальных этапах (алгоритм будет представлен в докладе).

5. Другие типы индикаторов

Другие типы индикаторов для объектов НТП, разработанные к настоящему моменту, объединены в следующие модели:

- «Исследования и разработки»;
- «Инновации»;
- «Внедрение и производство»;
- «Технологии»;
- «Лакуны в НТР»;
- «Ценности и потребности»;
- «Стандарты и регулирование»;
- «Уполномоченные органы, инстанции»;
- «Коммерческие аспекты»;
- «Мероприятия»;
- «Абстрактные объекты» (элементы процесса рассуждения, объяснения, анализа, такие как причина, следствие, цель, влияние, неизвестность, противопоставления, ограничения и др.).

Примеры фрагментов текста, обрабатываемых системой в соответствии с представленными системами индикаторов, представлены ниже (фрагменты, являющиеся основой для применения индикатора той или иной группы, выделены подчеркиванием):

- The use of adhesive bonds in engineering and marine structures is currently hindered by a lack of knowledge of joint reliability;
- Initial experimental work is presented on… Tests were performed…;
- An extensive study on the capabilities of marine animals has been conducted in relation to the equivalent functionalities in AUVs;
- The principal focus is on biological solutions to depth, speed, agility and endurance capabilities;
- PPT has shown promise as an effective near surface non-destructive evaluation (NDE) technique in a range of applications; focuses on the implementation of PPT on solid materials with artificial defects.
- Many of these have capabilities and functionality which have much in common with the engineered capabilities required for underwater vehicles e.g. propulsion/locomotion, manoeuvrability/agility and the ability & resilience to operate at depth;
- Indeed, in many examples, it appears the biological solutions exhibit superior performance compared to the technological alternative, yet by different and diverse means;
- Results from PPT are compared to other NDE methods such as water coupled ultrasound to ensure the PPT could produce results that are comparable to a more established technique;
- To be ready for production by next year, the company will offer the NM82/1500 designed specifically for low and medium wind zone onshore installations. With noise becoming an issue for turbines installed near built-up areas, and regulations already in place to limit that noise, the new design will have the ability to switch automatically to a quieter mode of operation to meet limits at certain periods of the day;
- The UK and the World can no longer afford to neglect the massive potential of wave and tidal energy, said Dr. Desmond Turner announcing the publication of a report on ocean energy prepared by the Science and Technology Committee of the British Parliament

Результаты исследования показывают, что различные типы индикаторов коррелируют с различными жанрами текстов. При этом ряд индикаторов имеют универсальный характер (например, лингвистические шкалы, ценности), другие являются жанрово зависимыми.

В дополнение к основным моделям вводятся вспомогательные, которые, в зависимости от решаемой задачи и конкретного шага алгоритма, могут быть использованы или для анализа широкого контекста объектов НТП (например, для элементов глобального контекста для дорожных карт, таких как «Барьер», «Окно возможностей» и т.п.; для нетехнологического – экономического, политического, социального – контекста технологических трендов), или в качестве источника стоп-слов и выражений (семантические стоп-поля). Примерами таких моделей являются:

- «Экономика»;
- «Политика»;
- «Общество» («Роли»);
- «Время»;
- «Единицы измерения».

Кроме того, при выявлении объектов НТП, относящихся к определенной стадии НТР, в качестве стоп-полей могут быть использованы множества концептов основных моделей, характеризующих другую стадию НТР (например, «Внедрение и производство» для «Исследований и разработок»).

Из вышесказанного следует, что удаление терминов, относящихся к стоп-полям, является опциональным.

Они могут быть также использованы позднее как фильтр для готовых результатов. При этом их включение в результаты предоставляет дополнительные возможности, в частности, их наличие:

- Потенциально свидетельствует о стадии НТР;
- Является вспомогательной информацией для развития модели, поскольку помогает решить задачу отображения индикаторов на этапы НТР (например, с использованием кривых Гартнера);
- Позволяет идентифицировать явления пограничного характера (см. выше).

Базовая версия разрабатываемого программного обеспечения основана на использовании индикаторов, являющихся общими для различных

Библиотека БГТУ им

предметных областей. При этом в качестве опциональной стадии можно ввести предобработку перечня индикаторов экспертами в предметной области, выбранной для анализа, с целью:

- Введения дополнительных, специфичных для предметной области терминов, которые могут играть роль индикаторов (включение на этом этапе более сложных конструкций, чем отдельные термины, требует лингвистической компетенции и представляется нецелесообразным; при этом допустимы многословные термины, для которых будет обеспечен морфологический анализ);
- Удаления индикаторов, являющихся общеупотребительными, но имеющих дополнительное, специфичное значение в рамках соответствующей предметной области, характеризуемых высоким уровнем многозначности.

В составе алгоритма также предусмотрены методы, обеспечивающие автоматизированный анализ синонимических рядов и родственных объектов (например, терминов, относящихся к различным явлениям, но характеризующих одну и ту же технологию).

В зависимости от особенностей коллекций могут быть применены отдельные типы индикаторов или их совокупности. При этом роль играет не только жанр, но и другие аспекты, например, объем коллекции. Так, на больших коллекциях представляется целесообразным анализ документов по отдельным группам индикаторов с последующим поиском пересечений (внутри типа и между типами).

Заключение

В качестве направлений дальнейшего исследования и разработки для развития методов и средств, представленных в статье (в т.ч., с выходом за рамки работ по созданию программного комплекса «Интерактивная дорожная карта с обратной связью »), планируются следующие:

- Анализ групп индикаторов с точки зрения их эффективности и значимости для различных жанров документов;
- Частотный анализ словосочетаний и шаблонов индикаторов применительно к различным объектам НТП, в частности, для различных слоев ДК;
- Доработка онтологических моделей, задающих семантические поля индикаторов и стоп-слов;
- Автоматизация анализа динамики изменения объектов НТП;
- Соотнесение типов индикаторов с этапами (стадиями) НТР, Рис. 3;

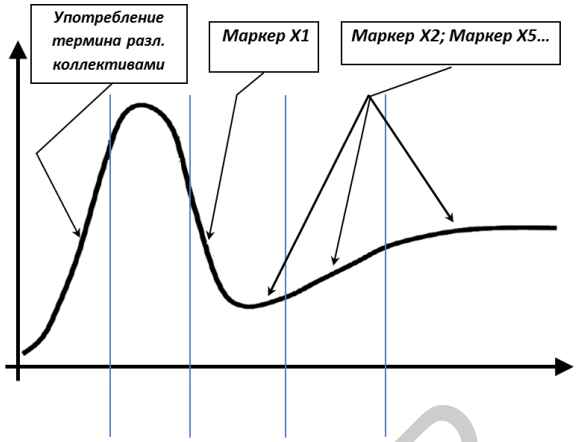

Рисунок 3 – Соотнесение индикаторов с этапами НТР (на примере кривой Гартнера)

- Уточнение формальной модели комплексных, слабо формализуемых объектов в области НТП, в т.ч., с учетом возможности формирования иерархий, декомпозиции, циклического характера развития;
- Разработка методов квантификации качественных оценок для времени, вероятности (plausibility), других типов оценок.

Благодарности. Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту от 16.05.2012 г. № 07.524.12.4018 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

Библиографический список

[Ефименко, 2012] Ефименко И.В. Модели использования и интерпретации оценочной информации в прогнозировании: время, состояние, вероятность. //Сборник трудов 13-й национальной конференции по искусственному интеллекту с международным участием. Том. 1, Белгород, 2012, С. 244-251.

[Хорошевский, 2013] Хорошевский В. Ф. Автоматизация процессов выявления технологических трендов в системе АРМ «Тренд» (OSTIS-2013).

HYBRID APPROACH FOR IDENTIFYING COMPEX CONCEPTS IN TECHNOLOGICAL FORESIGHT: BLACKBOX MODEL

Efimenko I. V.

Center for Information Intelligence Applications of Institute for Statistical Studies and Economics of Knowledge, NRU HSE
Moscow, Russia

iefimenko@hse.ru

The paper presents methods and algorithms for identifying complex concepts relevant for the domain of technological foresight within text collections. An approach based on the so called “black box” principle and combination of statistical and linguistic methods is proposed.

Библиотека БГУИР