УДК 519.25

АВТОМАТИЗИРОВАННОЕ ОЦЕНИВАНИЕ ФОРМУЛИРОВОК НАУЧНОЙ НОВИЗНЫ ПУБЛИКАЦИЙ

© **В. О. Толчеев**[1]

Статья поступила 15 ноября 2016 г.

Показаны возможности использования методов интеллектуального анализа текстовых данных (Text Mining) для задач, которые ранее решались с помощью экспертных процедур. Рассмотрены вопросы определения качества научных публикаций, выявления научной новизны и идентификации зарождающихся технологических трендов средствами Text Mining (обнаружение «слабых» сигналов). Отмечена специфика этих проблем и способов их решения. Проведен обзор используемых методов выявления научной новизны. Исследованы способы автоматизированного обнаружения слов-маркеров, характеризующих научную новизну, и на их основе составлены специализированные шаблоны. Сформированы и описаны выборки, содержащие экспертно размеченные документы (авторефераты и научные статьи). Описаны результаты экспериментов по практическому использованию полученных шаблонов (на примере публикаций в области информатики).

Ключевые слова: интеллектуальный анализ данных; выявление маркеров научной новизны; построение шаблонов; библиографические описания; классификация научных статей.

Методы интеллектуального анализа текстовых данных (Text Mining) традиционно применяются для решения широкого круга задач классификации и кластеризации документов, фильтрации спама, выявления плагиата, создания рекомендательных систем, извлечения из информационных массивов важных фактов и ключевых слов, мониторинга социальных сетей, оценки тональности высказываний [1, 2].

В последнее время прикладное использование методов Text Mining значительно расширилось и распространилось на области, в которых ранее анализ данных практически полностью осуществлялся экспертами (определение качества научных публикаций, оценка эффективности деятельности ученых и преподавателей, построение различных рейтингов научных и образовательных учреждений) [3 – 5]. В частности, средства Text Mining начали активно применять для решения такой чрезвычайно неформализованной задачи, как выявление научной новизны (НН) в журнальных публикациях. Эти исследования выполняют по нескольким направлениям.

1. Проводят работы по обнаружению в больших объемах научно-технических данных так называемых «слабых сигналов» (Weak Signals), способных диагностировать зарождающиеся научные и технологичные тренды [6, 7]. Для этого из текстов выделяют специальные маркеры, отражающие новизну, неожиданность и значимость (интенсивность) происходящих изменений. Отслеживают наличие скачкообразного увеличения встречаемости терминов и словосочетаний, которые до этого относились к низко- или среднечастотным, анализируют появление в предметной области новых слов или неожиданное употребление известных терминов вне традиционного контекста, оценивают темпы роста числа профильных публикаций с новыми и «неожиданными» терминами. Идентификацию назревающих технологических прорывов в быстроразвивающихся предметных областях (прежде всего в био-, нано- и информационных технологиях) осуществляют на основе обработки и анализа больших массивов многоязычной информации с помощью специально разрабатываемых онтологий. Необходимо отметить, что исследуемый массив должен в первую очередь включать еще неопубликованные материалы, появляющиеся в режиме on-line на сайтах электронных журналов и конференций, Web-страницах известных ученых и профессиональных сообществ, в социальных сетях, блогах, форумах. Анализ таких гетерогенных «сырых» данных и позволяет проводить упреждающее обнаружение технологических прорывов до того, как они станут очевидными трендами благодаря росту высокоцитируемых публикаций в авторитетных изданиях и окажутся в поле зрения наукометрии.

Использование методов Text Mining в форсайт-исследованиях преследует сразу несколько целей: автоматизация обработки чрезвычайно больших массивов научной информации на различных языках; снижение степени субъективности оценок; увеличение точности и достоверности обнаружения точек технологического роста; снижение затрат на организацию экспертных процедур. Вместе с тем окончательную интерпретацию и верификацию «слабых сигналов», выявленных с помощью средств Text Mining, по-прежнему проводят высококвалифицированные специалисты-предметники с использованием экспертных процедур [8].

[1] Национальный исследовательский университет «МЭИ», Москва, Россия; e-mail: tolcheevvo@mail.ru

2. В области интеллектуального анализа текстов начали активно проводить работы по изучению качества научных текстов, в частности, идентификации наличия научной новизны в журнальных публикациях и докладах на конференциях [3, 5, 9 – 11]. Эта постановка задачи существенно отличается от проблемы обнаружения «слабых сигналов». При выявлении Weak Signals интеллектуальная программа-радар должна детектировать в больших документальных массивах возможно имеющиеся (а чаще всего отсутствующие) принципиально новые знания. Фактически необходимо решить хорошо известную в Text Mining проблему обнаружения редких событий в документальном потоке. При анализе качества научных текстов, как представляется, отсутствие научной новизны в статье является редким событием, т.е. предполагается, что согласно требованиям редколлегий журналов и оргкомитетов конференций все работы, возможно, кроме обзорных, включают элементы НН.

Анализ качества научных текстов также основан на исследовании терминологического словаря публикаций. Однако акцент сделан не на оценке новизны, неожиданности и интенсивности появления слов, а на исследовании научной составляющей, заданной специальными маркерами (дескрипторами, индикаторами). В частности, к таким маркерам (словам и словосочетаниям) могут быть отнесены хорошо известные формулировки, встречающиеся, к сожалению, преимущественно в авторефератах и диссертациях, например, «предложен новый метод...», «впервые получены оценки...», «разработан подход, который, в отличие от известных, позволяет...». Такие маркеры достаточно просто выявляются средствами Text Mining. Однако очевидных признаков НН очень мало и они весьма редко встречаются в научных публикациях даже высокорейтинговых журналов. В статьях, в отличие от авторефератов, отсутствуют унифицированные требования по обязательному написанию отдельного раздела (или параграфа) «Научная новизна». Поэтому абзацы и предложения, в которых автор формулирует НН, могут располагаться в различных частях текста и формулироваться в неявном виде без употребления однозначно трактуемых маркеров. Это существенно затрудняет использование методов интеллектуального анализа данных для надежного определения наличия НН в публикациях.

Необходимо отметить, что НН может включать не только принципиально инновационные решения (для которых обосновано применять слова «новый», «впервые» и т.п.), но и основываться на известных идеях, их углублении, модификации, конкретизации. Элементами НН являются также разработка оригинальных подходов на стыке научных направлений и использование ранее созданных методов в других областях знаний. Таким образом, научная новизна является «зонтичным» термином, который включает разнородные составляющие. Это затрудняет ее идентификацию с помощью отдельных слов-маркеров и делает актуальным построение специализированных шаблонов.

Далее под шаблоном будем понимать список слов и словосочетаний, являющихся индикаторами научной новизны. Такие индикаторы могут извлекаться из текстов с помощью алгоритмов машинного обучения (Machine Learning) или задаваться специалистами-предметниками. К сожалению, при решении задач выявления НН средствами Machine Learning весьма сложно экспертно проанализировать и разметить большой массив статей для обучения. В связи с этим шаблоны, выделяемые в ходе машинного обучения на достаточно небольших выборках, чаще всего не обладают высокой точностью. Для улучшения диагностирующей способности шаблонов целесообразно привлекать экспертов, которые на основе собственных знаний и опыта добавляют информативные термины, словосочетания или исключают «слабые» дескрипторы. В данной работе построены гибридные шаблоны на основе использования алгоритмов Machine Learning и рекомендаций специалистов-предметников.

Постановка задачи и методы выявления научной новизны

Постановка задачи выявления научной новизны формулируется в рамках теории классификации [1, 12]: имеются множество документов $\{X\}$ и два класса — Q_1 и Q_2, причем Q_1 соответствует документам с НН, а Q_2 — без НН. Каждый документ представляется в виде вектора **X**, содержащего термины и словосочетания. Имеется неизвестная целевая функция (решающее правило, классификатор) J, J: $\mathbf{X} \to Q$.

На этапе обучения необходимо построить классификатор J^*, максимально близкий к J, на выбранной системе признаков (терминов и словосочетаний). Под правильным определением НН понимается отнесение документа с помощью классификатора J^* к тому же классу, на который указали эксперты (класс Q_1). При этом не исключается ситуация, что один и тот же документ **X** ($\mathbf{X} \in \{X\}$) может быть отнесен на основе своего терминологического состава сразу к двум классам (Q_1 и Q_2) одновременно.

Для представления текста используют частично структурированную модель, которая описывает документ в виде вектора [13, 14]:

$$\mathbf{X}_j = [x_j^{(1)}, \ldots, x_j^{(i)}, \ldots, x_j^{(M)}]^{\mathrm{T}}, \qquad (1)$$

где $x_j^{(i)}$ — вес терминов или (словосочетаний), являющихся словами-маркерами НН ($j = 1, \ldots, N$, N — количество документов в выборке; $i = 1, \ldots, M$, M — количество слов-маркеров). Формирование множества маркеров проводят на основе машинного обучения с последующим экспертным уточнением.

Словосочетания выявляют с помощью расчета частот совместного появления терминов [13]:

$$\omega = \frac{\omega_{kj}}{\omega_k \omega_j},$$

где ω_{kj} — совместная встречаемость терминов k и j в исследуемой выборке документов, ω_k и ω_j — частота терминов k и j в выборке, ω — вес словосочетания.

При расчете ω_k не обязательно, чтобы оба термина следовали друг за другом, их могут разделять случайные или незначащие термины. В таких случаях вес словосочетания ω корректируется по формуле

$$\omega = \frac{1}{2^t} \omega_k \omega_j,$$

где ω_k и ω_j — веса терминов k и j в словосочетании; t — количество незначимых слов между ними.

Особенностью научных документов является наличие библиографических описаний (БО), которые составляются в обязательном порядке для журнальных статей, докладов на конференциях, монографий, отчетов по НИОКР и т.д. БО могут быть представлены в виде кортежа $<T, A, K>$ [T — название (title), A — аннотация (abstract), K — ключевые слова (key words)]. На рисунке приведен пример БО, где подчеркиванием указаны термины и словосочетания, которые относятся к маркерам НН, выявленным на стадии обучения (в примерах приводятся только первые буквы фамилии, имени и отчества авторов).

Необходимо отметить, что в научных публикациях именно БО несут наибольшую информацию о НН, поскольку авторы включают в аннотации соответствующие формулировки. Кроме того, использование БО позволяет существенно сократить вычислительную сложность обработки научной информации. Далее под анализируемыми текстами будем понимать библиографические описания.

БО является короткой версией статьи и содержит незначительное число слов, поэтому веса маркеров НН (терминов или словосочетаний) $x_j^{(i)}$ целесообразно определять как $x_j^{(i)} = 1$ (или $x_j^{(i)} = w^{(i)}$, где $w^{(i)}$ — экспертно задаваемое значение веса дескриптора, которое отражает его ценность для выявления НН).

Для оценки качества определения научной новизны наиболее подходящим является показатель полнота – точность, хорошо известный в теории классификации и информационного поиска [13, 15].

Коэффициент полноты (Recall) характеризует долю найденных (с помощью построенного шаблона) статей с НН среди их общего количества в выборке:

$$R = \frac{a}{a + c}. \qquad (2)$$

Здесь a — количество выявленных в выборке публикаций с НН; $a + c$ — общее число статей с НН в выборке, c — количество документов с научной новизной (класс Q_1), которые не идентифицированы построенным шаблоном.

Коэффициент точности (Precision) характеризует долю публикаций с НН среди документов, в которых шаблон обнаружил научную новизну:

$$P = \frac{a}{a + b}. \qquad (3)$$

Здесь $a + b$ — общее количество документов с НН, определенных шаблоном, b — количество документов, не содержащих НН, но отнесенных к классу Q_1.

К сожалению, одновременная максимизация полноты и точности невозможна. Поэтому необходимо выбрать целевой критерий, который содержал бы требования к P и R. В нашей постановке задачи наиболее важно найти все статьи с НН. Причем максимизировать точность, на наш взгляд, более простая задача, решаемая за счет задания коротких шаблонов, состоящих из наиболее «сильных» маркеров (примеры таких маркеров были приведены ранее). Что касается повышения полноты, то для этого требуется включить в шаблон большее число маркеров, которые должны быть выбраны на этапе машинного обучения и экспертного оценивания. Вместе с тем концентрация усилий на получении высокой полноты может существенно снизить практическую ценность автоматизации процесса выявления НН и заставить потенциальных пользователей программно-алгоритмических средств (редколлегии журналов, оргкомитеты конференций, экспертные комиссии по выдаче грантов и т.п.) практически заново осуществлять их классификацию. При больших объемах текстовых данных такая перепроверка может оказаться очень дорогостоящей и трудозатратной.

Принимая во внимание вышеизложенное, сформулируем целевой критерий: добиться максимально возможной полноты при условии выполнения ограничения на минимально допустимое значение показателя точности:

$$\max\{R(\theta)\},\ P(\theta) \geq 80\ \%\ (\text{параметр } \theta \in [0; 1]). \qquad (4)$$

Как отмечалось ранее, в настоящее время активизировались работы по созданию автоматизированных процедур выявления НН. При этом применяют два основных подхода к извлечению информативных слов и словосочетаний из текстов. Эти подходы опираются на статистический и лингвистический анализ [1 – 3, 5, 10, 13, 16 – 18].

Статистический подход основан на оценке частоты встречаемости терминов (и словосочетаний) и предполагает, что наиболее важные маркеры многократно используются в документах. Часто употребляемые слова и специализированные термины рассматриваются как потенциальные кандидаты для включения в шаблон. Однако такие дескрипторы не всегда способны обеспечить приемлемое качество

решения сложной и неформализованной задачи идентификации НН.

Лингвистический подход направлен на выявление значимых словосочетаний с учетом синтаксических связей между словами. К числу таких процедур, в частности, относятся метод морфологических шаблонов и метод лексико-синтаксических шаблонов [9, 10]. В методе морфологических шаблонов проводят определение ключевых словосочетаний: П + С — согласованные прилагательное + существительное; С + $С_{род.п}$ — существительное + существительное в родительном падеже; С + $С_{тв.п}$ — существительное + существительное в творительном падеже; П + П + С — согласованные прилагательное + прилагательное + существительное; С + П + $С_{род.п}$ — существительное + согласованное прилагательное + существительное в родительном падеже; С + П + $С_{тв.п}$ — существительное + согласованное прилагательное + существительное в творительном падеже. Метод лексико-синтаксических шаблонов заключается в выделении конкретных лексем из словаря общенаучной речи, сокращений и знаков препинания, определении части речи, грамматической формы и составлении устойчивых конструкций-индикаторов НН.

Методика автоматизированного построения шаблонов выявления НН

В работе для построения шаблона использовали средства машинного обучения, экспертные оценки, а также проанализирована структура предложений, в которых формулируется НН. Необходимо отметить, что важную роль при составлении аннотации играют глаголы, которые авторы подбирают так, чтобы наилучшим образом описать полученный результат. Такие глаголы указывают на действие, направленное на объект исследования. Обычно используют безличные глаголы или глаголы 1-го лица (т.е. дается изложение материала от имени автора) [19].

Структура ключевого предложения (предложений) в аннотации, которое несет сведения о наличии элементов научной новизны, может быть представлена двумя схемами.

1. *Схема на основе безличных глаголов. Глагол действия в безличной форме* (что сделано) → *определение* (уточнение, что делается с объектом, какие свойства ему придаются) → *предмет* (объект, на который направлено действие) → *дополнительная информация* (например, причастный оборот, характеризующий объект или поясняющий, для чего надо было реализовать заявленное действие). Пример такого построения аннотации приведен на рисунке.

2. *Схема на основе глаголов в форме 1-го лица. Лицо*, осуществляющее действие (автор), → *глагол действия* → определение (уточнение, что делается с предметом, какие свойства ему придаются) → *предмет* (объект, на который направлено действие) → *дополнительная информация* (например, причастный оборот, характеризующий объект или поясняющий, для чего надо было реализовать заявленное действие). Дополнительная информация может также размещаться в начале статьи.

На наш взгляд, наиболее часто при составлении аннотаций используется первая схема, предполагающая безличное изложение информации.

В обеих схемах словами, указывающими на новизну, могут являться или глаголы, описывающие действие, или прилагательные и наречия, вводящие определения, а также существительные. Различные авторы имеют свои предпочтения для выражения НН и применяют разные лексико-семантические конструкции. Например, разработан (метод) *или* (авторы) разработали (метод), *или* разработанный (метод), *или* разработка (метода). Такая вариативность существенным образом затрудняет анализ текстов и требует построения специальных процедур для решения проблемы оценки наличия научной новизны в журнальных публикациях.

Исследования показали, что выявлять маркеры НН (термины и словосочетания) практически невозможно на основе анализа текстовых массивов, состоящих исключительно из научных статей. Из-за многообразия средств представления новых результатов ни статистический, ни лингвистический подходы не позволяют сформировать при обработке такой выборки множество индикаторов, применимых для диагностики НН. В связи с этим приято решение определять маркеры на основе анализа выборки авторефератов по специальностям 05.13.** (по техническим и физико-математическим наукам). Основное преимущество авторефератов заключается в хорошей структурированности изложения и наличия специального раздела «Научная новизна», содержащего информативные дескрипторы.

Таким образом, разрабатываемое программно-алгоритмическое обеспечение не является универсальным и зависит от предметной области (в нашем случае — это Информатика — «Computer Science»). В этой предметной области под НН обычно понимают следующие основные элементы [20]: разработка (и применение) новых (или модифицированных) критериев и математических (логических, эвристических, имитационных, экспертных, нейро-нечетких) моделей, методов и алгоритмов; применение известных методов в новых предметных областях для более эффективного решения задач, которые ранее решались другими способами; введение новых концепций, систематизаций, формулирование и анализ гипотез, доказательство теорем; создание программно-алгоритмического обеспечения.

Исходная выборка для проведения исследований состояла из 175 авторефератов. Проводили обработку и анализ разделов «Научная новизна» и выделяли маркеры, состоящие из наиболее часто встречающихся

Т. Г.А., А. А.В., Т. А.Г.

Математическое моделирование процессов формирования наноструктур легирующих примесей в базовом материале

Проведено математическое моделирование физико-химических процессов, лежащих в основе одного из сегментов технологического цикла создания новых полупроводниковых материалов для наноэлектроники. Этот этап производства – отжиг подложки базового материала (Si, Ti или Ge) в кислороде – предназначен для формирования особых наноструктур донорных (P, As или Sb) и акцепторных (B, Ga или Al) легирующих примесей, равномерно распределенных в базовом материале до начала отжига. В работе для одного из вариантов применяющихся конфигураций поверхности подложки ("траншея"), частично закрытой защитными масками, предохраняющими участки поверхности от воздействия окислителя, проведено исследование динамики роста пленки окисла и изучение перераспределения примесей вследствие физико-химического процесса сегрегации на фронте волны "окисел/материал". Получены и проанализированы распределения концентраций примесей, с образованием различных доменов, в том числе специфических наноструктур – узколокализованных зон (размерами 40-60 нм) повышенной концентрации донорных и акцепторных примесей. Подобные наноструктуры донорных и акцепторных примесей в подложке обеспечивают требуемые полупроводниковые электрофизические свойства материала.

Ключевые слова: нанотехнологии, конструирование новых материалов, математическое моделирование, окисление кристаллического кремния, сегрегация легирующих примесей.

Пример библиографического описания

терминов и словосочетаний (не более трех слов в словосочетании и не более двух слов между информативными терминами). В результате обучения на этой выборке составлен шаблон, в который было отобрано 125 терминов и словосочетаний, регулярно появлявшихся в документах обучающей выборки. При построении шаблона учитывали различные словоформы одного и того же термина (разработан — разработали — разработка — разработанный). Маркеры, найденные в ходе машинного обучения, уточняли с помощью экспертного оценивания (эксперты могли добавлять, исключать и модифицировать слова-дескрипторы). Окончательная длина шаблона после экспертизы составила 121 маркер. Эксперты также составили шаблон, в котором дескрипторам был присвоен различный вес. Все дескрипторы были разделены на три категории: сильные дискриминаторы (например, «впервые», «принципиально новый», «неизвестный ранее»); средние (например, «разработан», «предложен», «усовершенствован»); слабые (например, «проанализировано», «показано», «исследовано»). В результате обучения и экспертного оценивания было составлено два шаблона: № 1 — без учета весов маркеров и № 2 — с учетом различной дискриминирующей силы маркеров.

Построенные шаблоны были применены к выборке из научных статей, полученных из двух цифровых библиотек — eLibrary (http://elibrary.ru) и КиберЛенинка (http://cyberleninka.ru) — в области Информатики. Для проведения исследований три эксперта вручную разметили 112 библиографических текстов, разнеся их по трем классам: Q_1 (статьи с НН); Q_2 (статьи без НН); Q_3 (эксперты не достигли согласованного решения — отказ от классификации). В итоге в анализируемой выборке выявлено 52 статьи с НН и 46 статей без НН (14 публикаций отнесено к классу Q_3). На основе этого результата можно сделать вывод, что отсутствие в публикациях явных (однозначно трактуемых) формулировок НН не является таким уж редким событием, как это предполагалось ранее при постановке задачи, и достаточно часто встречается по крайней мере в библиографических описаниях.

Экспериментальные исследования показали, что для статей с НН совпадают как минимум три маркера из построенных шаблонов. Чем больше таких совпадений, тем выше уверенность в том, что в публикации присутствует НН. Значения показателя полнота-точность для шаблонов оказались достаточно близкими и составили: $R = 84$, $P = 82$ (шаблон № 1) и $R = 80$, $P = 84$ (шаблон № 2). Несущественные различия можно объяснить тем, что сильные дескрипторы НН достаточно редко встречаются в научных статьях и докладах на конференциях, поэтому введение весов не позволило заметно увеличить показатель полнота – точность.

Дальнейшие модификации шаблонов приводили к нарушению целевого критерия. В частности, дообучение шаблонов с помощью Machine Learning и экс-

пертного оценивания позволили увеличить полноту за счет увеличения длины шаблона (числа маркеров), но при этом не удалось выполнить ограничение по точности ($R = 94$, $P = 68$). В свою очередь, усечение шаблона и исключение из него наименее значимых дескрипторов приводило к существенному ухудшению полноты при высокой точности ($R = 65$, $P = 89$).

На основе проведенных исследований составлена следующая методика автоматизированного построения шаблонов выявления научной новизны в публикациях:

1. Задание целевого критерия качества обнаружения НН на основе показателя полнота – точность.

2. Формирование выборки, состоящей из авторефератов по специальностям анализируемой предметной области.

3. Проведение обработки текстов (удаление стоп-слов), составление списка маркеров, упорядоченных по частоте встречаемости, учет различных словоформ одного и того же термина и разных схем представления НН в документах.

4. Экспертное доопределение информативных слов (и словосочетаний), характеризующих научную новизну в указанной предметной области.

5. Составление шаблона на основе маркеров, выявленных в пп. 3, 4.

6. Формирование выборок из цифровых библиотек eLibrary.ru и КиберЛенинка по исследуемым тематикам. Проведение экспертной разметки статей.

7. Проверка шаблона на соответствие требованиям целевого критерия.

8. В случае невыполнения целевого критерия — переход к п. 2, увеличение выборки авторефератов, повторение пп. 3 – 5 и уточнение шаблона.

В настоящее время расширяются области использования интеллектуального анализа текстовых данных. В частности, разрабатываются подходы для решения слабоструктурированных и неформализованных задач автоматизированного выявления научной новизны в публикациях (ранее подобные задачи требовали проведения дорогостоящих и трудоемких экспертных процедур). Вместе с тем полностью заменить экспертов в данной области не представляется возможным и наиболее эффективными являются гибридные методы, которые используют сильные стороны интеллектуального анализа данных и экспертных оценок. Составленные на базе этих двух подходов шаблоны имеют прикладное значение и предназначены для использования редколлегиями журналов, организаторами конференций, экспертными комиссиями по выдаче грантов для предварительного программно-алгоритмического анализа наличия НН в материалах, предоставляемых авторами.

ЛИТЕРАТУРА

1. **Маннинг К. Д., Рагхаван П., Шютце Х.** Введение в информационный поиск. — М.: Вильямс, 2014. — 528 с.
2. **Большакова Е. И., Клышинский Э. С., Ландэ Д. В. и др.** Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. — М.: МИЭМ, 2011. — 272 с.
3. **Кузнецова Ю. М., Осипов Г. С., Чудова Н. В., Швец А. В.** Автоматическое установление соответствия статей требованиям к научным публикациям / Труды ИСА РАН. 2012. Т. 62. № 3. С. 132 – 138.
4. **Орлов А. И.** Критерии выбора показателей эффективности научной деятельности / Контроллинг. 2013. № 3(49). С. 72 – 78.
5. **Герасимов С. В., Курынин Р. В., Машечкин И. В., Петровский М. И., Царёв Д. В., Шестимеров А. А.** Инструментальные средства оценки качества научно-технических документов / Труды ИСА РАН. 2013. Т. 24. С. 359 – 379.
6. **Yoon J.** Detecting weak signals for long-term business opportunities using text mining of Web news / Expert Systems with Applications. 2012. Vol. 39. P. 12543 – 12550.
7. **Микова Н., Соколова А.** Мониторинг глобальных технологических трендов: теоретические основы и лучшие практики / Форсайт. 2014. Т. 8. № 4. С. 64 – 83.
8. **Орлов А. И.** Экспертные оценки / Заводская лаборатория. Диагностика материалов. 1996. Т. 62. № 1. С. 54 – 60.
9. **Леонова Ю. В., Федотов А. М.** Извлечение знаний и фактов из текстов диссертаций и авторефератов для изучения связей научных сообществ / XV Всероссийская научная конференция RCDL. — Ярославль: ЯрГУ, 2013. С. 135 – 144.
10. **Большакова Е. И., Васильева Н. Э., Морозов С. С.** Лексикосинтаксические шаблоны для автоматического анализа научно-технических текстов / X Национальная конференция по искусственному интеллекту. Т. 2. — М.: Физматлит, 2006. С. 506 – 524.
11. **Дербенёв Н. В., Толчеев В. О.** Что можно улучшить в наукометрическом анализе — учет наличия дубликатов и заимствований в научных публикациях / Управление большими системами. 2013. № 44. С. 366 – 380.
12. **Новиков Д. А., Орлов А. И.** Математические методы классификации / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 4. С. 3 – 5.
13. **Сэлтон Г.** Автоматическая обработка, хранение и поиск информации. — М.: Советское радио, 1973. — 560 с.
14. **Толчеев В. О.** Модели и методы классификации текстовой информации / Информационные технологии. 2004. № 5. С. 6 – 14.
15. **Powers D.** Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation / J. Machine Learning Technol. 2011. Vol. 2(1). P. 37 – 63.
16. **Liakata M., Thompson P., de Waard A., Nawaz R., Maat H. P., Ananiadou S.** A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction / Proc. Workshop DSSD. 2012. P. 37 – 46.
17. **Барахнин В. Б., Ткачев Д. А.** Классификация математических документов с использованием составных ключевых терминов / Материалы Всероссийской конференции «ЗОНТ». — Новосибирск, 2009. С. 16 – 23.
18. **Швец А. В.** Экспериментальный метод автоматического определения уровня качества научных публикаций / Труды пятой международной конференции «Системный анализ и информационные технологии». Красноярск, 2013. Т. 1. С. 304 – 312.
19. **Валеева Н. Г.** Жанрово-стилистическая характеристика научных текстов. Введение в переводоведение. — М.: РУДН, 2006. — 85 с.
20. **Ярская В. Н.** Методология диссертационного исследования: Методическое пособие. — Саратов: ПМУЦ, 2002. — 189 с.

REFERENCES

1. **Manning C. D., Raghavan P., Schutze H.** Introduction to information retrieval. — Moscow: Vil'yams, 2014. — 528 p. [Russian translation].
2. **Bol'shakova E. I., Klyshinskii É. S., Landé D. V., et al.** The automatic processing of texts in natural language and computational linguistics. — Moscow: MIÉM, 2011. — 272 p. [in Russian].
3. **Kuznetsova Yu. M., Osipov G. S., Chudova N. V., Shvets A. V.** Automatic determination of compliance with the requirements of articles for scientific publications / Trudy ISA RAN. 2012. Vol. 62. N 3. P. 132 – 138 [in Russian].
4. **Orlov A. I.** The selection criteria for indicators of efficiency of scientific activities / Kontrolling. 2013. N 3(49). P. 72 – 78 [in Russian].
5. **Gerasimov S. V., Kurynin R. V., Mashechkin I. V., Petrovskii M. I., Tsarev D. V., Shestimerov A. A.** Tools assessing the quality of scientific-technical documents / Trudy ISA RAN. 2013. Vol. 24. P. 359 – 379 [in Russian].

6. **Yoon J.** Detecting weak signals for long-term business opportunities using text mining of Web news / Expert Systems with Applications. 2012. Vol. 39. P. 12543 – 12550.
7. **Mikova N., Sokolova A.** Monitoring of global technology trends: theoretical foundations and best practices / Forsait. 2014. Vol. 8. N 4. P. 64 – 83 [in Russian].
8. **Orlov A. I.** Expert analysis (conclusions) / Zavod. Lab. Diagn. Mater. 1996. Vol. 62. N 1. P. 54 – 60 [in Russian].
9. **Leonova Yu. V., Fedotov A. M.** Extracting knowledge and facts from texts of dissertations and abstracts to examine the relationships of the scientific communities / XV All-Russian Scientific Conference RCDL. — Yaroslavl': YarGU, 2013. P. 135 – 144 [in Russian]
10. **Bol'shakova E. I., Vasil'eva N. É., Morozov S. S.** Lexico-syntax patterns for the automatic analysis of scientific and technical texts / X National Conference on Artificial Intelligence. Vol. 2. — Moscow: Fizmatlit, 2006. P. 506 – 524 [in Russian].
11. **Derbenev N. V., Tolcheev V. O.** What can be improved in the scientometric analysis — accounting for the presence of duplicates and borrowing in scientific publications / Upravl. Bol'sh. Sist. 2013. N 44. P. 366 – 380 [in Russian].
12. **Novikov D. A., Orlov A. I.** Mathematical methods of classification / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 4. P. 3 – 5 [in Russian].
13. **Salton G.** Automatic information organization and retrieval. — Moscow: Sovetskoe radio, 1973. — 560 p. [Russian translation].
14. **Tolcheev V. O.** Models and methods of classification of text information / Inform. Tekhnol. 2004. N 5. P. 6 – 14 [in Russian].
15. **Powers D.** Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation / J. Machine Learning Technol. 2011. Vol. 2(1). P. 37 – 63.
16. **Liakata M., Thompson P., de Waard A., Nawaz R., Maat H. P., Ananiadou S.** A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction / Proc. Workshop DSSD. 2012. P. 37 – 46.
17. **Barakhnin V. B., Tkachev D. A.** Classification of mathematical documents using a composite key terms / Materials of the All-Russian Conference "ZONT." — Novosibirsk, 2009. P. 16 – 23 [in Russian].
18. **Shvets A. V.** An experimental method for automatically determining the quality level of scientific publications / Proceedings of the Fifth International Conference "System Analysis and Information Technologies." Krasnoyarsk, 2013. Vol. 1. P. 304 – 312 [in Russian].
19. **Valeeva N. G.** Genre and stylistic characteristics of scientific texts. Introduction to translation studies. — Moscow: Izd. RUDN, 2006. — 85 p. [in Russian].
20. **Yarskaya V. N.** The methodology of the dissertation research. — Saratov: Izd. PMUTs, 2002. — 189 p. [in Russian].